



# Low-flow estimation beyond the mean – expectile loss and extreme gradient boosting for spatiotemporal low-flow prediction in Austria

Johannes Laimighofer<sup>1</sup>, Michael Melcher<sup>2</sup>, and Gregor Laaha<sup>1</sup>

<sup>1</sup>Department of Landscape, Spatial and Infrastructure Sciences, Institute of Statistics, University of Natural Resources and Life Sciences, Vienna, Peter-Jordan-Strasse 82/I, 1190 Vienna, Austria

<sup>2</sup>Institute of Information Management, FH JOANNEUM – University of Applied Sciences, Graz, Austria

**Correspondence:** Johannes Laimighofer (johannes.laimighofer@boku.ac.at)

Received: 8 April 2022 – Discussion started: 10 May 2022

Revised: 1 August 2022 – Accepted: 29 August 2022 – Published: 13 September 2022

**Abstract.** Accurate predictions of seasonal low flows are critical for a number of water management tasks that require inferences about water quality and the ecological status of water bodies. This paper proposes an extreme gradient tree boosting model (XGBoost) for predicting monthly low flow in ungauged catchments. Particular emphasis is placed on the lowest values (in the magnitude of annual low flows and below) by implementing the expectile loss function to the XGBoost model. For this purpose, we test expectile loss functions based on decreasing expectiles (from  $\tau = 0.5$  to 0.01) that give increasing weight to lower values. These are compared to common loss functions such as mean and median absolute loss. Model optimization and evaluation are conducted using a nested cross-validation (CV) approach that includes recursive feature elimination (RFE) to promote parsimonious models. The methods are tested on a comprehensive dataset of 260 stream gauges in Austria, covering a wide range of low-flow regimes. Our results demonstrate that the expectile loss function can yield high prediction accuracy, but the performance drops sharply for low expectile models. With a median  $R^2$  of 0.67, the 0.5 expectile yields the best-performing model. The 0.3 and 0.2 perform slightly worse, but still outperform the common median and mean absolute loss functions. All expectile models include some stations with moderate and poor performance that can be attributed to some systematic error, while the seasonal and annual variability is well covered by the models. Results for the prediction of low extremes show an increasing performance in terms of  $R^2$  for smaller expectiles (0.01, 0.025, 0.05), though leading to the disadvantage of classifying too many extremes for each station. We found that the application of different expectiles

leads to a trade-off between overall performance, prediction performance for extremes, and misclassification of extreme low-flow events. Our results show that the 0.1 or 0.2 expectiles perform best with respect to all three criteria. The resulting extreme gradient tree boosting model covers seasonal and annual variability nicely and provides a viable approach for spatiotemporal modeling of a range of hydrological variables representing average conditions and extreme events.

## 1 Introduction

Prediction of low flow in ungauged basins is a basic requirement for many water management tasks (Smakhtin, 2001). Current estimation procedures aim to estimate some long-term average low-flow characteristics, such as the low-flow quantile  $Q_{95}$  or the 7 d minimum flow, often calculated for a return period of 10 years ( $Q_{7,10}$ ) (Salinas et al., 2013). These signatures are either predicted by physically based models (Euser et al., 2013) or statistical methods, such as geostatistical methods (e.g., Castiglioni et al., 2009, 2011; Laaha et al., 2014) and regression-based models (e.g., Laaha and Blöschl, 2006, 2007; Tyrallis et al., 2021b; Worland et al., 2018; Ferreira et al., 2021; Laimighofer et al., 2022). Recently, the prediction of seasonal low-flow characteristics has gained increasing interest. Knowing the seasonal (e.g., monthly) distribution of low flows is necessary, e.g., when assessing the water quality or ecological status of water bodies, since low discharges combined with temperature can yield a cascade of hydrochemical processes that vary with the season. Such temporal low-flow characteristics require a

new class of models that take temporal signals of predictors into account. Assessment of low flow on a temporal scale is mostly based on empirical characteristics of the modeled hydrograph (e.g., Shrestha et al., 2014; Huang et al., 2017; Lees et al., 2021), with some exceptions where the accuracy of observations below a specific threshold are considered (e.g., Onyutha, 2016). All these approaches show a much higher bias for low flows than for high flows. This is often a consequence of a loss function that emphasizes high flows while giving too little weight to the low-flow events (Staudinger and Seibert, 2014; Staudinger et al., 2011). Although several approaches for modeling monthly or annual streamflow records exist (e.g., Vandewiele and Elias, 1995; Steinschneider et al., 2015; Yang et al., 2017; Ossandón et al., 2022; Pumo et al., 2016; Vicente-Guillén et al., 2012; Cutore et al., 2007; Lima and Lall, 2010; Roksvåg et al., 2020), we noticed a significant research gap in modeling monthly low flow, which to our knowledge has not been investigated to date.

Recently, data-driven models have gained interest for prediction of daily discharge in ungauged basins because of their fast implementation and good prediction performance. These approaches consider a wide range of models, e.g., long short-term memories (LSTMs; Kratzert et al., 2019a, b; Lees et al., 2021) or artificial neural networks (ANNs; Solomatine and Ostfeld, 2008; Dawson and Wilby, 2001; Abrahart et al., 2012). Similar methods are applied to lower temporal resolutions (monthly or annual) of streamflow data, where either parameters of hydrological models are interpolated in space (Yang et al., 2017; Vandewiele and Elias, 1995; Steinschneider et al., 2015), or different statistical methods are applied. Considering the data-driven models, a common approach is to fit independent models to each station (e.g., Shortridge et al., 2016; Parisouj et al., 2020; Chang and Chen, 2018). Such an approach does not seem efficient because spatial correlations of time series at neighboring stations are not considered in parameter estimation. This may lead to spatially inconsistent predictions that also have lower accuracy. Few approaches have been proposed that treat spatiotemporal flow indices in a single, spatiotemporal framework (e.g., Ossandón et al., 2022; Vicente-Guillén et al., 2012; Lima and Lall, 2010; Roksvåg et al., 2020; Pumo et al., 2016; Cutore et al., 2007). These studies either used a combination of deterministic models with kriging (Vicente-Guillén et al., 2012), time-series models (Pumo et al., 2016), or Bayesian hierarchical models (Ossandón et al., 2022; Lima and Lall, 2010; Roksvåg et al., 2020). Surprisingly, little efforts have been made to use statistical learning models for spatiotemporal flow patterns. One exception is Cutore et al. (2007), who tested an ANN model for mean monthly flow and compared it against various regression approaches. They found that a single ANN outperforms methods where regression parameters are interpolated in space, or single multivariable regression.

In this study, we propose to use the extreme gradient boosting model (XGBoost; Chen and He, 2015; Chen and Guestrin, 2016) for modeling spatiotemporal patterns of monthly low flows. The XGBoost model is an ensemble of boosted regression trees and a common model in hydrology (Zounemat-Kermani et al., 2021). Applications range from e.g., modeling water quality (e.g., Lu and Ma, 2020) to estimating groundwater salinity (Sahour et al., 2020). Additionally, XGBoost has shown to be a suitable model for streamflow forecasting (e.g., Yu et al., 2020; Tyralis et al., 2021a; Ni et al., 2020) and its fast implementation is beneficial in our spatiotemporal context. We further explore the use of the expectile loss function as a fitting criterion to give more weights on more extreme flows. Expectile regression (Aigner et al., 1976; Newey and Powell, 1987; Kneib, 2013; Kneib et al., 2021) has rarely been applied in hydrology (Tyralis et al., 2022), but the use of asymmetric weights appear well-suited for estimating flow quantiles beyond the mean (e.g., Toth, 2016). Our model shall also incorporate a variable selection procedure to obtain models that are more parsimonious and easier to interpret.

The objective of our study is to develop such a spatiotemporal low-flow model and to evaluate its performance when predicting at ungauged sites. The following research questions will be addressed: (i) To what extent can spatiotemporal monthly low flow be modeled by one single gradient boosting model? (ii) How does expectile regression perform compared to traditional loss functions (mean absolute error, median absolute error)? (iii) How accurate are low extremes modeled by different expectiles? (iv) Which spatial and spatiotemporal variables are used for different expectiles? Our analysis will be performed on a comprehensive Austrian streamflow dataset representing a range of seasonal low-flow regimes.

## 2 Data and methods

### 2.1 Data

#### 2.1.1 Hydrological data

Our study area covers 260 gauging stations in Austria with different low-flow seasonality. The dataset was already used in a wide range of studies (e.g., Laaha and Blöschl, 2006, 2007; Laaha et al., 2014; Laimighofer et al., 2022). Some stations included in previous studies had to be discarded as the gauging stations were removed, relocated, or the data included too many missing values. The hydrological data are available from the Hydrographic Service of Austria (HZB) and all stations have a continuous daily streamflow record from 1982 to 2018. From these data, we calculated the monthly low-flow index series for each station, using the monthly  $Q_{95}$  ( $P(Q > Q_{95}) = 0.95$ ) to characterize the low-flow regime (Fig. 1). The index is very similar to the

monthly 7 d minimum flow  $MM(7)$ , which has been used in an earlier study to assess the timing of low-flow events on a pan-European scale (Laaha et al., 2017). For one station, a monthly value had to be inserted using smoothed empirical orthogonal functions (Lindström et al., 2014) to preserve the station despite single missing values in the daily discharge record. The monthly  $Q95$  was standardized by the catchment area, resulting in the monthly specific low-flow ( $q95$ ) time series ( $1\text{ s}^{-1}\text{ km}^{-2}$ ) for each station, which constitutes the target variable of our study. The monthly  $q95$  series were further transformed by the square root to give less weight to the high low flows, according to preliminary evaluations. Finally, model evaluation was performed using the predictions after transforming back to the original scale.

Additionally, we used specific discharge quantiles with 0.95, 0.98, and 0.99 exceedance probability from the daily discharge series ( $q95_d$ ,  $q98_d$ ,  $q99_d$ ) to identify extreme events in our monthly series. This threshold selection procedure classifies about 11 % ( $q95_d$ ), 5 % ( $q98_d$ ), and 3 % ( $q99_d$ ) of observations at each station as extreme low-flow events.

### 2.1.2 Predictor variables

Spatiotemporal modeling requires predictor variables, of which two types can be distinguished. The first type consists of climate and catchment characteristics representing the long-term average hydrological conditions. This type corresponds to typical predictors in low-flow regionalization models such as regional regression approaches (Laaha et al., 2013). The second type of predictors consists of spatiotemporal covariates that capture the climate drivers of low-flow generation. These dynamic predictors are needed to extend the regional model with a temporal dimension so that spatiotemporal patterns of low flow can be represented. The former type will be referred to as static predictors, the latter as spatiotemporal covariates throughout the paper.

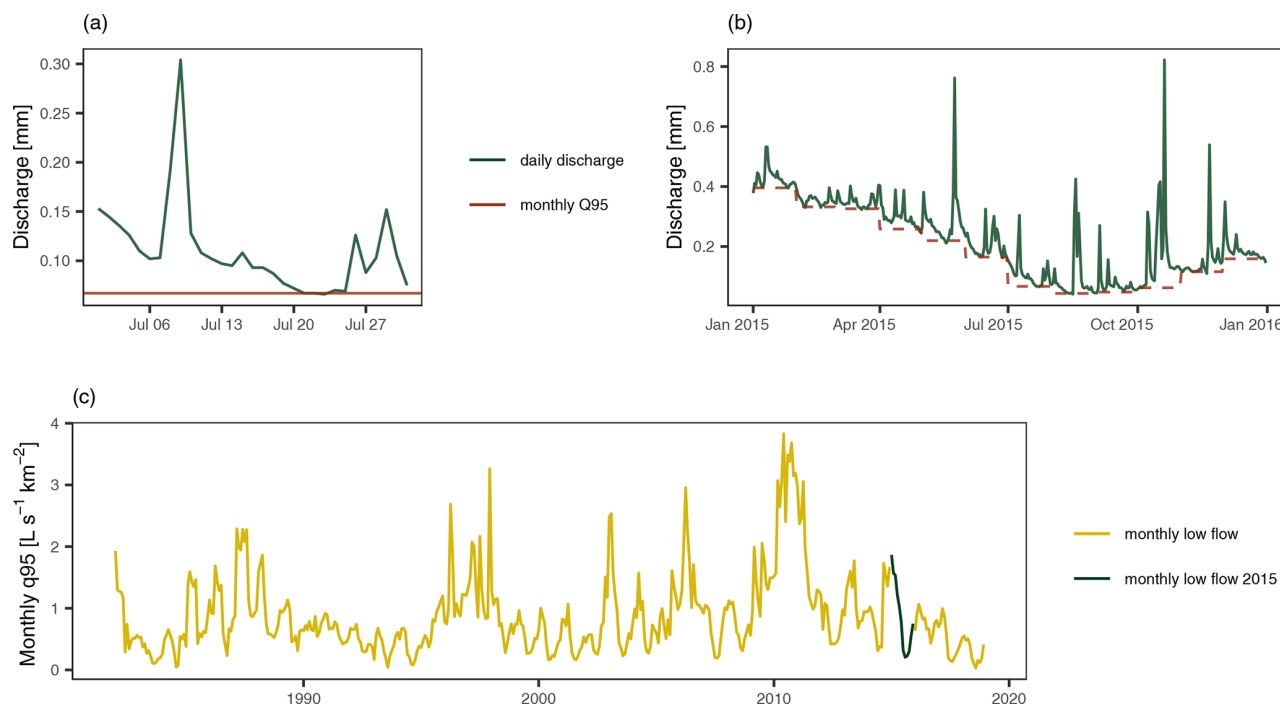
For the static predictors, we used a set of climate and catchment characteristics of precedent rationalization studies in Austria (e.g., Laaha and Blöschl, 2006; Laimighofer et al., 2022). They consist of topological and land-use variables, geological classes, and long-term average meteorological characteristics such as precipitation ( $P$ ), mean climatic water balance (MCWB), and others (Table 1). All these variables are aggregated on an annual basis (no subscript: e.g.,  $P$ , MCWB), for the summer half-year from April to September (e.g.,  $P_{\text{sum}}$ ,  $\text{MCWB}_{\text{sum}}$ ), and for the winter half-year from October to March (e.g.,  $P_{\text{win}}$ ,  $\text{MCWB}_{\text{win}}$ ). For more details about their calculation, see Laimighofer et al. (2022) and Laaha and Blöschl (2006).

For spatiotemporal covariates, the initial choice of the variables is also important, as it can affect model performance and interpretability of results. In a preliminary assessment (not shown in this paper), we tested several combinations of spatiotemporal covariates, including monthly

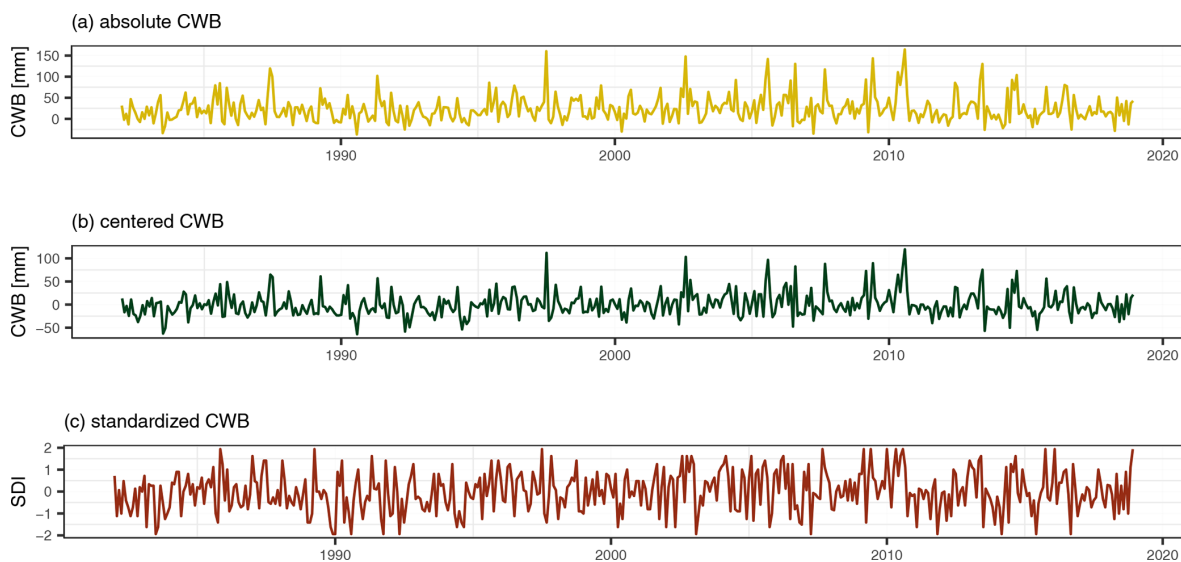
precipitation, climatic water balance (CWB), temperature, snowmelt, solid precipitation, or soil moisture characteristics. All these combinations were tested by a nested 10-fold cross validation (CV; see Sect. 2.2.3 for more detail) and compared to a range of error metrics (Sect. 2.2.4). The results showed that the monthly CWB calculated as the difference between monthly precipitation and potential evapotranspiration performs equivalently to the combination of its components provided as individual model terms. Since our study focuses more on a methodological assessment, we decided to only include the CWB as a spatiotemporal predictor for the sake of simplicity. Note that the CWB enters the model as a static variable (MCWB) and as a spatiotemporal covariate on a monthly basis (CWB). The former serves as an intercept in the model, the latter as a temporal signal that determines the monthly low-flow series.

The monthly CWB series are further coded as time lags ( $l$ ) from 1 to 12 months ( $\text{CWB}_l$ ) before each value in the target series. These lags are assumed to represent antecedent conditions in low-flow generation. In addition to using “raw” CWB values, the CWB was additionally transformed in order to test whether standardization has an effect on the performance of the predictor. First, each spatiotemporal variable ( $\text{CWB}$ ,  $\text{CWB}_l$ ) is centered by month ( $m$ ) and station ( $s$ ) via  $\text{CWB}_{\text{center},s,m} = \text{CWB}_{s,m} - \overline{\text{CWB}}_{s,m}$ , where  $\text{CWB}_{s,m}$  is the monthly climatic water balance at a station, and  $\overline{\text{CWB}}_{s,m}$  its monthly mean. Second, we transform the climatic water balance ( $\text{CWB}$ ,  $\text{CWB}_l$ ) to a non-parametric standardized drought index (SDI), similar to the SPEI (Beguería et al., 2014). Instead of fitting a parametric distribution, we estimate the empirical probability of the  $\text{CWB}_{s,m}$ . The empirical probabilities are then transformed to quantiles of a standard normal distribution. A visualization of these two transformations is given in Fig. 2 and a short overview of the variables in Table 2. These two transformations are highly correlated to the initial CWB, but can provide additional information for our model. All these lags and transformations are used simultaneously as input variables for our model, resulting in 39 spatiotemporal predictor variables.

Apart from the static predictions and the spatiotemporal covariates, some variables were added to capture the temporal periodicity. The numeric variable of the month ( $m = 1, 2, \dots, 12$ ), was converted to a categorical variable and additionally transformed to a sine and cosine curve so that their two-dimensional overlay gives a representation of the annual circle. In addition, the year was added as a numeric variable, and finally a cumulative sum of months ( $m_{\text{cum}} = 1, 2, \dots, 444$ ) was added in order to represent long-term trends of low flows. In total, this results in an initial set of 116 predictor variables that are used for the variable selection procedure.



**Figure 1.** Example of the calculation of the monthly  $q_{95}$  time series for the gauging station Hollenstein in lower Austria. Panel (a) shows the daily discharge in July 2015, (b) represents the monthly  $q_{95}$  and daily discharge for the full year of 2015, and (c) displays the full monthly  $q_{95}$  time series for the station.



**Figure 2.** Example of the transformations of the climate water balance (CWB) with no lag at station Hollenstein. Panel (a) shows the absolute values of the CWB. Panel (b) represents the CWB centered for each month and (c) is a computation of a non-parametric standardized drought index (SDI) for each month.

## 2.2 Methods

### 2.2.1 Extreme gradient tree boosting

Extreme gradient tree boosting (Chen and He, 2015; Chen and Guestrin, 2016) is a fast implementation of gradient tree

boosting (Friedman, 2001) and based on the general boosting algorithm (Friedman et al., 2000). The method is robust to overfitting and beneficial when predictors are collinear. Gradient tree boosting consists of an ensemble of additive trees that are each fitted by a greedy algorithm to minimize

**Table 1.** Descriptions of static predictors used in the study, structured in topological, land-use, geological, and meteorological characteristics. Abbreviations are further used in plots. Precipitation, climatic water balance, potential evapotranspiration, aridity index, snowmelt, and temperature variables are used on an annual and a summer/winter half-year basis. These different accumulation periods are indicated in the subscript: no subscript for annual characteristics (e.g.,  $P$ ), win for winter (e.g.,  $P_{\text{win}}$ ), sum for summer (e.g.,  $P_{\text{sum}}$ ).

Variable	Description	Unit
$A$	Catchment area	km <sup>2</sup>
Lat, Long	Latitude and longitude of gauging station	Decimal degrees
$H_+$ , $H_0$ , $H_M$ , $H_R$	Maximum, minimum, mean, and range of catchment altitude	m
$E$	Altitude of gauging station	m
$S_M$	Mean catchment slope	%
$S_{SL}$ , $S_{MO}$ , $S_{ST}$	Fraction of slight (< 5 %), moderate (5 % to 20 %), and steep slope (> 20 %) in the catchment	%
$M_S$	Major class of fraction of slope (slight, moderate, or steep)	–
$L_U$ , $L_A$ , $L_C$ , $L_F$ , $L_G$ , $L_R$ , $L_W$ , $L_{WA}$ , $L_{GL}$	Fraction of urban areas, agricultural areas, permanent crop, forest, grassland, wasteland, wetlands, water surfaces, glacier in catchment	%
$M_L$	Major land-use class in the catchment	–
$G_B$ , $G_G$ , $G_T$ , $G_F$ , $G_L$ , $G_C$ , $G_{GS}$ , $G_{GD}$ , $G_{SO}$	Fraction of Bohemian Massif, quaternary sediments, tertiary sediments, flysch, limestone, crystalline rock, shallow and deep groundwater table, source region in catchment	%
$M_G$	Major geological class in the catchment	–
$D$	Stream network density	10 <sup>2</sup> m km <sup>−2</sup>
$P$	Precipitation	mm
$ET_P$	Potential evapotranspiration	mm
$AI$	Aridity index	–
$AI_{\text{min}}$	Half year with lower AI	–
$MCWB$	Mean climatic water balance	mm
$S$	Snowmelt	mm
$T_+$ , $T_0$ , $T_M$ , $T_R$	Maximum, minimum, mean, and range of temperature	°C
$P_0$	Average number of days without precipitation (< 1 mm)	days
$P_H$	Average number of days with precipitation > 5 times the mean	days

**Table 2.** Description of the different lags and transformations for the CWB. Center means centering per station and month and the standardized drought index (SDI) is transforming the CWB to an SDI per station and month.

Variable	Description	Unit
CWB	Monthly climatic water balance	mm
$CWB_l$	Monthly climatic water balance at time lag $l$	mm
$CWB_{l,\text{center}}$	Centered $CWB_l$ with respect to station and month	mm
$CWB_{l,\text{SDI}}$	Standardized drought index of the $CWB_l$ with respect to station and month	–

a predefined loss function. Let  $\mathbf{y}$  be a vector of our response variable (monthly specific low flow  $q_{95}$ ) of length  $n = s \cdot t$ , where  $s$  is the number of stations and  $t$  the length of each individual time series per station, and index  $i$  referring to its  $i$ th element. Further, let  $\mathbf{X}$  be our predictor matrix with  $n \times p$  elements, wherein  $p$  is the number of predictor variables. We can then write the regression equation as

$$\hat{y}_i = \sum_{k=1}^K f_k(X_i), f_k \in F, \quad (1)$$

where  $f_k$  ( $k = 1, \dots, K$ ) is the ensemble of regression trees, and  $K$  the number of trees used. The regression trees are fit-

ted in an additive manner, where an objective function  $L^k$  is minimized:

$$L^k = \sum_{i=1}^n L(y_i, \hat{y}_i^{k-1} + \eta f_k(x_i)) + \Omega(f_k). \quad (2)$$

The  $k$ th iteration loss is denoted as  $L^k$ ,  $\hat{y}_i^{k-1}$  is the prediction of the regression tree in the previous iteration,  $f_k$  is the tree that most improves our model considering the predefined loss function,  $\Omega(f_k)$  is an additional penalization parameter for the complexity of the model and  $\eta$  is a shrinkage parameter. The shrinkage parameter  $\eta$  is set to 0.1, where a small value of  $\eta$  minimizes the risk of overfitting and the possibility of

finding only local minima. To reduce the computational burden, we tuned only the following hyperparameters for the final predictions: the maximum depth (the final XGBoost was optimized for a sequence from 6–8) of each additive tree, subsampling of predictor columns (0.25–1), which only uses a fraction of all predictors to search for the optimal split and subsampling of the observations (0.5–1), equivalent to random forests. Subsampling of the observations and the predictor variables is used to decorrelate the trees. The maximum depth can be described as the order of interaction used in the model. Finally, we introduced an early stopping rule for the boosting iterations, where the algorithm stops when the error does not decrease for  $K = 50$  iterations. The range of hyperparameters was set based on experience with the XGBoost model and our individual dataset. The final XGBoost model was optimized in a 10-fold CV by using all parameter combinations and tuning the number of boosting iterations (number of trees). A detailed description of our validation scheme is given in Sect. 2.2.3).

## 2.2.2 Loss function

One crucial point in our study is the application of a suitable loss function. The loss function has to be a twice differentiable convex function. Since our main aim is to model the low flows in the range of annual minima corresponding to the lower tail of the monthly  $q95$  series, we propose using the expectile loss ( $L_{EL}^\tau$ ) for model fitting. Expectile regression (Aigner et al., 1976; Newey and Powell, 1987) is a squared variant of quantile regression (Koenker and Bassett, 1978), where the absolute deviations are substituted by the squared deviations (Kneib, 2013). Expectile regression was already implemented in a boosting framework (Sobotka and Kneib, 2012) and the resulting expectile loss can be defined as follows:

$$L_{EL,i}^\tau = \begin{cases} \tau \cdot (y_i - \hat{y}_i)^2, & (y_i - \hat{y}_i) \geq 0 \\ (1 - \tau) \cdot (y_i - \hat{y}_i)^2, & (y_i - \hat{y}_i) < 0, \end{cases} \quad (3)$$

where  $L_{EL,i}^\tau$  is summed up to  $L_{EL}^\tau$  ( $L_{EL}^\tau = \sum_{i=1}^n L_{EL,i}^\tau$ ) for minimizing the error over all observations. If  $\tau = 0.5$ , the expectile loss is a scaled variant of a squared loss function ( $L_{\text{squared}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ ) resulting in a least squares regression. Figure 3 shows that altering  $\tau$  leads to an asymmetric weighting of the squared loss, where smaller  $\tau$  give more emphasis on negative residuals. Expectiles cannot directly be interpreted as a flow quantile, as this would be possible for quantile regression, but studies show that using transfer functions can do this very accurately (Waltrup et al., 2015) and differences mainly arise in the tail of the distribution. Estimating a full distribution of  $\tau$  values is not a very practical approach for our large dataset. Therefore, we will assess a sequence of  $\tau$  values ( $\tau = 0.01, 0.025, 0.05, 0.1, 0.2, 0.3$ ) and the special case of a least squares regression ( $\tau = 0.5$ ). Within this sequence, the  $\tau$  values of 0.1, 0.05,

and 0.025 give good approximations of our three thresholds:  $q95_d$ ,  $q98_d$ , and  $q99_d$ . Finally, we will compare the expectile loss function to the absolute loss (Fig. 3):

$$L_{\text{abs},i} = |y_i - \hat{y}_i|. \quad (4)$$

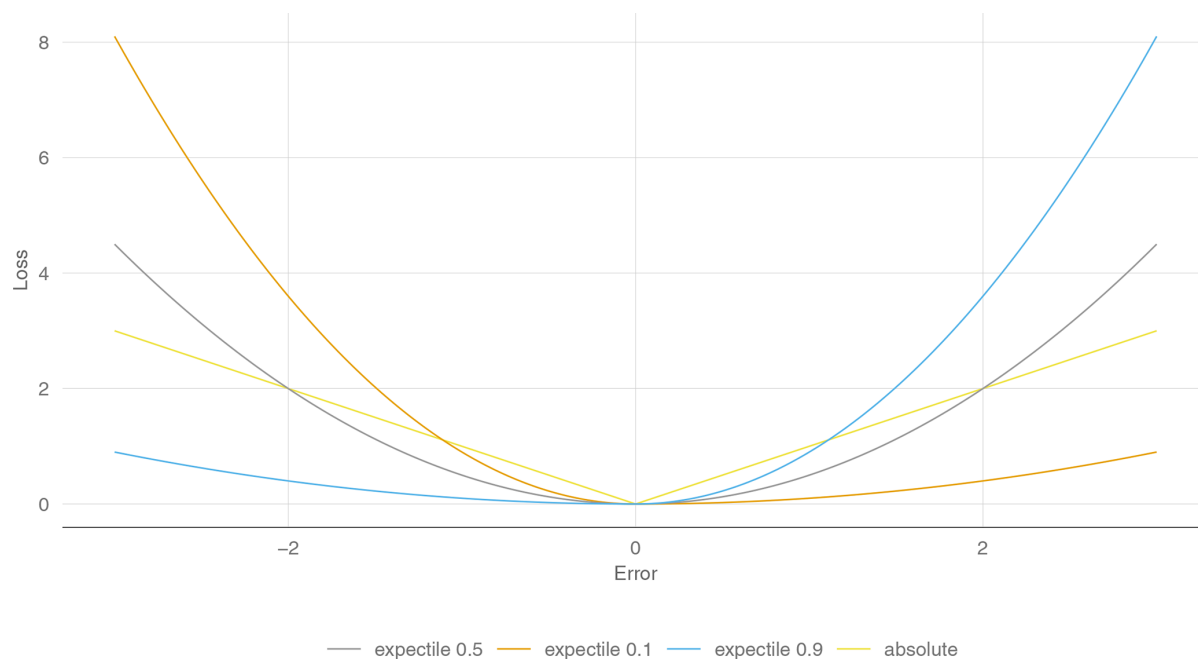
In case of the absolute loss, the mean absolute error will be minimized by the mean ( $\text{MAE} = \frac{\sum_{i=1}^n L_{\text{abs},i}}{n}$ ), and in case of the median absolute loss by the median ( $\text{MDAE} = \text{median}(L_{\text{abs},i})$ ). In both cases, the second derivative is approximated by a vector of 1 s.

## 2.2.3 Variable selection and validation scheme

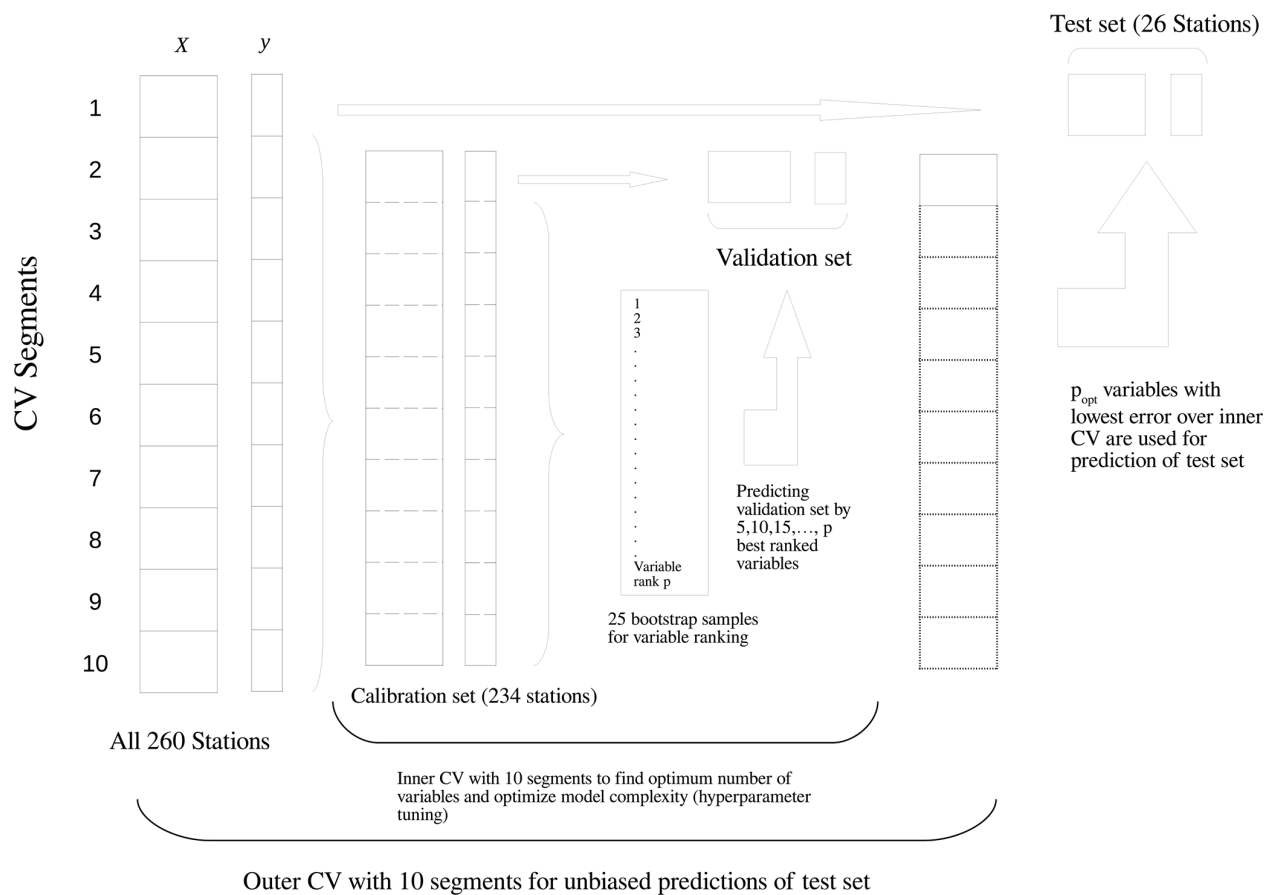
Model evaluation is performed by a nested 10-fold CV (Varma and Filzmoser, 2016). A nested CV contains two loops: an inner loop which is used for tuning of hyperparameters and variable selection and an outer loop for evaluating the predictive performance of the models (Laimighofer et al., 2022). In each inner loop, we include a variable selection by recursive feature elimination (RFE; Granitto et al., 2006). The RFE algorithm consists of an initial variable ranking and a backward variable selection. The initial variable ranking is computed by using the XGBoost algorithm with 500 boosting iterations and default hyperparameters (maximum depth = 8, subsample = 1, fraction of  $p = 0.5$ ). The variables are ranked after their additive gain in minimizing the loss function over the 500 boosting steps. For a more robust approach, the initial variable ranking is averaged over 25 bootstrap samples. The gain of each variable for the final variable ranking is the ratio of the individual additive gain to the total gain over all variables. In a next step, we are fitting an XGBoost model to a sequence from 5 to the maximum number of variables ( $p$ ), but only with a step length of 5 to reduce the computational burden. For each number of variable, the error is calculated and averaged over all 10-fold CV runs of the inner loop. The number of variables are then determined by using a threshold of 1.05 times the minimum error to produce parsimonious models. This approach yields models with fewer variables at only a small loss in predictive accuracy. Our method ensures that predictors are only selected when they increase the predictive performance of the model. In addition, the outer loop evaluates the predictive performance at ungauged sites independently of model fitting. The applied method is fully described in Laimighofer et al. (2022) and to increase clarity we added a visual explanation of the full CV procedure (Fig. 4).

## 2.2.4 Model evaluation

Model evaluation is performed by several error metrics. First, we quantify the overall performance of the model by means of four error metrics, the median absolute error (MDAE), the mean absolute error (MAE), the root mean squared er-



**Figure 3.** Comparison of different loss functions. Expectile loss functions are shown for various  $\tau$  parameters and the absolute loss function.



**Figure 4.** The nested cross-validation (CV) procedure as adapted from the double CV-scheme of Varmuza and Filzmoser (2016).

ror (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}, \quad (5)$$

and the coefficient of determination  $R^2$ , which – by our definition – is the same as the Nash–Sutcliffe efficiency (NSE) (Blöschl et al., 2013):

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \quad (6)$$

All measures are based on CV and therefore are indicators of the predictive performance when estimating at ungauged sites.

Second, we assess the predictive performance for individual stations by calculating the  $R^2$  for each station separately. The station-by-station performance is summarized by the distribution and the median  $R^2$  over all stations, which will be referred to as  $R_{\text{med}}^2$  throughout the paper.

For a more comprehensive evaluation of the performance, we perform a decomposition of the station-wise prediction errors on different time scales. The purpose is to assess the extent to which the model errors occur at the annual, seasonal, and monthly level, and which part of the error is due to a systematic error (i.e., bias). This will allow us to get insight into structural strength and weaknesses of the models. This decomposition is done by a three-way ANOVA, which was applied by e.g., Parajka et al. (2016) for assessing uncertainty contributions of climate projections. The basic linear model for our ANOVA can be defined as follows:

$$y_{m, \text{yr}} - \hat{y}_{m, \text{yr}} = \mu + \alpha_m + \beta_{\text{yr}} + \epsilon_{m, \text{yr}}, \quad (7)$$

where the left-hand side (predictand) are the prediction errors of a model at a single station. These are decomposed into the terms  $\mu$  representing the mean error (i.e., bias),  $\alpha_m$  the seasonal effects ( $m = 1, 2, \dots, 12$ ),  $\beta_{\text{yr}}$  the mean annual effects ( $\text{yr} = 1, 2, \dots, 37$ ), and the residual term  $\epsilon_{m, \text{yr}}$  corresponding to the model errors at the monthly level. Note that  $\mu$  is not a classical intercept but enters as a constant factor (coded as a vector of 1 s) which allows the bias to be considered as a separate effect in the error decomposition. In the ANOVA framework, the total variance of the prediction errors is characterized by the total sum of squares  $\text{SS}_T$ , and is decomposed into additive variance components of individual effects:

$$\text{SS}_T = \text{SS}_{\text{bias}} + \text{SS}_{\text{season}} + \text{SS}_{\text{year}} + \text{SS}_E. \quad (8)$$

The variance contributions of each term are estimated by the measure  $\omega^2$ , which is an analogue to the coefficient of determination. The measure  $\omega^2$  of, e.g., the seasonal effect is defined as:

$$\omega_{\text{season}}^2 = \frac{\text{SS}_{\text{season}} - \text{df}_{\text{season}} \cdot \text{MS}_E}{\text{SS}_T - \text{MS}_E}, \quad (9)$$

where  $\text{df}_{\text{season}}$  are the degrees of freedom (i.e., number of factor levels – 1), and  $\text{MS}_E = \text{SS}_E / \text{df}_E$  the residual mean squared error, which can be seen from the ANOVA table. The contribution of the mean annual effect ( $\omega_{\text{year}}^2$ ) and the fraction of the bias ( $\omega_{\text{bias}}^2$ ) can be determined analogously and a more detailed description can be found in Parajka et al. (2016). Finally, the residual term  $\omega_E^2$  can be calculated as follows:

$$\omega_E^2 = 1 - \omega_{\text{season}}^2 - \omega_{\text{year}}^2 - \omega_{\text{bias}}^2. \quad (10)$$

The last part of our model evaluation specifically emphasizes the extreme low flows. This is assessed by three performance metrics. First, we filter the observations at each station by a specific quantile and calculate the overall  $R^2$  based on the residuals of the filtered cases. Second, we calculate the relative expectile error ( $\text{EL}_\tau$ ) at each station:

$$\text{EL}_\tau = 1 - \frac{L_{\text{EL}}^\tau(y_i, \hat{y}_i)}{L_{\text{EL}}^\tau(y_i, q(y, \tau))}. \quad (11)$$

The error metric is defined in analogy to the expectile loss function to give more weight to the fraction of  $\tau$  lowest values. The denominator ( $L_{\text{EL}}^\tau$ ) is the naive estimate for each station. For this purpose, we are using the  $\tau$ -quantile of the observations ( $q(y, \tau)$ ) as a simple prediction at the specific station, assuming that quantiles and expectiles give similar estimates. The numerator, on the other hand, is the expectile error ( $L_{\text{EL}}^\tau$ ) for the predictions of our model. Similar to the  $R^2$ , the  $\text{EL}_\tau$  is bounded at a maximum of 1, representing perfect model fit. Values below 0 indicate that the naive prediction of the  $\tau$ -quantile would result in a better prediction than our model. One advantage of this metric is that it is based on the full data, but gives more weight to the values below our selected  $\tau$ .

Finally, we want to capture the model ability of classifying extreme low-flow events. For this purpose, the specific low-flow quantiles  $q_{95d}$ ,  $q_{98d}$ , and  $q_{99d}$  are calculated from the daily discharge records and used as thresholds for identifying drought events in the monthly low-flow series. Based on these thresholds, the drought/no drought cases of observed and predicted monthly time series are binary coded, and we calculate the hit score ( $H_S$ ) and the precision ( $P_{\text{rec}}$ ) for evaluating the performance at each station. The hit score (also termed recall, sensitivity, or true positive rate) is calculated by

$$H_S = \text{ET}_{\text{pred}} / E_{\text{obs}}, \quad (12)$$

where  $E_{\text{obs}}$  is the number of low-flow events in the original data and  $\text{ET}_{\text{pred}}$  is the number of low-flow events correctly classified by the predictions. The precision is computed as follows:

$$P_{\text{rec}} = \frac{\text{ET}_{\text{pred}}}{\text{ET}_{\text{pred}} + \text{EF}_{\text{pred}}}, \quad (13)$$

where  $\text{EF}_{\text{pred}}$  is the number of low flows falsely predicted by each monthly time series.



**Table 3.** Overview of the error metrics for all loss functions. Shown are the MDAE, MAE, RMSE, and  $R^2$ , all representing the overall predictive performance of the model for the study area.

Loss function	MDAE	MAE	RMSE	$R^2$
Absolute	2.01	3.95	8.28	0.74
Median absolute	1.99	3.95	8.42	0.73
Expectile 0.5	1.81	3.50	6.98	0.81
Expectile 0.3	1.85	3.58	7.20	0.80
Expectile 0.2	1.90	3.80	7.72	0.77
Expectile 0.1	2.09	4.26	8.77	0.70
Expectile 0.05	2.72	5.61	11.21	0.52
Expectile 0.025	3.36	6.63	12.87	0.36
Expectile 0.01	4.21	7.79	14.45	0.20

The data analysis was performed in R (R Core Team, 2021) and we want to acknowledge the use of the following packages: XGBoost (Chen et al., 2021), dplyr (Wickham et al., 2021), purrr (Henry and Wickham, 2020), tidyr (Wickham, 2021), ggplot2 (Wickham, 2016), caret (Kuhn, 2021), glmnet (Friedman et al., 2010), tidysselect (Henry and Wickham, 2021), tibble (Müller and Wickham, 2021), receipes (Kuhn and Wickham, 2021), ggthemes (Arnold, 2021), lubridate (Grolemund and Wickham, 2011), wesanderson (Ram and Wickham, 2018).

### 3 Results

#### 3.1 Global model performance

Table 3 presents the results for the overall performance of our spatiotemporal model. Generally, most loss functions show a good overall performance. The 0.5 expectile yields the best  $R^2$  of 0.81, which is substantially higher than 0.73 for the median absolute loss and 0.74 for the mean absolute loss. Regarding the MDAE, which gives more focus on low values, there is no change in the overall ranking of methods, with a MDAE of 1.81 for the 0.5 expectile, 1.99 for the median absolute loss, and 2.01 for the mean absolute loss. Regarding the results of the expectiles with  $\tau$  smaller than 0.5, we can identify a decreasing performance towards smaller expectiles over all performance metrics. The  $R^2$  is stable until the 0.3 expectile and then suddenly drops from 0.8 (0.3 expectile) to 0.2 for the 0.01 expectile. A similar loss in performance can also be observed with the MDAE, MAE, or RMSE. Expectiles below 0.05 show an insufficient performance on these global error metrics. For example, the RMSE of the 0.01 expectile is twice as high as the RMSE of the 0.5 expectile. Nevertheless, expectiles such as 0.2 and 0.3 show better overall metrics than the median absolute loss and the absolute loss, and even the 0.1 expectile demonstrates only a somewhat lower  $R^2$  of 0.7.

**Table 4.** Performance per station summarized by the median  $R^2$  over all stations ( $R^2_{\text{med}}$ ) and the fractions of stations with an  $R^2$  below 0.5.

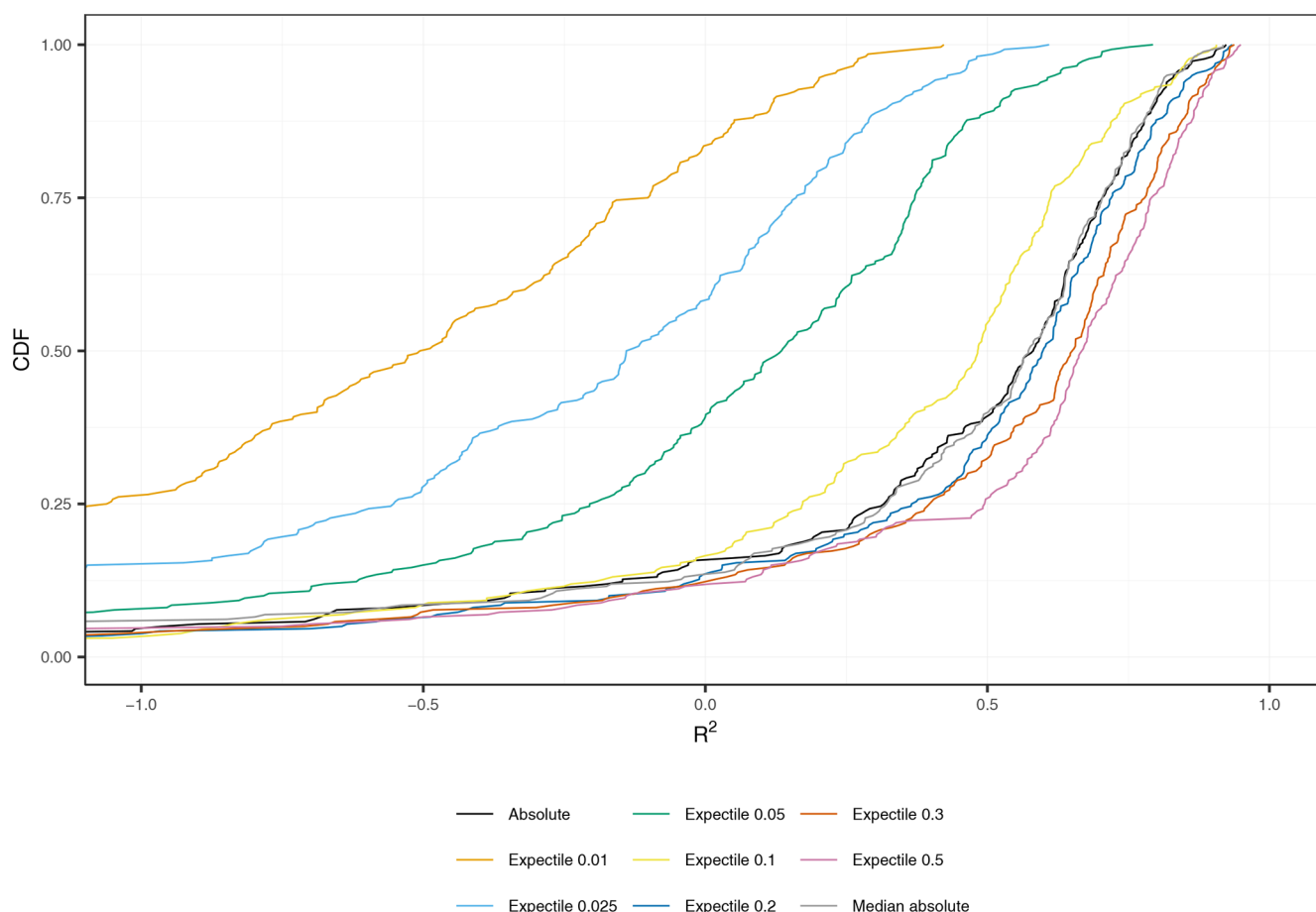
Loss function	$R^2_{\text{med}}$	$R^2 < 0.5$
Absolute	0.58	0.39
Median absolute	0.57	0.40
Expectile 0.5	0.67	0.26
Expectile 0.3	0.65	0.32
Expectile 0.2	0.60	0.36
Expectile 0.1	0.48	0.54
Expectile 0.05	0.14	0.89
Expectile 0.025	−0.13	0.98
Expectile 0.01	−0.50	1.00

#### 3.2 Station-by-station performance

A more detailed examination of our results is realized by analyzing the performance per station. Table 4 gives an overview of the results and Fig. 5 shows the empirical cumulative distribution of the station-specific  $R^2$ . From the graphs, the 0.5, 0.3, and 0.2 expectiles outperform the other models across all stations. The 0.5 expectile has the highest  $R^2_{\text{med}}$  (0.67) and shows a far better performance than the median absolute loss (0.57) and the mean absolute loss (0.58). This generally good performance is further underlined by the finding that only 26 % of the stations have a  $R^2$  below 0.5 for the 0.5 expectile. In contrast, for the mean and median absolute losses, about 40 % of the stations yield a  $R^2$  below 0.5. For the expectiles lower than 0.5, we can identify a decrease of the  $R^2_{\text{med}}$ , from 0.65 for the 0.3 expectile to 0.14 for the 0.05 expectile. Generally, low expectiles (0.01, 0.025, 0.05) yield a high number of inadequate models, where almost 90 % of the stations obtain a  $R^2$  below 0.5 for the 0.05 expectile. In case of the 0.01 expectile, no station has a  $R^2$  greater than 0.5. However, the 0.2 and 0.3 expectiles still yield a higher  $R^2_{\text{med}}$  of 0.65 and 0.6, compared to the mean absolute loss (0.58) and median absolute loss (0.57). Further, the fraction of stations with a weak performance ( $R^2 < 0.5$ ) is also lower for the 0.2 and 0.3 expectiles as shown in Table 4. These findings suggest, in response to our first research question, that a single model can provide very accurate results for most stations (0.5, 0.3, 0.2 expectiles), but 26 % to 36 % of the stations have inadequate performance. The origin of these errors will be analyzed in more depth in the subsequent section. As the results of the mean and the median absolute loss could not compete with the performance of the 0.2, 0.3, and 0.5 expectiles, we will not further include these two loss functions in our analysis.

##### 3.2.1 Error decomposition

For a better understanding of the error at the individual stations, we decompose the model error at each station into



**Figure 5.** Empirical cumulative distribution function of station-wise  $R^2$  by loss function across all stations. For improving visual clarity, the  $x$  axis is bounded at  $-1$ .

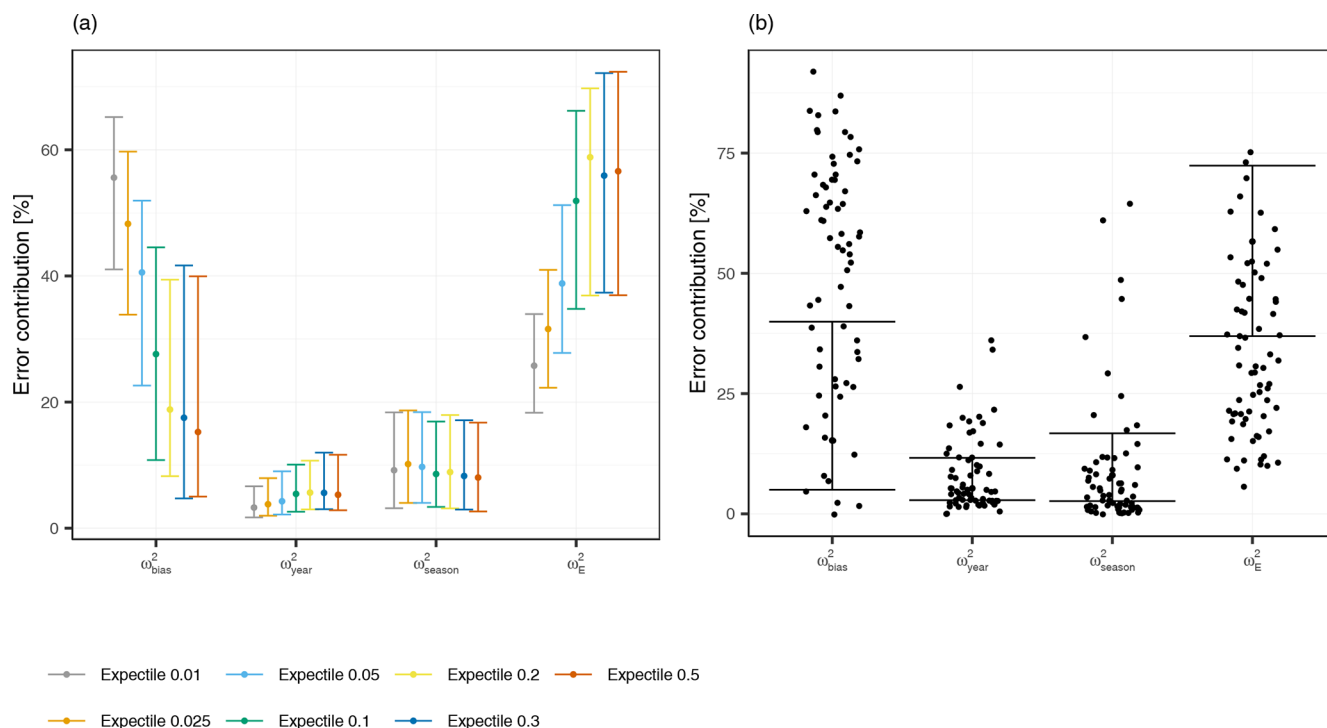
monthly, seasonal, annual fractions and a component representing the average error (prediction bias). Figure 6a gives an overview of the error components. For all expectiles,  $\omega_{\text{bias}}^2$  and  $\omega_{\text{E}}^2$  are the most important components. For the expectiles from 0.1 to 0.5, the  $\omega_{\text{E}}^2$  is the main error contribution with median values between 52 % (0.1 expectile) and 59 % (0.2 expectile). The median values of  $\omega_{\text{bias}}^2$  increases with decreasing  $\tau$ , from 15 % to 19 % for the 0.5 to 0.2 expectiles, and 28 % for the 0.1 expectile. This trend continues for the smaller expectiles where the fraction of the mean error rises up to 56 % for the 0.01 expectile. At the same time,  $\omega_{\text{E}}^2$  decreases for smaller expectiles, showing that expectiles with small  $\tau$  are better fitted to the monthly values but predictions are generally less accurate. With median values around 10 % and between 3 % and 6 %, respectively, the seasonal and annual errors are much smaller than the mean error component and show a good performance of the models in predicting annual and seasonal low-flow variability.

An additional perspective comes from analyzing the share of stations showing only moderate or weak performance, as indicated by an  $R^2$  of less than 0.5. As an example, Fig. 6b

shows such ill-performing stations for the case of the 0.5 expectile model, with 67 stations (or 26 %) having an  $R^2$  below 0.5. The main error contribution for these stations is the  $\omega_{\text{bias}}^2$  with a median value of 56 %, which is much higher than the  $\omega_{\text{bias}}^2$  of the well-performing stations (median value of 11 %). This again re-emphasizes the earlier finding that the main shortcoming of our modeling approach is an error in the mean, which reduces the predictive accuracy of the models. Seasonal and monthly variability, though, are well covered by the models, which is a strength of our spatiotemporal modeling approach.

### 3.3 Prediction of extremes

One objective of this study was the application of different expectiles for improving predictions at low extremes. This section will consequently focus on their evaluation. In a first step, we will filter our observations by only considering the observations at each station below a specific quantile. The filtered observations are then used to calculate an overall  $R^2$ , which is shown in Fig. 7. We can identify a huge performance advantage for low expectiles for predicting low extremes. For



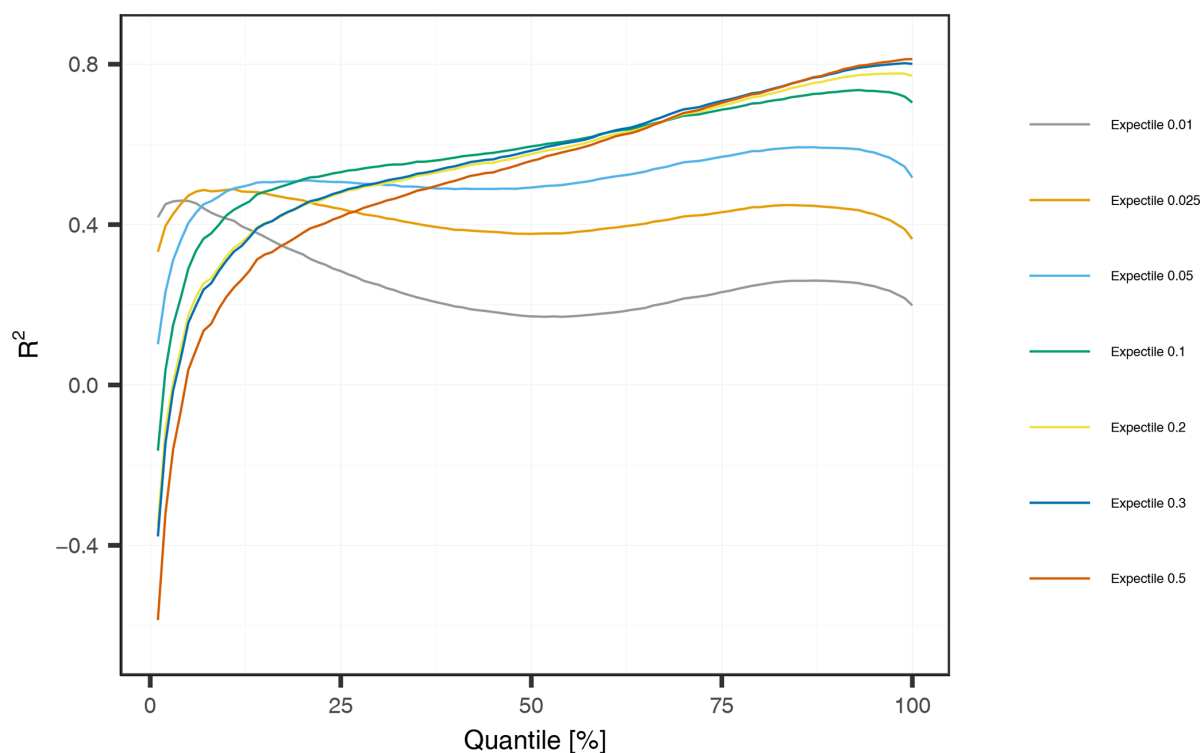
**Figure 6.** Error components of various expectile models. Panel (a) shows the relative error contributions for the season, year, month, and the bias part across all stations. The median (point) together with the 25 % quantile and the 75 % quantile (whiskers) of relative errors for each expectile loss function are shown. Panel (b) displays the error components for the 0.5 expectile in greater detail. The points are the stations which have a  $R^2$  lower than 0.5.

example, if we assess the accuracy of our models for cases below the 1 % quantile, the 0.01 expectile is yielding at least a  $R^2$  of 0.42, where larger expectiles (0.05–0.5) show inefficient models with  $R^2$  values below 0.1. What becomes apparent is that the performance is dropping at some point for the low expectiles (0.01, 0.025, 0.05), but is monotonically increasing for all other expectiles. Considering the results at the 25 % quantile, the  $R^2$  of the 0.01 expectile decreased to 0.28, so that the best performance is at the 0.1 (0.53) and 0.05 expectiles (0.51). Further, we found that the 0.5 expectile has a lower performance than the 0.1, 0.2, and 0.3 expectiles, even for estimating the lower 50 % of the data.

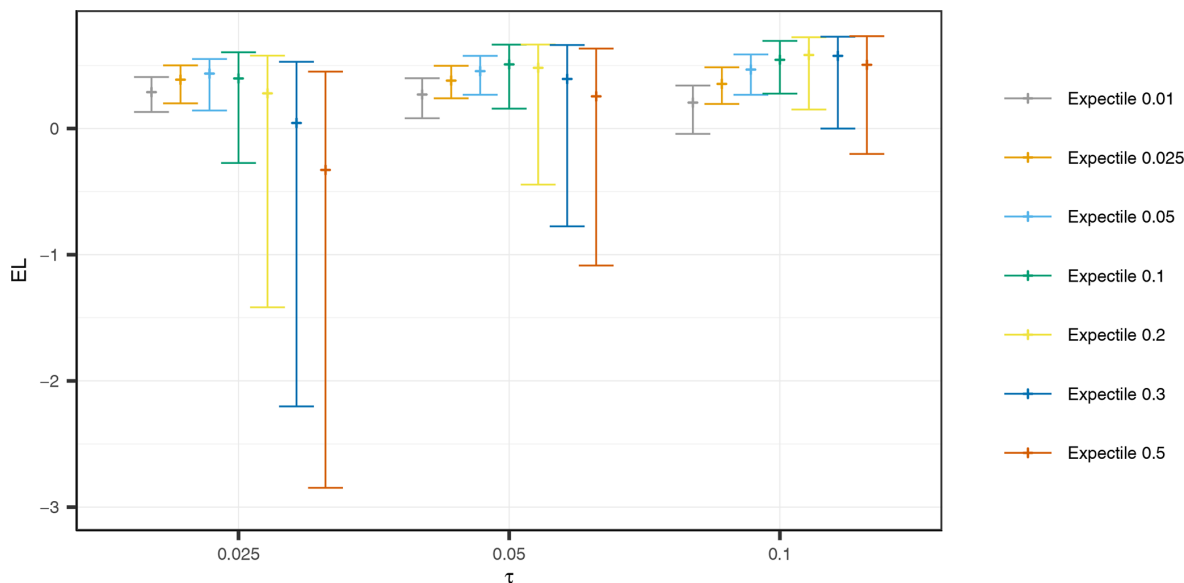
As a second assessment of the predictive performance for low extremes, we compute the relative expectile error ( $\text{EL}_\tau$ ), a weighted error metric that distributes weights by expectile functions that have the greatest weight around  $\tau$ . We choose  $\tau$  values of 0.1, 0.05, and 0.025, as these approximately correspond to the  $q_{95_d}$ ,  $q_{98_d}$ , and  $q_{99_d}$  thresholds used for drought identification. Figure 8 shows the expectile errors  $\text{EL}_\tau$  calculated for every station. Generally, the 0.1 expectile demonstrates a good performance by all  $\text{EL}_\tau$  metrics. Smaller expectiles (0.01, 0.025, 0.05) perform as good or better than the 0.1 expectile for the  $\text{EL}_{\tau=0.025}$  (which gives most weight to most extreme events), but not for the other  $\text{EL}_\tau$  metrics. Interestingly, the expectiles used for optimizing our loss functions do not perform best considering the respective

$\text{EL}_\tau$  metric. For example, we would assume that the best-performing expectile for the  $\text{EL}_{\tau=0.025}$  would be the 0.025 expectile, but on median, this expectile yields a  $\text{EL}_{\tau=0.025}$  of 0.39, which is somewhat lower than the 0.1 expectile (0.4) or the 0.025 expectile (0.44). Furthermore, the smallest expectile (0.01) only obtains a median  $\text{EL}_{\tau=0.025}$  of 0.29, which is only slightly higher than the 0.2 expectile (0.28). Higher expectiles (0.2, 0.3, 0.5) show an increasing performance with higher  $\tau$ , but their performance varies more across all stations. This suggests that one should choose an expectile function that gives most weight to somewhat higher values than the drought threshold of interest, and  $\tau$  values around 0.1 appear most accurate for  $q_{95_d}$  and  $q_{98_d}$  drought events.

In a final assessment of the model performance, we analyze the skill of the models to classify extreme events. For this purpose, we focus on the hit score ( $H_S$ ) and the precision ( $P_{\text{rec}}$ ) at each station. Table 5 shows the median of these two metrics over all stations and for all expectiles. Both skill scores decrease from less extreme ( $q_{95_d}$ ) to more extreme ( $q_{99_d}$ ) low flows for all expectiles. Concerning the effect of expectiles on  $P_{\text{rec}}$  and  $H_S$ , we see an opposite behavior, with  $H_S$  increasing towards the lower expectiles, and  $P_{\text{rec}}$  decreasing. For the lowest expectile (0.01), the  $H_S$  reaches 0.98, 0.94, and 0.9 for the three low-flow thresholds, respectively, showing that almost every observed drought event is identified. On the other hand, the  $P_{\text{rec}}$  is 0.2 for the  $q_{95_d}$  and



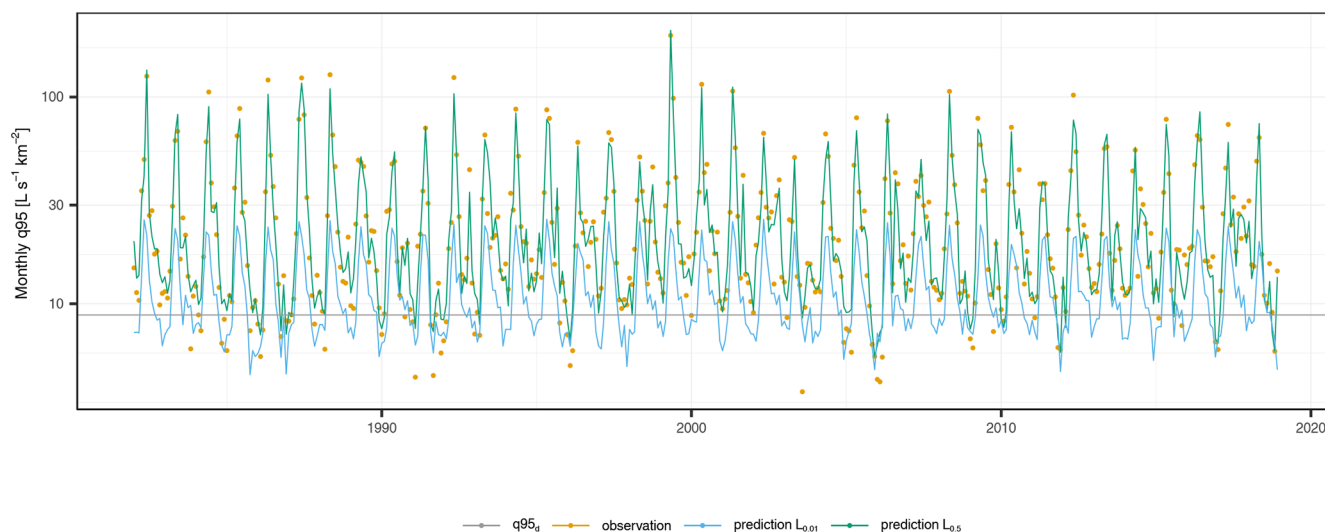
**Figure 7.** Global  $R^2$  conditional to observations below low-flow thresholds. The thresholds are (station-wise) quantiles of the monthly time series with 1 % to 100 % non-exceedance probability.



**Figure 8.**  $EL_\tau$  for  $\tau = 0.025, 0.05, 0.1$  presented for all expectiles. The median, 25 % quantile, and 75 % quantile over all stations and each expectile are displayed.

only 0.07 on median for the  $q_{99d}$ , which means that only 7 % of the predicted extreme events per station were actually extreme events. This points to an underestimation of extreme low flows when using low expectiles for model fitting. In contrast, using the 0.5 expectile would lead to an increase in

the precision rate, but to a very low hit score. Figure 9 shows the contrasting properties of a high and low expectile model on low-flow predictions. The low expectile shifts the predictions to match the most extreme events but underestimates the moderate events. Clearly, such low expectiles result in



**Figure 9.** The predictions for the 0.01 and 0.5 expectiles are shown for station Au/Bregenzzerach. The y axis is on a log scale for better visualization of the low extremes. Error metrics for the 0.5 expectile:  $R^2 = 0.95$ ,  $H_S = 0.51$ ,  $P_{\text{rec}} = 0.79$ ; 0.01 expectile:  $R^2 = -0.07$ ,  $H_S = 0.94$ ,  $P_{\text{rec}} = 0.25$ .

**Table 5.** The precision  $P_{\text{rec}}$  and hit rate  $H_S$  are computed per station and the median is shown in the table for  $q95_d$ ,  $q98_d$ , and  $q99_d$ .

Loss function	$q95_d$		$q98_d$		$q99_d$	
	$P_{\text{rec}}$	$H_S$	$P_{\text{rec}}$	$H_S$	$P_{\text{rec}}$	$H_S$
Expectile 0.5	0.61	0.50	0.37	0.33	0.21	0.18
Expectile 0.3	0.59	0.54	0.41	0.37	0.21	0.24
Expectile 0.2	0.53	0.59	0.35	0.44	0.21	0.29
Expectile 0.1	0.49	0.72	0.30	0.52	0.16	0.36
Expectile 0.05	0.32	0.84	0.19	0.65	0.10	0.45
Expectile 0.025	0.26	0.93	0.15	0.82	0.10	0.64
Expectile 0.01	0.20	0.98	0.11	0.94	0.07	0.90

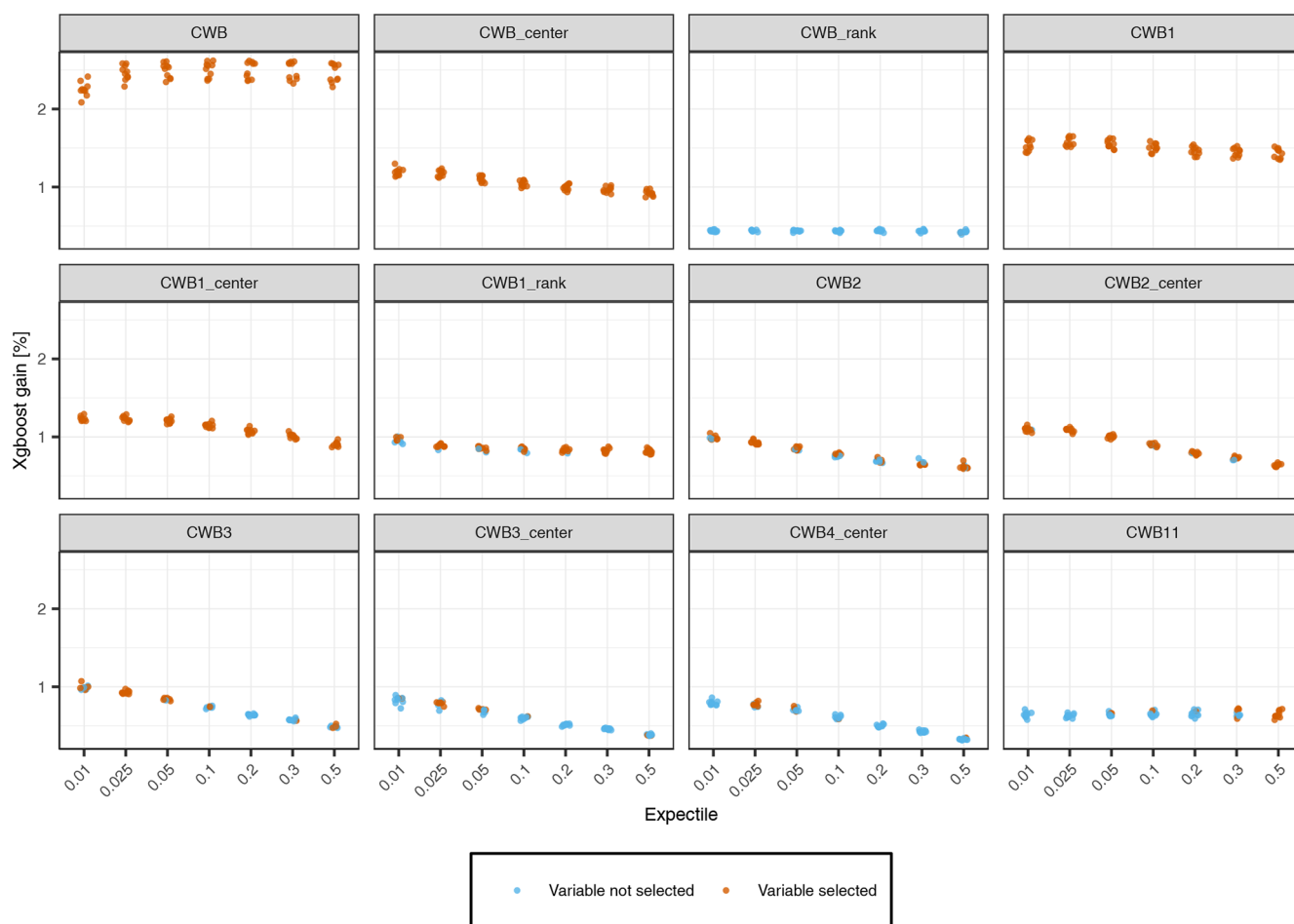
strongly biased models with little practical relevance. These findings suggest a trade-off between accurate prediction of extremes, overall prediction accuracy and correct classification of extreme events, where the user needs to find some optimum.

### 3.4 Variable selection

The implemented RFE algorithm leads to a substantial reduction of variables. On median, the different expectiles require between 20 and 30 variables, except the 0.5 expectile that has a median of 35 variables. A closer examination of the static predictors (an overview of the selection of the static variables is given in Appendix A) shows that geological variables are never selected for any expectiles, and also from the land-use variables only the fraction of forest is selected frequently. However, a larger number of meteorological variables were selected over all expectiles: aridity index (AI, AI<sub>sum</sub>), annual climatic water balance (MCWB),

annual precipitation ( $P$ ), and days with zero precipitation in the summer months ( $P_{0,\text{sum}}$ ). Despite their frequent appearance in models, the performance gain of using meteorological variables for predictions is low, with a typical variable importance of less than 2 % for all models. In contrast, topological variables emerge as the dominant predictors in the model. The three topological variables yielding the largest performance gain in the models are the mean and maximum catchment altitude ( $H_M$ ,  $H_+$ ) and the mean catchment slope  $S_M$ . The variable importance for  $H_M$  is on average between 3.8 % and 4.4 % for all expectiles. A slightly increasing trend towards larger expectiles can be observed for  $S_M$ , with a variable importance of 10 % for the 0.01 expectile and on average 11.6 % for the 0.5 expectile. An opposed effect can be observed for  $H_+$ , where the variable importance decreases from 7.2 % for the 0.01 expectile to 4.3 % for the 0.5 expectile. Other topological predictors whose inclusion only had a minor influence on model performance are catchment area, latitude, longitude, and altitude of the gauging station, stream network density, and the fractions of flat, moderate, and steep slopes in the catchment.

Figure 10 shows an overview of the variable importance for the most relevant spatiotemporal predictors. The CWB is the most important variable with more than 2 % performance gain over all expectiles. The CWB<sub>center</sub>, CWB<sub>1</sub>, and CWB<sub>1,center</sub> are also included in all models, but their variable importance is only 1 % to 1.5 %. Higher lags of the CWB are mainly included by the lower expectiles, but their performance gain is somewhat smaller and decreases with the lag of the variable. Comparing raw and standardized climate variables, the transformation of the absolute values to centered values seem to be beneficial in the case of the CWB, but this is not the case for the standardized drought indices,



**Figure 10.** Selection of spatiotemporal variables that are used for different expectiles. For each expectile it is shown whether the variable is used for the final prediction of the fold or not.

which were rarely selected and exhibited a low importance score in all cases.

## 4 Discussion

### 4.1 Value of expectile regression tree models for low-flow estimation

In this paper, we extended the extreme gradient boosting model XGBoost with the expectile loss function to develop a spatiotemporal model for low-flow predictions. We applied different  $\tau$  values from 0.01 to 0.5 and evaluated their predictive quality in terms of overall performance, accuracy at extremes, and potential to classify extreme events in a time series. Our findings showed two contrasting behaviors. Larger expectiles such as 0.2, 0.3, and 0.5 showed better performance in respect to  $R^2$  and  $R^2_{\text{med}}$ , whereas low expectiles ( $\tau = 0.01, 0.025, 0.05$ ) led to a sharp decline in  $R^2_{\text{med}}$  but increasing prediction accuracy at extreme low flows. However, optimization of the model cannot be reduced to these

two criteria, as was shown in the further assessments. Low expectiles lead to a reduced precision at extreme low-flow events as a result of too many events being predicted. This suggests an overfitting of the lower tail of the low-flow distribution by low expectiles that reduces the predictive performance at the entire distribution. In contrast, the hit score drops sharply for higher expectiles, resulting in a low detection rate of extreme events. Therefore, the application of the expectile loss has to be considered carefully and adjusted to the specific research question. Figure 11 gives a synopsis of key performance metrics with respect to predicting  $q_{95d}$ ,  $q_{98d}$ , and  $q_{99d}$  low-flow events. The synoptic representation nicely shows the trade-off between different performance metrics and puts the conclusions from their individual assessment into context. Depending on the low-flow event, a different optimum can be observed. When the focus of the study is on annual low flows in the order of  $q_{95d}$ , we see that the 0.2 expectile yields the optimal model fit, indicated by the crossing performance lines of the various metrics. However, when the purpose of the study is on predicting more

extreme events ( $q_{98d}$  and  $q_{99d}$ ), the 0.1 expectile is the optimal choice. These optima take into account the overall predictive performance ( $R^2_{med}$ ) and emphasize the performance at the considered extreme, by the expectile loss, the hit rate, and precision at the event of interest. The optimized model shows a good performance in all metrics and appears well suited for spatiotemporal predictions of low-flow events.

## 4.2 Performance compared to literature

Performance evaluation in light of the existing literature is not straightforward, as we are not aware of any studies evaluating monthly low-flow models. Nevertheless, some studies did model the mean monthly streamflow, which we will use for comparison. Comparisons will be made on the NSE reported in these studies, which is an analogue measure to the  $R^2$  in our study (Blöschl et al., 2013; Parajka et al., 2013). To put our monthly  $q_{95}$  models into context, we run an XGBoost model on monthly mean flow using the mean expectile regression (0.5) for our dataset. This leads to a median performance  $R^2_{med}$  of 0.77 with only 15 % of the stations having an  $R^2$  below 0.5. As expected, modeling monthly low flow shows weaker performance than modeling the mean monthly streamflow. Nevertheless, our results for  $q_{95}$  are still in the range of published studies on mean monthly streamflow models. For example, Cutore et al. (2007) found NSE values ranging from 0.57 to 0.78 by modeling 9 basins in Italy using ANNs. A similar performance has been reported by Pumo et al. (2016) for 59 stations in Sicily, Italy. They modeled monthly streamflow by interpolating regression parameters to ungauged basins and their NSE values ranged from 0.7 to 0.8 for 6 selected validation stations. A comparative assessment of regionalization approaches for hydrological models at 22 stations in the USA (Steinschneider et al., 2015) showed NSE values ranging from 0.6 to 0.85. Note that these results were obtained for relatively small datasets under quite homogeneous hydrological conditions. In our study, low-flow predictions were evaluated on a larger, diverse hydrological dataset and we found similar performance metrics to those in the aforementioned studies.

A more qualitative embedding to the scientific literature can be made by integrating our findings into the comparative evaluation of regionalization procedures of Parajka et al. (2013). Parajka et al. (2013) evaluated the performance of runoff-hydrograph studies, mostly performed on a daily time step, as a function of aridity, elevation, and catchment area to assess the extent to which performance depends on main climate and catchment characteristics. Figure 12 shows our finding in context of these three catchment characteristics. Parajka et al. (2013) found a decreasing performance with higher aridity. In our study, the aridity index is below 0.6 for all our catchments, but shows a similar decrease in performance. In our study, the performance increases with catchment elevation, with catchments below 450 m a.s.l. having the lowest performance ( $R^2_{med}$  of 0.5) and catchments above

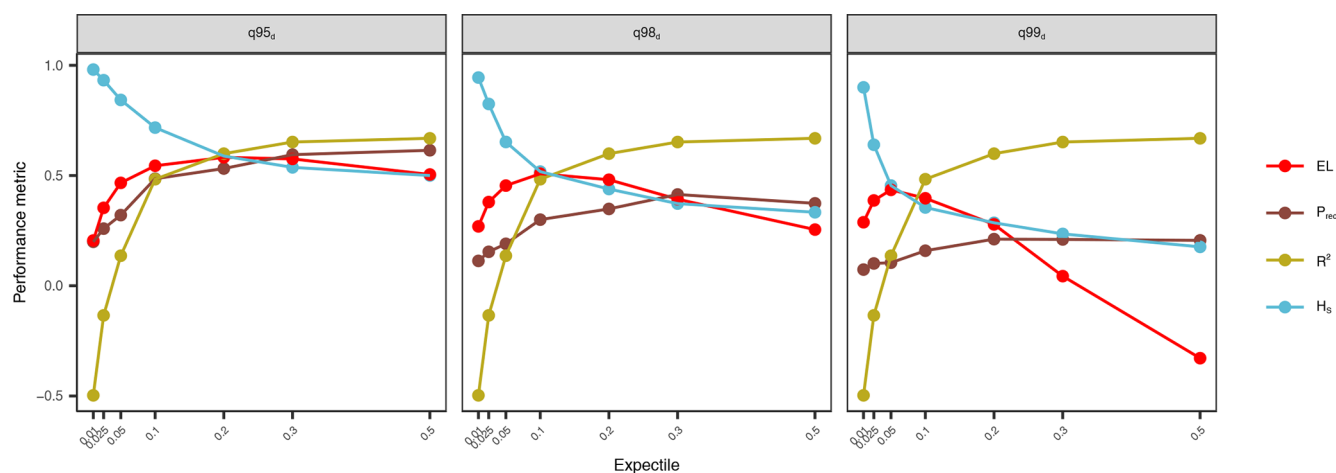
1350 m a.s.l. the highest performance ( $R^2_{med}$  of 0.8). The increase in performance with elevation points to a higher performance for catchments with dominant winter low-flow seasonality, which can be explained by their more regular and thus better predictable regime. We can indeed observe a lower performance for stations with a summer low-flow regime ( $R^2_{med} = 0.63$ ) than for stations with a winter low-flow regime ( $R^2_{med} = 0.73$ ). However, in a stratified assessment of summer and winter months, we could not find any difference in performance, neither in the entire study area, nor for summer- and winter-dominated catchments separately. This is in contrast to Staudinger et al. (2011), who found a reduction in model performance for summer months compared to winter months for a Norwegian study area with conditions comparable to the mountainous catchments in our study. Finally, we also found more accurate model predictions for larger basins. These are not due to a simple size effect of catchment area on discharges, since we use specific low flows as target variable to make catchments of different size more comparable. Many of our small catchments are located in a karstic environment or moors, which exhibit highly irregular regimes that are particularly hard to regionalize. Larger catchments, in contrast, have a more regular regime and show a high prediction performance.

## 5 Conclusions

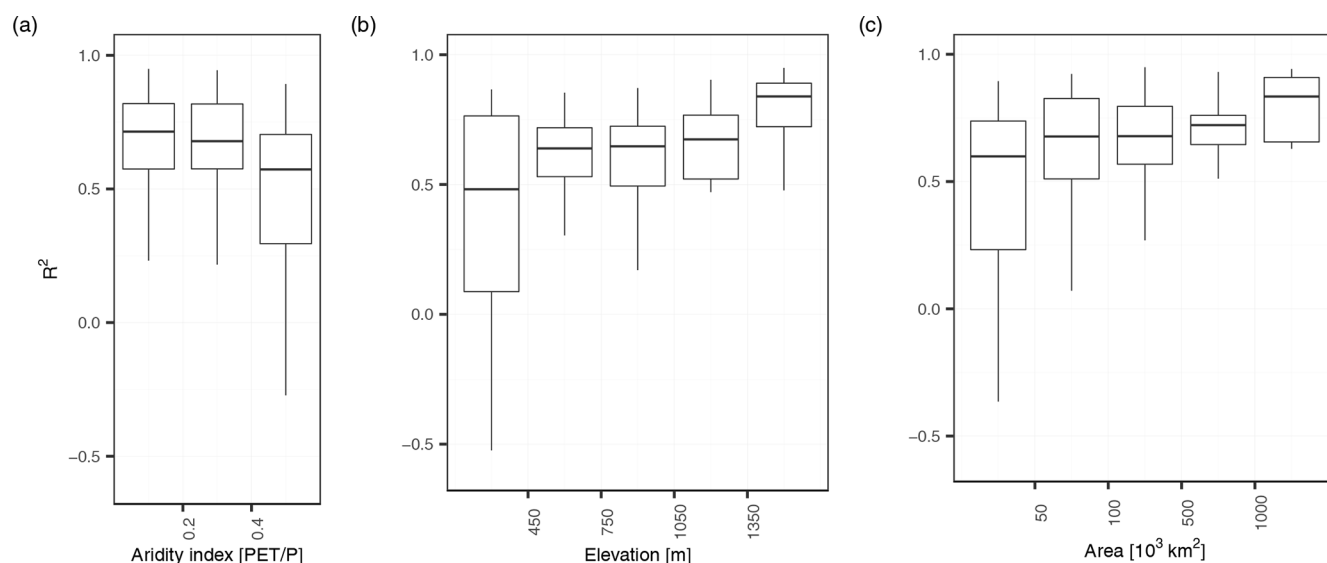
In this paper, we analyzed the performance of a single spatiotemporal XGBoost model on the prediction of monthly low flow for a comprehensive dataset of 260 gauging stations in Austria. We paid particular attention to the estimation of low extremes, by applying the expectile loss function as a fitting criterion. Our results show that the expectile loss yields a high prediction accuracy, but the performance decreases strongly for small expectiles. The best-performing model is the 0.5 expectile with a  $R^2_{med}$  of 0.67, but also the 0.2 and 0.3 expectile reach a higher  $R^2_{med}$  than the mean and median absolute loss. Small expectiles such as 0.01 or 0.025 already yield a negative  $R^2_{med}$ , resulting in a high number of poor-performing stations.

Weak-performing stations can also be found for the 0.5 expectile, where 26 % of the stations have an  $R^2$  below 0.5. A decomposition of the model error revealed that the monthly error is the main error component for large expectiles, while the seasonal and annual components are negligible for most stations. Considering the weak-performing stations of the 0.5 expectile, the dominant error component is a systematic error or bias. With a median value of 56 %, the bias is much larger at these stations compared to well-performing stations with a median value of only 11 %. This underlines the finding that the main shortcoming of our approach is a systematic error, which impairs the predictive accuracy. However, the strength of our spatiotemporal model is the good approximation of





**Figure 11.** Overview of the error metrics  $EL_{\tau}$ ,  $P_{\text{rec}}$ ,  $H_S$ , and  $R^2_{\text{med}}$ , plotted against every expectile. In each panel,  $H_S$  and  $P_{\text{rec}}$  are computed for the thresholds  $q95_d$ ,  $q98_d$ , and  $q99_d$ . The  $EL_{\tau}$  is calculated for the  $\tau$  values of 0.025 ( $q95_d$ ), 0.05 ( $q98_d$ ), and 0.1 ( $q99_d$ ).



**Figure 12.** Panels (a), (b), and (c) show  $R^2$  per station of the 0.5 expectile plotted against the aridity index, elevation and the area. In all plots, outliers are not included for a better visualization.

the seasonal and annual variability of monthly low flows, as shown by the low seasonal and annual error component.

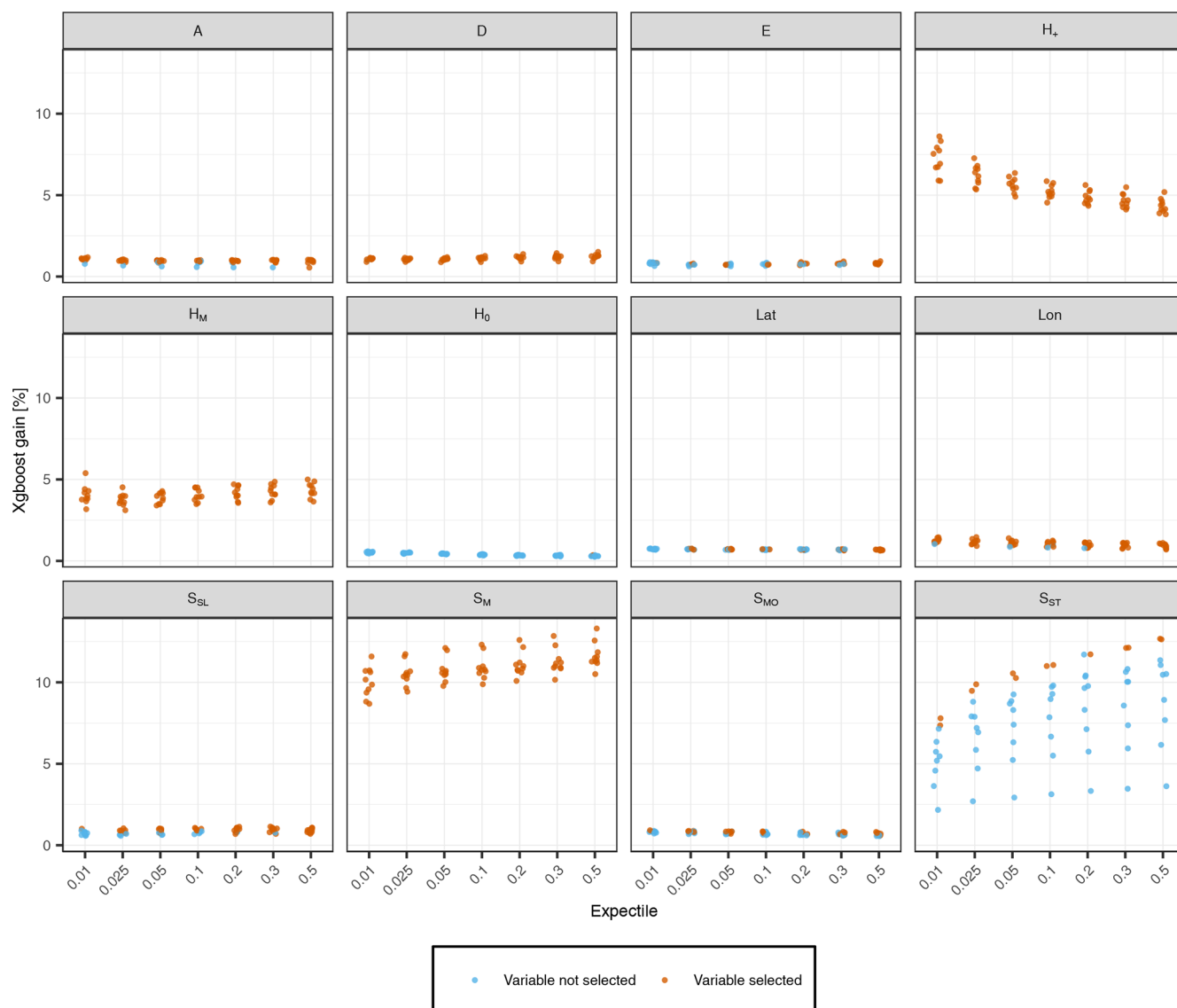
Despite the low global performance of small expectiles ( $\tau = 0.05, 0.025, 0.01$ ), they demonstrate an increasing accuracy when focusing on low extremes. This improvement in predictive performance has the simultaneous disadvantage that extreme low flows are increasingly misclassified. We found that the application of the expectile loss results in a trade-off between global performance, prediction accuracy of extremes, and misclassification rate of extreme events. The implementation of the 0.1 or 0.2 expectiles, depending on whether we want to optimize predictions for annual or more extreme low flow, appears to be optimal with respect to all criteria. The resulting extreme gradient tree boosting model

covers seasonal and annual variability nicely and provides a viable approach for spatiotemporal modeling of a range of hydrological variables that represent average conditions and extreme events.

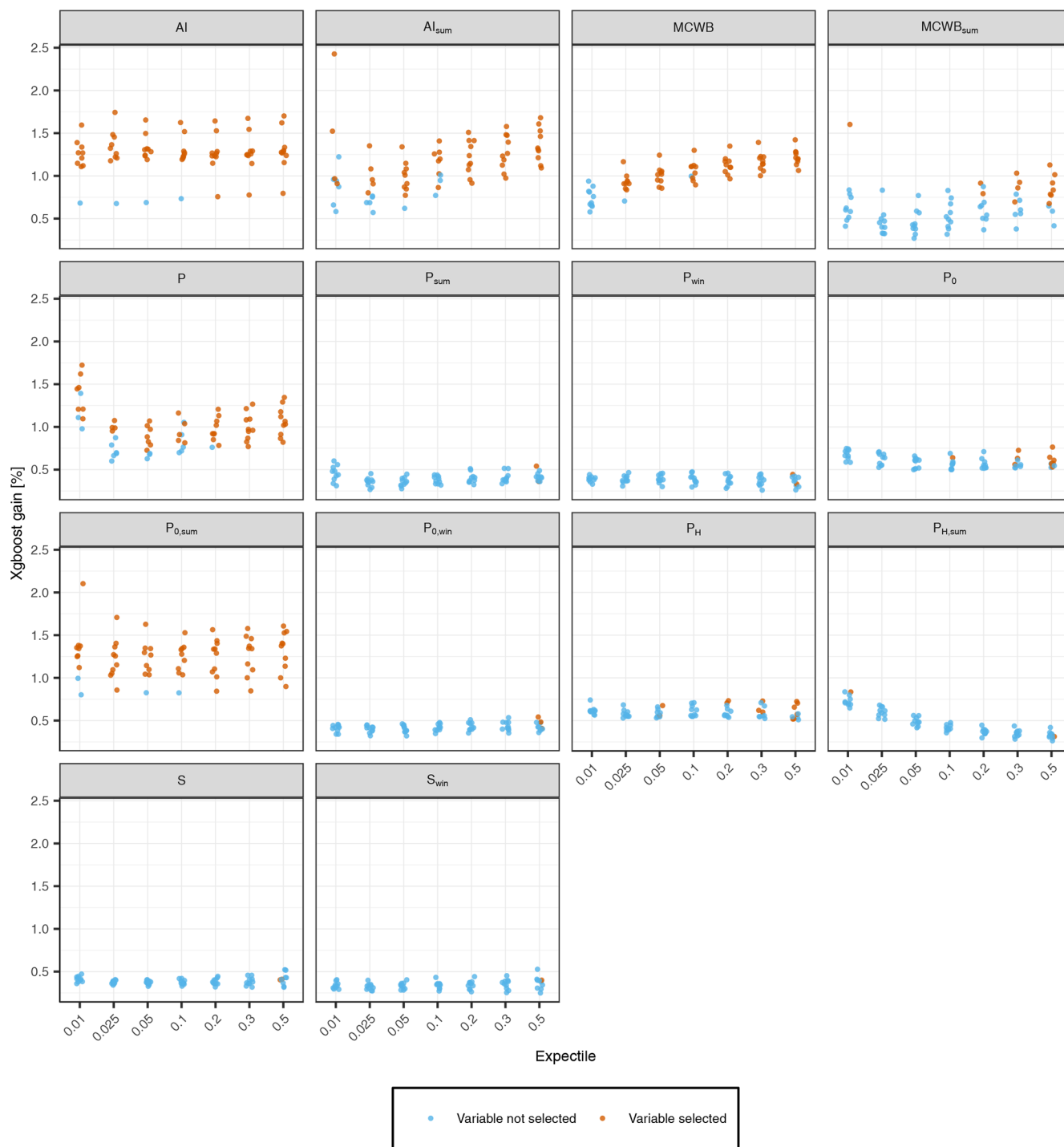
We demonstrate that the expectile loss is a suitable alternative to common loss functions in spatiotemporal low-flow models. However, its application is not limited to statistical learning models such as XGBoost, but can also be considered for hydrological models when their focus is on predicting hydrological extremes. As with all models, there is a trade-off between overall predictive performance, accuracy on the tails of the distribution, and identification of extreme events that should be considered in model applications.



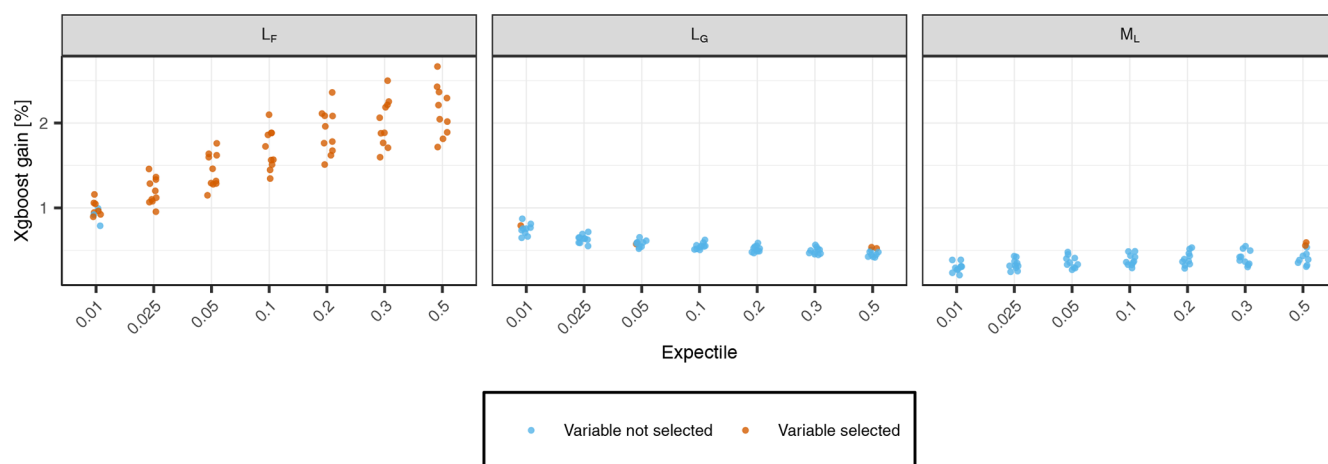
## Appendix A



**Figure A1.** Predictive power of static topological predictors obtained by the backward variable selection procedure (Sect. 2.2.3). The performance gains of the 10 prediction folds per expectile model are shown, with colors indicating whether the variable is used for the final prediction of the fold or not. Variables that were never selected by our model are not shown.



**Figure A2.** Predictive power of static meteorological predictors obtained by the backward variable selection procedure (Sect. 2.2.3). The performance gains of the 10 prediction folds per expectile model are shown, with colors indicating whether the variable is used for the final prediction of the fold or not. Variables that were never selected by our model are not shown.



**Figure A3.** Predictive power of static land-use predictors obtained by the backward variable selection procedure (Sect. 2.2.3). The performance gains of the 10 prediction folds per expectile model are shown, with colors indicating whether the variable is used for the final prediction of the fold or not. Variables that were never selected by our model are not shown.

*Code and data availability.* Data and code can be made available on personal request to johannes.laimighofer@boku.ac.at.

*Author contributions.* JL designed the research layout and GL contributed to its conceptualization. JL performed the formal analyses, and JL and GL prepared the draft paper. MM supported the analyses. GL supervised the overall study. All the authors contributed to the interpretation of the results and writing of the paper.

*Competing interests.* The contact author has declared that none of the authors has any competing interests.

*Disclaimer.* Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

*Acknowledgements.* Johannes Laimighofer is a recipient of a DOC fellowship (grant number 25819) of the Austrian Academy of Sciences, which is gratefully thanked for financial support.

Data provision by the Central Institute for Meteorology and Geodynamics (ZAMG) and the Hydrographic Service of Austria (HZB) was highly appreciated. This research supports the work of the UNESCO-IHP VIII FRIEND-Water program (FWP).

*Financial support.* This research has been supported by the Österreichischen Akademie der Wissenschaften (grant no. 25819).

*Review statement.* This paper was edited by Elena Toth and reviewed by three anonymous referees.

## References

- Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., and Wilby, R. L.: Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting, *Prog. Phys. Geog.*, 36, 480–513, <https://doi.org/10.1177/0309133312444943>, 2012.
- Aigner, D. J., Amemiya, T., and Poirier, D. J.: On the Estimation of Production Frontiers: Maximum Likelihood Estimation of the Parameters of a Discontinuous Density Function, *Int. Econ. Rev.*, 17, 377–396, <https://doi.org/10.2307/2525708>, 1976.
- Arnold, J. B.: ggthemes: Extra Themes, Scales and Geoms for “ggplot2”, r package version 4.2.4, <https://CRAN.R-project.org/package=ggthemes> (last access: 8 April 2022), 2021.
- Beguería, S., Vicente-Serrano, S. M., Reig, F., and Latorre, B.: Standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring, *Int. J. Climatol.*, 34, 3001–3023, <https://doi.org/10.1002/joc.3887>, 2014.
- Blöschl, G., Sivapalan, M., Wagener, T., Savenije, H., and Viglione, A.: Runoff prediction in ungauged basins: synthesis across processes, places and scales, Cambridge University Press, <https://doi.org/10.1017/CBO9781139235761>, 2013.
- Castiglioni, S., Castellarin, A., and Montanari, A.: Prediction of low-flow indices in ungauged basins through physiological space-based interpolation, *J. Hydrol.*, 378, 272–280, <https://doi.org/10.1016/j.jhydrol.2009.09.032>, 2009.
- Castiglioni, S., Castellarin, A., Montanari, A., Skøien, J. O., Laaha, G., and Blöschl, G.: Smooth regional estimation of low-flow indices: physiological space based interpolation and top-kriging, *Hydrol. Earth Syst. Sci.*, 15, 715–727, <https://doi.org/10.5194/hess-15-715-2011>, 2011.
- Chang, W. and Chen, X.: Monthly Rainfall-Runoff Modeling at Watershed Scale: A Comparative Study of Data-Driven and Theory-Driven Approaches, *Water*, 10, 1116, <https://doi.org/10.3390/w10091116>, 2018.

- Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, KDD '16, Association for Computing Machinery, New York, NY, USA, 785–794, <https://doi.org/10.1145/2939672.2939785>, 2016.
- Chen, T. and He, T.: Higgs Boson Discovery with Boosted Trees, in: Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning, 8–13 December 2014, Montreal, Canada, edited by: Cowan, G., Germain, C., Guyon, I., Kégl, B., and Rousseau, D., vol. 42 of Proceedings of Machine Learning Research, PMLR, Montreal, Canada, 69–80, <https://proceedings.mlr.press/v42/chen14.html> (last access: 8 April 2022), 2015.
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y.: xgboost: Extreme Gradient Boosting, r package version 1.4.1.1, <https://CRAN.R-project.org/package=xgboost> (last access: 8 April 2022), 2021.
- Cutore, P., Cristaudo, G., Campisano, A., Modica, C., Cancelliere, A., and Rossi, G.: Regional models for the estimation of streamflow series in ungauged basins, *Water Resour. Manag.*, 21, 789–800, <https://doi.org/10.1007/s11269-006-9110-7>, 2007.
- Dawson, C. and Wilby, R.: Hydrological modelling using artificial neural networks, *Prog. Phys. Geog.*, 25, 80–108, <https://doi.org/10.1177/030913330102500104>, 2001.
- Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G.: A framework to assess the realism of model structures using hydrological signatures, *Hydrol. Earth Syst. Sci.*, 17, 1893–1912, <https://doi.org/10.5194/hess-17-1893-2013>, 2013.
- Ferreira, R. G., da Silva, D. D., Elesbon, A. A. A., Fernandes-Filho, E. I., Veloso, G. V., de Souza Fraga, M., and Ferreira, L. B.: Machine learning models for streamflow regionalization in a tropical watershed, *J. Environ. Manage.*, 280, 111713, <https://doi.org/10.1016/j.jenvman.2020.111713>, 2021.
- Friedman, J., Hastie, T., and Tibshirani, R.: Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors), *Ann. Stat.*, 28, 337–407, <https://doi.org/10.1214/aos/1016218223>, 2000.
- Friedman, J., Hastie, T., and Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent, *J. Stat. Softw.*, 33, 1–22, <https://doi.org/10.18637/jss.v033.i01>, 2010.
- Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine, *Ann. Stat.*, 29, 1189–1232, <http://www.jstor.org/stable/2699986> (last access: 8 April 2022), 2001.
- Granitto, P. M., Furlanello, C., Biasioli, F., and Gasperi, F.: Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products, *Chemometr. Intell. Lab.*, 83, 83–90, <https://doi.org/10.1016/j.chemolab.2006.01.007>, 2006.
- Grolemund, G. and Wickham, H.: Dates and Times Made Easy with lubridate, *J. Stat. Softw.*, 40, 1–25, <https://doi.org/10.18637/jss.v040.i03>, 2011.
- Henry, L. and Wickham, H.: purrr: Functional Programming Tools, r package version 0.3.4, <https://CRAN.R-project.org/package=purrr> (last access: 8 April 2022), 2020.
- Henry, L. and Wickham, H.: tidyselect: Select from a Set of Strings, r package version 1.1.1, <https://CRAN.R-project.org/package=tidyselect> (last access: 8 April 2022), 2021.
- Huang, S., Kumar, R., Flörke, M., Yang, T., Hundecha, Y., Kraft, P., Gao, C., Gelfan, A., Liersch, S., Lobanova, A., Strauch, M., van Ogtrop, F., Reinhardt, J., Haberlandt, U., and Krysanova, V.: Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide, *Climatic Change*, 141, 381–397, <https://doi.org/10.1007/s10584-016-1841-8>, 2017.
- Kneib, T.: Beyond mean regression, *Stat. Model.*, 13, 275–303, <https://doi.org/10.1177/1471082X13494159>, 2013.
- Kneib, T., Silbersdorff, A., and Säfken, B.: Rage against the mean – a review of distributional regression approaches, *Econometrics and Statistics*, <https://doi.org/10.1016/j.ecosta.2021.07.006>, online first, 2021.
- Koenker, R. and Bassett, G.: Regression Quantiles, *Econometrica*, 46, 33–50, <https://doi.org/10.2307/1913643>, 1978.
- Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, *Water Resour. Res.*, 55, 11344–11354, <https://doi.org/10.1029/2019WR026065>, 2019a.
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, *Hydrol. Earth Syst. Sci.*, 23, 5089–5110, <https://doi.org/10.5194/hess-23-5089-2019>, 2019b.
- Kuhn, M.: caret: Classification and Regression Training, r package version 6.0-88, <https://CRAN.R-project.org/package=caret> (last access: 8 April 2022), 2021.
- Kuhn, M. and Wickham, H.: recipes: Preprocessing Tools to Create Design Matrices, r package version 0.1.16, <https://CRAN.R-project.org/package=recipes> (last access: 8 April 2022), 2021.
- Laaha, G. and Blöschl, G.: A comparison of low flow regionalisation methods–catchment grouping, *J. Hydrol.*, 323, 193–214, <https://doi.org/10.1016/j.jhydrol.2005.09.001>, 2006.
- Laaha, G. and Blöschl, G.: A national low flow estimation procedure for Austria, *Hydrolog. Sci. J.*, 52, 625–644, <https://doi.org/10.1623/hysj.52.4.625>, 2007.
- Laaha, G., Demuth, S., Hisdal, H., Kroll, C. N., van Lanen, H. A. J., Nester, T., Rogger, M., Sauquet, E., Tallaksen, L. M., Woods, R. A., and Young, A.: Prediction of low flows in ungauged basins, Cambridge University Press, 163–188, <https://doi.org/10.1017/CBO9781139235761.011>, 2013.
- Laaha, G., Sköien, J., and Blöschl, G.: Spatial prediction on river networks: comparison of top-kriging with regional regression, *Hydrol. Process.*, 28, 315–324, <https://doi.org/10.1002/hyp.9578>, 2014.
- Laaha, G., Gauster, T., Tallaksen, L. M., Vidal, J.-P., Stahl, K., Prudhomme, C., Heudorfer, B., Vlnas, R., Ionita, M., Van Lanen, H. A. J., Adler, M.-J., Caillouet, L., Delus, C., Fendekova, M., Gailliez, S., Hannaford, J., Kingston, D., Van Loon, A. F., Mediero, L., Osuch, M., Romanowicz, R., Sauquet, E., Stagge, J. H., and Wong, W. K.: The European 2015 drought from a hydrological perspective, *Hydrol. Earth Syst. Sci.*, 21, 3001–3024, <https://doi.org/10.5194/hess-21-3001-2017>, 2017.
- Laimighofer, J., Melcher, M., and Laaha, G.: Parsimonious statistical learning models for low-flow estimation, *Hydrol. Earth Syst. Sci.*, 26, 129–148, <https://doi.org/10.5194/hess-26-129-2022>, 2022.
- Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped con-

- ceptual models, *Hydrol. Earth Syst. Sci.*, 25, 5517–5534, <https://doi.org/10.5194/hess-25-5517-2021>, 2021.
- Lima, C. H. and Lall, U.: Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow, *J. Hydrol.*, 383, 307–318, <https://doi.org/10.1016/j.jhydrol.2009.12.045>, 2010.
- Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., and Sheppard, L.: A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates, *Environ. Ecol. Stat.*, 21, 411–433, 2014.
- Lu, H. and Ma, X.: Hybrid decision tree-based machine learning models for short-term water quality prediction, *Chemosphere*, 249, 126169, <https://doi.org/10.1016/j.chemosphere.2020.126169>, 2020.
- Müller, K. and Wickham, H.: tibble: Simple Data Frames, *r* package version 3.1.2, <https://CRAN.R-project.org/package=tibble> (last access: 8 April 2022), 2021.
- Newey, W. K. and Powell, J. L.: Asymmetric Least Squares Estimation and Testing, *Econometrica*, 55, 819–847, <https://doi.org/10.2307/1911031>, 1987.
- Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., and Liu, J.: Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model, *J. Hydrol.*, 586, 124901, <https://doi.org/10.1016/j.jhydrol.2020.124901>, 2020.
- Onyutha, C.: Influence of hydrological model selection on simulation of moderate and extreme flow events: a case study of the Blue Nile basin, *Adv. Meteorol.*, 2016, 7148326, <https://doi.org/10.1155/2016/7148326>, 2016.
- Ossandón, Á., Brunner, M. I., Rajagopalan, B., and Kleiber, W.: A space-time Bayesian hierarchical modeling framework for projection of seasonal maximum streamflow, *Hydrol. Earth Syst. Sci.*, 26, 149–166, <https://doi.org/10.5194/hess-26-149-2022>, 2022.
- Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins – Part 1: Runoff-hydrograph studies, *Hydrol. Earth Syst. Sci.*, 17, 1783–1795, <https://doi.org/10.5194/hess-17-1783-2013>, 2013.
- Parajka, J., Blaschke, A. P., Blöschl, G., Haslinger, K., Hepp, G., Laaha, G., Schöner, W., Trautvetter, H., Viglione, A., and Zessner, M.: Uncertainty contributions to low-flow projections in Austria, *Hydrol. Earth Syst. Sci.*, 20, 2085–2101, <https://doi.org/10.5194/hess-20-2085-2016>, 2016.
- Parisouj, P., Mohebzadeh, H., and Lee, T.: Employing machine learning algorithms for streamflow prediction: a case study of four river basins with different climatic zones in the United States, *Water Resour. Manage.*, 34, 4113–4131, <https://doi.org/10.1007/s11269-020-02659-5>, 2020.
- Pumo, D., Viola, F., and Noto, L. V.: Generation of natural runoff monthly series at ungauged sites using a regional regressive model, *Water*, 8, 209, <https://doi.org/10.3390/w8050209>, 2016.
- Ram, K. and Wickham, H.: wesanderson: A Wes Anderson Palette Generator, *r* package version 0.3.6, <https://CRAN.R-project.org/package=wesanderson> (last access: 8 April 2022), 2018.
- R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/> (last access: 8 April 2022), 2021.
- Roksvåg, T., Steinsland, I., and Engeland, K.: Estimation of annual runoff by exploiting long-term spatial patterns and short records within a geostatistical framework, *Hydrol. Earth Syst. Sci.*, 24, 4109–4133, <https://doi.org/10.5194/hess-24-4109-2020>, 2020.
- Sahour, H., Gholami, V., and Vazifedan, M.: A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer, *J. Hydrol.*, 591, 125321, <https://doi.org/10.1016/j.jhydrol.2020.125321>, 2020.
- Salinas, J. L., Laaha, G., Rogger, M., Parajka, J., Viglione, A., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins – Part 2: Flood and low flow studies, *Hydrol. Earth Syst. Sci.*, 17, 2637–2652, <https://doi.org/10.5194/hess-17-2637-2013>, 2013.
- Shortridge, J. E., Guikema, S. D., and Zaitchik, B. F.: Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds, *Hydrol. Earth Syst. Sci.*, 20, 2611–2628, <https://doi.org/10.5194/hess-20-2611-2016>, 2016.
- Shrestha, R. R., Peters, D. L., and Schnorbus, M. A.: Evaluating the ability of a hydrologic model to replicate hydro-ecologically relevant indicators, *Hydrol. Process.*, 28, 4294–4310, 2014.
- Smakhtin, V.: Low flow hydrology: a review, *J. Hydrol.*, 240, 147–186, [https://doi.org/10.1016/S0022-1694\(00\)00340-1](https://doi.org/10.1016/S0022-1694(00)00340-1), 2001.
- Sobotka, F. and Kneib, T.: Geoadditive expectile regression, *Comput. Stat. Data An.*, 56, 755–767, <https://doi.org/10.1016/j.csda.2010.11.015>, 2012.
- Solomatine, D. P. and Ostfeld, A.: Data-driven modelling: some past experiences and new approaches, *J. Hydroinform.*, 10, 3–22, <https://doi.org/10.2166/hydro.2008.015>, 2008.
- Staudinger, M. and Seibert, J.: Predictability of low flow – An assessment with simulation experiments, *J. Hydrol.*, 519, 1383–1393, <https://doi.org/10.1016/j.jhydrol.2014.08.061>, 2014.
- Staudinger, M., Stahl, K., Seibert, J., Clark, M. P., and Tallaksen, L. M.: Comparison of hydrological model structures based on recession and low flow simulations, *Hydrol. Earth Syst. Sci.*, 15, 3447–3459, <https://doi.org/10.5194/hess-15-3447-2011>, 2011.
- Steinschneider, S., Yang, Y.-C. E., and Brown, C.: Combining regression and spatial proximity for catchment model regionalization: a comparative study, *Hydrolog. Sci. J.*, 60, 1026–1043, <https://doi.org/10.1080/02626667.2014.899701>, 2015.
- Toth, E.: Estimation of flood warning runoff thresholds in ungauged basins with asymmetric error functions, *Hydrol. Earth Syst. Sci.*, 20, 2383–2394, <https://doi.org/10.5194/hess-20-2383-2016>, 2016.
- Tyralis, H., Papacharalampous, G., and Langousis, A.: Super ensemble learning for daily streamflow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms, *Neural Comput. Appl.*, 33, 3053–3068, <https://doi.org/10.1007/s00521-020-05172-3>, 2021a.
- Tyralis, H., Papacharalampous, G., Langousis, A., and Papalexioiu, S. M.: Explanation and probabilistic prediction of hydrological signatures with statistical boosting algorithms, *Remote Sens.*, 13, 333, <https://doi.org/10.3390/rs13030333>, 2021b.
- Tyralis, H., Papacharalampous, G., and Khatami, S.: Expectile-based hydrological modelling for uncertainty estimation: Life after mean, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.2201.05712>, 14 January 2022.
- Vandewiele, G. and Elias, A.: Monthly water balance of ungauged catchments obtained by geographical regionalization, *J. Hydrol.*,

- 170, 277–291, [https://doi.org/10.1016/0022-1694\(95\)02681-E](https://doi.org/10.1016/0022-1694(95)02681-E), 1995.
- Varmuza, K. and Filzmoser, P.: Introduction to multivariate statistical analysis in chemometrics, CRC press, <https://doi.org/10.1201/9781420059496>, 2016.
- Vicente-Guillén, J., Ayuga-Telléz, E., Otero, D., Chávez, J., Ayuga, F., and García, A.: Performance of a monthly Streamflow prediction model for Ungauged watersheds in Spain, *Water Resour. Manage.*, 26, 3767–3784, <https://doi.org/10.1007/s11269-012-0102-5>, 2012.
- Waltrup, L. S., Sobotka, F., Kneib, T., and Kauermann, G.: Expectile and quantile regression—David and Goliath?, *Stat. Model.*, 15, 433–456, <https://doi.org/10.1177/1471082X14561155>, 2015.
- Wickham, H.: *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York, <https://ggplot2.tidyverse.org> (last access: 8 April 2022), 2016.
- Wickham, H.: *tidyr: Tidy Messy Data*, r package version 1.1.3, <https://CRAN.R-project.org/package=tidyr> (last access: 8 April 2022), 2021.
- Wickham, H., François, R., Henry, L., and Müller, K.: *dplyr: A Grammar of Data Manipulation*, r package version 1.0.7, <https://CRAN.R-project.org/package=dplyr> (last access: 8 April 2022), 2021.
- Worland, S. C., Farmer, W. H., and Kiang, J. E.: Improving predictions of hydrological low-flow indices in ungaged basins using machine learning, *Environ. Model. Softw.*, 101, 169–182, <https://doi.org/10.1016/j.envsoft.2017.12.021>, 2018.
- Yang, X., Magnusson, J., Rizzi, J., and Xu, C.-Y.: Runoff prediction in ungauged catchments in Norway: comparison of regionalization approaches, *Hydrol. Res.*, 49, 487–505, <https://doi.org/10.2166/nh.2017.071>, 2017.
- Yu, X., Wang, Y., Wu, L., Chen, G., Wang, L., and Qin, H.: Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting, *J. Hydrol.*, 582, 124293, <https://doi.org/10.1016/j.jhydrol.2019.124293>, 2020.
- Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R.: Ensemble machine learning paradigms in hydrology: A review, *J. Hydrol.*, 598, 126266, <https://doi.org/10.1016/j.jhydrol.2021.126266>, 2021.