Received: 4 September 2019 | Accepted: 27 February 2020
DOI: 10.1002/hyp.13740

RESEARCH ARTICLE

WILEY

Machine learning based identification of dominant controls on runoff dynamics

Henning Oppel[1,2] | Andreas H. Schumann[1]

[1]Ruhr-University Bochum, Institute for Hydrologic Engineering and Water Management, Bochum, Germany

[2]University of Kassel, Center for Environmental System Research, Kassel, Germany

Correspondence
Henning Oppel, Ruhr-University Bochum, Institute for Hydrologic Engineering and Water Management, Bochum, Germany.
Email: henning.oppel@rub.de

Funding information
Bavarian Ministry of the Environment

Abstract

Hydrological models used for flood prediction in ungauged catchments are commonly fitted to regionally transferred data. The key issue of this procedure is to identify hydrologically similar catchments. Therefore, the dominant controls for the process of interest have to be known. In this study, we applied a new machine learning based approach to identify the catchment characteristics that can be used to identify the active processes controlling runoff dynamics. A random forest (RF) regressor has been trained to estimate the drainage velocity parameters of a geomorphologic instantaneous unit hydrograph (GIUH) in ungauged catchments, based on regionally available data. We analyzed the learning procedure of the algorithm and identified preferred donor catchments for each ungauged catchment. Based on the obtained machine learning results from catchment grouping, a classification scheme for drainage network characteristics has been derived. This classification scheme has been applied in a flood forecasting case study. The results demonstrate that the RF could be trained properly with the selected donor catchments to successfully estimate the required GIUH parameters. Moreover, our results showed that drainage network characteristics can be used to identify the influence of geomorphological dispersion on the dynamics of catchment response.

KEYWORDS
catchment classification, catchment similarity, drainage velocity, geomorphologic unit hydrograph, machine learning, ungauged catchments

1 | INTRODUCTION

It is argued that floods caused by extreme precipitation are becoming more frequent due to climatic changes. In the last two decades, such flood events have been the main reason for major flood damages in Germany (Uhlemann, Thieken, & Merz, 2010). The increased relevance of rainfall induced floods is caused by significant changes of precipitation extremes (Murawski, Zimmer, & Merz, 2016), and requires reconsideration of flood forecasting systems. Especially the lead time of forecast becomes more crucial for public safety and has to be prolongated by improved hydrological models.

Inside the hydrological model, processes are numerically reproduced in the chosen model structure and by the parameters that have been fitted to data of past events. Because streamflow records are only available as single point records, the required data base is rarely available at the desired locations where flood risk assessments are needed.

A common strategy to overcome this problem is to utilize regionalized data from stream gauging sites for ungauged locations. The requirement for utilizing regionally transferred data is that the respective catchments have to be hydrological similar, i.e., the active processes within these catchments are concurrent (Blöschl & Sivapalan, 1995). Although several studies have developed a framework for defining hydrologic similarity in different climatic regions and spatial scales (Wagener, Sivapalan, Troch, & Woods, 2007; Winter, 2001), to date a clear definition for similarity and for a clear catchment classification scheme has not been developed.

On large scale (continental to global), several studies concluded that classifications based on climate signatures, for example, annual sum of precipitation or aridity index, were useful regarding runoff signatures or parameter transfers (Addor et al., 2018; Bárdossy, Huang, & Wagener, 2016; Beck et al., 2016; Carrillo et al., 2011; Kuentz, Arheimer, Hundecha, & Wagener, 2017; Ragettli, Zhou, Wang, Liu, & Guo, 2017; Sawicz, Wagener, Sivapalan, Troch, & Carrillo, 2011; Singh, Archfield, & Wagener, 2014; Zhang, Chiew, Li, & Post, 2018). Kuentz et al. (2017) amended these findings by stating that the link between climate and runoff signatures was especially relevant for long-term runoff signatures like flood marks. They also found that flood flashiness indicators were linked to basin shape and baseflow indices to soil or geologic characteristics.

Studies at smaller scales, performing parameter or runoff signature transfer within a defined region, found more variability than what the results from the larger scale had suggested (Singh, Archfield, & Wagener, 2014). However, on the smaller scale, several studies confirmed that runoff dynamics were connected to basin shape (flow paths, concentration times, drainage density, etc.), land cover and soil properties within climatically homogenous regions (Bárdossy, 2007; Brunner et al., 2018; Drouge et al., 2002; Dunn & Lilly, 2001; Grimaldi, Petroselli, Alonso, & Nardi, 2010; Singh, Archfield, & Wagener, 2014; Soulsby, Tetzlaff, & Hrachowitz, 2010; Wigington, Leibowitz, Comeleo, & Ebersole, 2013; Yadav, Wagener, & Gupta, 2007). Although studies from Steinschneider, Yang, and Brown (2015) revealed that spatial proximity could be used to define reasonable catchment groups, it was concluded by several other studies that proximity is just a proxy for other driving factors (Sawicz et al., 2011). This also confirmed the key statements of Winter (2001) and Wagener et al. (2007) that a general classification framework should take climate, geology, surface form and runoff signatures into account. Although catchment characteristics can be linked to active hydrological processes at the catchment scale by utilizing tracer data (Klaus & McDonnell, 2013), the availability of such data is even more restricted and its recording more expensive compared to streamflow gauging data. At the basin scale, a clear connection between hydrologic processes and catchment characteristics, derived from analysis of streamflow data, remains thus to be identified.

This might go back to the fact that most previous studies either transferred parameters for sophisticated hydrological models or regionalized runoff signatures. As a result, findings were to a large extent restricted by the procedure used, that is, the a priori perception of the hydrologic process. Shen et al. (2018) and Mount et al. (2016) specified an alternative approach for analysis. They suggested using data-driven methods, Deep Learning in particular, as a new method of scientific analysis. Key benefits were supposed to be design free of preconception and the ability to adapt to specific problems. The learning procedure was basically a procedure of hypothesis-testing to provide new insights into hydrological processes. Mount et al. (2016) complemented these findings by demonstrating that data-driven models could be a useful supplement to classical physics-based models. Even though data-driven models, especially machine learning (ML) algorithms, were applied in many hydrological applications (see Elshorbagy, Corzo, Srinivasulu, & Solomatine, 2010a; Elshorbagy, Corzo, Srinivasulu, & Solomatine, 2010b; Solomatine & Ostfeld, 2008; Yaseen, El-shafie, Jaafar, Afan, & Sayl, 2015 for reviews), their focus was mainly on regression and classification results (Brunner et al., 2018 and Heřmanovský, Havlíček, Hanel, & Pech, 2017), or on trained model structures, for example, Singh, Archfield, and Wagener (2014) analyzed the trained structures of a decision tree. These ways of using ML for knowledge extraction are a first step towards exploiting the potential of ML for process analysis.

The trained algorithms inherit a hypothesis about the processes they were trained to be reiterated. This hypothesis emerged from a competition between numerous other hypotheses within the training phase. Herein lies the true power of ML-enforced analysis of processes. Because the algorithm has no constrains in its ability to build process abstractions, it will be able to test and discard more hypothesis than a single human researcher could in a comparable amount of time.

However, it has to be recognized that the result of a fitted algorithm might not be transferrable in the most cases, due to the inherit process uncertainty. ML-algorithms are usually trained for a single purpose; therefore, results tend to overfit the problem. A solution for this issue is the use of ensemble techniques and validation of the results based on large data sets. Though analyses based on trained ML-algorithms will benefit from the performed learning procedure, they not only considered the regression and classification results from a trained algorithm. In addition, they gained knowledge from analyzing the internal structures of their decision trees. Such decision tree algorithms are compelling for because they are easy to interpret. While there are more powerful algorithms like artificial neural networks or ensemble techniques as the random forest (RF), their internal structures are not understandable for the human logic (Kelleher, MacNamee, & D'Arcy, 2015).

Here, we pursued a new approach for data-driven process research. We studied the training phase of an RF that was trained to characterize the runoff dynamics in a (pseudo)-ungauged catchment. The training data was taken from gauged catchments in regional neighborhood of the target catchment. The questions we asked was “Which data set minimizes the parameter estimation error?” Using a step-by-step analysis of the learning procedure, we assessed which data sets should be used for an optimal training of the RF. The selection of ideal donor catchments, based on model performance, was followed by an analysis of the catchments characteristics. The groups of donor catchments as defined by ML were analyzed for homogenous grouping of various catchment characteristics, that is, basin shape (area, Horton ratios), land cover and soil properties. We thus recreated the ML-derived classification with catchment characteristics that could be used to identify donor catchments for an ungauged target catchment.

We used drainage velocity as a proxy for flow dynamics. In the concept of the geomorphologic instantaneous unit hydrograph (GIUH), drainage velocity defines the variance of catchment response while the shape of catchment response is defined by the geomorphology of the basin (Rodríguez-Iturbe & Valdés, 1979). Consequentially drainage velocity is directly connected to catchment response, that is, the process of runoff concentration. Using drainage velocity also offered other benefits that we subsequently implemented. First, this parameter can be derived analytically from runoff and precipitation data as well as it can be calibrated to a GIUH-model. Second, the GIUH directly relates catchment properties to discharge characteristics and is therefore a valuable tool for predictions in ungauged basins (Hrachowitz et al., 2013; Rigon, Bancheri, Formetta, & de Lavenne, 2016). Yet, its application had been limited due to missing solution for event-wise parametrization. The use of machine learning presents a major step forward to overcome this problem.

Two basins in in south-east Germany were used in this study. Uhlemann et al. (2010) found that the frequency of floods caused by heavy rainfall significantly increased in this region. The basins of the rivers Regen and the Upper Main are both situated in a mid-range mountainous area of Bavaria.

2 | METHODS AND DATA

2.1 | Case study catchments

We used rainfall-runoff events in hourly temporal resolution from 33 gauges. Twenty-two gauges are located in the basin of the Upper Main (Figure 1, upper left). The Upper Main basin covers an area of 4,223 km^2. Runoff from the entire basin is observed at gauge Kemmern at the outlet of the basin. Other gauges observe catchments with an area from 11.1 km^2 up to 500.4 km^2. The basin is dominated by agricultural land use, only the northern parts of the basin possess larger forested areas. The Regen basin (Figure 1, lower left), is located in the mid-range mountainous Bavarian forest. The largest catchment, Marienthal covers the entire basin area of 2,590 km^2. Data for 11 gauges were available in this basin.

For each gauge, continuous time series of hourly discharge were available from 1999 to 2012. Additionally, interpolated time series of hourly precipitation and temperature for each sub-basin were calculated by means of Thiessen polygons. We applied Thiessen polygon interpolation in this study to preserve the natural variance of observed precipitation measures. Possible precipitation volume errors were accepted, as our analyses focused on runoff dynamics. The data were provided by the Bavarian Ministry of the Environment (2018) and data from German Weather Service (Deutscher Wetterdienst DWD [German Weather Service], 2019). We extracted the five highest flood events per year from the continuous discharge time series. Once the flood time stamps of the highest flood values were identified, we extracted the discharge and precipitation events manually. The number of events we looked at (five per year) was considered adequate considering the quantity of data needed for our data base and the heights of the event discharge. A total of 831 events have been extracted, with an average of 25 events per gauge. The number of events varied from gauge to gauge because some events were removed from the data set, as they were influenced by snow impact. Floods caused or influenced by snow melt were excluded due to our focus on rainfall-induced floods.

2.2 | Catchment characteristics

For each sub-basin, several catchment characteristics were calculated in order to determine the catchment classification scheme (summarized in Table 1). From Corine land cover data (Bossard, Feranec, & Otahel, 2000), the percentage of agricultural (AGR) and forested areas (FOR) were determined to characterize the land cover of the

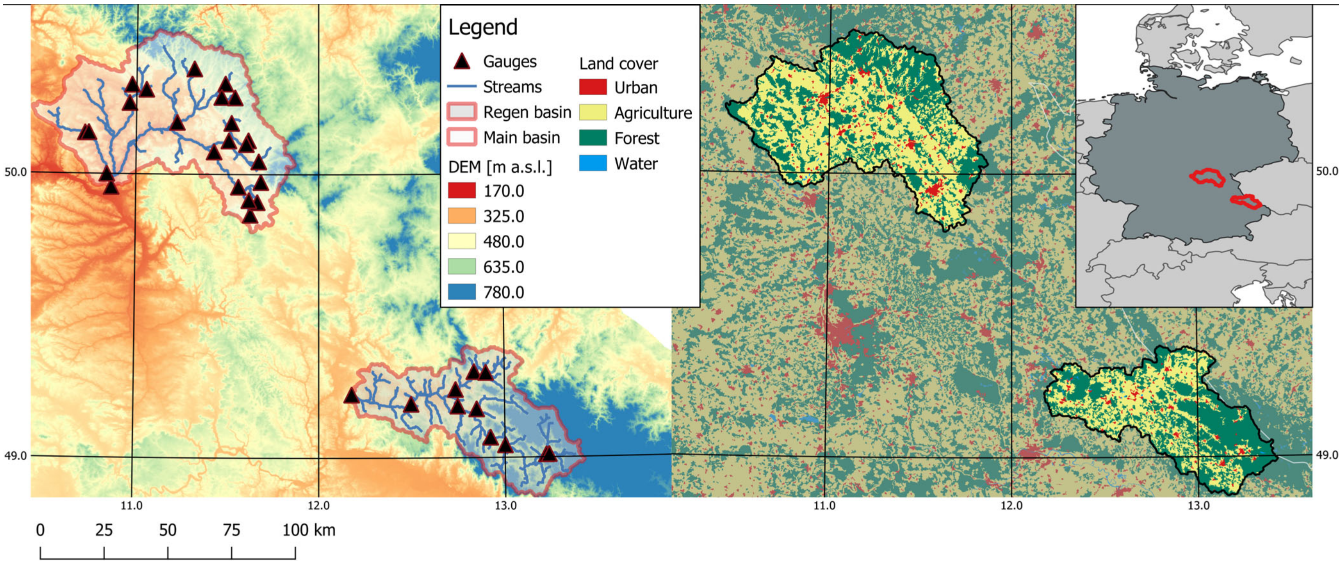

FIGURE 1 (Left) Digital elevation model (from SRTM data) and gauges (triangles) in the case study basins Upper Main (upper left) and Regen (lower right) located in south-east Germany. (Right) Land cover classes of the case study basins derived from CORINE land cover

TABLE 1 Catchment characteristics of the Regen and Upper Main catchment

Characteristic	ARE	ELE	SLO	AGR	FOR	TPV	R_A	R_L	R_B	L_{MAX}
Unit	[km^2]	[m a.s.l.]	[°]	[%]	[%]	[mm]	[–]	[–]	[–]	[km]
Regen										
Chamerau	1,356.5	707.0	8.71	33	64	241.2	1.90	2.91	1.93	0.9
Eschelkam	178.0	531.0	6.02	62	37	174.7	2.52	3.57	1.23	8.8
Furth im Wald	276.6	538.9	6.47	61	37	193.9	2.50	3.56	1.39	8.7
Kienhof	2,174.1	623.7	7.55	43	54	250.8	1.81	2.80	2.11	0.9
Koetzing	224.4	701.4	10.30	28	69	238.6	1.78	3.03	1.51	3.6
Kothmaissling	405.0	520.7	6.45	60	38	216.8	2.51	3.57	1.57	8.7
Lohmannmuehle	115.9	858.1	8.79	12	84	260.1	1.72	2.36	1.16	1.1
Marienthal	2,590.4	595.9	7.03	44	53	258.5	1.69	2.61	2.19	0.9
Sägemuehle	839.3	753.4	8.54	28	69	246.9	1.53	2.58	1.42	8.4
Teisnach	626.6	781.8	8.42	24	73	250.0	2.07	3.16	1.77	4.3
Zwiesel	293.4	873.2	9.65	9	88	263.2	2.13	3.21	1.59	4.3
Upper Main										
Adlerhuette	33.39	551.1	4.87	59	39	232.4	1.77	2.95	0.46	3.7
Bad Berneck	99.79	710.3	8.13	18	80	238.0	3.28	4.30	1.00	20.4
Bayreuth	340.28	488.0	5.02	56	35	232.1	3.10	4.20	1.90	16.8
Bernstein a.W.	35.38	611.5	7.88	29	68	186.8	3.23	4.27	0.90	20.2
Coburg	346.34	457.2	6.17	49	44	310.3	2.05	3.05	1.35	6.6
Friedersdorf	11.14	612.1	6.42	32	66	184.0	2.64	3.68	0.40	10.4
Gampelmuehle	62.2	455.0	3.93	70	28	204.7	2.76	3.90	1.18	8.6
Kauerndorf	246.23	516.8	5.74	61	36	259.3	2.39	3.44	1.32	12.2
Kemmern	4,223.84	432.9	5.39	54	41	330.7	2.22	3.27	2.40	8.2
Leucherhof	380.52	351.2	4.56	49	50	345.6	1.70	2.99	1.38	6.9
Lohr	165.3	366.0	4.18	46	54	417.7	1.67	2.96	1.14	6.9
Moenchroeden	70.7	433.3	6.50	37	51	216.3	0.74	2.44	1.00	1.5
Oberhammer	64.25	582.5	6.62	51	49	194.8	2.29	3.31	0.65	7.3
Oberlauter	31.56	457.9	6.48	41	59	207.2	2.10	3.20	0.70	2.4
Pfarrweisach	36.67	363.0	4.66	54	42	439.5	3.42	4.45	0.85	29.0
Rieblich	118.22	608.7	7.37	21	77	183.7	2.51	3.63	1.15	9.9
Schlehenmuehle	70.95	479.8	3.56	55	42	306.5	3.11	4.22	1.15	18.4
Unterlangenstadt	713.87	528.6	7.46	37	59	196.7	2.41	3.45	1.69	10.1
Untersteinach	73.52	635.7	8.42	33	63	203.3	3.03	4.12	1.28	14.0
Unterzettlitz	500.35	454.2	4.81	57	37	258.9	2.51	3.65	1.75	8.4
Wallenfels	96.45	576.5	9.27	22	76	186.0	2.62	3.66	1.02	10.2
Wirsberg	76.86	549.0	5.18	59	38	224.9	1.90	3.19	0.79	7.2

Note: Area (ARE), elevation (ELE) and slope (SLO) derived from DEM, share of agricultural (AGR) and forested areas (FOR) from land cover data, total pore volume in the upper 2 m soil layer (TPV) from soil data. Horton rations (R_A, R_L, R_B) and L_{Max} derived from DEM and stream network.

sub-basins. Topographical characteristics slope (SLO) and mean elevation (ELE) as well as the basin area (ARE) were derived from the DEM (Jarvis, Reuter, Nelson, & Guevara, 2008). The soil was characterized with the total pore volume in the upper 2 m of the soil layer (TPV) (Federal Institute for Geosciences and Natural Resources, 2006). Horton ratios of the drainage area R_A, stream lengths R_L, the bifurcation ratio R_B as well as the maximum flow length of the highest order stream within the basin L_{MAX} were determined as characteristics of the drainage system. The drainage network and the Horton rations were calculated following the methods proposed by Grimaldi, Petroselli, and Nardi (2011) and Moussa (2009). The Horton ratios are the only catchment characteristics affecting the simulation of the time series directly, as they were used to parametrize the GIUH (Section 2.2). All other characteristics were solely used for catchment classification.

2.3 | Geomorphologic instantaneous unit hydrograph model

A hydrological model was used to reproduce the flood events at hourly temporal resolution. The modelling study was performed in a leave-one-out procedure to test the ability of the model for runoff prediction in ungauged basins. From the variety of existing GIUH models (Rigon et al. (2016) and Singh, Mishra, and Jain (2014) provided a thorough reviews), we chose the rather simplistic approach by Rosso (1984) due to its common application and its quick and robust results. The ordinates of the GIUH were calculated depending on time step t:

$$\mathrm{GIUH}(t) = \frac{1}{k\Gamma(n)} \cdot \left(\frac{t}{k}\right)^{n-1} \cdot e^{-t/k} \tag{1}$$

with

$$n = 3.29 \cdot R_{\mathrm{B}}^{0.72} \cdot R_{\mathrm{A}}^{-0.78} \cdot R_{\mathrm{L}}^{0.07} \tag{2}$$

$$k = 0.70 \cdot \frac{L_{\mathrm{Max}}}{v_{\mathrm{D}}} R_{\mathrm{B}}^{-0.48} \cdot R_{\mathrm{A}}^{0.48} \cdot R_{\mathrm{L}}^{-0.48} \tag{3}$$

with v_{D} being the drainage velocity in [m/s] of the considered event. The constant numbers in Equations 2 and 3 are part of the model proposed by Rosso (1984) and have been determined for multiple combinations of R_{B}, R_{A} and R_{L}. Although the model and the constants derived by Rosso (1984) are commonly applied, the uncertainty concerning the constants has to be noted. However, the Horton ratios of the majority of catchments used in this study were within the parameter boundaries tested by Rosso (1984). As Grimaldi et al. (2011) stated, the drainage velocity should be considered as a calibration parameter, although it is connected to the concentration time. In order to focus on runoff dynamics, we omitted a runoff generation procedure and used empirical runoff ratios to define effective precipitation values. Please note that the application of the used GIUH-model is restricted to the simulation of rainfall-induced flood events, based on the effective precipitation as no other storage components for snow or groundwater are considered.

The performance of the GIUH-model was evaluated with the Kling-Gupta Efficiency (KGE). This efficiency criterion combines three aspects of hydrograph recreation: ratios α and β, of simulated and observed standard deviation σ and averages μ, as well as the linear correlation coefficient of simulated and observed hydrograph r (Gupta, Kling, Yilmaz, & Martinez, 2009):

$$\mathrm{KGE} = 1 - \sqrt{(1-\alpha)^2 + (1-\beta)^2 + (1-r)^2} \tag{4}$$

with

$$\alpha = \frac{\sigma_{\mathrm{Sim}}}{\sigma_{\mathrm{Obs}}}; \beta = \frac{\mu_{\mathrm{Sim}}}{\mu_{\mathrm{Obs}}}; r = \frac{\mathrm{Cov}_{\mathrm{Sim;Obs}}}{\sigma_{\mathrm{Sim}} \cdot \sigma_{\mathrm{Obs}}} \tag{5}$$

Due to the use of observed rainfall-runoff ratios to describe the runoff generation process, the volume of the simulated and observed hydrographs are identical. Hence, the performance criterion β of Equation 5 is equal to 1 in all cases. The other components of Equation 5 are not affected by this decision, rather they are influenced by v_{D} and the Horton ratios.

2.4 | Drainage velocity and process indicators

In their original description of the GIUH, Rodríguez-Iturbe and Valdés (1979) stated that the drainage velocity v_{D} needed to be estimated for each event individually. This requirement was one of the main restricting factors for operational applicability of GIUH-models. In this study, we performed parametrization of the GIUH by event with ML. The algorithms were trained to predict v_{D} from several process indicators that will be introduced in the following.

To train the algorithm, a set of known v_{D} values was required. For each event of our data base, v_{D} was calibrated with the BOBYQA (Bound optimization by quadratic approximation) algorithm (Johnson, 2018; Powell, 2009). We used the KGE as the target function with a minimizing target. The drainage velocity was scaled to a minimum of 0.01 m/s and a maximum of 2.0 m/s.

To estimate v_{D} of an upcoming flood event, the following climatic characteristics of the meteorological event have been used as predictors. From the precipitation data of an upcoming event we calculated the sum of precipitation VP [mm], the duration of the precipitation DP [h], the average precipitation intensity of all values greater than zero I_0 [mm/h] and the maximum intensity I_{MAX} [mm/h]. A moving sum was applied to assess the minimum inter-event duration to sum 50% of VP. The duration, normalized by DP provided the indicator D_{50} which we used to describe the temporal distribution of the precipitation event. In addition to these basic parameters, the antecedent precipitation index (API) for 30 days ahead of the event, was calculated. The API was calculated as the weighted sum of hourly intensities, weighted by the inverse of their temporal distance to the event beginning. The API was an indicator for the antecedent condition of the catchment. Although other indicators (temperature, snow cover, etc.) were derived, the performance of the RF did not increase significantly, compared to the use of precipitation indicators only. Consequentially, we only used the five indicators introduced as predictors in the ML-study.

2.5 | Machine learning algorithm

From available ML-algorithms, we chose a random forest regressor (RF) to estimate v_{D} for each event based on the precipitation indicators. The applied RF consisted of 1000 regression trees, each trained with a randomly chosen subset of the given training data. In training phase, a sequence of splitting rules is derived to minimize the regression error in each state. At each split, the training data (predictors and target variables) are divided into two subsets, based on the attribute of a certain predictors. The average of the given target variables in the subsets are the prediction of the branches after the split. For each split, the mean squared error of the prediction is minimized. Due to the randomly chosen subsets, each tree has different criteria for its splits. Please note that the number of successive splits is limited to five splits, in order to keep the trees simple. The limitation results in the RF being immune to overfitting (Breiman, 2001).

As an alternative to the random forest procedure, we applied the adaptive boosting strategy. Here the base estimators, again regression trees, were trained in steps based on the errors of the preceding base estimator. Due to the inferior results of adaptive boosting to the RF with randomly chosen subsets, this strategy was discarded. For detailed description and the theoretical background of the algorithms, see (Breiman, 2001) details on implementation that were provided by (Pedregosa et al., 2011).

We selected the RF due to its common application in hydrological studies (Addor et al., 2018; Brunner et al., 2018) and their ability to reduce process uncertainty from overfitting (Breiman, 2001). The use of multiple base estimators trained to different data sets (i.e. subsets of the complete training data) created a large ensemble of process perceptions. This decreased the tendency of single-estimator algorithms (like a neuronal network or a single decision tree) to overfit a problem.

The ML-algorithms were evaluated with the mean absolute error (MAE) as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|v_{\mathrm{D};i} - \hat{v}_{\mathrm{D};i}\right| \tag{6}$$

Since differences between all N true values $v_{\mathrm{D};i}$ (calibrated) and the estimates $\hat{v}_{\mathrm{D};i}$ were considered as absolute values, negative and positive errors did not eliminate each other. Contrasting other measures that unify different aspects of model performance (e.g. mean squared error [MSE] combining BIAS and error variance), the MAE allowed a direct interpretation (Ramsay & Silverman, 2005).

2.6 | k-means algorithm

The k-means algorithm is a common tool used for supervised classification of M objects into k clusters. Each object m is a vector comprising several characteristics. In this study, catchments were clustered based on selected characteristics taken from Table 1.

The k-means minimizes the intra-cluster variance, while maximizing the inter-cluster variance, giving a k disjoint clusters. The target function Z of the algorithm is (Pedregosa et al., 2011)

$$Z = \sum_{i=1}^{k}\sum_{x_j \in S_i}\left\|x_j - \mu_i\right\|^2 \rightarrow \min \tag{7}$$

The sum of the Euclidean distances between each object x to its assigned cluster center μ within all clusters S are to be minimized. The classification of the objects is iterated in order to minimize Equation 7. At the beginning of each iteration step, the cluster centers μ are calculated as the average of all assigned objects (note that in the first iteration step, all objects are randomly classified). Then the distance between each object and all available clusters is calculated. In the last step of the iteration, the objects are assigned to the nearest cluster center. This procedure is repeated until the classification is stable or a maximum number of iterations (in this case 100 steps) is reached.

2.7 | Silhouette coefficient

The Silhouette coefficient was applied to search for dominant controls on catchment grouping. For a known catchment classification, in this case derived from the analysis of the algorithm training, the catchment characteristic was assessed maximizing the following equation of the silhouette coefficient s (Pedregosa et al., 2011; Rousseeuw, 1987):

$$s = \frac{1}{N}\sum_{i=1}^{N}\frac{b(i) - a(i)}{\max(b(i); a(i))} \tag{8}$$

with a being the mean distance between a characteristic x of the sub-basin i and all other sub-basins belonging to the same cluster A. The second variable b was the mean distance between the characteristics x of sub-basin i and the sub-basins of the next nearest cluster (calculated over all clusters C, different from A):

$$a(i) = \frac{1}{|A| - 1}\sum_{\substack{x(j) \in A \\ j \neq i}}\sqrt{(x(i) - x(j))^2} \tag{9}$$

$$b(i) = \min_{C \neq A}\frac{1}{|C| - 1}\sum_{x(k) \in C}\sqrt{(x(i) - x(k))^2} \tag{10}$$

The silhouette coefficient is defined within the boundaries −1 and 1, with 1 being the ideal outcome and indicating a dense and exclusive structure of the classification relative to the chosen catchment characteristic x. Values around 0 indicate overlapping clusters.

3 | DEVELOPMENT OF A MACHINE LEARNING BASED CATCHMENT CLASSIFICATION SCHEME

Trained dependencies and structures of predictor interaction within the machine learning algorithms are incomprehensible (Han & Kamber, 2010), with the exception of the CART algorithm. Hence, it is not reasonable to evaluate the algorithm parameters, rather its functionality offers insight into underlying processes.

The focus of this study was put on catchment grouping, that is, hydrologic similarity. Following the argumentation of Blöschl and Sivapalan (1995), we assumed that an ML-algorithm trained with data from the most similar donor catchment (within the training data) will show a superior model performance, in comparison to ML-algorithms trained with less-similar donor catchments. In this case, model performance is a proxy for catchment similarity. A ranking of catchments lowering model performance can be interpreted as a ranking of similarity.

As first step, we analyzed the progress of predictive capability of an RF with increasing amount of training data in the Regen basin. Based on this assessment, we determined an empirical ranking of donor catchments for each target catchment that minimized the model error. Groups of catchments that served each other as donors were merged into groups. In a second step, we analyzed the connection between the empirical classification and catchment characteristics and developed the classification scheme. The findings were then tested in the basin of the Upper Main.

3.1 | Analysis of the ML learning procedure

To determine the rankings for each catchment, we performed the following analysis (Figure 2): One catchment was selected as target and was removed from the training data set. Then each remaining catchment was used as donor to individually train an RF. Each model was evaluated with data withheld from the training data set. The best performing model determined the most similar catchment.

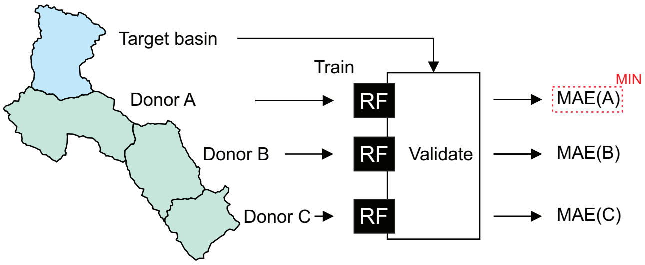

FIGURE 2 Schematic depiction of the learning steps. Data from each donor catchment were used to train an individual RF. All RFs were validated with data from the target catchment. The minimal model error MAE defined the best donor catchment

Next, each of the remaining catchments was used to train a new RF model in addition to the already determined donors. Again, the best performing model indicated the next catchment in the empirical catchment ranking. This procedure was repeated until a full ranking of all available catchments was determined.

In order to evaluate and benchmark the results two alternative rankings of donor catchments, based on common distance measures, were determined. One was based on the distance of the sub-basin centroids, assuming that catchment similarity could be defined by spatial proximity. In order to take catchment nesting into account, we chose a similarity measure based on the Top-Kriging method (Skøien, Merz, & Blöschl, 2006) as second alternate ranking.

To visualize the learning procedure of the RF, the regressor was trained step by step with an increasing amount of data. The data used were either defined by the empirical order of donor catchments (EMP), the centroid-order (CENT) or the Top-Kriging order (TOPK). After each step, that is, after additional data were added set to the training data, the RF was evaluated with the data withheld from the target basin. Note that the data for evaluation were identical in all steps of the analysis, that is, the validation was always performed on the same data set. The development of the model error, expressed as the MAE, is shown exemplary for two catchments in the Regen basin in Figure 3, dependent on the number of used data sets for training.

The development of the model error shown in Figure 3 can be considered representative for all catchments used. The uncertainty belts were computed by repeated fitting of the RF to randomly selected subsets of the chosen data. Results for catchment *Saegemuehle* (short: *Saegem*), shown on the left of Figure 3, were confirmed to be representative for the majority of catchments used in this study. The minimal MAE was achieved with the empirical ranking with only 1–2 data sets for training. In catchment Saegemuehle, the minimal MAE is ≤0.075 m/s, in Eschelkam ≤0.04 m/s, both catchment have an average v_D of ~1.4 m/s, giving a relative error lower than 5%. With increasing training data, the model error increased. Because the

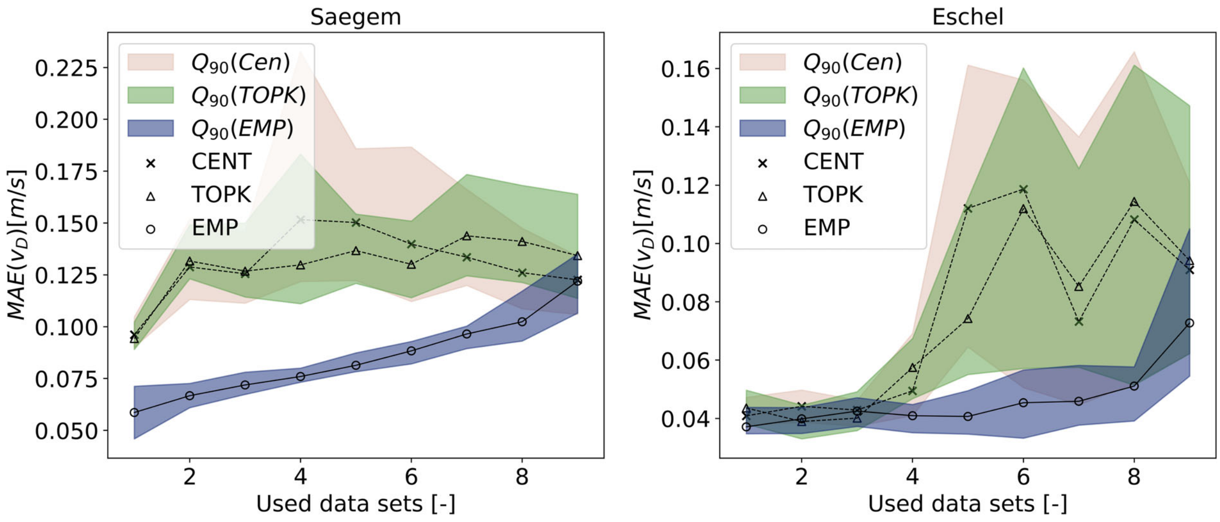

FIGURE 3 Progress of RF model error with increasing amount of data sets. Ranking of the data determined empirically (EMP), by centroid-distance (CENT) and top-kriging weights (TOPK). Results shown for sub-basin Saegemuehle (left) and Eschelkam (right)

validation data were unchanged in each step, the only possible explanation for the decrease is that redundant or misleading information was added to the training data. In this context, connections between precipitation indices and v_D caused other processes than those active in the target basin are misleading information. If such information are present in the training data, the performance of the ML-algorithm decreases.

Results for CENT or TOPK were very similar and led to higher model errors. The increase of error as a function of increasing data availability was not visible at first sight in most cases. Results from sub-basin *Eschelkam* (short: *Eschel*) displayed in the right diagram of Figure 3 showed a different result. In this particular case, all determined rankings led to the same result compared to the first three steps. From there the ranking of donor catchment as well as the model error diverged. This indicated that for *Eschelkam* donor catchments in close spatial proximity delivered the best training data and could be thus considered as similar. However, this result was only obtained for three nested sub-basins, located at a tributary stream of the Regen, which indicates that spatial proximity is proxy of another factor defining the similarity of these three catchments.

3.2 | Catchment groups in the Regen basin

Based on the results obtained from the analysis of the learning processes, we derived a catchment classification. Due to the noted increase of the MAE as a function of increasing data quantities, we considered two donor catchments as the optimum quantity for RF training. This quantity was considered a compromise between sub-basins like *Saegemuehle* (Figure 3, left side) that showed an increasing model error with each added data set and other sub-basins like *Eschelkam* (Figure 3, right side) that showed a stable, in some cases decreasing, MAE for a data base of up to 4 basins.

The first explorative analysis of the results was to localize the favored donor catchments. We created a map for each target catchment and its two donors. These maps showed that three affinity groups were present in the obtained empirical rankings. An RF to estimate v_D in a randomly chosen target catchment within one of these groups was likely to yield minimum MAE if the training data were taken from the remaining catchments within this particular group. Following the assumption that the empirical ranking of donor catchments was correspondent to their similarity, these groups could be interpreted as similarity groups (Figure 4, left panel). Note that the catchments are shown as sub-basin, yet the RF-regressors were fitted to represent entire catchments.

Please note that catchment group 3 differed from the two other groups. An RF for basins in cluster 3 were trained best with data taken from cluster 1. Additionally, these sub-basins were not preferred as training data for any other target sub-basin. The reason for this obvious anomaly will be discussed below.

The grouping showed that two headwater catchment groups and a downstream catchment groups were present, although they were not fully distinct. Group 1 contained mainly headwater catchments in the eastern parts of the Regen basin. Group 2 mainly consisted of catchments from the river Chamb (a tributary coming from the north), for example, sub-basins Eschelkam (short: *Eschel*) and Furth im Wald (short: *Furth*). Sub-basins that were not used as donors were mainly located at downstream positions, merging both tributaries. It has to be noted that the small headwater catchment *Lohmannmuehle* (short: *Lohman*) was assigned to group 3 and *Saegemuehle* to group 2, although their location suggested otherwise. This led to the conclusion that the location at a certain tributary was not a sufficient criterion for the reproducibility of the empirical catchment grouping obtained.

In order to determine the dominant controls on catchment similarity, we calculated the silhouette coefficient following Equation 8, for different characteristics (Table 1). Based on findings in the literature (Section 1), we selected the following characteristics: catchment properties such as area (ARE), mean elevation (ELE) and slope (SLO),

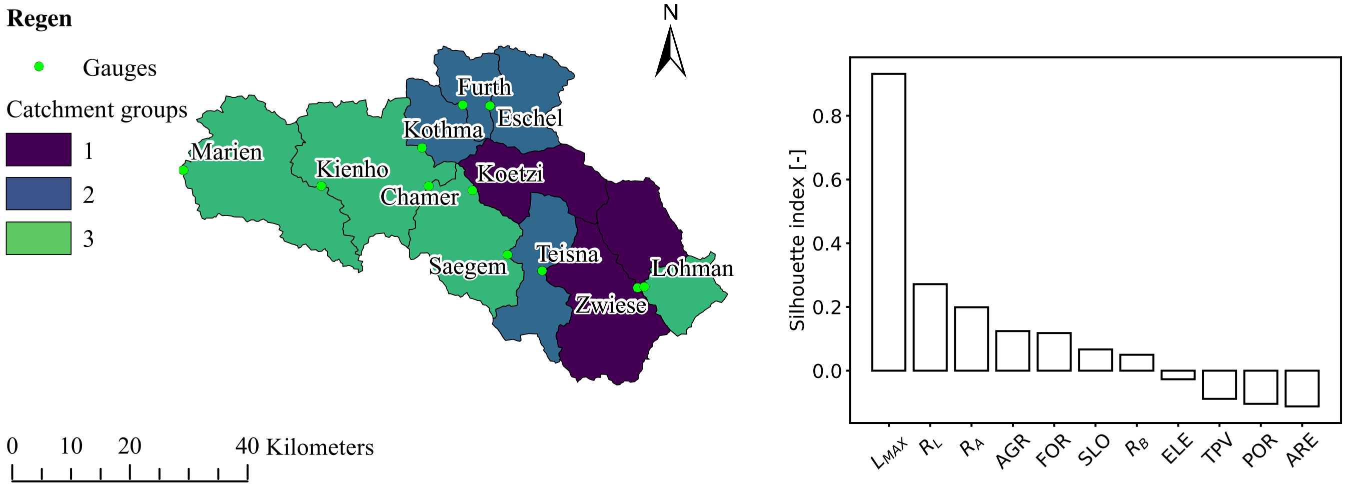

FIGURE 4 (Left) Spatial distribution of catchment groups in the Regen catchment (left) manually defined following the ML analysis. (Right) Silhouette coefficient of the groups for different catchment characteristics

as well as land cover shares (agricultural area AGR, forested area FOR). Additionally, we took into account the characteristics of the drainage system, that is, the Horton rations introduced in Section 2.2. Silhouette scores for each catchment characteristics were summarized in the right panel of Figure 4. The positive values of the Silhouette coefficients indicate that the obtained catchment grouping inherited a strong distinction of L_{Max} and weak structures for R_L and R_A. The physical reality behind these statistical values is that the members of the groups differ significantly in term of the maximum flow paths lengths and drainage network density. Values close to and below zero indicate an overlapping classification, that is, members of different groups have similar elevation, pore volume, etc. To clarify these results, the values of L_{MAX} were plotted as a function of R_L in Figure 5 including a marking of its catchment grouping. Note that the shown cluster boundary lines represent the perpendicular bisectors of cluster center distances. The distinction of the catchment classification derived is obvious for L_{MAX} (Figure 5). The distinction for R_L is significantly weaker.

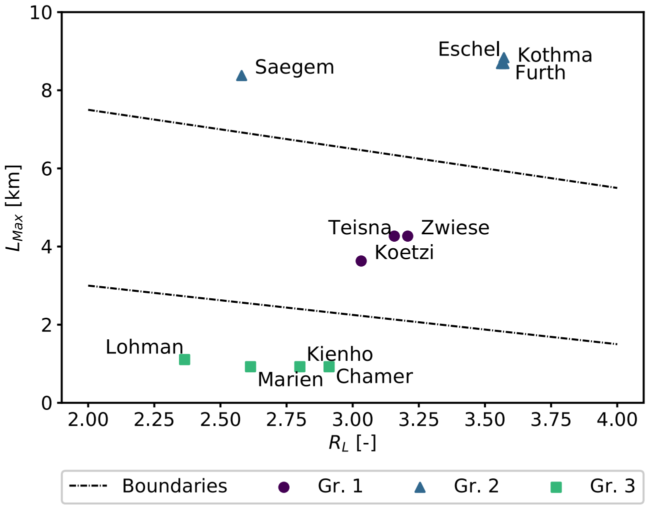

FIGURE 5 Drainage system characteristics L_{Max} and R_L for sub-basins in the Regen catchment. Groups in the Regen basin obtained empirically by ML analysis

Recall that the reproduction of the empirically defined classification uses only catchment characteristics (Figure 5). It is also visible that the catchments of group 3 possess a significant lower L_{Max} than all other catchments, with all values <2 km.

3.3 | Validation case study

To prove the validity of our findings, we applied the L_{Max}–R_L classification scheme in a second case study. We transferred our findings to the basin of the Upper Main and identified four catchment groups. The members of each group have again been defined using the k-means algorithm (Pedregosa et al., 2011). The results of catchment grouping were summarized in the right panel of Figure 6. Spatial arrangement of the catchments and groups in the Upper Main are shown additionally in the left panel of Figure 6. Note that we added an additional group to maximize the silhouette coefficient and to reduce the number of catchments per group. Without the additional cluster, catchment groups 2 and 3 would have been merged.

In the next step, we trained an RF for each sub-basin and used the data from the remaining sub-basins within the particular cluster. Cluster 4 was handled differently, due to its resemblance to group 3 of the Regen catchment. Analogous to the empirical ranking data from these catchments, with L_{Max} lower than 2–4 km, were not used for training. The RF for these basins used data from the nearest cluster for training.

In order to evaluate the model performance, we calculated the MAE for the prediction of v_D for the data withheld from the target catchment (Figure 7). For comparison, we trained three additional RFs with donor catchments defined by Top-kriging (TOPK), centroid-distance (CENT) and the empirical ranking (EMP). The number of data sets used for training was equal to the number of data sets as used

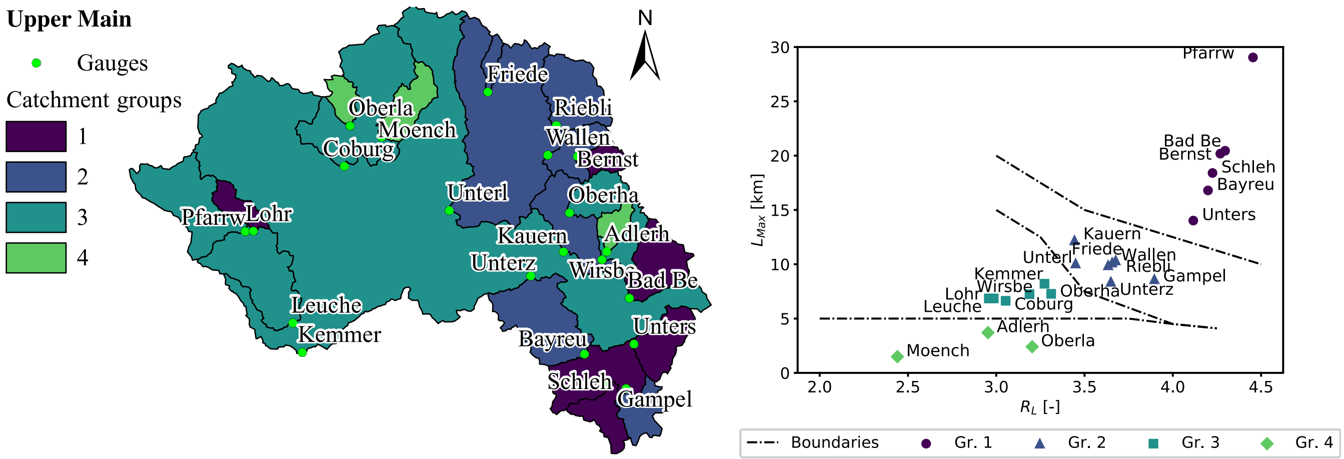

FIGURE 6 (Left) Spatial distribution of catchment groups in the Upper Main basin based on similarity of R_L and L_{Max}. (Right) Drainage system characteristics L_{Max} and R_L with grouping obtained by k-means algorithm

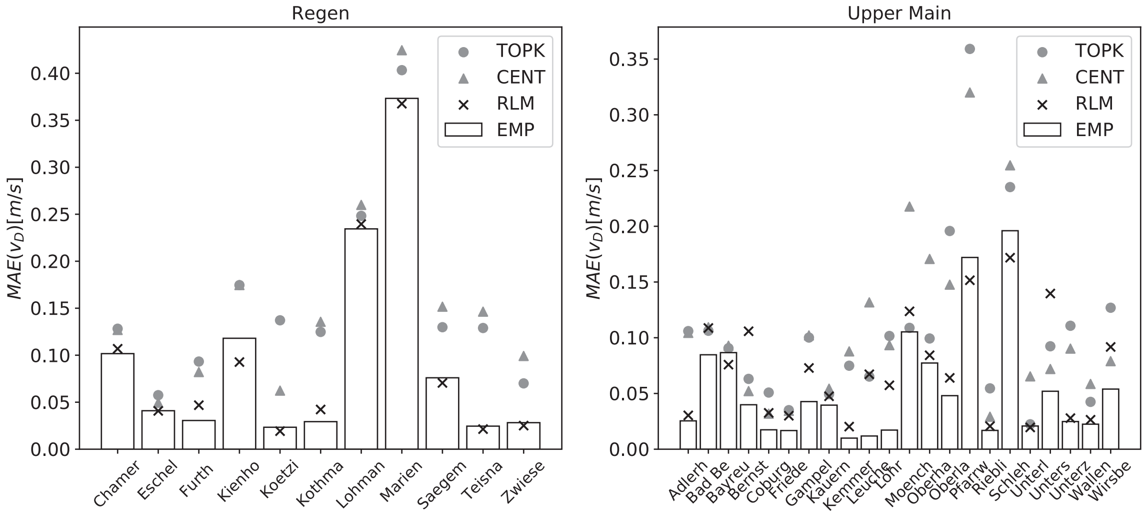

FIGURE 7 MAE of RF prediction of v_D in the Regen (left) and the Upper Main (right) basins. Algorithms trained with an equal amount of training data but different donor catchment selection. Selection determined empirically (EMP), Top-kriging (TOPK), centroid distance (CENT) and based on L_{MAX} and R_L (RLM) classification

with the R_L–L_{MAX} (RLM) similarity measure. Model error obtained with the empirical ranking has been interpreted as a benchmark and was plotted as bars. The competing similarity measures, resulting in different model errors, are shown as scatters for each sub-basin.

In the Regen basin, we were able to reproduce the empirical catchment classification with no exception. As a result, the MAE of the RF trained with data defined by empirical order and by RLM were closely related. Additionally the RLM catchment classification outperformed models trained with data determined by CENT and TOPK (Figure 7). The results for the catchment of the Upper Main looked different, though, in most cases (16 out of 22) the RF trained with data defined by RLM resulted in MAE values comparable to the EMP benchmark. However, in two catchments, *Bernstein* (short: *Bernst*) and *Untersteinach* (short: *Unters*), RLM trained models performed worse than all other models. In four other catchments, the RLM classifications were outperformed either by TOPK or CENT.

The decrease of performance in the validation case study showed that we did not fully encode the learning process of the machine learning algorithm. Nevertheless, the classification schemes produced satisfactory results in both case studies. It is important to note that we were able to reproduce a catchment grouping that was determined by a modelling procedure with catchment characteristics. Hence, we conclude that the characteristics of the drainage system R_L and L_{Max} are closely linked to the drainage velocity, a proxy for short-term runoff dynamics.

4 | RESULTS OF ML SUPPORTED GIUH MODELLING

In the previous sections, drainage system characteristics were identified as dominant runoff controls for runoff dynamics. The classification scheme we derived was used then to train an RF for each catchment included in this study. We also evaluated its capability to estimate drainage velocity parameters. Yet, the impact of the error on runoff simulation results remained unknown. Therefore, we subsequently applied the GIUH-model (Section 2.2) to all catchments. Runoff volumes were determined analytically from observed discharge, due to the focus of this study on runoff dynamics. The number of cascades n was constant for all events within a single catchment (cf. Equation 2) while the storage coefficient k changed as a function of the estimated drainage velocity v_D for each event (Equation 3).

The average performance of the GIUH per catchment with RF-estimated drainage velocities were summarized for the catchments of the Regen (Figure 8) and for the Upper Main (Figure 9). The ratio of simulated and observed standard deviation α (Equation 5), the coefficient of correlation (Equation 5) and the KGE (Equation 4) were given. The ratio of simulated and observed runoff volumes was not displayed due to the use of observed flood volumes for runoff generation.

In the Regen basin, two dependencies of model errors became visible. The comparison of the performances of the RF (MAE in Figure 7) and the GIUH (KGE-components in Figure 8) clearly shows an inheritance between the models. GIUH-simulations for sub-basins with the highest MAE performed worst in term of hydrograph reproduction. Additionally, a dependence on drainage area became visible (with exception to *Lohmannmuehle*). The inferior simulations in sub-basins *Chamerau*, *Kienhof*, *Lohmannmuehle* and *Marienthal* were mainly caused by RF estimation errors.

Results in the Upper Main basin (Figure 9) differed significantly from the Regen results. First of all, the dependence on drainage area vanished in this basin. Moreover, a clear connection to the MAE was not visible. However, the connection to the parametrization error of the RF became discernable, taking into account the average drainage velocity of each sub-basin, that is, when the relative MAE was

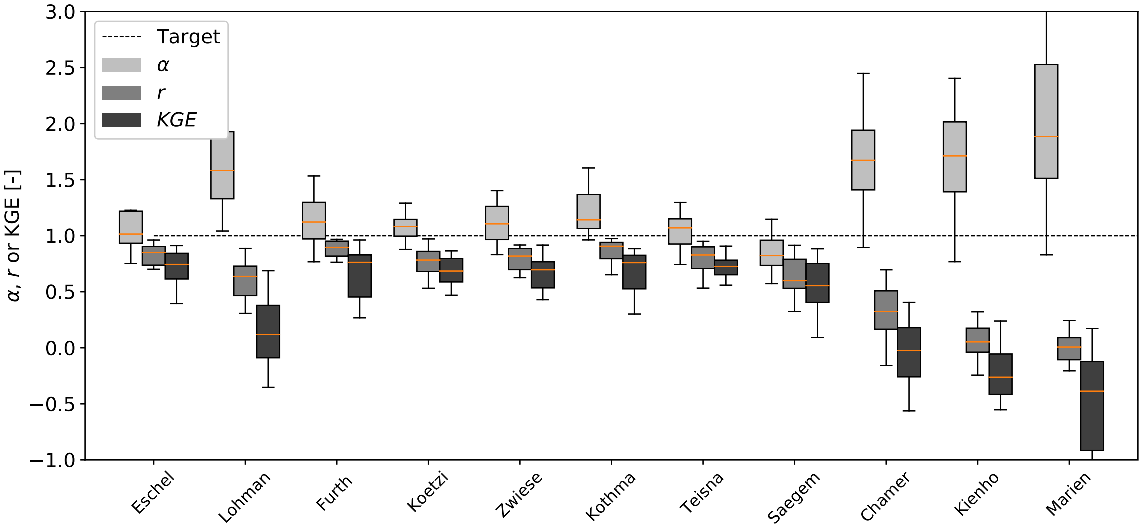

FIGURE 8 Average performance criteria α, r and KGE of the GIUH model application in the catchments of the Regen basin. Catchments sorted by drainage area from smallest to largest

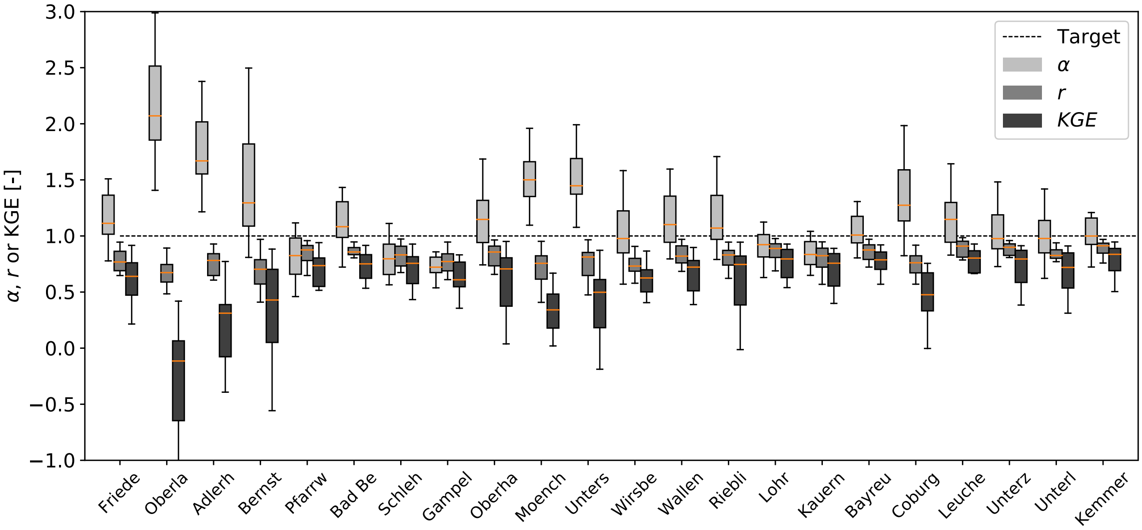

FIGURE 9 Average performance criteria α, r and KGE of the GIUH model application in the catchments of the Upper Main basin. Catchments sorted by drainage area from smallest to largest

evaluated. Sub-basins with the highest variance of KGE values (*Oberlauter*, *Untersteinach* and *Pfarrweisach*) exhibited the highest relative error. While the average relative error ranged from 10 to 30% in the other basins, the error exceeded 100% in these catchments. These results were a result of a mismatch between target and training v_D values. The average v_D in these catchments were significantly lower than in their assigned training data which indicated that these catchments were incorrectly classified.

The catchments with the lowest performance were located at the boundaries of the catchment groups (Figure 5). *Oberlauter* (short: *Oberl*) is a member of Cluster 3, a cluster that has been treated differently than the other. All members of this cluster displayed noticeably low KGE values. *Untersteinach* (short: *Unters*) and *Pfarrweisach* (short: *Pfarrw*) were both members of Cluster 1, although, their catchment characteristics, especially L_{Max} differed significantly. While *Untersteinach* defined the lower boundary of L_{Max} in cluster 1, *Pfarrweisach* defined the upper boundary. It might be reasonable to ask if *Pfarrweisach* was a single sample of another cluster that is otherwise not present in the used data set.

Overall, for the majority of catchments, the parametrization by event brought adequate simulation results. The classification scheme based on Horton's length ration R_L and the maximum flow path length L_{Max} allowed the selection of suitable donor catchments for each catchment. As an additional benefit, the classification allowed a preselection of required data, reducing the effort for data preprocessing (manual event separation, etc.)

5 | DISCUSSION

5.1 | Limitations of the ML predictor

In previous sections (Sections 3.1 and 3.3), the performance of the RF has been used as a foundation for process-related conclusions. From the analysis of the learning curves, we concluded that with increasing data from new catchments, redundant or misleading information was added to training data set hence the performance decreased. In our validation study, a decrease of the predictive performance has been connected to an insufficient classification scheme.

However, another explanation for these outcomes could be an insufficient capability of the model used, respectively the RF, to reproduce the full range of variation within the runoff dynamics. Hence, further analyses of the RF performance were required to eliminate this uncertainty.

As a consequence, the RF has been applied in a local application at each gauge. Each RF has been trained with a randomly chosen subset of 50% the available data. The RF has been evaluated using the data withheld from the same gauge. With this procedure we analyzed if the RF was capable to reproduce the variance of the observed v_D values. In order to reduce effect of the randomly chosen subset, this procedure has been repeated 10 times.

If the results are compared to the range of observed values (Figure 10), it can be concluded that the RF was generally able to reproduce the variance of occurring v_D values. A slight underestimation of the full range of v_D in sub-basins of Group 1 and 2 (left and middle panels) is visible (Figure 10). This means that these sub-basins offer a slightly wider range of variance than the RF could reproduce. It has to be noted that the Box-Whisker plots of the observed and estimated values shows different data sets. While the observed box contains all available data points, the observed box contains RF-estimated of 10 iterations, each with 25% of the available data. Hence, the observed Box-Whisker contains more data points than the estimated. But this only explains a small part of the lower variance. Our findings indicate that a single algorithm is not sufficient to reproduce the full range of hydrologic process heterogeneity, which is in concordance with findings of Elshorbagy et al. (2010b).

Our results showed that the RF was capable to reproduce a large amount of the natural heterogeneity. Additionally, the results showed that with training data from the correct catchment, the RF is able to reproduce the average of v_D values, that is, would result in low MAE-values. From this observation we can exclude limits of the ML-predictor as the source of MAE-increase in Section 3.1. New data sets, describing the connection of precipitation and v_D, were added to the training data in each step. The validation data set, that is, the variance of v_D, was kept unchanged. An increase of the MAE, that is, the prediction error is, hence, solely connected to wrong training data. In this case, misleading connections exist between precipitation and drainage velocity, that is, different active processes defining the hydrograph from the catchment.

The RF has been tested in the Main basin as well. Results show a comparable reproduction of variance by RF as in the Regen basin (Figure 11). We conclude that we can exclude limitations of RF as an explanation for the decreased performance of the model in the validation case study. However, there is another explanation that supports our earlier conclusions (Section 3.3) (Figure 11). The variance of v_D values within the groups, which were used for the definition of

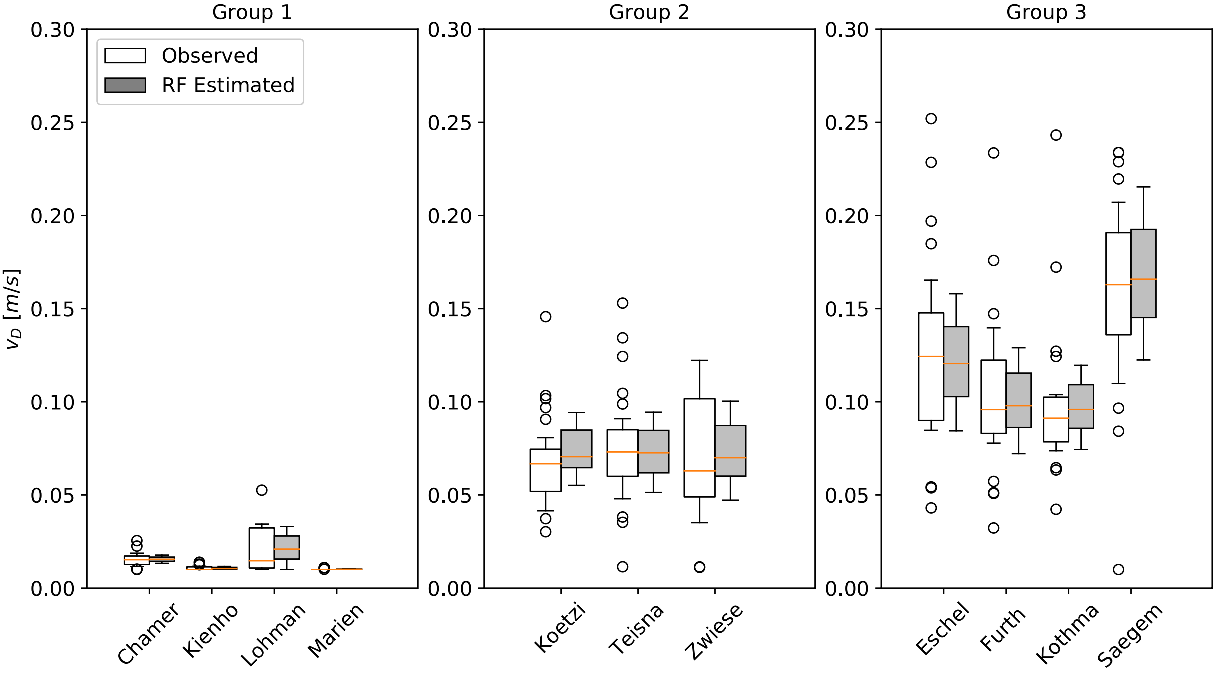

FIGURE 10 Observed (blank boxes) and RF estimated (grey boxes) v_D values in the catchments of the Regen basin. RF trained with 50% of available data of each sub-basin, validation (shown in grey boxes) for withheld data. Catchments grouped by ML analysis

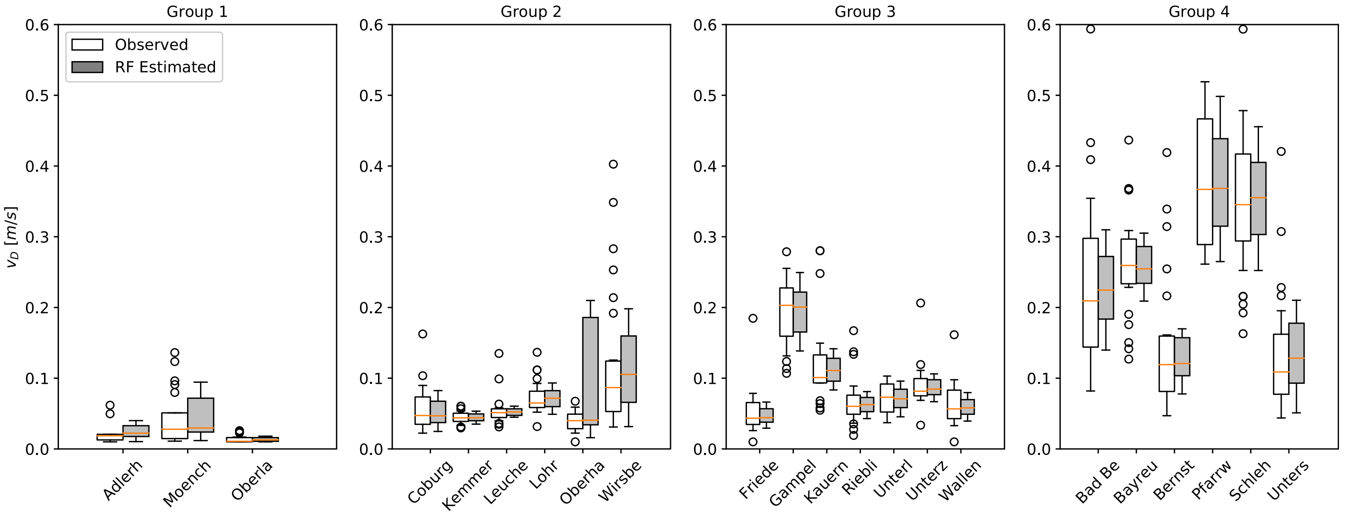

FIGURE 11 Observed (blank boxes) and RF estimated (grey boxes) v_D values in the catchments of the Upper Main basin. RF trained with 50% of available data of each sub-basin, validation (shown in grey boxes) for withheld data. Catchments grouped by RLM classification scheme

training data sets, is significantly larger in the Upper Main than in the Regen basin. It is not the range of v_D values solely that defines the training result of RF, but rather the connection of precipitation indicators and v_D. The inter-group variance is another indicator for an insufficient classification scheme (as stated in Section 3.3).

5.2 | Process implications

We showed that an ML-algorithm could be trained to estimate runoff dynamics of a flood, based on characteristics of the upcoming precipitation event. Moreover, we were able to show that data from selected neighboring catchments could be used to train such a model. Catchment selection was based on characteristics of the drainage system, L_{Max} and R_B, respectively. Robinson, Sivapalan, and Snell (1995) showed that catchments response is either governed by hillslope response or by network geomorphology. The latter is connected to process of network dispersion. They defined a transition zone between these types of response governance based on catchment area. White, Kumar, Saco, Rhoads, and Yen (2004) found that the network dispersion, that is, governance of the network geomorphology, increases with higher Strahler-order rivers in the basin. The findings in our study agree with these indications. In the Regen basin, we were able to derive three groups that show different catchment responses to similar precipitation inputs. Each group represents a different degree of network dispersion. If the same range of hillslope process heterogeneity is assumed, that is, the same range of hillslope drainage velocity, the inter-group variance of v_D values for the entire catchments indicates a different degrees of network dispersion. Group 1 showed the highest influence of network dispersion, while Group 3 showed the least influence. Group 2 was located in transition.

The analysis of the Silhouette coefficients showed that the groups differed significantly in terms of L_{Max} and R_L. Both characteristics describe the channel network. Group 1 was associated with the lowest L_{Max} and R_L values. The characteristics indicate that the highest order streams are particularly short in these catchments and the average length of all order streams are more equally distributed than in the other groups. Vice versa, catchments in Group 3 have a long highest order stream and the average length of the streams increases with the Strahler order. These findings show that confluences close to the outlet of the basins, as well as an equal distribution of flow length within the basin, increase network dispersion. A low L_{Max} and R_L indicate that the runoff from a catchment is governed by dispersion and hillslope processes are of minor importance. With an increasing L_{MAX} and R_L, the relevance of hillslope processes increases.

Note that our results are limited to rainfall-induced flood events. Flood events with other influences such as snow melt were excluded from this study. Results might hence not be valid for these types of events. Especially frozen precipitation has a significant influence on hillslope processes and increases the residence time of the water on the hillslope, giving lower v_D values. To account for snow melt and the related processes, the procedure needs to be repeated with an extended hydrological model (snow routine) and additional data (such as temperature, snow depths, etc.).

6 | CONCLUSIONS

It was the target of this study to identify dominant controls for runoff dynamics on the basin scale. Therefore, we performed a leave-one-out machine learning study in two basins in southeast Germany. Our findings demonstrated that runoff dynamics for an upcoming flood event could be estimated with a random forest solely based on precipitation data. Additionally, the learning procedure of the used ML-algorithm demonstrated that catchments with similar drainage systems characteristics could be considered as hydrologically similar (in terms of runoff dynamics). The similarity of catchments response was caused by an increased influence of channel network dispersion. We found that the distribution of flow length, as well as the presence of larger confluences close to the outlet defined the influence of channel network dispersion.

Our findings were consistent with findings in the literature. The transition from catchment response governed by hillslope processes to dispersion governed response was beforehand explained by drainage area. Our study showed that the transition can be described better with characteristics of the drainage system, L_{Max} and R_{L}, respectively. Our findings are supported by studies on hydrologic similarity in meso-scale catchments which identified topography or drainage characteristics as the relevant indicators for hydrologic similarity. Although we showed in a validation case study in the Upper Main basin that our classification scheme, based on L_{Max} and R_{L}, was transferrable, our results are restricted to the natural conditions of the basins used in this study. The Upper Main basin as well as the Regen basin are located in a mid-range mountainous area and share the same climatic conditions. Therefore, our findings on dominant controls on runoff dynamics are, at this point, restricted to these specific conditions. In future research, the proposed analysis of the ML-learning procedure will be applied to a wider set of basins in different natural and climatic regions. With this step, we will analyze the dependency of local similarity on these conditions and identify the respective dominant controls on runoff dynamics.

Beside our, locally and to rainfall induced floods constrained results, we proposed a new way of process research. In this study we gained knowledge from the analysis of a random forest and its training procedure. More specifically we followed the question: how did the algorithm learn and what data sources did it prefer? With a step-by-step analysis of the training data and the performance of the ML-based regression models, we drew conclusions about catchment groups. Subsequently these groups were related to clusters of catchment characteristics and we were able to build a catchment classification scheme. The derived scheme proved to be valid in the validation case study. Hence, we proved that our process assumptions, gained through ML-model analysis, were valid.

We also demonstrated that two other benefits of ML as a supplement for physically based hydrological models. On one hand, we obtained an operational benefit because the RF performed a calibration of the GIUH-model by events in ungauged catchments with sufficient performance, this being a problem that had been unresolved to this point. On the other hand, its learning procedure allowed to draw conclusions on runoff dynamics and catchment similarity. This dual benefit, operational applicability of the hydrological model and process analysis without a-priori process assumptions, showed the power of ML-application in hydrologic analysis. We therefore propose the use of machine learning and related analysis schemes, as applied in this work, as new way of interpreting data and process research.

In ongoing and future research, we will apply the presented technique to a larger set of basins to test our results in different topographic and climatic regions. We will analyze the dependency of dominant controls on runoff dynamics on catchment conditions and locations. Another focus will be laid on the input data. In this study we excluded snow-influenced floods, due to different active processes. A classification of the remaining rainfall-induced flood events will be introduced to consider flood types, that is, different active processes (comparable to Oppel, 2019) which will reduce the estimation errors of the RF. Additionally, we will include runoff generation as target variable. In this study, we focused on runoff dynamics, a parameter of limited complexity compared to runoff generation parameters. Furthermore future studies will be based on the analysis of a larger set of ML algorithm. While we relied solely on RF in this study, we will take other structures, like deep learning artificial neuronal networks, into account in future. A larger ensemble will raise the probability of finding the suited model for the represented catchment processes.

ACKNOWLEDGEMENT

The authors would like to thank the Bavarian Ministry of the Environment for providing the data used in this study.

DATA AVAILABILITY STATEMENT

Discharge and precipitation data used in this study can be retrieved from the Bavarian ministry of the Environment:

- Upper Main: https://www.gkd.bayern.de/de/fluesse/wasserstand/elbe/tabellen and https://www.gkd.bayern.de/de/meteo/niederschlag/elbe
- Regen: https://www.gkd.bayern.de/de/fluesse/wasserstand/naab_regen/tabellen?method=tabellen and https://www.gkd.bayern.de/de/meteo/niederschlag/naab_regen

ORCID

Henning Oppel https://orcid.org/0000-0002-2761-8674

REFERENCES

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. *Water Resources Research*, *54*(11), 8792–8812.

Bárdossy, A. (2007). Calibration of hydrological model parameters for ungauged catchments. *Hydrology and Earth System Sciences*, *11*(2), 703–710.

Bárdossy, A., Huang, Y., & Wagener, T. (2016). Simultaneous calibration of hydrological models in geographical space. *Hydrology and Earth System Sciences*, *20*(7), 2913–2928.

Bavarian Ministry of the Environment (2018). *Bavarian Institute of Hydrology*. Augsburg. Retrieved October 30, 2018, from https://www.lfu.bayern.de/umweltdaten/nutzungsbedingungen/index.htm.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel, L. A. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, *52*(5), 3599–3622.

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *9*(3–4), 251–290.

Bossard, M., Feranec, J., & Otahel, J. (2000). *CORINE land cover technical guide: Addendum 2000*. Copenhagen: EEA. Retrieved from http://www.eea.eu.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brunner, M. I., Furrer, R., Sikorska, A. E., Viviroli, D., Seibert, J., & Favre, A.-C. (2018). Synthetic design hydrographs for ungauged catchments: a comparison of regionalization methods. *Stochastic Environmental Research and Risk Assessment*, *32*(7), pp. 1993–2023, from https://doi.org/10.1007/s00477-018-1523-3.

Carrillo, G., Troch, P. A., Sivapalan, M., Wagener, T., Harman, C., & Sawicz, K. (2011). Catchment classification: Hydrological analysis of catchment behavior through process-based modeling along a climate gradient. *Hydrology and Earth System Sciences*, *15*(11), 3411–3430.

Deutscher Wetterdienst DWD (German Weather Service) (2019). Climate Data Center (CDC). Offenbach am Main. Retrieved March 27, 2019, from https://opendata.dwd.de/climate_environment/CDC/.

Drouge, G., Leviander, T., Pfister, L., Idrissi, A. E. L., Iffly, J.-F., Hoffmann, L., ..., & Humbert, J. (2002). The applicability of a parsimonious model for local and regional prediction of runoff. *Hydrological Sciences Journal*, *47*(6), 905–920.

Dunn, S. M., & Lilly, A. (2001). Investigating the relationship between a soils classification and the spatial parameters of a conceptual catchment-scale hydrological model. *Journal of Hydrology*, *252*(1–4), 157–173.

Elshorbagy, A., Corzo, G., Srinivasulu, S., & Solomatine, D. P. (2010a). Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and methodology. *Hydrology and Earth System Sciences*, *14*(10), 1931–1941.

Elshorbagy, A., Corzo, G., Srinivasulu, S., & Solomatine, D. P. (2010b). Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 2: Application. *Hydrology and Earth System Sciences*, *14*(10), 1943–1961.

Federal Institute for Geosciences and Natural Resources (2006). Soil map of Germany: (BÜK1000). Hannover, from https://www.bgr.bund.de.

Grimaldi, S., Petroselli, A., & Nardi, F. (2011). A parsimonious geomorphological unit hydrograph for rainfall–runoff modelling in small ungauged basins. *Hydrological Sciences Journal*, *57*(1), 73–83.

Grimaldi, S., Petroselli, A., Alonso, G., & Nardi, F. (2010). Flow time estimation with spatially variable hillslope velocity in ungauged basins. *Advances in Water Resources*, *33*(10), 1216–1223.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91.

Han, J., & Kamber, M. (2010). Data mining: Concepts and techniques (2. ed., [Nachdr.]). *The Morgan Kaufmann series in data management systems*. Amsterdam: Elsevier/Morgan Kaufmann.

Heřmanovský, M., Havlíček, V., Hanel, M., & Pech, P. (2017). Regionalization of runoff models derived by genetic programming. *Journal of Hydrology*, *547*, 544–556.

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

Jarvis, A., Reuter, H., Nelson, A., & Guevara, E. (2008). Hole-filled seamless SRTM data V4. Retrieved October 31, 2018, from http://srtm.csi.cgiar.org.

Johnson, S. (2018). The NLopt nonlinear-optimization package. Retrieved December 13, 2018, from http://ab-initio.mit.edu/nlopt.

Kelleher, J. D., MacNamee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. Cambridge, Massachusetts, London, England: The MIT Press.

Klaus, J., & McDonnell, J. J. (2013). Hydrograph separation using stable isotopes: Review and evaluation. *Journal of Hydrology*, *505*, 47–64.

Kuentz, A., Arheimer, B., Hundecha, Y., & Wagener, T. (2017). Understanding hydrologic variability across Europe through catchment classification. *Hydrology and Earth System Sciences*, *21*(6), 2863–2879.

Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.-J., & Abrahart, R. J. (2016). Data-driven modelling approaches for socio-hydrology: Opportunities and challenges within the Panta Rhei Science Plan. *Hydrological Sciences Journal*, *8*(5), 1–17.

Moussa, R. (2009). Definition of new equivalent indices of Horton-Strahler ratios for the derivation of the Geomorphological Instantaneous Unit Hydrograph. *Water Resources Research*, *45*(9), 15.

Murawski, A., Zimmer, J., & Merz, B. (2016). High spatial and temporal organization of changes in precipitation over Germany for 1951-2006. *International Journal of Climatology*, *36*(6), 2582–2597.

Oppel, H. (2019). *Entwicklung eines selbstkalibrierenden Niederschlag-Abfluss-Modells auf Basis der geomorphologischen Einheitsganglinie und Methoden des Machine Learning. Schriftenreihe Hydrologie/Wasserwirtschaft* (Vol. 30). Bochum: Lehrstuhl für Hydrologie, Wasserwirtschaft und Umwelttechnik.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ..., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives.

Ragettli, S., Zhou, J., Wang, H., Liu, C., & Guo, L. (2017). Modeling flash floods in ungauged mountain catchments of China: A decision tree learning approach for parameter regionalization. *Journal of Hydrology*, *555*, 330–346.

Ramsay, J. O., & Silverman, B. W. (2005). Functional data analysis. In *Springer Series in Statistics* (2nd ed.). New York, NY: Springer Science +Business Media Inc..

Rigon, R., Bancheri, M., Formetta, G., & de Lavenne, A. (2016). The geomorphological unit hydrograph from a historical-critical perspective. *Earth Surface Processes and Landforms*, *41*(1), 27–37.

Robinson, J. S., Sivapalan, M., & Snell, J. D. (1995). On the relative roles of hillslope processes, channel routing, and network geomorphology in the hydrologic response of natural catchments. *Water Resources Research*, *31*(12), 3089–3101.

Rodríguez-Iturbe, I., & Valdés, J. B. (1979). The geomorphologic structure of hydrologic response. *Water Resources Research*, *15*(6), 1409–1420.

Rosso, R. (1984). Nash model relation to Horton order ratios. *Water Resources Research*, *20*(7), 914–920.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., & Carrillo, G. (2011). Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences*, *15*(9), 2895–2911.

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., ..., & Tsai, W.-P. (2018). HESS opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11), 5639–5656.

Singh, R., Archfield, S. A., & Wagener, T. (2014). Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments – A comparative hydrology approach. *Journal of Hydrology*, *517*, 985–996.

Singh, P. K., Mishra, S. K., & Jain, M. K. (2014). A review of the synthetic unit hydrograph: From the empirical UH to advanced geomorphological methods. *Hydrological Sciences Journal*, *59*(2), 239–261.

Skøien, J. O., Merz, R., & Blöschl, G. (2006). Top-kriging - Geostatistics on stream networks. *Hydrology and Earth System Sciences*, *10*(2), 277–287.

Solomatine, D. P., & Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics*, *10*(1), 3–22.

Soulsby, C., Tetzlaff, D., & Hrachowitz, M. (2010). Are transit times useful process-based tools for flow prediction and classification in ungauged basins in montane regions? *Hydrological Processes*, *24*(12), 1685–1696.

Steinschneider, S., Yang, Y.-C. E., & Brown, C. (2015). Combining regression and spatial proximity for catchment model regionalization: A comparative study. *Hydrological Sciences Journal*, *60*(6), 1026–1043.

Uhlemann, S., Thieken, A. H., & Merz, B. (2010). A consistent set of trans-basin floods in Germany between 1952 & 2002. *Hydrology and Earth System Sciences*, *14*(7), 1277–1295.

Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment classification and hydrologic similarity. *Geography Compass*, *1*(4), 901–931.

White, A. B., Kumar, P., Saco, P. M., Rhoads, B. L., & Yen, B. C. (2004). Hydrodynamic and geomorphologic dispersion: Scale effects in the Illinois River Basin. *Journal of Hydrology*, *288*(3–4), 237–257.

Wigington, P. J., Leibowitz, S. G., Comeleo, R. L., & Ebersole, J. L. (2013). Oregon hydrologic landscapes: A classification framework. *Journal of the American Water Resources Association*, *49*(1), 163–182.

Winter, T. C. (2001). The concept of hydrological landscapes. *Journal of the American Water Resources Association*, *37*(2), 335–349.

Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, *30*(8), 1756–1774.

Yaseen, Z. M., El-shafie, A., Jaafar, O., Afan, H. A., & Sayl, K. N. (2015). Artificial intelligence based models for stream-flow forecasting: 2000–2015. *Journal of Hydrology*, *530*, 829–844.

Zhang, Y., Chiew, F. H. S., Li, M., & Post, D. (2018). Predicting runoff signatures using regression and hydrological modeling approaches. *Water Resources Research*, *54*(10), 7859–7878.

How to cite this article: Oppel H, Schumann AH. Machine learning based identification of dominant controls on runoff dynamics. *Hydrological Processes*. 2020;34:2450–2465. https://doi.org/10.1002/hyp.13740