Lennart Schmidt | Falk Heße | Sabine Attinger | Rohini Kumar

Challenges in applying machine learning models for hydrological inference: a case study for flooding events across Germany

Suggested citation referring to the original publication:
Water Resources Research 56 (2019) 5, Art. e2019WR025924
DOI https://doi.org/10.1029/2019wr025924

AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

TECHNICAL REPORTS: METHODS
10.1029/2019WR025924

Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

Challenges in Applying Machine Learning Models for Hydrological Inference: A Case Study for Flooding Events Across Germany

Lennart Schmidt[1,2], **Falk Heße**[2,3], **Sabine Attinger**[2,3], **and Rohini Kumar**[2]

[1]Department of Monitoring and Exploration Technologies, Helmholtz Centre for Environmental Research—UFZ, Leipzig, Germany, [2]Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research—UFZ, Leipzig, Germany, [3]Institute of Earth and Environmental Sciences, University of Potsdam, Potsdam, Germany

Key Points:
- We investigate the use of machine learning methods for flood forecasting
- Predictive accuracy of ML methods is generally very high
- Using ML methods for inference, however, is very elusive and potentially error prone

Supporting Information:
- Supporting Information S1

Correspondence to:
F. Heße and R. Kumar,
falk.hesse@ufz.de;
rohini.kumar@ufz.de

Citation:
Schmidt, L., Heße, F., Attinger, S., & Kumar, R. (2020). Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany. *Water Resources Research*, *56*, e2019WR025924. https://doi.org/10.1029/2019WR025924

Received 4 JUL 2019
Accepted 17 APR 2020
Accepted article online 23 APR 2020

Abstract Machine learning (ML) algorithms are being increasingly used in Earth and Environmental modeling studies owing to the ever-increasing availability of diverse data sets and computational resources as well as advancement in ML algorithms. Despite advances in their predictive accuracy, the usefulness of ML algorithms for inference remains elusive. In this study, we employ two popular ML algorithms, artificial neural networks and random forest, to analyze a large data set of flood events across Germany with the goals to analyze their predictive accuracy and their usability to provide insights to hydrologic system functioning. The results of the ML algorithms are contrasted against a parametric approach based on multiple linear regression. For analysis, we employ a model-agnostic framework named Permuted Feature Importance to derive the influence of models' predictors. This allows us to compare the results of different algorithms for the first time in the context of hydrology. Our main findings are that (1) the ML models achieve higher prediction accuracy than linear regression, (2) the results reflect basic hydrological principles, but (3) further inference is hindered by the heterogeneity of results across algorithms. Thus, we conclude that the problem of equifinality as known from classical hydrological modeling also exists for ML and severely hampers its potential for inference. To account for the observed problems, we propose that when employing ML for inference, this should be made by using multiple algorithms and multiple methods, of which the latter should be embedded in a cross-validation routine.

1. Introduction

The rapid progress made in the field of machine learning (ML) is arguably the most relevant current development for the field of hydrology. Employing ML methods is vital to make use of the increasing amount of data and to cope with the challenges of climate change and an ever-increasing human impact on the environment. ML models like random forest (RF) algorithm and artificial neural network (ANN) are promising candidates for that end. It is no wonder that ML has, therefore, received a lot of attention (Shen, 2018), with the majority of studies applying ML models for prediction or classification purposes. Examples are forecasting of urban water demand (Herrera et al., 2010), estimation of flow duration at ungauged sites (Booker & Snelder, 2012), streamflow classification (Peñas et al., 2014), and simulation (Gudmundsson & Seneviratne, 2015; Shortridge et al., 2016). Regarding the latter, Kratzert et al. (2019) have recently demonstrated the high potential of ML models for rainfall-runoff modeling, even when applied to ungauged basins.

When compared to traditional statistical models like multiple linear regression, ML models are recognized to have superior predictive performance (e.g., Elshorbagy et al., 2010; Lima et al., 2015; Shortridge et al., 2016). However, due to the substantial complexity of a fully trained ML model, their usefulness for understanding relevant relationships contained in the data is less clear. To address this challenge, the field of Interpretable Machine Learning has seen rapid advancement in recent years, and several methods have been developed (Molnar, 2019). These include, for example, methods based on Permuted Feature Importance (PFI) (Breiman, 2001a) or Shapley values (Lundberg & Lee, 2017; Shapley, 1988). Note that the two terms "interpretable" and "explainable" are being used interchangeably in the field of ML and that no unambiguous distinction between the two terms has yet established in literature (compare Gilpin et al., 2018; Rudin, 2019). Therefore, we adopt the more widely used term of "interpretable" ML throughout this work.

The development of interpretable ML may be particularly important in hydrology due to the generally nonunique nature of typical hydrological model descriptions. Nonuniqueness is also referred to as *equifinality*, meaning that in many situations, very different model structures and/or parameter sets are able to describe some observed behaviors with equal accuracy (Beven, 2006). As a result, the task of identifying a single best model representation or estimating a unique parameterization is often not possible. This severely hampers our ability to gain in-depth knowledge about an underlying functioning of the hydrologic system based on such models (Clark et al., 2011). Causes of this problem include the substantial uncertainties/errors in input/output data sets, the presence of nonlinear, interacting processes of internal complexity, and the inflexibility in the structure of classic hydrological models (Nearing et al., 2016). This problem exists despite the fact that hydrological models can be constrained by additional physical information provided through a priori knowledge of hydrologic system functioning—encoded within both model structure and states and flux relationships (Clark et al., 2015a, 2015b). Data-driven ML models, on the other hand, typically do not account for such additional information, which may exacerbate the problem of nonuniqueness. It is, however, this very flexibility that facilitates ML models to exploit the full information content within input/output data sets, and thereby makes possible to eliminate, or at least greatly reduce, the model structural and parameterization errors (Nearing et al., 2016). ML methods are generally considered to be universal function approximators, meaning they can, in principle, approximate any complex relationship observed between inputs and outputs (Cybenko, 1989; Hornik, 1991). It is therefore possible that ML methods suffer less from the aforementioned problem of nonuniqueness.

To investigate this possibility, we present and discuss here the analysis of three data-driven models to a large hydrological data set of flood events in Germany. We use two ML algorithms, namely, RF and ANN. To provide better context to the study, we also use a simple statistical model based on multiple linear regression or linear model (LM) on the same data set. We analyze the magnitudes of around 30,000 flood events across a range of German basins over the period 1950–2010. The example analysis of flood events was chosen due to the high relevance of the topic (Jongman et al., 2014; Paprotny et al., 2018), aiming at contributing to the ongoing discussion about what constitutes the most informative predictors of flood events (Berghuijs et al., 2019). Understanding of the underlying relevant predictors is particularly important due to the expected changes in flood patterns under climate change (Arnell & Gosling, 2016; Blöschl et al., 2017; Milly et al., 2002; Winsemius et al., 2015). While precipitation is generally established to have a dominant impact (Froidevaux et al., 2015; Keller et al., 2018), many recent studies emphasize the role of antecedent soil moisture conditions in shaping the overall magnitude of flood events (Bennett et al., 2018; Ivancic & Shaw, 2015; Nied et al., 2013). To investigate whether data-driven ML models can help to elucidate the varying role of different predictor variables, we use here the following methodology: First, we train the aforementioned data-driven models on the data set of floods across Germany and evaluate their predictive accuracy. Then, we use the trained models to perform hydrologic inference and check the consistency among different models in determining the most important predictors of flood magnitude. The latter is done through an approach based on Permutation Feature Importance (PFI), introduced by Breiman (2001a). The PFI method allows to infer the importance of a predictor in a selected model configuration by computing the drop in prediction accuracy after shuffling its values while keeping the other predictors untouched. Once this has been repeated for all selected predictors, the drop in accuracy can be used to rank the importance of predictor variables.

2. Methods

2.1. Data Set

We used a data set comprising 29,248 flood events that were sampled from time series of daily streamflow of 374 catchments across Germany over the period 1950–2010 (Figure 1). The flood events were identified by applying a peak-over-threshold approach for each catchment, separately: Subsequent days on which streamflow exceeded the 98th percentile value of the respective catchment were grouped as one flood event, and the flood magnitude corresponding to the maximum value (Q_f) was extracted for each event. A set of predictors was then assigned to all events that includes: first, catchment average preconditions of daily precipitation (P; mm/day), daily mean soil moisture (SM; %), and daily mean temperature (T; °C)—that is, the mean of these predictors over multiple time periods $\delta t = [0, 1, 3, 5, 7]$, going back in time from the day of Q_f. Thus, the hydrometeorological conditions from the day of Q_f up to 7 days prior to the flood event were

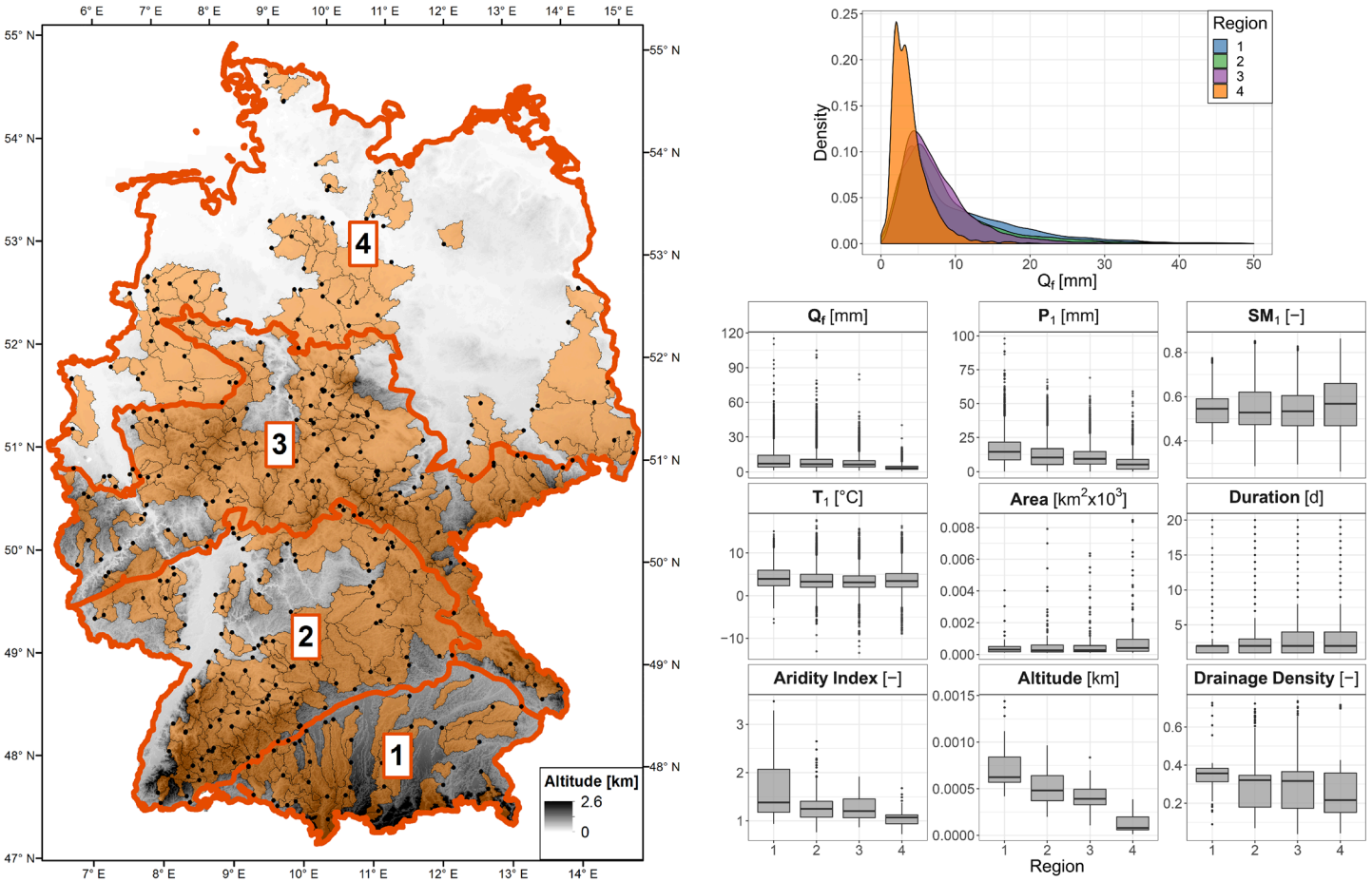

Figure 1. Summary characteristics of case study data sets. Left: map of the study area including the study catchments (transparent orange) and corresponding gauging stations (black dots) as well as the borders of Regions 1 to 4 (solid orange). The shaded background depicts terrain elevation (source: BKG, 2019). Top right: density distribution of flood magnitude (Q_f) summarized over the four main natural regions. Note that the value range of the x axis is limited from [0, 120] to [0, 50] for readability. Bottom right: box plots of Q_f and relevant predictors across the regions. P_1, SM_1, and T_1 represent the average estimates of precipitation, soil moisture, and temperature 1 day prior to and on the day of Q_f, respectively.

covered. In the following, these dynamic predictors are denoted by their name and the respective time interval, for example, P_0, P_1, …, P_7. Here, P refers to the sum of rainfall and snow melt as simulated by the mesoscale hydrologic model (mHM) (Kumar et al., 2013; Samaniego et al., 2010). Similarly, *SM* is based on mHM simulations (Zink et al., 2017)—reflecting the antecedent root-zone soil moisture conditions over the catchment for an approx. 2-m soil depth (see Zink et al., 2017, for more details). *SM* is normalized based on soil porosity, that is, the actual soil water content divided by porosity, to account for the varying soil textural properties. Details on the mHM parameterization and modeling concept can be found in Kumar et al. (2013), Samaniego et al. (2010), and Zink et al. (2017) as well as on the website (www.ufz.de/mhm). Second, to complement the above dynamic predictors, a set of 10 different static predictors was extracted for each catchment. These include average climatic conditions (aridity index and mean annual precipitation), topography (slope and elevation), geomorphology (channel slope, drainage density, and flow path lengths), and land cover. For more details on the predictor variables, we refer to supporting information Table S1.

The selected study catchments cover most of Germany except the southwest and some parts in the north (Figure 1). Catchment size varies between 100 and 8,469 km^2, with the majority of the study catchments (85%) falling below 1,000 km^2. The mean catchment size is close to 750 km^2. All variables with volumetric units (e.g., m^3/s) were area adjusted and converted to mm/day. The flood data set was further grouped into four regions to account for spatial variations in the flood generation mechanisms. Grouping followed the classification of *Naturräumliche Gliederung* (Natural Regions) as produced by the German Federal Institute

of Regional Studies (Meynen et al., 1962). This grouping accounts for spatial variation in geomorphological, geological, hydrological, ecological, and pedological criteria.

In addition to analyzing the difference in spatial variations of flood magnitude, we examined the seasonal variations by splitting the flood events according to their occurrence during hydrological summer (April to September) and winter (October to March) season. The resulting data sets are hereafter referred to as "regional-seasonal data sets," labeled according to region number and season such as 1S, 1W, or 2S.

Figure 1 presents information on the spatial and density distributions of Q_f and relevant predictors. Generally, Q_f is highly left skewed in all regions with decreasing average values when moving from south to north. This gradient corresponds well to a topographical gradient: from alpine conditions in the southernmost region, topography varies from mid-elevation mountains in the middle to the lowlands in the northernmost region. This pattern is also reflected in the corresponding catchment characteristics, that is, average catchment elevation, preevent precipitation P_1, aridity index, and drainage density are on average on the higher end for catchments in the southernmost region as compared to those in the northern regions.

2.2. Model Calibration

On each of the eight regional-seasonal data sets, three data-driven algorithms were calibrated: RF, ANN, and LM, resulting in a total of 24 model combinations. For LM and RF, a feature selection algorithm was applied to identify the model structure that produces the best goodness of fit while reducing model complexity. We split the data set into a training and test subset with a ratio of $\frac{3}{4}$ versus $\frac{1}{4}$, following the procedure as proposed by Hastie et al. (2009). Details on the resulting sample sizes of these split sets can be found in Table S3. The final prediction accuracy on test data was measured by the coefficient of determination (R^2) to ensure the comparability of efficiency measures between different model combinations.

The LM was implemented in the form of a principal component (PC) regression, where PC analysis (PCA) was performed on each of the training data sets prior to model calibration. The resulting PCs that explained 95% of the variance in data sets were kept as predictors for the LM. This approach was used to account for multicollinearity present among the predictor variables (see Figure S4 for more details). The LM model structures included first- and second-order terms including the interaction components, and corresponding parameters were identified through a maximum likelihood estimation approach. A stepwise feature selection approach by Bayesian information criterion (BIC) was used to select the suitable PCs included in the final model (see Table S2 for more details).

The RF model was calibrated using a recursive feature elimination approach (Guyon et al., 2002): Starting from a full model that included all 40 predictors, the five least relevant predictors by PFI were excluded at each iteration. Model performance of the resulting models was assessed by a 10-fold cross-validation scheme (Stone, 1974), using root-mean-square error (RMSE) as the objective function. To limit model complexity, we considered only those models that had an RMSE value within a prescribed range of $[RMSE_{\min}, RMSE_{\min} * 1.01]$, where $RMSE_{\min}$ denotes the lowest RMSE achieved by any of the RF models. The most parsimonious one of these, as measured by number of predictors, was selected for interpretation (see Table S2). We used standard algorithmic parameters as described in the R package "randomForest" (Liaw & Wiener, 2002).

The ANN model was calibrated by virtue of the well-established multilayer perceptron (see Murtagh, 1991), using a stochastic gradient descent backpropagation with mean square error (MSE) as objective function. A hold-out sampling approach was applied during model calibration, that is, one fifth of the training data were used as evaluation data for an early stopping routine to avoid overfitting. Training data were fed into the ANN model a maximum of 800 times, the so-called epochs, with a batch size of 50. When the objective function values on the evaluation data stabilized (i.e., 40 epochs passed without any further decrease in evaluation loss), the step of model calibration was terminated. We considered a range of neural network architectures varying from one to three hidden layers and different number of nodes (see Table S2 for details). For each network architecture, the tuning parameters (i.e., learning rate and momentum) were optimized as a grid-search hyperparameter optimization problem (Claesen & De Moor, 2015). This resulted in 336 candidate models for each of the eight regional-seasonal data sets. Finally, all models that fell within $[RMSE_{\min}, RMSE_{\min} * 1.01]$ of the independent test data sets were collected, and the most parsimonious one—measured by number of algorithmic parameters (e.g., ANN nodes/layers)—was selected for further interpretation.

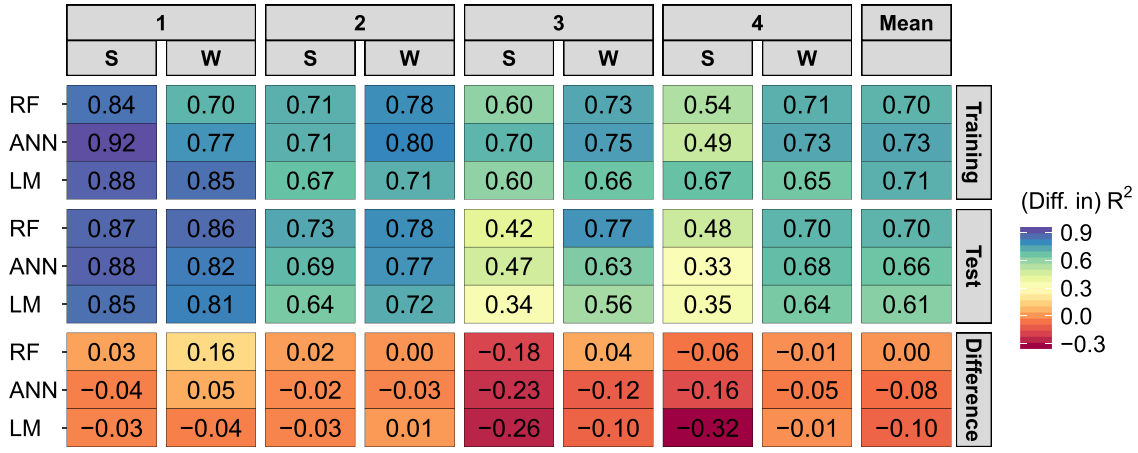

Figure 2. Prediction accuracy in terms of the coefficient of determination (R^2) for all three algorithms on training (top) and test (middle) data sets, estimated across four regions (1 to 4) and two seasons (S = summer; W = winter). Also shown are the corresponding differences in R^2 from training to test data sets (bottom).

2.3. Model Interpretation by PFI

Given the differences in the underlying algorithmic structure, a direct interpretation and comparison of the three algorithms is elusive. The inherent structure is simply too complex for the ML models (RF and ANN), and even for LM, a meaningful retransformation of interactions among PCs to the original input space is not straightforward. Thus, a method based on PFI was employed to explore the influence of individual predictors in each of the algorithms. PFI was first introduced by Breiman (2001a) for interpreting RF models and has recently been extended to the model-agnostic case by Fisher et al. (2019). The underlying principle of PFI is to shuffle the values of one predictor at a time, while leaving the others untouched. In this way, the ties between the target and the respective predictor are broken. The more a predictor contributes to the models' prediction, the stronger the prediction accuracy deteriorates as a result of the shuffling. Thus, the degree of deterioration serves as a measure of the predictors' importance. For a trained model $\hat{f}$ with p predictors, predictor matrix X, target vector Y, vector of predictions $\hat{Y}$, and an error measure $L(Y, \hat{Y})$, estimation of PFI follows these steps (Molnar, 2019):

1. Estimate the original model error $e_{\text{orig}}(\hat{f}) = L(Y, \hat{f}(X))$.
2. For each predictor $j \in 1, \ldots, p$, do
   - generate permuted predictor matrix $X_{\text{perm},j}$ by duplicating X and shuffling the values of feature X_j,
   - estimate error $e_{\text{perm},j} = L(Y, \hat{f}(X_{\text{perm},j}))$, and
   - calculate PFI of predictor j as $PFI_j = e_{\text{perm},j}(\hat{f})/e_{\text{orig}}(\hat{f})$.
3. Sort predictors by descending PFI.

In this study, we used MSE as error measure $L(Y, \hat{Y})$, following the original implementation for PFI in RF. The above algorithm was repeated 50 times for each regional-seasonal model on the test data sets to ensure robustness of results. The PFI values were rescaled as relative feature importance (RFI in %) to ensure the comparability of results across the different regional-seasonal models in the following way: First, the predictors aridity index and mean annual precipitation were excluded from the PFI analysis, as they only served as control variables to account for the spatial gradient of average catchment wetness. Second, the PFIs corresponding to the remaining predictors were rescaled such that the sum of each models' RFI was 100%. For the sake of readability, from hereafter, we refer to RFI as the "importance" of specific predictors.

3. Results and Discussion

Generally, all algorithms delivered reasonable prediction accuracy on the test data sets across all regions and seasons except for two cases (Figure 2). Both ML algorithms outperformed the LM in terms of prediction accuracy, and of the two ML algorithms, RF achieved higher prediction accuracy in most of the analyzed regions/seasons. The superiority of RF stems from the fact that it generalized best on the given data set—in contrast to ANN and LM, it delivered similar mean prediction accuracy on the training and test data sets. For LM, the model accuracy dropped significantly from training to testing—which highlights the issue of

overfitting. While the training for RF and ANN was based on a (cross-)validation procedure (i.e., independent evaluation data being used during the calibration), the LM calibration was based on the training data set, only. Some model combinations exhibit higher accuracy on the testing data sets as compared to the training sample, which is most likely the result of unbalanced sampling of split sets. This may occur when sampling from a highly skewed distribution, as it is the case here. Despite these differences, in general, we find a strong correspondence in predicted flood magnitudes among the three models (see Figure S5).

The analysis of residuals revealed that all algorithms were subject to similar challenges. The predictions of low and medium flood magnitudes were generally more accurate than those of higher magnitude events (see Figure S6). This is likely due to several reasons, such as the distribution of flood magnitudes generally being highly (left) skewed. In our analysis, all models had only limited success in capturing the tails of the distribution. Also worth noting is the issue of measurement uncertainty in discharge observations, with observational errors generally being higher during the medium to large flood events (Baldassarre & Montanari, 2009; McMillan et al., 2012).

The analysis resulted in a distribution of importance over all 39 individual predictors of each of the eight regional-seasonal models for each of the three algorithms (see Figures S7 and S8). To enable an evaluation in an aggregated manner, Figure 3 (top) depicts the mean PFI grouped according to predictor types (dynamic: *P*, *SM*, and *T*; static) for each of the applied algorithms. The derived mean PFI patterns differed substantially depending on the algorithm used. While RF and LM mainly relied on precipitation (*P*) and catchment attributes for predictions, ANN mostly employed soil moisture (*SM*). Moreover, we did not see commonalities with respect to the seasonal signal. Figure 3 (bottom) depicts the importance aggregated to the regional-seasonal scale, that is, averaged by the predictor group, algorithm, and region. Here, we found some similarities and also distinct differences in the importance patterns between the algorithms. For RF and LM, the influence of precipitation (*P*) decreased from south to north, which was generally in agreement with the trend observed for ANN. Also, for soil moisture (*SM*), all three algorithms produced a similar pattern of importance that peaked toward Region 3. With regard to static variables, RF produced a distinct pattern of increasing importance from south to north while there was an inverse and patchy distribution for ANN and LM, respectively. For temperature, there was no distinct pattern in any of the analyzed cases. Similar to before, there was no identifiable pattern in the seasonal signals across the three algorithms. In the distributions over all individual predictors (see Figures S7 and S8), the only commonality was found between RF and ANN. There was a gradient in response time, that is, P_0 and P_1 being assigned higher values in the southern regions than those in the northern ones.

To summarize, the agreement between the three algorithms is as follows: There is a gradient from south to north as to how much influence precipitation (*P*) has on flood magnitude (decreasing) and at which time lag prior to the respective flood event (increasing). These gradients follow the pattern of decreasing model accuracy toward the north. This indicates a gradient in system complexity from the southern regions where flood magnitudes are mostly governed directly by precipitation events, to the northern ones, where the signal of precipitation is attenuated and flood generation is more complex. Regarding soil moisture (*SM*), Region 3 is identified to stand out from the other regions.

When contrasting the above findings with the hydrological a priori knowledge as derived in section 2.1, they are compatible with a basic hydrological interpretation. As presented, there is a topographical gradient from south to north. Consequently, one of the main drivers of runoff generation and concentration in the southern regions is the steepness of hillslopes. This results in a direct translation of precipitation events into flood events, small catchments, and high drainage density that lead to short response times. In the northern regions, larger catchment area, lower slope, and lower drainage density result in a stronger influence of infiltration processes and thus longer response times. The uniqueness of Region 3 regarding the influence of soil moisture, however, cannot be directly connected to common hydrological understanding.

Even though parts of the results can be linked to common hydrological knowledge of the effect that topography has on flood magnitude, this interpretation remains tentative in the light of the heterogeneity of results both in aggregated means and at the scale of individual predictors. Consequently, a more detailed (hydrologic) interpretation is not possible based on our study results. It is striking how much the importance of predictors differs between the algorithms, which lets us conclude that the nonuniqueness, or equifinality, known from classic hydrological modeling concepts, remains a relevant challenge for ML models, too.

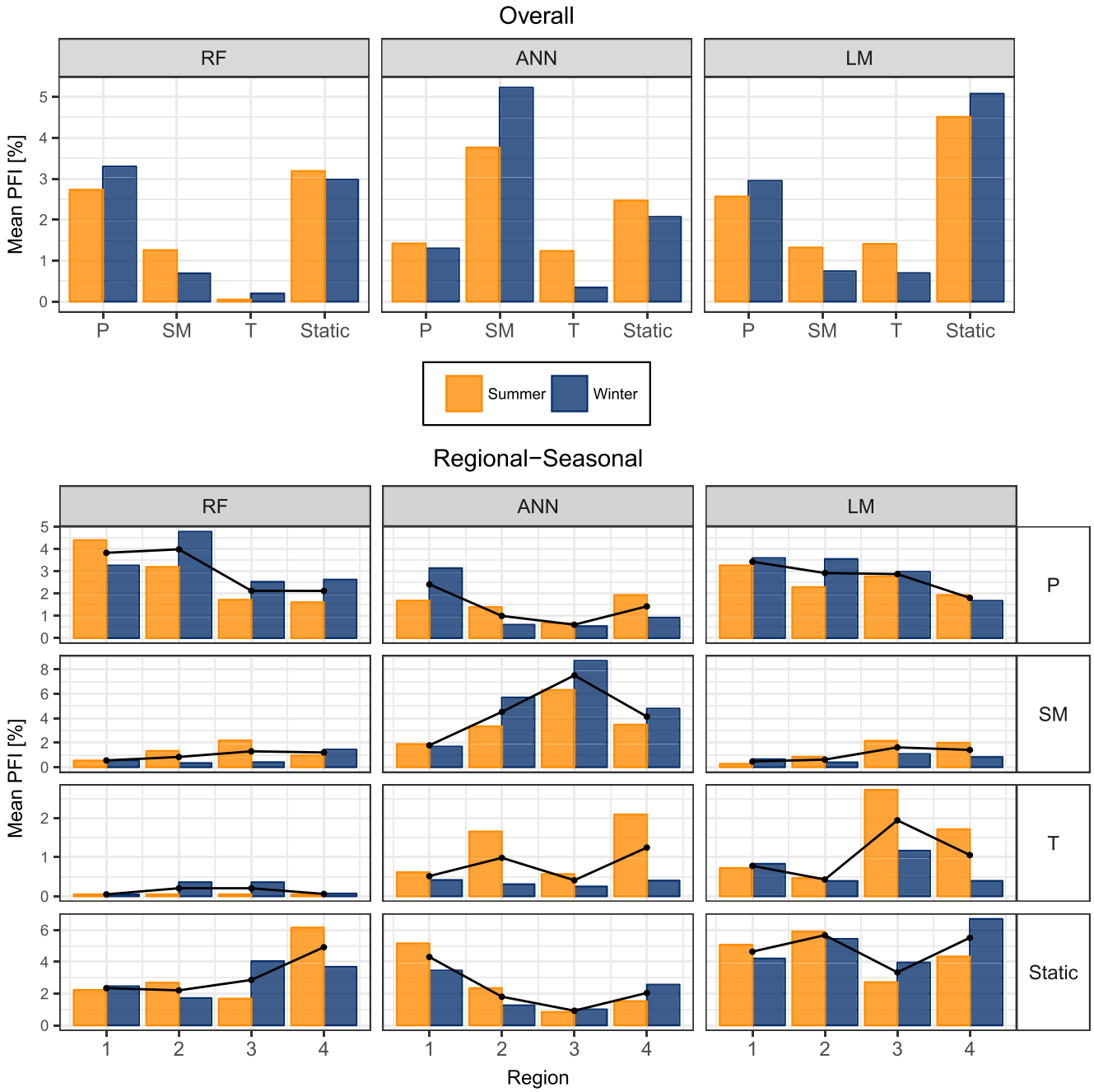

Figure 3. Permutation Feature Importance (PFI) of the predictor groups: precipitation (*P*), soil moisture (*SM*), air temperature (*T*), and static predictors by season and region. Top: mean values of all predictors in a predictor group across all eight regional-seasonal models, plotted by algorithm. Bottom: mean values of all predictors in a predictor group, plotted by algorithm and region. The black line depicts the mean of both summer and winter for each of the predictor groups. Note that the *y* axis is scaled identically by row, only, to allow for interpretation of the patterns.

This is to say that different model structures yield equally acceptable representations of the observed natural processes, but these are based on, sometimes strikingly, different parameterizations. In fact, in ML theory, it is acknowledged that the aim of parameter optimization is and can technically only be to find one of multiple local minima that are equally close to the global minimum. Finding this global minimum at all costs makes the optimization prone to overfitting, that is, a poor performance on unseen data, and is thus not a good strategy (Choromanska et al., 2015). As most ML applications aim at prediction only, this does generally not pose a problem. However, as this study illustrates, similarly high prediction accuracy from different

models does not guarantee a similar underlying inference, that is, equally close local minima may still represent substantial differences in parameterizations. In the ML literature, this phenomenon is referred to as the *Rashomon effect*, a term coined by Breiman (2001b) to capture the often elusive nature of interpretative endeavors. Despite this, further research into the topic has rarely been transferred from theoretical considerations to practical applications. One exception is Semenova and Rudin (2019), who introduced and investigated the Rashomon set of models with similar accuracy but different degrees of complexity. Likewise, Fisher et al. (2019) present model class reliance as a theoretical extension of PFI to the Rashomon set. And while the field of interpretable ML has started to blossom in the recent past, little attention has been directed to this topic.

To account for these observed problems, we propose the following strategies to infer knowledge based on ML models: (1) Multiple algorithms should be calibrated to yield an understanding of a generalized relationship between input and output data sets. (2) PFI, as a statistical tool, maps the patterns that the models detected in the data instead of physical principles and might be subject to inherent bias. Therefore, other model-agnostic methods that have been made available in the recent past like Shapley values or Local Interpretable Model-agnostic Explanations (LIME) (Lundberg & Lee, 2017; Ribeiro et al., 2016) should be applied additionally to rule out any bias that is specific to the respective method. (3) Adding to this, it may be beneficial to derive an estimate of the results' robustness like the local Lipschitz continuity as presented in Alvarez-Melis and Jaakkola (2018) or by computing the importance not after but during cross-validation. In the case of, say, a 10-fold cross-validation, this can yield an ensemble of 10 potential importance distributions, which can be interpreted as a measure of robustness of the obtained results.

Our study has been mostly geared toward providing a demonstration of the challenges in applying data-driven models for inference and thereby raising awareness in the hydrological community not to blindly rely on a single ML algorithm. Future studies should focus at gaining a more profound understanding of the underlying processes that promote nonuniqueness of inferential results among different ML algorithms. Possible extensions of this study could thus be the evaluation of multiple data sets, including simulated ones, to link data set characteristics to the occurrence of the observed problem. In addition, we consider two more fields within the topic of statistical learning to be promising candidates for future inference processes: physics-informed neural networks (see, e.g., Raissi et al., 2020) and causal inference (see, e.g., Runge et al., 2019), both of which are yet to be transferred to the hydrological community.

4. Conclusion

Interpretable ML is a field that has received significant attention in the recent past and several tools for the interpretation of these so-called black boxes have been proposed. Of these, PFI has been the most popular approach to quantify predictor influence. While, previously, PFI was limited to the RF model, this study makes use of a recently published extension of PFI to the model-agnostic case for the first time in the context of hydrology. This allows us to compare the results of three different (data-driven) models: RF, ANN, and multiple linear regression (LM). These models were employed to analyze the influence of dynamic factors such as precipitation, soil moisture, air temperature, and static catchment attributes on flood magnitudes in a range of catchments located across Germany.

The predictive efficiency of all three employed algorithms was in a same range—though the two ML algorithms in general exhibited better predictive power, especially on the test data sets and in regions with relatively lower skill (i.e., northern catchments). Our results reflect the topographical gradient that is present in the study area insofar as the transformation of precipitation into flood streamflow was detected to be more direct and comparatively faster in the south German catchments. In the north—where the topographical differences are less pronounced, prediction of flood events was proven to be more complex, and the streamflow response to precipitation events is attenuated and slower.

While the combination of all three models' results does allow for inference that corresponds to basic hydrological concepts, the extent to which the importance of different predictors differs across models is substantial. Relevant differences in importance can be found across all investigated regions, even though the models' prediction accuracy is fairly similar. This shows that in ML, too, predictive accuracy does not necessarily guarantee valid inference. In addition, this demonstrates that the fundamental challenge of equifinality in hydrological process representation also exists for ML models. This can severely hamper inference

based from these algorithms as the causal relationships between input and outputs remain elusive. In analogy to the classic, mechanistic modeling paradigm where model ensembles are often deployed, based on the results presented here, we suggest the application of multiple ML algorithms for hydrological system understanding. Also, we advise to assess the reliability of the results, for example, by the use of cross-validation and by comparison against other model-agnostic methods for interpretation of ML algorithms.

Acknowledgments
We kindly acknowledge our data providers: the German Meteorological Service (DWD), the Joint Research Center of the European Commission, the European Environmental Agency, the Federal Institute for Geosciences and Natural Resources (BGR), the Federal Agency for Cartography and Geodesesy (BKG), the European Water Archive, and the Global Runoff Data Centre. References to all underlying data sets used in this study are provided in Table S1 and also in Zink et al. (2017). Furthermore, all algorithmic setups and modeling results as well as the flood data sets that were compiled in this study are available from the website (https://doi.org/10.5281/zenodo.3538206). We provide at the website (https://gitlab.com/lennartschmidt/floodmagnitude) the numerical codes used to obtain the results. This work was partially supported by funding from the Helmholtz Association within the research project "Reduced Complexity Models." During the preparation of the manuscript, Falk Heße was financially supported by the Deutsche Forschungsgemeinschaft via Grant HE-7028-1/2. The authors would like to sincerely thank Thorsten Wagener, Hoshin Gupta, and Nans Addor for their helpful and critical comments on the manuscript. This project relies heavily on open-source software. All programming was done in R (R Core Team, 2019) and associated libraries, including dplyr (Wickham et al., 2019), ggplot (Wickham, 2016), randomForest (Liaw & Wiener, 2002), and tensorflow (Abadi et al., 2015). The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Retrieved from https://www.tensorflow.org/ (Software available from tensorflow.org).

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049.

Arnell, N. W., & Gosling, S. N. (2016). The impacts of climate change on river flood risk at the global scale. *Climatic Change*, *134*(3), 387–401. https://doi.org/10.1007/s10584-014-1084-5

BKG (2019). Bundesamt für kartographie und geodäsie: Digitales geländemodell dgm1000. *GeoBasis-DE/BKG*. Retrieved from https://www.geodatenzentrum.de/

Baldassarre, G. D., & Montanari, A. (2009). Uncertainty in river discharge observations: A quantitative analysis. *Hydrology and Earth System Sciences*, *13*(6), 913–921.

Bennett, B., Leonard, M., Deng, Y., & Westra, S. (2018). An empirical investigation into the effect of antecedent precipitation on flood volume. *Journal of Hydrology*, *567*, 435–445. https://doi.org/10.1016/j.jhydrol.2018.10.025

Berghuijs, W. R., Harrigan, S., Molnar, P., Slater, L. J., & Kirchner, J. W. (2019). The relative importance of different flood-generating mechanisms across Europe. *Water Resources Research*, *55*, 4582–4593. https://doi.org/10.1029/2019WR024841

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Blöschl, G., Hall, J., Parajka, J., Perdigão, R. A. P., Merz, B., Arheimer, B., et al. (2017). Changing climate shifts timing of European floods. *Science*, *357*(6351), 588–590. https://doi.org/10.1126/science.aan2506

Booker, D. J., & Snelder, T. H. (2012). Comparing methods for estimating flow duration curves at ungauged sites. *Journal of Hydrology*, *434*(Supplement C), 78–94. https://doi.org/10.1016/j.jhydrol.2012.02.031

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

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., & LeCun, Y. (2015). The loss surfaces of multilayer networks. In *Artificial intelligence and statistics* (pp. 192–204). Brookline, MA: JMLR, Inc. and Microtome.

Claesen, M., & De Moor, B. (2015). Hyperparameter search in machine learning. arXiv preprint arXiv:1502.02127.

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*, W09301. https://doi.org/10.1029/2010WR009827

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015a). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, *51*, 2498–2514. https://doi.org/10.1002/2015WR017198

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015b). A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resources Research*, *51*, 2515–2542. https://doi.org/10.1002/2015WR017200

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Elshorbagy, A., Corzo, G., Srinivasulu, S., & Solomatine, D. (2010). 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.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variables importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Froidevaux, P., Schwanbeck, J., Weingartner, R., Chevalier, C., & Martius, O. (2015). Flood triggering in Switzerland: The role of daily to monthly preceding precipitation. *Hydrology and Earth System Sciences*, *19*(9), 3903–3924. https://doi.org/10.5194/hess-19-3903-2015

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, (pp. 80–89). New York: Institute of Electrical and Electronics Engineers (IEEE).

Gudmundsson, L., & Seneviratne, S. I. (2015). Towards observation-based gridded runoff estimates for Europe. *Hydrology and Earth System Sciences*, *19*(6), 2859–2879. https://doi.org/10.5194/hess-19-2859-2015

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1-3), 389–422.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). New York: Springer. Retrieved from https://www-stat.stanford.edu/tibs/ElemStatLearn/

Herrera, M., Torgo, L., Izquierdo, J., & Pérez-Garcia, R. (2010). Predictive models for forecasting hourly urban water demand. *Journal of Hydrology*, *387*(1), 141–150. https://doi.org/10.1016/j.jhydrol.2010.04.005

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257. https://doi.org/10.1016/0893-6080(91)90009-T

Ivancic, T. J., & Shaw, S. B. (2015). Examining why trends in very heavy precipitation should not be mistaken for trends in very high river discharge. *Climatic Change*, *133*(4), 681–693. https://doi.org/10.1007/s10584-015-1476-1

Jongman, B., Hochrainer-Stigler, S., Feyen, L., Aerts, J. C. J. H., Mechler, R., Botzen, W. J. W., et al. (2014). Increasing stress on disaster-risk finance due to large floods. *Nature Climate Change*, *4*. https://doi.org/10.1038/nclimate2124

Keller, L., Rössler, O., Martius, O., & Weingartner, R. (2018). Delineation of flood generating processes and their hydrological response. *Hydrological Processes*, *32*(2), 228–240.

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*, 11,344–11,354. https://doi.org/10.1029/2019WR026065

Kumar, R., Samaniego, L., & Attinger, S. (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, *49*, 360–379. https://doi.org/10.1029/2012WR012195

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18–22.
Lima, A. R., Cannon, A. J., & Hsieh, W. W. (2015). Nonlinear regression in environmental sciences using extreme learning machines: A comparative evaluation. *Environmental Modelling & Software*, *73*, 175–188.
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, (pp. 4765–4774). San Diego, CA: Neural Information Processing Systems Foundation Inc.
McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078–4111. htpps://doi.org/10.1002/hyp.9384
Meynen, E., Schmitdhüsen, J., Gellert, J., Neef, E., Müller-Miny, H., & Schultze, J. H. (1962). Handbuch der naturräumlichen gliederung deutschlands: unter mitwirkung des zentralausschusses für deutsche landeskunde (No. Bd. 1-9). Bundesanstalt für Landeskunde und Raumforschung.
Milly, P. C. D., Wetherald, R. T., Dunne, K. A., & Delworth, T. L. (2002). Increasing risk of great floods in a changing climate. *Nature*, *415*, 514–517. htpps://doi.org/10.1038/415514a
Molnar, C. (2019). Interpretable machine learning. A guide for making black box models explainable.
Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, *2*(5-6), 183–197.
Nearing, G. S., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V., & Xia, Y. (2016). Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. *Journal of Hydrometeorology*, *17*(3), 745–759 English. htpps://doi.org/10.1175/JHM-D-15-0063.1
Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijs, S. V. (2016). A philosophical basis for hydrologic uncertainty. *Hydrological Sciences Journal*, *61*, 1666–1678. htpps://doi.org/10.1080/02626667.2016.1183009
Nied, M., Hundecha, Y., & Merz, B. (2013). Flood-initiating catchment conditions: A spatio-temporal analysis of large-scale soil moisture patterns in the Elbe River basin. *Hydrology and Earth System Sciences*, *17*(4), 1401–1414. htpps://doi.org/10.5194/hess-17-1401-2013
Paprotny, D., Sebastian, A., Morales-Nápoles, O., & Jonkman, S. N. (2018). Trends in flood losses in Europe over the past 150 years. *Nature Communications*, *9*, 1985. htpps://doi.org/10.1038/s41467-018-04253-1
Peñas, F. J., Barquín, J., Snelder, T. H., Booker, D. J., & Álvarez, C. (2014). The influence of methodological procedures on hydrological classification performance. *Hydrology and Earth System Sciences*, *18*(9), 3393–3409. htpps://doi.org/10.5194/hess-18-3393-2014
R Core Team (2019). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org
Raissi, M., Yazdani, A., & Karniadakis, G. E. (2020). Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, *367*(6481), 1026–1030.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135–1144). New York: Association for Computing Machinery.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.
Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., et al. (2019). Inferring causation from time series in Earth system sciences. *Nature communications*, *10*(1), 1–13.
Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*, W05523. htpps://doi.org/10.1029/2008WR007327
Semenova, L., & Rudin, C. (2019). A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. arXiv preprint arXiv:1908.01755.
Shapley, L. S. (1988). A value for n-person games. In A. E. Roth (Ed.), *The Shapley value: Essays in honor of Lloyd S. Shapley* (pp. 3140). Cambridge, UK: Cambridge University Press. htpps://doi.org/10.1017/CBO9780511528446.003
Shen, C. (2018). A trans-disciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*, 8558–8593. htpps://doi.org/10.1029/2018WR022643
Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, *20*(7), 2611–2628. htpps://doi.org/10.5194/hess-20-2611-2016
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133.
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. Retrieved from https://ggplot2.org
Wickham, H., Franois, R., Henry, L., & Müller, K. (2019). dplyr: A grammar of data manipulation [computer software manual]. Retrieved from https://CRAN.R-project.org/package=dplyr (R package version 0.8.0.1).
Winsemius, H. C., Aerts, J., van Beek, L., Bierkens, M., Bouwman, A., Jongman, B., et al. (2015). Global drivers of future river flood risk. *Nature Climate Change*, *6*, 381–385. htpps://doi.org/10.1038/nclimate2893
Zink, M., Kumar, R., Cuntz, M., & Samaniego, L. (2017). A high-resolution dataset of water fluxes and states for Germany accounting for parametric uncertainty. *Hydrology and Earth System Sciences*, *21*(3), 1769–1790. htpps://doi.org/10.5194/hess-21-1769-2017