Development of a Wilks Feature Importance Method with Improved Variable Rankings for Supporting Hydrological Inference and Modelling

Kailong Li[1], Guohe Huang[1], Brian Baetz[2]

[1]Faculty of Engineering, University of Regina, Regina, Saskatchewan, Canada S4S 0A2
[2]Department of Civil Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4L8.
Correspondence to: Guohe Huang (huangg@uregina.ca)

Abstract. Feature importance has been a popular approach for machine learning models to investigate the relative significance of model predictors. In this study, we developed a Wilks feature importance (WFI) method for hydrological inference. Compared with conventional feature importance methods such as permutation feature importance (PFI) and mean decrease impurity (MDI), the proposed WFI aims to provide more reliable variable rankings for hydrological inference. To achieve this, WFI measures the importance scores based on Wilk's Λ (a test statistic that can be used to distinguish the differences between two or more groups of variables) throughout an inference tree. Compared with PFI and MDI methods, WFI does not rely on any performance measures to evaluate variable rankings, which can thus result in less biased criteria selection during the tree deduction process. The proposed WFI was tested by simulating monthly streamflows for 673 basins in the United States and applied to three interconnected irrigated watersheds located in the Yellow River Basin, China, through concrete simulations for their daily streamflows. Our results indicated that the WFI could generate stable variable rankings in response to the reduction of irrelevant predictors. In addition, the WFI selected predictors helped RF achieve its optimum predictive accuracy, which indicates the proposed WFI could identify more informative predictors than other feature importance measures.

1 Introduction

Machine learning (ML) has been used for hydrological forecasting and examining modeling processes underpinned by statistical and physical relationships. Due to the rapid progress in data science, increased computational power, and the recent advances in ML, the predictive accuracy of hydrological processes has been greatly improved (Reichstein et al., 2019; Shortridge et al., 2016). Yet, the explanatory power of ML models for hydrological inference has not increased apace with their predictive power for forecasting (Konapala and Mishra, 2020). Previous studies have indicated that purely pursuing predictive accuracy may not be a sufficient reason for applying

Deleted: Gaining Hydrological Insights Through Wilk's
Deleted: : A Test-Statistic Interpretation method
Deleted: Reliable and Robust
Formatted: Snap to grid
Deleted: Wilk's
Deleted: in
Deleted: importance scores that could partially address the equifinality problem in hydrology.
Deleted: the
Deleted: -
Deleted: a decision
Deleted: The WFI has an advantage over PFI and MDI as it does not account for predictive accuracy so the risk of overfitting will be greatly reduced.
Deleted: . By employing the recursive feature elimination approach, our
Deleted: more
Deleted: relative importance scores
Deleted: , as compared with PFI and MDI embedded in three different machine learning algorithms.
Deleted: comparative study also shows that the
Deleted: identified by WFI achieved the highest
Deleted: on the testing dataset
Deleted: among many irrelevant ones. We also extended the WFI to the local importance scores for reflecting the varying characteristics of a predictor in the hydrological processes. The related findings could help to gain insights into different hydrological behaviours
Deleted: hydrological
Deleted: descriptive power (also known as interpretability)
Deleted: hydrological

a certain hydrological model to a given problem (Beven, 2011). The ever-increasing data sources allow ML models to incorporate potential driven forces that cannot be easily considered in physically-based hydrological models (Kisi et al., 2019). The increasing volume of input information has left one challenge as "how to extract interpretable information and knowledge from the model." Even though obtaining exact mappings from data input to prediction is technically infeasible for ML models, previous research has shown opportunities to understand the model decisions through either post-hoc explanations or statistical summaries of model parameters (Murdoch et al., 2019). Nevertheless, the reliability of the interpretable information is still less studied. Therefore, quality interpretable information from ML models is much desired for evolving our understanding of nature's laws (Reichstein et al., 2019).

The main idea of model interpretation is to understand the model decisions, including the main aspects of (i) identifying the most relevant predictor variables (i.e., predictors) leading to model predictions and (ii) reasoning why certain predictors are responsible for a particular model response. Interpretability can be defined as the degree to which a human can understand the cause of a decision (Miller, 2019). The model interpretation for ML is mainly achieved through feature importance, which relies on techniques that quantify and rank the variable importance (i.e., a measure of the influence of each predictor to predict the output) (Scornet, 2020). The obtained importance scores can be used to explain certain predictions through relevant knowledge. Moreover, Gregorutti et al. (2017) pointed out that some irrelevant predictors may have a negative effect on the model accuracy. Therefore, eliminating irrelevant predictors might improve the predictive accuracy. Feature importance methods can be categorized as model-agnostic and model-specific (Molnar, 2020). The model-agnostic methods refer to extracting post-hoc explanations by treating the trained model as a black box (Ribeiro et al., 2016a). Such methods usually follow a process of learning an interpretable model based on the outputs of the black-box model (Craven and Shavlik, 1996) and perturbing inputs and seeing the response of the black-box model (Ribeiro et al., 2016b). Such methods mainly include permutation feature importance (PFI) (Breiman, 2001a), partial dependence (PD) plots (Friedman, 2001), individual conditional expectation (ICE) plots (Goldstein et al., 2015), accumulated local effects (ALE) plots (Apley and Zhu, 2016), local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016b), Morries method (Morris, 1991) and Shapley values (Lundberg and Lee, 2017; Shapley, 1953). In hydrology, Yang and Chui

Deleted: Recent studies have addressed the importance of extracting interpretable information and knowledge from big

Deleted: ,

Deleted: evolve our understanding of nature's laws behind the ML modeling processes (Murdoch et al., 2019;Reichstein et al., 2019). In hydrology, studies have indicated that the hydrological inference through ML has the

Deleted: to deal with the problem of equifinality that exists in most physically-based hydrological model descriptions (Schmidt et al., 2020;Shortridge et al., 2016). In hydrology, equifinality means different hydrological model structures and/or parameter sets describe similar observed behaviors with similar accuracy (Beven, 2011). As a result, the same hydrological behavior can thus

Deleted: described by the non-unique parameter sets associated with different physical laws. This problem could severely hamper our understanding of the underlying functioning in the hydrologic system (Clark et al., 2011). One possible cause of this problem is that the

Deleted: are constrained by additional physical information provided through a priori knowledge of hydrologic functioning encoded within both model structure and states and flux relationships (Clark et al., 2015;Schmidt et al., 2020). The ML models on another hand, are more flexible than physically-based hydrological models as they can approximate any complex relationships without relying on additional physical

Deleted: , thus structural and parameterization errors can be greatly reduced (Nearing et al., 2016). The above motivations have led us to improve the interpretability of ML models to gain reliable

Deleted: for hydrological inference.

Formatted: Font: LMRoman12-Regular, English (United States)

Deleted: includeing

Deleted: (i.e., independent variables

Deleted: statistical significance of input predictors and their effect on the model response. The obtained importance scores (i.e., statistical significance of input predictors) can be used to explain certain predictions through relevant knowledge…

Deleted: original

Deleted: (Ribeiro et al., 2016b). This can be achieved by learning an interpretable model based on the outputs of the black …

Deleted: ,

Deleted: (Ribeiro et al., 2016a).

Deleted: (Breiman, 2001)

Deleted: (Ribeiro et al., 2016a)

(2020) used Shapley values to explain individual predictions of hydrological response in sustainable drainage systems at fine temporal scales. Kratzert et al. (2019a) used Morries method to estimate the rankings of predictors for a long short-term memory (LSTM) model. Worland et al. (2019) used the LIME to infer the relation between basin characteristics and the predicted flow duration curves. Konapala and Mishra (2020) used partial dependence plots to understand the role of climate and terrestrial components in the development of hydrological drought. Compared with the above model-agnostic methods, PFI is more widely used in hydrological inference due to its high efficiency and ability to take global insights into model behaviors (Molnar, 2020). Recent applications of PFI include inferring the relationship between basin characteristics and predicted low flow quantiles (Ahn, 2020) and comparing the interpretability among multiple machine learning models in the context of flood events (Schmidt et al., 2020). The above model-agnostic methods are handy for comparative studies of ML models with exceedingly complex (such as deep neuron networks) algorithmic structures to extract the interpretable information.

On the other hand, the model-specific methods (also known as interpretable models), such as decision trees and sparse regression models, can inspect model components directly (Ribeiro et al., 2016a). For instance, the weights (or coefficients) of a linear regression model can directly reflect how the predictions are produced, thus can provide critical information for ranking the model predictors. Due to the oversimplified input-output relationships, linear regression models may be inadequate to approximate the complex reality. As a consequence, these models may hardly achieve satisfactory predictive accuracy and obtain quality interpretable information. As one of the essential branches of interpretable models, tree-structured models such as classification and regression trees (CART) (Breiman et al., 1984) have been an excellent alternative to linear regression models for solving complex non-linear problems. The principle of CART is to successively split the training data space (i.e., predictors and response) into many irrelevant subspaces. These subspaces and the splitting rules will form a decision/regression tree, which asks each of the new observations a series of "Yes/No" questions and guides it to the corresponding subspaces. The model prediction for a new observation shares the same value as the average value for the training responses in that particular subspace. Mean decrease impurity (MDI) is the feature importance method for CART, and it summarizes how much a predictor can improve the model performance through the paths of a tree. Compared with linear regression models, trees are more

Deleted: owing

Deleted: its

Deleted: particularly useful

Deleted: where the underlying algorithmic structure is

Deleted: for direct extraction of

Deleted: from big data

Formatted: Font color: Auto, English (Canada)

Formatted: Font color: Custom Color(RGB(39,37,37)), English (United States)

Deleted: When compared to the model-agnostic methods, the model-specific methods (also known as interpretable models) such as decision trees and sparse regression models, can inspect model components directly (e.g., through the paths in a decision tree or the weight of a specific predictor in a linear model) (Ribeiro et al., 2016b). In fact, regression tree ensembles (RTEs) as one of the important branches in ML, are composed of hundreds of interpretable models (i.e., decision trees). As long as the predictive performance is satisfied, a reasonable inference can be achieved through statistical summaries (e.g., mean decrease in node impurity (Breiman, 2001) or how often a predictor has been used for node splitting (Chen et al., 2015)) of the decision trees. In hydrology, RTEs have been receiving increasing attention for hydrological forecasting owing to their superior predictive accuracy, yet their usefulness for hydrological inference in terms of using such a transparent inference process is still limited. Such studies can be found from Worland (2018) and Lawson et al. (2017). A possible reason causing the interpretable models to be less preferrable than model-agnostic interpretation methods is that people believe higher predictive accuracy can potentially lead to more faithful inference (Murdoch et al., 2019). Nevertheless, interpretable models such as decision trees are still considered understandable tools for inferring a particular model behavior because the transparent decision-making process functions similarly to how the human brain makes decisions for a series of questions.¶
¶
Even though both types of interpretation methods have been applied for hydrological inference, they possess several drawbacks. The model-agnostic interpretation methods assume that the same predictive accuracy will lead to the same, or at least similar inferences (i.e., importance scores). However, Schmidt et al. (2020) disclosed that such an assumption may not be valid since the problem of equifinality (which exists in conventional hydrological model descriptions) also exists for ML model inferences (i.e., different importance scores can be observed from different ML models). Such inconsistency in the inference may hamper effective reasoning for hydrological processes. The interpretable models, on the other hand, suffer less from the equifinality problem since the importance scores can be inspected internally (e.g., through paths of a decision tree). However, Scornet (2020) revealed that the interpretability of the mean decrease in impurity (MDI) (for measuring the importance scores of the classification and regression trees (CART) (Breiman et al., 1984)) is strongly affected by the ...

understandable for inferring a particular model behavior because the transparent decision-making process functions similarly to how the human brain makes decisions for a series of questions (Murdoch et al., 2019). Based on CART, Breiman (2001a) proposed an ensemble of trees named random forest (RF), which significantly improved the predictive accuracy compared with CART. Previous studies reported that RF could outperform many other ML models in predictive accuracy (Fernández-Delgado et al., 2014; Galelli and Castelletti, 2013; Schmidt et al., 2020). The high predictive accuracy allowed RF to become very useful in interpretation, especially in hydrology (Lawson et al., 2017; Worland, 2018). As Murdoch et al. (2019) argued, higher predictive accuracy can lead to a more reliable inference.

Owing to its widespread success in prediction and interpretation, Breiman's RF has been under active development during the last two decades. For instance, Athey et al. (2019) presented generalized random forests for solving heterogeneous estimating equations. Friedberg et al. (2020) proposed a local linear forest model to improve the conventional RF in terms of smooth signals. Ishwaran et al. (2008) introduced random survival forests, which can be used for the analysis of right-censored survival data. Wager and Athey (2018) developed a nonparametric causal forest for estimating heterogeneous treatment effects (HTE). Du et al. (2021) proposed another variant of random forests to help HTE inference through estimating some key conditional distributions. Katuwal et al. (2020) proposed several variants of heterogeneous oblique random forest employing several linear classifiers to optimize the splitting point at the internal nodes of the tree. These new variants of RF are primarily focused on handling various regression and classification tasks or improving the predictive accuracy, yet the usefulness for interpretation is still less studied.

In fact, many studies have reported that the feature importance methods used in Breiman's RF (including PFI and MDI) are unstable (i.e., a small perturbation of training data may significantly change the relative importance of predictors) (Bénard et al., 2021; Breiman, 2001b; Gregorutti et al., 2017; Strobl et al., 2007). Such instability has become one of the critical challenges for the practical use of current feature importance measures. Yu (2013) defined that statistical stability holds if statistical conclusions are robust or stable to appropriate perturbations. In hydrology, stability is critical in terms of interpretation and prediction. For interpretation, if a distinctive set of variable rankings was observed after a small perturbation of training data, it thus unable to conclude realistic reasonings of hydrological processes. For prediction, there is no guarantee that the predictors with low rankings do not bear more valuable information than the higher ones. This problem challenges the selection of a subset of predictors for the optimum predictive accuracy (Gregorutti et al., 2017). Strobl et al. (2008) and Scornet (2020) disclosed that positively correlated predictors would lead to biased criteria selection during the tree deduction process, which further amplifies such instability. To address the issues mentioned above, Hothorn et al. (2006) proposed an unbiased node splitting rule for criteria selection. The proposed method showed that the predictive performance of the resulting trees is as good as the performance of established exhaustive search procedures used in CART. Strobl et al. (2007) examined Hothorn's method under the RF framework, which was called Cforest. They found that the bias of criteria selection can be further reduced if their method is applied using subsampling without replacement. Nevertheless, Xia (2009) found that Cforest only outperformed Breiman's RF in some extreme cases and concluded that RF was able to provide more accurate predictions and more reliable PFI compared to Cforest. A similar finding was also achieved by Fernández-Delgado et al. (2014), who reported RF was likely to be the best among 179 ML algorithms (including Cforest) in terms of predictive accuracy based on 121 data sets. More recently, Epifanio (2017) proposed a feature importance method called intervention in prediction measure (IPM), which was reported as a competitive alternative to other PFI and MDI. Since the proposed IPM was specifically designed for high-dimensional problems (i.e., the number of predictor is much larger than the number of observed samples), which thus is not suitable for most hydrological problems. Bénard et al. (2021) proposed a stable rule learning algorithm (SIRUS) based on RF. The algorithm (which aimed to remove the redundant paths of a decision tree) has indicated stable behavior when data is perturbed, while the predictive accuracy was not as good as the Breiman's RF. To sum up, the existing approaches do not guarantee stable and reliable variable ranking for robust interpretability and optimum predictive accuracy.

Therefore, as an extension of the previous efforts, the objective of this study is to develop a Wilks feature importance (WFI) method with improved variable rankings for supporting hydrological inference and modelling. WFI is based on an advanced splitting procedure, stepwise cluster analysis (SCA) (Huang, 1992), which employed statistical significance of *F*-test, instead of least square fitting (used in CART), to determine the optimum splitting points. These points, in combination with the subsequent sub-cluster mergence, can eventually lead to the desired inference tree for variable rankings. The importance scores of predictors can then be obtained according to the values of Wilk's Λ for reflecting the significance of differences between two or more groups of response variables. Compared with MDI and PFI, WFI does not rely on any performance measures (e.g., least-square errors in MDI or mean square errors in PFI), and can thus result in less biased criteria selection during the tree deduction process. Comparative assessment of WFI, PFI and MDI performances under the RF framework will then be undertaken through efforts in simulating monthly streamflows for 673 basins in the United States. With a finer temporal resolution, the proposed approach has also been applied to three irrigated watersheds in the Yellow River Basin, China, through concrete simulations for their daily streamflows.

2 Related Works

2.1. Random Forest

RF is an ensemble of decision trees, each of which is grown in accordance with a random subset of predictors and a bootstrapped version of the training set. As the ensemble members (trees) increase, the non-linear relationships between predictors and responses become increasingly stable. The prediction can thus be more robust and accurate (Breiman, 2001a; Zhang et al., 2018). The training set for building each tree is drawn randomly from the original training dataset with replacement. Such bootstrap sampling process will leave about 1/3 of the training dataset as out-of-bag (OOB) data, which thus can be used as a validation dataset for the corresponding tree.

There are many variants of RF according to the types of trees (e.g., CART). Based on splitting rules equipped in different types of trees, the resulting RF may use various feature importance measures. In this study, Breiman's RF is selected as the benchmark algorithm to investigate the feature importance measures. The algorithm is implemented using the R package "randomForest" (Liaw and Wiener, 2002). There are three hyperparameters in RF as the number of trees (***Ntree***), the minimum number of samples in a node (***Nmin***) for a splitting action, and the number/ratio of predictors in a subspace (***Mtry***). In addition, Breiman's RF has two feature importance measures: permutation feature importance (PFI) and mean decrease impurity (MDI).

2.2. Permutation Feature Importance

PFI was initially proposed by Breiman (2001a) and can be described as follows: Assume a trained decision tree $\boldsymbol{t}$ (where $\boldsymbol{t} \in \{1,..., \boldsymbol{ntree}\}$; $\boldsymbol{ntree}$ is the total number of decision trees in the forest) with a subset of predictor $\boldsymbol{u}$ (where $\boldsymbol{u} \in \boldsymbol{p}$; and $\boldsymbol{p}$ is complete set of predictors), predictor matrix $\boldsymbol{X}$ (with full predictors), response vector $\boldsymbol{Y}$, predicted vector $\boldsymbol{Y'}$, and an error measure $\boldsymbol{L(Y, Y')}$; (1) calculate the original model error based on the OOB dataset of the $\boldsymbol{t_{th}}$ decision tree: $t(e_{orginal}) = L(Y, t(X^u))$ (where X^u is a subset of predictor matrix $\boldsymbol{X}$); (2) for each predictor $\boldsymbol{j}$ (where $\boldsymbol{j} \in \{1,..., \boldsymbol{p}\}$), (i) generate permuted predictor matrix $\boldsymbol{X_{perm, j}}$ by duplicating $\boldsymbol{X}$ and shuffling the values of predictor $\boldsymbol{X_j}$, (ii) estimate error for the permuted dataset $t(e_{perm,\ j}) = L(Y, t(X^u_{perm,j}))$; and (iii) calculate variable importance of predictor $\boldsymbol{j}$ for the $\boldsymbol{t_{th}}$ decision tree as $PFI(t)_j = t(e_{perm,\ j}) - t(e_{orginal})$; (note that $PFI(t)_j = 0$ if predictor $\boldsymbol{j}$ is not in $\boldsymbol{u}$); (3) calculate the variable importance for the forest by averaging the variable importance over all trees: $\boldsymbol{PFI_j} = \frac{1}{ntree}\sum_{t=1}^{ntree} PFI(t)_j$. The error measure $\boldsymbol{L(Y, Y')}$ used in this study is mean squared error (MSE), given by:

$$MSE = \frac{1}{n}\sum_{n=1}^{N}\left(y_n - y_n^*\right)^2 \qquad (1)$$

where y_n and y_n^* are the $\boldsymbol{n^{th}}$ observed and predicted quantities, respectively; $\boldsymbol{N}$ is the total number of quantities.

2.3. MDI feature importance

The MDI importance measure is based on the CART decision tree, which is illustrated using a hydrological dataset (Figure 1) including 20 instances and 3 predictors as $\boldsymbol{X_1}$ (i.e., precipitation), $\boldsymbol{X_2}$ (i.e., 3-day cumulative precipitation) and $\boldsymbol{X_3}$ (i.e., temperature), and a response $\boldsymbol{Y}$ (i.e., streamflow). It starts by sorting the value of $\boldsymbol{X_j}$ in ascending order ($\boldsymbol{j}$ indicates the column index of the predictors so that $\boldsymbol{j} \in \{1, 2, 3\}$), and the $\boldsymbol{Y}$ will be reordered accordingly. Then we go through each instance of $\boldsymbol{X_j}$ from the top to examine each candidate split point. For a sample set with $\boldsymbol{k}$ instances, the total number of split points for $\boldsymbol{X_j}$ will be $\boldsymbol{k-1}$. Any instance $\boldsymbol{z}$ (where $\boldsymbol{z} \in \{1, \ldots, \boldsymbol{k}\}$) in $\boldsymbol{X_j}$ can split the predictor space into two subspaces as $\boldsymbol{X_1(i, j)} = \{\boldsymbol{X_{1,j}}, \boldsymbol{X_{2,j}}, \ldots, \boldsymbol{X_{z,j}}\}$(where $\boldsymbol{i} \in$

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Deleted: three

Deleted: $j \in 1, 2, 3$),

Deleted: It should be noted that if there are $\boldsymbol{K}$

Deleted: ($\boldsymbol{K}$=20 in this case),

Deleted: $\boldsymbol{K}$

Deleted: ($z \in 1, 2, \ldots, K$)

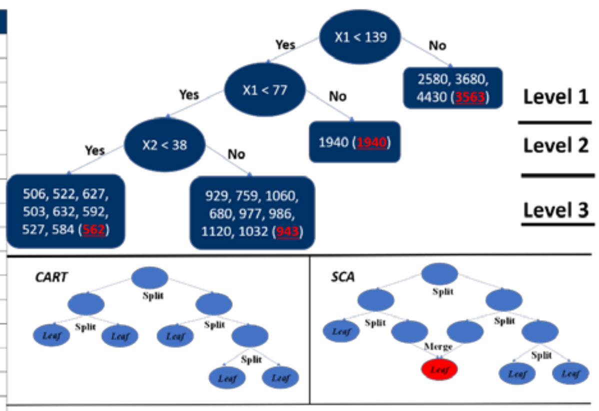
Deleted: $X_1(j, z) = \{X_{j,1}, X_{j,2}, \ldots, X_{j,z}\}$; $j \in 1, 2, 3$

{1, …, z}); and $X_2(i, j)$ = {$X_{z+1,j}$, $X_{z+2,j}$, …, $X_{k,j}$} (where $i \in$ {z+1, …, k}). The response space Y will be correspondingly divided into two subspaces as $Y_1(i)$ = {Y_1, Y_2, …, Y_z} (where $i \in$ {1, …, z}); and $Y_2(i)$ = { Y_{z+1}, Y_{z+2}, …, Y_k} (where $i \in$ {z+1, …, k}). To maximize the predictive accuracy, the objective of the splitting process is to find the split point (based on the row and column coordinate z and j, respectively) with the minimum squared errors (SE) of Y_1 and Y_2:

$$SE(z, j) = \sum_{i=1}^{z}\left(Y_1(i) - \overline{Y}_1\right)^2 + \sum_{i=z}^{k}\left(Y_2(i) - \overline{Y}_2\right)^2 \;;\; \forall\, z \text{ in } 1, \ldots, k-1 \;;\; \forall\, j \text{ in } 1, \ldots, 3 \qquad (2)$$

where $\overline{Y}_1$ and $\overline{Y}_2$ indicate the mean value of Y_1 and Y_2, respectively.

Instance	X1	X2	X3	Y
1	244.5	244.5	16.9	4430
2	23.4	23.4	6.3	592
3	58.6	144.3	12.5	1032
4	5.6	8.6	14.5	627
5	38.6	52.2	12.1	1060
6	183.2	183.2	17.4	2580
7	94.6	128.8	21.3	1940
8	0	1.2	5.2	506
9	18.5	18.5	23	632
10	0	27.6	9.6	527
11	43.2	97.3	19.6	986
12	4.3	4.3	9.3	522
13	38.4	38.4	14.8	929
14	37.5	37.5	35.8	584
15	16.5	136.2	8.6	1120
16	56.1	56.1	36.7	681
17	43.6	43.6	18.4	759
18	2.3	15.3	12.6	503
19	198.6	211.5	6.9	3680
20	46.5	77.2	20.1	977



Tree level	Decrease in impurity	Wilk's Λ
1	20946255	0.17
2	1059237	0.36
3	969815	0.24

Figure 1: Table on the left is a numeric hydrological dataset; figure on the top right is the tree deduction process for both CART and SCA with the dataset (note: the highlighted numbers in brackets of the leaf-nodes are the mean response values of those nodes; in this particular case, the two algorithms share the same node splitting rules, however, for most real-world cases, they lead to different decision trees); figure on the middle right illustrates the distinct difference of deduction process between CART and SCA (not related to the case); the bottom-right table is the statistic summaries for CART and SCA of this synthetic case.

Deleted: $X_2(j, z) = \{X_{j,z+1}, X_{j,z+2}, \ldots, X_{j,K}\}$; $j \in 1, 2, 3$.

Deleted: $Y_1(z) = \{Y_1, Y_2, \ldots, Y_z\}$

Deleted: $Y_2(z) = \{Y_{z+1}, Y_{z+2}, \ldots, Y_K\}$. We also define $R_1 = X_1 \cup Y_1$ and $R_2 = X_2 \cup Y_2$ as illustrated in Figure 1.

Deleted: i.e., j

Deleted: instances in

Formatted: Font: Bold, Italic

Deleted: z) that minimize

Deleted: square error

Formatted: Font: Not Bold, Not Italic

Deleted: $\sum_{\substack{i \in 1,2,\ldots,z;\, j \in 1,2,3 \\ x_i \in X_1(j,z)}} \left(y_i - \overline{y}_{Y_1}\right)^2 + \sum_{\substack{i \in z+1,z+2,\ldots,20;\, j \in 1,2,3 \\ x_i \in X_2(j,z)}} ($ (1)¶
where i indicates a particular instance either for R_1 or R_2; x_i indicates a particular instance in subspace $X_1(j, z)$ or $X_2(j, z)$, y_i indicates a particular instance y in subspace Y_1 or Y_2, $\overline{y}_{Y_1}$ and $\overline{y}_{Y_2}$ indicate the mean value of y_i in subspaces Y_1 and Y_2, respectively. Therefore, the split point of the predictor space (i.e., j and z) can be obtained by minimizing the equation (1). To remove ties in the argmax, the best split value for X_j is obtained as the average of $X_{j,z}$ and $X_{j,z+1}$ (Scornet, 2020).¶
Insert Figure 1 here¶
¶

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

After each split, each of the newly generated subspaces can be further split using the same process as long as the number of instances in a subspace is greater than a threshold. This process will be repeated until reaching a stopping criterion, such as a threshold value by which the square errors must be reduced after each split.

The importance score of a particular predictor is measured based on how effective this predictor can reduce the square error in Eq. (1) for the entire tree deduction process (i.e., MDI). In the case of regression, "impurity" reflects the square error of the sample in a subspace (e.g., the larger the square error, the more "impure" the subspace is). The decrease in node impurity (DI) for splitting a particular space $\boldsymbol{s}$ is calculated as:

$$DI(z,j,s)=\sum_{i\in 1,2,\ldots,k}\left(Y(i)-\overline{Y}\right)^2-\frac{z}{k}\cdot\sum_{i\in 1,2,\ldots,z}\left(Y_1(i)-\overline{Y}_1\right)^2-\frac{k-z}{k}\cdot\sum_{i\in z+1,z+2,\ldots,k}\left(Y_2(i)-\overline{Y}_2\right)^2 \qquad (3)$$

where $\boldsymbol{z}$ and $\boldsymbol{j}$ are the coordinates for the optimum splitting point of space $\boldsymbol{s}$, $\boldsymbol{k}$ is the number of instances in space $\boldsymbol{s}$ and $\overline{Y}$ is the mean value of $\boldsymbol{Y(i)}$ in space $\boldsymbol{s}$. Therefore, the Mean Decrease Impurity (MDI) for the variable $\boldsymbol{X_j}$ computed via a decision tree is defined as:

$$MDI\left(X_j\right)=\sum_{s\in S;\, j=j} P_s\cdot DI(z,j,s) \qquad (4)$$

where $\boldsymbol{S}$ is the total spaces in a tree, $\boldsymbol{P_s}$ is the fraction of instances falling into $\boldsymbol{s}$. In other words, the MDI of $\boldsymbol{X_j}$ computes the weighted DI related to the splits using the $\boldsymbol{j_{th}}$ predictor. MDI computed via RF is simply the average of the MDI computed via each tree of the forest. The ensemble (i.e., average) of important scores from the forest is assumed to be more robust than the individual tree.

3. Wilks Feature Importance

WFI is based on the stepwise cluster analysis (SCA) algorithm (Huang, 1992). The fundamental difference between WFI and MDI comes from the split criterion and the tree deduction process. Let us recall the split criterion of CART, in which the optimum split point for $\boldsymbol{X_j}$ is located based on the minimum squared errors of $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$ as shown in Eq (1). In WFI, this function is achieved by comparing the two subspaces' (i.e., $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$) likelihood, which is measured through the Wilks' Λ statistics (Nath and Pavur, 1985; Wilks, 1967). It is defined as $\Lambda = Det(W)/Det(B+W)$, where $Det(W)$ is the determinant of a matrix, W and B are the within- and between-group sums of squares

Deleted: splitted

Deleted: $DI(j,z,s)=\sum_{i\in 1,2,\ldots,k}\left(y_i-\overline{y}_Y\right)^2-\frac{z}{k}\cdot\sum_{i\in 1,2,\ldots;\, x_i\in X_1(}$ (2)

Field Code Changed

Deleted: *j*

Deleted: *z*

Deleted: $\overline{y}_Y$

Deleted: y_i

Deleted: in

Field Code Changed

Deleted: $MDA\left(X_j\right)=\sum_{s\in T;\, j=j} P_s\cdot DI(j,z,s)$ (3)

Field Code Changed

Deleted: *T*

Deleted: ¶
In Eq. (2), DI is reduced as long as the tree level goes down (i.e., from the top to the bottom level of the decision tree (shown in Figure 1)). Such treatment naturally assumes that the predictors considered (for splitting spaces

and cross-product matrices in a standard one-way analysis of variance, respectively. The W and B can be given by:

$$W = \frac{z \cdot (k-z)}{k} (\overline{Y_1} - \overline{Y_2})' \cdot (\overline{Y_1} - \overline{Y_2}) \quad (5)$$

$$B = \sum_{i=1}^{z} \left[Y_1(i) - \overline{Y_1} \right]' \cdot \left[Y_1(i) - \overline{Y_1} \right] + \sum_{i=1}^{k-z} \left[Y_2(i) - \overline{Y_2} \right]' \cdot \left[Y_2(i) - \overline{Y_2} \right] \quad (6)$$

The value of Λ is a measure of how effective $\boldsymbol{X_j}$ can differentiate between $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$. The smaller Λ value representing a larger difference between $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$. The distribution of Λ is approximated by Rao's ***F***-approximation (***R***-statistic), which is defined as:

$$R = \frac{1-\Lambda^{1/S}}{\Lambda^{1/S}} \cdot \frac{Z \cdot S - d \cdot (m-1)/2 + 1}{d \cdot (m-1)} \quad (7)$$

$$Z = k - 1 - (d+m)/2 \quad (8)$$

$$S = \frac{d^2 \cdot (m-1)^2 - 4}{d^2 + (m-1)^2 - 5} \quad (9)$$

where statistic ***R*** is distributed approximately as an ***F***-variate with $n_1 = d \cdot (m-1)$ and $n_2 = d \cdot (m-1)/2 + 1$ degrees of freedom; m is the number of groups. Since the number of groups is two in this study, an exact ***F***-test is possibly performed based on the following Wilks' Λ criterion be:

$$F(d, k-d-1) = \frac{1-\Lambda}{\Lambda} \cdot \frac{k-d-1}{d} \quad (10)$$

Therefore, the two subspaces can be compared for examining significant differences through the ***F***-test. The null hypothesis would be H_0: $\boldsymbol{\mu}(\boldsymbol{Y_1}) = \boldsymbol{\mu}(\boldsymbol{Y_2})$ versus the alternative hypothesis H_1: $\boldsymbol{\mu}(\boldsymbol{Y_1}) \neq \boldsymbol{\mu}(\boldsymbol{Y_2})$, where $\boldsymbol{\mu}(\boldsymbol{Y_1})$ and $\boldsymbol{\mu}(\boldsymbol{Y_2})$ are population means of $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$, respectively. Let the significance level be $\boldsymbol{\alpha}$, the split criterion would be: $F_{cal} \geq F_{\alpha}$ and H_0 are false, which implies that the difference between two subspaces is significant thus they should be split.

The second difference between the CART and SCA algorithms lies in the tree deduction procedure. In CART, the splitting process will be repeated until any newly generated subspace can no longer be split. In SCA, once all the nodes in the current stage have been examined for splitting, merging will be followed in the next stage, as illustrated in Figure 1. The merging process will compare any pairs of nodes based on the value of Wilks' Λ to test if they can be merged (i.e., for $F_{cal} < F_{\alpha}$

and H_0 are true, which indicates that these two subspaces have no significant difference thus should be merged). Such splitting and merging processes are iteratively performed until no node can be further split or merged. Once an SCA tree is built, the WFI for the variable $\boldsymbol{X_j}$ computed via an SCA tree is defined as:

Moved (insertion) [5]

$$WFI\left(X_j\right) = \sum_{s \in S;\, j=j} P_s \cdot \left(1\text{-}\Lambda(z, j, s)\right) \qquad (11)$$

Field Code Changed

where $\boldsymbol{S}$ is the total spaces in a tree, $\boldsymbol{P_s}$ is the fraction of instances falling into $\boldsymbol{s}$, Λ ($\boldsymbol{z, j, s}$) denotes the value of Λ obtained at the optimum splitting point of space $\boldsymbol{s}$ with row and column coordinates $\boldsymbol{z}$ and $\boldsymbol{j}$, respectively. Similar to the calculation of MDI in Eq. (3), the WFI for $\boldsymbol{X_j}$ computes the weighted (1-Λ) value related to the splits using the $\boldsymbol{j_{th}}$ predictor.

Moved (insertion) [6]

According to the law of large numbers, WFI is expected to perform better under the RF framework since the randomized predictors ensure enough tree diversity, leading to more balanced importance scores. Therefore, we name the ensemble of SCA as the stepwise clustered ensemble (SCE). In addition to the three hyperparameters (i.e., ***Ntree***, ***Nmin*** and ***Mtry***) for Breiman's RF, SCE also requires significance level ($\boldsymbol{\alpha}$), which is used for the *F*-test during the node splitting process.

There could be two potential advantages of WFI over MDI. First, the decrease in node impurity (DI) will become smaller and smaller as long as the tree level goes down (as shown in the bottom-right table in Figure 1). Such a mechanism naturally assumes that the predictors considered (for node splitting) in lower levels of the tree are less significant than those in upper levels. This effect is even aggravated by the existence of predictor dependence, which will depress the importance scores of independent predictors and increase the positively dependent ones (Scornet, 2020). As a consequence, some critical predictors may only receive small importance scores. In comparison, Wilk's Λ is a measure of the separateness of two subspaces, which could avoid the above-mentioned issue for MDI because values of (1-Λ) do not necessarily decline as long as the tree level goes down (as shown in the bottom-right table in Figure 1). Therefore, the predictors that are primarily considered in latter splits still possible to own higher importance scores than those in early splits. As a consequence, some critical predictors might be identified by WFI but overlooked by MDI. Second, the node splitting mechanism of WFI is based on *F*-test, which, therefore, may significantly reduce the probabilities that the two child-nodes are split due to chance. Such a

Deleted: as

Deleted: also

Deleted: As a consequence, predictors considered in lower levels of the tree will only receive small importance scores and may be neglected by decision-makers. Therefore, the importance scores obtained from CART is biased towards the predictors considered in early-cut spaces for obtaining a highest square error reduction. This will mislead the hydrological inference since some predictors that are considered as less significant in CART (with a limited contribution in reducing the square error) could be vital for explaining some concerned hydrological behaviors such as streamflow peaks.

mechanism could be helpful to build more robust input-output relationships for prediction and inference by reducing overfitting. The above-mentioned potential advantages of WFI will be tested with a large number of hydrological simulations in the following two sections.

4. Comparative studies over the NCAR CAMELS dataset

4.1. Dataset description

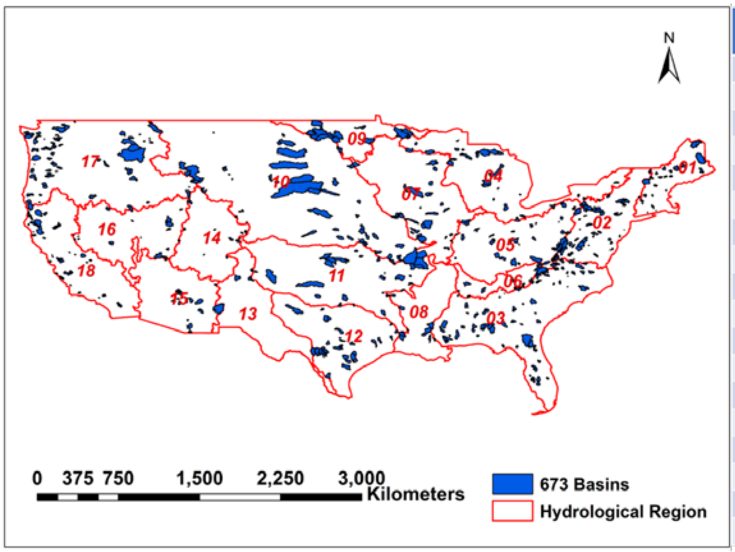
Catchment Attributes and Meteorological (CAMELS) dataset (version 1.2) (Addor et al., 2017; Newman et al., 2015) was used to evaluate the WFI performance. The dataset contains daily forcing and hydrologic response data for 673 basins across the contiguous United States that spans a very wide range of hydroclimatic conditions (Figure 2) (Newman et al., 2015). These basins range in size between 4 and 25,000 km^2 (with a median basin size of 336 km^2) and have relatively low anthropogenic impacts (Kratzert et al., 2019b).

In attempting to demonstrate the relative importance of meteorological data and large-scale climatic indices on streamflow, we used monthly mean values of meteorological data in CAMELS dataset and 4 commonly used large-scale climatic indices (including Nino3.4 (Trenberth, 1997), Pacific decadal oscillation (PDO) (Mantua et al., 1997), interdecadal Pacific oscillation (IPO) (Mantua et al., 1997) and Pacific North American index (PNA) (Leathers et al., 1991)) to simulate the monthly streamflows. To refect the initial catchment conditions and lagged impact of climatic indices, the 2-month moving average meteorological data and climatic indices of the preceding 2 months were incorporated as model predictors. Therefore, the input-output structure (with 22 predictors) for each of these basins can be written as follows:

$$Q_t = f\begin{pmatrix} Pr_t, Rad_t, Tmax_t, Tmin_t, Vp_t, (Pr_t + Pr_{t-1})/2, (Rad_t + Rad_{t-1})/2, \\ (Tmax_t + Tmax_{t-1})/2, (Tmin_t + Tmin_{t-1})/2, (Vp_t + Vp_{t-1})/2, \\ Nino3.4_t, Nino3.4_{t-1}, Nino3.4_{t-2}, PDO_t, PDO_{t-1}, PDO_{t-2}, \\ IPO_t, IPO_{t-1}, IPO_{t-2}, PNA_t, PNA_{t-1}, PNA_{t-2} \end{pmatrix} \qquad (12)$$

Field Code Changed

where Q_t represents streamflow of month ***t***. *Pr*, *Rad*, *Tmax*, *Tmin* and *Vp* represent monthly values of precipitation, short-wave radiation, maximum temperature, minimum temperature and vapor pressure, respectively.



Hydrological Region	Number of basins
1	27
2	75
3	92
4	31
5	45
6	17
7	33
8	12
9	9
10	72
11	31
12	37
13	7
14	17
15	19
16	18
17	91
18	40

Figure 2. Overview of the basin location and corresponding hydrological region. This map was created using ArcGIS software (Esri Inc. 2020).

4.2. Evaluation procedures and metrics

The model training was performed based on January 1980 to December 2005, while the testing was done based on the period of January 2006 to December 2014. The hyperparameters for both RF and SCE were set as follows: ***Ntree*** was set as 100, ***Nmin*** was set as 5, and ***Mtry*** was set as 0.5 as suggested by Barandiaran (1998), indicating half of the predictors were selected in each tree. In addition, the significance level (***α***) was set as 0.05 for the *F*-test in SCE.

Moved (insertion) [7]

The performance of WFI will be evaluated and compared against PFI (applied to RF and SCE) and MDI (applied to RF). To improve the stability of the PFI results, previous studies have suggested repeating and average the PFI over repetitions (Molnar, 2020). In this study, the PFI process was repeated 10 times and then averaged for stabilizing the results. To facilitate the comparisons among different variable rankings, importance scores from the three feature importance methods were scaled into the [0,1] range. All the feature importance methods will be evaluated through recursive feature elimination (RFE) (Guyon et al., 2002) as follows: (1) train

Moved (insertion) [8]

SCE and RF models with all predictors; (2) calculate the importance scores using the three interpretation methods embedded in their corresponding models; (3) exclude three least relevant predictors for each set of the importance scores obtained in step 2; (4) retrain the models using the remaining predictors in step 3; (5) repeat step 2 to 4 until the number of predictors less or equals to a threshold (set to 4 in this case study). To directly compare the quality of variable rankings from different feature importance measures, the selected predictors from WFI (after every RFE iteration) were also used to train RF. This procedure allows the effects of different variable rankings to be solely from feature importance methods (i.e., removed the effects from different node splitting algorithms). The same procedure was also performed for SCE-based PFI (i.e., SCE-PFI) to examine whether the differences in variable rankings are from the WFI method or the tree deduction process in SCE.

Two error metrics (i.e., adjusted R^2 and RMSE) were used to evaluate the model performance. Adjusted R^2 has been used instead of R^2 because adjusted R^2 can consider the number of predictors. Adjusted R^2 is defined as:

$$adj\ R^2 = 1 - \frac{(1-R^2)(N-1)}{N-P-1} \qquad (13)$$

where $\boldsymbol{P}$ is the number of predictors and $\boldsymbol{N}$ is the number of instances.

RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{n=1}^{N}\left(y_n - y_n^*\right)^2} \qquad (14)$$

where y_n and y_n^* are the $\boldsymbol{n^{th}}$ observed and predicted streamflow values, respectively.

To evaluate the stability of a feature importance method, we consider reducing predictors during the RFE iterations as a form of perturbation in the dataset. Suppose the obtained importance score for a dominant predictor indicates an irregular changing pattern during the RFE iterations. In that case, the method thus is not stable because it can lead to many versions of inferences for such a predictor. On the other hand, if such a changing pattern is predictable (e.g., monotonically

Moved (insertion) [9]

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

increasing trend), a stable inference can be achieved among interactions because the predictable pattern can help analyze how a predictor reacts to the change in the dataset. In this study, the monotonicity is examined by using the Spearman's rank correlation coefficient (i.e., Spearman's ρ), which is commonly used to test the statistical dependence between the rankings of two variables and is defined as:

$$\rho = \frac{\sum_i \left(RX_i - \overline{RX}\right)\left(RY_i - \overline{RY}\right)}{\sqrt{\sum_i \left(RX_i - \overline{RX}\right)^2 \left(RY_i - \overline{RY}\right)^2}} \quad (15)$$

where $\boldsymbol{RX_i}$ is the ranks of variables $\boldsymbol{X}$ for the $\boldsymbol{i}_{th}$ RFE iteration and $\boldsymbol{RY_i}$ is the number of selected predictors for the $\boldsymbol{i}_{th}$ RFE iteration; $\overline{RX}$ and $\overline{RY}$ are the means of $\boldsymbol{RX_i}$ and $\boldsymbol{RY_i}$, respectively. A larger Spearman's $\boldsymbol{\rho}$ indicates the importance score for a predictor will increase along with the reduction of irrelevant predictors, leading to stable importance scores.

4.3. Predictive Accuracy and Interpretation Stability

Figure 3 shows the model testing performances (adjusted R^2) for 18 hydrological regions with all 22 predictors. The results show that SCE and RF significantly outperform SCA and CART, respectively. When taking a close look at these two pairs of model performance, SCA and CART are close to each other, while SCE outperforms RF in most hydrological regions (except the 9th region).

The pairwise comparisons of these four algorithms over 673 basins show a high coefficient of determination (0.913) of adjusted R^2 between SCE and RF, and an even higher coefficient of determination (0.965) between SCE and RF (Figure 4). This result indicates that, in general, it is not likely to have a distinct performance gap for a particular simulation task either between SCE and RF, or between SCE and RF. Therefore, SCE can be a good substitute for RF.

Moved (insertion) [10]

Field Code Changed

Moved (insertion) [11]

Deleted: 3 Methodology¶
3.1. Wilks Feature Importance¶
For an unbiased estimation of the importance scores for decision trees, the WFI is developed and illustrated using the same dataset as CART in Figure 1. The fundamental difference between WFI and MDI comes from the split criterion and procedures used for the tree deduction process. We will talk about the split criterion first: Recalling the procedure of CART, any possible splits of $\boldsymbol{X_j}$ are examined to find the optimum split point that can minimize the square errors of $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$ as shown in Eq (1). In WFI, the function for finding the optimum split point is achieved by comparing the two subspaces' likelihood ratio, which is measured through the Wilks' Λ statistics. The optimum value of Λ can be used to measure how effective the $\boldsymbol{X_j}$ can differentiate $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$. ¶
¶
The calculation process of WFI employs the tree deduction processes of stepwise cluster analysis (SCA) (Huang, 1992). In SCA, the Wilks' Λ statistics (Wilks, 1967;Nath and Pavur, 1985) is used as the criterion for node splitting, and ...

Moved up [1]: The distribution of Λ is approximated by Rao's $\boldsymbol{F}$-approximation ($\boldsymbol{R}$-statistic), which is defined as:¶

Deleted: 6)¶
$Z = k - 1 - (d + m)/2$ (7)¶ ...

Moved up [2]: Since the number of groups is two in this study, an exact $\boldsymbol{F}$-test is possibly performed based on the following Wilks' Λ criterion be:¶

Deleted: 9)¶
¶
Therefore, the sample means of

Moved up [3]: the two subspaces can be compared for examining significant differences through the $\boldsymbol{F}$-test. The null hypothesis would be H_0: $\boldsymbol{\mu}(\boldsymbol{Y_1}) = \boldsymbol{\mu}(\boldsymbol{Y_2})$ versus the

Moved up [5]: Such splitting and merging processes are iteratively performed until no node can be further split or

Deleted: Let the significance level be $\boldsymbol{\alpha}$ (which is set as 0.05 in this study), the split criterion would be: $F_{cal} \geq F_\alpha$ ar ...

Moved up [4]: to test if they can be merged (i.e., for $F_{cal} < F_\alpha$ and H_0 are true, which indicates that these two subspaces have no significant difference thus should be merged). Such

Deleted: $WFI\left(X_j\right) = \sum_{\substack{s \in T \\ j=j}} P_s \cdot \left(1\text{-}\Lambda(j,z,s)\right)$ ...

Moved up [6]: Similar to the calculation of MDI in Eq. (3), the WFI for $\boldsymbol{X_j}$ computes the weighted (1-Λ) value related to the splits using the $\boldsymbol{j_{th}}$ predictor. ¶

Deleted: The major advantage of WFI over MDI is that every spliting and merging action by WFI is evaluated based on Wilk's test-statistics with the significance level $\boldsymbol{\alpha}$ set ...

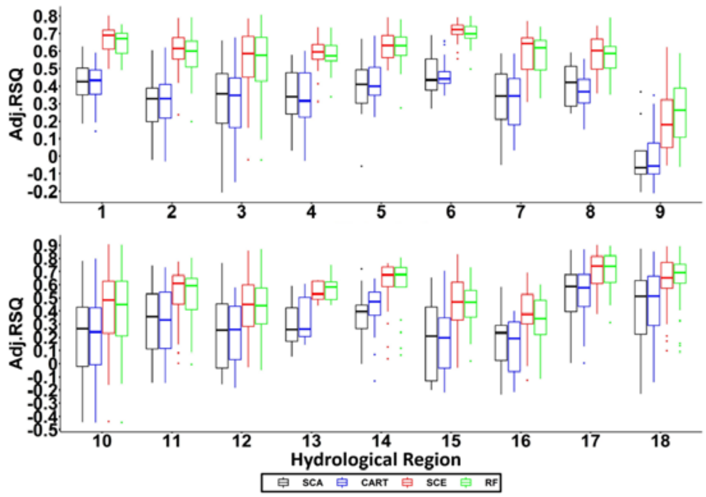


Figure 3. Adjusted R^2 for 18 hydrological regions. Each box indicates statistical summaries (i.e., the bars represent median value; the lower and upper boundaries of a box represent 1st and 3rd quantiles, respectively; dots represent outliers) of adjusted R^2 for all the basins in a particular hydrological region.

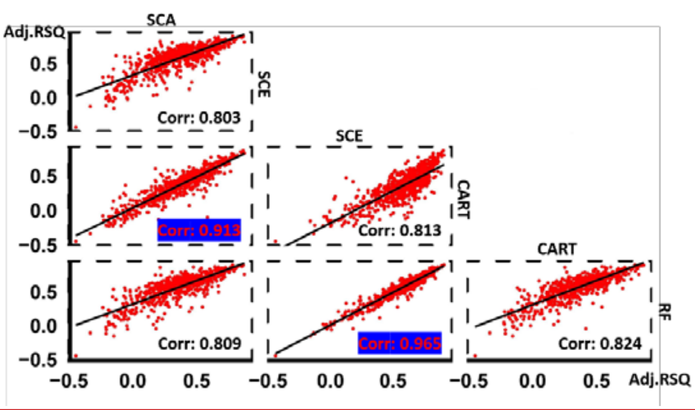


Figure 4. Pairwise comparison for adjusted R^2 over 673 US basins.

The left column in Figure 5 shows simulation performances based on RFE iterations for three feature importance measures embedded in SCE and RF. In general, both models can improve their simulation performance by eliminating irrelevant predictors. When the number of predictors reduces to 7 (i.e., at 5th iteration), both models reach their highest predictive accuracy over the OOB and testing dataset. This result indicates that it is plausible to use the OOB dataset to identify the optimum subset of predictors. Comparing the simulation performance for the training period, the simulation performance for SCE is much lower than it for RF, while an opposite result is observed for the testing period. This result highlights the issue of overfitting for RF. One exceptional that RF outperforms SCE (for the testing period) happens to the last (i.e., 6th) iteration, where RF with MDI selected predictors outperforms SCE with WFI selected ones. We can assume that RF may have a better chance to outperform SCE with insufficient predictors. Nevertheless, SCE owns the overall best performance with PFI-selected predictors.

The upper left panel in Figure 6 shows that from 0th to 5th iterations, over 55% to 60% of basins (as indicated in yellow diamonds) simulated by SCE with WFI selected predictors outperforms those simulated by RF with MDI selected ones. In comparison, the number drops to about 40% at the 6th iteration. This result agrees with the results for Figure 5. The lower left panel in Figure 6 shows that from 1st to 5th iterations, there is a higher chance that SCE with PFI selected predictors outperforms RF with MDI selected ones for over 75% of the hydrological regions (as we can see, the black boxes are above the blue line).

To further investigate the solo effect of variable rankings, the WFI and SCE-PFI selected predictors in each RFE iterations were used for RF simulations. The results are shown in the right column in Figure 5. The RF simulations with WFI selected predictors owned the highest predictive accuracy in most RFE iterations over the training, OOB validation and testing datasets. In particular, the WFI selected predictors have shown significant strength in the last two iterations and facilitated RF to improve its predictive accuracy. It is worth mentioning that even though SCE-PFI selected predictors allowed SCE to achieve its optimum performance, they did not deliver optimum performance for RF. This result shows WFI selected predictors provide a better universal solution than the PFI-selected ones.

The upper left panel in Figure 7 shows that a majority of basins simulated by RF with WFI selected predictors outperform those simulated by RF with MDI selected ones. In particular, at the 6th

iteration, basins in 16 (out of 18) hydrological regions may probably own better performance with WFI selected predictors than the MDI-selected ones. In addition, as the number of predictors decreases, there are increasing chances that WFI selected predictors could generate higher performance than the MDI-selected ones. Based on a two-sided Mann-Kendall (M-K) trend test (Kendall, 1948; Mann, 1945), such increasing trend is significant with the $\boldsymbol{Z}$ score equals 2.63 and $\boldsymbol{p}$-value smaller than 0.01. Another significant increasing trend (with the $\boldsymbol{Z}$ score equals 1.88 and $\boldsymbol{p}$-value equals 0.06) also can be observed for the paired studies of WFI and RF-PFI. In contrast, no significant increasing trend can be observed for the pairs of SCE-PFI and MDI, as well as SCE-PFI and RF-PFI. This finding indicates WFI could generate robust variable rankings, based on which informative predictors are more likely to be kept for optimum simulation performance. In contrast, other feature importance measures may lose critical predictors during the RFE process.

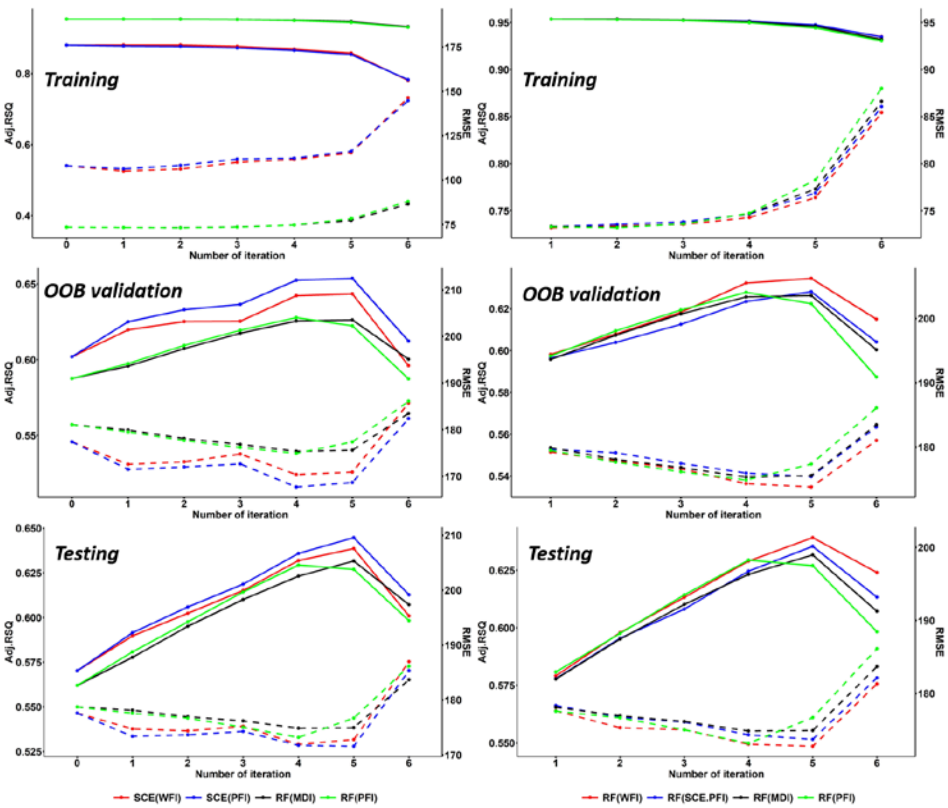


Figure 5. Iterative change in accuracy in mean values of 673 US basins. The solid lines indicate adjusted R^2, while the dashed lines represent RMSE. Figures on the left column show SCE and RF performances based on variables selected by themselves, while figures on the right show RF model performances based on variables selected by SCE and RF. The models with an iteration number of 0 represent the model with all 22 predictors.

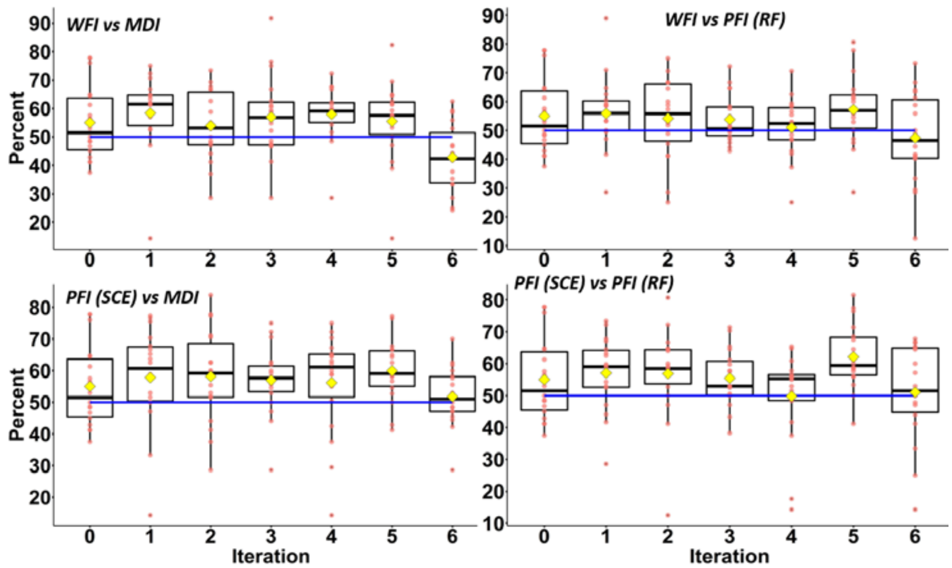


Figure 6. Pairwise comparisons of model performance with different feature importance measures. Each of these red points represents the percentage of basins simulated by model A outperform model B (based on adjusted R^2), in one particular hydrological region. The blue line represents 50% percent, and the yellow square represents the mean percentage of 18 hydrological regions.

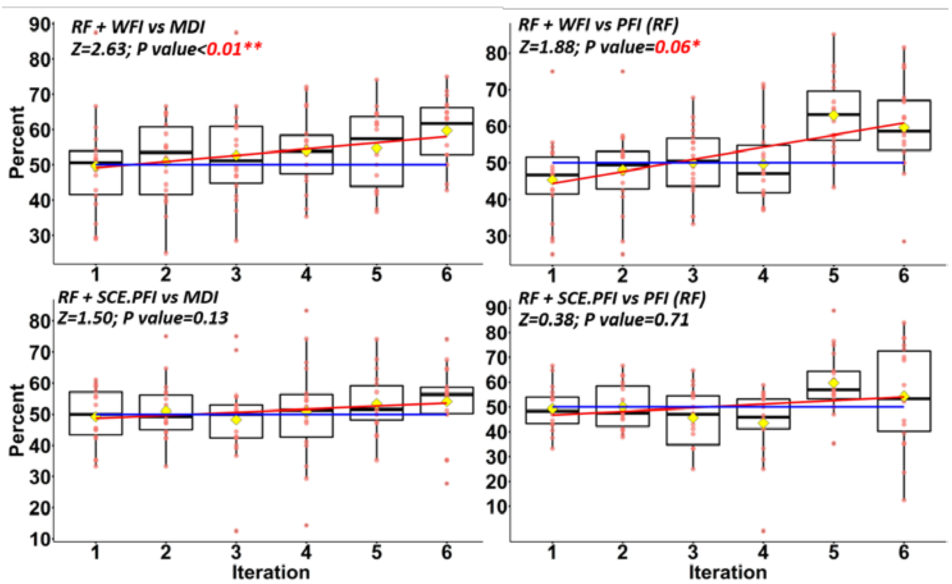


Figure 7. Pairwise comparisons of RF model performance with different feature importance measures. ***Z*** scores and ***P*** values are calculated based on the two-sided Mann-Kendall trend test. If the ***Z*** score greater than ***1.96***, an increasing trend can be assumed with a significance level of ***0.05***. If the ***Z*** score greater than ***1.645***, an increasing trend can be assumed with a significance level of ***0.1***. Other notations are the same as those in Figure 6.

Figure 8 shows the summaries of selected predictors (in the last iteration) with different feature importance measures. *Pr* (monthly precipitation at time step *t*) and *Pr2* (mean values for monthly precipitation at time step *t* and *t-1*) are considered the two most important predictors for the SCE algorithm with WFI selected predictors. In contrast, MDI considers *Tmax2* (mean values for the monthly maximum temperature at time step *t* and *t-1*) as the most important predictor for monthly streamflow simulation. It is acknowledged that streamflow is more responsive to precipitation than air temperature. Therefore, we can assume that RF may capture more acuate responses of streamflow with WFI selected features than MDI or PFI selected ones. This assumption could be one of the reasons that RF with WFI selected predictors outperforms the others. It should be noted that IPO is considered as an important predictor for 56 out of 673 basins with WFI, while this predictor has only employed in 21, 5 and 10 basins with SCE-PFI, MDI and RF-PFI methods, respectively.

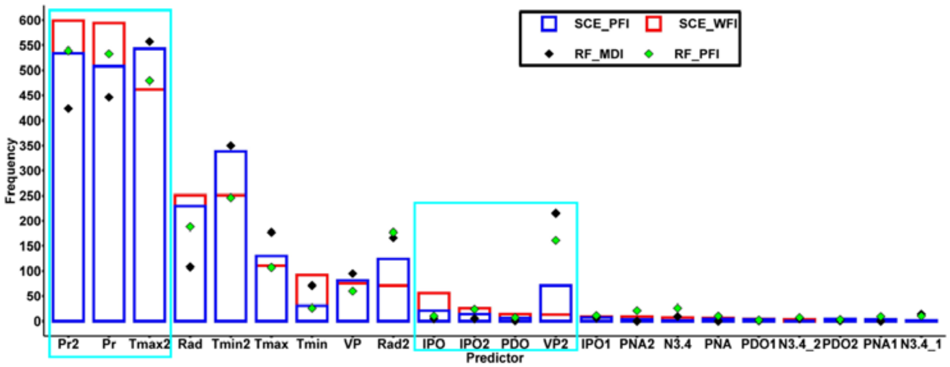


Figure 8. Summaries of the predictors used in last iterations for 673 US basins.

The Spearman's Rho (ρ) values for the predictor with the highest importance score at the last RFE iteration illustrate the stability of all three interpretation methods embedded in SCE and RF (Figure 9). The results indicate the importance score for the predominant predictor increase in response to the reduction of irrelevant predictors. Compared with other feature importance measures, WFI owns the highest ρ values in general with the *p*-value less than 0.01, indicating a significant correlation between the importance score and the reduction of irrelevant features. In comparison, eliminating irrelevant predictors will significantly influence the importance score of predominant predictors obtained by PFI and MDI. This fact challenges the application of the PFI and MDI since the removal of irrelevant predictors cannot guarantee the same or similar level of hydrological inference because the importance score may vary distinctly according to the reduction of irrelevant predictors. In contrast, the WFI method provides more stable importance scores and will lead to more consistent hydrological inferences.

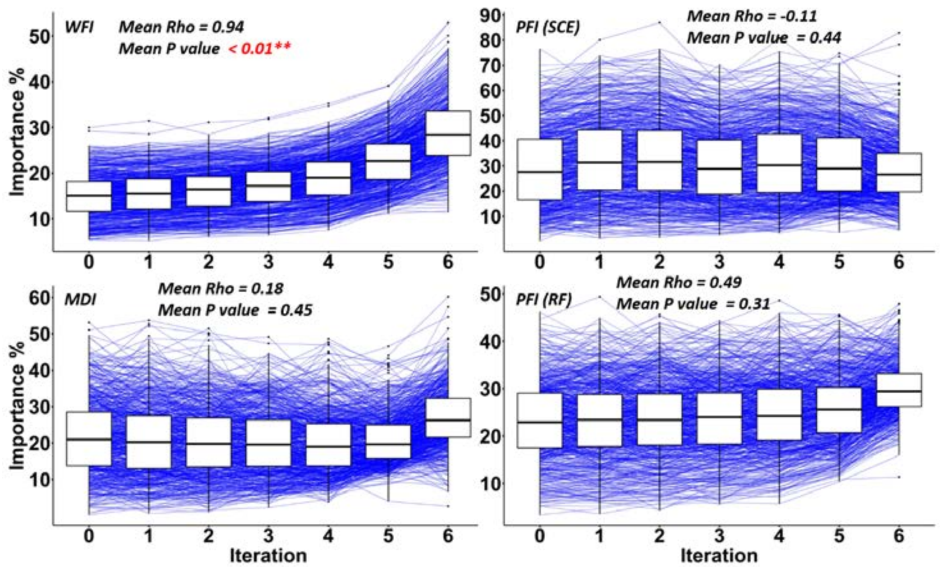


Figure 9. Mean Spearman's ρ values for the most important features. The mean p-value means how likely it is that the observed correlation is due to chance. Small *p*-values indicate strong evidence for the observed correlations.

5. Application of WFI over irrigated watersheds in the Yellow River Basin, China

5.1. Study Area and Data

Daily streamflow simulations for three irrigated watersheds located in the alluvial plain of the Yellow River in China were conducted to test the capability of the proposed WFI method at a finer temporal resolution. These watersheds share a total area of 4,905 km^2, consisting of 52% irrigated land, 17% residential area, 15% desert, 12% forested land, and 4% water surface (Figure 10). The landscape of the study area is characterized by an extremely flat surface with an average slope ranging from 1:4000 to 1:8000, with mostly highly permeable soil (sandy loam). The climatic condition of the study area is characterized by extreme arid environments with annual precipitation ranging from 180 to 200 mm, and annual potential evaporation ranging from 1,100 to 1,600 mm (Yang et al., 2015).

Deleted: 4

Deleted: Three

Deleted: selected

Deleted: (Figure 2).

Deleted: .

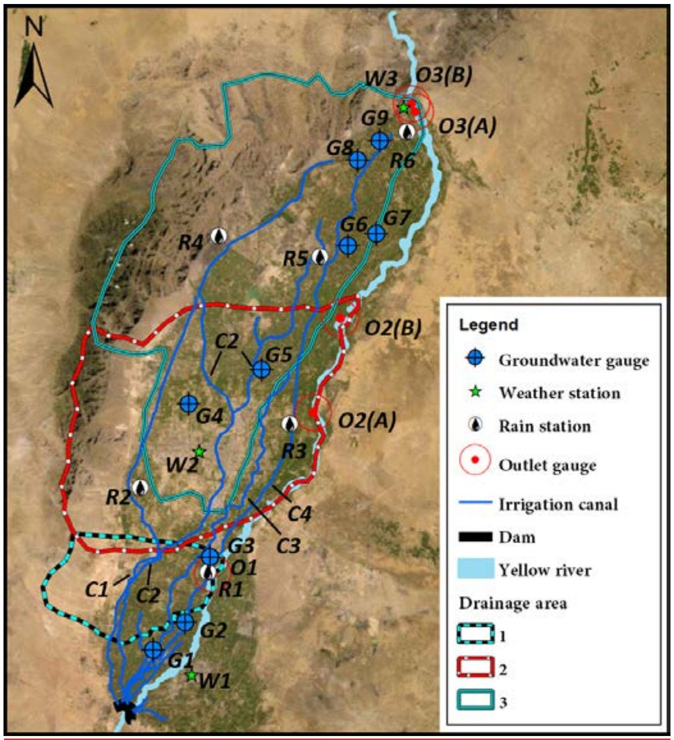


Figure 10: Map of the study area. Note: due to the extremely flat surface, three interconnected irrigated watersheds are approximately delineated. In this map, ***G*** indicates groundwater gauges, ***W*** indicates weather stations, ***R*** indicates rain stations, ***C*** indicates irrigation canals and ***O*** indicates drainage outlets. Both 2nd and 3rd irrigated watersheds contain two crisscrossed drainages with strong hydrological connections. The map was created using ArcGIS software (Esri Inc. 2020).

Initial catchment conditions were also considered in this case study to improve the model performance. Specifically, moving sums of daily precipitation, temperature and evaporation time series over multiple time periods $\delta_{P,T,E}$ = [1, 3, 5] prior to the date of predictions were set as predictors to reflect the antecedent watershed conditions. Similarly, the moving window for daily irrigation time series δ_I = [1, 3, 5, 7, 15, 30]. In addition, daily groundwater level data are used as additional predictors to reflect the baseflow conditions of the catchments. The daily time-series data were divided into two subsets; one from 2001/01/01 to 2011/12/31 for model training and

Moved (insertion) [12]

Moved (insertion) [13]

Formatted: Font: Bold, Italic

Formatted: Centered, Line spacing: 1.5 lines

Deleted: ***Insert Figure 2 here***¶
¶

Deleted: timeseries

Deleted: Similary

Deleted: timeseries

Deleted: timeseries

Deleted: , including

OOB validation and the other from 2012/01/01 to 2015/12/31 for model testing. Table 1 list the weather, rain and groundwater stations used for each basin. The streamflow processes show distinct behaviors in terms of flow magnitude and duration due to the different irrigation schedules in spring and winter. To analyze such temporal variations, daily streamflow for spring-summer (April to September) and autumn-winter (October to March) were examined separately. In this case study, the same hyperparameters for RF and SCE are used as in Section 4.

Table 1: Weather, rain and groundwater gauges, and irrigation canals used in each irrigation basin.

Watershed ID	Stations/canals	outlets
1	*C1, C2, C3, W1, R1, G1, G2, G3*	*O1*
2	*C1, C2, C3, C4, W2, R2, R3, R5, G4, G5*	*O2(A)+ O2(B)*
3	*C1, C2, C4, W2, W3, R4, R5, R6, G4, G5, G6, G7 G8, G9*	*O3(A)+ O3(B)*

Note: Streamflow for each watershed is integrated as the sum of the gauged streamflows within this area.

5.2. Results Analysis

Generally, SCE and RF delivered reasonable predictive accuracy (using all considered predictors) across all watersheds and seasons (Table 2). The SCE approaches the best overall predictive accuracy for the testing dataset. Compared with RF, the SCE has a smaller drop in predictive accuracy from the training to testing period, indicating the SCE algorithm captured a more robust input-output relationship during the training period. This result agrees with those for the large-scale dataset in Section 4. The convergence tests for training, OOB validation, and testing datasets were shown in Figures S1, S2 and 11, respectively. The results from the testing period (Figure 11) show that SCE always outperforms RF as the number of trees increases.

Table 2: The adjusted R^2 for SCE and RF with all considered predictors.

Basin	Season	Training		OOB		Testing	
		SCE	RF	SCE	RF	SCE	RF
1st	spring	0.94	0.98	0.87	0.88	**0.82**	0.81
1st	winter	0.98	0.99	0.94	0.95	**0.91**	0.90
2nd	spring	0.94	0.98	0.86	0.89	**0.77**	0.76
2nd	winter	0.98	0.99	0.95	0.96	**0.66**	0.65
3rd	spring	0.94	0.98	0.85	0.88	**0.69**	0.68
3rd	winter	0.98	0.99	0.95	0.95	**0.83**	0.82

- **Deleted:** training and prediction, respectively.
- **Deleted:** Owing to the different irrigation schedules in spring and winter, the
- **Deleted:** .
- **Deleted:** the hydrological processes for Spring-Summer
- **Deleted:** Autumn-Winter
- **Formatted:** English (United States)
- **Moved up [7]:** as suggested by Barandiaran (1998), indicating half of the predictors
- **Deleted:** *Insert*
- **Deleted:** the number/ratio of predictors in a subspace (*Mtry*). In this study, we set the *α* value for 0.05 as suggested by Huang (1992). The *Ntree* was set as 200, after which no further improvement in model validation accuracy can be achieved. The *Nmin* was set as 5 to ensure the rare events can be identified. The *Mtry* was set as 50%
- **Formatted:** Left, Line spacing: Multiple 1.08 li
- **Formatted:** Font: Not Italic
- **Deleted:** *here*¶ ¶
- **Formatted:** English (Canada)
- **Deleted:** are selected
- **Deleted:** SCA tree. Similar to the RF, cross-validation is
- **Deleted:** 4.2. WFI Evaluation ¶
- **Moved up [8]:** methods will be evaluated through
- **Deleted:** three models (i.e., SCE, RF and XGB) with all
- **Moved up [9]:**)¶
- **Deleted:** By evaluating the RMSE and adjusted R^2 after
- **Moved up [10]:** In this study, the monotonicity is
- **Deleted:** , and is defined as:¶
- **Moved up [11]:** where RX_i is the ranks of variables X for
- **Deleted:** which therefore will lead to relatively more robu
- **Deleted:** 5.1. Interpretation Accuracy and Robustness¶
- **Deleted:** all three algorithms
- **Deleted:** by
- **Deleted:** irrigated
- **Formatted:** English (United States)
- **Deleted:** approached
- **Deleted:** When compared
- **Deleted:** /validation
- **Deleted:** , which indicates
- **Deleted:** can better address overfitting. ¶
- **Deleted:** dataset (Figure 3).

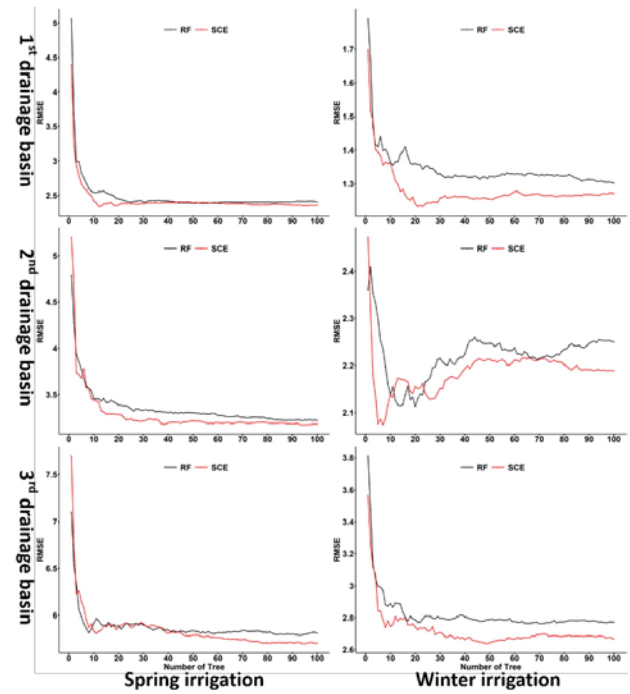


Figure 11: Convergence of the SCE and RF model based on RMSE over the testing period.

The iterative reductions in accuracy for training, OOB validation and test datasets are listed in Figure S3, S4 and S5, respectively. The summary (Figure 12) shows that WFI owns the smallest reduction in accuracy (for both adjusted R^2 and RMSE) over the testing period, followed by SCE-PFI, MDI and RF-PFI. A smaller reduction in accuracy means the selected predictors are more informative in describing the complex relationships of hydrological processes. As a consequence, WFI can identify the most informative predictors compared with other methods. Figure 12 also shows that over the training period, RF receives a much smaller impact from RFE in terms of adjusted R^2 compared with SCE, which is because the least-square fittings employed in the CART training process pursue the highest R^2 over the training period.

Deleted: S1, S2

Deleted: Figure 4

Deleted: Surprisingly, we found that the predictors selected by SCE-WFI lead to even higher predictive accuracy on the testing dataset than the SCE-PFI selected ones. The possible reason is that the PFI method can only consider the effect of one predictor at a time, thus the interactions (i.e., quadratic terms) between the considered predictor and the rest predictors are overlooked. The WFI method on the other hand naturally considers all the interactions among predictors in the tree deduction process, thus the importance scores are considered to be more comprehensive than those generated by the PFI method. Similar evidence can also be found between RF-Purity (i.e., MDI method) and RF-PFI methods: the average reduction in accuracy for RF-Purity is less than that for RF-PFI under the testing dataset. ¶
Insert

Formatted: Font: Not Bold, Not Italic

Deleted: ***3 here***¶
Insert Figure 4 here¶
¶
Comparative studies among the three interpretation methods illustrate overfitting can greatly affect predictive and interpretation accuracy. For instance, RF-PFI

Deleted: lowest average

Deleted: among all models on the training dataset, while its value becomes the highest on

Deleted: dataset.

Formatted: Not Superscript/ Subscript

Deleted: retained

Deleted: Based on this, SCE-

Deleted: provide

Deleted: among all considered models because it suffers the least

Deleted: overfitting. The results also indicate that the XGB algorithm suffers less from overfitting compared with RF. Nevertheless, the predictive accuracy for the XGB algorithm is not as good as it for RF or SCE (Table 2).

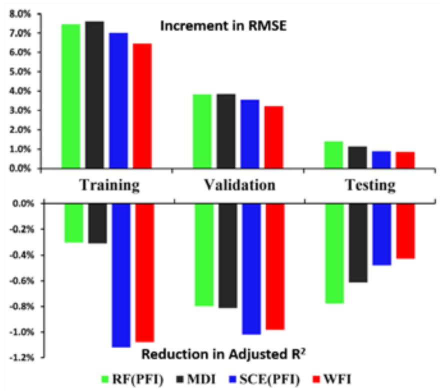


Figure 12: Change in predictive accuracy averaged across three watersheds and two seasons. Note: the change in predictive accuracy for a particular case is calculated as the accuracy for the last iteration minus it for the full predictors.

Figure 13 shows the Spearman's ρ values of the most relevant predictor (i.e., with the highest importance score in the last RFE iteration). The result indicates that WFI owns the highest absolute ρ values for the majority of the cases. This result agrees with those demonstrated in section 4. In fact, the highest absolute Spearman's ρ values for the rest of the relevant predictors (selected for the last RFE iteration) mainly belong to the WFI method (as shown in Figure 14), which further illustrates that WFI could provide stable relative importance among essential predictors for hydrological inference.

Moved (insertion) [14]

Formatted: Font: Times New Roman, 12 pt

Formatted: Justified, Automatically adjust right indent when grid is defined, Line spacing: 1.5 lines, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Formatted: Font: Not Bold, English (Canada)

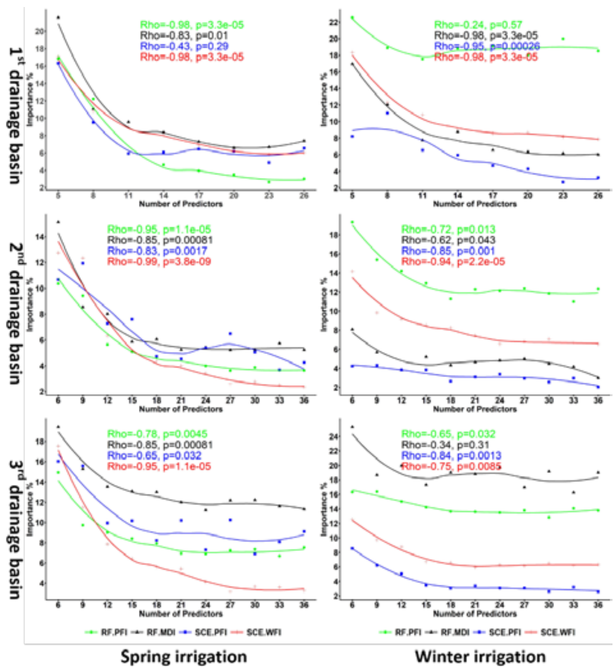


Figure 13: Spearman's ρ values for the most important predictor. Note: the most important predictor is the predictor with the highest importance score in the last RFE iteration. The *p*-value means how likely it is that the observed correlation is due to chance. Small p-values indicate strong evidence for the observed correlations.

Moved (insertion) [15]

Formatted: Font: Italic

Moved (insertion) [16]

Deleted: ***Insert Table 2 here¶***
¶
The Spearman's ρ values for the most relevant predictors (i.e., selected by the last iteration of RFE) across all drainage basins and seasons illustrate the robustness of all three interpretation methods embedded in different models. The results indicate the relative importance of a particular predictor increase in response to the reduction of irrelevant predictors (Figure 5 and Figure S3-S6). Compared with other interpretation methods and ML algorithms, the SCE-WFI owns the highest absolute Spearman's ρ values for the majority of the cases (Figure 6). This indicates the reduction of irrelevant predictors would greatly influence the importance scores obtained by PFI and MDI. This challenges the application of the PFI and MDI since the removal of irrelevant predictors cannot guarantee the same or similar level of hydrological inference (i.e., the relative importance scores may vary distinctly according to the reduction of irrelevant predictors). In contrast, the WFI method provides more stable relative importance scores and will lead to more consistent hydrological inferences.¶
Insert Figure 5 here¶
Insert Figure 6 here¶
5.2. Insights Toward Hydrological Processes¶
To explore the relationships between the hydrological responses and their driving forces, the

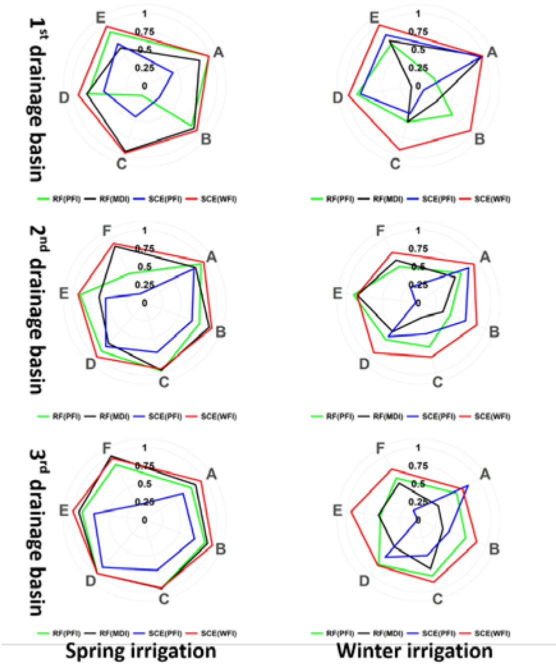


Figure 14: Spearman's ρ values for three watersheds and seasons. Note: the RFE process of this case study keeps at least five and up to seven of the most relevant predictors in the last iteration, according to the remainder of the total considered predictors divided by three. Capital letters from A to F represent the most relevant predictors identified by different feature importance methods.

The importance scores were aggregated and analyzed according to different types (i.e., precipitation, irrigation, evaporation, etc.) to explore the relationships between the hydrological responses and their driving forces. We chose the models with the smallest RMSE (among all the RFE iterations) on the testing dataset for the hydrological inference. The results indicate the importance scores differed significantly according to the algorithms and interpretation methods used (Figure 15). In particular, the aggregated predictor P1 (i.e., daily precipitation for timestep ***t*** from all spatial locations) owns positive contributions (in reducing the RMSE) for WFI in the Spring irrigations. At the same time, it has merely no contribution for other feature importance methods. To investigate whether the predictors identified by WFI are also meaningful to other

Formatted: Automatically adjust right indent when grid is defined, Space After: 8 pt, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Deleted: .).

Deleted: to investigate

Deleted: relationships between importance scores and

Deleted: processes

Deleted: 7

Deleted: of the current

Deleted: SCE-

Deleted: , while

Deleted: ML algorithms and interpretation algorithms, we reinserted the predictors in P1 into the best RF model (in which the set of predictors reaches the smallest RMSE over the testing dataset). Indeed, we found the RF with reinserted predictors showing slightly improved predictive accuracy (i.e., RMSE and adjusted R^2) for Spring irrigations across all watersheds on the testing dataset (Table 3). This result illustrates that even though the predictors in P1 have no contribution in improving the predictive accuracy on the training dataset, it can potentially distinguish different hydrological behavior (i.e., with a small Wilk's Λ value) and lead to improved model performance on the testing dataset. In fact, the time of concentration for these basins is usually less than one day if the storm falls near the outlets of the irrigation basins. This fact proves the above hydrological inference is reasonable.

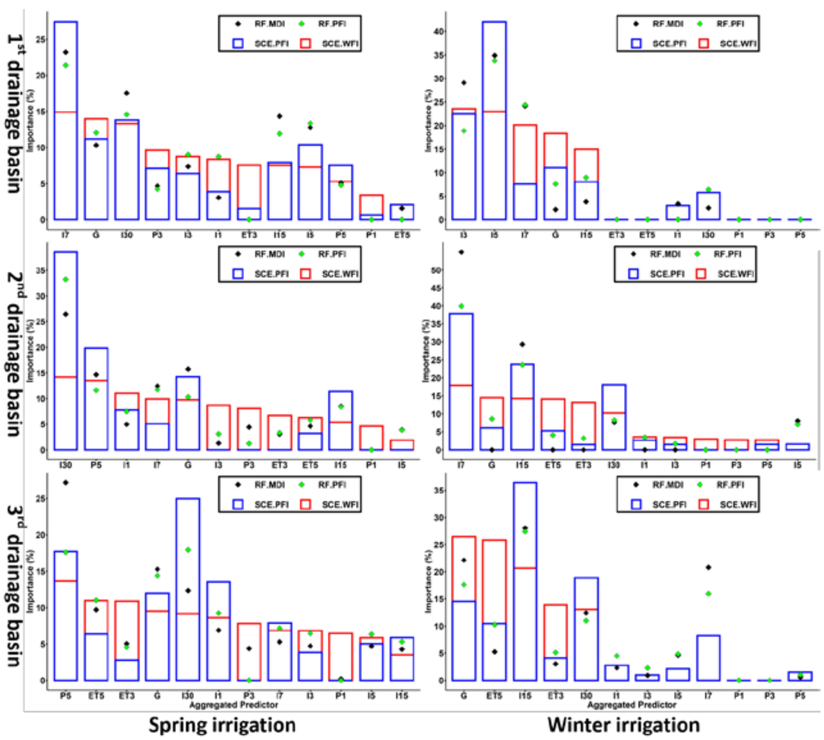


Figure 15: Importance scores aggregated by predictor types. Note: each type of predictor includes predictors from all considered spatial locations. For example, P1 includes predictors for all the

Deleted: assoicated with

Deleted: -performance

Deleted: and XGB models. Surprisingly

Deleted: both models

Deleted: drainage basins

Formatted: English (United States)

Deleted: finding reveals

Deleted: has

Deleted: still

Deleted: have the potential

Deleted: improve the

Deleted: are

Deleted: ***Insert Figure 7 here¶***
Insert Table 3 here¶
¶
It should be noted that the variation of importance scores among predictors for the WFI method is much smaller than it for other feature importance methods. This is caused by the nature of Wilk's Λ: In

Moved (insertion) [17]

considered climatic stations with 1-day precipitation. Therefore, the importance score of P1 is the average of the importance score from the predictors of P1.

Table 3: Predictive accuracy for reinserting the predictors in P1 to the RF model (Spring irrigation).

	Basin	RF with P1	RF without P1
RMSE	1st	**2.42**	2.44
	2nd	**3.16**	3.17
	3rd	**5.81**	5.81
Adjusted R^2	1st	**0.81**	0.81
	2nd	**0.77**	0.76
	3rd	**0.69**	0.69

Note: The RF model was based on the optimum set of predictors in RFE iterations.

6. Discussion

There could be several reasons why WFI can have more robust variable rankings than other feature importance measures. First, WFI does not rely on performance measures to evaluate the variable importance. Instead, it depends on Wilk's Λ, which prevent any splitting that due to chance. In fact, in the node splitting process, a predictor that significantly increases the predictive accuracy may not necessarily have the ability to differentiate two potential sub-spaces. Therefore, the WFI method (which evaluates every splitting and merging action based on Wilk's test-statistics with the predefined significance level $\boldsymbol{\alpha}$) is expected to generate more robust variable rankings. Second, WFI considers all the interactions among predictors in the tree deduction process, while PFI can only consider the effect of one predictor at a time. Thus the interactions between the target predictor and the rest predictors are overlooked. For example, in section 4, the SCE-PFI selected predictors achieved higher performance (over the testing dataset) than the WFI selected ones. However, these SCE-PFI selected predictors are model-specific, which means when transferring these predictors to the other model (e.g., RF), they may not deliver the optimum performance. In contrast, the WFI selected predictors have good transferability: they helped RF achieve optimum predictive accuracy. Similar evidence was also found by Schmidt et al. (2020), who reported that the variable rankings from PFI might vary significantly according to different algorithms. This fact has been considered a major challenge for hydrological inference because one cannot reach the same reasoning with different algorithms. Based on the results above, we can conclude that the WFI could produce more robust variable rankings, which enables a universal solution rather than a specific one for hydrological inference.

Deleted: the node splitting process, a predictor that significantly increases the predictive accuracy may not necessarily have a strong separative power (i.e., a small Λ value) to differentiate two potential sub-spaces. As a consequence, such a predictor could gain a relatively higher importance score for accuracy-based interpretation methods than the WFI. However, predictors identified by accuracy-based interpretation methods maybe subject to overfitting, which does not guarantee a valid inference on the testing dataset. The WFI method (which evaluates every spliting and merging action based on Wilk's test-statistics with the predefined significance level $\boldsymbol{\alpha}$) is less likely to be overfitted and expected to generate more reliable importance scores. ¶
¶
Figure 8 and S8 depict the varying roles of a predictor played in the hydrological processes, that is, streamflows at 25%, 50%, 75%, 90%, 95%, 99% and 100% percentiles. We found the P3 and P5 (i.e., 3 and 5-day accumulative precipitation) have higher importance scores for peak flows than those for low flows during the spring irrigation periods, while no significant trends can be observed for P1. This is probably because the accumulative precipitation bears the information of both antecedent watershed conditions and the storms, which are the keys to the formation of streamflow peaks. Significant trends on importance scores also can be found for some irrigation-related factors. For instance, the I1 and I3 (i.e., one- and three-day accumulative irrigation) in the first drainage area share relatively higher importance scores on low flows than those for the high and peak flows. This probably due to the majority of the irrigated lands in the first drainage area being paddy fields, which require flood irrigation to soak the rice fields (which are usually bunded) for a few days. Once the unsaturated zone of the soil becomes saturated (which usually happens two weeks after the beginning of irrigation), the groundwater table will be elevated and the irrigation factors such as I15 and I30 will increase their dominance on peak flows. ¶
***Insert Figure 8 here*¶**
¶
Owing to the different characteristics of the study watersheds, the same factor may behave distinctively according to the landuse, irrigation schedule and cropping pattern. The third drainage basin, for example, shows a decreasing (or increasing) trend of importance scores for I30 (or P5) as the flow magnitude increases. This indicates that the peak flows in this area are probably caused by excessive rainfall rather than long-term irrigation. This is quite different from the first irrigated area. In fact, most of the irrigated lands in the third drainage area grow corn which does not require to be soaked as rice does. Therefore, the irrigated water in this area will drain faster than that in the first irrigated area. Moreover, the third drainage baisn includes more mountainous area than other basins, which allows a shorter time of concentration, making the precipitation the dominant factor in the peak flows.¶
6. Discussion¶
Previous studies indicated that equifinality is a major challenge for hydrological inference using conventional machine learning approaches such as RF and MLP (Schmidt et al., 2020). Studies from Schmidt et al. (2020) mentioned that the patterns of importance scores (achieved by PFI) may vary significantly according to different algorithms. Their ...

RFE was used to identify the most relevant predictors for optimum predictive accuracy. This approach could be quite useful in real-world practice, especially in hydrology, where the simulation problem may involve hundreds of inputs (from climate models, observations or remote sensing, etc.) describing the spatial and temporal variabilities of the system. Each of these inputs may contain useful information, while it also contains noise that will mislead the model (e.g., increase the simulation errors). Therefore, it is critical to eliminate those variables that cannot improve the predictive accuracy. WFI, in combination with the RFE process, can thus be used for facilitating hydrological inference and modelling.

7. Conclusions

WFI was developed to improve the robustness of variable rankings for tree-structured statistical models. Our results indicate that the proposed WFI can provide more robust variable rankings than well-known PFI and MDI methods. In addition, we found that the predictors selected by WFI can replace those selected by RF with its default methods to improve the model predictive accuracy.

The achievements of the proposed WFI approach could be two-fold: firstly, robust variable rankings are provided for a sound hydrological inference. In specific, some critical predictors that may be overlooked by conventional feature importance methods (PFI and MDI) can be captured through WFI. Secondly, the enhanced variable rankings combined with RFE process can help identify the most important predictors for the optimum model predictive accuracy.

The proposed WFI could be a step closer for earth system scientists to get a preliminary understanding of the hydrological process through ML. Future studies may focus on the development of tree-structured hydrological models that not only be viewed as black-box heuristics but also can be used for rigorous hydrological inference. Even though the focus of this paper is hydrological inference, WFI can also be applied to a variety of important applications. Moreover, current applications of importance scores are still limited. As interpretable ML continues to mature, its potential benefits for hydrological inference could be promising.

Deleted: The Wilk's feature importance

Deleted: interpretability

Deleted: decision trees and regression

Deleted: ensembles.

Deleted: provides a

Deleted: hydrological inference, compared with the

Deleted: method

Deleted: method

Deleted: WFI can identify more informative

Deleted: , compared with PFI and MDI in terms of predictive accuracy (i.e., adjusted R^2 and RMSE). With the provisions of the BMA algorithm, the posterior information allowed the

Deleted: downscale the global importance scores to local ones. The localized importance scores can reflect the varying characteristics of a predictor involved in the hydrological processes.

Deleted: There are three main

Deleted: in hydrology

Deleted: the issue of equifinality that exists in conventional statistical models can be partially addressed by using the proposed WFI method; secondly

Deleted: the WFI; thirdly, the posterior-informed WFI can

Deleted: to gain insights into some hydrological behaviours

Deleted: Although a complete description of all the decision trees within the model is infeasible, the

Deleted: hydrologists

Deleted: machine learning. However, several challenges still exist in the current interpretation approach such as how to find the best balance amonst the model complexity, performance, and interpretability. A complex model

Deleted: yield higher performance skills than a simple model, but at the same time, will introduce the multicollinearity problem, which in turn will hamper the model interpretability.

Deleted: machine learning models continue

Deleted: the

Deleted: of

Deleted: if the importance scores can be associated with physically-based hydrological models.

Code and data availability. The climatic data are available on the data repository of China meteorological data service center (http://data.cma.cn/en). The hydrological data and code for the numeric case can be accessed from Zenodo repository (https://doi.org/10.5281/zenodo.4387068). The entire model code for this study can be obtained upon email request to the corresponding author.

Author contribution. Kailong Li designed the research under the supervision of Guohe Huang. Kailong Li carried out the research, developed the model code and performed the simulations. Kailong Li prepared the manuscript with contributions from Guohe Huang and Brian Baetz.

Competing interests: The authors declare that they have no conflict of interest.

Acknolwdgement. We appreciate Ningxia Water Conservancy for offering the streamflow, groundwater and irrigation data, as well as related help. We are also very grateful for the helpful inputs from the Editor and anonymous reviewers.

Financial support. This research was supported by Canada Research Chair Program, Natural Science and Engineering Research Council of Canada, Western Economic Diversification (15269), and MITACS.

Deleted: and

8. References

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences, 21(10): 5293-5313.

Ahn, K.-H., 2020. A neural network ensemble approach with jittered basin characteristics for regionalized low flow frequency analysis. Journal of Hydrology, 590: 125501.

Apley, D.W., Zhu, J., 2016. Visualizing the effects of predictor variables in black box supervised learning models. arXiv preprint arXiv:1612.08468.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. Annals of Statistics, 47(2): 1148-1178.

Barandiaran, I., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell, 20(8): 1-22.

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines
Deleted: .:
Deleted: ,
Deleted: ,
Deleted: , 2020.
Deleted:
Deleted: and
Deleted: .:
Deleted: ,
Deleted: , 2016.
Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines
Deleted: .:
Deleted: ,
Deleted: ,
Deleted: , 1998.

Bénard, C., Biau, G., Veiga, S., Scornet, E., 2021. Interpretable random forests via rule extraction, International Conference on Artificial Intelligence and Statistics. PMLR, pp. 937-945.

Beven, K.J., 2011. Rainfall-runoff modelling: the primer John Wiley & Sons

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.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees CRC press

Craven, M., Shavlik, J.W., 1996. Extracting tree-structured representations of trained networks, Advances in neural information processing systems, pp. 24-30

Du, Q., Biau, G., Petit, F., Porcher, R., 2021. Wasserstein Random Forests and Applications in Heterogeneous Treatment Effects, International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1729-1737.

Epifanio, I., 2017. Intervention in prediction measure: a new approach to assessing variable importance for random forests. BMC bioinformatics, 18(1): 1-16.

Esri. "Topographic" [basemap]. Scale Not Given. "World Topographic Map". December 19, 2020.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research, 15(1): 3133-3181.

Friedberg, R., Tibshirani, J., Athey, S., Wager, S., 2020. Local linear forests. Journal of Computational and Graphical Statistics: 1-15.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine Annals of statistics: 1189-1232

Galelli, S., Castelletti, A., 2013. Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. Hydrology and Earth System Sciences, 17(7): 2669.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation Journal of Computational and Graphical Statistics, 24(1): 44-65

Deleted:

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted: .:

Deleted: ,

Deleted: , 2011.

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted:

Deleted: and

Deleted:

Deleted: .:

Deleted: ,

Deleted: , 1984.

Deleted: Breiman, L.: Random forests, Machine learning, 45, 5-32, 2001.¶
Chen, T., He, T., Benesty, M., Khotilovich, V., and Tang, Y.: Xgboost: extreme gradient boosting, R package version 0.4-2, 1-4, 2015.¶
Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, 2011.¶
Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., ...

Deleted: and

Deleted:

Deleted: .:

Deleted: 1996,

Deleted: ,

Deleted: Duan, Q., Ajami, N. K., Gao, X., and Sorooshia ...

Formatted ...

Deleted:

Formatted ...

Deleted:

Deleted: .:

Deleted: ,

Deleted: ,

Deleted: , 2001.

Formatted ...

Deleted: and

Deleted: .:

Deleted: ,

Deleted: ,

Deleted: , 2015.

Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. Statistics and Computing, 27(3): 659-678.

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.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics, 15(3): 651-674.

Huang, G., 1992. A stepwise cluster analysis method for predicting air quality in an urban environment. Atmospheric Environment. Part B. Urban Atmosphere, 26(3): 349-357.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival forests. Annals of Applied Statistics, 2(3): 841-860.

Katuwal, R., Suganthan, P.N., Zhang, L., 2020. Heterogeneous oblique random forest. Pattern Recognition, 99: 107078.

Kendall, M.G., 1948. Rank correlation methods.

Kisi, O., Choubin, B., Deo, R.C., Yaseen, Z.M., 2019. Incorporating synoptic-scale climate signals for streamflow modelling over the Mediterranean region using machine learning models. Hydrological Sciences Journal, 64(10): 1240-1252.

Konapala, G., Mishra, A., 2020. Quantifying climate and catchment control on hydrological drought in the continental United States. Water Resources Research, 56(1): e2018WR024620.

Kratzert, F. et al., 2019a. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resources Research, 55(12): 11344-11354.

Kratzert, F. et al., 2019b. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology & Earth System Sciences, 23(12).

Lawson, E., Smith, D., Sofge, D., Elmore, P., Petry, F., 2017. Decision forests for machine learning classification of large, noisy seafloor feature sets. Computers & Geosciences, 99: 116-124.

Leathers, D.J., Yarnal, B., Palecki, M.A., 1991. The Pacific/North American teleconnection pattern and United States climate. Part I: Regional temperature and precipitation associations. Journal of Climate, 4(5): 517-528.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3): 18-22.

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted: and

Deleted: .:

Deleted: ,

Deleted: ,

Deleted: , 2002.

Deleted: .:

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted: ,

Deleted: ,

Deleted: , 1992.

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted: and

Deleted: .:

Deleted: ,

Deleted: ,

Deleted: , 2020.

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted: and

Deleted: .:

Deleted: ,

Deleted: ,

Deleted: , 2017.

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted: and

Deleted: .:

Deleted: ,

Deleted: ,

Deleted: , 2002.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions, Advances in neural information processing systems, pp. 4765-4774.

Mann, H.B., 1945. Nonparametric tests against trend. Econometrica: Journal of the Econometric Society: 245-259.

Mantua, N.J., Hare, S.R., Zhang, Y., Wallace, J.M., Francis, R.C., 1997. A Pacific interdecadal climate oscillation with impacts on salmon production. Bulletin of the american Meteorological Society, 78(6): 1069-1080.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267: 1-38.

Molnar, C., 2020. Interpretable Machine Learning. Lulu. com.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics, 33(2): 161-174.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44): 22071-22080.

Nath, R., Pavur, R., 1985. A new statistic in the one-way multivariate analysis of variance. Computational Statistics & Data Analysis, 2(4): 297-315.

Newman, A. et al., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrology and Earth System Sciences, 19(1): 209.

Reichstein, M. et al., 2019. Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743): 195-204.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016a. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016b. " Why should I trust you?" Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144.

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(5): e2019WR025924.

Deleted:

Deleted: and

Deleted: .:

Deleted: 2017,

Deleted: ,

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted: .:

Deleted: ,

Deleted: , 2020.

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted:

Deleted: and

Deleted: .:

Deleted: ,

Deleted: ,

Deleted: , 2019.

Deleted: and

Deleted: .:

Deleted: ,

Deleted: ,

Deleted: , 1985.

Deleted: Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M ...

Formatted ...

Deleted: ., Camps-Valls, G., Stevens, B., Jung, M., ...

Deleted: ,

Deleted: ,

Deleted: , 2019.

Deleted:

Formatted ...

Deleted: and

Deleted: .:

Deleted: 2016a,

Deleted: ,

Deleted: Ribeiro, M. T., Singh, S., and Guestrin, C.: ...

Deleted: and

Deleted: .:

Deleted: ,

Deleted: ,

Deleted: , 2020.

Scornet, E., 2020. Trees, forests, and impurity-based variable importance. arXiv preprint arXiv:2001.04295.

Shapley, L.S., 1953. A value for n-person games. Contributions to the Theory of Games, 2(28): 307-317.

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 & Earth System Sciences, 20(7).

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC bioinformatics, 9(1): 1-11.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1): 25.

Trenberth, K.E., 1997. The definition of el nino. Bulletin of the American Meteorological Society, 78(12): 2771-2778.

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523): 1228-1242.

Wilks, S.S., 1967. Collected papers; contributions to mathematical statistics. Wiley.

Worland, S.C., 2018. Data-driven methods for hydrologic inference and discovery. Vanderbilt University.

Worland, S.C., Steinschneider, S., Asquith, W., Knight, R., Wieczorek, M., 2019. Prediction and Inference of Flow Duration Curves Using Multioutput Neural Networks. Water Resources Research, 55(8): 6850-6868.

Xia, R., 2009. Comparison of Random Forests and Cforest: Variable Importance Measures and Prediction Accuracies.

Yang, J. et al., 2015. Drought adaptation in the Ningxia Hui Autonomous Region, China: Actions, planning, pathways and barriers. Sustainability, 7(11): 15029-15056.

Yang, Y., Chui, T.F.M., 2020. Modeling and interpreting hydrological responses of sustainable urban drainage systems with explainable machine learning methods. Hydrology and Earth System Sciences Discussions: 1-41.

Yu, B., 2013. Stability. Bernoulli, 19(4): 1484-1500.

Zhang, Y., Chiew, F.H., Li, M., Post, D., 2018. Predicting Runoff Signatures Using Regression and Hydrological Modeling Approaches. Water Resources Research, 54(10): 7859-7878.

Deleted: .:…, 2020. Trees, forests, and impurity-based variable importance,… arXiv preprint arXiv:2001.04295, 2020.

Deleted: ….:…, 1953. A value for n-person games,… Contributions to the Theory of Games, 2,…28): 307-317, 1953.

Deleted: …., Guikema, S. …., and …aitchik, B. ….:…, 2016. Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds,… Hydrology & Earth System Sciences, 20, 2016.

Deleted: and

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted: .:…, 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution,… BMC bioinformatics, 8,…1): 25, 2007.

Deleted: ….:…, 1967. Collected papers; contributions to mathematical statistics,… Wiley, 1967.

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted: ….:…, 2018. Data-driven methods for hydrologic inference and discovery,… Vanderbilt University, 2018.

Deleted: …., Steinschneider, S., Asquith, W., Knight, R., and …ieczorek, M.:…, 2019. Prediction and Inference of Flow Duration Curves Using Multioutput Neural Networks,… Water Resources Research, 55,…8): 6850-6868, 2019.

Deleted: Yang, J., Tan, C., Wang, S., Wang, S., Yang, Y., and Chen, H.:

Formatted: Indent: Left: 0", Hanging: 0.5", Line spacing: 1.5 lines

Deleted: ,… Sustainability, 7,…11): 15029-15056, 201

Deleted: and …hui, T. …. ….:…, 2020. Modeling and interpreting hydrological responses of sustainable urban drainage systems with explainable machine learning methods,… Hydrology and Earth System Sciences Discussions,… 1-41, 2020.

Deleted: …., Li, M., and …ost, D.:…, 2018. Predicting Runoff Signatures Using Regression and Hydrological Modeling Approaches,… Water Resources Research, 54,…10): 7859-7878, 2018.¶

Formatted: EndNote Bibliography, Left, Indent: Left: 0", Hanging: 0.5", Automatically adjust right indent when grid is defined, Line spacing: 1.5 lines, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Moved up [12]: In this map, ***G*** indicates groundwater gauges, ***W*** indicates weather stations, ***R*** indicates rain stations, ***C*** indicates irrigation canals and ***O*** indicates drainage outlets.

1980