Statistical learning and topkriging improve spatio-temporal low-flow estimation

Johannes Laimighofer[1] and Gregor Laaha[2]

[1]Institute of Statistics, University of Natural Resources and Life Sciences
[2]University of 4 5 Natural Resources and Life Sciences, Vienna

June 11, 2023

Abstract

This study assesses the potential of a hierarchical space-time model for monthly low-flow prediction in Austria. The model decomposes the monthly low-flows into a mean field and a residual field, where the mean field estimates the seasonal low-flow regime augmented by a long-term trend component. We compare four statistical (learning) approaches for the mean field, and three geostatistical methods for the residual field. All model combinations are evaluated using a hydrological diverse dataset of 260 stations in Austria, covering summer, winter, and mixed regimes. Model validation is performed by a nested 10-fold cross-validation. The best model for monthly low-flow prediction is a combination of a model-based boosting approach for the mean field and topkriging for the residual field. This model reaches a median R2 of 0.73. Model performance is generally higher for stations with a winter regime (best model yields median R2 of 0.84) than for summer regimes (R2 = 0.7), and lowest for the mixed regime type (R2 = 0.68). The model appears especially valuable in headwater catchments, where the performance increases from 0.56 (median R2 for simple topkriging routine) to 0.67 for the best model combination. The favorable performance results from the hierarchical model structure that effectively combines different types of information: average low-flow conditions estimated from climate and catchment characteristics, and information of adjacent catchments estimated by spatial correlation. The model is shown to provide robust estimates not only for moderate events, but also for extreme low-flow events where predictions are adjusted based on synchronous local observations.

Statistical learning and topkriging improve spatio-temporal low-flow estimation

J. Laimighofer[1], G. Laaha[1]

[1]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

Key Points:

- Model-based boosting of the seasonal low-flow regime and topkriging for the residual field improve monthly low-flow predictions.
- Model accuracy is particularly high in the alpine areas, where low-flow occurs predominantly in winter.
- The hierarchical model structure is especially valuable in headwater catchments, and shows good performance for extreme events.

Corresponding author: Johannes Laimighofer, `johannes.laimighofer@boku.ac.at`

Abstract

This study assesses the potential of a hierarchical space-time model for monthly low-flow prediction in Austria. The model decomposes the monthly low-flows into a mean field and a residual field, where the mean field estimates the seasonal low-flow regime augmented by a long-term trend component. We compare four statistical (learning) approaches for the mean field, and three geostatistical methods for the residual field. All model combinations are evaluated using a hydrological diverse dataset of 260 stations in Austria, covering summer, winter, and mixed regimes. Model validation is performed by a nested 10-fold cross-validation. The best model for monthly low-flow prediction is a combination of a model-based boosting approach for the mean field and topkriging for the residual field. This model reaches a median R^2 of 0.73. Model performance is generally higher for stations with a winter regime (best model yields median R^2 of 0.84) than for summer regimes ($R^2 = 0.7$), and lowest for the mixed regime type ($R^2 = 0.68$). The model appears especially valuable in headwater catchments, where the performance increases from 0.56 (median R^2 for simple topkriging routine) to 0.67 for the best model combination. The favorable performance results from the hierarchical model structure that effectively combines different types of information: average low-flow conditions estimated from climate and catchment characteristics, and information of adjacent catchments estimated by spatial correlation. The model is shown to provide robust estimates not only for moderate events, but also for extreme low-flow events where predictions are adjusted based on synchronous local observations.

1 Introduction

Droughts and low-flows are significant hydrological and environmental hazards that threaten a wide range of water-related sectors, such as navigation, hydropower production and water management in general. Currently, prediction of low-flow is mainly focused on the spatial scale (Euser et al., 2013; Salinas et al., 2013; Castiglioni et al., 2009; Laaha et al., 2014; Tyralis et al., 2021; Worland et al., 2018; Laimighofer et al., 2022a), whereby deterministic models, or statistical models are applied. Spatio-temporal low-flow prediction is still rare, although space-time information on monthly low-flow is crucial for assessing ecological impacts on water quality, or estimating the risk of navigation disruptions. Space-time models are currently used in a wide range of environmental research fields (Kyriakidis & Journel, 1999), e.g. soil moisture modelling (Rodríguez-Iturbe et al., 2006), distribution of atmospheric pollution (Szpiro et al., 2010; Sampson et al., 2011; Lindström et al., 2014; Lindstrom et al., 2019; Mercer et al., 2011), downscaling meteorological variables (Wilby et al., 1998), or risk of wildfire outbreaks (Opitz et al., 2020). Transferring these space-time models to streamflow poses a particular challenge due to the tree-like structure of river catchments. Nevertheless, space-time models for streamflow are of particular interest, as they can be used for prediction in ungauged basins (Hrachowitz et al., 2013, PUB). This study aims to transfer an existing approach, originally adapted for air pollution modelling (Szpiro et al., 2010), to the space-time prediction of monthly low-flow.

Conceptually, statistical space-time models can be divided into individual space-time models, models that use temporal functions (deterministic or stochastic) that are correlated in space, or spatial functions that are correlated in time (Kyriakidis & Journel, 1999). The latter are less common for streamflow. Individual space-time models for prediction in ungauged basins (PUB) are mainly based on data-driven approaches such as long short-term memories (Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019; Lees et al., 2021, LSTM), artificial neural networks (Solomatine & Ostfeld, 2008; Cutore et al., 2007, ANN), or other machine learning methods such as tree-based models (Laimighofer et al., 2022a). These models typically use auxiliary space-time information on precipitation or evapotranspiration for streamflow estimation. In contrast, spatio-temporal geostatistical approaches exploit the similarity of hydrographs from nearby catchments. The simplest case is to apply ordinary kriging to the runoff time series, neglecting temporal correlations. In this context, Farmer (2016) found that such a simple model requires only a single (pooled) variogram to yield a median Nash-Sutcliffe efficiency of 0.7 for daily streamflow predictions on 182 stations in the United States. Ordinary kriging may not be the best choice for runoff, due to the nested and tree-like structure of the catchments. Therefore, other methods have been developed to take into account the peculiarities of catchment runoff. For example methods constraining the spatial covariance function by the water balance (Müller & Thompson, 2015), or methods that incorporate the river network hierarchy (Gottschalk, 1993; Sauquet et al., 2000), such as topkriging (Skøien et al., 2006; Skøien & Blöschl, 2007, TK). Farmer (2016) compared ordinary kriging to topkriging and showed a similar performance for both approaches. This is in contrast to studies in Austria and France (Skøien & Blöschl, 2007; Viglione et al., 2013; de Lavenne et al., 2016), which showed a favorable performance of topkriging also for daily and hourly runoff. Skøien and Blöschl (2007) additionally found, that in their topkriging application it was sufficient to estimate each time step separately, and no temporal dependency structure needed to be considered to achieve adequate performance.

Space-time models of the type where a temporal function (stochastic or deterministic) is correlated in space, are more common for runoff applications. They can be used, for instance, to improve the predictions of a hydrological model, when considering the output of a hydrological model as a deterministic function, which is interpolated in space by its model parameters. This regionalization of model parameters is performed on different temporal and spatial resolutions (Guo et al., 2021; Razavi & Coulibaly, 2013). Applications that use a stochastic temporal function are less frequent. For instance, Pumo et al. (2016) used a time series model for estimating monthly runoff in 59 basins in Sicily, with NSE values ranging from 0.7 to 0.8, but the model was validated only on a small subset of catchments. The time series model of Pumo et al. (2016) was determined a priori and only the coefficients of the parameters were estimated in space. A more flexible approach, that involves less information loss, is to use empirical ortoghonal functions (EOF). Gottschalk et al. (2015) and Li et al. (2018) applied EOFs for filling gaps in monthly discharge time series and Sauquet et al. (2008) tested spatially weighted EOFs for prediction of normalized mean monthly runoff in France. Studies, intended to model air pollutants, extended the approach of weighted EOFs, by adding a residual field (Szpiro et al., 2010), altering the methods for estimating the weights of the EOFs (Sampson et al., 2011; Mercer et al., 2011), or including spatio-temporal variables (Lindström et al., 2014; Lindstrom et al., 2019). All these studies analysed air pollutants in the United States, and reported cross-validated R^2 from 0.6 to about 0.75. The flexible model structure and the already highlighted use of EOFs for streamflow variables (Gottschalk, 1993; Li et al., 2018; Sauquet et al., 2008) demonstrate the potential for transferring this model to monthly low-flow. Such a transfer would involve incorporating both the average low-flow regime and the nested structure of river networks into the model.

The main objective of this study is to develop a hierarchical spatio-temporal model for monthly low-flow in Austria. The model consists of a mean field which should capture the seasonal cycle and the long-term trend of monthly low-flow and a residual field where geostatistical approaches are deployed. We test four different models for the mean field: (i) spatially weighted smoothed EOFs, (ii) a model-based boosting approach, which only estimates the seasonal cycle, (iii) a model-based boosting approach, which estimates the seasonal cycle and the long-term trend and (iv) a combination of model (ii) and (i). For the residual field we compare three kriging approaches - ordinary kriging (OK), physiographic kriging (PK) and topkriging (TK). The models are evaluated on a comprehensive Austrian dataset by 10-fold nested cross validation (CV) to emulate prediction in ungauged basins. The following research questions will be addressed:

1. Can a combination of statistical learning approaches and kriging methods improve spatio-temporal low-flow prediction in Austria?

2. What approach is best suited to model the seasonal low-flow regime?
3. Which kriging variant is best suited to model the space-time residual field?
4. How does prediction performance vary between headwater and non-headwater catchments?
5. What is the performance for summer, winter and mixed low-flow regimes?

2 Data

2.1 Hydrological data

This study is performed on a hydrological diverse dataset in Austria. We use 260 stations with a continuous daily streamflow record between 1982 to 2018. The same dataset was already used in a study on spatial low-flow prediction (Laimighofer et al., 2022b) and spatio-temporal low-flow prediction in Austria (Laimighofer et al., 2022a). The hydrological data can be downloaded from the Hydrographic Service of Austria (HZB). Our study focuses on a space-time model for low-flow. Hence, the daily streamflow time series is used to calculate the 0.05 quantile of discharge for every month (444 months at every station). We will refer to this index as monthly Q95 ($P(Q > Q95) = 0.95$). The monthly Q95 was standardized by catchment area, which results in the monthly specific low-flow (q95) time series (l s^{-1} km^{-2}). For all modelling approaches q95 is transformed by the square root, to approximate a normal distribution.

Occurrence of low-flow in Austria is more dominant in the winter half-year (November to April, winter regime type) for alpine catchments, where summer discharge is increased by snowmelt and increasing precipitation (Laaha & Blöschl, 2006; Laaha et al., 2017). In the northern parts of Austria and the Eastern low-lands low-flow mainly is present in the summer half-year (May to October, summer regime type). Nevertheless, not all catchments have this strong seasonality, and the occurrence of low-flow is alternating between winter and summer. This type of low-flow regime will be referred to as mixed regime type (Laaha & Blöschl, 2006; Laaha, 2023). The regime types are defined based on the seasonality ratio (SR)

$$SR = Q95_{summer}/Q95_{winter}, \tag{1}$$

where $Q95_{summer}$ is the 0.05 quantile of daily discharge for the summer period (May to November), and $Q95_{winter}$ the corresponding 0.05 quantile for the winter period of the respective station. A SR below 0.8 indicates a summer regime, a SR above 1.25 (1/0.8) determines a winter regime, and a SR between 0.8 and 1.25 is defined as a mixed regime. A graphical illustration of the defined regime types is given in Fig. 1. Despite the models developed here are on monthly time scale and thus not restricted to a particular regime type, we will use the seasonality regime types for an in-depth analysis of the results.

2.2 Catchment characteristics

In this study we apply several geostatistical and statistical learning methods, which all rely on catchment characteristics, that are supposed to be static over time in our approach. Ordinary kriging uses the geographic coordinates of the gauging stations, top-kriging requires the river network as input, and physiographic kriging is based on a principal component analysis of all catchment characteristics. The catchment characteristics can be subdivided into landuse variables, topographic descriptors, geological predictors and climatic characteristics. An overview of all variables is given in Table 1. For a more detailed description of the computation of the catchment characteristics we refer to Laaha and Blöschl (2006) and Laimighofer et al. (2022b). How the temporal information is added to the space-time models will be explained in Sect. 3.2.

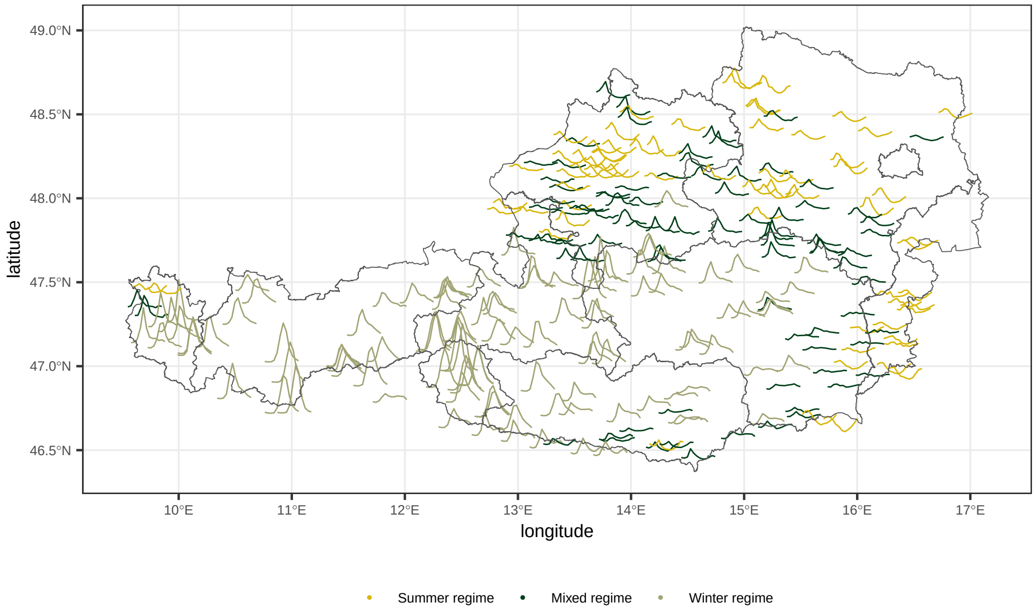

Figure 1. Overview of the study area. The colours indicate the seasonality regime type of the station, defined by the SR. The curves of each station is the scaled seasonal low-flow at each station for illustration of the different regime types.

3 Methods

3.1 Model structure

The basic model structure is given by

$$y(s,t) = \mu(s,t) + \upsilon(s,t), \tag{2}$$

where $y(s,t)$ is the monthly low-flow at a station s and time point t, $\mu(s,t)$ is defined as the mean field and $\upsilon(s,t)$ is the residual field of our model. Similar model designs were used by e.g. Szpiro et al. (2010), Lindstrom et al. (2019) or Sampson et al. (2011). In this model conceptualization the mean field should capture the seasonal cycle and the long-term trend of the response variable. Szpiro et al. (2010) used ordinary kriging for prediction of the space-time residual field, where only one variogram is estimated for all timesteps. A graphical overview specific to low-flow is shown in Fig. 2. In this study we extend the model introduced by Szpiro et al. (2010) to capture the nested structure of river catchments. We employ a hierarchical modeling framework, that (i) considers four different modeling approaches for the mean field, and (ii) three forms of kriging for the space-time residual field, to find the best-performing model combination for monthly low-flow prediction.

3.2 Mean field

The objective for modelling the mean field is to estimate the seasonal cycle and the long-term trend in the spatio-temporal model. In the context of low flows, the seasonal cycle corresponds to the average monthly low-flow regime (seasonal low-flow regime), which is augmented to transient conditions by the long-term trend component. Szpiro et al. (2010) or Lindström et al. (2014) used weighted empirical orthogonal functions (EOF), which were initially proposed by Fuentes et al. (2006), for estimating the mean field. Their

Table 1. Description of the catchment characteristics used in this study. The climatic characteristics as precipitation, climatic water balance, potential evapotranspiration, aridity index, snowmelt and temperature are computed 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_{win}), sum for summer (e.g. P_{sum}).

Variable	Description	Unit
A	catchment area	km^2
Lat, Lon	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	%
L_U, L_A, L_C, L_F, L_G, L_R, L_W, L_{WA}	Fraction of urban areas, agricultural areas, permanent crop, forest, grassland, wasteland, wetlands, water surfaces	%
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	%
D	Stream network density	10^2 m km^{-2}
P	Precipitation	mm
ET_P	Potential Evapotranspiration	mm
AI	Aridity index	-
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

approach can be written as

$$\mu(s,t) = \sum_{i=1}^{m} \beta_i(s) f_i(t). \tag{3}$$

The $f_i(t)$ are smoothed empirical orthogonal functions, which are spatially weighted by regression coefficients (β_i), so that the temporal structure can vary in space (Lindström et al., 2014). The number of EOFs is given by m, whereas $f_1(t)$ is always an intercept term. In this study, we compare four different methods for estimating the mean field. First, we will use the basic approach from Szpiro et al. (2010), by estimating the mean field with spatially weighted smoothed EOFs. This approach will be referred to as EOF_{simple}, and will serve as a benchmark for the other three methods. The second and third method use a model-based boosting approach for estimating the mean field. One implementation will only estimate the seasonal cycle of low-flow at each station (Boost_{SC}), while the other implementation will further include the long-term trend of low-flow at each station (Boost_{ST}). Finally, we combine the two approaches Boost_{SC} and EOF_{simple}, by first

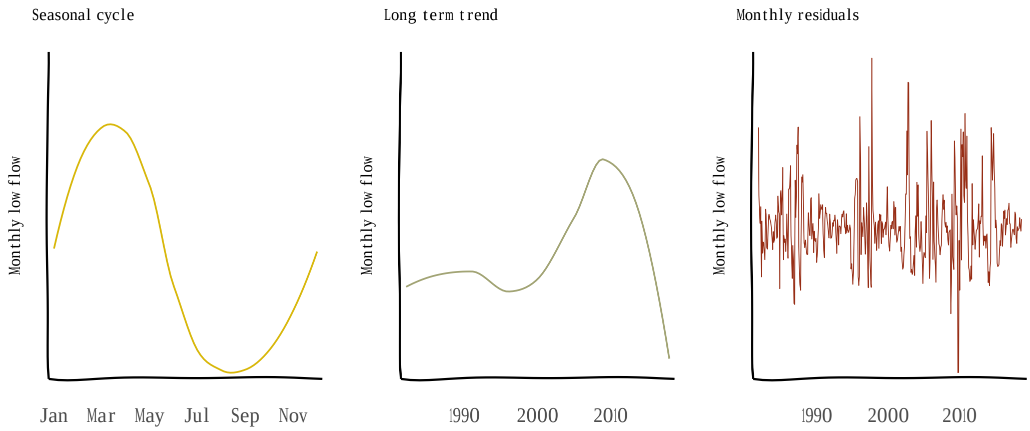

Figure 2. Model structure, exemplified for monthly Q95.

predicting the seasonal cycle and using the residuals for estimating the long-term effect by spatially weighted EOFs (Boost_{EOF}).

3.2.1 Smoothed empirical orthogonal functions

In our hierarchical model framework EOFs are used for estimating the mean field (EOF_{simple}), and in combination with seasonal boosting (Boost_{EOF}). In both cases, the first step is to build a matrix $\mathbf{x}_{EOF}$ ($T \times S$), where each column either corresponds to the monthly low-flow (EOF_{simple}), or to the residuals (Boost_{EOF}) at station s. The dimension T ($T = 444$) is the length of each monthly low-flow series at each station, and S ($S = 260$) is the number of stations. The matrix $\mathbf{x}_{EOF}$ is centered and scaled before applying a singular value decomposition. The smoothed EOFs are then calculated by fitting a spline on each singular value vector.

The number of EOFs (m) is determined by fitting a linear model (as in Eq. 3, for each s) to each column of $\mathbf{x}_{EOF}$ against m EOFs, where m is ranging from 1 (only an intercept term) to a maximum of 50 EOFs. For each single model the Bayesian information criterion (BIC) is calculated and averaged over all stations, resulting in a vector of BIC values (BIC_m) for each number of EOFs. As this approach would give only one realization for the entire set of stations and thereby lead to overfitting, we perform a bootstrap procedure with 25 repetitions (where a fraction of 70 % of the stations is sampled) to optimize the parameter m for the prediction at ungauged sites. The final number of EOFs is then determined by averaging every BIC_m over all 25 bootstrap samples and for a more parsimonious model we add 1 standard deviation to the minimized BIC value, which then serves as the threshold. The minimum number of EOFs with an average BIC below this threshold are then selected as the final number of EOFs (m). A graphical description of this selection is shown in Fig. 3. For any number of EOFs, f_1 is an intercept term which is a vector of 1s, with length S.

The f_i are then weighted in space by the regression coefficients β_i, where each β_i is a vector of regression coefficients for all stations. To obtain predictions at ungauged locations, every β_i is estimated by a linear model, which can be formulated as

$$\beta_i = \alpha_{0i} + \sum_{j=1}^{J} \mathbf{x_j} \alpha_{ij}, \tag{4}$$

where α_{0i} is an intercept term, $\mathbf{x}$ is the matrix of the spatial predictors presented in Sect. 2, and α_i are regression coefficients. J is the number of predictor variables that needs to be optimized. As it is a priori not clear which variables to include in the β_i-regression

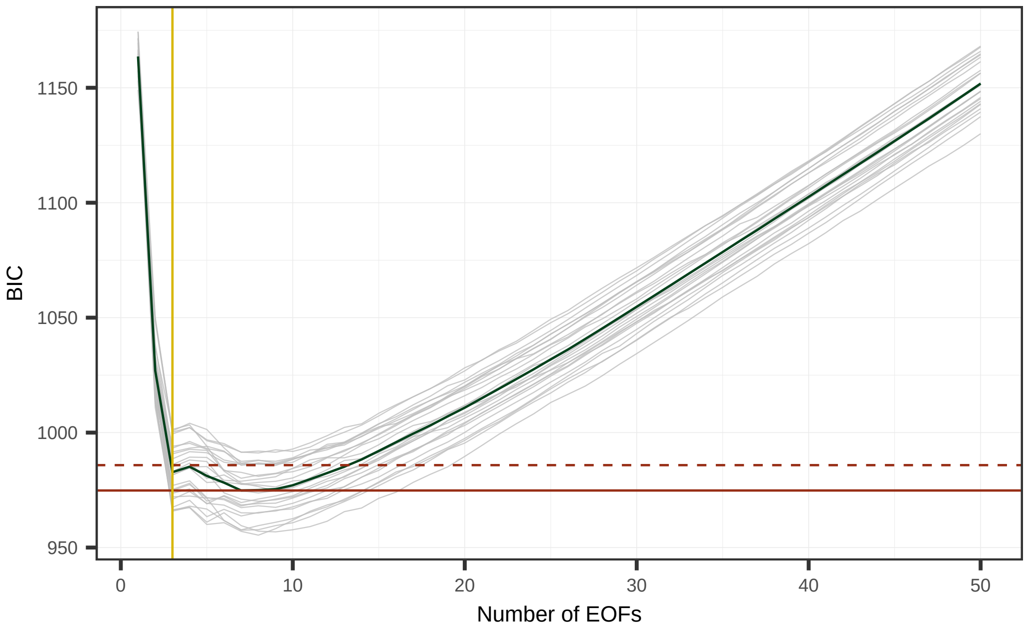

Figure 3. The number of EOFs are selected by a bootstrap procedure - with 25 samples. For each of the bootstrap samples the average BIC is calclulated. The number of EOFs is selected (the yellow line indicates the number of EOFs) by adding 1 standard deviation (shown by the red dashed line) to the minimum BIC value (shown by the red solid line).

model, possible approaches are to use shrinkage approaches as Lasso (Mercer et al., 2011), or dimension reduction methods as partial least-squares (Sampson et al., 2011, PLS). In this study we apply an approach that has already been shown to be useful for low-flow estimation in Austria (Laimighofer et al., 2022b). The variable selection is based on a recursive feature elimination (Granitto et al., 2006, RFE), which consists of an initial variable ranking and a backward variable selection. The initial variable ranking is estimated by a linear model-based boosting approach (a description of model based-boosting follows in Sect. 3.2.2). The variables are ranked after their absolute coefficients, and to obtain more robust results, the variable ranking is repeated 25-times by bootstrapping. For each β_i, a linear model is fitted to the first p ($p = 1, 2, 3, ..., 59$) ranked variables and the error is calculated and averaged over 25-bootstrap samples. The final number of variables (J) is defined by using 1.05 times the minimum error as a threshold to produce parsimonious models. The variable selection is performed for each β_i individually.

3.2.2 Model-based boosting

Model-based boosting (Bühlmann & Hothorn, 2007) is an iterative algorithm, where in each step a baselearner is selected, which best minimizes a predefined loss function (squared error in this study). To avoid overfitting the boosting algorithm uses a learning rate, to slowly approximate the final coefficients of the model. A baselearner can be e.g. a linear, a non-linear, random or spatial effect. Model-based boosting provides an intrinsic variable selection (Hofner et al., 2011), supports penalization of the effects and is robust against multicollinearity (Mayr & Hofner, 2018). The only parameter of the model that was tuned in this study was the number of boosting iterations, which was optimized using a 10-fold cross validation (CV) approach.

Based on this framework, the model for seasonal boosting (Boost_{SC}) can be formulated as

$$\mu(s,t) = \beta_0 + f_1(month) + \sum_{k=2}^{K} f_k(\mathbf{x}) + f_1(month)\mathbf{x}. \tag{5}$$

The model captures the average monthly low-flow regime. In the equation, β_0 is the intercept of the model and $\mathbf{x}$ is the predictor matrix with all spatial predictor variables. The spatial predictors can be parameterized by $f_k(.)$, either as a linear or a non-linear effect. We decomposed all non-linear effects into a linear and a non-linear part, as proposed by Kneib et al. (2009), to distinguish between linear and non-linear effects for each spatial variable. Further, a cyclic B-spline $f_1(month)$ according to Hofner et al. (2016) was added, which should represent the seasonal cycle of monthly low-flow. Finally, the term $f_1(month)\mathbf{x}$ was added to allow the seasonal cycle to vary in predictor variable space, in analogy to a varying-coefficient model (Hastie & Tibshirani, 1993; Fahrmeir et al., 2004). This leads to a total of $3p+1$ (178) baselearners for Boost_{SC}. For faster computation this model was not fitted to the full data, but only to the monthly averages at each station (seasonal low-flow cycle with 12 values per station).

In case of our spatio-temporal boosting approach (Boost_{ST}), the before mentioned model is extended by a long-term trend component to account for a transient seasonal low-flow regime. This trend component is captured by adding a sequence of all months ($T = 1, 2, 3, ..., 444$) as effect $f_2(time)$ to the model, which then can be written as

$$\mu(s,t) = \beta_0 + f_1(month) + f_2(time) + \sum_{k=3}^{K} f_k(\mathbf{x}) + f_1(month)\mathbf{x} + f_2(time)\mathbf{x}. \tag{6}$$

The long-term trend is modeled by a constant term and a spatially varying term, as it is done for the seasonal cycle. This results in $4p+2$ (238) baselearners for Boost_{ST}.

3.3 Residual field

Following Szpiro et al. (2010) and Lindström et al. (2014), the residual field $\upsilon(s,t)$ is estimated by a kriging structure,

$$\hat{\upsilon}_{st} = \sum_{s=1}^{S} \lambda_s \upsilon_{st}, \tag{7}$$

where υ_{st} are the fitted residuals at location s and time t and λ_s are the kriging weights. Note that the kriging weights (λ_s) are static over all timepoints. Hence, only one variogram model is used across time. The original approach employs an ordinary kriging estimator that is based spatial proximity, which appears well suited for air-quality models, in which context the proposed model was first introduced (Szpiro et al., 2010; Sampson et al., 2011; Lindström et al., 2014).

Considering river discharge, geographic kriging may not be fully appropriate, as it does not include the nested structure of catchments. Therefore, we estimate the residual field not only by ordinary kriging (OK), but additionally use physiographic kriging (PK) and topkriging (TK). Physiographic kriging was introduced by Castiglioni et al. (2011) and computes the first two principal components (PC) on a set of catchment characteristics. These two PCs then span up the physiographic space for the kriging structure. Topkriging (Skøien et al., 2006; Laaha et al., 2014) takes into account not only the size and distance of the catchments, but also their nested structure along the river network. This makes the method particularly well suited for interpolation of river discharge.

To obtain the kriging weights λ_s, we need to estimate a variogram model for each of the three kriging approaches. Lindström et al. (2014) proposes to use a maximum likelihood approach for estimation of the residual field, that includes variogram estimation. As it is not straightforward to estimate a topkriging variogram through a maximum likelihood approach, we introduce a simple framework for the optimization of the variogram for all three kriging methods. The procedure starts by calculating the coefficient of determination R_t^2 for every timestep:

$$R_t^2 = 1 - \frac{\sum_{s=1}^{S}(y_{st} - \hat{\mu}_{st})^2}{\sum_{s=1}^{S}(y_{st} - \overline{y_t})^2}, \tag{8}$$

where $\hat{\mu}_{st}$ at this point is the prediction of one of the four models for the mean field, y_{st} are the observations, and $\overline{y_t}$ is the spatial average at the specific timepoint. If only a kriging approach is used alone (no estimation of the mean field), $\hat{\mu}_{st}$ is simply the average low-flow at every station. Next, we compute the average $\overline{R_t^2}$ over all R_t^2 and select the timestep (t_{step}) in which the deviation of R_t^2 is minimal to $\overline{R_t^2}$. The variogram is then optimized at the unique residual timeslice $\upsilon_{st_{step}}$. For the optimization of the variogram we use a 10-fold CV and a grid search over the parameter space. For each combination of the parameters the R_{CV}^2 of $\upsilon_{st_{step}}$ is calculated and the parameters with the highest R^2 are used for the final prediction.

3.4 Model validation

Model evaluation is performed by a nested 10-fold cross validation (Varmuza & Filzmoser, 2016, CV). A nested CV consists of an inner and an outer loop. The inner loop in this study is used for tuning the boosting model, select the number of EOFs, variable selection for the regression coefficients of the EOFs, and optimizing the variogram parameters. The outer loop is solely used for assessing the model performance. This nested CV-scheme was already applied in two studies for low-flow in Austria (Laimighofer et al., 2022a, 2022b). However, in this study some parts of the inner loop are altered, due to the hierarchical structure of the model. An illustration of the scheme is given in Fig. 4.

3.4.1 Performance metrics

We assess performance using three main metrics. They are calculated using cross validated predictions and should therefore provide an unbiased estimate of the model error. First, we compute the root mean squared error (RMSE) by

$$RMSE = \sqrt{1/N \sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{9}$$

where N is the total number of observations ($N = S * T$), y_i are the observations and $\hat{y}_i$ are the predictions. Further, we calculate the median absolute error (MDAE):

$$MDAE = median(|y_i - \hat{y}_i|), \tag{10}$$

and the R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y_i})^2}. \tag{11}$$

The RMSE, MDAE and R^2 are computed based on all data points. Since we are particularly interested in how well the models can reproduce the mean field and thus provide an estimate of the mean seasonal low-flow regime, we additionally calculate all three metrics ($RMSE_{month}$, $MDAE_{month}$, R^2_{month}) for the seasonal predictions. This is shown exemplarily for the RMSE:

$$RMSE_{month} = \sqrt{1/SM \sum_{s=1}^{S}\sum_{m=1}^{M}(\overline{y}_{sm} - \overline{\hat{y}}_{sm})^2}, \tag{12}$$

where M is the number of months (12) and $\overline{y}_{sm}$ is:

$$\overline{y}_{sm} = 1/(N/M)\sum_{m=1}^{M} y_{sm}. \tag{13}$$

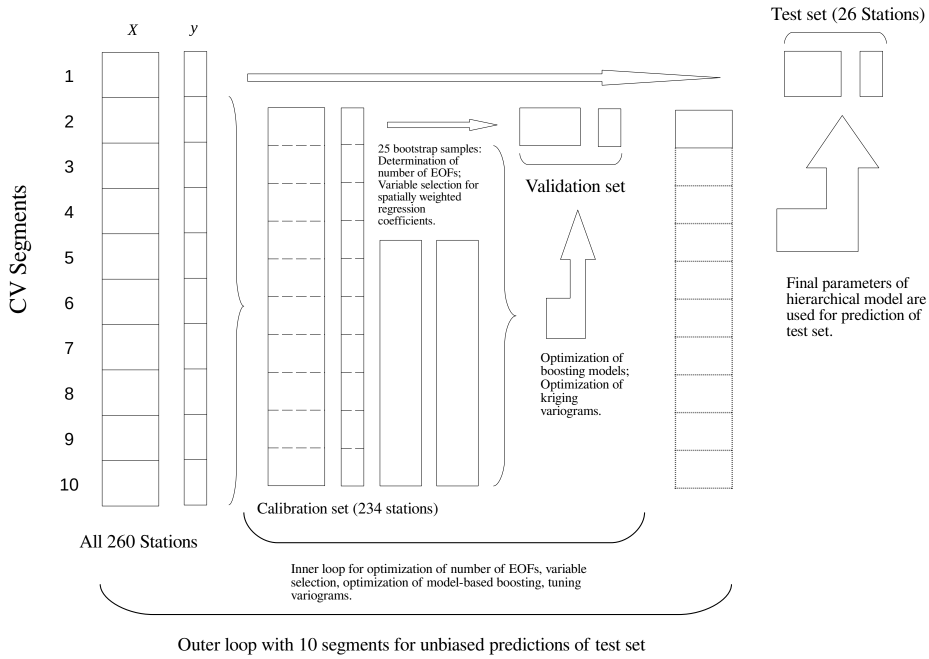

Figure 4. Schematic overview of the nested CV, that is used for model validation. We use different bootstrap samples for determination of number of EOFs and the selection of the β_i. Additionally the inner 10-fold CV is altered between optimization of the boosting models and the optimization of the variograms.

The ratio N/M can also be specified by the number of years (37 years of observations) at each station, and y_{sm} is every monthly low-flow value in month m at station s. The equation can be written accordingly for the predictions $\widehat{y}_{sm}$. Finally, we are interested in the performance of our models at each station, hence the R^2 is calculated for each station separately. Note that the equation of R^2 is equivalent to the formulation of the Nash–Sutcliffe efficiency (NSE, including the bias) in many hydrological studies (Blöschl et al., 2013).

4 Results

4.1 Mean field model components

Before proceeding with an overview of model performance, we shortly discuss some intrinsic features of the individual mean field model components - the inherent variable selection for the two boosting approaches, the weighted coefficients for the empirical orthogonal functions, and the determination of the number of EOFs.

4.1.1 Seasonal and spatiotemporal boosting

Model-based boosting includes an inherent variable selection procedure. Hence, we can analyse the selected variables and compare the structure of the two boosting models - seasonal boosting (Boost$_{SC}$) and spatiotemporal boosting (Boost$_{ST}$). Both boosting approaches used the maximum number of boosting steps over all ten folds that were predefined for each model (3000 for Boost$_{SC}$, 5000 for Boost$_{ST}$). In the seasonal boosting approach 45 baselearners were added on average to the model, whereas Boost$_{ST}$ exploited 67 baselearners on average. For both models the monthly cyclic spline ($f_1(month)$) was the most important variable. In both cases spatial covariabes were not added as single linear or non-linear baselearners, but solely as interaction effect of the cyclic spline, or the long-term trend. Figure 5 displays a graphical overview of the most important interaction effects for Boost$_{SC}$. The main important spatial predictors for Boost$_{SC}$ and Boost$_{ST}$ were topographic variables such as average catchment altitude and stream network density, landuse variables such as the fraction of wasteland, grassland and forest and, finally, meteorological conditions such as summer precipitation or snowmelt in winter. The long-term trend in the Boost$_{ST}$ model was added as a linear and non-linear effect and also weighted by spatial variables, but was generally negligible over all folds.

4.1.2 Smoothed empirical orthogonal functions

The smoothed empirical orthogonal functions (EOFs) were used as a single spatiotemporal framework (EOF$_{simple}$) and in combination with the seasonal boosting approach, where the EOFs (Boost$_{EOF}$) were estimated on the residuals of the seasonal predictions. In both cases, the number of EOFs were selected by a bootstrap procedure. EOF$_{simple}$ used 5 EOFs over all ten folds (Fig. 3 shows the selection of the number of EOFs), whereas the number of EOFs was slightly higher for the Boost$_{EOF}$ approach, ranging from 6 to 8 EOFs.

The EOFs were weighted by the meteorological, geological, landuse and topographic predictor variables in space. Our initially described variable selection (Sect. 3.2.1) reduced the number of variables for EOF$_{simple}$ to 6 predictors for the intercept term and 7 to 23 variables (from 59) for the other EOFs. In contrast, the Boost$_{EOF}$ approach produced more parsimonious models with only 2 spatial variables for the intercept, and 2 to 18 variables for the other EOFs.

Interpreting the selected variables is only straightforward in the case of the weighted intercept, which can be described as the mean low-flow for every station. This also explains the low number of variables used in the Boost$_{EOF}$ method, where the mean low-flow should already have been approximated by the seasonal boosting model. The left-over variables were the fraction of quarternary sediments, source region or stream net-

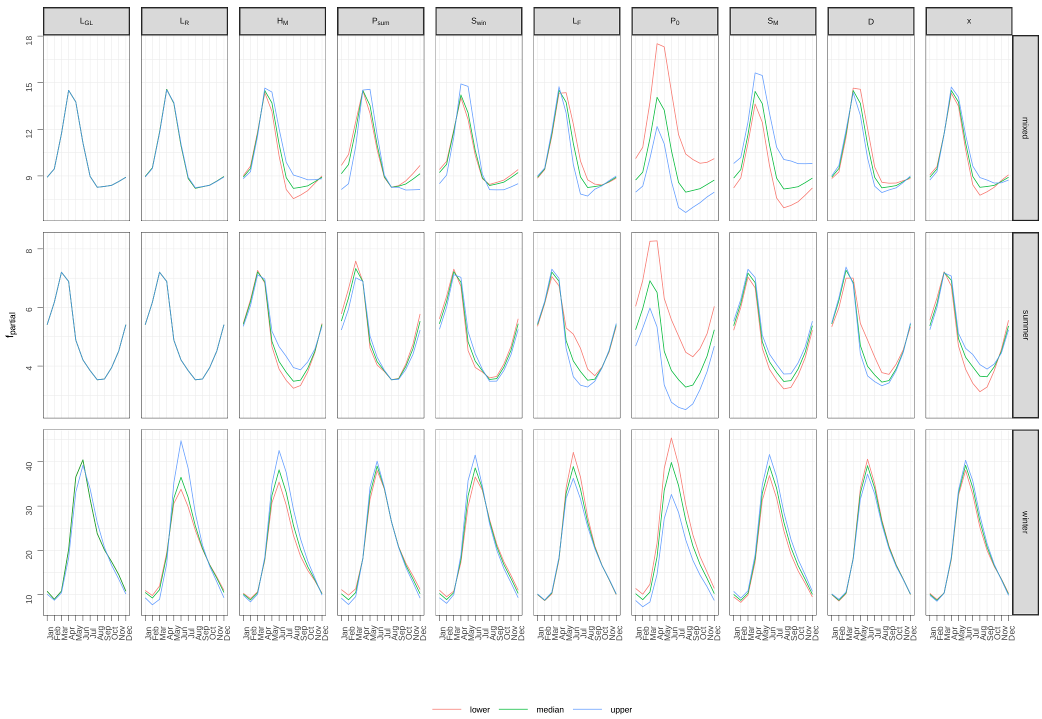

Figure 5. Partial predictions of the mean field with interaction effects of spatial predictors and the cyclic spline in the Boost$_{SC}$ model, stratified by low-flow regime type. Shown are the partial predictions for the 20%, 50% and 80% quantile of each spatial predictor variable within the considered regime type. The variable with the highest range in the spline coefficients is shown on the left, with a decreasing range to the right. Only the ten most important spatial variables are shown. As each fold leds to different results, the underlying model is the equivalent to the model produced by the first cross-validation run.

work density. The intercept of EOF$_{simple}$ was mainly modeled by topographic descriptors as maximum and average catchment altitude, average slope and meteorological conditions as the aridity index in summer and days without precipitation in summer.

4.2 Model performance

This section assesses model performance from different perspectives. In a first step, we investigate how well the mean seasonal low-flow regime is represented by the individual mean field models. This is followed by an analysis of the predictive performance of the individual components of the hierarchical model, i.e. the four mean-field models and the three kriging approaches when they are used on their own. Finally, we evaluate the full hierarchical models composed of these components.

4.2.1 Representation of the seasonal low-flow regime

Table 2. Three different error measures (RMSE, R^2, MDAE) for the mean seasonal low-flow regime are presented. For the calculation the predicted and observed low-flow is averaged for each month and station.

Model structure	R^2_{month}	RMSE$_{month}$	MDAE$_{month}$
Boost$_{SC}$	0.84	5.76	1.56
Boost$_{ST}$	0.82	6.15	1.65
EOF$_{simple}$	0.74	7.47	1.87
Boost$_{EOF}$	0.85	5.70	1.57

In a first step of assessing model performance, we evaluate the four approaches used for modelling the mean field and how well they can estimate the seasonal low-flow regimes across Austria. Table 2 presents the RMSE$_{month}$, the R^2_{month} and the MDAE$_{month}$ for the four approaches. Generally, the seasonal low-flow regime was well predicted by all four methods, but the EOF$_{simple}$ approach showed a weaker performance on all three error metric with a RMSE$_{month}$ of 7.47, compared to 6.15 (Boost$_{ST}$) and 5.76 (Boost$_{SC}$) for the two boosting approaches. The best performance was reached by the stacked model of seasonal boosting and the use of EOFs for the residuals (RMSE$_{month}$ = 5.7), albeit the differences to the seasonal boosting approach is almost negligible and also the spatiotemporal boosting approach yields only slightly weaker performance metrics.

Examining the estimates for the three different regime types (Fig. 6) gives a more detailed picture of model performance. The weaker performance of EOF$_{simple}$ was apparent for all three regime types, with a R^2_{month} ranging from 0.59 to 0.66, but is negligible for the mixed low-flow regime. EOF$_{simple}$ resulted in smoother estimates of the seasonal cycle, which probably led to the lower performance especially for the summer and winter regime. Assessing the performance of the three other methods, the winter regime was best predicted with a R^2_{month} ranging from 0.78 to 0.81. For the summer regime the R^2_{month} was between 0.74 to 0.76, where for the mixed regime it dropped to 0.62 (0.6 for Boost$_{ST}$).

4.2.2 Performance of individual components

In a next step of model evaluation we assess the individual performances of the components of the hierarchical model framework: models for the mean field and the sole use of the three different kriging structures without considering the mean field. Table 3 gives an overview of the results. The individual model components yielded a RMSE of 8.42 (Boost$_{EOF}$) to 9.79 (EOF$_{simple}$). What is striking is that the spatiotemporal boosting

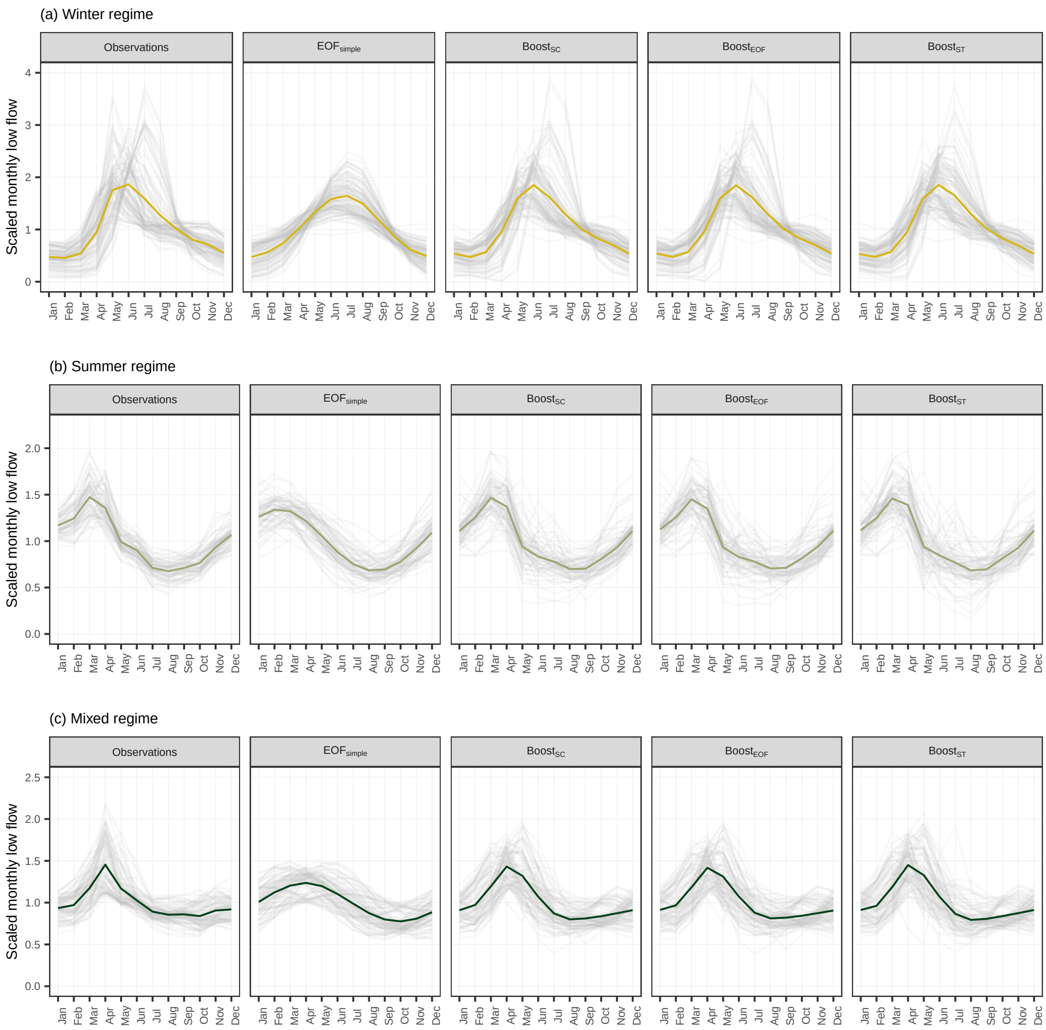

Figure 6. Predictions of the mean seasonal low-flow regime by various mean field models, stratified by regime type. The seasonal low-flow cycle is scaled by the mean at each station, for a better visualization. Each transparent line presents the seasonal cycle of one station, where the colored thick line is the average over all stations.

(Boost_{ST}) approach has a weaker overall performance on all metrics compare to the seasonal boosting approach (Boost_{SC}). Both approaches only yield a $R^2_{0.5}$ of 0.15. In case of Boost_{SC} this is not surprising, as the model can only capture the seasonal cycle at each station. However, the additional long-term trend in the Boost_{ST} approach is only adding noise to the model and leads to no improvement in terms of model performance. The long-term trend is better approximated by the stacked model of seasonal boosting and EOFs (Boost_{EOF}), which obtain the best results comparing the four mean field components. Generally, all three kriging approaches yielded a higher $R^2_{0.5}$, ranging from 0.48 for physiographic kriging (PK), to 0.63 for ordinary kriging (OK) and 0.75 for topkriging (TK). These results show, that TK already provides very accurate predictions without taking any spatio-temporal information into account.

Table 3. Overview of the error for the individual model components. R^2, RMSE and MDAE refer to the performance for all data points. $R^2 < 0.5$ ($R^2 < 0$) is the fraction of stations that yield a R^2 below 0.5 (0) and $R^2_{0.5}$ is the median of all R^2 computed per each station.

Model structure	R^2	RMSE	MDAE	$R^2 < 0.5$	$R^2 < 0$	$R^2_{0.5}$
Boost_{SC}	0.68	9.06	2.59	0.77	0.33	0.15
Boost_{ST}	0.67	9.29	2.61	0.79	0.34	0.15
EOF_{simple}	0.63	9.79	2.49	0.77	0.18	0.37
Boost_{EOF}	0.73	8.42	2.24	0.64	0.17	0.41
OK	0.75	8.02	2.09	0.40	0.22	0.63
PK	0.69	8.96	2.34	0.52	0.26	0.48
TK	0.80	7.25	1.62	0.30	0.15	0.75

4.2.3 Performance of hierarchical models

Table 4. Overview of the overall error for all hierarchical models. The Kriging column identifies the kriging approach which was used for the residual field. The Mean field column distinguishes between the different approaches for estimating the mean field of the model.

Kriging	Mean field	R^2	RMSE	MDAE	$R^2 < 0.5$	$R^2 < 0$	$R^2_{0.5}$
OK	Boost_{SC}	0.83	6.72	1.78	0.30	0.10	0.69
OK	EOF_{simple}	0.81	6.99	1.86	0.31	0.11	0.67
OK	Boost_{EOF}	0.82	6.77	1.79	0.32	0.10	0.68
OK	Boost_{ST}	0.81	6.96	1.86	0.33	0.11	0.66
PK	Boost_{SC}	0.79	7.37	1.94	0.41	0.13	0.58
PK	EOF_{simple}	0.77	7.73	2.00	0.38	0.15	0.59
PK	Boost_{EOF}	0.79	7.42	1.92	0.39	0.13	0.58
PK	Boost_{ST}	0.77	7.77	1.99	0.43	0.13	0.56
TK	Boost_{SC}	0.84	6.35	1.56	0.25	0.09	0.73
TK	EOF_{simple}	0.83	6.56	1.60	0.25	0.09	0.73
TK	Boost_{EOF}	0.84	6.35	1.60	0.24	0.09	0.72
TK	Boost_{ST}	0.84	6.44	1.64	0.27	0.08	0.70

In a next step we want to assess the prediction performance of the full hierarchical models that combine the component models evaluated before. Table 4 gives an overview of the cross-validated error of all models. We can observe that the application of differ-

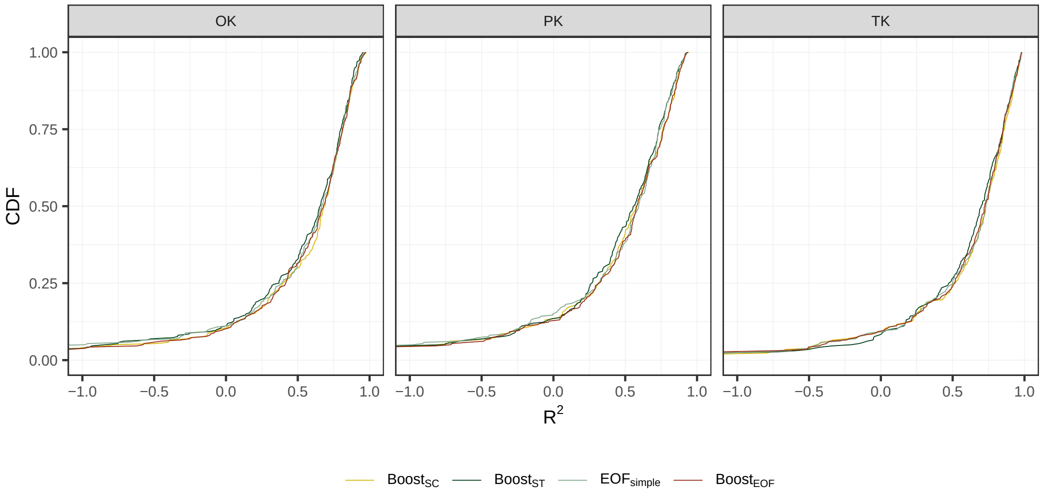

Figure 7. Cumulative distribution of station-wise R^2 stratified by kriging-method. Stations with a R^2 below -1 are omitted for clarity.

ent kriging methods led to the main variation in model performance, with a better performance of TK than OK and PK. The model combinations with TK yielded a RMSE from 6.35 to 6.56, whereas OK resulted in a somewhat higher RMSE between 6.72 and 6.99 and the use of PK for the residual field led to a RMSE of 7.37 to 7.77. This overall trend was also visible for all other performance measures.

In contrast, the different approaches for estimating the mean field only slightly altered the prediction performance of the models. For all kriging approaches the use of seasonal boosting, or Boost_{EOF} yielded similar results. The models showed a somewhat weaker performance when the mean field was estimated by spatiotemporal boosting or EOF_{simple}, but when we look at the distribution of the R^2 over all stations (Fig. 7), these differences almost disappear. For instance, the $R^2_{0.5}$ for topkriging ranged only from 0.7 to 0.73 and the number of low-performing stations with a R^2 below 0 was between 8 % and 9 %.

4.2.4 Performance of hierarchical models grouped by seasonal regime types

For a deeper performance analysis of the hierarchical models, we again focus on the three low-flow regime types (winter, summer, mixed). Figure 8 gives an overview of the distribution of the R^2 for all three regimes. Regarding the kriging structure, hierarchical model with TK show the best performance over all three regimes. Highest prediction accuracy is reached for winter regime, where hierarchical models with TK yield a median R^2 of 0.8 to 0.84. Performance of OK is only slightly lower with a median R^2 from 0.78 to 0.81, but only 0.72 to 0.76 for PK. The performance is somewhat smaller for summer regimes for all models, and is lowest for mixed regimes, where combinations with TK still reach a median R^2 of 0.68 (lowest R^2 of 0.62), but median R^2 values for OK are only ranging from 0.5 to 0.57.

A further stratification of the results by the mean field model did not reveal a systematic picture of the performance. For example, EOF_{simple} in combination with OK, resulted in the worst performance for summer regimes, but for physiographic kriging and topkriging the combination with EOF_{simple} led to the best performance. Focusing on the mixed regime, the Boost_{ST} method seemed to be disadvantageous for all kriging structures, but this was not apparent in the results of the winter or summer regime.

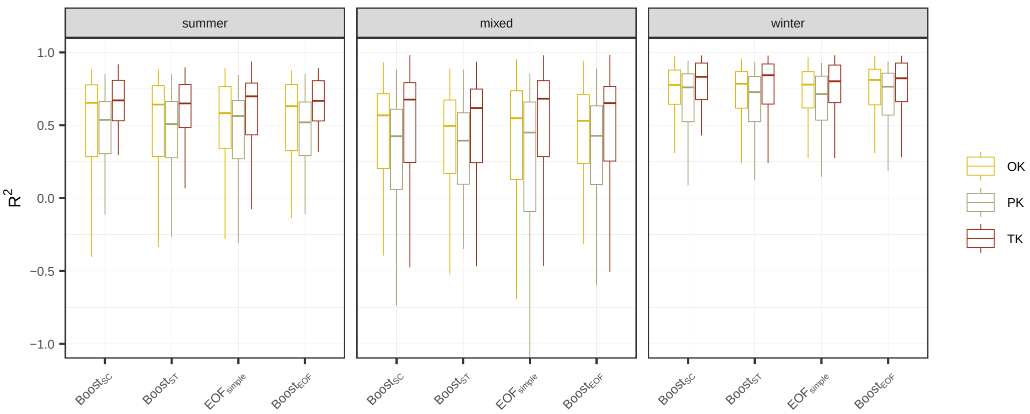

Figure 8. Comparison of the overall performance of the hierarchical models stratified by kriging method and regime type. Each boxplot shows the distribution of the R^2 over all stations. Outliers are removed from the plot for better visualization.

5 Discussion

5.1 Comparison of performance

In this paper, we extended an existing hierarchical model, initially proposed by Szpiro et al. (2010), for performing spatio-temporal predictions of monthly low-flow index series in Austria. We tested four models to approximate the seasonal cycle and the long-term trend, and compared three geostatistical approaches for the residual field. Comparison to existing literature is mainly limited to the study by Laimighofer et al. (2022a), where results can directly be compared as stations, temporal resolution, and even the used cross validation folds are equivalent to this study. In Laimighofer et al. (2022a) a single spatio-temporal framework was applied, where the best model yielded a median R^2 of 0.67 and an overall RMSE of 6.98. In this study these measures could be improved to a RMSE of 6.35 and a median R^2 of 0.73, for our best model ($Boost_{SC}$ and TK). Performance comparison to other literature is somehow difficult, as prediction studies on monthly streamflow data is mainly performed on monthly mean values and results are partially not evaluated by cross validation (Gottschalk et al., 2015; Sauquet et al., 2008; Pumo et al., 2016), which can best capture the error of prediction in ungauged basins.

In a more qualitative embedding of our results, we can highlight that hierarchical model combinations with topkriging yield the highest prediction accuracy. This is in line with studies for spatial low-flow prediction (Laaha et al., 2014), or spatio-temporal streamflow prediction in Austria (Skøien & Blöschl, 2007; Viglione et al., 2013), where also TK reaches high prediction performance. In contrast, Farmer (2016) shows that OK can perform as well as TK in a spatio-temporal framework, and suggests that ordinary kriging should be preferred over TK, due to the lower model complexity. Our results could paint a similar picture, as the performance metrics are only slightly improved by TK, but this is only true if we consider the full hierarchical model structure, where the between-model differences are reduced. Studies as Farmer (2016) or Skøien and Blöschl (2007) considered no additional seasonal cycle or long-term trend in their models. Focusing on our results for a single kriging structure (Table 3), the median R^2 for OK is only 0.63, but the median R^2 for TK is 0.75. However, the single TK approach only yields a RMSE of 7.25, which is substantially higher to the RMSE of 6.35 of the combination of $Boost_{SC}$ and topkriging. We will discuss these performance issues of topkriging in more depth in the next section.

Prediction accuracy of PK is generally lower for all hierarchical model combinations and for the single kriging approach. Results for spatial low-flow prediction in Italy (Castiglioni et al., 2011) showed similar performance of PK and TK, but this is not reflected in our space-time framework. The lower performance of PK may be caused by the similar information used by the mean field models and the first two principal components covering the physiographic space for PK.

5.2 Effect of headwater vs. non-headwater on topkriging performance

Albeit, several studies demonstrated the good performance of topkriging (Skøien & Blöschl, 2007; de Lavenne et al., 2016; Laaha et al., 2014; Farmer, 2016; Viglione et al., 2013), accuracy of TK is altered as a function of catchment area (Viglione et al., 2013), station density (Parajka et al., 2015), or the hierarchical position in the river network (Laaha et al., 2014; de Lavenne et al., 2016). Laaha et al. (2014) found that the R^2 for TK in headwater catchments for spatial low-flow prediction is 0.59, whereas in non-headwater catchments performance increased to a R^2 of 0.91. A similar trend was shown by de Lavenne et al. (2016), where the performance of TK increased with higher Strahler order. This is consistent with our results (displayed in Fig. 9), where we can see a general trend for all model combinations that a higher Strahler order increases the prediction performance. Considering the performance of each model combination, we observe that a simple topkriging routine is not sufficient for headwater catchments (Strahler order 1 - 2). For example the median R^2 for simple TK is 0.56 for catchments with a Strahler order 1. Adding seasonal predictions (Boost_{SC}) to the model structure enhances prediction to a median R^2 of 0.67. Differences between the models almost disappear when considering catchments with Strahler order 2. Here the median R^2 is between 0.67 and 0.7, but simple TK shows a much higher variance in the results. In catchments with a Strahler order of 3 or more, the simple TK routine provides the most accurate predictions compared to the hierarchical model combinations. However, we can show that the lower performance of topkriging in headwater catchments can be improved by a hierarchical framework that that exploits the seasonal cycle in advance.

5.3 Case study - extreme events

So far our model assessment focused on global model performance. In a last step, we want to consider a concrete discharge time series, to demonstrate the potential of our modeling approach. As our main interest is to predict low-flows we will focus on two drought years 2003 and 2015 (Ionita et al., 2017; Laaha et al., 2017). We selected the hydrograph Altschlaining at the river Tauchenbach in eastern Austria, which already was investigated by Laaha et al. (2017). The Tauchenbach is a small (upstream) catchment with 89.2 km^2, which experienced a particularly extreme low-flow event in 2003 (Fig. 10). The event of 2003 started with an early onset and continued over the whole year, whereas in 2015 wetter preconditions in spring led to a later onset and prevented a more severe low-flow event in summer.

The seasonal boosting approach in combination with TK yields a cross-validated R^2 of 0.45 at Altschlaining, which is lower than about 80 % of all stations. Nevertheless, the development of the low-flow events is captured quite well by model predictions, which can be decomposed to the mean field component and the residual field component. Figure 10 illustrates the complementary behaviour of these two components. In extreme events like 2003 and 2015, the observed low flows deviate strongly from the seasonal low-flow regime. For this reason, the mean field component of the hierarchical model would provide a biased estimate. The TK of the residual field, however, performs an adjustment of the predictions to the respective event conditions, as can be seen for both events. It uses synchronous information of adjacent stations to achieve enhanced space-time predictions. Such adjustment would indeed be much smaller in a 'normal' year, where the low-flow conditions are similar to the average regime.

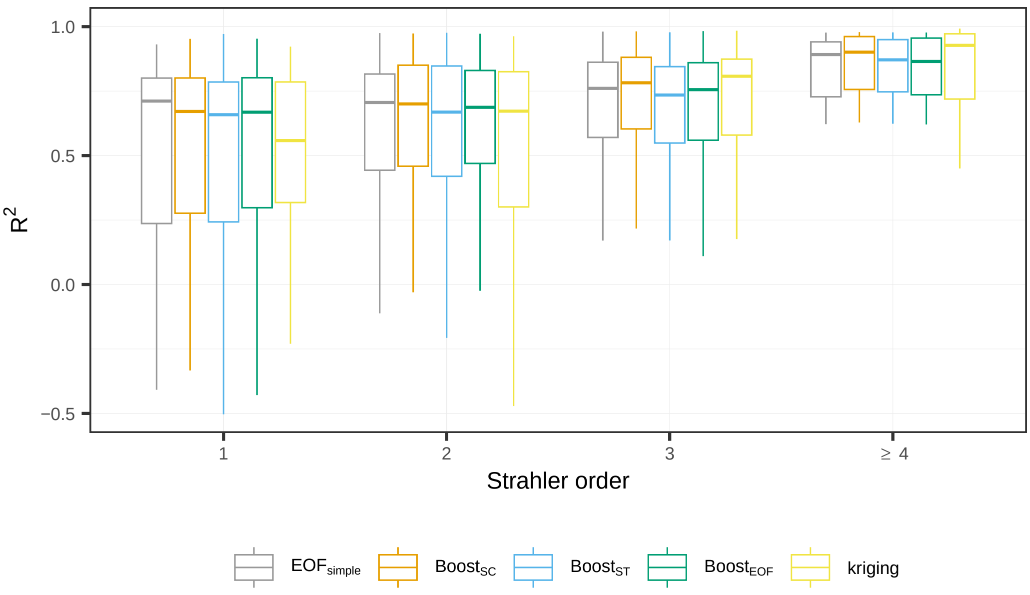

Figure 9. The boxplots show all possible estimation of the mean field in combination with topkriging, and a simple topkriging routine in which only one variogram is estimated for the full spatio-temporal domain. The catchments are further stratified by their Strahler order (x-axis). Due to the limited stations with Strahler order ≥ 4, these stations are condensed in one group.

Despite these favorable properties, some below-average performance can be observed in spring 2003, where discharges reflect the very dry preconditions that led to the severe low-flow event. This seasonal anomaly can be explained by a particular weather situation where the Tauchenbach experienced a precipitation deficit over several years due to lee-effects behind alpline and pre-alpine mountain ranges (Laaha et al., 2017). Since this is a local singularity, the anomaly cannot be adjusted by information from neighboring stations, so a residual TK does not significantly improve the estimates. Further on, the (regionally more consistent) atmospheric water deficit of the summer drought event gets increasingly important. This leads to enhanced residual TK, which is reflected in steadily improving predictions during the ongoing low-flow event.

6 Conclusions

In this study we adopted a hierarchical model framework for spatio-temporal modelling of monthly low-flow in Austria. The best performing model is a combination of model-based boosting for the mean field, which estimates the seasonal low-flow regime, and topkriging for predicting the residuals. It gives a median R^2 of 0.73 over all stations, demonstrating the high potential of the hierarchical model.

Generally, stations with a strong winter seasonality of low-flows show a higher prediction accuracy than summer or mixed regimes. The drivers of monthly low-flow in winter regime catchments are mainly high sums of precipitation and snowmelt in the summer months, and freezing and low sums of precipitation in the winter. The signal of monthly low-flow in mixed or summer regimes is more noisy, which slightly weakens the prediction performance.

Regardless of regime type or mean field methods used, topkriging shows the best performance for all model combinations, followed by ordinary kriging and physiographic kriging. It is striking that even a simple topkriging routine without an additional mean

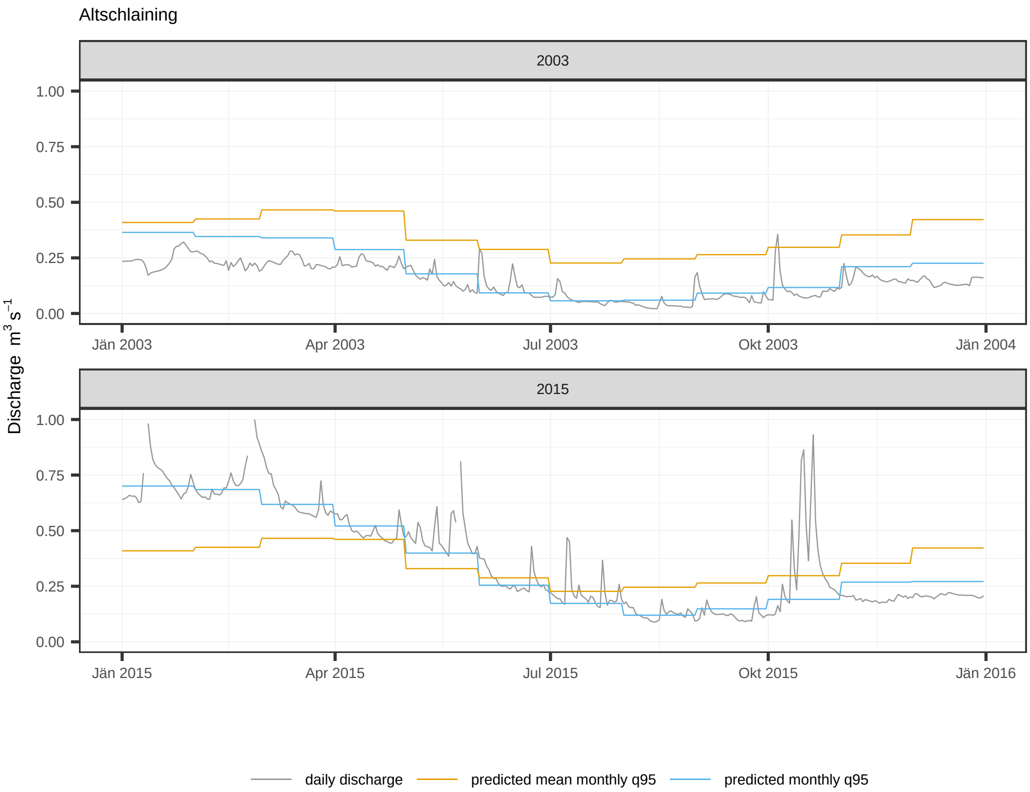

Figure 10. Comparison of two drought years (2003 and 2015), for the station Altschlaining, river Tauchenbach. Each plot shows the daily discharge, predicted mean monthly q95 and predicted monthly q95 - both are transformed back to discharge values ($m^3 s^{-1}$).

field achieves a median R^2 of 0.75, but has a higher number of poorly performing stations ($R^2 < 0.5$). It shows a lack of prediction accuracy, especially in headwater catchments. In these catchments the hierarchical model framework is particularly beneficial, whereas in catchments of Strahler order ≥ 3 the simple topkriging routine is sufficient.

Overall, the favorable performance of the model results from its specific structure, which seems well suited to combine different types of information: average low flow conditions estimated from climate and catchment characteristics, and information of neighbouring catchments estimated by spatial correlation. This combination provides accurate results not only for average years, where the high prediction accuracy for the seasonal low-flow regime comes into play, but also for extreme years, where top-kriging adapts to the anomalous conditions of the low-flow event and can also capture the preconditions. The model is shown to provide robust estimates for a range of conditions, including headwater catchments and extreme events. It demonstrates a high degree of suitability for predicting gaps in the low-flow series, and for providing estimates at ungauged sites.

7 Open Research

Modelling and data analysis was performed in R version 4.2.2 (R Core Team, 2022). We want to acknowledge the use of the following packages: caret (Kuhn, 2022), cubble (Zhang et al., 2022), gridExtra (Auguie, 2017), lubridate (Grolemund & Wickham, 2011), mboost (Hothorn et al., 2022), Metrics (Hamner & Frasco, 2018), rtop (Skoien et al., 2014), sf (Pebesma, 2018), tidyverse (Wickham et al., 2019), wesanderson (Ram & Wickham, 2018). Model output and code to produce the figures is available at zenodo (Laimighofer & Laaha, 2023).

Acknowledgments

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. This research has been supported by the Climate and Energy Fund under the programme "ACRP" (grant no. C265154).

The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC).

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).

References

Auguie, B. (2017). gridextra: Miscellaneous functions for "grid" graphics [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=gridExtra` (R package version 2.3)

Blöschl, G., Sivapalan, M., Wagener, T., Savenije, H., & Viglione, A. (2013). *Runoff prediction in ungauged basins: synthesis across processes, places and scales.* Cambridge University Press. doi: 10.1017/CBO9781139235761

Bühlmann, P., & Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, *22*(4), 477 – 505. doi: 10.1214/07-STS242

Castiglioni, S., Castellarin, A., & Montanari, A. (2009). Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. *Journal of hydrology*, *378*(3-4), 272–280. doi: 10.1016/j.jhydrol.2009.09.032

Castiglioni, S., Castellarin, A., Montanari, A., Skøien, J. O., Laaha, G., & Blöschl, G. (2011). Smooth regional estimation of low-flow indices: physiographi-

cal space based interpolation and top-kriging. *Hydrology and Earth System Sciences*, *15*(3), 715–727. doi: 10.5194/hess-15-715-2011
Cutore, P., Cristaudo, G., Campisano, A., Modica, C., Cancelliere, A., & Rossi, G. (2007). Regional models for the estimation of streamflow series in ungauged basins. *Water resources management*, *21*(5), 789–800. doi: 10.1007/s11269-006-9110-7
de Lavenne, A., Skøien, J. O., Cudennec, C., Curie, F., & Moatar, F. (2016). Transferring measured discharge time series: Large-scale comparison of top-kriging to geomorphology-based inverse modeling. *Water Resources Research*, *52*(7), 5555-5576. doi: https://doi.org/10.1002/2016WR018716
Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, *17*(5), 1893–1912. doi: 10.5194/hess-17-1893-2013
Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression for space-time data: a bayesian perspective. *Statistica Sinica*, 731–761.
Farmer, W. H. (2016). Ordinary kriging as a tool to estimate historical daily streamflow records. *Hydrology and Earth System Sciences*, *20*(7), 2721–2735. doi: 10.5194/hess-20-2721-2016
Fuentes, M., Guttorp, P., & Sampson, P. D. (2006). Using transforms to analyze space-time processes. *Monographs on Statistics and Applied Probability*, *107*, 77.
Gottschalk, L. (1993). Interpolation of runoff applying objective methods. *Stochastic hydrology and hydraulics*, *7*, 269–281.
Gottschalk, L., Krasovskaia, I., Dominguez, E., Caicedo, F., & Velasco, A. (2015). Interpolation of monthly runoff along rivers applying empirical orthogonal functions: Application to the upper magdalena river, colombia. *Journal of Hydrology*, *528*, 177–191.
Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and intelligent laboratory systems*, *83*(2), 83–90.
Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from `https://www.jstatsoft.org/v40/i03/`
Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *Wiley Interdisciplinary Reviews: Water*, *8*(1), e1487.
Hamner, B., & Frasco, M. (2018). Metrics: Evaluation metrics for machine learning [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=Metrics` (R package version 0.1.4)
Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(4), 757–779.
Hofner, B., Hothorn, T., Kneib, T., & Schmid, M. (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, *20*(4), 956–971.
Hofner, B., Kneib, T., & Hothorn, T. (2016). A unified framework of constrained regression. *Statistics and Computing*, *26*, 1–14.
Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2022). mboost: Model-based boosting [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=mboost` (R package version 2.9-7)
Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., ... others (2013). A decade of predictions in ungauged basins (pub)—a review. *Hydrological sciences journal*, *58*(6), 1198–1255.
Ionita, M., Tallaksen, L. M., Kingston, D. G., Stagge, J. H., Laaha, G., Van Lanen, H. A. J., ... Haslinger, K. (2017). The european 2015 drought from a clima-

tological perspective. *Hydrology and Earth System Sciences*, *21*(3), 1397–1419. doi: 10.5194/hess-21-1397-2017

Kneib, T., Hothorn, T., & Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics*, *65*(2), 626–634.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344–11354. doi: 10.1029/2019WR026065

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. doi: 10.5194/hess-23-5089-2019

Kuhn, M. (2022). caret: Classification and regression training [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=caret` (R package version 6.0-92)

Kyriakidis, P. C., & Journel, A. G. (1999). Geostatistical space–time models: a review. *Mathematical geology*, *31*, 651–684.

Laaha, G. (2023). A mixed distribution approach for low-flow frequency analysis – part 1: Concept, performance, and effect of seasonality. *Hydrology and Earth System Sciences*, *27*(3), 689–701. doi: 10.5194/hess-27-689-2023

Laaha, G., & Blöschl, G. (2006). Seasonality indices for regionalizing low flows. *Hydrological Processes*, *20*(18), 3851-3878. doi: https://doi.org/10.1002/hyp.6161

Laaha, G., & Blöschl, G. (2006). A comparison of low flow regionalisation methods—catchment grouping. *Journal of Hydrology*, *323*(1), 193-214. doi: https://doi.org/10.1016/j.jhydrol.2005.09.001

Laaha, G., Gauster, T., Tallaksen, L. M., Vidal, J.-P., Stahl, K., Prudhomme, C., . . . Wong, W. K. (2017). The european 2015 drought from a hydrological perspective. *Hydrology and Earth System Sciences*, *21*(6), 3001–3024. doi: 10.5194/hess-21-3001-2017

Laaha, G., Skøien, J., & Blöschl, G. (2014). Spatial prediction on river networks: comparison of top-kriging with regional regression. *Hydrological Processes*, *28*(2), 315-324. doi: https://doi.org/10.1002/hyp.9578

Laimighofer, J., & Laaha, G. (2023, June). *Code and model output to "Statistical learning and topkriging improve spatio-temporal low-flow estimation".* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.8007772` doi: 10.5281/zenodo.8007772

Laimighofer, J., Melcher, M., & Laaha, G. (2022a). Low-flow estimation beyond the mean–expectile loss and extreme gradient boosting for spatiotemporal low-flow prediction in austria. *Hydrology and Earth System Sciences*, *26*(17), 4553–4574.

Laimighofer, J., Melcher, M., & Laaha, G. (2022b). Parsimonious statistical learning models for low-flow estimation. *Hydrology and Earth System Sciences*, *26*(1), 129–148.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall–runoff models in great britain: a comparison of long short-term memory (lstm)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, *25*(10), 5517–5534. doi: 10.5194/hess-25-5517-2021

Li, L., Gottschalk, L., Krasovskaia, I., & Xiong, L. (2018). Conditioned empirical orthogonal functions for interpolation of runoff time series along rivers: Application to reconstruction of missing monthly records. *Journal of Hydrology*, *556*, 262-278. doi: https://doi.org/10.1016/j.jhydrol.2017.11.014

Lindstrom, J., Szpiro, A., Sampson, P. D., Bergen, S., & Oron, A. P. (2019). Spatiotemporal: Spatio-temporal model estimation [Computer software manual].

Retrieved from https://CRAN.R-project.org/package=SpatioTemporal (R package version 1.1.9.1)

Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., & Sheppard, L. (2014). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and ecological statistics*, *21*, 411–433.

Mayr, A., & Hofner, B. (2018). Boosting for statistical modelling-a non-technical introduction. *Statistical Modelling*, *18*(3-4), 365–384.

Mercer, L. D., Szpiro, A. A., Sheppard, L., Lindström, J., Adar, S. D., Allen, R. W., ... Kaufman, J. D. (2011). Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (nox) for the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Atmospheric Environment*, *45*(26), 4412-4420. doi: https://doi.org/10.1016/j.atmosenv.2011.05.043

Müller, M., & Thompson, S. (2015). Topreml: a topological restricted maximum likelihood approach to regionalize trended runoff signatures in stream networks. *Hydrology and Earth System Sciences*, *19*(6), 2925–2942.

Opitz, T., Bonneu, F., & Gabriel, E. (2020). Point-process based bayesian modeling of space–time structures of forest fire occurrences in mediterranean france. *Spatial Statistics*, *40*, 100429.

Parajka, J., Merz, R., Skøien, J. O., & Viglione, A. (2015). The role of station density for predicting daily runoff by top-kriging interpolation in austria. *Journal of Hydrology and Hydromechanics*, *63*(3), 228–234.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. Retrieved from https://doi.org/10.32614/RJ-2018-009 doi: 10.32614/RJ-2018-009

Pumo, D., Viola, F., & Noto, L. V. (2016). Generation of natural runoff monthly series at ungauged sites using a regional regressive model. *Water*, *8*(5), 209. doi: 10.3390/w8050209

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Ram, K., & Wickham, H. (2018). wesanderson: A wes anderson palette generator [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=wesanderson (R package version 0.3.6)

Razavi, T., & Coulibaly, P. (2013). Streamflow prediction in ungauged basins: review of regionalization methods. *Journal of hydrologic engineering*, *18*(8), 958–975.

Rodríguez-Iturbe, I., Isham, V., Cox, D. R., Manfreda, S., & Porporato, A. (2006). Space-time modeling of soil moisture: Stochastic rainfall forcing with heterogeneous vegetation. *Water Resources Research*, *42*(6). doi: https://doi.org/10.1029/2005WR004497

Salinas, J., Laaha, G., Rogger, M., Parajka, J., Viglione, A., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins–part 2: Flood and low flow studies. *Hydrology and Earth System Sciences*, *17*(7), 2637–2652. doi: 10.5194/hess-17-2637-2013

Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindström, J., & Kaufman, J. D. (2011). Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment*, *45*(36), 6593-6606. doi: https://doi.org/10.1016/j.atmosenv.2011.04.073

Sauquet, E., Gottschalk, L., & Krasovskaia, I. (2008, 10). Estimating mean monthly runoff at ungauged locations: an application to France. *Hydrology Research*, *39*(5-6), 403-423. doi: 10.2166/nh.2008.331

Sauquet, E., Gottschalk, L., & Leblois, E. (2000). Mapping average annual runoff: a hierarchical approach applying a stochastic interpolation scheme. *Hydrological*

sciences journal, *45*(6), 799–815.
Skøien, J. O., & Blöschl, G. (2007). Spatiotemporal topological kriging of runoff time series. *Water Resources Research*, *43*(9).
Skoien, J. O., G. Bloschl, G. Laaha, E. Pebesma, J. Parajka, & A. Viglione. (2014). Rtop: An r package for interpolation of data with a variable spatial support, with an example from river networks. *Computers & Geosciences*.
Skøien, J. O., Merz, R., & Blöschl, G. (2006). Top-kriging - geostatistics on stream networks. *Hydrology and Earth System Sciences*, *10*(2), 277–287. doi: 10.5194/hess-10-277-2006
Solomatine, D. P., & Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of hydroinformatics*, *10*(1), 3–22. doi: 10.2166/hydro.2008.015
Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S. D., & Kaufman, J. D. (2010). Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics*, *21*(6), 606-631. doi: https://doi.org/10.1002/env.1014
Tyralis, H., Papacharalampous, G., Langousis, A., & Papalexiou, S. M. (2021). Explanation and probabilistic prediction of hydrological signatures with statistical boosting algorithms. *Remote Sensing*, *13*(3), 333. doi: 10.3390/rs13030333
Varmuza, K., & Filzmoser, P. (2016). *Introduction to multivariate statistical analysis in chemometrics.* CRC press.
Viglione, A., Parajka, J., Rogger, M., Salinas, J. L., Laaha, G., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins; part 3: Runoff signatures in austria. *Hydrology and Earth System Sciences*, *17*(6), 2263–2279. doi: 10.5194/hess-17-2263-2013
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686
Wilby, R. L., Wigley, T., Conway, D., Jones, P., Hewitson, B., Main, J., & Wilks, D. (1998). Statistical downscaling of general circulation model output: A comparison of methods. *Water resources research*, *34*(11), 2995–3008.
Worland, S. C., Farmer, W. H., & Kiang, J. E. (2018). Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. *Environmental modelling & software*, *101*, 169–182. doi: 10.1016/j.envsoft.2017.12.021
Zhang, H. S., Cook, D., Laa, U., Langrené, N., & Menéndez, P. (2022). cubble: A vector spatio-temporal data structure for data analysis [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=cubble` (R package version 0.1.1)

Statistical learning and topkriging improve spatio-temporal low-flow estimation

J. Laimighofer[1], **G. Laaha**[1]

[1]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

Key Points:

- Model-based boosting of the seasonal low-flow regime and topkriging for the residual field improve monthly low-flow predictions.
- Model accuracy is particularly high in the alpine areas, where low-flow occurs predominantly in winter.
- The hierarchical model structure is especially valuable in headwater catchments, and shows good performance for extreme events.

Corresponding author: Johannes Laimighofer, `johannes.laimighofer@boku.ac.at`

Abstract

This study assesses the potential of a hierarchical space-time model for monthly low-flow prediction in Austria. The model decomposes the monthly low-flows into a mean field and a residual field, where the mean field estimates the seasonal low-flow regime augmented by a long-term trend component. We compare four statistical (learning) approaches for the mean field, and three geostatistical methods for the residual field. All model combinations are evaluated using a hydrological diverse dataset of 260 stations in Austria, covering summer, winter, and mixed regimes. Model validation is performed by a nested 10-fold cross-validation. The best model for monthly low-flow prediction is a combination of a model-based boosting approach for the mean field and topkriging for the residual field. This model reaches a median R^2 of 0.73. Model performance is generally higher for stations with a winter regime (best model yields median R^2 of 0.84) than for summer regimes ($R^2 = 0.7$), and lowest for the mixed regime type ($R^2 = 0.68$). The model appears especially valuable in headwater catchments, where the performance increases from 0.56 (median R^2 for simple topkriging routine) to 0.67 for the best model combination. The favorable performance results from the hierarchical model structure that effectively combines different types of information: average low-flow conditions estimated from climate and catchment characteristics, and information of adjacent catchments estimated by spatial correlation. The model is shown to provide robust estimates not only for moderate events, but also for extreme low-flow events where predictions are adjusted based on synchronous local observations.

1 Introduction

Droughts and low-flows are significant hydrological and environmental hazards that threaten a wide range of water-related sectors, such as navigation, hydropower production and water management in general. Currently, prediction of low-flow is mainly focused on the spatial scale (Euser et al., 2013; Salinas et al., 2013; Castiglioni et al., 2009; Laaha et al., 2014; Tyralis et al., 2021; Worland et al., 2018; Laimighofer et al., 2022a), whereby deterministic models, or statistical models are applied. Spatio-temporal low-flow prediction is still rare, although space-time information on monthly low-flow is crucial for assessing ecological impacts on water quality, or estimating the risk of navigation disruptions. Space-time models are currently used in a wide range of environmental research fields (Kyriakidis & Journel, 1999), e.g. soil moisture modelling (Rodríguez-Iturbe et al., 2006), distribution of atmospheric pollution (Szpiro et al., 2010; Sampson et al., 2011; Lindström et al., 2014; Lindstrom et al., 2019; Mercer et al., 2011), downscaling meteorological variables (Wilby et al., 1998), or risk of wildfire outbreaks (Opitz et al., 2020). Transferring these space-time models to streamflow poses a particular challenge due to the tree-like structure of river catchments. Nevertheless, space-time models for streamflow are of particular interest, as they can be used for prediction in ungauged basins (Hrachowitz et al., 2013, PUB). This study aims to transfer an existing approach, originally adapted for air pollution modelling (Szpiro et al., 2010), to the space-time prediction of monthly low-flow.

Conceptually, statistical space-time models can be divided into individual space-time models, models that use temporal functions (deterministic or stochastic) that are correlated in space, or spatial functions that are correlated in time (Kyriakidis & Journel, 1999). The latter are less common for streamflow. Individual space-time models for prediction in ungauged basins (PUB) are mainly based on data-driven approaches such as long short-term memories (Kratzert, Klotz, Herrnegger, et al., 2019; Kratzert, Klotz, Shalev, et al., 2019; Lees et al., 2021, LSTM), artificial neural networks (Solomatine & Ostfeld, 2008; Cutore et al., 2007, ANN), or other machine learning methods such as tree-based models (Laimighofer et al., 2022a). These models typically use auxiliary space-time information on precipitation or evapotranspiration for streamflow estimation. In contrast, spatio-temporal geostatistical approaches exploit the similarity of hydrographs from nearby catchments. The simplest case is to apply ordinary kriging to the runoff time series, neglecting temporal correlations. In this context, Farmer (2016) found that such a simple model requires only a single (pooled) variogram to yield a median Nash-Sutcliffe efficiency of 0.7 for daily streamflow predictions on 182 stations in the United States. Ordinary kriging may not be the best choice for runoff, due to the nested and tree-like structure of the catchments. Therefore, other methods have been developed to take into account the peculiarities of catchment runoff. For example methods constraining the spatial covariance function by the water balance (Müller & Thompson, 2015), or methods that incorporate the river network hierarchy (Gottschalk, 1993; Sauquet et al., 2000), such as topkriging (Skøien et al., 2006; Skøien & Blöschl, 2007, TK). Farmer (2016) compared ordinary kriging to topkriging and showed a similar performance for both approaches. This is in contrast to studies in Austria and France (Skøien & Blöschl, 2007; Viglione et al., 2013; de Lavenne et al., 2016), which showed a favorable performance of topkriging also for daily and hourly runoff. Skøien and Blöschl (2007) additionally found, that in their topkriging application it was sufficient to estimate each time step separately, and no temporal dependency structure needed to be considered to achieve adequate performance.

Space-time models of the type where a temporal function (stochastic or deterministic) is correlated in space, are more common for runoff applications. They can be used, for instance, to improve the predictions of a hydrological model, when considering the output of a hydrological model as a deterministic function, which is interpolated in space by its model parameters. This regionalization of model parameters is performed on different temporal and spatial resolutions (Guo et al., 2021; Razavi & Coulibaly, 2013). Applications that use a stochastic temporal function are less frequent. For instance, Pumo et al. (2016) used a time series model for estimating monthly runoff in 59 basins in Sicily, with NSE values ranging from 0.7 to 0.8, but the model was validated only on a small subset of catchments. The time series model of Pumo et al. (2016) was determined a priori and only the coefficients of the parameters were estimated in space. A more flexible approach, that involves less information loss, is to use empirical ortoghonal functions (EOF). Gottschalk et al. (2015) and Li et al. (2018) applied EOFs for filling gaps in monthly discharge time series and Sauquet et al. (2008) tested spatially weighted EOFs for prediction of normalized mean monthly runoff in France. Studies, intended to model air pollutants, extended the approach of weighted EOFs, by adding a residual field (Szpiro et al., 2010), altering the methods for estimating the weights of the EOFs (Sampson et al., 2011; Mercer et al., 2011), or including spatio-temporal variables (Lindström et al., 2014; Lindstrom et al., 2019). All these studies analysed air pollutants in the United States, and reported cross-validated R^2 from 0.6 to about 0.75. The flexible model structure and the already highlighted use of EOFs for streamflow variables (Gottschalk, 1993; Li et al., 2018; Sauquet et al., 2008) demonstrate the potential for transferring this model to monthly low-flow. Such a transfer would involve incorporating both the average low-flow regime and the nested structure of river networks into the model.

The main objective of this study is to develop a hierarchical spatio-temporal model for monthly low-flow in Austria. The model consists of a mean field which should capture the seasonal cycle and the long-term trend of monthly low-flow and a residual field where geostatistical approaches are deployed. We test four different models for the mean field: (i) spatially weighted smoothed EOFs, (ii) a model-based boosting approach, which only estimates the seasonal cycle, (iii) a model-based boosting approach, which estimates the seasonal cycle and the long-term trend and (iv) a combination of model (ii) and (i). For the residual field we compare three kriging approaches - ordinary kriging (OK), physiographic kriging (PK) and topkriging (TK). The models are evaluated on a comprehensive Austrian dataset by 10-fold nested cross validation (CV) to emulate prediction in ungauged basins. The following research questions will be addressed:

1. Can a combination of statistical learning approaches and kriging methods improve spatio-temporal low-flow prediction in Austria?

2. What approach is best suited to model the seasonal low-flow regime?
3. Which kriging variant is best suited to model the space-time residual field?
4. How does prediction performance vary between headwater and non-headwater catchments?
5. What is the performance for summer, winter and mixed low-flow regimes?

2 Data

2.1 Hydrological data

This study is performed on a hydrological diverse dataset in Austria. We use 260 stations with a continuous daily streamflow record between 1982 to 2018. The same dataset was already used in a study on spatial low-flow prediction (Laimighofer et al., 2022b) and spatio-temporal low-flow prediction in Austria (Laimighofer et al., 2022a). The hydrological data can be downloaded from the Hydrographic Service of Austria (HZB). Our study focuses on a space-time model for low-flow. Hence, the daily streamflow time series is used to calculate the 0.05 quantile of discharge for every month (444 months at every station). We will refer to this index as monthly Q95 ($P(Q > Q95) = 0.95$). The monthly Q95 was standardized by catchment area, which results in the monthly specific low-flow (q95) time series (l s^{-1} km^{-2}). For all modelling approaches q95 is transformed by the square root, to approximate a normal distribution.

Occurrence of low-flow in Austria is more dominant in the winter half-year (November to April, winter regime type) for alpine catchments, where summer discharge is increased by snowmelt and increasing precipitation (Laaha & Blöschl, 2006; Laaha et al., 2017). In the northern parts of Austria and the Eastern low-lands low-flow mainly is present in the summer half-year (May to October, summer regime type). Nevertheless, not all catchments have this strong seasonality, and the occurrence of low-flow is alternating between winter and summer. This type of low-flow regime will be referred to as mixed regime type (Laaha & Blöschl, 2006; Laaha, 2023). The regime types are defined based on the seasonality ratio (SR)

$$SR = Q95_{summer}/Q95_{winter}, \tag{1}$$

where $Q95_{summer}$ is the 0.05 quantile of daily discharge for the summer period (May to November), and $Q95_{winter}$ the corresponding 0.05 quantile for the winter period of the respective station. A SR below 0.8 indicates a summer regime, a SR above 1.25 (1/0.8) determines a winter regime, and a SR between 0.8 and 1.25 is defined as a mixed regime. A graphical illustration of the defined regime types is given in Fig. 1. Despite the models developed here are on monthly time scale and thus not restricted to a particular regime type, we will use the seasonality regime types for an in-depth analysis of the results.

2.2 Catchment characteristics

In this study we apply several geostatistical and statistical learning methods, which all rely on catchment characteristics, that are supposed to be static over time in our approach. Ordinary kriging uses the geographic coordinates of the gauging stations, top-kriging requires the river network as input, and physiographic kriging is based on a principal component analysis of all catchment characteristics. The catchment characteristics can be subdivided into landuse variables, topographic descriptors, geological predictors and climatic characteristics. An overview of all variables is given in Table 1. For a more detailed description of the computation of the catchment characteristics we refer to Laaha and Blöschl (2006) and Laimighofer et al. (2022b). How the temporal information is added to the space-time models will be explained in Sect. 3.2.

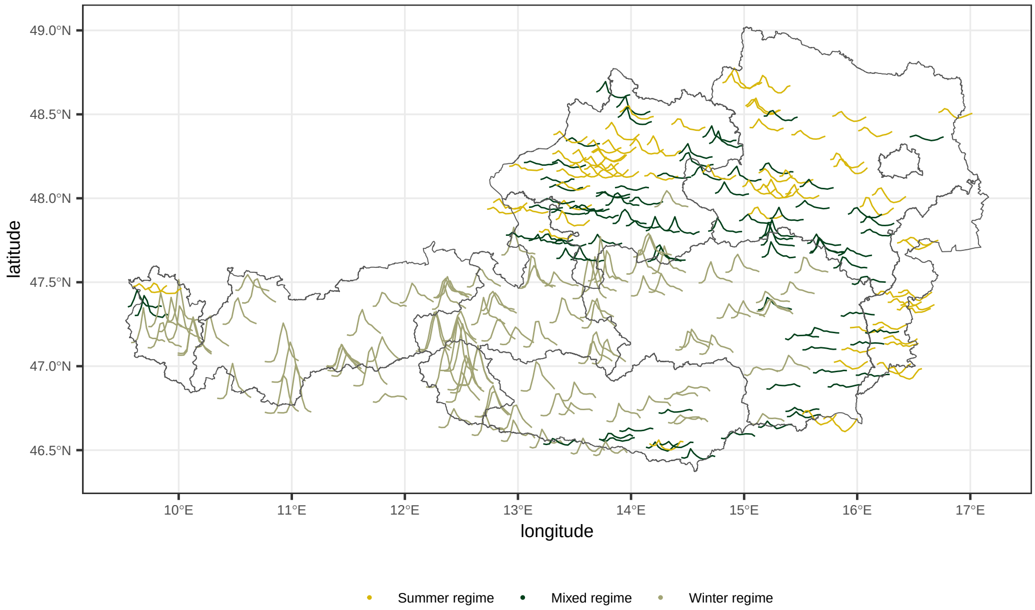

Figure 1. Overview of the study area. The colours indicate the seasonality regime type of the station, defined by the SR. The curves of each station is the scaled seasonal low-flow at each station for illustration of the different regime types.

3 Methods

3.1 Model structure

The basic model structure is given by

$$y(s,t) = \mu(s,t) + \upsilon(s,t), \tag{2}$$

where $y(s,t)$ is the monthly low-flow at a station s and time point t, $\mu(s,t)$ is defined as the mean field and $\upsilon(s,t)$ is the residual field of our model. Similar model designs were used by e.g. Szpiro et al. (2010), Lindstrom et al. (2019) or Sampson et al. (2011). In this model conceptualization the mean field should capture the seasonal cycle and the long-term trend of the response variable. Szpiro et al. (2010) used ordinary kriging for prediction of the space-time residual field, where only one variogram is estimated for all timesteps. A graphical overview specific to low-flow is shown in Fig. 2. In this study we extend the model introduced by Szpiro et al. (2010) to capture the nested structure of river catchments. We employ a hierarchical modeling framework, that (i) considers four different modeling approaches for the mean field, and (ii) three forms of kriging for the space-time residual field, to find the best-performing model combination for monthly low-flow prediction.

3.2 Mean field

The objective for modelling the mean field is to estimate the seasonal cycle and the long-term trend in the spatio-temporal model. In the context of low flows, the seasonal cycle corresponds to the average monthly low-flow regime (seasonal low-flow regime), which is augmented to transient conditions by the long-term trend component. Szpiro et al. (2010) or Lindström et al. (2014) used weighted empirical orthogonal functions (EOF), which were initially proposed by Fuentes et al. (2006), for estimating the mean field. Their

Table 1. Description of the catchment characteristics used in this study. The climatic characteristics as precipitation, climatic water balance, potential evapotranspiration, aridity index, snowmelt and temperature are computed 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_{win}), sum for summer (e.g. P_{sum}).

Variable	Description	Unit
A	catchment area	km^2
Lat, Lon	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	%
L_U, L_A, L_C, L_F, L_G, L_R, L_W, L_{WA}	Fraction of urban areas, agricultural areas, permanent crop, forest, grassland, wasteland, wetlands, water surfaces	%
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	%
D	Stream network density	10^2 m km^{-2}
P	Precipitation	mm
ET_P	Potential Evapotranspiration	mm
AI	Aridity index	-
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

approach can be written as

$$\mu(s,t) = \sum_{i=1}^{m} \beta_i(s) f_i(t). \tag{3}$$

The $f_i(t)$ are smoothed empirical orthogonal functions, which are spatially weighted by regression coefficients (β_i), so that the temporal structure can vary in space (Lindström et al., 2014). The number of EOFs is given by m, whereas $f_1(t)$ is always an intercept term. In this study, we compare four different methods for estimating the mean field. First, we will use the basic approach from Szpiro et al. (2010), by estimating the mean field with spatially weighted smoothed EOFs. This approach will be referred to as EOF_{simple}, and will serve as a benchmark for the other three methods. The second and third method use a model-based boosting approach for estimating the mean field. One implementation will only estimate the seasonal cycle of low-flow at each station (Boost_{SC}), while the other implementation will further include the long-term trend of low-flow at each station (Boost_{ST}). Finally, we combine the two approaches Boost_{SC} and EOF_{simple}, by first

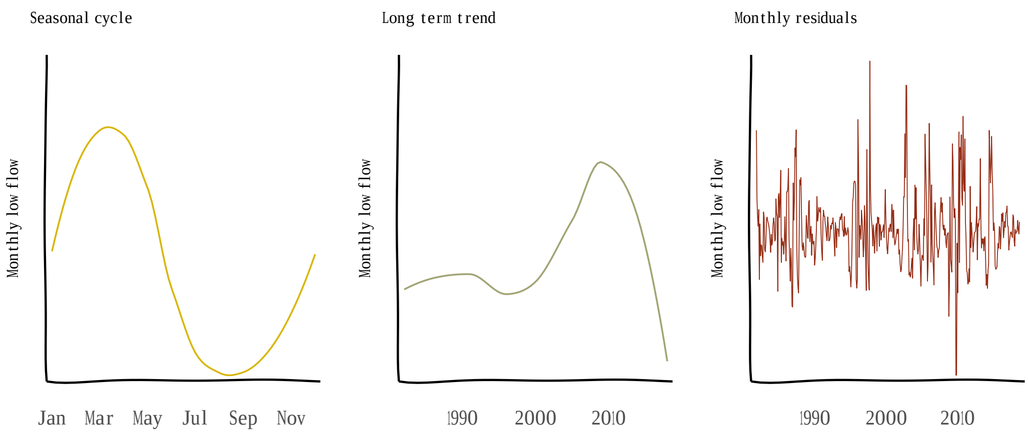

Figure 2. Model structure, exemplified for monthly Q95.

predicting the seasonal cycle and using the residuals for estimating the long-term effect by spatially weighted EOFs (Boost$_{EOF}$).

3.2.1 Smoothed empirical orthogonal functions

In our hierarchical model framework EOFs are used for estimating the mean field (EOF$_{simple}$), and in combination with seasonal boosting (Boost$_{EOF}$). In both cases, the first step is to build a matrix $\mathbf{x}_{EOF}$ ($T \times S$), where each column either corresponds to the monthly low-flow (EOF$_{simple}$), or to the residuals (Boost$_{EOF}$) at station s. The dimension T ($T = 444$) is the length of each monthly low-flow series at each station, and S ($S = 260$) is the number of stations. The matrix $\mathbf{x}_{EOF}$ is centered and scaled before applying a singular value decomposition. The smoothed EOFs are then calculated by fitting a spline on each singular value vector.

The number of EOFs (m) is determined by fitting a linear model (as in Eq. 3, for each s) to each column of $\mathbf{x}_{EOF}$ against m EOFs, where m is ranging from 1 (only an intercept term) to a maximum of 50 EOFs. For each single model the Bayesian information criterion (BIC) is calculated and averaged over all stations, resulting in a vector of BIC values (BIC$_m$) for each number of EOFs. As this approach would give only one realization for the entire set of stations and thereby lead to overfitting, we perform a bootstrap procedure with 25 repetitions (where a fraction of 70 % of the stations is sampled) to optimize the parameter m for the prediction at ungauged sites. The final number of EOFs is then determined by averaging every BIC$_m$ over all 25 bootstrap samples and for a more parsimonious model we add 1 standard deviation to the minimized BIC value, which then serves as the threshold. The minimum number of EOFs with an average BIC below this threshold are then selected as the final number of EOFs (m). A graphical description of this selection is shown in Fig. 3. For any number of EOFs, f_1 is an intercept term which is a vector of 1s, with length S.

The f_i are then weighted in space by the regression coefficients β_i, where each β_i is a vector of regression coefficients for all stations. To obtain predictions at ungauged locations, every β_i is estimated by a linear model, which can be formulated as

$$\beta_i = \alpha_{0i} + \sum_{j=1}^{J} \mathbf{x_j}\alpha_{ij}, \tag{4}$$

where α_{0i} is an intercept term, $\mathbf{x}$ is the matrix of the spatial predictors presented in Sect. 2, and α_i are regression coefficients. J is the number of predictor variables that needs to be optimized. As it is a priori not clear which variables to include in the β_i-regression

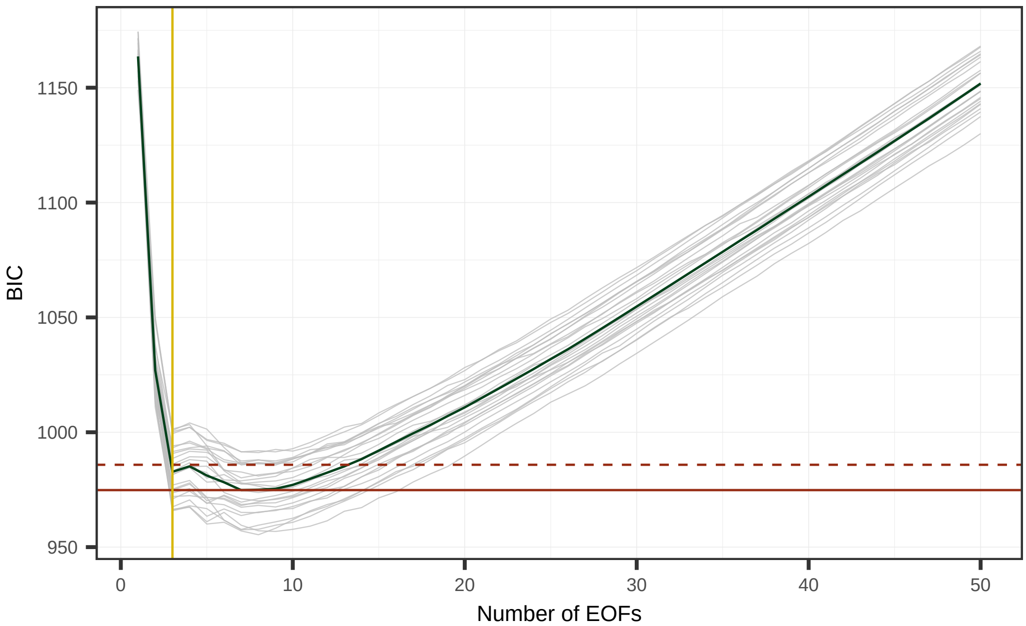

Figure 3. The number of EOFs are selected by a bootstrap procedure - with 25 samples. For each of the bootstrap samples the average BIC is calclulated. The number of EOFs is selected (the yellow line indicates the number of EOFs) by adding 1 standard deviation (shown by the red dashed line) to the minimum BIC value (shown by the red solid line).

model, possible approaches are to use shrinkage approaches as Lasso (Mercer et al., 2011), or dimension reduction methods as partial least-squares (Sampson et al., 2011, PLS). In this study we apply an approach that has already been shown to be useful for low-flow estimation in Austria (Laimighofer et al., 2022b). The variable selection is based on a recursive feature elimination (Granitto et al., 2006, RFE), which consists of an initial variable ranking and a backward variable selection. The initial variable ranking is estimated by a linear model-based boosting approach (a description of model based-boosting follows in Sect. 3.2.2). The variables are ranked after their absolute coefficients, and to obtain more robust results, the variable ranking is repeated 25-times by bootstrapping. For each β_i, a linear model is fitted to the first p ($p = 1, 2, 3, ..., 59$) ranked variables and the error is calculated and averaged over 25-bootstrap samples. The final number of variables (J) is defined by using 1.05 times the minimum error as a threshold to produce parsimonious models. The variable selection is performed for each β_i individually.

3.2.2 Model-based boosting

Model-based boosting (Bühlmann & Hothorn, 2007) is an iterative algorithm, where in each step a baselearner is selected, which best minimizes a predefined loss function (squared error in this study). To avoid overfitting the boosting algorithm uses a learning rate, to slowly approximate the final coefficients of the model. A baselearner can be e.g. a linear, a non-linear, random or spatial effect. Model-based boosting provides an intrinsic variable selection (Hofner et al., 2011), supports penalization of the effects and is robust against multicollinearity (Mayr & Hofner, 2018). The only parameter of the model that was tuned in this study was the number of boosting iterations, which was optimized using a 10-fold cross validation (CV) approach.

Based on this framework, the model for seasonal boosting (Boost$_{SC}$) can be formulated as

$$\mu(s,t) = \beta_0 + f_1(month) + \sum_{k=2}^{K} f_k(\mathbf{x}) + f_1(month)\mathbf{x}. \tag{5}$$

The model captures the average monthly low-flow regime. In the equation, β_0 is the intercept of the model and $\mathbf{x}$ is the predictor matrix with all spatial predictor variables. The spatial predictors can be parameterized by $f_k(.)$, either as a linear or a non-linear effect. We decomposed all non-linear effects into a linear and a non-linear part, as proposed by Kneib et al. (2009), to distinguish between linear and non-linear effects for each spatial variable. Further, a cyclic B-spline $f_1(month)$ according to Hofner et al. (2016) was added, which should represent the seasonal cycle of monthly low-flow. Finally, the term $f_1(month)\mathbf{x}$ was added to allow the seasonal cycle to vary in predictor variable space, in analogy to a varying-coefficient model (Hastie & Tibshirani, 1993; Fahrmeir et al., 2004). This leads to a total of $3p+1$ (178) baselearners for Boost$_{SC}$. For faster computation this model was not fitted to the full data, but only to the monthly averages at each station (seasonal low-flow cycle with 12 values per station).

In case of our spatio-temporal boosting approach (Boost$_{ST}$), the before mentioned model is extended by a long-term trend component to account for a transient seasonal low-flow regime. This trend component is captured by adding a sequence of all months ($T = 1, 2, 3, ..., 444$) as effect $f_2(time)$ to the model, which then can be written as

$$\mu(s,t) = \beta_0 + f_1(month) + f_2(time) + \sum_{k=3}^{K} f_k(\mathbf{x}) + f_1(month)\mathbf{x} + f_2(time)\mathbf{x}. \tag{6}$$

The long-term trend is modeled by a constant term and a spatially varying term, as it is done for the seasonal cycle. This results in $4p + 2$ (238) baselearners for Boost$_{ST}$.

3.3 Residual field

Following Szpiro et al. (2010) and Lindström et al. (2014), the residual field $\upsilon(s,t)$ is estimated by a kriging structure,

$$\hat{\upsilon}_{st} = \sum_{s=1}^{S} \lambda_s \upsilon_{st}, \tag{7}$$

where υ_{st} are the fitted residuals at location s and time t and λ_s are the kriging weights. Note that the kriging weights (λ_s) are static over all timepoints. Hence, only one variogram model is used across time. The original approach employs an ordinary kriging estimator that is based spatial proximity, which appears well suited for air-quality models, in which context the proposed model was first introduced (Szpiro et al., 2010; Sampson et al., 2011; Lindström et al., 2014).

Considering river discharge, geographic kriging may not be fully appropriate, as it does not include the nested structure of catchments. Therefore, we estimate the residual field not only by ordinary kriging (OK), but additionally use physiographic kriging (PK) and topkriging (TK). Physiographic kriging was introduced by Castiglioni et al. (2011) and computes the first two principal components (PC) on a set of catchment characteristics. These two PCs then span up the physiographic space for the kriging structure. Topkriging (Skøien et al., 2006; Laaha et al., 2014) takes into account not only the size and distance of the catchments, but also their nested structure along the river network. This makes the method particularly well suited for interpolation of river discharge.

To obtain the kriging weights λ_s, we need to estimate a variogram model for each of the three kriging approaches. Lindström et al. (2014) proposes to use a maximum likelihood approach for estimation of the residual field, that includes variogram estimation. As it is not straightforward to estimate a topkriging variogram through a maximum likelihood approach, we introduce a simple framework for the optimization of the variogram for all three kriging methods. The procedure starts by calculating the coefficient of determination R_t^2 for every timestep:

$$R_t^2 = 1 - \frac{\sum_{s=1}^{S}(y_{st} - \hat{\mu}_{st})^2}{\sum_{s=1}^{S}(y_{st} - \overline{y_t})^2}, \tag{8}$$

where $\hat{\mu}_{st}$ at this point is the prediction of one of the four models for the mean field, y_{st} are the observations, and $\overline{y_t}$ is the spatial average at the specific timepoint. If only a kriging approach is used alone (no estimation of the mean field), $\hat{\mu}_{st}$ is simply the average low-flow at every station. Next, we compute the average $\overline{R_t^2}$ over all R_t^2 and select the timestep (t_{step}) in which the deviation of R_t^2 is minimal to $\overline{R_t^2}$. The variogram is then optimized at the unique residual timeslice $\upsilon_{st_{step}}$. For the optimization of the variogram we use a 10-fold CV and a grid search over the parameter space. For each combination of the parameters the R_{CV}^2 of $\upsilon_{st_{step}}$ is calculated and the parameters with the highest R^2 are used for the final prediction.

3.4 Model validation

Model evaluation is performed by a nested 10-fold cross validation (Varmuza & Filzmoser, 2016, CV). A nested CV consists of an inner and an outer loop. The inner loop in this study is used for tuning the boosting model, select the number of EOFs, variable selection for the regression coefficients of the EOFs, and optimizing the variogram parameters. The outer loop is solely used for assessing the model performance. This nested CV-scheme was already applied in two studies for low-flow in Austria (Laimighofer et al., 2022a, 2022b). However, in this study some parts of the inner loop are altered, due to the hierarchical structure of the model. An illustration of the scheme is given in Fig. 4.

3.4.1 Performance metrics

We assess performance using three main metrics. They are calculated using cross validated predictions and should therefore provide an unbiased estimate of the model error. First, we compute the root mean squared error (RMSE) by

$$RMSE = \sqrt{1/N\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{9}$$

where N is the total number of observations ($N = S * T$), y_i are the observations and $\hat{y}_i$ are the predictions. Further, we calculate the median absolute error (MDAE):

$$MDAE = median(|y_i - \hat{y}_i|), \tag{10}$$

and the R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y_i})^2}. \tag{11}$$

The RMSE, MDAE and R^2 are computed based on all data points. Since we are particularly interested in how well the models can reproduce the mean field and thus provide an estimate of the mean seasonal low-flow regime, we additionally calculate all three metrics (RMSE_{month}, MDAE_{month}, R^2_{month}) for the seasonal predictions. This is shown exemplarily for the RMSE:

$$RMSE_{month} = \sqrt{1/SM\sum_{s=1}^{S}\sum_{m=1}^{M}(\overline{y}_{sm} - \overline{\hat{y}}_{sm})^2}, \tag{12}$$

where M is the number of months (12) and $\overline{y}_{sm}$ is:

$$\overline{y}_{sm} = 1/(N/M)\sum_{m=1}^{M} y_{sm}. \tag{13}$$

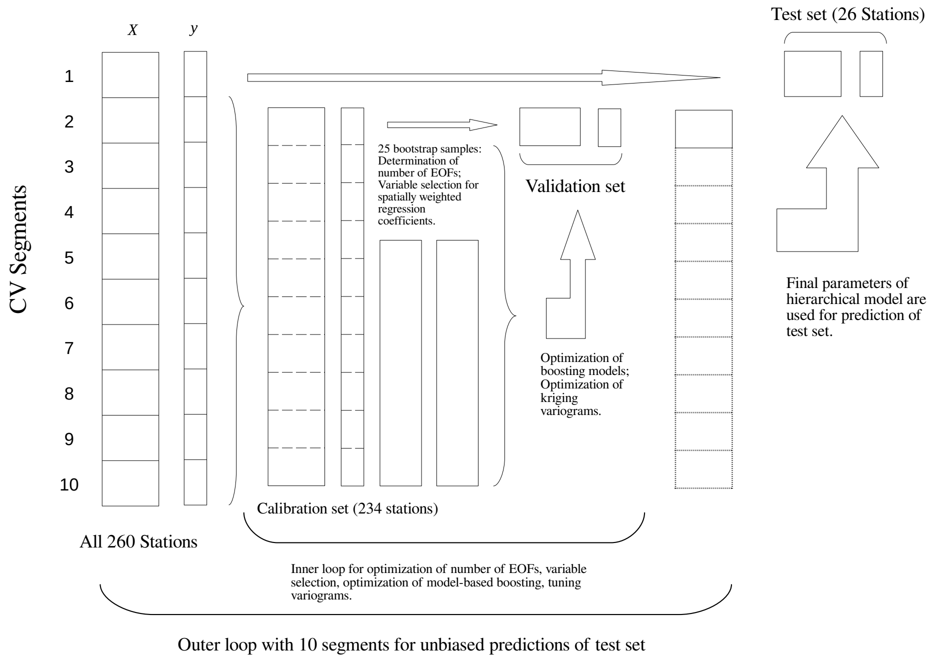

Figure 4. Schematic overview of the nested CV, that is used for model validation. We use different bootstrap samples for determination of number of EOFs and the selection of the β_i. Additionally the inner 10-fold CV is altered between optimization of the boosting models and the optimization of the variograms.

The ratio N/M can also be specified by the number of years (37 years of observations) at each station, and y_{sm} is every monthly low-flow value in month m at station s. The equation can be written accordingly for the predictions $\overline{\hat{y}}_{sm}$. Finally, we are interested in the performance of our models at each station, hence the R^2 is calculated for each station separately. Note that the equation of R^2 is equivalent to the formulation of the Nash–Sutcliffe efficiency (NSE, including the bias) in many hydrological studies (Blöschl et al., 2013).

4 Results

4.1 Mean field model components

Before proceeding with an overview of model performance, we shortly discuss some intrinsic features of the individual mean field model components - the inherent variable selection for the two boosting approaches, the weighted coefficients for the empirical orthogonal functions, and the determination of the number of EOFs.

4.1.1 Seasonal and spatiotemporal boosting

Model-based boosting includes an inherent variable selection procedure. Hence, we can analyse the selected variables and compare the structure of the two boosting models - seasonal boosting (Boost$_{SC}$) and spatiotemporal boosting (Boost$_{ST}$). Both boosting approaches used the maximum number of boosting steps over all ten folds that were predefined for each model (3000 for Boost$_{SC}$, 5000 for Boost$_{ST}$). In the seasonal boosting approach 45 baselearners were added on average to the model, whereas Boost$_{ST}$ exploited 67 baselearners on average. For both models the monthly cyclic spline ($f_1(month)$) was the most important variable. In both cases spatial covariabes were not added as single linear or non-linear baselearners, but solely as interaction effect of the cyclic spline, or the long-term trend. Figure 5 displays a graphical overview of the most important interaction effects for Boost$_{SC}$. The main important spatial predictors for Boost$_{SC}$ and Boost$_{ST}$ were topographic variables such as average catchment altitude and stream network density, landuse variables such as the fraction of wasteland, grassland and forest and, finally, meteorological conditions such as summer precipitation or snowmelt in winter. The long-term trend in the Boost$_{ST}$ model was added as a linear and non-linear effect and also weighted by spatial variables, but was generally negligible over all folds.

4.1.2 Smoothed empirical orthogonal functions

The smoothed empirical orthogonal functions (EOFs) were used as a single spatiotemporal framework (EOF$_{simple}$) and in combination with the seasonal boosting approach, where the EOFs (Boost$_{EOF}$) were estimated on the residuals of the seasonal predictions. In both cases, the number of EOFs were selected by a bootstrap procedure. EOF$_{simple}$ used 5 EOFs over all ten folds (Fig. 3 shows the selection of the number of EOFs), whereas the number of EOFs was slightly higher for the Boost$_{EOF}$ approach, ranging from 6 to 8 EOFs.

The EOFs were weighted by the meteorological, geological, landuse and topographic predictor variables in space. Our initially described variable selection (Sect. 3.2.1) reduced the number of variables for EOF$_{simple}$ to 6 predictors for the intercept term and 7 to 23 variables (from 59) for the other EOFs. In contrast, the Boost$_{EOF}$ approach produced more parsimonious models with only 2 spatial variables for the intercept, and 2 to 18 variables for the other EOFs.

Interpreting the selected variables is only straightforward in the case of the weighted intercept, which can be described as the mean low-flow for every station. This also explains the low number of variables used in the Boost$_{EOF}$ method, where the mean low-flow should already have been approximated by the seasonal boosting model. The left-over variables were the fraction of quarternary sediments, source region or stream net-

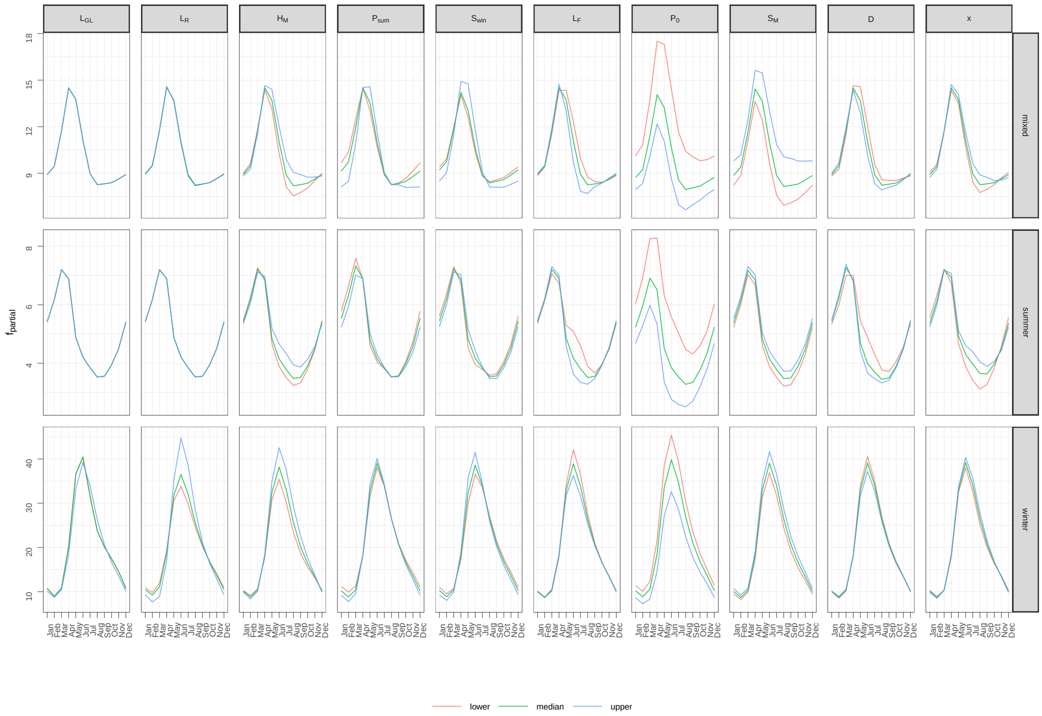

Figure 5. Partial predictions of the mean field with interaction effects of spatial predictors and the cyclic spline in the Boost$_{SC}$ model, stratified by low-flow regime type. Shown are the partial predictions for the 20%, 50% and 80% quantile of each spatial predictor variable within the considered regime type. The variable with the highest range in the spline coefficients is shown on the left, with a decreasing range to the right. Only the ten most important spatial variables are shown. As each fold leds to different results, the underlying model is the equivalent to the model produced by the first cross-validation run.

work density. The intercept of EOF_{simple} was mainly modeled by topographic descriptors as maximum and average catchment altitude, average slope and meteorological conditions as the aridity index in summer and days without precipitation in summer.

4.2 Model performance

This section assesses model performance from different perspectives. In a first step, we investigate how well the mean seasonal low-flow regime is represented by the individual mean field models. This is followed by an analysis of the predictive performance of the individual components of the hierarchical model, i.e. the four mean-field models and the three kriging approaches when they are used on their own. Finally, we evaluate the full hierarchical models composed of these components.

4.2.1 Representation of the seasonal low-flow regime

Table 2. Three different error measures (RMSE, R^2, MDAE) for the mean seasonal low-flow regime are presented. For the calculation the predicted and observed low-flow is averaged for each month and station.

Model structure	R^2_{month}	RMSE_{month}	MDAE_{month}
Boost_{SC}	0.84	5.76	1.56
Boost_{ST}	0.82	6.15	1.65
EOF_{simple}	0.74	7.47	1.87
Boost_{EOF}	0.85	5.70	1.57

In a first step of assessing model performance, we evaluate the four approaches used for modelling the mean field and how well they can estimate the seasonal low-flow regimes across Austria. Table 2 presents the RMSE_{month}, the R^2_{month} and the MDAE_{month} for the four approaches. Generally, the seasonal low-flow regime was well predicted by all four methods, but the EOF_{simple} approach showed a weaker performance on all three error metric with a RMSE_{month} of 7.47, compared to 6.15 (Boost_{ST}) and 5.76 (Boost_{SC}) for the two boosting approaches. The best performance was reached by the stacked model of seasonal boosting and the use of EOFs for the residuals ($\text{RMSE}_{month} = 5.7$), albeit the differences to the seasonal boosting approach is almost negligible and also the spatiotemporal boosting approach yields only slightly weaker performance metrics.

Examining the estimates for the three different regime types (Fig. 6) gives a more detailed picture of model performance. The weaker performance of EOF_{simple} was apparent for all three regime types, with a R^2_{month} ranging from 0.59 to 0.66, but is negligible for the mixed low-flow regime. EOF_{simple} resulted in smoother estimates of the seasonal cycle, which probably led to the lower performance especially for the summer and winter regime. Assessing the performance of the three other methods, the winter regime was best predicted with a R^2_{month} ranging from 0.78 to 0.81. For the summer regime the R^2_{month} was between 0.74 to 0.76, where for the mixed regime it dropped to 0.62 (0.6 for Boost_{ST}).

4.2.2 Performance of individual components

In a next step of model evaluation we assess the individual performances of the components of the hierarchical model framework: models for the mean field and the sole use of the three different kriging structures without considering the mean field. Table 3 gives an overview of the results. The individual model components yielded a RMSE of 8.42 (Boost_{EOF}) to 9.79 (EOF_{simple}). What is striking is that the spatiotemporal boosting

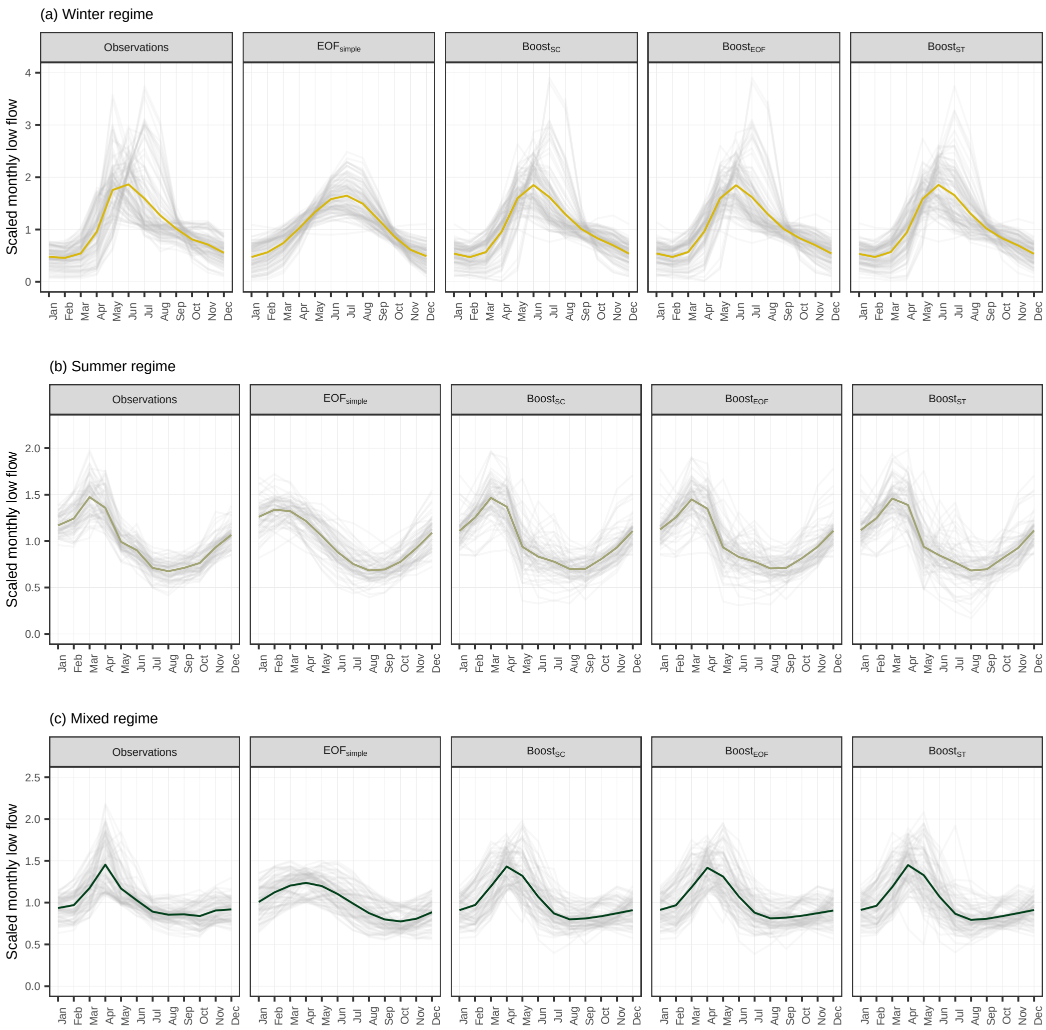

Figure 6. Predictions of the mean seasonal low-flow regime by various mean field models, stratified by regime type. The seasonal low-flow cycle is scaled by the mean at each station, for a better visualization. Each transparent line presents the seasonal cycle of one station, where the colored thick line is the average over all stations.

(Boost$_{ST}$) approach has a weaker overall performance on all metrics compare to the seasonal boosting approach (Boost$_{SC}$). Both approaches only yield a $R^2_{0.5}$ of 0.15. In case of Boost$_{SC}$ this is not surprising, as the model can only capture the seasonal cycle at each station. However, the additional long-term trend in the Boost$_{ST}$ approach is only adding noise to the model and leads to no improvement in terms of model performance. The long-term trend is better approximated by the stacked model of seasonal boosting and EOFs (Boost$_{EOF}$), which obtain the best results comparing the four mean field components. Generally, all three kriging approaches yielded a higher $R^2_{0.5}$, ranging from 0.48 for physiographic kriging (PK), to 0.63 for ordinary kriging (OK) and 0.75 for topkriging (TK). These results show, that TK already provides very accurate predictions without taking any spatio-temporal information into account.

Table 3. Overview of the error for the individual model components. R^2, RMSE and MDAE refer to the performance for all data points. $R^2 < 0.5$ ($R^2 < 0$) is the fraction of stations that yield a R^2 below 0.5 (0) and $R^2_{0.5}$ is the median of all R^2 computed per each station.

Model structure	R^2	RMSE	MDAE	$R^2 < 0.5$	$R^2 < 0$	$R^2_{0.5}$
Boost$_{SC}$	0.68	9.06	2.59	0.77	0.33	0.15
Boost$_{ST}$	0.67	9.29	2.61	0.79	0.34	0.15
EOF$_{simple}$	0.63	9.79	2.49	0.77	0.18	0.37
Boost$_{EOF}$	0.73	8.42	2.24	0.64	0.17	0.41
OK	0.75	8.02	2.09	0.40	0.22	0.63
PK	0.69	8.96	2.34	0.52	0.26	0.48
TK	0.80	7.25	1.62	0.30	0.15	0.75

4.2.3 Performance of hierarchical models

Table 4. Overview of the overall error for all hierarchical models. The Kriging column identifies the kriging approach which was used for the residual field. The Mean field column distinguishes between the different approaches for estimating the mean field of the model.

Kriging	Mean field	R^2	RMSE	MDAE	$R^2 < 0.5$	$R^2 < 0$	$R^2_{0.5}$
OK	Boost$_{SC}$	0.83	6.72	1.78	0.30	0.10	0.69
OK	EOF$_{simple}$	0.81	6.99	1.86	0.31	0.11	0.67
OK	Boost$_{EOF}$	0.82	6.77	1.79	0.32	0.10	0.68
OK	Boost$_{ST}$	0.81	6.96	1.86	0.33	0.11	0.66
PK	Boost$_{SC}$	0.79	7.37	1.94	0.41	0.13	0.58
PK	EOF$_{simple}$	0.77	7.73	2.00	0.38	0.15	0.59
PK	Boost$_{EOF}$	0.79	7.42	1.92	0.39	0.13	0.58
PK	Boost$_{ST}$	0.77	7.77	1.99	0.43	0.13	0.56
TK	Boost$_{SC}$	0.84	6.35	1.56	0.25	0.09	0.73
TK	EOF$_{simple}$	0.83	6.56	1.60	0.25	0.09	0.73
TK	Boost$_{EOF}$	0.84	6.35	1.60	0.24	0.09	0.72
TK	Boost$_{ST}$	0.84	6.44	1.64	0.27	0.08	0.70

In a next step we want to assess the prediction performance of the full hierarchical models that combine the component models evaluated before. Table 4 gives an overview of the cross-validated error of all models. We can observe that the application of differ-

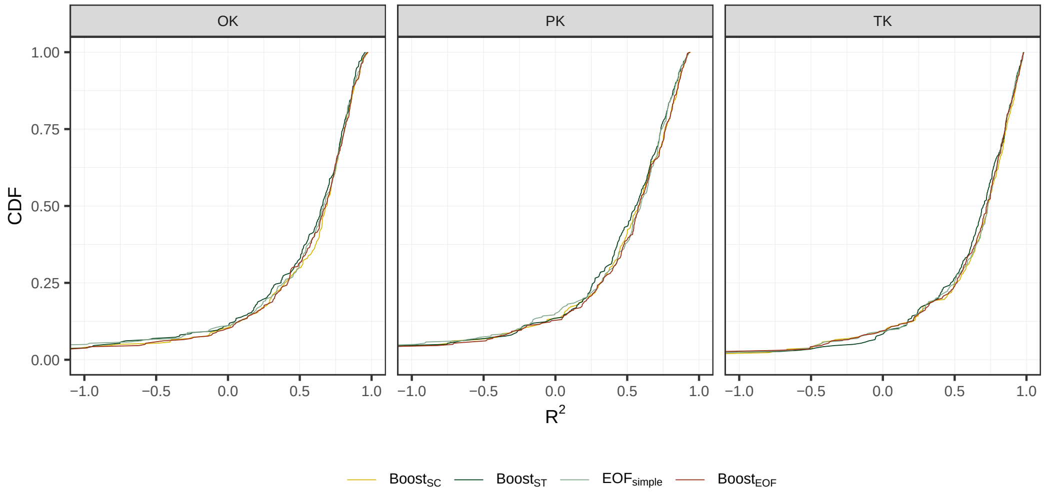

Figure 7. Cumulative distribution of station-wise R^2 stratified by kriging-method. Stations with a R^2 below -1 are omitted for clarity.

ent kriging methods led to the main variation in model performance, with a better performance of TK than OK and PK. The model combinations with TK yielded a RMSE from 6.35 to 6.56, whereas OK resulted in a somewhat higher RMSE between 6.72 and 6.99 and the use of PK for the residual field led to a RMSE of 7.37 to 7.77. This overall trend was also visible for all other performance measures.

In contrast, the different approaches for estimating the mean field only slightly altered the prediction performance of the models. For all kriging approaches the use of seasonal boosting, or Boost$_{EOF}$ yielded similar results. The models showed a somewhat weaker performance when the mean field was estimated by spatiotemporal boosting or EOF$_{simple}$, but when we look at the distribution of the R^2 over all stations (Fig. 7), these differences almost disappear. For instance, the $R^2_{0.5}$ for topkriging ranged only from 0.7 to 0.73 and the number of low-performing stations with a R^2 below 0 was between 8 % and 9 %.

4.2.4 Performance of hierarchical models grouped by seasonal regime types

For a deeper performance analysis of the hierarchical models, we again focus on the three low-flow regime types (winter, summer, mixed). Figure 8 gives an overview of the distribution of the R^2 for all three regimes. Regarding the kriging structure, hierarchical model with TK show the best performance over all three regimes. Highest prediction accuracy is reached for winter regime, where hierarchical models with TK yield a median R^2 of 0.8 to 0.84. Performance of OK is only slightly lower with a median R^2 from 0.78 to 0.81, but only 0.72 to 0.76 for PK. The performance is somewhat smaller for summer regimes for all models, and is lowest for mixed regimes, where combinations with TK still reach a median R^2 of 0.68 (lowest R^2 of 0.62), but median R^2 values for OK are only ranging from 0.5 to 0.57.

A further stratification of the results by the mean field model did not reveal a systematic picture of the performance. For example, EOF$_{simple}$ in combination with OK, resulted in the worst performance for summer regimes, but for physiographic kriging and topkriging the combination with EOF$_{simple}$ led to the best performance. Focusing on the mixed regime, the Boost$_{ST}$ method seemed to be disadvantageous for all kriging structures, but this was not apparent in the results of the winter or summer regime.

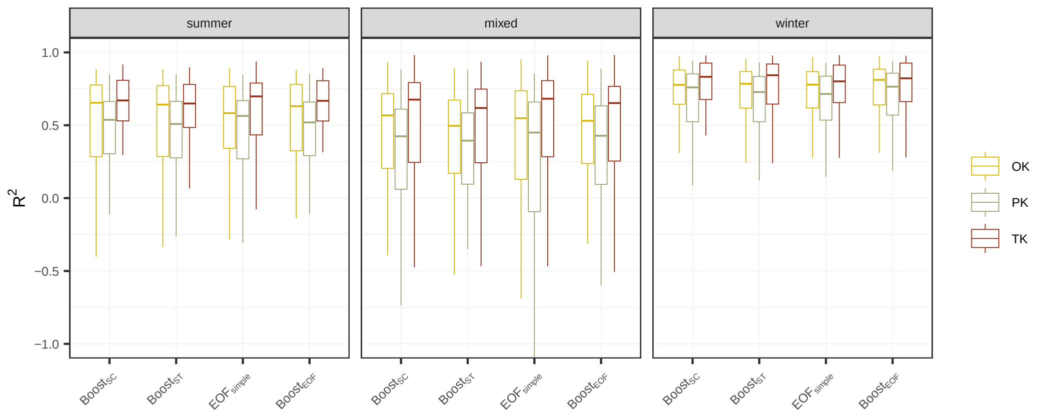

Figure 8. Comparison of the overall performance of the hierarchical models stratified by kriging method and regime type. Each boxplot shows the distribution of the R^2 over all stations. Outliers are removed from the plot for better visualization.

5 Discussion

5.1 Comparison of performance

In this paper, we extended an existing hierarchical model, initially proposed by Szpiro et al. (2010), for performing spatio-temporal predictions of monthly low-flow index series in Austria. We tested four models to approximate the seasonal cycle and the long-term trend, and compared three geostatistical approaches for the residual field. Comparison to existing literature is mainly limited to the study by Laimighofer et al. (2022a), where results can directly be compared as stations, temporal resolution, and even the used cross validation folds are equivalent to this study. In Laimighofer et al. (2022a) a single spatio-temporal framework was applied, where the best model yielded a median R^2 of 0.67 and an overall RMSE of 6.98. In this study these measures could be improved to a RMSE of 6.35 and a median R^2 of 0.73, for our best model ($Boost_{SC}$ and TK). Performance comparison to other literature is somehow difficult, as prediction studies on monthly streamflow data is mainly performed on monthly mean values and results are partially not evaluated by cross validation (Gottschalk et al., 2015; Sauquet et al., 2008; Pumo et al., 2016), which can best capture the error of prediction in ungauged basins.

In a more qualitative embedding of our results, we can highlight that hierarchical model combinations with topkriging yield the highest prediction accuracy. This is in line with studies for spatial low-flow prediction (Laaha et al., 2014), or spatio-temporal streamflow prediction in Austria (Skøien & Blöschl, 2007; Viglione et al., 2013), where also TK reaches high prediction performance. In contrast, Farmer (2016) shows that OK can perform as well as TK in a spatio-temporal framework, and suggests that ordinary kriging should be preferred over TK, due to the lower model complexity. Our results could paint a similar picture, as the performance metrics are only slightly improved by TK, but this is only true if we consider the full hierarchical model structure, where the between-model differences are reduced. Studies as Farmer (2016) or Skøien and Blöschl (2007) considered no additional seasonal cycle or long-term trend in their models. Focusing on our results for a single kriging structure (Table 3), the median R^2 for OK is only 0.63, but the median R^2 for TK is 0.75. However, the single TK approach only yields a RMSE of 7.25, which is substantially higher to the RMSE of 6.35 of the combination of $Boost_{SC}$ and topkriging. We will discuss these performance issues of topkriging in more depth in the next section.

Prediction accuracy of PK is generally lower for all hierarchical model combinations and for the single kriging approach. Results for spatial low-flow prediction in Italy (Castiglioni et al., 2011) showed similar performance of PK and TK, but this is not reflected in our space-time framework. The lower performance of PK may be caused by the similar information used by the mean field models and the first two principal components covering the physiographic space for PK.

5.2 Effect of headwater vs. non-headwater on topkriging performance

Albeit, several studies demonstrated the good performance of topkriging (Skøien & Blöschl, 2007; de Lavenne et al., 2016; Laaha et al., 2014; Farmer, 2016; Viglione et al., 2013), accuracy of TK is altered as a function of catchment area (Viglione et al., 2013), station density (Parajka et al., 2015), or the hierarchical position in the river network (Laaha et al., 2014; de Lavenne et al., 2016). Laaha et al. (2014) found that the R^2 for TK in headwater catchments for spatial low-flow prediction is 0.59, whereas in non-headwater catchments performance increased to a R^2 of 0.91. A similar trend was shown by de Lavenne et al. (2016), where the performance of TK increased with higher Strahler order. This is consistent with our results (displayed in Fig. 9), where we can see a general trend for all model combinations that a higher Strahler order increases the prediction performance. Considering the performance of each model combination, we observe that a simple topkriging routine is not sufficient for headwater catchments (Strahler order 1 - 2). For example the median R^2 for simple TK is 0.56 for catchments with a Strahler order 1. Adding seasonal predictions (Boost_{SC}) to the model structure enhances prediction to a median R^2 of 0.67. Differences between the models almost disappear when considering catchments with Strahler order 2. Here the median R^2 is between 0.67 and 0.7, but simple TK shows a much higher variance in the results. In catchments with a Strahler order of 3 or more, the simple TK routine provides the most accurate predictions compared to the hierarchical model combinations. However, we can show that the lower performance of topkriging in headwater catchments can be improved by a hierarchical framework that that exploits the seasonal cycle in advance.

5.3 Case study - extreme events

So far our model assessment focused on global model performance. In a last step, we want to consider a concrete discharge time series, to demonstrate the potential of our modeling approach. As our main interest is to predict low-flows we will focus on two drought years 2003 and 2015 (Ionita et al., 2017; Laaha et al., 2017). We selected the hydrograph Altschlaining at the river Tauchenbach in eastern Austria, which already was investigated by Laaha et al. (2017). The Tauchenbach is a small (upstream) catchment with 89.2 km^2, which experienced a particularly extreme low-flow event in 2003 (Fig. 10). The event of 2003 started with an early onset and continued over the whole year, whereas in 2015 wetter preconditions in spring led to a later onset and prevented a more severe low-flow event in summer.

The seasonal boosting approach in combination with TK yields a cross-validated R^2 of 0.45 at Altschlaining, which is lower than about 80 % of all stations. Nevertheless, the development of the low-flow events is captured quite well by model predictions, which can be decomposed to the mean field component and the residual field component. Figure 10 illustrates the complementary behaviour of these two components. In extreme events like 2003 and 2015, the observed low flows deviate strongly from the seasonal low-flow regime. For this reason, the mean field component of the hierarchical model would provide a biased estimate. The TK of the residual field, however, performs an adjustment of the predictions to the respective event conditions, as can be seen for both events. It uses synchronous information of adjacent stations to achieve enhanced space-time predictions. Such adjustment would indeed be much smaller in a 'normal' year, where the low-flow conditions are similar to the average regime.

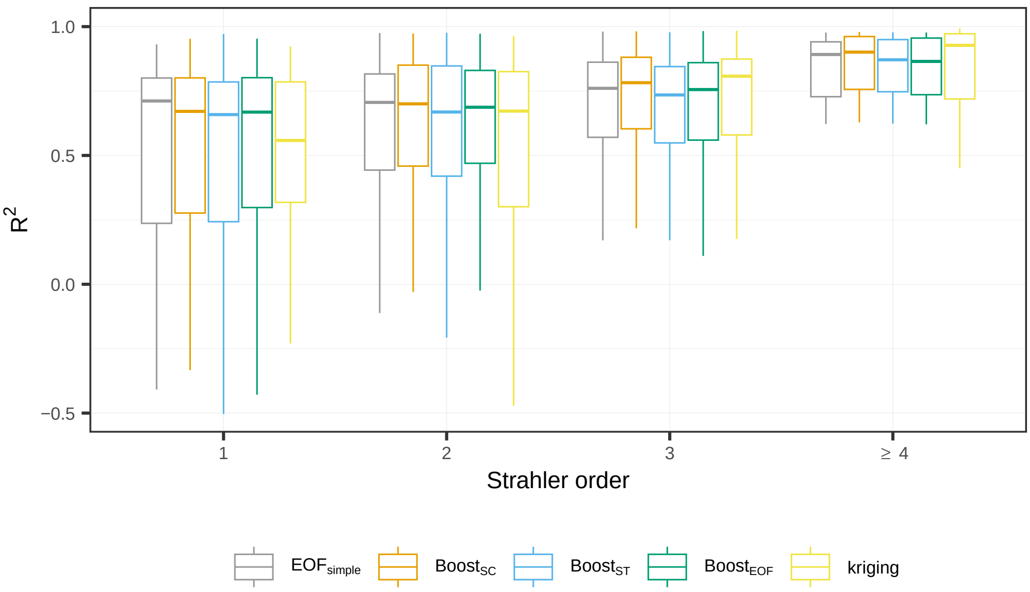

Figure 9. The boxplots show all possible estimation of the mean field in combination with topkriging, and a simple topkriging routine in which only one variogram is estimated for the full spatio-temporal domain. The catchments are further stratified by their Strahler order (x-axis). Due to the limited stations with Strahler order ≥ 4, these stations are condensed in one group.

Despite these favorable properties, some below-average performance can be observed in spring 2003, where discharges reflect the very dry preconditions that led to the severe low-flow event. This seasonal anomaly can be explained by a particular weather situation where the Tauchenbach experienced a precipitation deficit over several years due to lee-effects behind alpline and pre-alpine mountain ranges (Laaha et al., 2017). Since this is a local singularity, the anomaly cannot be adjusted by information from neighboring stations, so a residual TK does not significantly improve the estimates. Further on, the (regionally more consistent) atmospheric water deficit of the summer drought event gets increasingly important. This leads to enhanced residual TK, which is reflected in steadily improving predictions during the ongoing low-flow event.

6 Conclusions

In this study we adopted a hierarchical model framework for spatio-temporal modelling of monthly low-flow in Austria. The best performing model is a combination of model-based boosting for the mean field, which estimates the seasonal low-flow regime, and topkriging for predicting the residuals. It gives a median R^2 of 0.73 over all stations, demonstrating the high potential of the hierarchical model.

Generally, stations with a strong winter seasonality of low-flows show a higher prediction accuracy than summer or mixed regimes. The drivers of monthly low-flow in winter regime catchments are mainly high sums of precipitation and snowmelt in the summer months, and freezing and low sums of precipitation in the winter. The signal of monthly low-flow in mixed or summer regimes is more noisy, which slightly weakens the prediction performance.

Regardless of regime type or mean field methods used, topkriging shows the best performance for all model combinations, followed by ordinary kriging and physiographic kriging. It is striking that even a simple topkriging routine without an additional mean

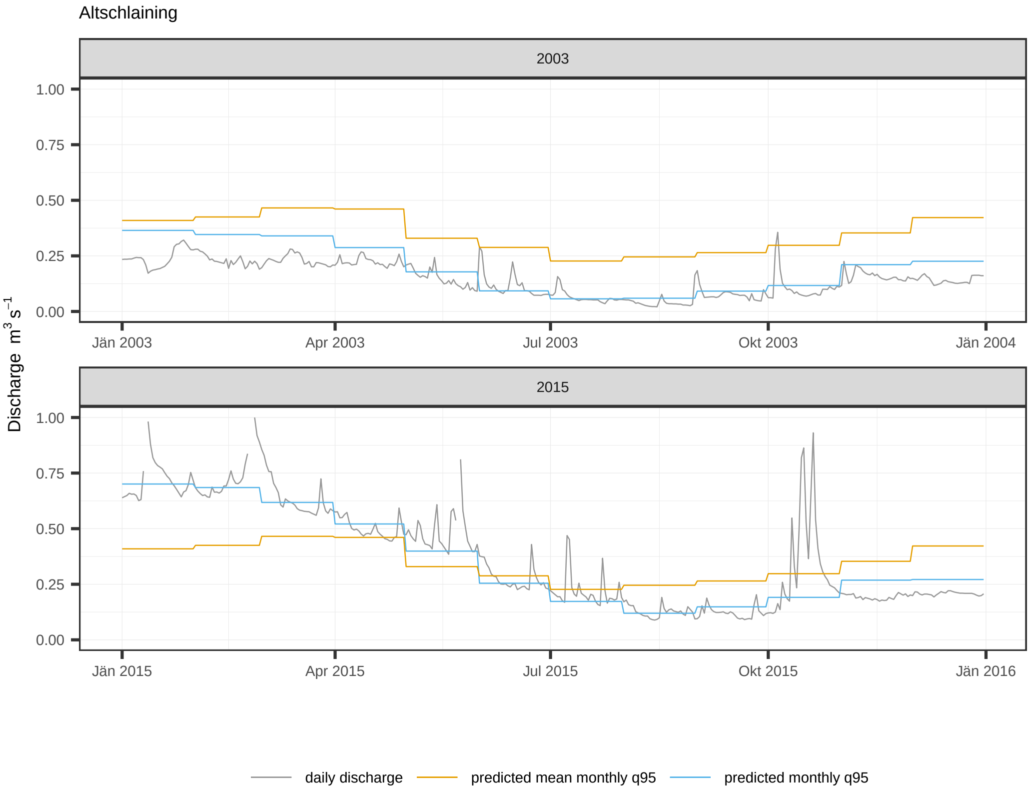

Figure 10. Comparison of two drought years (2003 and 2015), for the station Altschlaining, river Tauchenbach. Each plot shows the daily discharge, predicted mean monthly q95 and predicted monthly q95 - both are transformed back to discharge values (m^3s^{-1}).

field achieves a median R^2 of 0.75, but has a higher number of poorly performing stations ($R^2 < 0.5$). It shows a lack of prediction accuracy, especially in headwater catchments. In these catchments the hierarchical model framework is particularly beneficial, whereas in catchments of Strahler order ≥ 3 the simple topkriging routine is sufficient.

Overall, the favorable performance of the model results from its specific structure, which seems well suited to combine different types of information: average low flow conditions estimated from climate and catchment characteristics, and information of neighbouring catchments estimated by spatial correlation. This combination provides accurate results not only for average years, where the high prediction accuracy for the seasonal low-flow regime comes into play, but also for extreme years, where top-kriging adapts to the anomalous conditions of the low-flow event and can also capture the preconditions. The model is shown to provide robust estimates for a range of conditions, including headwater catchments and extreme events. It demonstrates a high degree of suitability for predicting gaps in the low-flow series, and for providing estimates at ungauged sites.

7 Open Research

Modelling and data analysis was performed in R version 4.2.2 (R Core Team, 2022). We want to acknowledge the use of the following packages: caret (Kuhn, 2022), cubble (Zhang et al., 2022), gridExtra (Auguie, 2017), lubridate (Grolemund & Wickham, 2011), mboost (Hothorn et al., 2022), Metrics (Hamner & Frasco, 2018), rtop (Skoien et al., 2014), sf (Pebesma, 2018), tidyverse (Wickham et al., 2019), wesanderson (Ram & Wickham, 2018). Model output and code to produce the figures is available at zenodo (Laimighofer & Laaha, 2023).

Acknowledgments

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. This research has been supported by the Climate and Energy Fund under the programme “ACRP” (grant no. C265154).

The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC).

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).

References

Auguie, B. (2017). gridextra: Miscellaneous functions for "grid" graphics [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=gridExtra` (R package version 2.3)

Blöschl, G., Sivapalan, M., Wagener, T., Savenije, H., & Viglione, A. (2013). *Runoff prediction in ungauged basins: synthesis across processes, places and scales.* Cambridge University Press. doi: 10.1017/CBO9781139235761

Bühlmann, P., & Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, *22*(4), 477 – 505. doi: 10.1214/07-STS242

Castiglioni, S., Castellarin, A., & Montanari, A. (2009). Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. *Journal of hydrology*, *378*(3-4), 272–280. doi: 10.1016/j.jhydrol.2009.09.032

Castiglioni, S., Castellarin, A., Montanari, A., Skøien, J. O., Laaha, G., & Blöschl, G. (2011). Smooth regional estimation of low-flow indices: physiographical space based interpolation and top-kriging. *Hydrology and Earth System Sciences*, *15*(3), 715–727. doi: 10.5194/hess-15-715-2011

Cutore, P., Cristaudo, G., Campisano, A., Modica, C., Cancelliere, A., & Rossi, G. (2007). Regional models for the estimation of streamflow series in ungauged basins. *Water resources management*, *21*(5), 789–800. doi: 10.1007/s11269-006-9110-7

de Lavenne, A., Skøien, J. O., Cudennec, C., Curie, F., & Moatar, F. (2016). Transferring measured discharge time series: Large-scale comparison of top-kriging to geomorphology-based inverse modeling. *Water Resources Research*, *52*(7), 5555-5576. doi: https://doi.org/10.1002/2016WR018716

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, *17*(5), 1893–1912. doi: 10.5194/hess-17-1893-2013

Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression for space-time data: a bayesian perspective. *Statistica Sinica*, 731–761.

Farmer, W. H. (2016). Ordinary kriging as a tool to estimate historical daily streamflow records. *Hydrology and Earth System Sciences*, *20*(7), 2721–2735. doi: 10.5194/hess-20-2721-2016

Fuentes, M., Guttorp, P., & Sampson, P. D. (2006). Using transforms to analyze space-time processes. *Monographs on Statistics and Applied Probability*, *107*, 77.

Gottschalk, L. (1993). Interpolation of runoff applying objective methods. *Stochastic hydrology and hydraulics*, *7*, 269–281.

Gottschalk, L., Krasovskaia, I., Dominguez, E., Caicedo, F., & Velasco, A. (2015). Interpolation of monthly runoff along rivers applying empirical orthogonal functions: Application to the upper magdalena river, colombia. *Journal of Hydrology*, *528*, 177–191.

Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and intelligent laboratory systems*, *83*(2), 83–90.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from `https://www.jstatsoft.org/v40/i03/`

Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *Wiley Interdisciplinary Reviews: Water*, *8*(1), e1487.

Hamner, B., & Frasco, M. (2018). Metrics: Evaluation metrics for machine learning [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=Metrics` (R package version 0.1.4)

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(4), 757–779.

Hofner, B., Hothorn, T., Kneib, T., & Schmid, M. (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, *20*(4), 956–971.

Hofner, B., Kneib, T., & Hothorn, T. (2016). A unified framework of constrained regression. *Statistics and Computing*, *26*, 1–14.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2022). mboost: Model-based boosting [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=mboost` (R package version 2.9-7)

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., … others (2013). A decade of predictions in ungauged basins (pub)—a review. *Hydrological sciences journal*, *58*(6), 1198–1255.

Ionita, M., Tallaksen, L. M., Kingston, D. G., Stagge, J. H., Laaha, G., Van Lanen, H. A. J., … Haslinger, K. (2017). The european 2015 drought from a climatological perspective. *Hydrology and Earth System Sciences*, *21*(3), 1397–1419. doi: 10.5194/hess-21-1397-2017

Kneib, T., Hothorn, T., & Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics*, *65*(2), 626–634.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, *55*(12), 11344–11354. doi: 10.1029/2019WR026065

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. doi: 10.5194/hess-23-5089-2019

Kuhn, M. (2022). caret: Classification and regression training [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=caret` (R package version 6.0-92)

Kyriakidis, P. C., & Journel, A. G. (1999). Geostatistical space–time models: a review. *Mathematical geology*, *31*, 651–684.

Laaha, G. (2023). A mixed distribution approach for low-flow frequency analysis – part 1: Concept, performance, and effect of seasonality. *Hydrology and Earth System Sciences*, *27*(3), 689–701. doi: 10.5194/hess-27-689-2023

Laaha, G., & Blöschl, G. (2006). Seasonality indices for regionalizing low flows. *Hydrological Processes*, *20*(18), 3851-3878. doi: https://doi.org/10.1002/hyp.6161

Laaha, G., & Blöschl, G. (2006). A comparison of low flow regionalisation methods—catchment grouping. *Journal of Hydrology*, *323*(1), 193-214. doi: https://doi.org/10.1016/j.jhydrol.2005.09.001

Laaha, G., Gauster, T., Tallaksen, L. M., Vidal, J.-P., Stahl, K., Prudhomme, C., ... Wong, W. K. (2017). The european 2015 drought from a hydrological perspective. *Hydrology and Earth System Sciences*, *21*(6), 3001–3024. doi: 10.5194/hess-21-3001-2017

Laaha, G., Skøien, J., & Blöschl, G. (2014). Spatial prediction on river networks: comparison of top-kriging with regional regression. *Hydrological Processes*, *28*(2), 315-324. doi: https://doi.org/10.1002/hyp.9578

Laimighofer, J., & Laaha, G. (2023, June). *Code and model output to "Statistical learning and topkriging improve spatio-temporal low-flow estimation"*. Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.8007772` doi: 10.5281/zenodo.8007772

Laimighofer, J., Melcher, M., & Laaha, G. (2022a). Low-flow estimation beyond the mean–expectile loss and extreme gradient boosting for spatiotemporal low-flow prediction in austria. *Hydrology and Earth System Sciences*, *26*(17), 4553–4574.

Laimighofer, J., Melcher, M., & Laaha, G. (2022b). Parsimonious statistical learning models for low-flow estimation. *Hydrology and Earth System Sciences*, *26*(1), 129–148.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall–runoff models in great britain: a comparison of long short-term memory (lstm)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, *25*(10), 5517–5534. doi: 10.5194/hess-25-5517-2021

Li, L., Gottschalk, L., Krasovskaia, I., & Xiong, L. (2018). Conditioned empirical orthogonal functions for interpolation of runoff time series along rivers: Application to reconstruction of missing monthly records. *Journal of Hydrology*, *556*, 262-278. doi: https://doi.org/10.1016/j.jhydrol.2017.11.014

Lindstrom, J., Szpiro, A., Sampson, P. D., Bergen, S., & Oron, A. P. (2019). Spatiotemporal: Spatio-temporal model estimation [Computer software manual].

Retrieved from https://CRAN.R-project.org/package=SpatioTemporal (R package version 1.1.9.1)
Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., & Sheppard, L. (2014). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and ecological statistics*, *21*, 411–433.
Mayr, A., & Hofner, B. (2018). Boosting for statistical modelling-a non-technical introduction. *Statistical Modelling*, *18*(3-4), 365–384.
Mercer, L. D., Szpiro, A. A., Sheppard, L., Lindström, J., Adar, S. D., Allen, R. W., ... Kaufman, J. D. (2011). Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (nox) for the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Atmospheric Environment*, *45*(26), 4412-4420. doi: https://doi.org/10.1016/j.atmosenv.2011.05.043
Müller, M., & Thompson, S. (2015). Topreml: a topological restricted maximum likelihood approach to regionalize trended runoff signatures in stream networks. *Hydrology and Earth System Sciences*, *19*(6), 2925–2942.
Opitz, T., Bonneu, F., & Gabriel, E. (2020). Point-process based bayesian modeling of space–time structures of forest fire occurrences in mediterranean france. *Spatial Statistics*, *40*, 100429.
Parajka, J., Merz, R., Skøien, J. O., & Viglione, A. (2015). The role of station density for predicting daily runoff by top-kriging interpolation in austria. *Journal of Hydrology and Hydromechanics*, *63*(3), 228–234.
Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. Retrieved from https://doi.org/10.32614/RJ-2018-009 doi: 10.32614/RJ-2018-009
Pumo, D., Viola, F., & Noto, L. V. (2016). Generation of natural runoff monthly series at ungauged sites using a regional regressive model. *Water*, *8*(5), 209. doi: 10.3390/w8050209
R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
Ram, K., & Wickham, H. (2018). wesanderson: A wes anderson palette generator [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=wesanderson (R package version 0.3.6)
Razavi, T., & Coulibaly, P. (2013). Streamflow prediction in ungauged basins: review of regionalization methods. *Journal of hydrologic engineering*, *18*(8), 958–975.
Rodríguez-Iturbe, I., Isham, V., Cox, D. R., Manfreda, S., & Porporato, A. (2006). Space-time modeling of soil moisture: Stochastic rainfall forcing with heterogeneous vegetation. *Water Resources Research*, *42*(6). doi: https://doi.org/10.1029/2005WR004497
Salinas, J., Laaha, G., Rogger, M., Parajka, J., Viglione, A., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins–part 2: Flood and low flow studies. *Hydrology and Earth System Sciences*, *17*(7), 2637–2652. doi: 10.5194/hess-17-2637-2013
Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindström, J., & Kaufman, J. D. (2011). Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment*, *45*(36), 6593-6606. doi: https://doi.org/10.1016/j.atmosenv.2011.04.073
Sauquet, E., Gottschalk, L., & Krasovskaia, I. (2008, 10). Estimating mean monthly runoff at ungauged locations: an application to France. *Hydrology Research*, *39*(5-6), 403-423. doi: 10.2166/nh.2008.331
Sauquet, E., Gottschalk, L., & Leblois, E. (2000). Mapping average annual runoff: a hierarchical approach applying a stochastic interpolation scheme. *Hydrological*

sciences journal, *45*(6), 799–815.
Skøien, J. O., & Blöschl, G. (2007). Spatiotemporal topological kriging of runoff time series. *Water Resources Research*, *43*(9).
Skoien, J. O., G. Bloschl, G. Laaha, E. Pebesma, J. Parajka, & A. Viglione. (2014). Rtop: An r package for interpolation of data with a variable spatial support, with an example from river networks. *Computers & Geosciences*.
Skøien, J. O., Merz, R., & Blöschl, G. (2006). Top-kriging - geostatistics on stream networks. *Hydrology and Earth System Sciences*, *10*(2), 277–287. doi: 10.5194/hess-10-277-2006
Solomatine, D. P., & Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of hydroinformatics*, *10*(1), 3–22. doi: 10.2166/hydro.2008.015
Szpiro, A. A., Sampson, P. D., Sheppard, L., Lumley, T., Adar, S. D., & Kaufman, J. D. (2010). Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics*, *21*(6), 606-631. doi: https://doi.org/10.1002/env.1014
Tyralis, H., Papacharalampous, G., Langousis, A., & Papalexiou, S. M. (2021). Explanation and probabilistic prediction of hydrological signatures with statistical boosting algorithms. *Remote Sensing*, *13*(3), 333. doi: 10.3390/rs13030333
Varmuza, K., & Filzmoser, P. (2016). *Introduction to multivariate statistical analysis in chemometrics.* CRC press.
Viglione, A., Parajka, J., Rogger, M., Salinas, J. L., Laaha, G., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins; part 3: Runoff signatures in austria. *Hydrology and Earth System Sciences*, *17*(6), 2263–2279. doi: 10.5194/hess-17-2263-2013
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686
Wilby, R. L., Wigley, T., Conway, D., Jones, P., Hewitson, B., Main, J., & Wilks, D. (1998). Statistical downscaling of general circulation model output: A comparison of methods. *Water resources research*, *34*(11), 2995–3008.
Worland, S. C., Farmer, W. H., & Kiang, J. E. (2018). Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. *Environmental modelling & software*, *101*, 169–182. doi: 10.1016/j.envsoft.2017.12.021
Zhang, H. S., Cook, D., Laa, U., Langrené, N., & Menéndez, P. (2022). cubble: A vector spatio-temporal data structure for data analysis [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=cubble` (R package version 0.1.1)