Data-driven Estimation of Groundwater Level Time-Series Using Comparative Regional Analysis

E. Haaf[1], M. Giese[2], T. Reimann[3], and R. Barthel[2]

[1]Department of Architecture and Civil Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden.

[2]Department of Earth Sciences, University of Gothenburg, Sweden.

[3]Institute for Groundwater Management, TU Dresden, Dresden, Germany.

Corresponding author: Ezra Haaf (ezra.haaf@chalmers.se)

Key Points:

- Presents method for estimation of daily groundwater levels through transfer of head duration curves based on similarity of site characteristics at monitored sites.
- Nonlinearity of controls on groundwater levels favors use of Machine Learning (e.g., regression trees) over multiple linear regression for prediction.
- Investigates the dynamic nature of controls on groundwater levels, which is central for studies of recharge seasonality, droughts and floods.

Abstract

A new method is presented to efficiently estimate daily groundwater level time series at unmonitored sites by linking groundwater dynamics to local hydrogeological system controls. The presented approach is based on the concept of comparative regional analysis, an approach widely used in surface water hydrology, but uncommon in hydrogeology. The method uses regression analysis to estimate cumulative frequency distributions of groundwater levels (groundwater head duration curves (HDC)) at unmonitored locations using physiographic and climatic site descriptors. The HDC is then used to construct a groundwater hydrograph using time series from distance-weighted neighboring monitored (donor) locations. For estimating times series at unmonitored sites, in essence, spatio-temporal interpolation, stepwise multiple linear regression, extreme gradient boosting, and nearest neighbors are compared. The methods were applied to ten-year daily groundwater level time series at 157 sites in alluvial unconfined aquifers in Southern Germany. Models of HDCs were physically plausible and showed that physiographic and climatic controls on groundwater level fluctuations are nonlinear and dynamic, varying in significance from "wet" to "dry" aquifer conditions. Extreme gradient boosting yielded a significantly higher predictive skill than nearest neighbor and multiple linear regression. However, donor site selection is of key importance. The study presents a novel approach for regionalization and infilling of groundwater level time series that also aids conceptual understanding of controls on groundwater dynamics, both central tasks for water resources managers.

1 Introduction

Groundwater head observations are the basis for most investigations in hydrogeology. However, boreholes for groundwater observation as well as corresponding groundwater level time series are often scarce and unevenly distributed in both space and time. This is a disadvantage for effective management of groundwater resources at the regional scale (Butler et al., 2021), where water managers assess the current and future status of groundwater resources (Lóaiciga & Leipnik, 2001). In consequence, methods are needed to estimate groundwater head time series at ungauged sites.

Two main approaches are commonly used by hydrogeologists to predict temporal changes in groundwater head at a given site, (a) numerical and (b) statistical models. The typical approach is to implement a process-based, numerical groundwater flow model. However, numerical models typically require large amounts of data and effort, while investigators commonly are confronted with a lack of comprehensive description and documentation of the subsurface. This results in significant uncertainty, both regarding conceptualization and parametrization (e.g. Enemark et al., 2019). Dealing with this uncertainty leads to a tedious and time-consuming process to construct, calibrate, and run these process-based models (Bakker & Schaars, 2019). Additionally, models for meaningful local projections at large spatial scales are not yet available (Berg & Sudicky, 2019). An alternative to regional scale modelling with less need for detailed subsurface description are lumped (rainfall-runoff) hydrological models with a groundwater component (Barthel & Banzhaf, 2016). However, these models are problematic as they usually imply oversimplification of the groundwater component, disregarding the local descriptors of hydrogeological systems and their 3-dimensional setup (Barthel & Banzhaf, 2016; Butler et al., 2021). Generally, lumped models may provide adequate descriptions of groundwater systems only for simple hydrogeological situations such as shallow, unconfined aquifers, but not for more complex systems, such as deep and confined aquifers.

A different type of approach requiring only measured groundwater level data for groundwater time series estimation are parametric or data-driven methods. This approach requires few data on local system descriptors, while often long and measurement-dense series of input signal and groundwater measurements are necessary to achieve good calibrations. In contrast to groundwater-gradient driven methods, data-driven methods either use spatio-temporal geostatistics (e.g. Ruybal et al., 2019; Varouchakis et al., 2022) or transfer net precipitation input into groundwater level changes (Z. Chen et al. (2002)). However, available methods predict groundwater level only at monthly or annual resolution and consequently do not capture the large intra-annual and intra-monthly variability of groundwater dynamics (e.g. Heudorfer et al., 2019). An approach to predict time series at higher temporal scales are transfer functions, that can be used to yearly, monthly and daily temporal resolutions, such as impulse-response functions (e.g. Collenteur et al., 2019; Marchant & Bloomfield, 2018; Von Asmuth, 2012) or artificial neural networks (c.f. Rajaee et al., 2019; Wunsch et al., 2022). However, no formal method is known to transfer information from such models from monitored to unmonitored aquifers, although recently attempted in streamflow (Kratzert et al., 2019). This means that these methods can only make predictions when sufficient local time series data are available (e.g., 10 years weekly data (Wunsch et al., 2021)).

In summary, neither numerical models nor the currently available data-driven tools provide a straightforward approach to estimate daily groundwater levels at unmonitored sites to aid regional scale management. Therefore, new and complementary methodologies are required to overcome scarcity and patchy data distribution. Such approaches should be less data hungry than numerical models, yet account for local hydrogeological conditions and allow prediction at high temporal resolution despite limited local data availability. In surface-water-orientated hydrology, data scarcity has been countered with approaches of classification and similarity analysis, embraced by the hydrological community particularly within the PUB initiative (Predictions in Ungauged Basins; (Blöschl et al., 2013; Hrachowitz et al., 2013; McDonnell & Woods, 2004; Sivakumar & Singh, 2012; Wagener et al., 2007). These concepts attempt to systematically link the physical form and structure of catchments to their functioning by comparative analysis. Such links can then be used to transfer information to similar systems for prediction, i.e., regionalization or spatio-temporal interpolation. However, such approaches are rarely considered in groundwater research, which is pointed out by various authors, e.g., Barthel et al. (2021); de Marsily et al. (2005); Green et al. (2011); Voss (2005). Recently, a number of studies initiated the implementation of these approaches in groundwater, quantitatively connecting groundwater response to physiographic and climatic descriptors (Boutt, 2017; Giese et al., 2020; Haaf & Barthel, 2018; E. Haaf et al., 2020; Heudorfer et al., 2019; M. Rinderer et al., 2017; M. Rinderer et al., 2019; M. Rinderer et al., 2014; Michael Rinderer et al., 2016). These approaches, however, have not yet been exploited to predict daily groundwater levels at unmonitored sites.

When looking for methodological inspiration in the body of literature within the surface water community, and more specifically the PUB initiative, a large majority of approaches use regionalization mainly as a tool to calibrate lumped rainfall-runoff models at unmonitored sites (He et al., 2011; Hrachowitz et al., 2013). As mentioned above, such lumped models are often not useful for describing groundwater dynamics and, when available, are time-consuming to set up and calibrate (Jackson et al., 2016; Mackay et al., 2014). Simpler statistical methods for regionalization of streamflow time series, however, have been proposed by e.g. Shu and Ouarda (2012) based on Hughes and Smakhtin (1996). These methods make use of the characteristic relationship between flow duration curve (FDC; cumulative frequency of time where a flow is equaled or exceeded) and physiographic and climatic site descriptors, a relationship that is well investigated (Yokoo & Sivapalan, 2011). FDCs in surface water hydrology are commonly used to study the flow regime throughout the range of discharges and integrate effects of climate, topography, geology, and also anthropogenic activity (Ridolfi et al., 2020; Sugiyama et al., 2003; Vogel & Fennessey, 1995). This implies that the shape of a specific FDC is theoretically inferable from site descriptors. The technique evaluated in this study takes advantage of this through estimation of duration curves at unmonitored (target) sites based on similarity to neighboring donor sites. Then, from the estimated duration curve, time series are reconstructed at the target site into a daily time series (Hughes & Smakhtin, 1996; Mohamoud, 2010; Shu & Ouarda, 2012; Smakhtin, 1999).

Cumulative frequency or duration curves of groundwater heads are not as broadly used for studying groundwater resources, except when for example analyzing the relative state of groundwater storage (e.g. Maxe, 2013). Giese et al. (2020) estimated aggregates (indices) of head duration curves (HDC) and linked differences in shapes to local, intermediate, and regional groundwater flow patterns. Ezra Haaf et al. (2020) found correlation between HDC indices and map-derivable physiographic and climatic site descriptors. These are indications that alike streamflow, system controls are integrated in groundwater level regimes and may be exploited by analysis of duration curves.

Accordingly, regionalization and subsequent estimation of daily time series at unmonitored sites through duration curves of groundwater head is evaluated in this paper. Hereby the approach is based on the methodology proposed by Shu and Ouarda (2012) for streamflow. It is adapted to groundwater, where groundwater head duration curves as well as groundwater-relevant and map-derivable site descriptors are used. Within surface-water, this method has only been tested using stepwise multiple linear regression (MLR). In this study, a comparison is carried out with estimation through averaging of the nearest neighbor sites (NN), MLR, and extreme gradient boosting (XGB). XGB can represent nonlinear relationships between groundwater dynamics and site descriptors and has shown to be powerful in e.g., recharge studies (Naghibi et al., 2020). In summary, a method is evaluated that may be used when aquifer and time series data at a site of interest are unmonitored. The regionalization approach is applied to unconfined, alluvial aquifers in a humid climate in Southern Germany at unmonitored sites using solely map-derivable site descriptors and data from neighboring locations.

2 Method and Data

2.1 General strategy

The methodology of estimating groundwater level time series at an unmonitored site , is based on information from donor sites and requires the steps as explained in Figure 1. In the beginning, donor sites are selected with a time series period that is of interest for target site estimation. Next, time series are transformed to HDCs, and at 15 fixed percentile levels, models are constructed based on multiple regression analysis and gradient boosted regression trees, and logarithmically inter- and extrapolated (section 2.4.1-2.4.2). Finally, time series at ungauged sites are then reconstructed with a distance-based weighting method using the sequence of records from donor sites (section 2.4.3). For performance comparison, time series are also evaluated using only a distance-based average of time series from donor sites, further called Nearest-Neighbour (NN). Then, the number of neighbors and the performance of daily groundwater level estimations at target sites are evaluated using leave-one-out cross-validation (2.5). The models that are used for estimation of time series are then checked for plausibility (section 2.6). In section 2.7 the case data set is described, which is further analyzed using cluster analysis to understand results with regard to different groundwater regimes and systems. All data analysis was carried out by using the programming language *R* (R Development Core Team, 2022).

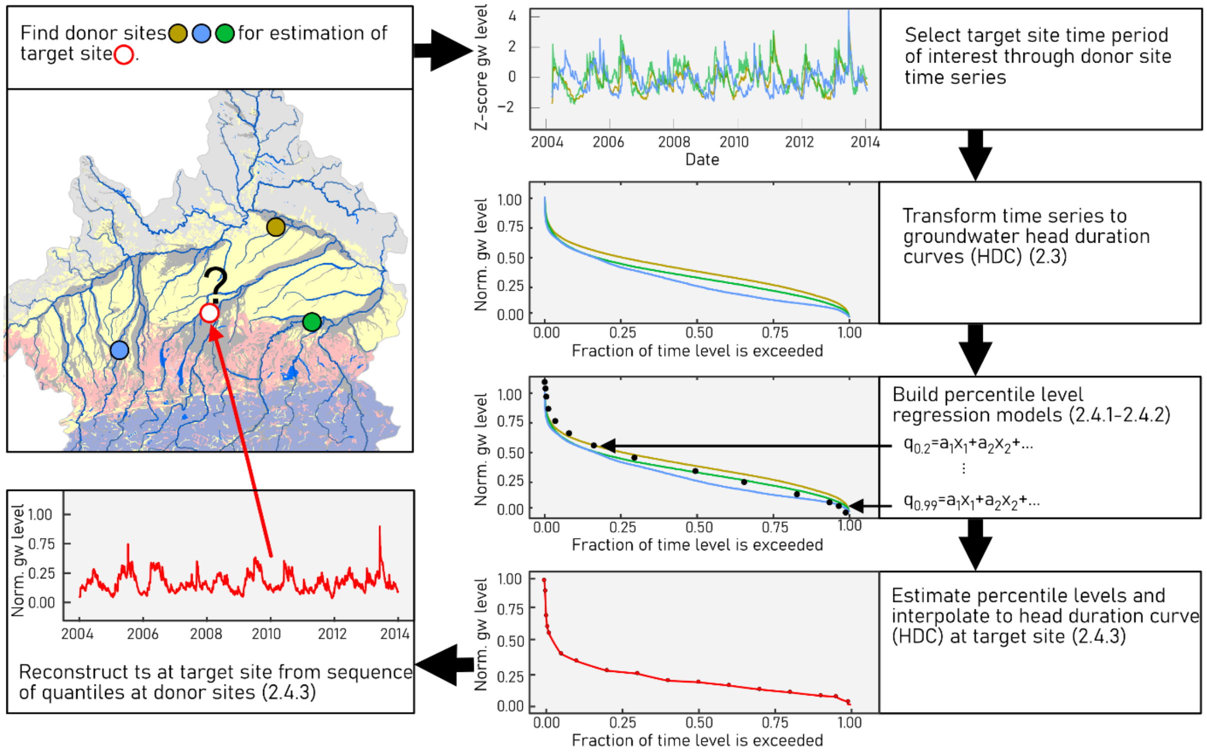
Find donor sites ● ● ● for estimation of target site ○.
?
Norm. gw level
1.00
0.75
0.25
0.00
2004
2006
2008
2010
2012
2014
Reconstruct ts at target site from sequence of quantiles at donor sites (2.4.3)
Z-score gw level
4
2
0
−2
Date
Select target site time period of interest through donor site time series
Norm. gw level
1.00
0.75
0.50
0.25
0.00
Fraction of time level is exceeded
Transform time series to groundwater head duration curves (HDC) (2.3)
Build percentile level regression models (2.4.1-2.4.2)
$q_{0.2}=a_1x_1+a_2x_2+...$
⋮
$q_{0.99}=a_1x_1+a_2x_2+...$
Estimate percentile levels and interpolate to head duration curve (HDC) at target site (2.4.3)

Figure 1. Principle steps to estimate groundwater level time series at unmonitored sites using the head duration curve methodology.

2.2 Data Selection and Processing

Groundwater level time series are selected from a data set described by E. Haaf et al. (2020). The data set contains groundwater level time series from the Upper Danube catchment in Bavaria, Southern Germany, with available geological information and absence of patterns of direct anthropogenic impact (for a more detailed explanation refer to Heudorfer et al. (2019)). From this data set observation wells were selected that come (1) with continuous daily time series and at least 10 year record length, (2) less than 1% missing data, which are (3) concurrent with a record period 2004–2014. The resulting set of 157 obervation wells are mostly located in shallow, quaternary sediments in river valleys and fluvial sand as well as in gravel deposits, with a few boreholes located in deeper tertiary sediments. All wells are classified as penetrating unconfined aquifers. Then, at each site, 47 physiographical and meteorological descriptors were derived, described in detail in Ezra Haaf et al. (2020). In addition to Ezra Haaf et al. (2020), percentage of land cover within a 3 km radius of each site was derived from the CORINE land cover data set (Bossard et al., 2000). Table 1 shows selected descriptors that are most important for models on this study and therefore discussed in more detail. Remaining descriptors can be found in the supporting information SI (Table S1). Descriptors are called predictors when in context of regression models.

Table 1. Descriptive statistics of physiographic and climatic descriptors, discussed in the paper. Class of variable in parenthesis: (G) Geology, (M) Morphology, (L) Land cover, (B) Boundaries and (C) Climate.

Variable	**Description**	**Range**		**Unit**
		Minimum	Maximum	
dist_stream (B) ‡	Estimated distance from well to nearest stream (main rivers)	6	10958	m
well_elevation (B)	Estimated Elevation of well	310	839	m asl.
P_avg (C)	Mean annual precipitation	675	1613	mm
T_avg (C)	Mean annual temperature	6.4	9.3	°C
SI (C)	Seasonality index of precipitation	.11	.31	-
A_thickness (G)	Average thickness of saturated zone	1	50.1	m
A_Depth (G)	Bottom of formation	3	110	m
Depth_to_GW (G)	Average depth to Water table	0.3	39.8	m
Broadleaved_forest (L)	% of 3 km buffer occupied by broadleaved forest	0	44.5	%
Coniferous_forest (L)	% of 3 km buffer occupied by coniferous forest	0	93.5	%
Urban (L)	% of 3 km buffer occupied by urban fabric	0	74.9	%
slp_sk (M) †	Mean slope	0/-0.1	1.95/2.6	-
twi (M)	Mean value of Topographic Wetness index	5.8	8.9	-

† skewness was calculated for local and regional scale respecitvely. For these, the ranges are given seperated by a slash l/r.

2.3 Transformation to head duration curves (HDCs)

In a first step, groundwater head time series were normalized. Subsequently, duration curves of groundwater levels were calculated at each site. This was done, by first ranking all n observed, normalized (on a 0-1 scale) groundwater levels $l_i, i = 1,2,\ldots,n$ in descending order, where i is the rank of an observation. The head duration curve (HDC) is then constructed following the Weibull plotting formula (Sugiyama et al., 2003):

$$p_i = P(L \geq l_i) = \frac{i}{n+1}, \tag{1}$$

where p_i is the percentage of time that a given level l_i is equaled or exceeded. Groundwater level or head duration curves are subsequently the plot of percentage level p_i against the corresponding level l_i (as seen in Figure 1).

2.4 Regression analysis for percentile models

To be able to estimate the duration curve at an ungauged site, forward stepwise regression (MLR, see section 2.4.1) and extreme gradient boosting (XGB, see section 2.4.2) were applied to build models from physiographic and climatic predictors at selected percentage level (0.1%, 0.5%, 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99%). Models are fit using a nested cross-validation approach on 80% training data with 20% hold-out data on which evaluation is performed. Models were trained 30 times by leaving out one group each time and then evaluating against predictions in the left-out group.

2.4.1 Construction of percentile models with MLR

Multiple linear regression models at selected percentage levels are built using a selective inference framework. Selective inference adjusts p-values for the effect of sequential selection of variables (Taylor & Tibshirani, 2015). This is necessary since conventional stepwise regression leads to an overestimation of the strength of apparent relations. The consequence of conventional models is therefore selection of non-significant predictors and therefore overfitting (Taylor & Tibshirani, 2015). Instead of using p-values based on the t-test for forward selection, the procedure is here stopped based on the false discovery rate (exceeding 0.1; (G'Sell et al., 2016). The selected variables are then used to build a regression relationship for the training data set with n observations (from well locations) and percentage levels, $p = 1,2 \dots n$, where H_p is the percentile of the normalized head H and x_p the selected climatic and physiographic descriptors with the following form:

$$H_p = \beta_0 + \sum_j x_{pj}\beta_j + \epsilon_p, \tag{2}$$

errors ϵ_p being independent and normally distributed and where β is a vector of model parameters that are estimated.

2.4.2 Construction of percentile models with XGB

Alternative models for each percentile were constructed using extreme gradient boosting, an implementation of boosted regression trees (Friedman, 2001). Hereby, the *xgb.train* function from the XGBoost R package (T. Chen & Guestrin, 2016) was used to predict H_p based on the entire set of climatic and physiographic descriptors. To optimize the model fit but reduce risk of overfitting, two further steps were carried out, after the 80/20 hold-out split mentioned above. After this, hyperparameters were determined on the training data using 5-fold cross validation, using the performance measure root mean square error (RMSE). Finally, after cross-validation, the risk for overfitting was reduced by stopping the ensemble at the number of decision trees, where the difference between training and evaluation error reaches a minimum.

2.4.3 From percentile models to estimated time series

Once percentile levels are predicted for a given target site using XGB and MLR models, logarithmic interpolation is used to estimate percentiles of groundwater levels between the percentage points in order to construct the entire duration curve. The percentile to be estimated is found by identifying the closest (modelled) fixed percentage levels p_i above and p_{i-1} below and their corresponding groundwater heads H_i and H_{i-1}. The groundwater level H can then be found using the following equation:

$$\ln(H) = \ln(H_i) + \frac{\ln(H_{i-1}) - \ln(H_i)}{p_{i-1} - p_i} \times (p - p_i) \tag{3}$$

In cases where percentiles are estimated that are larger than the highest percentage point or lower than the lowest (modelled) percentage point, logarithmic extrapolation is used. Hereby, the closest two percentage points are found (p_{n1}, p_{n2}) and the corresponding groundwater levels (H_{n1}, H_{n2}). Extrapolating to the percentile p is done using the equation below.

$$\ln(H) = \ln(H_{n1}) + \frac{\ln(H_{n1}) - \ln(H_{n2})}{p_{n1} - p_{n2}} \times (p - p_{n2}) \tag{4}$$

Reconstruction of the groundwater level time series from interpolated duration curves can then be carried out following the principle given by Smakhtin (1999) for streamflow estimation. Groundwater levels H_t at the target site are estimated by looking up the donor site's percentile of the duration curve at the first date to be estimated. Then the same percentile is found in the target site's duration curve and the corresponding groundwater level is chosen as the estimated level at the particular date. This process is repeated for all dates available within the record of the donor sites. However, not all donor sites are given the same weight for estimation at the target site. The estimated series of groundwater levels at the target site H_t are rather put together (equation 5) by weighting each source site's contribution based on the Euclidean distance d_t to the target.

$$H_t = \sum_{j=1}^{n} w_j H_{sj} \,/\, \sum_{j=1}^{n} w_j \tag{5}$$

The weights are calculated based on a dissimilarity measure:

$$w_j = \frac{1/d_t}{\sum_{j=1}^{n} 1/d_t} \tag{6}$$

Groundwater levels are also estimated at each target site using a straightforward nearest neighbor method (NN). Here, NN means that no duration curve is reconstructed but only the actual time series of each source site L_{tj} is used, however, weighted according to eq. 5 and 6.

2.5 Evalutation of Time Series Estimation

The performance of the daily groundwater level prediction was evaluated using leave-one-out cross validation as performed by Shu and Ouarda (2012). Using a leave-one-out cross validation procedure means that one (target) site is considered unmonitored and thus left out from the dataset. With the remaining data set ($n - 1$ sites), the groundwater level time series are estimated at the target site. Here, a maximum of n=20 sites were allowed as donor sites. Then, the performance at that site is evaluated by calculating the Kling-Gupta Efficiency (KGE), Pearson correlation coefficient (R), and Root-mean-square error (RMSE) as goodness of fit measures between estimated and observed time series. These steps are repeated at each of the n sites and the average (cross-validated) estimate is found by aggregating the goodness of fit-estimates from each sub-sample.

2.6 Plausibility Analysis of Models

To examine the plausibility of models used to predict percentile points along the HDC, the impact on model output is analyzed using standardized regression coefficients (MLR) and Shapley Additive Explanations values (SHAP) for XGB (Lundberg et al., 2020) using the *R* package *SHAPforxgboost* (Liu & Just, 2021). SHAP values quantify how much individual predictors, across the predictor's value range, contribute to the output variable (here the percentile point).

More specifically, the SHAP value gives the difference in the model output depending on if the model is fit with or without the predictor. Using scatterplots, SHAP values can then be interpreted locally which allows understanding of the dependence structure within each model for each predictor. Further, mean absolute SHAP of all data points for each model is estimated, yielding global feature importance across each percentile. This supports understanding of the dynamic changes of importance of controls across different aquifer states and allows qualitative comparison to standardized regression coefficients of MLR models.

2.7 Cluster Analysis

In order to get a better understanding of the dataset, regarding similarities in dynamics and subsequently site descriptors, hierarchical cluster analysis was performed. Prior to cluster analysis, the selected groundwater level time series are transformed to z-scores. As input into the clustering algorithm, Euclidean pairwise distances between time series were computed. Subsequently, hierarchical cluster analysis using Ward linkage is performed on the matrix of pairwise distances. The hierarchical relationship between the series can then be displayed in a dendrogram. From the dendrograms a scree plot is constructed, by sorting the heights of the dendrograms branches and plotting these against the number of nodes. The inflection point of the scree plot is then identified to select the number of clusters that sufficiently describes the patterns of member time series, while still generalizing the data set to a managable level.

3 Results and Discussion

3.1 Hydrogeological Description of Clusters

Cluster analysis of the data set based on similarity of groundwater level time series results in hydrogeologically meaningful groups. The six identified clusters (see SI, Figure S1-S2) are either made up of wells exclusively located in alluvial deposits or in alluvial deposits and outwash plains. Further, cluster separation can be linked to differences in distance to stream, depth to water table, size of aquifer, local hydrology and geographical location.

Figure 2A and B show that groundwater level time series in clusters C1 and C6 have similar groundwater regimes. Time series in C1 show a relatively fast response (flashy) and overprinting of high peaks to varying degree, which is seen to a slightly lesser degree in C6. Inter- and intra-annual patterns are mostly absent. Groundwater levels in these two clusters are shallow (75% < 5 m) and with the wells relatively close to groundwater basin boundaries and streams in medium size aquifers (Figure 2D). Presumably, these clusters represent wells tapping mainly local groundwater flow systems (Giese et al., 2020). The pronounced flashiness is linked to interaction with streams (E. Haaf et al., 2020) and can also be seen in the low percentiles of the duration curves that are significantly steeper in the flashier C1 and C6 than other clusters (Figure 2B). Differences between C1 and C6 can be attributed to the different geographical areas, with C1 located in more extensive aquifers far downstream of the headwater catchment in the South and C6 located mainly in smaller alluvial aquifers in the Salzach and Inn catchments at the foot of the Alps (Figure 2C and SI, Figure S3).

Flashiness in cluster C2 is like C6, however, exhibiting intra-annual variations and weak inter-annual seasonality. Like C1 and C6, C2 is characterized as local flow due to the very shallow wells, however, wells are in intermediate locations in large aquifers. Therefore, dynamics are not closely coupled to the major rivers, which are at larger distances, but presumably to (unmapped) smaller creeks and to vegetation considering the shallow groundwater table.

C3 is less flashy than C2, but shows a similar inter- and intra-annual pattern, which can also be seen in the similarity of the two cluster’s head duration curves (Figure 2B). C3 wells are, similar to C2, located in larger aquifers, but are deeper and closer to streams, likely representing local and intermediate flow systems.

C4 has dominant inter-annual variability, which is linked to the larger distance to groundwater level and streams (E. Haaf et al., 2020). The larger inter annual variability in C4 is also seen in the less steep lower percentiles of the duration curves (Figure 2B) and is linked to mainly intermediate and regional flow systems.

Groundwater hydrographs in cluster C5 show a very distinct pattern compared to the remaining clusters. The HDC falls steeply at lower percentiles, following the flashier C1 and C6, until stabilizing and resembling more the weakly intra-annual dominated HDCs of C2 and C3, before crossing back to C1 and C6 at higher percentiles, due to cluster’s weak intra-annual periodicity. The distinct pattern and in-group similarity of the 14 wells in C5 is explained by their locations, concentrated near the Inn, which is regulated by run-of-the-river hydroelectric plants with pondage (Figure 2C).

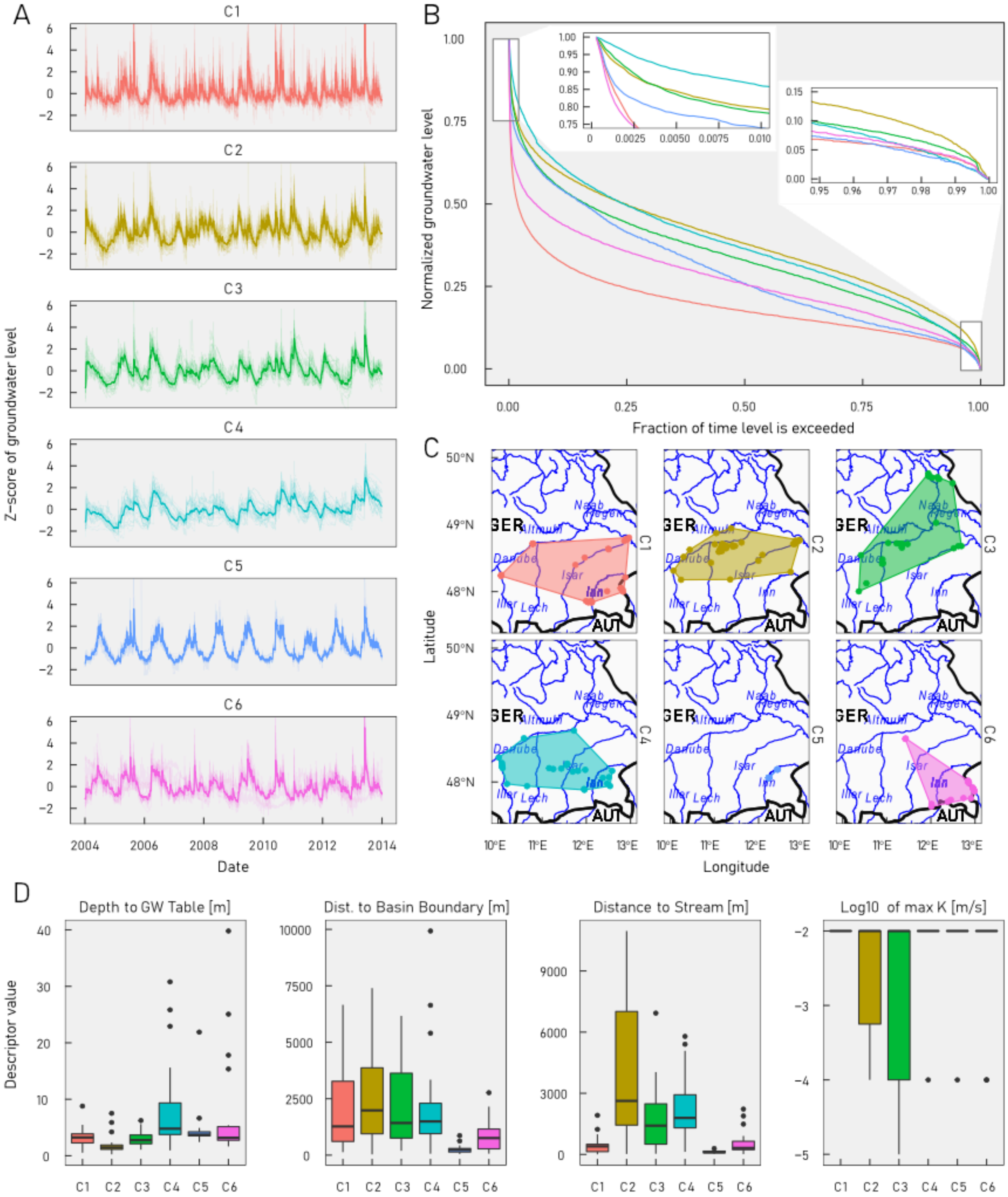

Figure 2. A. Time series within each cluster. B. Mean of groundwater level duration curve of color related to cluster in A. C. Location of cluster members with convex hull and stream network, ISO 3166-1 alpha-3 country codes. D. Hydrogeological descriptors of sites within each cluster.

3.2 Performance of HDC reconstruction

After regression analysis, models were found for all fifteen fixed percentage points. Regression models fitted on 30 different sets of hold-out data resulted in a distribution of results that are robust with regard to central tendency. Median XGB model performance on hold-out data expressed as R^2 is around 0.5, except for the lowest and upper percentiles (0.1%, 80-99%), i.e., wet and dry states, where goodness-of-fit declines (Figure 3). A lower fit at the extremes is expected since fewer data points make these values less robust compared to other percentiles. XGB models perform significantly better than MLR models that show a similar behavior across percentiles but with lower goodness-of-fit (median R^2: 0.3). Figure 3 also shows that the range of R^2 is large, which is very likely related to the size of the data set. The consequence of small data sets, when using hold-out data is that the evaluation data (here, n=32) may not be representative of the training data across sets of hold-out data. Further, when running models on the entire data set (training+evaluation), both XGB and MLR models show around 100% and 70% performance improvement from median R^2. Performance loss across hold-out data and against the entire data set indicates that generalization from the training set is moderate and likely to improve with larger data sets.

When comparing results to studies using an analogous methodology in streamflow, model results of R^2 between 0.72 and 0.99 are reported and analogous lower values in the extremes (Mohamoud, 2010; Shu & Ouarda, 2012). This study's performance is nearly 100% higher, however, neither hold-out data, cross-validation methods, or p-value adjustment for stepwise MLR is used. This means that models presented in these studies are likely overfitting and generalization outside of the data set could be questioned. The performance achieved on evaluation+training data by XGB and MLR models in this study would thus be more comparable and are in fact in parity with performance reported in streamflow studies.

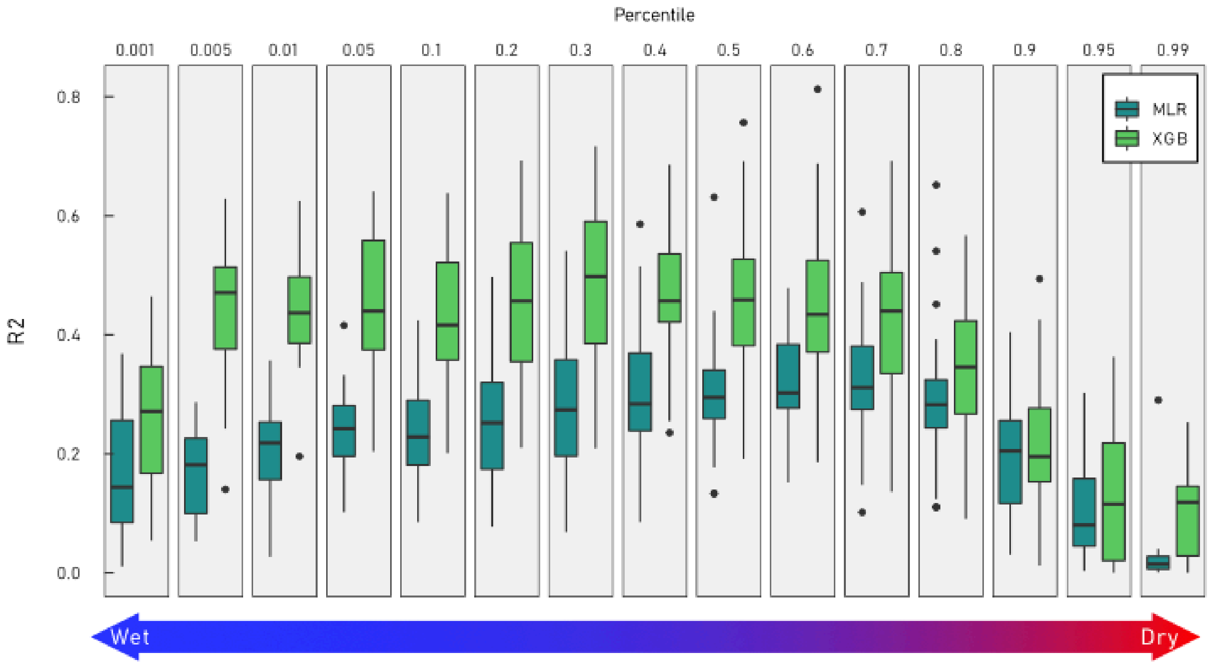

Figure 3. Performance of percentile regression models.

3.3 Dynamic Controls on Groundwater Levels

Relative predictor importance across percentage point models stratified by predictor class for MLR and XGB models respectively is shown in Figure 4. Standardized regression coefficients in MLR give both relative predictor importance (higher absolute value) but also the direction of the relationship between predictor and output variable (percentile level of HDC) through the sign of the coefficient (Figure 4A). Mean absolute SHAP value on the other hand, shows only relative predictor importance (Figure 4B). Further, for clarity of presentation, only the most salient variables are shown (MLR: variables are shown that are selected in at least 30% of hold-out data sets; XGB: only the top two predictors are shown per predictor class (based on overall mean absolute SHAP value).

The main result is that the importance of predictors varies across percentiles. This implies that different site (or system) descriptors to varying extents control the groundwater dynamics when the aquifer is moving from “wet” to “dry” states and vice versa. An example is distance to stream that is important through all aquifer states but dominating in wet states (both MLR and XGB, Figure 4A-B). Depth to the groundwater table, on the other hand, becomes more dominant when the aquifer is in dry states (only XGB, Figure 4B). A pattern that can be seen across all variables is that predictor strength declines significantly (approaches zero) at higher percentiles, which is also connected to lower goodness-of-fit at these percentiles (Figure 3). Consequently, predictability of percentiles coupled to groundwater drought is low.

Another important finding is that many of the most important predictors are consistently selected across both MLR and XGB as well as show a similar importance progression across percentiles (distance to stream, well elevation, average annual precipitation, broadleaved Forest and regional slope skewness). This means that many of the important variables have a sufficiently linear relationship with percentiles of groundwater head duration curves so that it can be picked up by MLR. For instance, MLR models show that percentage points of the HDC increases with distance to stream (the further away from streams, the less flashy the groundwater level). This is plausible and expected, since streams are the aquifer's given drainage boundary and known through previous regional scale empirical studies (e.g. Boutt, 2017; Giese et al., 2020; E. Haaf et al., 2020; Vidon, 2012). However, SHAP values of individual data points related to XGB prediction allows us to look more closely at linearity of relationships between HDC and predictor value ranges (Figure 5). The SHAP values reveal a more complex relationship, where the relationship between distance to stream and dynamics is constant up to about 500 m distance, turning into a linear relationship, where groundwater dynamics become less flashy with distance until reaching a plateau at about 3000 m distance. Here, presumably a decoupling between groundwater and stream occurs and a constant contribution to the HDC is reached (Figure 5). This effect is consistent across aquifer states, however weakens, when the groundwater level drops into dry states. The nonlinearity of relationships with threshold effects is common, as described below for variables selected in Figure 5:

- Average annual precipitation has relatively low impact on the HDC, which is also true for other climate predictors in this study. However, precipitation below approximately 800 mm leads to slightly less flashy dynamics in wet states. This can be coupled to less infiltration and recharge events. At higher precipitation rates, no systematic impact on HDC can be seen.
- Depth to groundwater table only affects the HDC when very shallow, approximately 2 m and above. Shallow water tables increase the percentile level accordingly, meaning that less flashiness may be expected. Sites, where groundwater levels are very shallow may be coupled to discharge zones. Here the aquifer is continuously replenished through recharge

from uplands with significant upward hydraulic gradients (Gribovszki et al., 2010; Winter, 2001). Generally, this effect increases in importance at higher percentiles, i.e., in a drier aquifer state

- If the percentage of broadleaved forests exceeds approximately 10%, groundwater levels become flashier in wet states, which can be linked to higher soil moisture, preferential flow and recharge than other land cover types, reducing surface runoff (Brinkmann et al., 2019; Dubois et al., 2021).
- If regional slopes are right skewed, sites are located in alluvial valley bottoms at the fringes of higher hill ranges (Ezra Haaf et al., 2020; Montgomery, 2001). In these locations amplitudes are expected to be higher due to front slope flow and mountain block recharge, which is also seen here particularly in wet aquifer states with lower SHAP values at higher slope skewness. Low slope skewness (<.0.3) on the other hand contributes to less flashy groundwater dynamics.

Overall, the progression of controls have implications not only for prediction but also conceptual understanding of groundwater dynamics in this region. The nonlinear relationships of groundwater dynamics and controls and the alternating dominance of these controls throughout different aquifer states are likely of interest, when studying e.g., vulnerability to drought events and climate change. Certainly, there is a need for a dedicated analysis of the dependence of controls on aquifer states, which was outside of the scope in this study.

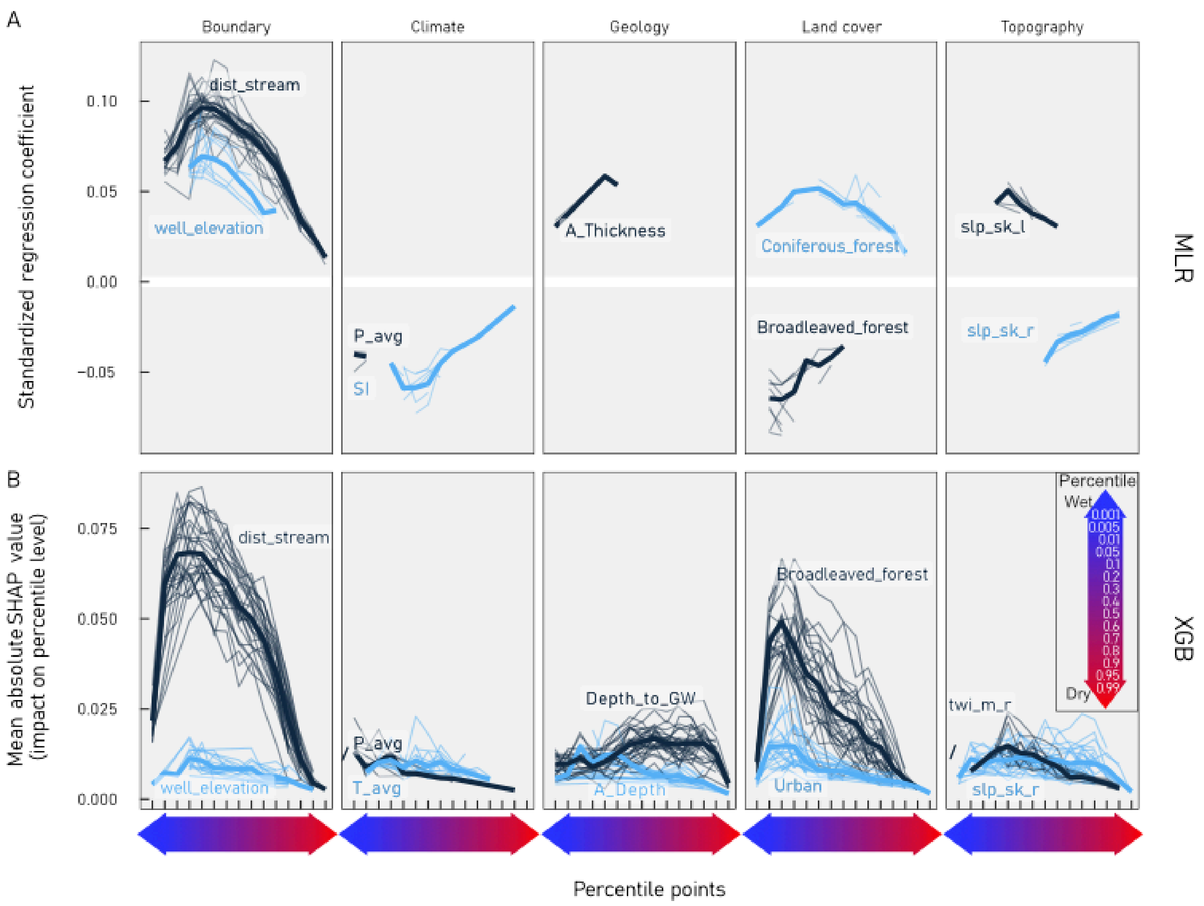

Figure 4. Relative predictor importance across percentage point models stratified by predictor class for MLR and XGB models (scales not comparable). Data from all hold-out datasets are plotted and fitted with a local polynomial regression to emphasize the central behavior of the data. A. Standardized regression coefficients show both relative predictor importance and direction of relationship between predictor and model output. B. Mean absolute SHAP value shows relative importance through impact on the output variable.

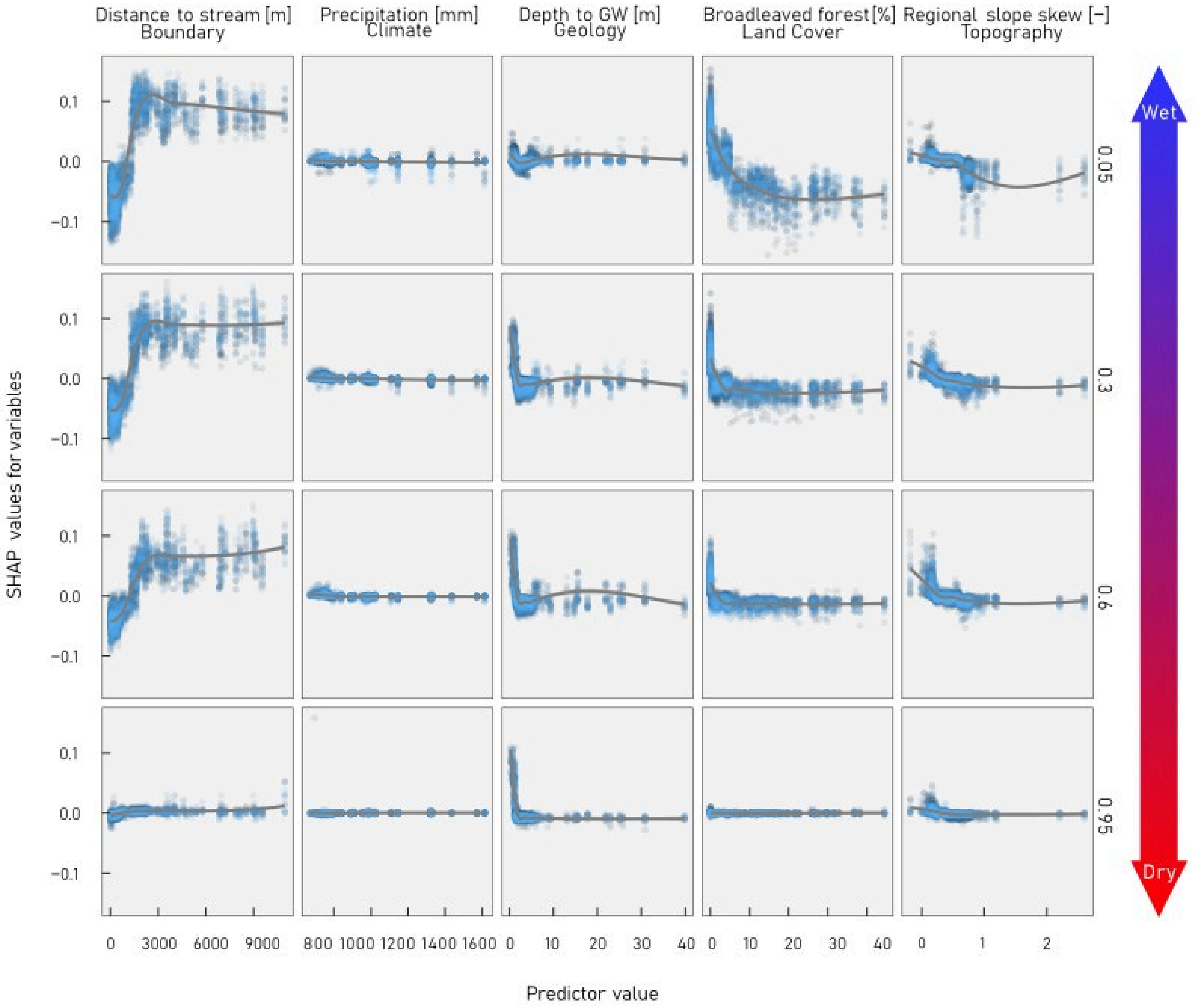

Figure 5. Relationship between feature value and impact on prediction for five selected variables across four percentiles. Each point represents an observation of the predictor variable and its SHAP value. Data from all hold-out datasets are plotted and fitted with a local polynomial regression to emphasize the central behavior of the data.

3.4 Performance of estimation techniques

Daily groundwater level time series were estimated at target sites, using representative models from each of MLR and XGB models as well as using the Nearest Neighbor method (NN). The XGB model had a higher KGE than NN at 120 of 157 (76%) sites, and a higher KGE than MLR at 136 of 157 (87%) sites. In consequence KGE is also significantly higher for XGB than NN and MLR (Figure 6A). Interestingly, MLR has a lower median KGE than NN, (slightly higher performance at the lower quartiles) which means that HDC modelling in the case of MLR deteriorates estimation on average, compared to the simple NN approach.

The higher performance of XGB can almost entirely be attributed to smaller amplitude errors between simulated and observed time series. Amplitude errors are expressed by the RMSE component of KGE, which is significantly improved when using XGB compared to NN and MLR (Figure 6B). The correlation component of the KGE on the other hand shows no significant differences between methods, meaning that timing errors between observed and simulated time series are not significantly improved through XGB or MLR (Figure 6C). As discussed by Mohamoud (2010), timing errors are coupled to the mismatch of time sequence in hydrograph events (here, e.g., recharge events) at donor and target sites. Still, from a water resources management perspective, the HDC estimation approach using XGB implies better estimation of the quantitative status of groundwater resources through significantly reduced amplitude errors.

Figure 6D shows that an optimal number of donor sites (neighbors) is generally reached with only 1-3 neighbors, as expressed by the maximum KGE. Sourcing more neighbors generally results in plateauing or even decrease of estimation performance across different groundwater regimes, as expressed by clusters C1 – C6. Although the number of optimal donor sites is consistent, C4 and C6 exhibit a sharp decline, when more than three or two source sites respectively are added. A possible reason for this is that these two clusters contain sites with significantly deeper groundwater tables (Figure 2D). This means that source sites with e.g., more shallow water table and therefore deviating groundwater response will be weighted in and cause a mismatch of time sequence, decreasing the quality of the predicted groundwater level time series at the target site.

Not only hydrogeological suitability of donor sites is important, but also proximity (Figure 6E). Performance decreases approximately with the natural logarithm of mean distance of neighbors. However, even at large mean distances to source sites (e.g. > 5 km), estimation performance at many sites may remain high. This is particularly the case for cluster C2 and C3. These cluster also show significantly higher performances by both HDC-based estimation techniques MLR and XGB. On the other hand, at sites with sufficient neighbors nearby (< 5 km), NN is preferred over MLR. Overall, however, XGB yields best performance independently of mean distance to neighbors.

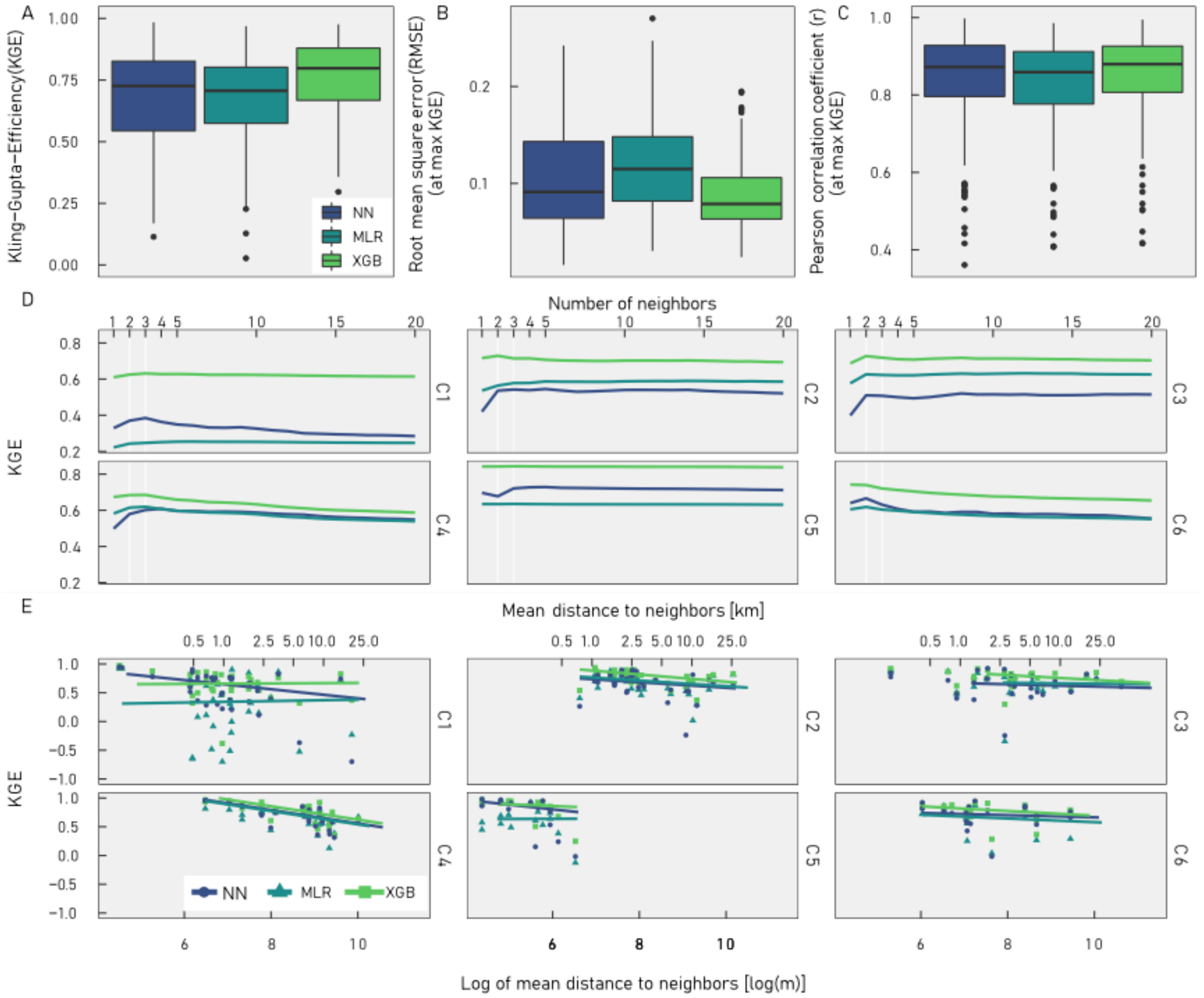

Figure 6. **A.-C.** Performance of estimation of daily groundwater level time series for the three approaches across all unmonitored sites, measured as KGE (A), RMSE (B), Pearson's r (C). **D**. Mean performance – measured as KGE - of the three estimation methods plotted against number of included neighboring sites, stratified by cluster. **E**. Performance of all sites – measured by KGE - plotted versus mean distance to neighbors, stratified by estimation method and cluster.

3.5 Hydrogeological Controls and Plausibility of Models

From a hydrogeological perspective, there are obviously missing descriptors to describe groundwater levels, such as aquifer properties, transmissivity and storativity. These are often not consistently available at the scale of this study (regional scale), or only with a low level of certainty at the level of 1-2 orders of magnitude (e.g., hydraulic conductivity in this study). However, it can be argued that the importance of storativity in this study is reduced, since normalization on a 0-1 scale of groundwater level time series reduce the importance of amplitude. Regarding hydraulic conductivity a relatively homogenous selection of sites is used (Figure 2D). When assuming order of magnitude similarity of hydraulic conductivity, the predictor aquifer thickness (A_thickness) may be considered a rough proxy. With these simplifications and proxy variables, model fits are acceptable, but still contain significant uncertainty, resulting in lower quality of time series prediction. Adding hydraulic properties, i.e., storativity values and less uncertainty regarding hydraulic conductivity to the set of predictors would likely improve the fit of regression models. It would further allow for use of more heterogeneous data sets. Different strategies to extract such hydraulic properties at wells from groundwater level time series of unconfined aquifers was recently proposed using transfer function noise models (Peterson & Fulton, 2019) and spectral analysis (Houben et al., 2022).

Apart from the missing hydraulic properties, other factors likely also play a role in explaining the moderate goodness-of-fit of the HDC models. Some of the uncertainty may be due to different hydraulic properties stratified within the zone of fluctuation. This is the case at only a few sites according to the borehole logs. Other sources of uncertainty may be found in data (groundwater level measurements, spatial resolution of DEM and climate data) or method of estimating physiographic and climatic descriptors.. Other reasons may be found in the overrepresentation of relatively shallow alluvial aquifers, particularly in the north-east of the study area. Using mean squared error as a loss function, regression models tend to better represent the bulk of the sites within the data set, which are mainly lowland riverine aquifers with shallow groundwater levels (local groundwater flow) and less so the peri-alpine river valleys in the north-east. A functional stratification of the data prior to HDC model building by e.g., the dominating predictor distance to stream, or more conceptually-based, using the hydrological landscape concept (Winter, 2001) may improve the predictive performance of the HDC models for sites that are less well represented. Using these functional pre-classifications should also improve transferability of methods to other study domains. For such an exercise, however, a data set would be necessary with sufficient data points that ensures robust models in each functional stratum.

3.6 Improvement of Donor Selection

The bias of the models towards well-represented hydrogeological settings as described above, also has consequences on donor-based reconstruction of time series at unmonitored sites. As discussed in section 3.4, differences in timing error between the three methods, NN, MLR and XGB, are very small and related to the similarity of time sequences between target and donor sites. A mismatch occurs, when inadequate donor sites are selected, which can be seen for example in cluster C4 and C6 (Figure 6D). Performance in these clusters declines with each additional donor and is presumably related to donors for intermediate/regional flow (C4) target sites being selected from (C6) sites that are located near rivers. In other words, donor sites have hydrological responses that differ from the target sites. Similar responses at sites with intermediate and regional flow systems can however be expected even at larger distances (Giese et al., 2020; Haaf & Barthel, 2018). In consequence, careful selection of donor sites is crucial to the performance of the method (also pointed out by authors applying the approach to streamflow: e.g., Hughes & Smakhtin, 1996; Shu & Ouarda, 2012; Smakhtin, 1999) and geographical proximity should not always be the main or sole selection criteria for source sites.

Likely, a cleverer approach than solely proximity for donor site selection, would surely improve the performance of the presented approach significantly. Such a strategy could be based on a distance metric that uses physiographic and climatic site descriptors for quantification of similarity between sites, as proposed for streamflow by Shu et al, 2012. However, after studying the nonlinearity of relationships between site descriptors and groundwater dynamics, a non-continuous approach may be more useful. Often, step changes could be seen, which indicates that a discrete classification approach may provide a more optimal pool of donor sites. Such classes of similar responses could be developed from the SHAP values in Figure 5, for example, that neighbors must be within the same distance to stream, i.e., within one of three classes (1-500m, 500-1500, > 1500m). For many of the sites, however, nearby sites still provide the most adequate timing of events. Therefore, any of the donor selection strategies discussed above must be combined with an approach that applies weights to donors within the similar class based on proximity.

4 Conclusions

Using the presented method, groundwater head duration curves can be transferred based on comparative regional analysis of map-derivable site descriptors from monitored to unmonitored sites. Neighboring donor sites can then be used to successfully reconstruct the daily groundwater level time series based on the transferred duration curve. Apart from time series estimation at unmonitored sites - in essence spatio-temporal interpolation - the modelling approach also gives insight into hydrological processes through identification of significant controls. Specifically, at the study site, controls on groundwater dynamics were nonlinear, which favors use of Machine

Learning (i.e., gradient boosted regression trees) over multiple linear regression and therefore makes possible improved conceptual hydrogeological understanding as well as higher predictive skill. The method and results were robust as tested through nested cross-validation, however, require thorough testing with larger data sets for application in other hydrogeological settings.

The study also showed that only 1-3 neighboring donor sites are generally necessary to optimally reconstruct time series of unmonitored sites. Further, the fewer nearby donor sites are available, the more benefit can be drawn from using the proposed comparative regional analysis approach, compared to nearest neighbor averaging of time series. Importantly, the selection of donor sites was identified as a key factor to improve predictive skill and should be expanded on using a combination of geographical proximity and functional classes of groundwater sites from which to draw appropriate neighbors. Finally, the study shows ways forward to investigate the dynamic nature of controls on groundwater levels, which may provide valuable insight to studies of recharge seasonality, droughts and floods.

Author Contributions

Haaf conceived the study with input from all co-authors. Haaf performed the statistical analysis and wrote the manuscript. All co-authors edited and revised the manuscript and approved the final version.

Acknowledgments

The authors would like to thank the German federal state agency Bayerisches Landesamt für Umwelt (LfU, https://www.lfu.bayern.de) for the provision of data and supporting information. Big thanks to Lars Rosén for valuable comments.

Open Research

Groundwater time series cannot be provided publicly by the authors based on the data usage agreement with the LfU, but can be downloaded from https://www.gkd.bayern.de/en/groundwater/upper-layer and https://www.gkd.bayern.de/en/groundwater/deeper-layer. The selected station names are provided in the Supplementary Information. Processed data will be made available on zenodo after

acceptance. Code for reproduction of results can be obtained from the corresponding author. All the analysis was performed in the statistical language R (R Development Core Team, 2022) using apart from the packages mentioned in the body “tidyverse”, “lubridate”, “rsample”, “vtreat,” and “selectiveInference” The authors thank the contributors of all these packages.

References

Bakker, M., & Schaars, F. (2019). Solving Groundwater Flow Problems with Time Series Analysis: You May Not Even Need Another Model. *Ground Water*. https://www.ncbi.nlm.nih.gov/pubmed/31347160

Barthel, R., & Banzhaf, S. (2016). Groundwater and Surface Water Interaction at the Regional-scale – A Review with Focus on Regional Integrated Models. *Water Resources Management, 30*(1), 1-32. journal article. http://dx.doi.org/10.1007/s11269-015-1163-z

Barthel, R., Haaf, E., Giese, M., Nygren, M., Heudorfer, B., & Stahl, K. (2021). Similarity-based approaches in hydrogeology: proposal of a new concept for data-scarce groundwater resource characterization and prediction. *Hydrogeology Journal*.

Berg, S. J., & Sudicky, E. A. (2019). Toward Large-Scale Integrated Surface and Subsurface Modeling. *Ground Water, 57*(1), 1-2. https://www.ncbi.nlm.nih.gov/pubmed/30513544

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., & Savenije, H. (2013). *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*: Cambridge University Press.

Bossard, M., Feranec, J., & Otahel, J. (2000). CORINE land cover technical guide: Addendum 2000.

Boutt, D. F. (2017). Assessing hydrogeologic controls on dynamic groundwater storage using long-term instrumental records of water table levels. *Hydrological Processes, 31*(7), 1479-1497.

Brinkmann, N., Eugster, W., Buchmann, N., & Kahmen, A. (2019). Species-specific differences in water uptake depth of mature temperate trees vary with water availability in the soil. *Plant Biology, 21*(1), 71-81.

Butler, J. J., Knobbe, S., Reboulet, E. C., Whittemore, D., Wilson, B. B., & Bohling, G. C. (2021). Water well hydrographs: An underutilized resource for characterizing subsurface conditions. *Groundwater*.

Chen, T., & Guestrin, C. (2016). *XGBoost : A Scalable Tree Boosting System*. http://doi.acm.org/10.1145/2939672.2939785

Chen, Z., Grasby, S. E., & Osadetz, K. G. (2002). Predicting average annual groundwater levels from climatic variables: an empirical model. *Journal of Hydrology, 260*(1), 102-117. http://www.sciencedirect.com/science/article/pii/S0022169401006060

Collenteur, R. A., Bakker, M., Calje, R., Klop, S. A., & Schaars, F. (2019). Pastas: open source software for the analysis of groundwater time series. *Ground Water*. https://www.ncbi.nlm.nih.gov/pubmed/31347164

de Marsily, G., Delay, F., Gonçalvès, J., Renard, P., Teles, V., & Violette, S. (2005). Dealing with spatial heterogeneity. *Hydrogeology Journal, 13*(1), 161-183.

Dubois, E., Larocque, M., Gagné, S., & Meyzonnat, G. (2021). Simulation of long-term spatiotemporal variations in regional-scale groundwater recharge: contributions of a water budget approach in cold and humid climates. *Hydrol. Earth Syst. Sci., 25*(12), 6567-6589. https://hess.copernicus.org/articles/25/6567/2021/

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2019). Hydrogeological conceptual model building and testing: A review. *Journal of Hydrology, 569*, 310-329. https://dx.doi.org/10.1016/j.jhydrol.2018.12.007

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189-1232, 1144. https://doi.org/10.1214/aos/1013203451

G'Sell, M. G., Wager, S., Chouldechova, A., & Tibshirani, R. (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 78*(2), 423-444. <Go to ISI>://WOS:000369136600005

Giese, M., Haaf, E., Heudorfer, B., & Barthel, R. (2020). Comparative hydrogeology – reference analysis of groundwater dynamics from neighbouring observation wells. *Hydrological Sciences Journal, (accepted)*.

Green, T. R., Taniguchi, M., Kooi, H., Gurdak, J. J., Allen, D. M., Hiscock, K. M., et al. (2011). Beneath the surface of global change: Impacts of climate change on groundwater. *Journal of Hydrology, 405*(3-4), 532-560.

Gribovszki, Z., Szilágyi, J., & Kalicz, P. (2010). Diurnal fluctuations in shallow groundwater levels and streamflow rates and their interpretation – A review. *Journal of Hydrology, 385*(1-4), 371-383.

Haaf, E., & Barthel, R. (2018). An inter-comparison of similarity-based methods for organisation and classification of groundwater hydrographs. *Journal of Hydrology, 559*, 222-237.

Haaf, E., Giese, M., Heudorfer, B., Stahl, K., & Barthel, R. (2020). Physiographic and Climatic Controls on Regional Groundwater Dynamics. *Water Resources Research, 56*(10).

Haaf, E., Heudorfer, B., Giese, M., Stahl, K., & Barthel, R. (2020). Physiographic and climatic controls on groundwater dynamics on the regional scale. *(under Review)*.

He, Y., Bárdossy, A., & Zehe, E. (2011). A review of regionalisation for continuous streamflow simulation. *Hydrology and Earth System Sciences, 15*(11), 3539-3553.

Heudorfer, B., Haaf, E., Stahl, K., & Barthel, R. (2019). Index-Based Characterization and Quantification of Groundwater Dynamics. *Water Resources Research, 55*(7), 5575-5592. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024418

Houben, T., Pujades, E., Kalbacher, T., Dietrich, P., & Attinger, S. (2022). From Dynamic Groundwater Level Measurements to Regional Aquifer Parameters— Assessing the Power of Spectral Analysis. *Water Resources Research, 58*(5).

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal, 58*(6), 1198-1255.

Hughes, D. A., & Smakhtin, V. (1996). Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. *Hydrological Sciences Journal, 41*(6), 851-871. https://www.tandfonline.com/doi/abs/10.1080/02626669609491555

Jackson, C. R., Wang, L., Pachocka, M., Mackay, J. D., & Bloomfield, J. P. (2016). Reconstruction of multi-decadal groundwater level time-series using a lumped conceptual model. *Hydrological Processes*, n/a-n/a.

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. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR026065

https://agupubs.onlinelibrary.wiley.com/doi/pdfdirect/10.1029/2019WR026065?download=true

Liu, Y., & Just, A. C. (2021). *SHAPforxgboost: SHAP Plots for 'XGBoost', R package version 0.1.1*. Retrieved from https://CRAN.R-project.org/package=SHAPforxgboost

Lóaiciga, H. A., & Leipnik, R. B. (2001). Theory of sustainable groundwater management: an urban case study. *Urban Water, 3*(3), 217-228. http://www.sciencedirect.com/science/article/pii/S1462075801000401

Lundberg, S. M., Erion, G., Chen, H., Degrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56-67. https://dx.doi.org/10.1038/s42256-019-0138-9

https://www.nature.com/articles/s42256-019-0138-9.pdf

Mackay, J., Jackson, C. R., & Wang, L. (2014). A lumped conceptual model to simulate groundwater level time-series. *Environmental Modelling and Software, 61*, 229-245.

Marchant, B. P., & Bloomfield, J. P. (2018). Spatio-temporal modelling of the status of groundwater droughts. *Journal of Hydrology, 564*, 397-413.

Maxe, L. (2013). Bedömningsgrunder för grundvatten. *Sveriges geologiska undersökning SGU-rapport 2013, 1.*

McDonnell, J. J., & Woods, R. (2004). On the Need for Catchment Classification. *Journal of Hydrology, 299*(1), 2-3.

Mohamoud, Y. M. (2010). Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves. *Hydrological Sciences Journal, 53*(4), 706-724.

Montgomery, D. (2001). Slope Distributions, Threshold Hillslopes, and Steady-state Topography. *American Journal of Science, 301*, 432-454.

Naghibi, S. A., Hashemi, H., Berndtsson, R., & Lee, S. (2020). Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors. *Journal of Hydrology, 589.*

Peterson, T. J., & Fulton, S. (2019). Joint Estimation of Gross Recharge, Groundwater Usage, and Hydraulic Properties within HydroSight. *Groundwater, 57*(6), 860-876.

R Development Core Team. (2022). R: A language and environment for statistical computing: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rajaee, T., Ebrahimi, H., & Nourani, V. (2019). A review of the artificial intelligence methods in groundwater level modeling. *Journal of Hydrology, 572*, 336-351. Review. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85062607865&doi=10.1016%2fj.jhydrol.2018.12.037&partnerID=40&md5=d92bcea98e88e59d453c5cb7fd5feddd

Ridolfi, E., Kumar, H., & Bárdossy, A. (2020). A methodology to estimate flow duration curves at partially ungauged basins. *Hydrology and Earth System Sciences, 24*(4), 2043-2060. https://dx.doi.org/10.5194/hess-24-2043-2020

https://hess.copernicus.org/articles/24/2043/2020/hess-24-2043-2020.pdf

Rinderer, M., McGlynn, B. L., & van Meerveld, H. J. (2017). Groundwater similarity across a watershed derived from time-warped and flow-corrected time series. *Water Resources Research, 53*(5), 3921-3940.

Rinderer, M., Meerveld, H. J., & McGlynn, B. L. (2019). From Points to Patterns: Using Groundwater Time Series Clustering to Investigate Subsurface Hydrological Connectivity and Runoff Source Area Dynamics. *Water Resources Research, 55*(7), 5784-5806.

Rinderer, M., van Meerveld, H. J., & Seibert, J. (2014). Topographic controls on shallow groundwater levels in a steep, prealpine catchment: When are the TWI assumptions valid? *Water Resources Research, 50*(7), 6067-6080.

Rinderer, M., van Meerveld, I., Stähli, M., & Seibert, J. (2016). Is groundwater response timing in a pre-alpine catchment controlled more by topography or by rainfall? *Hydrological Processes, 30*(7), 1036-1051.

Ruybal, C. J., Hogue, T. S., & McCray, J. E. (2019). Evaluation of Groundwater Levels in the Arapahoe Aquifer Using Spatiotemporal Regression Kriging. *Water Resources Research, 55*(4), 2820-2837. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR023437

Shu, C., & Ouarda, T. B. M. J. (2012). Improved methods for daily streamflow estimates at ungauged sites. *Water Resources Research, 48*(2).

Sivakumar, B., & Singh, V. P. (2012). Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework. *Hydrology and Earth System Sciences, 16*(11), 4119-4131.

Smakhtin, V. Y. (1999). Generation of natural daily flow time-series in regulated rivers using a non-linear spatial interpolation technique. *Regulated Rivers: Research & Management, 15*(4), 311-323. https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-1646%28199907/08%2915%3A4%3C311%3A%3AAID-RRR544%3E3.0.CO%3B2-W

Sugiyama, H., Vudhivanich, V., Whitaker, A. C., & Lorsirirat, K. (2003). STOCHASTIC FLOW DURATION CURVES FOR EVALUATION OF FLOW REGIMES IN RIVERS. *JAWRA*

Journal of the American Water Resources Association, 39(1), 47-58. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-1688.2003.tb01560.x

Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proc Natl Acad Sci U S A, 112*(25), 7629-7634. https://www.ncbi.nlm.nih.gov/pubmed/26100887

Varouchakis, E. A., Guardiola-Albert, C., & Karatzas, G. P. (2022). Spatiotemporal Geostatistical Analysis of Groundwater Level in Aquifer Systems of Complex Hydrogeology. *Water Resources Research, 58*(3), e2021WR029988. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021WR029988

Vidon, P. (2012). Towards a better understanding of riparian zone water table response to precipitation: surface water infiltration, hillslope contribution or pressure wave processes? *Hydrological Processes, 26*(21), 3207-3215.

Vogel, R. M., & Fennessey, N. M. (1995). FLOW DURATION CURVES II: A REVIEW OF APPLICATIONS IN WATER RESOURCES PLANNING. *JAWRA Journal of the American Water Resources Association, 31*(6), 1029-1039. Article. https://www.scopus.com/inward/record.uri?eid=2-s2.0-0029481769&doi=10.1111%2fj.1752-1688.1995.tb03419.x&partnerID=40&md5=d88f3813ee3ec385ff41ad795eb7319b

Von Asmuth, J. R. (2012). *Groundwater System Identification Through Time Series Analysis*.

Voss, C. I. (2005). The future of hydrogeology. *Hydrogeology Journal, 13*(1), 1-6.

Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment Classification and Hydrologic Similarity. *Geography Compass, 1*(4), 901-931.

Winter, T. C. (2001). The concept of hydrologic landscapes. *Journal of the American Water Resources Association, 37*(2), 335-349. <Go to ISI>://WOS:000171459900008

Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrol. Earth Syst. Sci., 25*(3), 1671-1687. https://hess.copernicus.org/articles/25/1671/2021/

Wunsch, A., Liesch, T., & Broda, S. (2022). Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. *Nat Commun, 13*(1), 1221. https://www.ncbi.nlm.nih.gov/pubmed/35264569

Yokoo, Y., & Sivapalan, M. (2011). Towards reconstruction of the flow duration curve: development of a conceptual framework with a physical basis. *Hydrology and Earth System Sciences, 15*(9), 2805-2819. https://dx.doi.org/10.5194/hess-15-2805-2011