https://doi.org/10.5194/hess-2022-213
Preprint. Discussion started: 2 September 2022

Hydrology and Earth System Sciences
Discussions
Open Access

Hydrologic Interpretation of Machine Learning Models for 10-daily streamflow simulation in Climate sensitive Upper Indus Catchments

Haris Mushtaq[1,2], Taimoor Akhtar[3], Muhammad Zia-ur-Rahman Hashmi[1], and Amjad Masood[1]

[1]Global Change Impact Studies Centre (GCISC), Islamabad, Pakistan
[2]Weather and Climate Services Limited, Islamabad, Pakistan
[3]School of Engineering, University of Guelph, ON, Canada

Correspondence: Haris Mushtaq (harismushtaq21@yahoo.com)

Abstract. Machine learning for hydrologic modeling has seen significant development and has been suggested as a valuable augmentation to physical hydrological modeling, especially in data-scarce catchments. In Pakistan, surface water flows predominantly originate from the transboundary Upper Indus sub-catchments of Chenab, Jhelum, Indus, and Kabul rivers. These are high elevation data scarce catchments and generated streamflows are highly seasonal and prone to climate change. Given the catchment characteristics, there is an utmost need to develop machine learning models that are hydrologically robust. Thus, the current study besides evaluating the potential of three machine learning models for streamflow simulation also focused on the hydrologic interpretation of machine learning models using SHapley Additive exPlananations(SHAP). XGBOOST, RandomForest, and Classification and Regression Trees(CART) were evaluated. All of these models performed well and the range of R ^{2} and Nasche-Efficiency for all three models lies between 0.61 to 0.90. Our study's most crucial contribution is SHapley Additive exPlananations(SHAP) method which gives extensive insights into the influence of each variable on simulated streamflow. SHAP analysis highlighted the significance of minimum temperature in high elevation zones for both Indus and Chenab catchments where streamflows are dominated by snow and glacier melt. We strongly believe that the findings of this study will promote the use of SHAP analysis for streamflow forecasting in data scarce and high elevation catchments in Pakistan.

1 Introduction

Streamflow modeling is critical to nearly every element of water resource management, water quality, climate risk assessment, and advanced detection of hazardous hydrological events. However, due to the evolving and dynamic character of runoff in response to stochastic precipitation, physical properties of the catchment, temporal and regional variability in temperature and evapotranspiration, and climate change regime, forecasting remains a difficult job. (Adnan et al., 2019; Meng et al., 2019; jing Niu et al., 2019). Hence, monthly and daily streamflow forecasting has received significant recognition in the past decades (Wang et al. 2019;(Parisouj et al., 2020; Hadi and Tombul, 2018). Thus, in recent years, the development of efficient and well-performing streamflow models has become an intriguing idea in hydrology and other relevant engineering fields. (Pham et al., 2020).

Since the nineteenth century, when the rational system was developed, water managers and hydrologists have utilized observable relationships between rains and runoff for streamflow modeling (Beven, 2011). Nevertheless, in the last decade, the emergence of highly complex machine learning algorithms, coupled with significant gains in processing power, has spurred considerable research into new approaches for data-driven streamflow forecasting (Shortridge et al., 2016). While distributed physical models that precisely depict hydrologic processes can still be considered a holy grail for rainfall-runoff modeling, empirical models can prove to be a beneficial tool for data-scarce physical watershed processes. Adequate streamflow and climate data have been more readily accessible in data-poor places due to the creation of historic data centers and more recent initiatives to blend satellite-generated information with in situ observations to track hydrologic and climatic conditions. Empirical models could be especially helpful in these situations because collecting measurement-based values of soil hydraulic characteristics or details on hydrologically suitable land management operations might be challenging.

The tremendous advancements in machine learning (ML) are undoubtedly the most important contemporary development in the subject of hydrology. Machine Learning is essential to deal with the issues posed by climate change and an ever-increasing human effect on the natural world (Schmidt et al., 2020). Owing to the substantial economic and health-related implications of floods, streamflow forecasting is a pillar of operational hydrology and an essential study area. Forecasting has also become a critical aspect of water resources planning in climate emergency times such as predicting summer low flows (Godsey et al., 2014), anticipating drought conditions (Kapnick et al., 2018), or managing reservoir systems with changing flow regimes(Tennant et al., 2020).

Despite the encouraging results reported in contemporary literature, the majority of ML streamflow forecast applications are restricted to watersheds where rainfall is the primary contributor. A blend of spring snowmelt and rainwater can boost streamflow in many areas, especially, non-alpine regions (Johnstone, 2011; Knowles et al., 2007). The amount of snow accumulation and its proportion to runoff varies greatly among watersheds (Knowles et al., 2006). The utilization of Data-driven methods has made it possible to construct implementable models with high predictive performance, using the available data, while relying on limited domain knowledge. At the same time, data-driven models have become a matter of contention, particularly in scientific domains including hydrology. They are frequently described as "black box" models, where the user has restricted information on the working of models that generate forecasts(Herath et al., 2020). However, failure to realize the potential of ML models in hydrological modeling, on the other hand, is regarded as a threat to hydrological modeling (Nearing et al., 2021). Nearing et al. (2021) argue that machine learning models can find similarities between catchments by producing good results even for watersheds that were not used to train the models. Thus, there is a potential for using a machine learning model to provide basin-scale hypotheses that conventional models cannot. The question of whether machine learning models yield similar outcomes in watersheds with a mix of rainfall and snowmelt contributions is a natural one. Snowpack management is critical, and quick snowmelt plays a significant role in flood disasters, which merits further investigation. In the field of nonlinear hydrology number of ML models, such as support vector regression (SVR) (Liang et al., 2018), adaptive neuro-fuzzy inference system (ANFIS) (Dehghani et al., 2019), artificial neural networks (ANN) (Wen et al., 2019), (Zhang et al., 2018), (Meng et al., 2019) and genetic programming (GP) (Ravansalar et al., 2017); (Yaseen et al., 2017) have been developed and shown acceptable performance in modeling nonlinear hydrological processes. These non-linear processes include drought forecasting (Rahmati et al., 2020), sediment transport modeling (Ebtehaj et al., 2016), and rainfall-runoff modeling. Deep learning (neural networks) have been extensively used for streamflow modeling, however, neural networks generally need a lot of training data to yield accurate predictions, while, tree-based models have been shown to work well when provided with limited training data (Gauch et al., 2021). It is a tree-based method in which any route from the root to the leaf node is characterized by a data separator sequence till a Boolean result at the leaf node is obtained. Furthermore, no assumptions about the predictor's distribution are required. DT is computationally less expensive than other ML models. It also has unique benefits, such as being simple to understand and visualize (Ni et al., 2020).

When it comes to Pakistan, more than 60 % of the total population is rural and mostly resides in the semi-arid to hyper-arid plains of the lower Indus basin (Akhtar et al., 2020). Agricultural activities of these rural dwellers are heavily dependent on streamflows originating from the Upper Indus catchments, that are diverted towards agricultural fields via the massive Indus Basin Irrigation System Siddiqi et al. (2018). These streamflows are highly seasonal, variant, and prone to extreme events and climate change Lutz et al. (2016); Wijngaard et al. (2017). This gives rise to the need for the development of a robust hydrological model that can simulate streamflows. The use of Machine Learning methods for streamflow forecasting in Upper Indus Catchments is very limited. The literature we studied is quite limited, it compared and evaluated the performance of machine learning models. ur Rauf et al. (2018) compared the performance of four SVR models and ANN to forecast streamflows in the upper Indus. The authors observed the long-term predictions were best made by the ANN model. Hussain (2020) compared the performance of multilayer perceptron (MLP), SVR, and random forest (RF) for streamflow forecasting in the Hunza river basin(sub-catchment of Upper Indus Catchment). The findings suggest that the performance of RF was the best followed by MLP and SVR. According to (Hassan and Hassan, 2021)to increase the performance of ANN-based streamflow forecasting, a pre-processing strategy can be used. Box-Cox transformation was used in the study and a comparison between original and transformed data revealed better performance in both training and testing datasets. There is no prior literature on the hydrological interpretation of Machine learning models. This, gives rise to a key research question, can ML Models encapsulate hydrologic behaviors in data-scarce catchments of the Upper Indus?

Another significant challenge in effectively using ML algorithms for hydrologic predictions in watersheds similar to the Upper Indus Catchments is the limited availability of observed streamflow and climate stations. These catchments are large, with high variability in elevation across each catchment, but have limited availability of streamflow and climate stations to adequately represent their diverse topography, land-use dynamics, and associated hydrologic response. Hence, feature engineering and selection, with due consideration of data scarcity challenges is critical for the development of data-driven hydrologic models.

To this end, we tend to evaluate the potential of three machine learning models namely Random Forest, XGBoost, and Classification And Regression Tree (CART) in making a short-term streamflow forecast at the 10-day lead time in four catchments of upper Indus catchments in the Indus Basin. All three selected models fall under the category of decision tree algorithms. Our selection of decision tree algorithms is based on the reason that tree-based models work well when provided with limited data set. To our knowledge, there are limited applications of classification and regression trees within hydrological forecasting. Vezza et al. (2010) applied multiple regressions with morphoclimatic catchment characteristics in North-Western Italy obtained through four classification methods: seasonality indices (SIs), classification and regression trees (CARTs), residual pattern approach (RPA), and weighted cluster analysis (WCA). Erdal and Karakurt (2013) found that the CART model outperforms

SVR and the results from this study indicate that the classification and regression trees (CARTs) are a promising technique for monthly streamflow forecasting. Choubin et al. (2018) focused on the application of the classification and regression trees (CART) model, for the prediction of the precipitation over a highly complex semiarid climate system of Kerman Province Iran, and obtained satisfactory results. Our selection of RF for streamflow forecasting is based on its excellent track record in short-term streamflow forecasting. Rasouli et al. (2012) forecasted streamflow at 1–7 d lead times using three ML models and data from combinations of climate indices and local meteo-hydrologic observations in southern British Columbia, Canada. Moreover, other studies also reported high performance of the random forest model in highly seasonal rivers in the highlands of Ethiopia (Shortridge et al., 2016), in the Wei River Basin, China (Li et al., 2019), and Current River at Doniphan, Missouri (Papacharalampous and Tyralis, 2018), and in the Pacific Northwest (Pham et al., 2021). In practice, RF can be trained to forecast streamflow at various timescales depending on the input variables provided. The selection of XGBoost is based on its success in ML challenges, which stems from the fact that it is a speedy, effective, and scalable approach that generates good results in a variety of domains. Unlike the "random forest" (RF) model, which uses a parallel ensemble, the XGBoost model is built on the concept of "boosting," which involves combining all of the predictions of "weak" learners in order to produce a "strong" learner using additive training procedures (Fan et al., 2019). Compared with other ensemble models, it can improve the model's robustness by introducing regular terms and column sampling; and when each tree selects the split point, a parallelization strategy will be adopted to improve the model's running speed. A few studies have investigated its performance in earth science fields and achieved state-of-art results. Fan et al. (2018a) compared the SVM and the XGBoost methods for daily solar radiation prediction in humid subtropical China. They found that XGBoost outperformed the studied empirical models, and recommended it as a promising model for solar radiation estimation due to better model stability, efficiency, and comparable prediction accuracy. (Zhang et al., 2019) applied XGBoost to predict the Standardized Precipitation Evapotranspiration Index (SPEI), meteorological measures, and climate signals from 32 stations. The results showed that the XGBoost predicted more accurately in 1–6 steps than ANN did. Despite its application in other fields (Le et al., 2019; Xia et al., 2017), the use of XGBoost for streamflow modeling is yet to be explored further (Ni et al., 2020).

1.1 Hydrologic Interpretation

While ML models and data-driven methods have shown immense potential in adequately simulating streamflow (even for data-scarce and ungauged catchments), detailed diagnosis and analysis of interpretation of the trends embedded within such models, is extremely important to create confidence in such models. Black-box use of such models, without intricate interpretation of how they simulate streamflow can be extremely dangerous. However, the recent emergence of the SHAP method seems to have reversed this situation. SHAP originates from "game theory" and constructs an additive interpretation model (Strumbelj and Kononenko, 2014), and all features are regarded as "contributors". The response of the machine learning output to input can be obtained through the calculated SHAP value (Lundberg and Lee, 2017). SHAP (Lundberg and Lee, 2017) is a game-theoretic approach to explain the performance of the machine learning model. It can quantify the contribution of the input feature to the prediction for an individual sample. In the study of feature analysis, SHAP technology has achieved satisfactory results in many fields, such as medicine (Lama et al., 2021), materials (Mangalathu et al., 2020), hydrology (Hu et al., 2021), water

https://doi.org/10.5194/hess-2022-213
Preprint. Discussion started: 2 September 2022

environment (Wang et al., 2021, 2022)l and confirmed its applicability in feature analysis. However, only a few studies have applied interpretable machine learning methods to the attribution analysis of variables in hydrology (Wang et al., 2022).

Applying SHAP analysis in data-scarce Upper Indus catchments and complex hydrologic dynamics helps in interpreting ML model in terms of physical linkages between different hydrometeorological parameters. We intend to develop ML models that are not only efficient in terms of evaluation metrics but also hydrologically robust.

Thus, the overall objective of our study is to determine a robust methodology for the interpretation of these models in the large snowmelt and rainfall-fed watersheds. Therefore, our study has two distinct objectives (1) ML model development for large snow dominant catchments to compare the observed streamflow with Machine learning model generated streamflows and 2)SHAP analysis to interpret ML models and explain physical linkages between hydrometeorological parameters.

2 Study Area

In this study, we focus on watersheds in the upper Indus catchments, namely, Kabul, Jhelum, Chenab, and Indus Catchment. This region covers a total area of 321,540 km2, while Kabul and Indus Catchment covers almost 82 % percent of the total area. Figure 1 delineates the four Indus basin catchments selected for employing machine learning algorithms for streamflow simulation, i.e., Upper Chenab, Upper Jhelum, Upper Indus, and Kabul. As mentioned earlier, these catchments account for more than 70 % (Karimi et al., 2013; Laghari et al., 2012; Yu et al., 2013) of Pakistan's renewable freshwater availability.

2.1 Catchment Characteristics

For the analysis of this study, a Digital Elevation Model (DEM) was produced for all four basins (Terbela, Mangla, Chenab, and Kabul) using the Shuttle Radar Topography Mission (SRTM) with a spatial resolution of 30 m. The watershed boundary was delineated using SRTM DEM, which was acquired from the following web page; "http://www. srtm.csi.cgiar.org/". Different elevation zones for different catchments were then defined Table 1. Hypsometric elevation of each zone was plotted between basin elevations and cumulative areas of the zones as shown in Figure 3. These elevation zones are made so that the gridded data of all the parameters [precipitation, temperature (max and min), snow water equivalent, and potential evapotranspiration] should fall within each zone of the basins for analysis.Figure 2 represents the DEM for each catchment.

3 Data

3.1 Streamflow

We used 10-Daily streamflow data of outlets of the four catchments discussed in the previous section and delineated in Fig. 2 for the Streamflow prediction. Daily streamflow records from 1979-2014 for the relevant flow stations,i.e., Chenab at Marala Barrage, Jhelum at Mangla Reservoir (upstream), Indus at Tarbela Reservoir (upstream), and Kabul at Nowshera gage (see Fig. 2) are used in our analysis.

3.2 Climate Data

3.2.1 Precipitation

As we are using the 10-Daily streamflow dataset, the subsequent dataset has also been transformed into 10-Daily time series. In this study (Yatagai et al., 2012) APHRODITE's (Asian Precipitation - Highly-Resolved Observational Data Integration Towards Evaluation) gridded precipitation was used. The dataset has a resolution of 0.25 degrees in space and one day in time. The temporal range of the data set is from January 1979 to December 2014. Figure 4 shows the correlations between APHRODITE and observed precipitation (mm) at six meteorological stations in the study area and their respective catchments, with R squared values ranging from 0.5-0.75. The stations that were looked at and their catchments are Jhelum (Jhelum), Gilgit (Indus), Kotli (Jhelum), and Muzaffarabad (Jhelum), Balakot (Jhelum).

3.2.2 Snow Water Equivalent(SWE)

Global Land Data Assimilation System (GLDAS) was used to collect the Snow Water Equivalent data set. The dataset spans the years January 1948 to December 2014 and has a resolution of 0.25 degrees. The dataset further includes various current and prospective weather and climate prediction, water resource applications, and water cycle studies. Princeton University's worldwide meteorological forcing data were used to constrain the simulations (Sheffield et al., 2006).

3.2.3 Temperature

A shift in temperature is deemed as the key trigger for snowmelt, daily minimum (Tmin), and maximum temperature (Tmax) from CPC Global Temperature data products. The spatial Resolution of a dataset is 0.50 Degrees while the temporal resolution is daily. The dataset covers from January 1979 to December 2014.

3.2.4 Potential Evaptranspiration (PET)

For Potential Evaptranspiration (PET), PyETo package was used. PyETo is a Python package for calculating Potential Evapotranspiration (PET). There are various features in the package for predicting missing meteorological data.

The package provides three methods for estimating ETo/PET namely; FAO-56 Penman-Monteith (Allen, R. G. et al., 1998),Hargreaves (Hargreaves and Samani, 1985) and Thornthwaite (THORNTHWAITE, 1948).

Hargreaves method (Hargreaves and Samani, 1985) was used in our study to derive the Potential Evapotranspiration. It is a simple evapotranspiration equation that only involves only a few easily obtainable metrics i-e: minimum, maximum and mean temperature, and extraterrestrial radiation.

If the available meteorological statistics for the Penman-Monteith approach are inadequate, the FAO recommends the Hargreaves equation (Allen, R. G. et al., 1998) as an alternative way for determining ETo.

https://doi.org/10.5194/hess-2022-213
Preprint. Discussion started: 2 September 2022

4 Methodology

Methodology of the study, presented in Figure 5, and discussed in forthcoming sections:

4.1 Predictor Selection

The response variable in our study is 10-daily mean streamflow. We used five predictors to try to elaborate on the variability in streamflow. The predictors were chosen after a thorough review of the literature along with our knowledge of the region's hydrology. Precipitation (PRECIP), in particular, is a fundamental source of streamflow. Temperature changes, and change in minimum and maximum temperature impacts the "Snow Water Equivalent" (SWE), which provides storage data on the quantity of stored snow accessible for runoff.(Table 2).

The gridded data sets were compared with the nearest grid point of the global elevation shapefile to sort out the data based on the elevation zones discussed in Section 2.1. The number of features corresponding to the elevation zones of each catchment is presented in Table 3.

Two categorical features were also incorporated into the models to understand the impact of seasonality on streamflow prediction. To this end, Months and Seasons (Early Kharif, Late Kharif, and Rabi) were included as categorical features. Values assigned to three seasons were; zero for Early Kharif (April to June), 1 for Late Kharif (June to September), and 2 for Rabi Season (October to March), while Months range from 1 to 12.

4.2 Models

4.2.1 Classification And Regression Trees (CART)

Breiman et al. (1984) introduced Classification and regression trees (CARTs), which gained popularity in recent years. It is a regression and classification algorithm that works with both numerical and categorical predictors. Every node of the tree has a splitting rule. The splitting rule is calculated by minimizing the "relative error" (RE) (Hancock et al., 2005).CART, on the other hand, has an increased likelihood of overfitting, due to its sensitivity to even minor changes in the training data (Erdal and Karakurt, 2013). While creating a CART model, hyper-parameters must be decided to direct the establishment of correlation. Key parameters used were: maximum depth, minimum samples, and minimum samples leaf.

4.2.2 Random Forest

It is a fairly new concept that was created as an extension of CART to boost prediction accuracy (Liaw and Wiener, 2002). Proposed by Breiman,(2001), RF is a non-parametric, quasi-unsupervised algorithm within the decision tree family that comprises a combination of statistically independent trees to forecast regression and classification tasks. Using the algorithm. various decision trees are combined and their forecasts are aggregated. The algorithm performs well when the number of observations is greater than the number of variables. Moreover, it is considered dynamic due to being easily adaptable to various ad-hoc learning tasks and effectively providing rankings of variable significance. Since even the slightest change in the dataset used for training causes a decision tree to bring a significant difference in output, several decision trees with bootstrap aggregation are utilized to counter this statistical instability. Breiman, (2001) thoroughly elaborated on the Random Forest and its significance. In our research, the randomForest Python has been used for training and testing the model.

Since hyperparameter tuning is based more on experimental data rather than theory, the ideal method to determine the optimal values is to try different combinations and assess each precision. However, assessing every model based solely on the training data can lead to overfitting (which is regarded as one of the most common problems in ML). Although the model will score well on the training data set for which it was optimized, it will fail to generalize its performance on the test dataset. Overfitting happens when a model performs well on the training data but poorly on the test data, resulting in a model that is well-versed in the training data but unable to solve new tasks. Cross-validation is used to account for overfitting in hyperparameter optimization.

K-fold cross-validation (CV) is a common way to choose which model to use. Folds are used to build models, and the hold-out fold is used to validate the models. This indicates that model development takes precedence over model validation (Jung, 2018). In our model, K = 10 was used to fit the model.

Because we usually only have a vague understanding of what the ideal hyperparameters are, examining a wide variety of values for each hyperparameter is the most effective technique to narrow down our search. Using Scikit-RandomizedSearchCV Learn's method, we set up a grid of hyperparameter ranges, took random samples from the grid, ran K-Fold CV with each set of values, and obtained the best parameter values.

4.2.3 XGBoost

Extreme Gradient Boosting (Fan et al., 2018b)is a quicker and more efficient version of gradient boosted decision trees (see for example (Natekin and Knoll, 2013)). It is an approach that predicts the errors of prior models by creating new models (in this instance decision trees). All fitting models are combined in the final model. When adding new decision trees, to make the loss function as small as possible, a gradient descent algorithm is used. To avoid overfitting, XGBoost employs a more regularized model formalization. This approach improves the accuracy of XGBoost over gradient boosting. The XGBoost package in Python was used for model training and validation in our study.

Using the selected predictors, the training dataset described in Section 3.1 has been used to develop the XGBoost model for predicting 10-daily stream flow. Hyper-parameters must be selected while creating an XGBoost model to promote correlation formation. Important parameters used were: maximum depth, n estimators, and gamma.

4.2.4 Model Development

The stream flow data has a temporal resolution of 35 years (1979-2014). Daily streamflow records were then converted into 10-daily averages i-e 36 values for each year and 1296 values for the complete dataset. Subsequent datasets are also converted into 10-daily timesteps. The reason for using 10-daily timesteps is based on the fact that operational management of Indus River Basin is based on 10 daily timesteps. The observed stream flow data is divided into two parts: namely training/model

https://doi.org/10.5194/hess-2022-213
Preprint. Discussion started: 2 September 2022

development and testing/model evaluation. The first 75 percent of the dataset is utilized for training, while the remaining 25 percent is used for testing.

4.3 SHAP Analysis

Louppe et al. (2013) found that the importance of variables is a useful tool for understanding both the underlying process of the current model as well as offering useful information for the selection of predictor variables in future research. The SHapley Additive exPlanations (SHAP) method proposed by Lundberg and Lee (2017) is used in this study to explain the feature importance in the predictions .In machine learning, a feature's SHAP value for a particular prediction describes its contribution to that prediction compared to the average prediction for the dataset (Molnar, 2020). The SHAP values, however, can be extremely computationally expensive to calculate (Lundberg et al., 2019b). The TreeExplainer algorithms developed in Lundberg et al. (2019a) are efficient in calculating the SHAP values for tree-based models. TreeExplainer is designed such that the unique structures of trees are utilized to simultaneously account for all possible subsets, allowing the SHAP values to be computed in polynomial time. More details on TreeExplainer can be found in Lundberg et al. (2019a).TreeExplainer is employed in this study to calculate the SHAP values for the XGBoost, RandomForest, and CART models to assess the contribution of hydro-environmental features to the prediction.

"Connection weight" (Garson, 1991), Partial derivative (Dimopoulos et al., 1995), or input variable disturbance (Qi et al., 2016) are some more approaches for analyzing the feature sensitivity of neural network models. These approaches are, nevertheless, "local sensitivity analysis" methods, which only assess the impact of one input variable on the outcome. In reality, when constructing a sensitivity analysis experiment, it is crucial to evaluate the interactions between different variables (Gevrey et al., 2006). SHAP can be used to quantify the impact of each model input on the model's final output. It is more accurate because it considers the impact of the interactions between the input variables on the result. The findings of SHAP are thus regarded as more accurate and credible owing to its robust theoretical underpinning. The contribution of an input parameter is determined by SHAP and it is centered on the marginal contribution:

$$\varnothing i = \sum_{S \subseteq N(i)} \frac{|S|!(n-|S|-1)!}{n!}[v(S \cup i) - v(S)] \tag{1}$$

where ∅i represents the impact of input factor i to the model output, n represents exactly how many input variables were involved, N denotes the entire set of inputs, S denotes the set of elements preceding I in the sequence, which is a subset of N, and v (S) denotes the value created by the interaction of the elements included in the S subset. By comparing the corresponding SHAP readings, the impact of every input variable on the model output could be readily seen.

4.4 Evaluation Metrics

Various model performance measures exist, and each one offers a specific understanding of the relationship between simulated and observed values of streamflow. Although the "Pearson correlation coefficient" (r) and its square, the "coefficient of determination" (R ^{2}), are frequently employed, Legates and McCabe (1999) discussed the limitation of these two measures. The authors suggested that a measure of goodness-of-fit, such as the Nash-Sutcliffe efficiency (NSE), could provide a more credible assessment of the models. Due to these factors, we decided to use the following three measures for model assessment:R^2 ,NSE, and hydrologic signatures. Insight into the streamflow model can be obtained by evaluating both error statistics and signatures of observed and simulated hydrologic time-series (Addor et al., 2018). While statistical error metrics typically focus on analyzing the agreement between observed and simulated time series, hydrologic signatures (e.g., bias, flow duration curve (FDC) slope etc.) are metrics that define the characteristics of a hydrologic time series. Consequently, it is imperative that signatures of streamflows simulated via ML methods closely resemble observed streamflow signatures. Moreover, incorporating hydrologic signatures in the assessment of ML algorithms is important in ascertaining if simulated streamflows adequately mimic observed streamflow characteristics or not. Typically 5-10 signatures are selected to summarize the flow regime(Euser et al., 2013). In this study, Relative Bias, Discharge peak, Auto-Correlation, Logarithmic Bias, and Variance were selected to obtain in-depth details of discharge at selected watersheds. These parameters cover a variety of characteristics of the model's performance and are easy to understand.

5 Results and Discussion

5.1 Hydrologic Signatures

The comparison of hydrological signatures for observed and simulated streamflows over the four catchments is presented in Figure 6. The 10-daily discharge data are used to calculate these signatures. The relative bias is used as an indicator to describe the error between actual and simulated discharge from the model. The values of relative bias help us to assess the change of total error of predicted discharge. The variance percentage indicates how well the model represents baseflow statistics like mean baseflow volume, median and mean daily flow, and index. In the same way, the model shows high-spell and low-spell statistics with a reasonable amount of variation in each sub-basin. The selected models are effective at reflecting local conditions, according to the assessment of performance evaluation parameters and hydrological signatures. Hence, the selected models can be suggested for streamflow projections for river basin-scale studies depending on the accuracy of climate datasets.

5.2 Variable importance using SHapley Additive exPlanations (SHAP)

We analyzed the variable importances and their interactions using SHAP analysis across each of the four catchments. Figures 10 11 and 12 represents SHAP summary plots for top 10 variables. It is comprised of colored dots, each observation in the dataset generates a dot for each variable for the test period 2006-2014. Note that the order of the variables is the same as the Feature Importance order, it is formed by summing all the weights of the color chart. The second dimension is the shape of the distribution, when you have a large mass of dots, it piles the dots in the format of a lying violin. The third dimension is the position of the points, note that there is a vertical axis at zero, dots on this axis indicate that the variable is not positively or negatively impacting, if the dot is to the right of this line, it positively impacts the target, and left, negatively, the farther from the axis, the greater the impact. The fourth and last dimension is the color of the balls, blue indicates low values and red indicates the high value this is done for each of the variables.

In all three models Tmin in the elevation zone 7 (5400-7103m) of Chenab and Tmin in the elevation zone 5 (3801-4600m), elevation zone 6 (4601-5400m), and elevation zone 7 (5400-8578m) of Indus has the most impact on streamflow prediction. As the values of Tmin increase higher will the predicted streamflow. The variable importance for Indus and Chenab captured in our SHAP analysis is logical considering the fact that, snowmelt is the most important contributor to flow in the upper Indus catchments, followed by glacier melt, with 80 % of the flow occurring between June and September. Two processes, namely, accumulation of snow, which is governed by winter precipitation and temperature, and meltwater generation, which is determined by summer temperatures, control interannual flow variability (Charles et al., 2018).In Karakoram, the highest point where meltwater can come from is between 5400 m and 6200 m. In the Western Greater Himalayas, meltwater comes from much higher elevations, between 4250 or 4750 m and 5500 or 5750 m. (Mukhopadhyay and Khan, 2015). So, the flow caused by snowmelt is a function of winter precipitation and temperature as well as summer temperature. On the other hand, glacier melt is a function of summer temperature, moreover snow cover also affects glacier melt (Charles, 2016). Thus this makes minimum Temperature a key driver for predicted streamflow. In Chenab Catchment Season and Precipitation also plays a significant role owing to the fact the streamflow in Chenab catchment is influenced by the monsoon, which becomes evident from the results the early Kharif and late Kharif season and precipitation play a significant role. This is further complemented by Shap dependence plots in Figure 13. For the Indus catchment, we can interpret this to say, in general, the higher value of Tmin in the highest elevation zone increases the streamflow Similarly, as Tmax increases Tmin also increases which in turn increases the predicted streamflow. The same is the case for the Chenab catchment.

When it comes to Kabul Catchment, it is a semi-arid catchment, which means higher evaporation rates than annual total precipitation. In line with Kabul catchment hydrology, our analysis presented in Figures 10 and 11 indicate Potential Evapotranspiration as the most significant variable in the prediction of streamflow. Although determining the causes of runoff generation has proven to be a difficult task, in semi-arid catchments highly variable rainfall and streamflow, high evaporation rates, and deep groundwater reservoirs may result in a highly complex hydrological process dynamics (Camacho Suarez et al., 2015). Our analysis does depict that Higher Potential Evapotranspiration means higher Temperature which in turn results in increased snow melt. This is further explained through the relation of PET and Months, where higher PET values correspond to a higher temperature which in turn reflects summer months in Figure 13.

The hydro-climatic nature of Jhelum Watershed is eccentric due to transboundary data scarcity issues and heavily impacted by two distinct climate trends (monsoon and westerlies circulation). Moreover, under the influence of the westerlies circulation pattern, this watershed receives significant runoff contributions from early summer (April to mid-June) snow- and glacier-melt that builds over the winter season, as well as heavy summer monsoon rains (end of June to August) (Azmat et al., 2018). Shap analysis rightly captures the bird-eye view of the catchment, where Early Kharif, Late Kharif seasons, and snow-melt from the highest elevation zone 5 of the catchment play a vital role in streamflow prediction. Figure 13 further illustrates a logical interaction of Snow water equivalent with Months for the Jhelum catchment.

The aforementioned analysis depicts that all the tree-based algorithms are hydrologically robust and capture the physical linkages among the hydrometeorological parameters logically. Thus, consistency of SHAP identification results with hydrology theory further supports the reliability of the results.

5.3 Model Performance

Except for 2 cases, Kabul and Jhelum Catchment, all algorithms produced adequate prediction accuracy on the test data across the selected study area (Figure 2). The models performed exceptionally well for Indus and Chenab Catchments. The range of R^2 and NSE for all three models lies between 0.61 to 0.90 for four catchments represented through Table 4. Among XGBoost and RandomForest, Random Forest performs relatively better which indicates its suitability for snow-dominated catchments.Figures 7 8 and 9 represents the simulated and actual hydrographs of the XGBoost, RandomForest and CART models.

6 Conclusion

We examined 3 models for data-driven rainfall-runoff modeling for their capacity for streamflow simulation in 4 highly seasonal and data-scarce catchments. Our research has primarily focused on demonstrating the difficulties of using data-driven models in large watersheds (with limited spatial coverage of streamflow observational data), where both rainfall and snowmelt are the sources of streamflow, and observational streamflow data is only accessible at catchment outlets. Our study presents a complete machine learning framework for catchments with a mix of snowmelt and rainfall, which includes hyperparameter optimization methods, performance evaluation methods, and model interpretation methods. The framework is particularly beneficial for high elevation large data scarce catchments where the information for setting up process-based models is insufficient. Further, it is recommended to use the SHAP method in more case studies and compare the representations learned by different models. Using model explanation methods, such as SHAP, provides guiding information in the process of predictor selection for data-scarce catchments by enabling visualization of the contribution of individual variables in a prediction. Moreover, SHAP analysis allows reasons for making each prediction to be explained and the learned representations to be analyzed, which increases the transparency of machine learning, and may eventually promote the application of machine learning methods in hydrological studies.

Author contributions. HM formulated the research questions for this study. TA and HM prepared the scripts for development of Machine learning models. HM performed exploratory analysis of SHAP analysis. AM prepared the catchments characteristics. TA, HM and ZH prepared and revised the manuscript.

Competing interests. The authors have no competing interests.

Acknowledgements. The authors thank the Surface Water Hydrology (SWH) group at Water and Power Development Authority (WAPDA), for providing the streamflow data for this study.

References

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., and Clark, M. P.: A Ranking of Hydrological Signatures Based on Their Predictability in Space, Water Resources Research, 54, 8792–8812, https://doi.org/10.1029/2018WR022606, 2018.

Adnan, R. M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., Li, B., and Kisi, O.: Daily streamflow prediction using optimally pruned extreme learning machine, Journal of Hydrology, 577, 123 981, https://doi.org/10.1016/j.jhydrol.2019.123981, 2019.

Akhtar, T., Mushtaq, H., and Hashmi, M. Z.-u.-R.: Drought monitoring and prediction in climate vulnerable Pakistan: Integrating hydrologic and meteorologic perspectives, Hydrology and Earth System Sciences Discussions, pp. 1–29, https://doi.org/10.5194/hess-2020-297, 2020.

Allen, R. G., Pereira, L., Raes, D., and Smith, M.: Crop evapotraspiration guidelines for computing crop water requirements., 1998.

Azmat, M., Qamar, M. U., Huggel, C., and Hussain, E.: Future climate and cryosphere impacts on the hydrology of a scarcely gauged catchment on the Jhelum river basin, Northern Pakistan, Science of the Total Environment, 639, 961–976, https://doi.org/10.1016/j.scitotenv.2018.05.206, 2018.

Camacho Suarez, V. V., Saraiva Okello, A. M., Wenninger, J. W., and Uhlenbrook, S.: Understanding runoff processes in a semi-arid environment through isotope and hydrochemical hydrograph separations, Hydrology and Earth System Sciences, 19, 4183–4199, https://doi.org/10.5194/hess-19-4183-2015, 2015.

Charles, S. P., Wang, Q. J., Ahmad, M. U. D., Hashmi, D., Schepen, A., Podger, G., and Robertson, D. E.: Seasonal streamflow forecasting in the upper Indus Basin of Pakistan: An assessment of methods, Hydrology and Earth System Sciences, 22, 3533–3549, https://doi.org/10.5194/hess-22-3533-2018, 2018.

Choubin, B., Zehtabian, G., Azareh, A., Rafiei-Sardooi, E., Sajedi-Hosseini, F., and Kişi, Ö.: Precipitation forecasting using classification and regression trees (CART) model: a comparative study of different approaches, Environmental Earth Sciences, 77, https://doi.org/10.1007/s12665-018-7498-z, 2018.

Dehghani, M., Seifi, A., and Riahi-Madvar, H.: Novel forecasting models for immediate-short-term to long-term influent flow prediction by combining ANFIS and grey wolf optimization, Journal of Hydrology, 576, 698–725, https://doi.org/10.1016/j.jhydrol.2019.06.065, 2019.

Ebtehaj, I., Bonakdari, H., Shamshirband, S., and Mohammadi, K.: A combined support vector machine-wavelet transform model for prediction of sediment transport in sewer, Flow Measurement and Instrumentation, 47, 19–27, https://doi.org/10.1016/j.flowmeasinst.2015.11.002, 2016.

Erdal, H. I. and Karakurt, O.: Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms, Journal of Hydrology, 477, 119–128, https://doi.org/https://doi.org/10.1016/j.jhydrol.2012.11.015, 2013.

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

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., and Xiang, Y.: Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China, Energy Conversion and Management, 164, 102–111, https://doi.org/10.1016/j.enconman.2018.02.087, 2018a.

Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X., and Xiang, Y.: Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China, Agricultural and Forest Meteorology, 263, 225–241, https://doi.org/10.1016/j.agrformet.2018.08.019, 2018b.

Fan, J., Wu, L., Zhang, F., Cai, H., Zeng, W., Wang, X., and Zou, H.: Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: A review and case study in China, Renewable and Sustainable Energy Reviews, 100, 186–212, https://doi.org/10.1016/j.rser.2018.10.018, 2019.

Gauch, M., Mai, J., and Lin, J.: The proper care and feeding of CAMELS: How limited training data affects streamflow prediction, Environmental Modelling and Software, 135, 104 926, https://doi.org/10.1016/j.envsoft.2020.104926, 2021.

Gevrey, M., Dimopoulos, I., and Lek, S.: Two-way interaction of input variables in the sensitivity analysis of neural network models, Ecological Modelling, 195, 43–50, https://doi.org/10.1016/j.ecolmodel.2005.11.008, 2006.

Godsey, S. E., Kirchner, J. W., and Tague, C. L.: Effects of changes in winter snowpacks on summer low flows: Case studies in the Sierra Nevada, California, USA, Hydrological Processes, 28, 5048–5064, https://doi.org/10.1002/hyp.9943, 2014.

Hadi, S. J. and Tombul, M.: Forecasting Daily Streamflow for Basins with Different Physical Characteristics through Data-Driven Methods, Water Resources Management, 32, 3405–3422, https://doi.org/10.1007/s11269-018-1998-1, 2018.

Hancock, T., Put, R., Coomans, D., Vander Heyden, Y., and Everingham, Y.: A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies, Chemometrics and Intelligent Laboratory Systems, 76, 185–196, https://doi.org/10.1016/j.chemolab.2004.11.001, 2005.

Hargreaves, G. H. and Samani, Z. A.: Reference Crop Evapotranspiration From Ambient Air Temperature., Paper - American Society of Agricultural Engineers, pp. 96–99, 1985.

Hassan, M. and Hassan, I.: Improving Artificial Neural Network Based Streamflow Forecasting Models through Data Preprocessing, KSCE Journal of Civil Engineering, 25, 3583–3595, https://doi.org/10.1007/s12205-021-1859-y, 2021.

Herath, H. M. V. V., Chadalawada, J., and Babovic, V.: Hydrologically Informed Machine Learning for Rainfall-Runoff Modelling: Towards Distributed Modelling, Hydrology and Earth System Sciences Discussions, pp. 1–42, https://doi.org/10.5194/hess-2020-487, 2020.

Hu, X., Shi, L., Lin, G., and Lin, L.: Comparison of physical-based, data-driven and hybrid modeling approaches for evapotranspiration estimation, Journal of Hydrology, 601, 126 592, https://doi.org/https://doi.org/10.1016/j.jhydrol.2021.126592, 2021.

jing Niu, W., kai Feng, Z., Zeng, M., fei Feng, B., wu Min, Y., tian Cheng, C., and zhong Zhou, J.: Forecasting reservoir monthly runoff via ensemble empirical mode decomposition and extreme learning machine optimized by an improved gravitational search algorithm, Applied Soft Computing Journal, 82, 105 589, https://doi.org/10.1016/j.asoc.2019.105589, 2019.

Jung, Y.: Multiple predicting K-fold cross-validation for model selection, Journal of Nonparametric Statistics, 30, 197–215, https://doi.org/10.1080/10485252.2017.1404598, 2018.

Kapnick, S. B., Yang, X., Vecchi, G. A., Delworth, T. L., Gudgel, R., Malyshev, S., Milly, P. C., Shevliakova, E., Underwood, S., and Margulis, S. A.: Potential for western US seasonal snowpack prediction, Proceedings of the National Academy of Sciences of the United States of America, 115, 1180–1185, https://doi.org/10.1073/pnas.1716760115, 2018.

Karimi, P., Bastiaanssen, W. G., Molden, D., and Cheema, M. J.: Basin-wide water accounting based on remote sensing data: An application for the Indus Basin, Hydrology and Earth System Sciences, 17, 2473–2486, https://doi.org/10.5194/hess-17-2473-2013, 2013.

Knowles, N., Dettinger, M. D., and Cayan, D. R.: Trends in snowfall versus rainfall in the western United States, Journal of Climate, 19, 4545–4559, https://doi.org/10.1175/JCLI3850.1, 2006.

Laghari, A. N., Vanham, D., and Rauch, W.: The Indus basin in the framework of current and future water resources management, Hydrology and Earth System Sciences, 16, 1063–1083, https://doi.org/10.5194/hess-16-1063-2012, 2012.

Lama, L., Wilhelmsson, O., Norlander, E., Gustafsson, L., Lager, A., Tynelius, P., Wärvik, L., and Östenson, C.-G.: Machine learning for prediction of diabetes risk in middle-aged Swedish people, Heliyon, 7, e07 419, https://doi.org/https://doi.org/10.1016/j.heliyon.2021.e07419, 2021.

Le, T., Vo, B., Fujita, H., Nguyen, N.-T., and Baik, S. W.: A fast and accurate approach for bankruptcy forecasting using squared logistics loss with GPU-based extreme gradient boosting, Information Sciences, 494, 294–310, https://doi.org/https://doi.org/10.1016/j.ins.2019.04.060, 2019.

Legates, D. R. and McCabe, G. J.: Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation, Water Resources Research, 35, 233–241, https://doi.org/10.1029/1998WR900018, 1999.

Li, X., Sha, J., and Wang, Z. L.: Comparison of daily streamflow forecasts using extreme learning machines and the random forest method, Hydrological Sciences Journal, 64, 1857–1866, https://doi.org/10.1080/02626667.2019.1680846, 2019.

Liang, Z., Li, Y., Hu, Y., Li, B., and Wang, J.: A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework, Theoretical and Applied Climatology, 133, 137–149, https://doi.org/10.1007/s00704-017-2186-6, 2018.

Liaw, A. and Wiener, M.: Classification and Regression by randomForest, R News, 2, 18–22, 2002.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P.: Understanding variable importances in Forests of randomized trees, Advances in Neural Information Processing Systems, pp. 1–9, 2013.

Lundberg, S. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, 2017.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. I.: Explainable AI for trees: From local explanations to global understanding, arXiv, 2, https://doi.org/10.1038/s42256-019-0138-9, 2019a.

Lundberg, S. M., Erion, G. G., and Lee, S.-I.: Consistent Individualized Feature Attribution for Tree Ensembles, 2019b.

Mangalathu, S., Hwang, S.-H., and Jeon, J.-S.: Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach, Engineering Structures, 219, 110 927, https://doi.org/https://doi.org/10.1016/j.engstruct.2020.110927, 2020.

Meng, E., Huang, S., Huang, Q., Fang, W., Wu, L., and Wang, L.: A robust method for non-stationary streamflow prediction based on improved EMD-SVM model, Journal of Hydrology, 568, 462–478, https://doi.org/10.1016/j.jhydrol.2018.11.015, 2019.

Mukhopadhyay, B. and Khan, A.: A reevaluation of the snowmelt and glacial melt in river flows within Upper Indus Basin and its significance in a changing climate, Journal of Hydrology, 527, 119–132, https://doi.org/10.1016/j.jhydrol.2015.04.045, 2015.

Natekin, A. and Knoll, A.: Gradient boosting machines, a tutorial, Frontiers in Neurorobotics, 7, https://doi.org/10.3389/fnbot.2013.00021, 2013.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, e2020WR028 091, https://doi.org/https://doi.org/10.1029/2020WR028091, e2020WR028091 10.1029/2020WR028091, 2021.

Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., and Liu, J.: Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model, Journal of Hydrology, 586, https://doi.org/10.1016/j.jhydrol.2020.124901, 2020.

Papacharalampous, G. A. and Tyralis, H.: Evaluation of random forests and Prophet for daily streamflow forecasting, Advances in Geosciences, 45, 201–208, https://doi.org/10.5194/adgeo-45-201-2018, 2018.

Parisouj, P., Mohebzadeh, H., and Lee, T.: Employing Machine Learning Algorithms for Streamflow Prediction: A Case Study of Four River Basins with Different Climatic Zones in the United States, Water Resources Management, 34, 4113–4131, https://doi.org/10.1007/s11269-020-02659-5, 2020.

Pham, L., Luo, L., and Finley, A.: Evaluation of Random Forest for short-term daily streamflow forecast in rainfall and snowmelt driven watersheds, Hydrology and Earth System Sciences Discussions, pp. 1–33, https://doi.org/10.5194/hess-2020-305, 2020.

Pham, L. T., Luo, L., and Finley, A.: Evaluation of random forests for short-term daily streamflow forecasting in rainfall- And snowmelt-driven watersheds, Hydrology and Earth System Sciences, 25, 2997–3015, https://doi.org/10.5194/hess-25-2997-2021, 2021.

Qi, M., Fu, Z., and Chen, F.: Research on a feature selection method based on median impact value for modeling in thermal power plants, Applied Thermal Engineering, 94, 472–477, https://doi.org/10.1016/j.applthermaleng.2015.10.104, 2016.

Rahmati, O., Falah, F., Dayal, K. S., Deo, R. C., Mohammadi, F., Biggs, T., Moghaddam, D. D., Naghibi, S. A., and Bui, D. T.: Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia, Science of the Total Environment, 699, 134 230, https://doi.org/10.1016/j.scitotenv.2019.134230, 2020.

Rasouli, K., Hsieh, W. W., and Cannon, A. J.: Daily streamflow forecasting by machine learning methods with weather and climate inputs, Journal of Hydrology, 414-415, 284–293, https://doi.org/https://doi.org/10.1016/j.jhydrol.2011.10.039, 2012.

Ravansalar, M., Rajaee, T., and Kisi, O.: Wavelet-linear genetic programming: A new approach for modeling monthly streamflow, Journal of Hydrology, 549, 461–475, https://doi.org/10.1016/j.jhydrol.2017.04.018, 2017.

Schmidt, L., Heße, F., Attinger, S., and Kumar, R.: Challenges in Applying Machine Learning Models for Hydrological Inference: A Case Study for Flooding Events Across Germany, Water Resources Research, 56, https://doi.org/10.1029/2019WR025924, 2020.

Shortridge, J. E., Guikema, S. D., and Zaitchik, B. F.: Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds, Hydrology and Earth System Sciences, 20, 2611–2628, https://doi.org/10.5194/hess-20-2611-2016, 2016.

Tennant, C., Larsen, L., Bellugi, D., Moges, E., Zhang, L., and Ma, H.: The Utility of Information Flow in Formulating Discharge Forecast Models: A Case Study From an Arid Snow-Dominated Catchment, Water Resources Research, 56, 1–21, https://doi.org/10.1029/2019WR024908, 2020.

THORNTHWAITE, C.: An Approach Toward a Rational, Geographical Review, 38, 55–94, 1948.

ur Rauf, A., Ghumman, A. R., Ahmad, S., and Hashmi, H. N.: Performance assessment of artificial neural networks and support vector regression models for stream flow predictions, Environmental Monitoring and Assessment, 190, https://doi.org/10.1007/s10661-018-7012-9, 2018.

Vezza, P., Comoglio, C., Rosso, M., and Viglione, A.: Low Flows Regionalization in North-Western Italy, Water Resources Management, 24, 4049–4074, https://doi.org/10.1007/s11269-010-9647-3, 2010.

Wang, H., Lv, X., and Zhang, M.: Sensitivity and attribution analysis based on the Budyko hypothesis for streamflow change in the Baiyangdian catchment, China, Ecological Indicators, 121, 107 221, https://doi.org/https://doi.org/10.1016/j.ecolind.2020.107221, 2021.

Wang, S., Peng, H., Hu, Q., and Jiang, M.: Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method, Journal of Hydrology: Regional Studies, 42, 101 139, https://doi.org/10.1016/j.ejrh.2022.101139, 2022.

Wen, X., Feng, Q., Deo, R. C., Wu, M., Yin, Z., Yang, L., and Singh, V. P.: Two-phase extreme learning machines integrated with the complete ensemble empirical mode decomposition with adaptive noise algorithm for multi-scale runoff prediction problems, Journal of Hydrology, 570, 167–184, https://doi.org/10.1016/j.jhydrol.2018.12.060, 2019.

Xia, Y., Liu, C., Li, Y., and Liu, N.: A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring, Expert Systems with Applications, 78, 225–241, https://doi.org/https://doi.org/10.1016/j.eswa.2017.02.017, 2017.

Yaseen, Z. M., Ebtehaj, I., Bonakdari, H., Deo, R. C., Mehr, A. D., Mohtar, W. H. M. W., Diop, L., El-shafie, A., and Singh, V. P.: Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model, Journal of Hydrology, 554, 263–276, https://doi.org/10.1016/j.jhydrol.2017.09.007, 2017.

Yatagai, A., Kamiguchi, K., Arakawa, O., Hamada, A., Yasutomi, N., and Kitoh, A.: Aphrodite constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges, Bulletin of the American Meteorological Society, 93, 1401–1415, https://doi.org/10.1175/BAMS-D-11-00122.1, 2012.

Yu, W., Yang, Y.-C., Savitsky, A., Alford, D., Brown, C., Wescoat, J., Debowicz, D., Robinson, S., Yu, W., Yang, Y.-C., Savitsky, A., Alford, D., Brown, C., Wescoat, J., Debowicz, D., and Robinson, S.: Modeling Water, Climate, Agriculture, and the Economy, The Indus Basin of Pakistan, pp. 95–118, https://doi.org/10.1596/9780821398746_ch05, 2013.

Zhang, J., Zhu, Y., Zhang, X., Ye, M., and Yang, J.: Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas, Journal of Hydrology, 561, 918–929, https://doi.org/10.1016/j.jhydrol.2018.04.065, 2018.

Zhang, R., Chen, Z. Y., Xu, L. J., and Ou, C. Q.: Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province, China, Science of the Total Environment, 665, 338–346, https://doi.org/10.1016/j.scitotenv.2019.01.431, 2019.

https://doi.org/10.5194/hess-2022-213
Preprint. Discussion started: 2 September 2022

Table 1. Elevation bands information for the watershed model of this study

Basin Name	Bands	Elevation Zones(meters)	Area(Km^2)	Zone Area %
Indus	1	0-1400	5097.04	6.90
	2	1401-2200	5675.26	7.68
	3	2201-3000	8342.17	11.29
	4	3001-3800	12922.16	17.49
	5	3801-4600	21446.88	29.03
	6	4601-5400	15685.52	21.23
	7	5401-8578	4721.55	6.39
Jhelum	1	0-1000	4490.99	13.39
	2	1001-2000	10062.41	30.01
	3	2001-3000	8329.57	24.84
	4	3001-4000	7491.24	22.34
	5	4001-6121	3157.10	9.42
Chenab	1	0-1300	1954.54	8.06
	2	1301-2100	2702.91	11.15
	3	2101-3000	3772.52	15.56
	4	3001-3900	4510.45	18.60
	5	3901-4800	6516.03	26.87
	6	4801-5400	3803.26	15.68
	7	5401-7103	988.22	4.08
Kabul	1	0-1000	12016.92	13.16
	2	1001-1667	12277.93	13.45
	3	1668-2335	15268.77	16.72
	4	2336-3002	16246.39	17.80
	5	3003-3669	15337.44	16.81
	6	3670-4337	10665.69	11.68
	7	4338-5003	7260.94	7.95
	8	5003-7701	2212.35	2.42

https://doi.org/10.5194/hess-2022-213
Preprint. Discussion started: 2 September 2022

Table 2. List of Primary Predictor variables

No.	Predictors	Index	Unit	Source
1	Precipitation	PRECIP	mm	APHRODITE
2	Maximum Temerature	Tmax	degree C	CPC
3	Minimum Temperature	Tmin	degree C	CPC
4	Snow Water Equivalent	SWE	mm	GLDAS
5	Potential Evapotranspiration	PET	mm	Hargreaves Equation

Table 3. List of predictor variables corresponding to elevation zones of Upper Indus catchments of Pakistan: a) Indus at Tarbela b)Chenab at Marala c) Kabul at Nowshera and d) Jhelum at Mangla

Catchments.	Features	Elevation Zones	Categorical Features
Indus	35	7	2
Jhelum	25	5	2
Chenab	35	7	2
Kabul	40	8	2

Table 4. Performance Matrix of Machine learning models

Model Type	Catcment Efficiencies (NSE/R^2)			
	Chenab	Jhelum	Indus	Kabul
CART	0.77/0.70	0.64/0.61	0.81/0.78	0.70/0.66
XGBOOST	0.85/0.87	0.63/0.66	0.9/0.9	0.6/0.67
RandomForest	0.83/0.78	0.69/0.63	0.89/0.89	0.72/0.67

https://doi.org/10.5194/hess-2022-213
Preprint. Discussion started: 2 September 2022

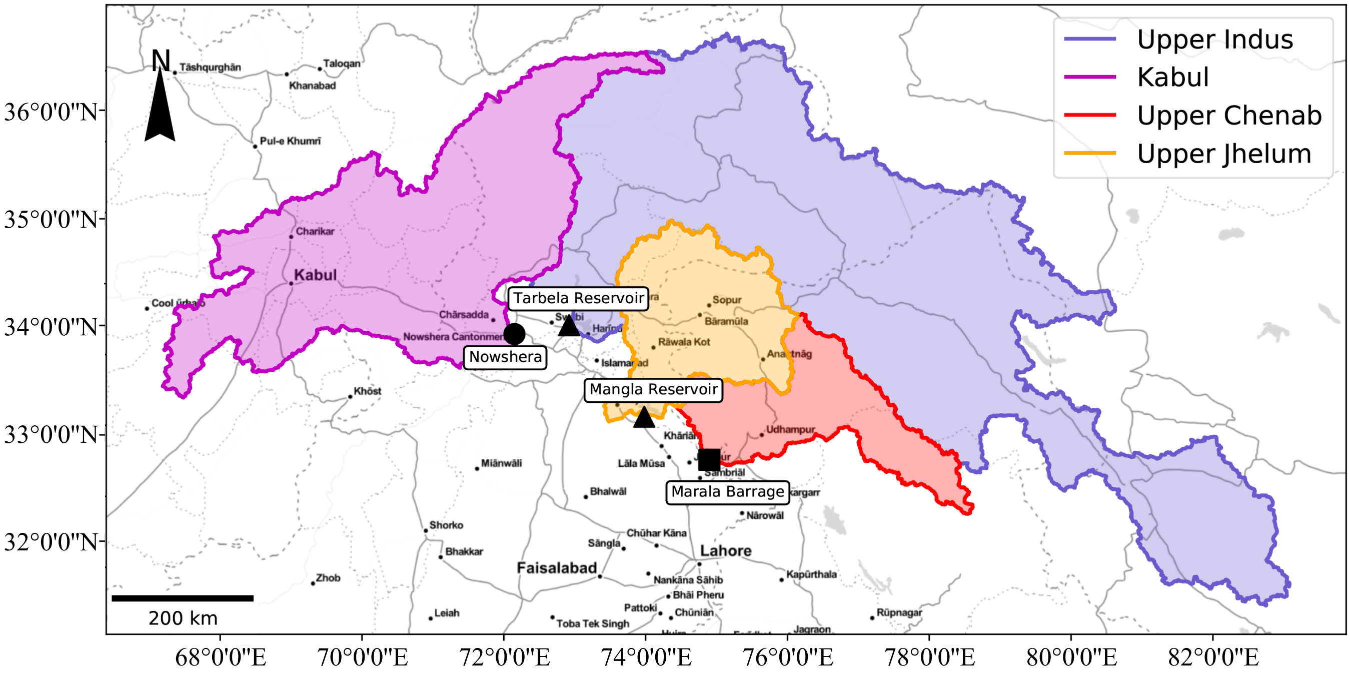

Figure 1. Overview of study area for meteorological and hydrological analysis, that includes the four western river catchments of the Upper Indus Basin, i.e., Upper Indus, Kabul Upper Chenab and Upper Jhelum, and the studied streamflow stations at respective catchment outlets, i.e., Nowshera, Tarbela, Mangla and Marala (Akhtar et al., 2020)

https://doi.org/10.5194/hess-2022-213
Preprint. Discussion started: 2 September 2022

Hydrology and Earth System Sciences
Open Access
Discussions

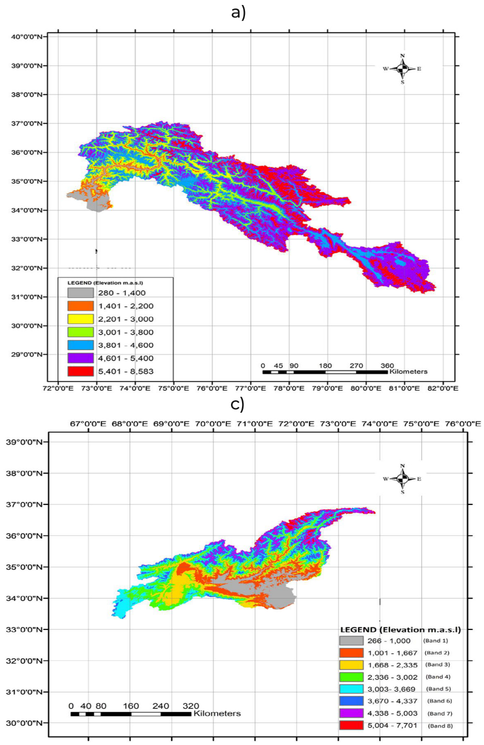

Figure 2. Elevation zones based Digital Elevation Model at outlets of Upper Indus catchments of Pakistan: a) Indus at Tarbela b)Chenab at Marala c) Kabul at Nowshera and d) Jhelum at Mangla

https://doi.org/10.5194/hess-2022-213
Preprint. Discussion started: 2 September 2022

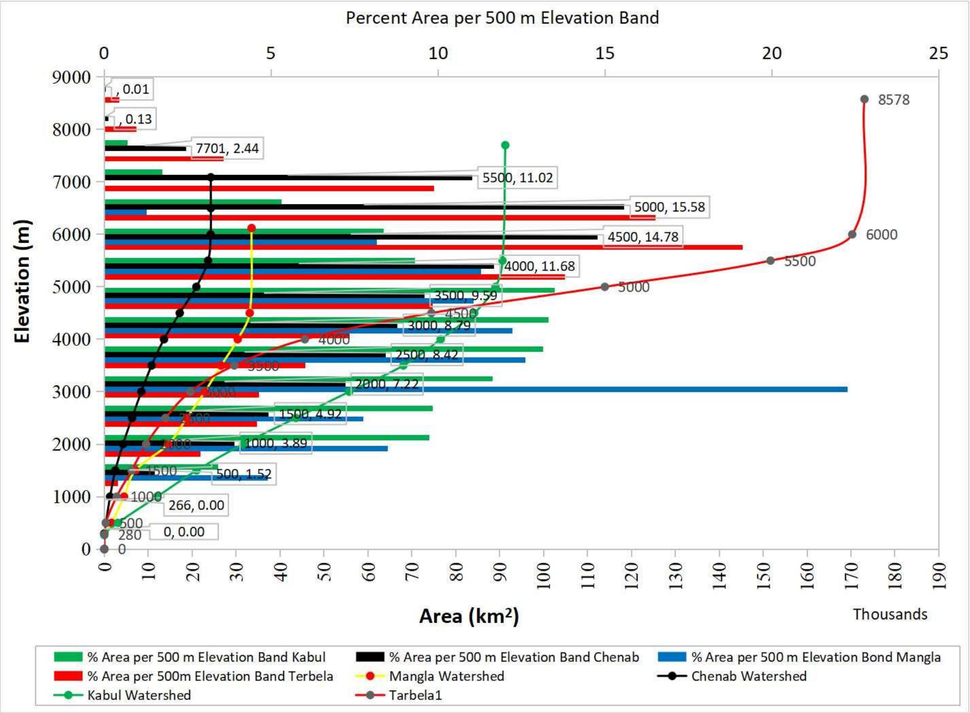

Figure 3. Hypsometric curve of the study watershed at four catchments namely: Indus at Tarbela, Jhelum at Mangla, Chenab at Marala and Kabul at Nowshera

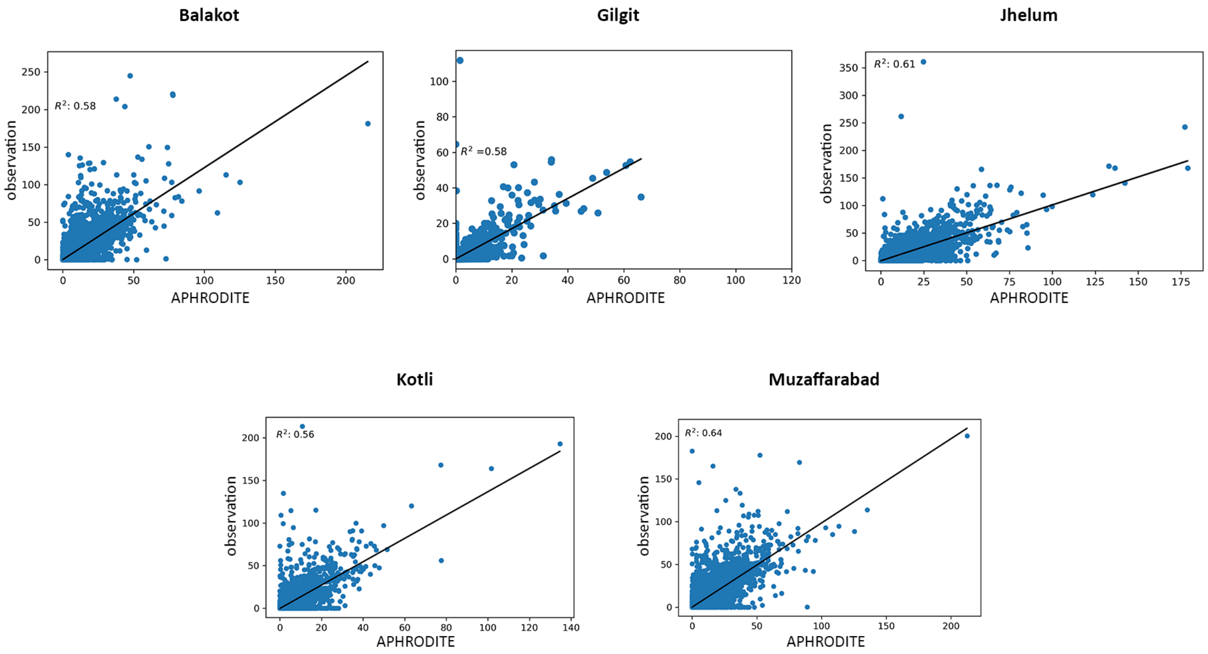

Figure 4. Comparison of daily APHRODITE precipitation with station observations

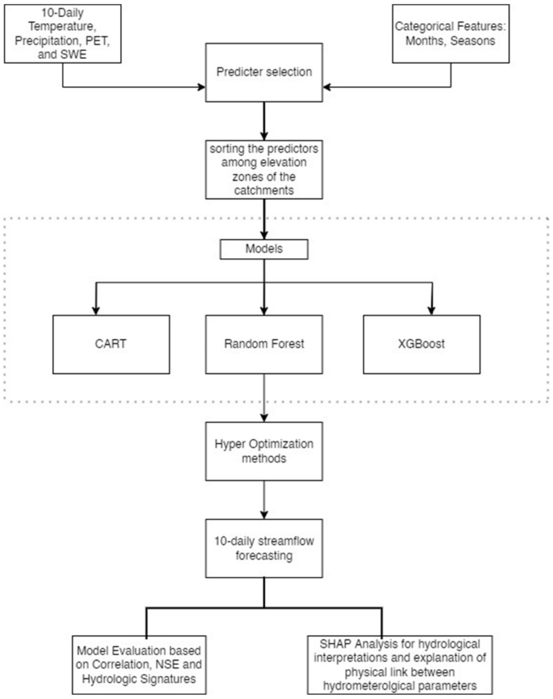
10-Daily Temperature, Precipitation, PET, and SWE
Categorical Features: Months, Seasons
Predicter selection
sorting the predictors among elevation zones of the catchments
Models
CART
Random Forest
XGBoost
Hyper Optimization methods
10-daily streamflow forecasting
Model Evaluation based on Correlation, NSE and Hydrologic Signatures
SHAP Analysis for hydrological interpretations and explanation of physical link between hydrometerolgical parameters

Figure 5. Flow chart of development of ML models for streamflow simulating and interpretation of ML models using SHAP analysis

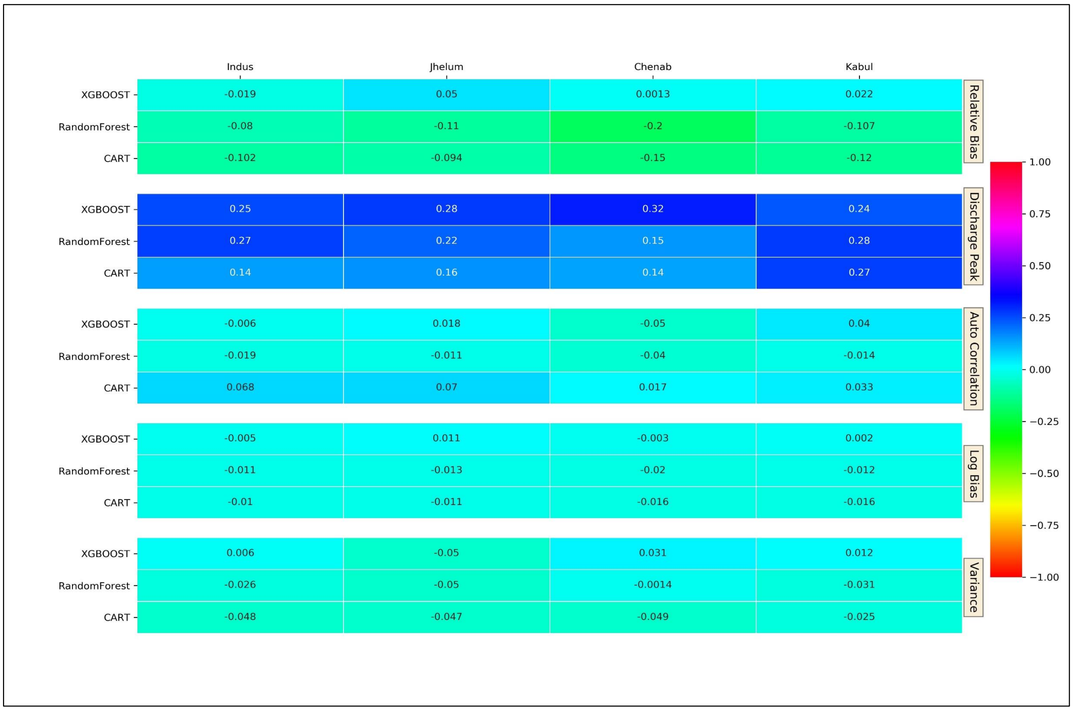

Figure 6. Deviation score of hydrologic signature for each ML algorithm for four catchments: Indus at Tarbela, Jhelum at Mangla, Chenab at Marala and Kabul at Nowshera

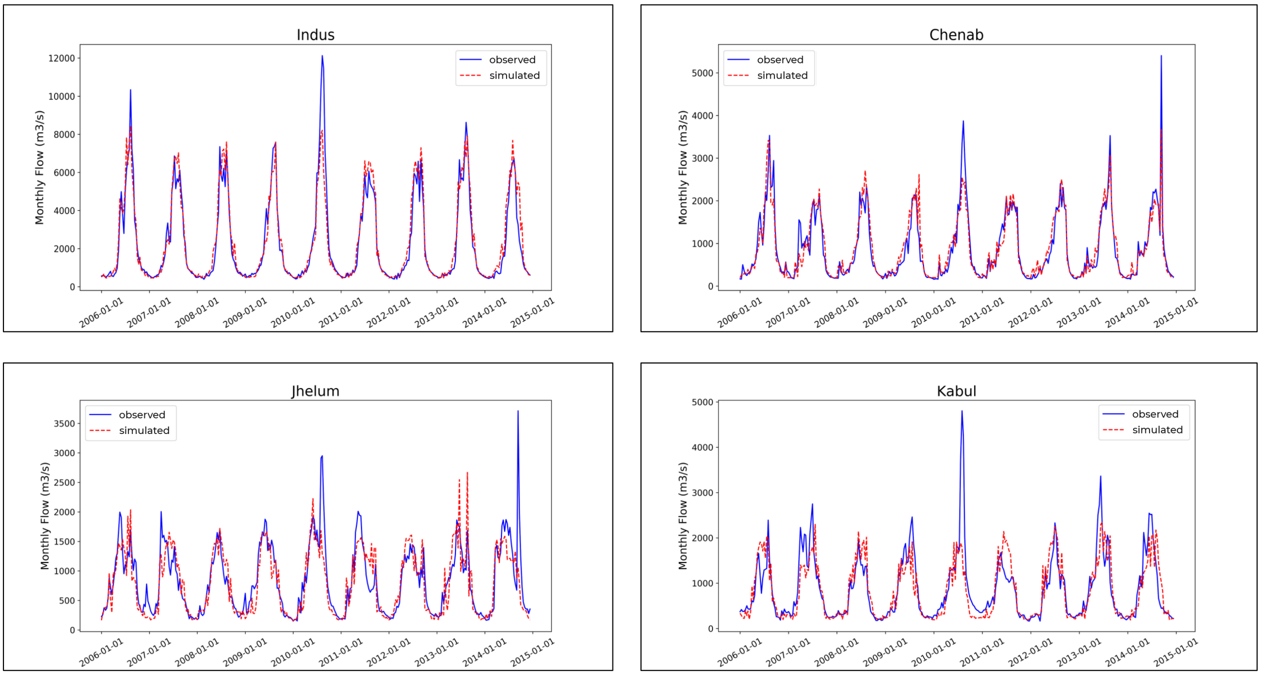

Figure 7. Simulated and observed hydrographs of XGBoost model at outlets of Upper Indus catchments of Pakistan: a) Indus at Tarbela b)Jhelum at Mangla c)Chenab at Marala and d)Kabul at Nowshera

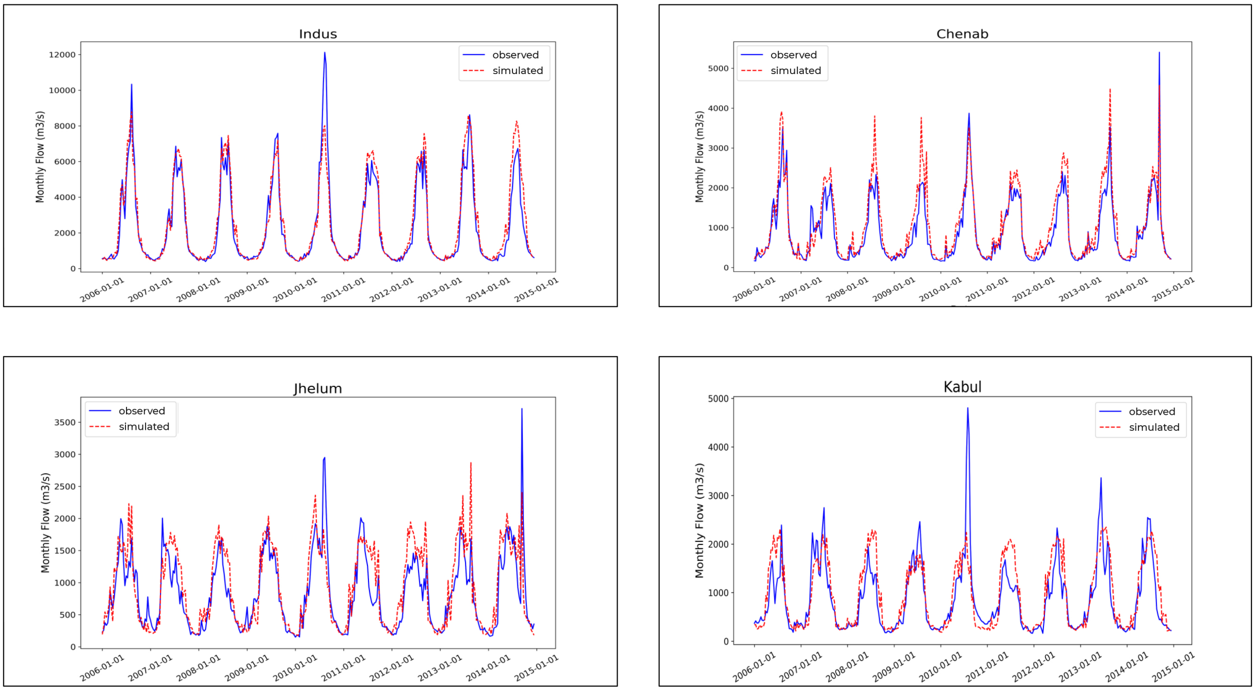

Figure 8. Simulated and observed hydrographs of Random Forest model at outlets of Upper Indus catchments of Pakistan: a) Indus at Tarbela b)Jhelum at Mangla c)Chenab at Marala and d)Kabul at Nowshera

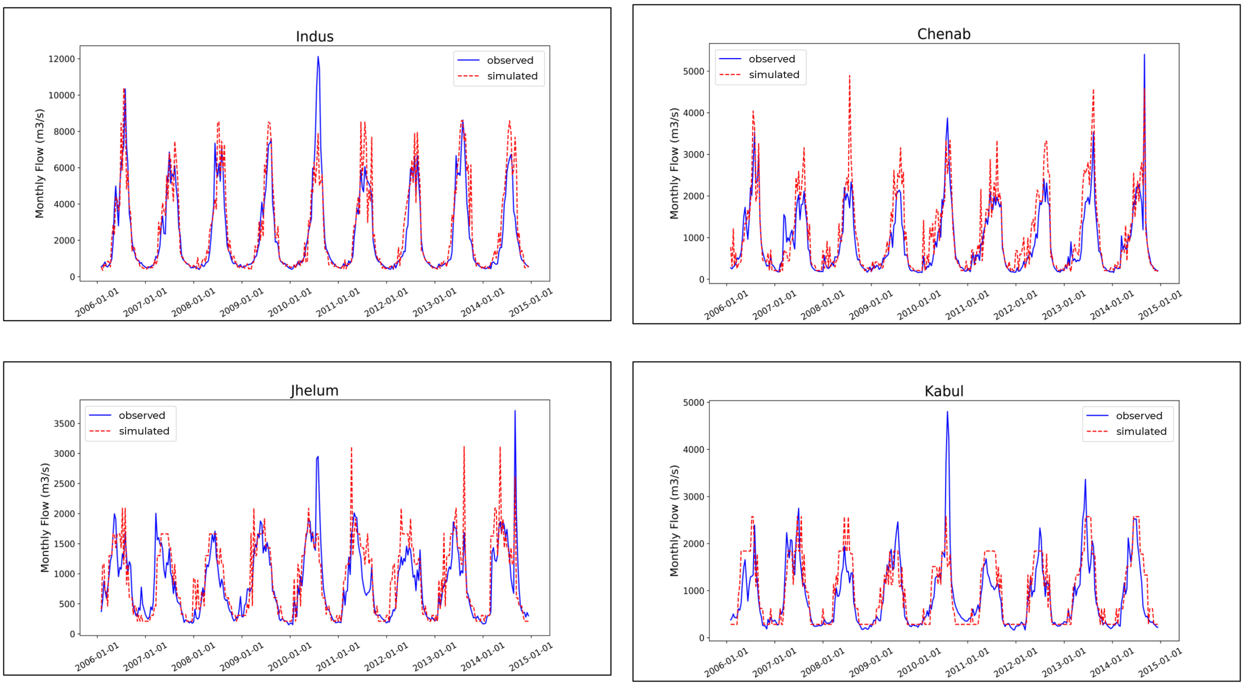

Figure 9. Simulated and observed hydrographs of CART model at outlets of Upper Indus catchments of Pakistan: a) Indus at Tarbela b)Jhelum at Mangla c)Chenab at Marala and d)Kabul at Nowshera

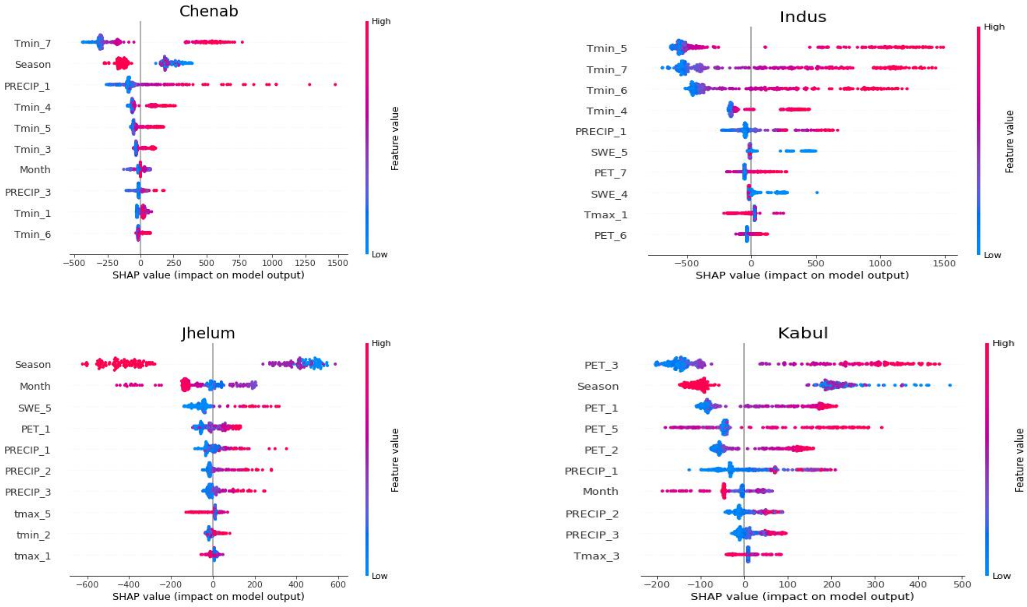

Figure 10. SHAP summary plot. Each dot corresponds to a sample, and its x position shows the impact a predictor variable has on the simulated streamflow for Random Forest Model testing period 2005-2014

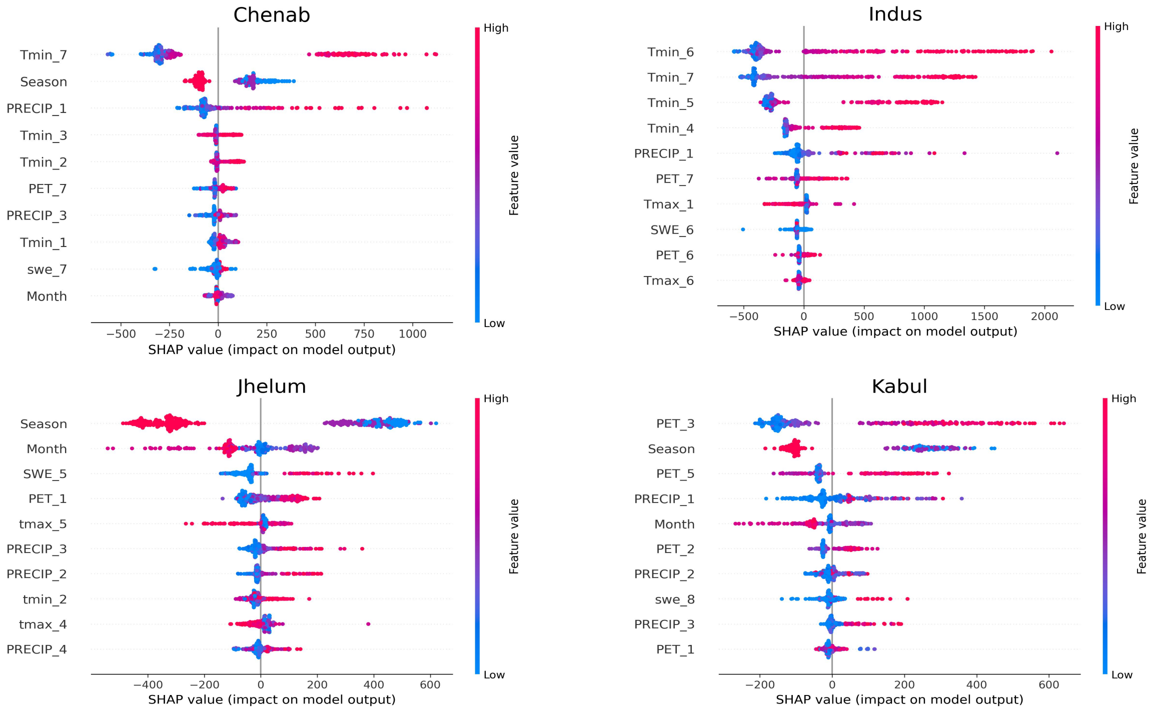

Figure 11. SHAP summary plot. Each dot corresponds to a sample, and its x position shows the impact a predictor variable has on simulated streamflow for XGBoost Model testing period 2005-2014

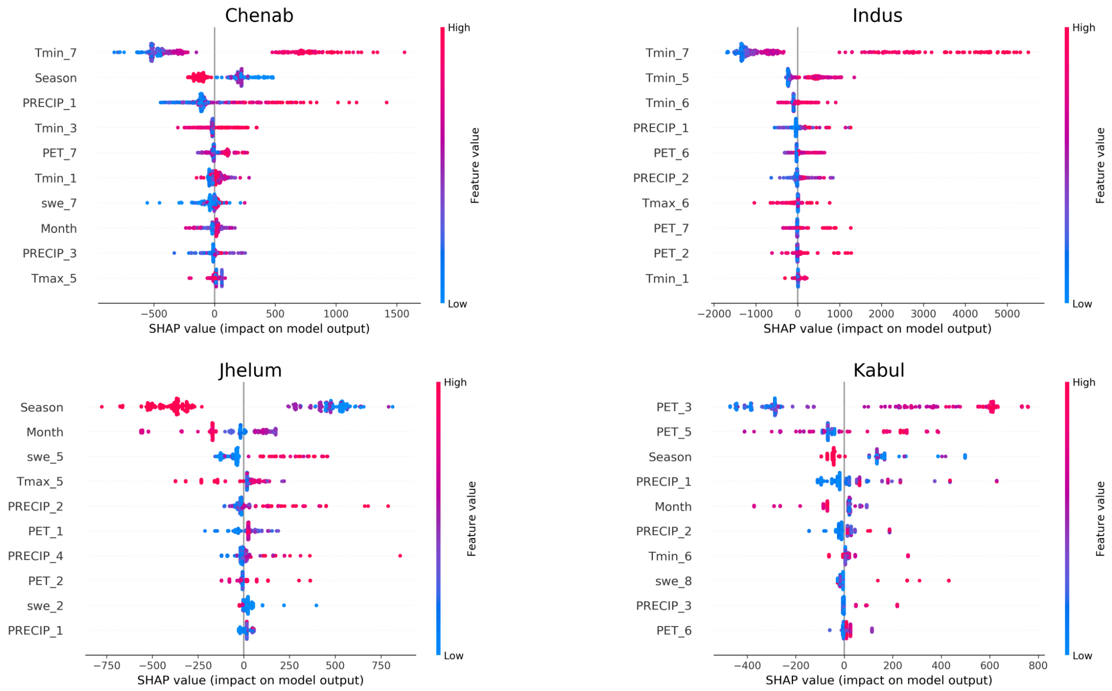

Figure 12. SHAP summary plot. Each dot corresponds to a sample, and its x position shows the impact a predictor variable has on simulated streamflow for CART Model testing period 2005-2014

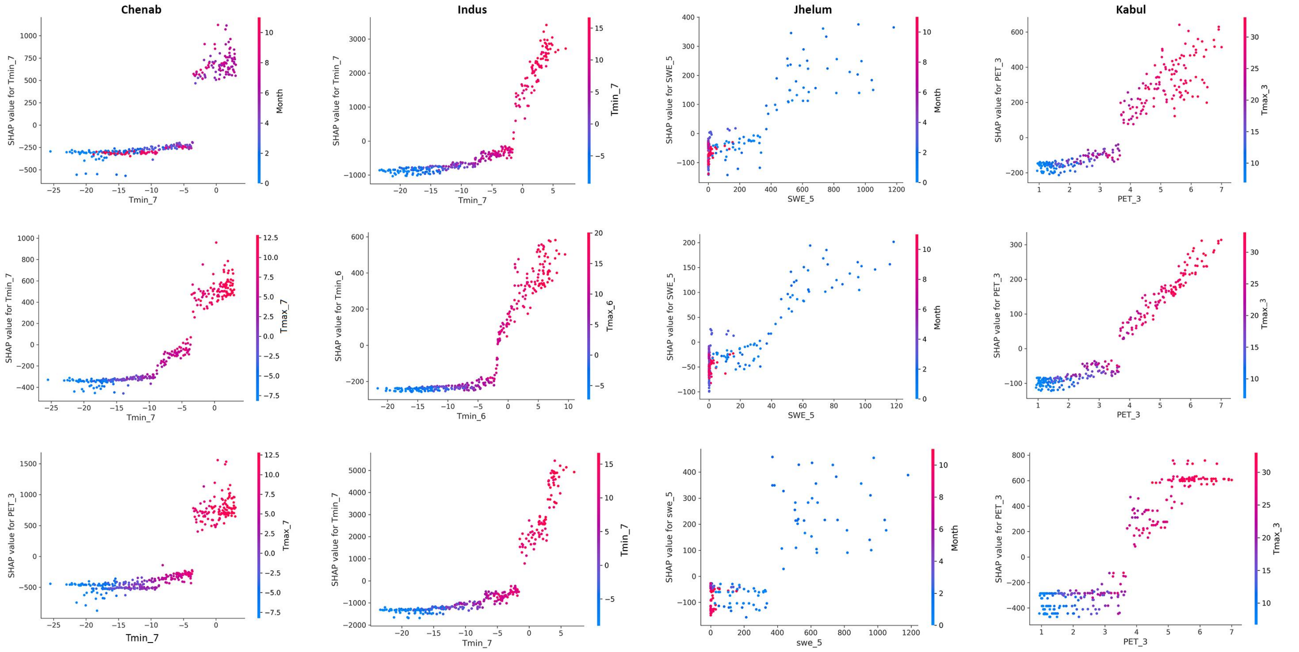

Figure 13. SHAP dependence plots for XGBoost (First Row), RandomForest(Second Row) and CART (Third Row). The variables are plotted against the SHAP values. Each dot corresponds to a sample