ORIGINAL PAPER

Feature importance measures to dissect the role of sub-basins in shaping the catchment hydrological response: a proof of concept

Francesco Cappelli[1] · Flavia Tauro[2] · Ciro Apollonio[3] · Andrea Petroselli[4] · Emanuele Borgonovo[1] · Salvatore Grimaldi[2]

Abstract
Understanding the response of a catchment is a crucial problem in hydrology, with a variety of practical and theoretical implications. Dissecting the role of sub-basins is helpful both for advancing current knowledge of physical processes and for improving the implementation of simulation or forecast models. In this context, recent advancements in sensitivity analysis tools could be worthwhile for bringing out hidden dynamics otherwise not easy to distinguish in complex data driven investigations. In the present work seven feature importance measures are described and tested in a specific and simplified proof of concept case study. In practice, simulated runoff time series are generated for a watershed and its inner 15 sub-basins. A machine learning tool is calibrated using the sub-basins time series for forecasting the watershed runoff. Importance measures are applied on such synthetic hydrological scenario with the aim to investigate the role of each sub-basin in shaping the overall catchment response. This proof of concept offers a simplified representation of the complex dynamics of catchment response. The interesting result is that the discharge at the catchment outlet depends mainly on 3 sub-basins that are consistently identified by alternative sensitivity measures. The proposed approach can be extended to real applications, providing useful insights on the role of each sub-basin also analyzing more complex scenarios.

Keywords Feature importance measures · Global sensitivity measures · Machine learning · Random Forest · Catchment hydrological response

1 Introduction

Storm hydrographs have been traditionally associated with physical portions of a catchment (Betson 1964; Hewlett 1974), whereby catchment runoff has been described as a threshold-driven interaction of phenomena (Ali et al. 2013;

Bonell 1998; Graham and McDonnell 2010; Graham et al. 2010; Lehmann et al. 2007; Uchida et al. 2005; Zehe et al. 2005), whose prominence has been associated with rainfall, seasonality, and connectivity (Detty and McGuire 2010; Hopp and McDonnell 2009; Iwasaki et al. 2020; Jencso and McGlynn 2011; Liu et al. 2019; McGuire and

✉ Francesco Cappelli
  francesco.cappelli@phd.unibocconi.it

  Flavia Tauro
  flavia.tauro@unitus.it

  Ciro Apollonio
  ciro.apollonio@unitus.it

  Andrea Petroselli
  petro@unitus.it

  Emanuele Borgonovo
  emanuele.borgonovo@unibocconi.it

  Salvatore Grimaldi
  salvatore.grimaldi@unitus.it

[1] Department of Decision Science, Università Bocconi, 20136 Milan, Italy

[2] Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia, 01100 Viterbo, Italy

[3] Department of Agriculture and Forest Sciences, University of Tuscia, 01100 Viterbo, Italy

[4] Department of Economics, Engineering, Society and Business Organization, University of Tuscia, 01100 Viterbo, Italy

McDonnell 2010; Scaife and Band 2017; Subagyono et al. 2005). Efforts to investigate the contribution of individual compartments to catchment-wide stormflow are limited (Asano et al. 2020; Beiter et al. 2020; Bergstrom et al. 2016; Demand et al. 2019; Guastini et al. 2019; Jencso et al. 2009). For instance, in Asano et al. (2020), the catchment-wide propagation of a stormflow peak was studied by quantifying flow paths in hillslopes and channels. According to this study, during intense storms, the hillslope response may be quicker than theoretically predicted, thus abruptly increasing stormflow. Despite several studies supporting the relevance of sub-basins in governing the catchment-wide storm hydrograph, a quantitative framework to describe their dynamics, and eventually, inform monitoring of critical sub-catchment compartments is still lacking. Investigating the hydrological response at the sub-catchment level involves coping with a large amount of hydrological data. In this vein, recent and rapid technological advancements are providing new instrumentation, impressive computational power and huge data storage opportunities to deal with big volumes of hydrological data (Butler 2014; Tauro et al. 2018). In turn, big data mandate advanced data analysis techniques (Chen and Han 2016; Chen and Wang 2018; Blöschl et al. 2019; Sun and Scanlon 2019; Papacharalampous et al. 2021).

Among emerging statistical and data mining methods, machine learning (ML) approaches have had an impressive diffusion in the environmental sciences and specifically in hydrology. Several ML techniques, such as ensemble and ordinary learning algorithms (i.e. Model Averaging, Stacking, Bagging, Boosting, Dagging) have been extensively tested, compared, and applied in river flow, river quality, sediment transport, rainfall-runoff, and groundwater modelling for simulation and forecasting applications at diverse time aggregation scales. The success of such approaches is due as well to the mentioned increasing data availability and to the complexity of hydrological phenomena, which are difficult to model with linear or simple non linear statistical methods. For a full overview on the use of ML methods in hydrology, the reader could refer to the following recent papers: Zounemat-Kermani et al. (2021), Gharib and Davies (2021), Rajaee et al. (2020), Tyralis et al. (2021).

Despite their popularity, ML models are often regarded as "black boxes" whose internal working is not transparent to the analyst (Molnar 2020). Towards correctly interpreting ML model findings, diagnostic tools (such as feature importance measures, marginal effect indicators, etc.) may be beneficial. Among ML diagnostic techniques, feature importance measures provide knowledge about the key-drivers of uncertainty that drive the response of the ML model. Several methods have been developed to assess feature importance. They can be distinguished in model-specific and model-agnostic methods (Molnar 2020). Model-specific methods can be used solely in conjunction with the ML model to which they are associated. Model-agnostic methods are applicable to general classes of models. In this class, popular approaches are permutation feature importance (PFI) measures (Breiman 2001a; Fisher et al. 2019).

In hydrology, Schmidt et al. (2020) use PFI measures to check whether the key-drivers in forecasting the flood magnitude match among different ML models. Thorslund et al. (2021) use conditional PFI measures to recognise key-drivers in predicting salinity levels.

Identifying influential features is also a crucial task in Sensitivity Analysis (SA) (Saltelli et al. 2008). More specifically, factor prioritization is the determination of the features that drive variability in the model output (see Saltelli et al. (2004); Borgonovo and Plischke (2016) for a review). This information can be obtained using variance-based sensitivity indices (Iman and Hora 1990; Saltelli 2002), density-based sensitivity indices (Borgonovo 2007) or cumulative distribution-based sensitivity indices (Gamboa et al. 2018). These indices quantify the degree of statistical dependence between the output and the features (Borgonovo 2007; Saltelli et al. 2008). The computation of these indices can be performed using a data-driven approach (Plischke et al. 2013), which enables us to estimate the corresponding measures directly from given data. In hydrology, Borgonovo et al. (2017) employ such an approach to identify the most important features in hydrological models of a river catchment generated using the Framework for Understanding Structural Errors (FUSE) (Clark et al. 2008).

In this work we test seven feature importance measures combining model-agnostic methods and global SA indices and, for the first time in hydrology, we employ *Shapley feature importance* (Casalicchio et al. 2018), *ALE-indices* (Borgonovo et al. 2022 (Unpublished data)) and *ALE-based feature importance* (Greenwell et al. 2018). Such testing is performed through a proof of concept that aims to understand a catchment hydrological response by investigating how the sub-basins of a selected natural watershed contribute to its storm response. More specifically, the aim of the proposed preliminary application is to verify if it is possible (with the current results and/or in future research applications) to answer the following questions:

1. Does one (or more) sub-basin exist that contributes more than others to the catchment-scale hydrological response?
2. Do eventually dominant sub-basins exhibit distinctive morpho-hydrological characteristics that control the feature importance measure analysis results?

To this end, we focus on a natural catchment divided in 15 sub-basins, and analyze their individual flow discharge signals along with the flow discharge at the catchment outlet. Given the nature of the proof of concept, in this preliminary work we opted for the well-known Hydrologic Modeling System (HEC-HMS) semi-distributed hydrological model for simulating runoff time series, and for a supervised ML model for forecasting the catchment outlet discharge. This simple model configuration (maybe the simplest) will help to verify if the feature importance measure could contribute to answer questions 1 and 2.

Addressing these outstanding questions bears remarkable implications for the comprehension of hydrological systems. In fact, identifying sub-basins within the catchment as critical for the whole hydrological response is expected to open new avenues in rainfall-runoff modeling as well as in environmental monitoring and engineering practice. For instance, the design of monitoring networks and the installation of sensors in the catchment may be optimized by insights on the areas that more significantly contribute to watershed stormflow. The manuscript is organized as follows: in Sect. 2 a full description of the seven feature importance measures is provided. In Sect.3 case study information is described. The results for the hydrological scenario and the ML forecasting tool are illustrated in Sects. 4.1 and 4.2. In Sect. 4.3 the importance measure analysis is reported. Comments and discussions are given in Sect. 4.4. Finally, Sect. 5 concludes the manuscript.

2 Feature importance methods

Here, we introduce the notation. In Section 2.1 we describe three sensitivity measures from SA: variance-based, density-based and distribution-based methods. In Section 2.2, we present some of the most relevant model-agnostic feature importance approaches applied in supervised machine learning.

Let Y and $\mathbf{X} = (X_1, ..., X_p)$ denote the random variables/vectors on the reference probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$, with $\mathbf{X} \in \mathcal{X}_P \subseteq \mathbb{R}^p$, $Y \in \mathcal{Y} \subseteq \mathbb{R}$ and push forward measures $\mathbb{P}_\mathbf{X}$ and $\mathbb{P}_Y$. We denote the observed value of the j-th feature as $\mathbf{x}_j = (x_j^{(1)}, ..., x_j^{(N)})'$ and the i-th observation as $\mathbf{x}^{(i)} = (x_1^{(i)}, ..., x_p^{(i)}) \in \mathcal{X}_P$ associated with the corresponding target value $y^{(i)} \in \mathcal{Y}$. In the remainder, it will be useful to write $\mathbf{X}$ as $\mathbf{X} = (X_j, \mathbf{X}_{-j})$, where $\mathbf{X}_{-j} = \{X_l : l = 1, ..., p, l \neq j\}$. We also have $\mathbf{x} = (x_j, \mathbf{x}_{-j})$. The data is divided into training data and testing data. We suppose to use a supervised ML model $\widehat{f}$ to learn the unknown mapping from an independent and identically distributed training sample $\{(\mathbf{x}_i, y_i) : i = 1, ..., N\}$. We

denote the loss function by $\mathcal{L} : \mathcal{Y} \times \mathbb{R}^p \to \mathbb{R}_{\geqslant 0}$, which quantifies the difference between the vector of observed target values y and the vector of predicted values $\widehat{y}$. It is used to compute the generalization error for a given fitted ML model on unseen test data, i.e., $ge(\widehat{f}) = \mathbb{E}_{\mathbf{X}Y}(\mathcal{L}(Y, \widehat{f}(\mathbf{X})))$.

2.1 Global sensitivity measures

2.1.1 Variance-based sensitivity index

The variance-based sensitivity measure of X_j is defined as (Iman and Hora 1990; Homma and Saltelli 1996)

$$\eta_j^2 = \frac{\mathbb{V}_{X_j}[\mathbb{E}_{\mathbf{X}_{-j}}[Y \mid X_j]]}{\mathbb{V}[Y]} = \frac{\mathbb{V}[Y] - \mathbb{E}_{\mathbf{X}_{-j}}[\mathbb{V}_{X_j}[Y \mid X_j]]}{\mathbb{V}[Y]}. \quad (1)$$

This sensitivity measure corresponds to the expected reduction in model output variance achieved by fixing X_j and coincides with the Pearson correlation ratio (Pearson 1905).

2.1.2 Density-based sensitivity index

This sensitivity measure quantifies the expected distance between the marginal output density p_Y and the conditional density $p_{Y|X_j}$ through the L_1-norm (Borgonovo 2007). It is given by

$$\delta_j = \frac{1}{2} \mathbb{E}_{X_j} \left[\int_\mathcal{Y} \left| p_Y(y) - p_{Y|X_j}(y) \right| dy \right], \quad (2)$$

where $p_{Y|X_j}(y)$ can be obtained by fixing X_j at a realization value.

2.1.3 Cdf-based sensitivity indices

The cumulative distribution-based sensitivity measure β^{KS} (Borgonovo et al. 2014) is based on the Kolmogorov-Smirnov distance between cumulative distribution functions ($\mathbb{P}_Y$, $\mathbb{P}_{Y|X_j}$, respectively). It is defined as

$$\beta_j^{KS} = \mathbb{E}_{X_j} \left[\sup_\mathcal{Y} \left| \mathbb{P}_Y(y) - \mathbb{P}_{Y|X_j}(y) \right| dy \right]. \quad (3)$$

The global sensitivity measures in Eqs. (1), (2), (3) can be estimated from the same dataset of features-forecast realizations. The computation is performed using the given-data (or one-sample) approach proposed in Plischke et al. (2013). In order to use this approach, the support $\mathcal{X}_j$ of the feature X_j is partitioned into mutually exclusive and collectively exhaustive classes. Formally, we denote the partition of $\mathcal{X}_j$ into M classes as $\mathcal{P} = \{\mathcal{C}_m : m = 1, ..., M\}$ with $\mathcal{C}_{m,j} \cap \mathcal{C}_{m',j} = \emptyset$, $\mathcal{X}_j = \bigcup_{m=1}^M \mathcal{C}_{m,j}$ for $m \neq m'$. Let $N_{m,j}$ be the

number of observations of the response variable Y in the m-th class and N be the total number of observations.

Then, an estimate of η_j^2 is given by

$$\widehat{\eta}_j^2 = \frac{\sum_{m=1}^M \frac{N_{m,j}}{N} \left(\overline{y}_{m,j} - \overline{y}\right)^2}{\frac{1}{N-1} \sum_{i=1}^N (y_i - \overline{y})^2}, \tag{4}$$

where $\overline{y}_{m,j}$ and $\overline{y}$ are an estimate of the conditional mean of Y given $X_j \in \mathcal{C}_{m,j}$ and an estimate of the mean of Y, respectively.

An estimate of the δ-measure in Eq. (2) is given by:

$$\widehat{\delta}_j = \sum_{m=1}^M \frac{N_{m,j}}{N} \int_{\mathcal{Y}} \left|\widehat{p}_Y(y) - \widehat{p}_{m,j}(y)\right| dy, \tag{5}$$

where $\widehat{p}_Y$ and $\widehat{p}_{m,j}$ are kernel smoothing functions of the output vector $\mathbf{y} = \{y_j : j = 1, ..., p\}$ and the within class output vector $\mathbf{y}_{m,j} = \{y_j : x_j \in \mathcal{C}_m\}$.

An estimate of β_j^{KS} is given by

$$\widehat{\beta}_j^{KS} = \sum_{m=1}^M \frac{N_{m,j}}{N} \max_{i \in \{1, ..., N\}} \left|\widehat{\mathbb{P}}_Y(y_i) - \widehat{\mathbb{P}}_{m,j}(y_i)\right| dy, \tag{6}$$

where $\widehat{\mathbb{P}}_Y$ and $\widehat{\mathbb{P}}_{m,j}$ correspond to the empirical cumulative distribution functions of $\mathbf{y}$ and $\mathbf{y}_{m,j}$, respectively.

2.2 Feature importance in supervised machine learning

In this subsection, we present importance measures specifically defined for machine learning applications.

2.2.1 Conditional permutation feature importance

The permutation feature importance (PFI) of (Breiman 2001a) a model $\widehat{f}$ is defined as

$$\text{PFI}_j = \mathbb{E}_{X^\pi \mathbf{X}_{-j} Y} \left[\mathcal{L}\left(Y, \widehat{f}\left(X_j^\pi, \mathbf{X}_{-j}\right)\right) \right] - \mathbb{E}_{\mathbf{X}Y} \left[\mathcal{L}\left(Y, \widehat{f}(X_j, \mathbf{X}_{-j})\right) \right], \tag{7}$$

where X_j^π follows the marginal distribution of X_j. This importance measure quantifies the variation in the accuracy of the ML model fitted on the (original) training data after permuting a feature of interest. A high value of PFI_j means that the predictive performance of the ML model drops significantly when the dependence between Y and X_j is broken as a result of the permutation of X_j. An estimate of the PFI of feature X_j is given by

$$\widehat{\text{PFI}}_j = \frac{1}{N} \sum_{i=1}^N \mathcal{L}\left(y^{(i)}, \widehat{f}\left(x_j^{\pi,i}, \mathbf{x}_{-j}^{(i)}\right)\right) - \frac{1}{N} \sum_{i=1}^N \mathcal{L}\left(y^{(i)}, \widehat{f}\left(x_j^{(i)}, \mathbf{x}_{-j}^{(i)}\right)\right). \tag{8}$$

Due to their popularity, PFI measures have been set under intensive scrutiny. The work of Hooker and Mentch (2019) shows that PFI measures may lead to misleading results when there is a strong statistical dependence among features. In order to overcome this drawback, numerous alternatives have been explored in the literature (Casalicchio et al. 2018; Strobl et al. 2007; Candes et al. 2018).

Strobl et al. (2007) suggest to rely on a conditional PFI defined as

$$\text{cPFI}_j = \mathbb{E}_{X^{C\pi} \mathbf{X}_{-j} Y} \left[\mathcal{L}\left(Y, \widehat{f}\left(X_j^{C\pi}, \mathbf{X}_{-j}\right)\right) \right] - \mathbb{E}_{\mathbf{X}Y} \left[\mathcal{L}\left(Y, \widehat{f}(X_j, \mathbf{X}_{-j})\right) \right], \tag{9}$$

where $X_j^{C\pi}$ follows the conditional distribution of X_j given $\mathbf{X}_{-j}$. This is equivalent to compute the PFI importance using a conditional permutation scheme. Specifically, the support of X_j is partitioned based on $\mathbf{X}_{-j}$ and then the values of X_j are conditionally permuted within each partition. This approach preserves the data dependence structure without breaking the relationship between the feature and the target variable: see also Debeer and Strobl (2020).

2.2.2 Shapley feature importance

Casalicchio et al. (2018) propose an extension of the PFI measure called Shapley PFI (SPFI). The Shapley PFI is based on the notion of Shapley value (Shapley 1952), a method from game theory that it is known for its attractive fairness properties (Lundberg and Lee 2017).

Consider a coalitional game with a payoff in which a group of p players, denoted by P plays by joining coalitions $K \subseteq P$. We denote the coalition value function by $v : 2^p \to \mathbb{R}_{\geqslant 0}$ with $v(\emptyset) = 0$, where $\emptyset$ denote the empty set. The Shapley value of the j-th player is given by

$$\phi_j(v) = \sum_{K \subseteq P \setminus \{j\}} \frac{|K|!(|P| - |K| - 1)!}{|P|!} [v(K \cup \{j\}) - v(K)], \tag{10}$$

where $v(K \cup \{j\}) - v(K)$ is the individual contribution of the j-th player in coalition K. Shapley values assign players a fraction of the overall value by averaging their contributions to all coalitions. Ribeiro et al. (2016) and Lundberg and Lee (2017) define the value function $v(K)$ as the conditional expectation of the target variable on a specific observation when the features in coalition K are known, that is

$$v(K) = \mathbb{E}[\widehat{f}(\mathbf{X}) \mid \mathbf{X}_K = \mathbf{x}_K] = \mathbb{E}_{\mathbf{X}_{-K} | X_K} \left[\widehat{f}(\mathbf{x}_K), \mathbf{X}_{-K} \right]. \tag{11}$$

Based on this result, Casalicchio et al. (2018) propose the SPFI measure as follows:

$$\text{SPFI}_j = \sum_{K \subseteq P \setminus \{j\}} \frac{|K|!(|P| - |K| - 1)!}{|P|!} \left[v_{ge}(K \cup \{j\}) - v_{ge}(K) \right], \tag{12}$$

where $v_{ge}(K) = ge_K(\widehat{f}) - ge_\emptyset(\widehat{f})$ is the value function associated to the predictive performance of a ML model. Note that $ge_K(\widehat{f})$ is the generalization error computed using features in coalition K and $ge_\emptyset(\widehat{f})$ is the error when no features are considered. SPFI is designed to quantify the individual contribution of each feature to the prediction on each observation $\mathbf{x}$. Casalicchio et al. (2018) show that an estimate of SPFI_j is given by

$$\widehat{\text{SPFI}}_j = \frac{1}{P!} \sum_\pi \left[\widehat{ge}_{B_j(\pi) \cup \{j\}}(\widehat{f}) - \widehat{ge}_{B_j(\pi)}(\widehat{f}) \right], \tag{13}$$

where π is a permutation of the features. Given a permutation π, $B_j(\pi)$ is the set of features preceding X_j. For instance, if we assume that $p = 5$, for $j = 3$ and $\pi = \{2, 5, 3, 4, 1\}$, we have that $B_3(\pi) = \{2, 5\}$.

2.2.3 ALE-index

Recently, Borgonovo et al. 2022 (Unpublished data) propose a new feature importance measure, called ALE-index. We recall that an ALE-plot is a powerful graphical tool that describes the relationship between the prediction of a supervised learning model and the feature of interest (Apley and Zhu 2020). However, it does not provide insights concerning the relative importance of features directly. Borgonovo et al. 2022 (Unpublished data) show that one can compute the importance of X_j exploiting the algorithm that produces these marginal effect indicators introduced by Apley (2018).

To estimate an ALE main effect from data, one needs to partition the support of X_j into K intervals, $\mathcal{X}_j^k = \left[z_j^{k-1}, z_j^k \right]$, with $k = 1, ..., K$, such that their union equals the support of X_j and their intersection is null. Here, N_j^k refers to the number of values of X_j in the k-th interval. Apley and Zhu (2020) propose the following finite difference estimator of $ALE_j(x_j)$

$$\widehat{ALE}_j(x_j) = \sum_{k=1}^K \frac{1}{N_j^k} \sum_{i: \mathbf{x}^{(i)} \in \mathcal{X}_j^k} \left[\widehat{f}\left(z_j^k, \mathbf{x}_{-j}^{(i)} \right) - \widehat{f}\left(z_j^{k-1}, \mathbf{x}_{-j}^{(i)} \right) \right], \tag{14}$$

for each $x_j \in \left[z_j^0, z_j^K \right]$, where z_j^0 is chosen slightly smaller than $x_j^{min} = \min\left\{ x_j^{(1)}, ..., x_j^{(N)} \right\}$ and $z_j^K = x_j^{max} =$ $\max\left\{ x_j^{(1)}, ..., x_j^{(N)} \right\}$. T_j' index is defined as (Borgonovo et al, 2022 (Unpublished data))

$$\text{T}_j' = \frac{1}{2} \frac{\mathbb{E}_{X_j^k \mathbf{X}_{-j} \mathbf{X}} \left[\left(\widehat{f}\left(X_j^k, \mathbf{X}_{-j} \right) - \widehat{f}(\mathbf{X}) \right)^2 \right]}{\sigma_Y^2}, \tag{15}$$

where σ_Y^2 is the output variance. The authors show that, under feature independence, T_j' coincides with the total order sensitivity index of Homma and Saltelli (1996). We recall that a total index captures the contribution of X_j to the output variance including its individual and interaction contributions. The ALE-based feature importance measure possesses the zero-independence property under both independence and dependence (Renyi 1959). This property states that the value of a feature importance measure is zero if and only if the target variable and the feature of interest are independent.

An estimate of T_j' is given by

$$\widehat{T}_{j,N}' = \frac{1}{2N\sigma_y^2} \sum_{i=1}^N \left(\widehat{f}\left(x_j'^{,(i)}, \mathbf{x}_{-j}^{(i)} \right) - \widehat{f}\left(\mathbf{x}^{(i)} \right) \right)^2, \tag{16}$$

where $x_j'^{,(i)}$ is the left endpoint of the interval in which $x_{i,j}$ falls and it is assumed to be sampled from the marginal distribution of X_j.

2.2.4 ALE-based feature importance

Another feature importance measure based on ALE-plots is proposed by Greenwell et al. (2020) and used also in Christensen et al. (2021). This importance measure is given by:

$$\text{ALE-IMP}_j = \sqrt{\mathbb{V}_{X_j}\left(\widehat{\text{ALE}}_j(x_j) \right)}. \tag{17}$$

An estimate of the ALE-based feature importance measure is given by

$$\widehat{\text{ALE-IMP}}_j = \sqrt{\frac{1}{N-1} \sum_{i=1}^N \left[\widehat{\text{ALE}}_j\left(x_j^i \right) - \overline{\widehat{\text{ALE}}_j(x_j)} \right]^2}, \tag{18}$$

where $\overline{\widehat{\text{ALE}}_j(x_j)} = \frac{1}{N} \sum_{i=1}^N \widehat{\text{ALE}}_j\left(x_j^i \right)$. It is defined by computing the sample standard deviation of $\widehat{\text{ALE}}_j$. So, this measure quantifies the variability of the $ALE_j(x_j)$ curve itself. It is defined exploiting the marginal relationship of the target variable and the feature of interest. For a flat ALE curve $\text{ALE-IMP}_j \approx 0$ meaning that X_j has a small influence on Y. Differently, a fluctuating ALE curve has a higher variability and so the value of ALE-IMP_j is larger. We use it for comparison purposes.

Table 1 Feature importance measures calculated in this work. The last column refers to the framework in which the importance measures are evaluated

Importance measure	Definition	Context
Conditional PFI	$\text{cPFI}_j = \mathbb{E}\left[\mathcal{L}\left(Y, \hat{f}\left(X_j^{C\pi}, \mathbf{X}_{-j}\right)\right)\right] - \mathbb{E}\left[\mathcal{L}\left(Y, \hat{f}(X_j, \mathbf{X}_{-j})\right)\right]$ (9)	ML
Shapley PFI	$\text{SPFI}_j = \sum \frac{|K|!(|P|-|K|-1)!}{|P|!}\left[v_{ge}(K \cup \{j\}) - v_{ge}(K)\right]$ (12)	ML
ALE-plot total index	$\text{T}_j^{\text{ALE}} = \frac{1}{2}\frac{\mathbb{E}_{X_j^k\mathbf{X}_{-j}\mathbf{X}}\left[\left(\hat{f}\left(X_j^k, \mathbf{X}_{-j}\right) - \hat{f}(\mathbf{X}_{-j})\right)^2\right]}{\sigma_Y^2}$ (15)	ML
ALE-based importance	$\text{ALE-IMP}_j = \frac{\sqrt{\mathbb{V}_{X_j}(\text{ALE}_j(x_j))}}{\sum_j \text{ALE-IMP}_j}$ (18)	ML
Variance-based measure	$\eta_j^2 = \frac{\mathbb{V}[Y] - \mathbb{E}_{\mathbf{X}_{-j}}[\mathbb{V}_{X_j}[Y|X_j]]}{\mathbb{V}[Y]}$ (1)	SA
Density-based measure	$\delta_j = \frac{1}{2}\mathbb{E}_{X_j}\left[\int_{\mathcal{Y}} |p_Y(y) - p_{Y|X_j}(y)|dy\right]$ (2)	SA
Cdf-based measure	$\beta_j^{KS} = \mathbb{E}_{X_j}\left[\sup_{\mathcal{Y}}|\mathbb{P}_Y(y) - \mathbb{P}_{Y|X_j}(y)|dy\right]$ (3)	SA

ML, Machine Learning; SA, Sensitivity Analysis

Note that, the SA measures are calculated using a data-driven approach of Plischke et al. (2013). So, such measures consider correlations among features and, thus, provide indications of their importance by taking into account the dependency structure in the hydrological system. Contrary, ML measures are selected to deal with the strong correlations among features. Moreover, the statistical importance measures derived from the SA are used to confirm the ML finding arising from the importance analysis.

Finally, in Table 1 we summarize the feature importance measures used in the case study in Section 4.4.

3 Materials and Methods

We consider a watershed with a dense hydrographic monitoring network that provides discharge measurements at n sub-basin outlets, and assume that an ML tool has been selected to forecast discharge values. Calibration is based on available observations. We aim to investigate whether the feature importance measures are able to distinguish the sub-basin influence identifying those that most affect the discharge time series at the outlet. With this general aim, in Sect. 3.1 we present the watershed selected for this application. In Sect. 3.2 we present the semi-distributed hydrology-hydraulic model HEC-HMS used to generate a synthetic hydrologic scenario. In Sect. 3.3, we describe the ML models used for forecasting the catchment outlet discharge and the performance measures used to quantify the accuracy of the ML models.

3.1 Watershed case study description

The selected study site is the Samoggia River basin, a tributary of the Reno River located in the Emilia Romagna

region, Italy (see Fig. 1). We use a digital elevation model at 20m resolution made available to the authors by the Italian Geographic Military Institute. Land cover data related to year 2018 are downloaded from the Coordination of Information on Environment (CORINE) database, and soil data are taken from the soil map provided by the local administration. The elevation of the investigated basin lies in the range 51–883 m a.m.s.l., the total contributing area is 178.5 km^2 and the basin average slope is approximately 19.1%. Regarding land cover, the site is characterized by valley bottoms that are mainly floodplains hosting farmland and urban areas, and by mountain areas in which there are mainly broadleaved woods. Regarding soil data, the catchment can be classified as a mix between loamy sand and sandy loam. Further details on the Reno River basin can be found in Castellarin et al. (2009) and Di Prinzio et al. (2011). Regarding the available hydrological data, rainfall observations are downloaded from Emilia Romagna regional agency for environmental protection website (https://simc.arpae.it/dext3r/), selecting three years (from 1st January 2014 to 31st December 2016) at 1 hour time resolution.

3.2 HEC-HMS model implementation

The synthetic hydrologic scenario is carried out using the software HEC-HMS by the Hydrologic Engineering Center of US Army Corps of Engineer (2017). HEC-HMS allows one to simulate hydrological processes using different options and modules (Chu and Steinman 2009; De Silva et al. 2014). In the present case study, we apply the HEC-HMS to the Samoggia watershed selecting 15 sub-basins as shown in Fig. 1 (panel d). Hereafter, we employ the simplest configuration that includes:

- Spatial homogeneous rainfall estimation through Thiessen Polygons;

Fig. 1 **a** and **b** Samoggia river basin, located in northern Italy, **c** Digital elevation model, Raingauge and drainage network, **d** Fiftteen sub-basins
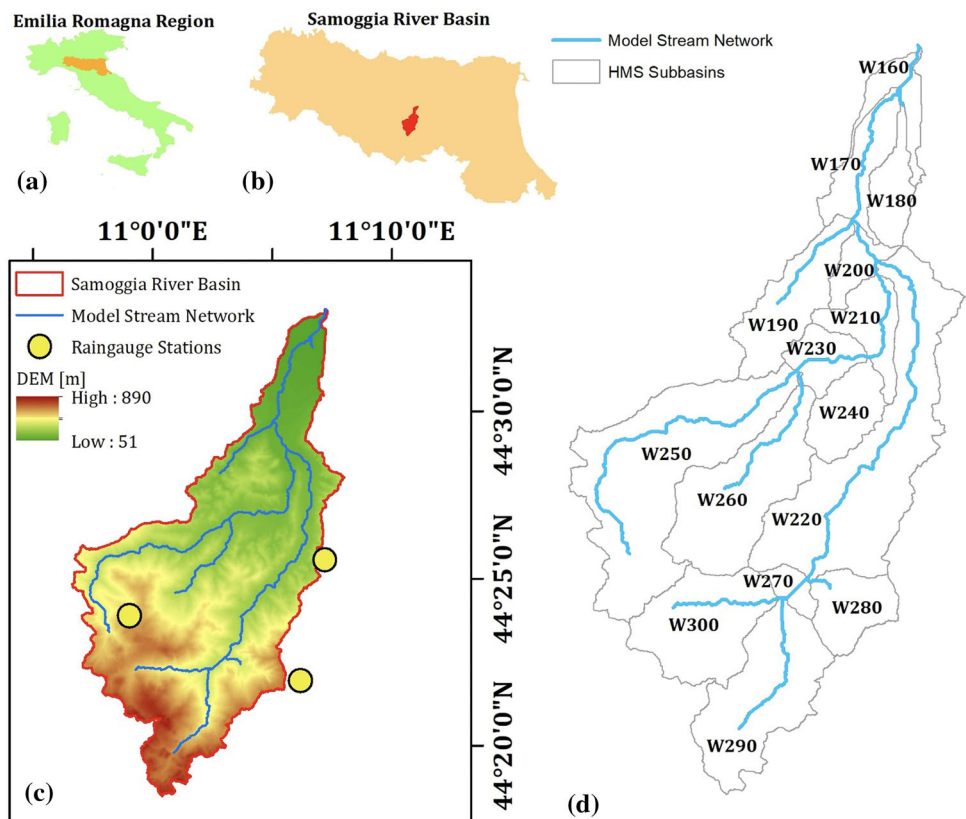
- Soil Conservation Service—Curve Number (CN) infiltration approach;
- Soil Conservation Service—Unit Hydrograph (UH) rainfall runoff model;
- Muskingum method for hydraulic propagation.

We use the physical and hydrological parameters for the sub-basins obtained from HEC-GeoHMS and available in previous literature (Ramly and Tahir 2016; Ramly et al. 2020; Mourato et al. 2021). As mentioned in Sect. 3.1, rainfall data are collected from three rain gauge stations (see panel c in Fig. 1). In order to emphasize the role of sub-basins, we assume a spatially homogeneous rainfall. Thus, the well known Thiessen method can be adopted for computing the gauge-weighting factors. The Soil Conservation Service dimensionless UH is used as rainfall-runoff model. It includes the CN as main parameter affecting infiltration and surface flow velocity defined using land use information. The dimensionless UH is shaped using the concentration time (Tc) and peak discharge (Qp). In particular, Tc is linked to the time lag (TL), calculated by Mockus Formula (Mockus 1964), that depends on the maximum flow length, the mean slope and CN value. The flow length is calculated as the sum of sheet flow, shallow concentrated flow and channel flow. Finally, we select the Muskingum model as flow routing model, setting its

parameters (X, dimensionless attenuation, and K, travel time) equal to 0.5 and 1, respectively (Gilcrest 1950).

3.3 Machine Learning models and performance measures

We employ the following ML Models: a ridge regression, a random forest, a gradient boosting machine and a neural network.

Ridge regression is a regularized version of the linear model, where the loss function includes a penalty term (Gruber 2017). The magnitude of the penalty term is regulated by the hyperparameter *lambda*. The introduction of the penalty term aims to reduce model complexity and prevent overfitting.

Random Forests (Breiman 2001a) and Gradient Boosting machines (Friedman 2001) are tree-based ensemble models. Random Forests consist of many decision trees. A decision tree presents a tree-like structure: it is composed of nodes (root node, decision nodes and leaf node) and branches. The root node represents the entire dataset. The decision nodes flow from the root node and may have several branches representing the decision rules applied to split these nodes. From the decision nodes flow the nodes leaf that are the outcome and have no branches (Hastie et al. (2009), for a broader review). In a Random Forest

each tree is trained on a randomized subset of features and provides separate predictions. By averaging the predictions resulting from the decision trees one obtains the final estimate of the response variable, see plot panel (a) in Fig. 2. This model includes two main hyperparameters: the number of trees (*n.trees*) and the number of features sampled for splitting at each node (*mtry*). For a full description see Liaw and Wiener (2002) and Desai and Ouarda (2021).

Gradient Boosting machines use weak learner models (usually decision trees) to iteratively build a strong learner (ensemble model). At each step, we train a decision tree on the residuals from the previous sequence of trees. The resulting ensemble model is built using an additive model defined through the contributions of each tree, see plot (b) in Fig. 2. The training of a gradient boosting machine requires analysts to set several hyperparameters: the number of trees (*n.trees*), the number of splits it has to perform on a tree (*interaction.depth*), the learning rate (*shrinkage*), the minimum number of observations terminal nodes of the trees (*n.minobsinnode*) and the sub-sampling fraction of the training set values randomly selected to propose the next tree (*bag.fraction*) (Kuhn 2008; Fienen et al. 2018). These tree-based ensemble models are able to manage nonlinear and complex relationships among features. Moreover, Breiman (2001b) shows that Random Forest is not affected by multicollinearity (Farrar and Glauber 1967).

Neural networks are a class of ML models well known for their versatility (Dreiseitl and Ohno-Machado 2002). For this case study, we focus on a single layer neural network H_n, several input neurons X_n and an output layer with the observed outcome O. We denote the connection weights from input to hidden layer by W_n and the connection weights from hidden to the output layer by W_n^{out}. In the hidden and output layer the output is computed as the weighted combination of the outputs of the neurons of the preceding layers processed by a predefined activation function σ, such as the sigmoid function or the softmax function. Specifically, we have $H_n = \sigma(\sum W_n)$ and $O_n = \sigma(\sum H_n W_n^{out})$, respectively (see panel (c) in Fig. 2). The hyperparameters of a single layer neural network are the number of units in the hidden layer (*size*) and the regularization parameter to avoid over-fitting (*decay*) (Teweldebrhan et al. 2020).

In order to achieve a high performance of the ML models, we combine hyperparameter tuning and cross validation. Hyperparameter tuning is a process to search for a set of optimal hyperparameters for a ML model to minimize the loss function (Hastie et al. 2009). We tune the ML models using grid search method (Agrawal 2021). This procedure builds a ML model for every combination of hyperparameters specified in a predefined grid by the analyst and evaluates each ML model through a performance measure using *k*-fold cross-validation. Among all evaluated ML model configurations, we select the hyperparameters that exhibits the smallest performance metric. In the *k*-fold cross-validation scheme (Stone 1974), the data is partitioned into *k* training and validation subsets. The process is repeated for different model configurations. The
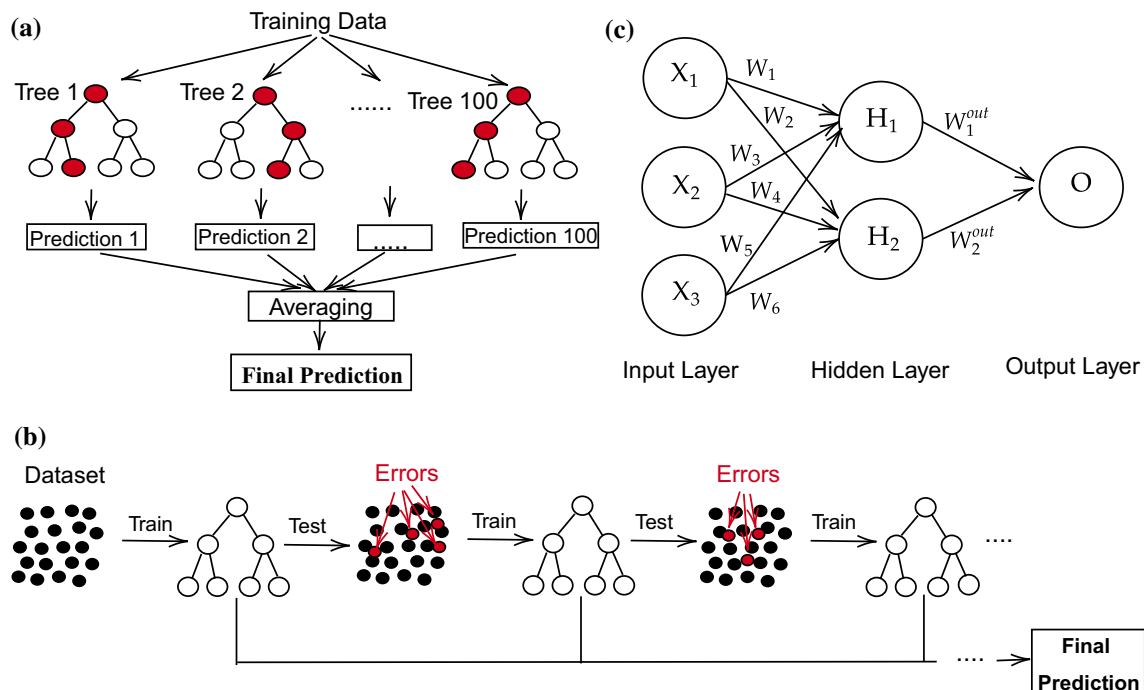
Fig. 2 ML algorithms: **a** Random Forest, **b** Gradient Boosting and **c** one single hidden-layer Neural Network

configuration that achieves the smallest validation error, computed averaging over all k subsets, is selected as optimal.

The accuracy of the ML models is evaluated on the testing data using three criteria: the root-mean-square error (RMSE), the mean absolute error (MAE) and the coefficient of model determination (R^2). The RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}, \tag{19}$$

where y is the vector of observed target values and $\widehat{y}$ is the vector of predicted values. The MAE is the mean of absolute values of differences between observed and predicted values. The MAE is estimated by:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i|. \tag{20}$$

Both performance measures range from 0 to ∞, where the value 0 indicates a perfect fit. RMSE and MAE are measured in the same units as the model output response. MAE is less sensitive to outliers compared to RMSE. The third performance measure is the coefficient of determination (R^2). It is equal to:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2}, \tag{21}$$

where $\overline{y}$ is the average value of y. R^2 is the proportion of variation in the response variable that is explained by the machine learning model forecasts. It ranges from 0 to 1,

where the value 0 indicates that the trained ML model does not explain any variability in the target variable. On the contrary, the value 1 indicates that the trained ML model explains all variability in the target variable.

4 Results and discussion

In this section, we report and discuss the case study results. Firstly, the HMS model implementation is presented in Sect. 4.1, where the characterization of the 15 sub-basins and the 15+1 discharge time series are provided. In Sect. 4.2, the comparison of ML tools is presented. Section 4.3 reports the feature importance measure analysis. Section 4.4 discusses the results of the feature importance analysis.

4.1 Hydrologic synthetic scenario

The watershed case study simulated using the HMS model consists of 15 sub-basins (see panel d in Fig. 1) characterized by heterogeneous geomorphological properties (Table 2). The contributing areas span from 3.5 km² (W200) to 34.5 km² (W220), while slope values are in the large range: 1.0% (W160)–22.9% (W240), reflecting the watershed characteristics shown in Fig. 1 (panel c). In particular, the watershed case study includes a mountainous area in the upper part and a flat area near the outlet. This is also confirmed by outlet elevations that vary from 51 m (W160) to 347 m (W300). The land use suggests a limited variability of CN values in the range 84.8 (W240)–

Table 2 Main hydro-morphological properties of the fifteen sub-basins in the case study

Sub-basin	Watershed area [km²]	Average slope [%]	Curve number [-]	Mean elevation [m]	Minimum elevation [m]	Outlet flow length [km]	Concentration time [min]
W160	4.4	1.0	88.6	61	51	0	253
W170	6.9	2.4	92.0	82	54	3	164
W180	5.2	3.8	92.0	96	54	3.1	134
W190	11.0	22.8	86.6	203	95	10.6	78
W200	3.5	15.1	91.1	175	95	11	53
W210	6.6	21.3	86.4	195	118	12.9	63
W220	34.5	19.8	90.1	303	118	13.1	131
W230	4.7	10.8	88.2	195	150	17.9	56
W240	7.1	22.9	84.8	250	150	18	60
W250	33.8	19.7	91.7	427	175	21.8	117
W260	19.2	19.8	91.7	419	175	21.5	81
W270	2.2	22.0	92.0	424	347	31	21
W280	7.1	23.6	92.0	550	347	31.1	37
W290	18.7	22.1	92.0	640	347	32.6	67
W300	15.7	20.9	92.0	645	347	32.7	71

Table 3 Main summary statistics of the simulated runoff time series [m³/s]

	Mean	SD	Min	Max	Median	P0.75	P0.9	P0.99	P0.999
Outlet	5.63	20.79	0	351.68	0	1.3	12.81	103.17	252.3
W160	0.14	0.64	0	18.91	0	0	0.19	3.08	7.96
W170	0.21	1.02	0	30.09	0	0	0.3	4.87	12.6
W180	0.16	0.78	0	23.12	0	0	0.21	3.69	10
W190	0.34	1.67	0	49.31	0	0	0.44	7.87	21.33
W200	0.11	0.51	0	14.83	0	0	0.16	2.45	6.31
W210	0.21	0.94	0	25.86	0	0	0.32	4.6	11.7
W220	1.07	5.16	0	152.41	0	0	1.44	24.58	65.52
W230	0.15	0.7	0	20.78	0	0	0.19	3.33	9.03
W240	0.22	0.98	0	26.19	0	0	0.36	4.84	12.23
W250	1.06	5.04	0	148.99	0	0	1.45	24.03	63.65
W260	0.6	2.9	0	85.57	0	0	0.78	13.68	37.06
W270	0.07	0.32	0	9.41	0	0	0.1	1.54	3.96
W280	0.22	1	0	26.35	0	0	0.37	4.89	12.38
W290	0.58	2.56	0	64.65	0	0.01	0.98	12.77	31.15
W300	0.49	2.04	0	44.81	0	0.04	0.93	10.24	24.24

SD is standard deviation; P0.x is the percentile at 75%, 90%, 99%, 99,9%

92 (for six sub-basins), defined in the Antecedent Moisture Condition (AMC) II, characterizing a soil in a moderate humidity condition. The hydrologic synthetic scenario is simulated applying the HEC-HMS model on the three years of rainfall observations at 1-h resolution, generating 15 discharge time series at the same time resolution in the outlet sub-basins and in the watershed outlet (hereinafter Outlet). An overview of the considered scenario is provided in Table 3 and Fig. 3. In particular, Table 3 reports the main summary statistics. The time series distributions of flow discharge signals are positively skewed due to the large proportion of zero values and exhibit sharp peaks. Note that summary statistics reflect the typical hydrological behavior of small sub-basins with low concentration times and high CN values. In fact, the discharge median value is zero and quantile values confirm the high time series intermittency. Figure 3 displays the individual flow discharge signals of the 15 sub-basins along with the flow discharge at the catchment outlet. Note that since rainfall is assumed spatially homogeneous, all recorded signals show similar behaviour over the considered time interval.

4.2 Optimal ML method selection

We divide the feature-output data into 80% training and 20% testing. All features are normalized, i.e., $0 \leq X_j \leq 1$ ($j = 1, ..., 15$). We use the following R-packages: `glmnet`, `randomForest`, `gbm`, `nnet` (Friedman et al. 2009; Liaw and Wiener 2002; Ridgeway 2007; Ripley et al. 2016) and `caret` (Kuhn 2009) to perform

hyperparameter optimization. After training the models, we obtain the following values of the hyperparameters:

- **Ridge regression**: *lambda* = 0.001;
- **Random Forest**: *mtry* = 15 and *n.trees* = 500;
- **Gradient Boosting**: *shrinkage* = 0.071, *n.trees* = 951, *interaction.depth* = 7, *n.minobsinnode* = 10 and *bag.fraction* = 0.65;
- **Neural Network**: *size* = 12 and *decay* = 0.1;

Note that in the Random Forest model, all features are used in each tree (*mtry* = 15). Hence, it can be regarded as a Bagging model (Breiman 1996). In Table 4 the estimates of the performance measures of the ML models are reported. Random Forest is the best performing model according to all three measures compared to all other models. Consequently, we select such ML model to carry out the discharge forecasting analysis.

Note that the results illustrated here and in the next sections refer to the case of lag equal to zero. In such a case, the machine learning tool and the measure importance (later described) investigate on the dependence among simultaneous flow discharge signals of the 15 sub-basins and the flow discharge at the outlet. For offering a more complete overview of the hydrological response the case of lag = 3 is reported in the Appendix. For such time response the results are in line with results for the lag = 0 case study.

4.3 Importance analysis

We recall that the first four feature importance measures (cPFI, SPFI, ALE-IMP, T′), reported in Table 1, are
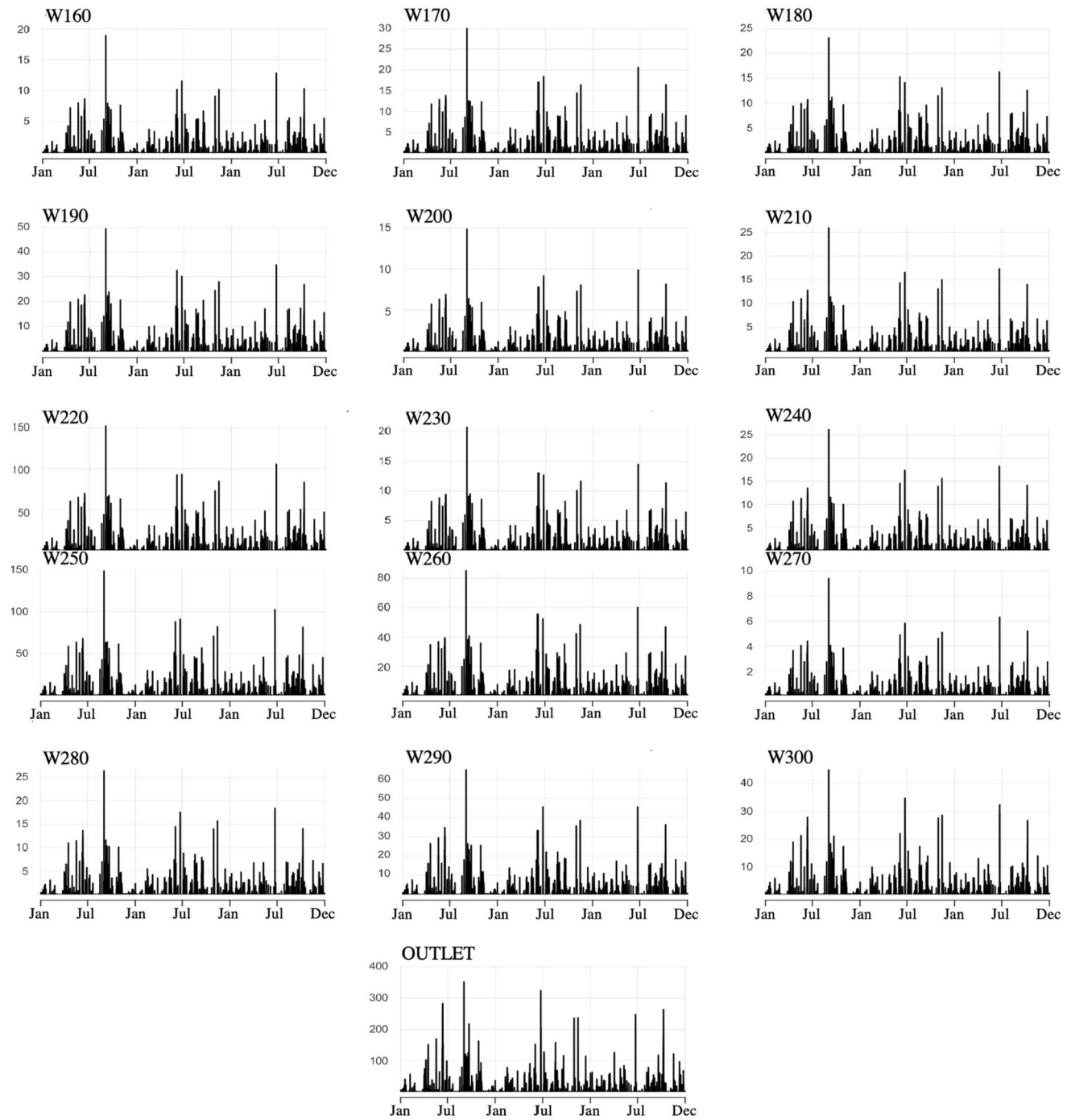
Fig. 3 The hydrologic synthetic scenario. Each plot displays the simulated runoff hourly time series. y-axis dimension $[m^3/s]$

Table 4 Performance measures estimated for the four ML models

Performance measures	Ridge regression	Random forest	Gradient boosting	Neural network
MAE (10^{-4})	100	55	59	86
RMSE (10^{-3})	25	21	23	25
R^2 (10^{-2})	87	89	86	82
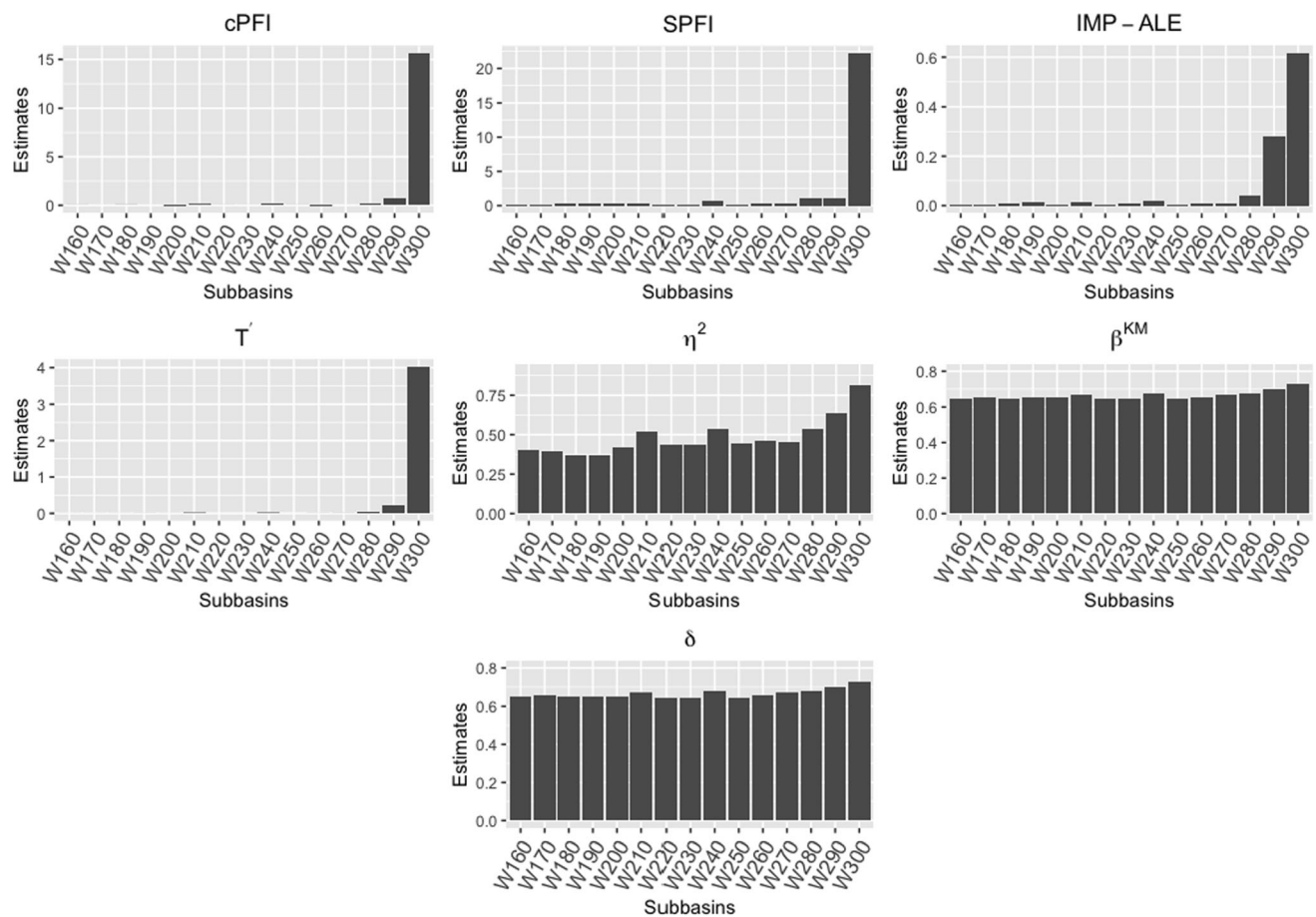
Fig. 4 Estimates of seven feature importance measure used in the case study

computed using the predictions of the optimal ML model and the remaining feature importance measures (η^2, β^{KS} and δ) are evaluated directly from the data. For the computation of the sensitivity measures we use betaKS3.m[1].

The conditional PFI measure is calculated using the algorithmic implementation of Debeer and Strobl (2020). Both performance-based measures (cPFI and SPFI) are computed using RMSE as a loss function. We use the R-packages permimp (Debeer et al. 2021) and featureImportance[2]. The variance-based measures (ALE-IMP and T′) are computed partitioning the support of the feature of interest into 100 equally-spaced intervals ($K = 100$). The ALE-IMP measure is calculated using the algorithmic implementation proposed by Christensen et al. (2021). For both measures we use the R-package ALE-Plot (Apley 2018).

Figure 4 displays the estimates of the feature importance measures used in the case study. The results of the ML feature importance measures show that only a few sub-basins are influential for forecasting the watershed outlet

discharge. Differently, the global SA indices assign a considerable importance to all sub-basins which is due to the presence of a strong correlation between sub-basins. This shows that all of them are active in the watershed dynamics.

From our analysis we have that some estimates of conditional PFI are close to zero. This means that permuting X_j does not produce a reduction in the performance of the RF model. Then, such feature has no impact on the predictive performance of the ML model. Therefore, the corresponding sub-basin might be unnecessary. Differently, a high cPFI value denotes that the sub-basin is important in the ML model. In order to have a better understanding of the results presented in Fig. 4, we provide the ranking for each feature importance measure and the mean ranking resulting from the ensemble of the importance measures used (Table 5). The latter is defined as the average ranking resulting from the ensemble of the importance measures used (Kuncheva 2014).

The results in Fig. 4 and Table 5 suggest that we can identify three groups of sub-basins based on their importance. The first group consists of sub-basins W300, W290 and W280. Note that the seven feature importance

---

[1] https://zenodo.org/record/885332#.XgoB-kdKiUk.

[2] https://github.com/giuseppec/featureImportance.

Table 5 Ranking for each feature importance measure and the mean ranking

Sub-basin	cPFI	SPFI	ALE-IMP	T'	η^2	β^{KS}	δ	Mean ranking
W300	1	1	1	1	1	1	1	*1*
W290	2	2	2	2	2	2	2	*2*
W280	4	3	3	3	3	3	3	*3*
W240	3	4	4	5	4	4	4	*4*
W210	5	8	5	4	5	6	6	*5*
W270	8	5	11	6	7	5	5	*6*
W260	14	7	8	7	6	7	7	**7**
W190	9	6	6	10	15	9	9	**8**
W180	6	9	7	11	14	12	12	**9**
W200	15	10	12	8	11	10	10	**10**
W170	7	11	13	13	13	8	8	**11**
W230	10	12	9	12	10	13	13	**12**
W160	13	13	10	9	12	11	11	**13**
W220	11	14	14	14	8	15	15	**14**
W250	12	15	15	15	9	14	14	**15**

Table 6 Estimates of the performance measures for the configurations defined using the mean ranking

Configuration	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MAE (10^{-4})	110	61	58	58	58	56	56	56	56	56	56	56	56	56	56
RMSE (10^{-3})	37	24	22	22	22	21	21	21	21	21	21	21	21	21	21
R^2 (10^{-2})	69	86	88	89	89	90	89	89	89	89	89	89	89	89	89

measures defined on different aspects (i.e. the predictive accuracy of the optimal ML model, the individual and total contribution to the output variance and the probabilistic effect on the output response) simultaneously identify W300, W290 and W280 as the most influential sub-basins. The second group consists of sub-basins W240, W210 and W270. Note that, almost all feature importance measures identify W240 and W210 as the fourth and fifth most important sub-basins. While the ranking of W270 varies across the importance measures. Note that there is a third group of sub-basins for which the estimates of all importance measures are generally much lower than the estimates of the first two classes, showing that such sub-basins are less (or not) influential in predicting the catchment outlet discharge. Interestingly, employing ML and SA feature importance measures one can obtain rankings that are in agreement with each other. Such correspondence produces more confidence about which sub-basins are important for forecasting the flow discharge at the catchment outlet.

In order to increase our confidence on the ranking reported in the last column in Table 5, we investigate the predictive accuracy of the optimal ML model fitting an incremental sequence of *Model Configurations* built by including one sub-basin at a time. The order of inclusion follows the ranking resulting from the importance analysis. To be more precise, the sequence of *Model Configurations*

is initialised including only the first ranked sub-basin (W300). Then, *Configuration 2* includes sub-basins W300 and W290; *Configuration 3* includes sub-basins W300, W290, and W280 and, finally *Configuration 15* includes all sub-basins. For each configuration we train a Random Forest model and evaluate the performance measures presented in Sect. 3.3. Based on predictive performances, we aim to identify how many sub-basins we need to include in the optimal ML model to achieve a desired high level of accuracy.

The results reported in Table 6 suggest that the first group of sub-basins (which includes the three most important ones) explains 88% of the variability of the output response. Including the second group produces only a slight improvement in the performance measures. Table 6 also shows that including the least relevant sub-basins does not improve accuracy further. Therefore, they can be excluded from the machine learning analysis.

Conversely, if we were to include only the non-relevant sub-basins, we would obtain the following values of the performance measures: MAE = 0.0164, RMSE = 0.0501 and R^2 = 0.2748. These values confirm that if we were to train the model using only the least relevant sub-basins as inputs, we would not achieve a desirable prediction accuracy.

4.4 Discussion

Let us now come to the questions posed in the introduction. Regarding the first question, feature importance measures have allowed us to identify the group of sub-basins that influence the catchment-scale hydrological response the most.

Regarding the second question, the discussion is a bit more elaborate and we focus on: (a) the watershed and the hydrological model characteristics shown in Table 3 and (b) the insights arising from the ranking of the importance analysis (Table 5). In particular, the sub-basin contributing areas do not allow us to distinguish the role of the sub-basins. In fact, the largest sub-basins (W220 and W250) are included in the uninfluential group (bold group in Table 5) and, interestingly, their contributing areas are twice those of the dominant sub-basins. Differently, slope values are more consistent with the importance ranking. Indeed, all six influential sub-basins are characterized by slope values higher than 20%. However, high values are observed also for W190, W220, W250, W260, which belong to the uninfluential group. The Curve Number is almost homogeneous among the sub-basins and it does not appear to be a distinguishing characteristic. Note that, although the dominant sub-basins have the highest CN values, the same value is also observed for W170 and W180 (bold group). Moreover, the lowest value (84.8) is registered for W240 which is in the italic group. The Average Elevation is also in partial agreement with the importance ranking. In particular, the dominant sub-basins present the highest values, nevertheless high values also characterize W220 and W260 (bold group). Conversely, we register an agreement between Minimum Elevation and the sub-basins ranking. In fact, the first three ranked sub-basins are characterized by the highest minimum elevation. High outlet elevation indicates that these three sub-basins are located in the upper part of the watershed, as confirmed by the values of the hydraulic distance to the watershed outlet listed in the sixth column of Table 2. The last comparison involves the concentration time parameter (Tc). This is estimated using several empirical equations which include the slope, the drainage network length, the contributing areas and the CN values. Such parameter offers a combination of the previously described topographic properties. Tc is responsible for the UH shape and then for the sub-basin response function: small Tc values refer to concentrated response functions while larger values refer to more spread functions. Comparing the Tc parameter with the feature importance measure ranking, one notes a good overall agreement, with all influential sub-basins having low Tc values.

In conclusion, even if the results do not suggest clear agreement between watershed ranking and specific hydromorphological characteristics, useful for answering the second paper question, it is possible to make some reasonable hypotheses. The dominant role of sub-basins W300, W290, and W280 is not surprising since (a) the watershed dimension is above the average and (b) they are located upstream and therefore they influence the downstream watersheds. Indeed, the outlet flow length shows the maximum values. Moreover, the sub-basin W260, characterized by the same distance to the outlet, is ranked in the italic group in Table 5. So, the contributing area and the upstream location could be relevant characteristics for discriminating the role of the sub-basins.

However, making hypotheses for the other two sub-basins located in the italic group in Table 5 (W240 and W210) is more challenging. In this case, the time of concentration could be the prominent concomitant characteristic, indeed, for both sub-basins it is very low due to the steep slopes, therefore, the more concentrated hydrological response could make their contribution more influential.

In order to properly answer the second question of the paper, a more descriptive modelling approach should be applied, as the simplified hydrological model scenario was only used here to investigate the potential of the importance measure approach. In future research, a fully distributed hydrological model will be applied to a large basin (< 5000 km^2), calibrating it with observed data and referring to very long synthetic rainfall scenarios (1000 years at 15 minutes temporal resolution). Such realistic and large case study will allow to investigate on the watershed role at different spatial scale shedding the light on the preliminary results here showed.

5 Conclusions

This work has investigated the use of feature importance measures in hydrology and, specifically, it provides some preliminary results on their use in dissecting the role of sub-basins in hydrological response.

Our goal, partially reached with the simplified proof of concept here presented, has been to verify: (a) whether such measures are able to identify sub-basins that contribute more than others to the outlet flow discharge and (b) whether such sub-basins exhibit distinctive morpho-hydrological characteristics that influence the feature importance analysis. We use a well-known hydrological model (HEC-HMS) to simulate flow discharge signals of the sub-basins along with the flow discharge at the catchment outlet in a watershed located in Italy. For this synthetic scenario, we have applied seven feature importance measures, three of them for the first time

in hydrology, from the machine learning and the global sensitivity analysis framework. The importance analysis allows us to identify 3 sub-basins as highly influential, 3 as moderately influential and 9 as uninfluential. The role of the three "dominant" sub-basins is confirmed and quantified comparing their prediction performances to the whole set of 15 sub-basins resulting in explaining the 88% of the variability of the output response. While the case study application is able to distinguish the sub-basins role, as expected, it only partially contributes to identify the factors that characterize influential sub-basins. Indeed, given the complex nature of the hydrological response, goal (b) is particularly challenging and difficult to reach with a simplified model. Comparing the resulting ranking to some morpho-hydrological properties we can only note that a combination of slope, CN, distance from the outlet and concentration time plays a prominent role for predicting the catchment outlet discharge. Surprisingly, the contributing area has a marginal role compared to the above mentioned parameters.

Overall, our study demonstrates that feature importance measures have a great potential for investigating the sub-basin role, thus positively contributing to a variety of possible investigation and applications: selecting "dominant" sub-basins for designing Early Warning Systems (based on discharge), selecting sub-basins where installing instrumentations, setting automatic procedures for sub-basin selection in semi-distributed models, calibrating machine learning tools, and offering another perspective to answer the theoretical question concerning the distinctive morpho-hydrological characteristics of sub-basins. A future research objective will include more complex hydrological modelling and simulation for supporting in a more general context the final goal here presented.

Appendices
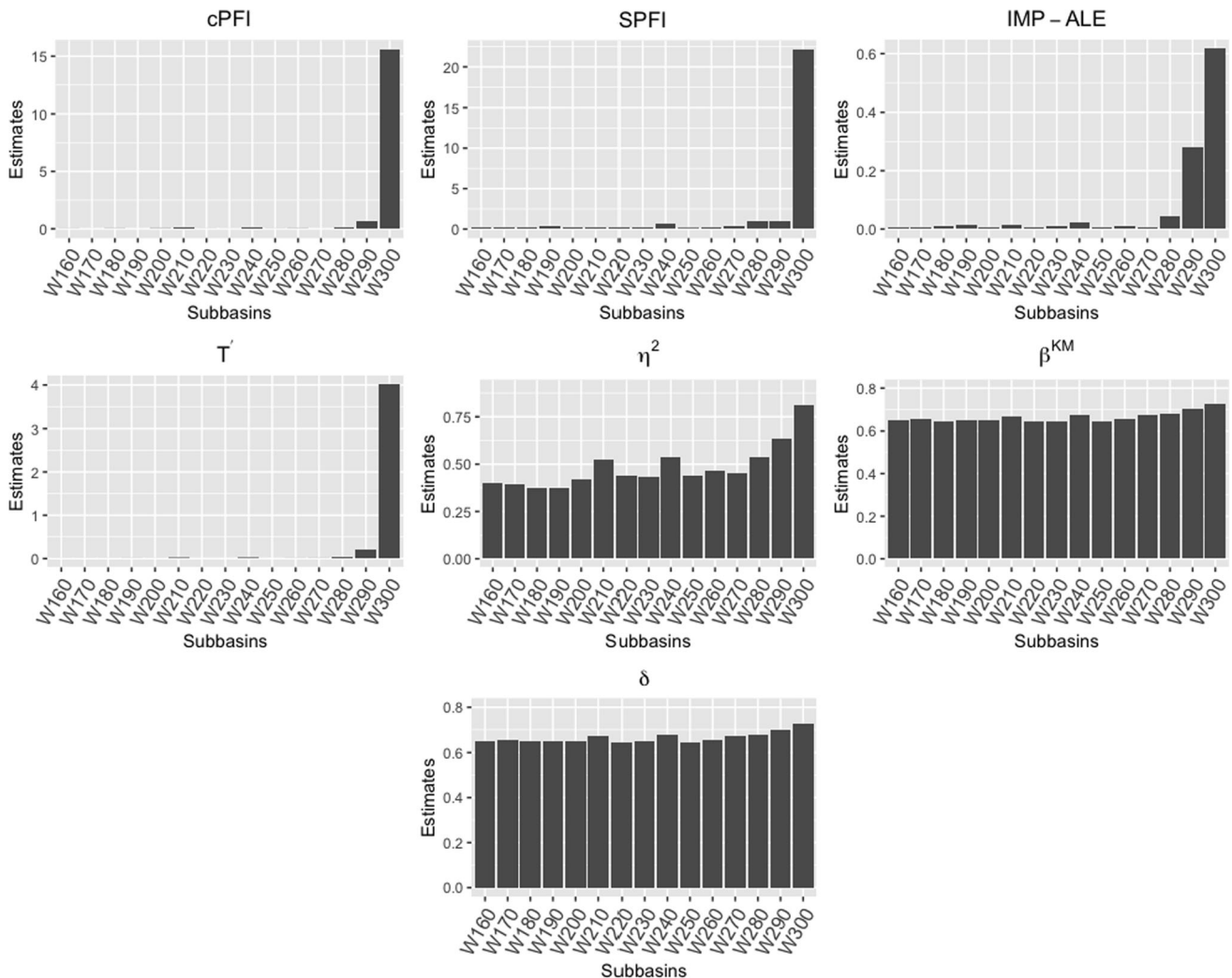
See Fig. 5 and Tables 7, 8, 9.

Fig. 5 Estimates of seven feature importance measure used in the case study for three hours time response (lag= 3)

Table 7 Performance measures estimated for the four ML models for three hours time response (lag = 3)

Performance measures	Ridge regression	Random forest	Gradient boosting	Neural network
MAE (10^{-4})	57	54	52	64
RMSE (10^{-3})	16	17	17	17
R^2 (10^{-2})	95	95	95	94

Table 8 Ranking for each feature importance measure and the mean ranking for three hours time response (lag = 3)

Sub-basin	cPFI	SPFI	ALE-IMP	T'	η^2	β^{KS}	δ	Mean Ranking
W300	1	1	1	1	1	1	1	*1*
W290	2	5	2	2	2	2	2	*2*
W280	3	3	3	3	4	3	3	*3*
W240	4	4	4	6	3	4	4	*4*
W210	5	2	5	5	5	5	6	*5*
W270	10	6	12	4	7	6	5	*6*
W160	9	8	14	8	8	9	11	**7**
W200	12	7	15	10	6	8	10	**8**
W260	14	10	9	7	12	10	7	**9**
W170	11	9	13	13	9	8	8	**10**
W190	8	12	6	9	14	12	10	**11**
W220	6	14	10	14	11	13	14	**12**
W250	7	15	11	15	10	12	15	**13**
W230	13	13	8	12	13	14	12	**14**
W180	15	11	7	11	15	15	13	**15**

Table 9 Estimates of the performance measures for the configurations defined using the mean ranking for three hours time response (lag = 3)

Configuration	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MAE (10^{-4})	65	62	60	60	57	54	52	52	52	52	52	52	52	52	52
RMSE (10^{-3})	21	20	19	19	18	18	18	18	18	18	18	18	18	18	18
R^2 (10^{-2})	92	93	94	94	94	95	95	95	95	95	95	95	95	95	95

References

Agrawal T (2021) Hyperparameter optimization in machine learning: make your machine learning and deep learning models more efficient. Apress

Ali G, Oswald CJ, Spence C et al. (2013) Towards a unified threshold-based hydrological theory: necessary components and recurring challenges. Hydrol Process 27(2):313–318

Apley D (2018) Aleplot: accumulated local effects (ale) plots and partial dependence (pd) plots. R package version 1

Apley D, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Ser B Stat Methodol 82:1059–1086

Asano Y, Uchida T, Tomomura M (2020) A novel method of quantifying catchment-wide average peak propagation speed in hillslopes: fast hillslope responses are detected during annual floods in a steep humid catchment. Water Resour Res 56(1):e2019WR025,070

Beiter D, Weiler M, Blume T (2020) Characterising hillslope-stream connectivity with a joint event analysis of stream and groundwater levels. Hydrol Earth Syst Sci 24(12):5713–5744

Bergstrom A, Jencso K, McGlynn B (2016) Spatiotemporal processes that contribute to hydrologic exchange between hillslopes, valley bottoms, and streams. Water Resour Res 52(6):4628–4645

Betson RP (1964) What is watershed runoff? J Geophys Res 69(8):1541–1552

Blöschl G, Bierkens MF, Chambel A et al (2019) Twenty-three unsolved problems in hydrology (uph)-a community perspective. Hydrol Sci J 64(10):1141–1158

Bonell M (1998) Selected challenges in runoff generation research in forests from the hillslope to headwater drainage basin scale 1. JAWRA J Am Water Resour Assoc 34(4):765–785

Borgonovo E (2007) A new uncertainty importance measure. Reliab Eng Syst Saf 92(6):771–784

Borgonovo E, Plischke E (2016) Sensitivity analysis: a review of recent advances. Eur J Oper Res 248(3):869–887

Borgonovo E, Tarantola S, Plischke E et al (2014) Transformations and invariance in the sensitivity analysis of computer experiments. J R Statist Soc Ser B (Statist Methodol) 76(5):925–947

Borgonovo E, Lu X, Plischke E et al (2017) Making the most out of a hydrological model data set: Sensitivity analyses to open the model black-box. Water Resour Res 53(9):7933–7950

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16(3):199–231

Butler D (2014) Earth observation enters next phase. Nature 508(7495):160–161

Candes E, Fan Y, Janson L et al (2018) Panning for gold: model-x knockoffs for high dimensional controlled variable selection. J R Statist Soc Ser B (Statist Methodol) 80(3):551–577

Casalicchio G, Molnar C, Bischl B (2018) Visualizing the feature importance for black box models. In: Joint European conference on machine learning and knowledge discovery in databases, Springer, pp 655–670

Castellarin A, Merz R, Blöschl G (2009) Probabilistic envelope curves for extreme rainfall events. J Hydrol 378(3–4):263–271

Chen L, Wang L (2018) Recent advance in earth observation big data for hydrology. Big Earth Data 2(1):86–107

Chen Y, Han D (2016) Big data and hydroinformatics. J Hydroinf 18(4):599–614

Christensen K, Siggaard M, Veliyev B (2021) A machine learning approach to volatility forecasting. Available at SSRN

Chu X, Steinman A (2009) Event and continuous hydrologic modeling with HEC-HMS. J Irrig Drain Eng 135(1):119–124

Clark MP, Slater AG, Rupp DE, Vrugt JA, Gupta HV, Wagener T, Hay LE (2008) Framework for understanding structural errors (fuse): A modular framework to diagnose differences between hydrological models. Water Resour Res, 44(12). https://doi.org/10.1029/2007wr006735

De Silva M, Weerakoon S, Herath S (2014) Modeling of event and continuous flow hydrographs with HEC-HMS: case study in the Kelani river basin, Sri Lanka. J Hydrol Eng 19(4):800–806

Debeer D, Strobl C (2020) Conditional permutation importance revisited. BMC Bioinformat 21(1):1–30

Debeer D, Hothorn T, Strobl C (2021) Permimp: conditional permutation importance. In (Version 1.0-2) [R package]. https://CRAN.R-project.org/package=permimp

Demand D, Blume T, Weiler M (2019) Spatio-temporal relevance and controls of preferential flow at the landscape scale. Hydrol Earth Syst Sci 23(11):4869–4889

Desai S, Ouarda TB (2021) Regional hydrological frequency analysis at ungauged sites with random forest regression. J Hydrol 594(125):861

Detty JM, McGuire KJ (2010) Threshold changes in storm runoff generation at a till-mantled headwater catchment. Water Resour Res 46, W07525. https://doi.org/10.1029/2009WR008102

Di Prinzio M, Castellarin A, Toth E (2011) Data-driven catchment classification: application to the pub problem. Hydrol Earth Syst Sci 15(6):1921–1935

Dreiseitl S, Ohno-Machado L (2002) Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 35(5–6):352–359

Farrar DE, Glauber RR (1967) Multicollinearity in Regression analysis: the problem revisited. Rev Econ Stat 49(1):92–107. https://doi.org/10.2307/1937887

Fienen MN, Nolan BT, Kauffman LJ et al (2018) Metamodeling for groundwater age forecasting in the lake Michigan basin. Water Resour Res 54(7):4750–4766

Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177):1–81

Friedman J, Hastie T, Tibshirani R et al (2009) glmnet: lasso and elastic-net regularized generalized linear models. R package version 1(4):1–24

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of statistics 29, pp 1189–1232

Gamboa F, Klein T, Lagnoux A (2018) Sensitivity analysis based on cramèr von mises distance. SIAM/ASA J Uncert Quantif 6(2):522–548

Gharib A, Davies EG (2021) A workflow to address pitfalls and challenges in applying machine learning models to hydrology. Adv Water Resour 152(103):920

Gilcrest BR (1950) Flood routing. In: Ronse H (ed) Engineering hydraulics, vol X. Wiley, New York, pp 635–710

Graham CB, McDonnell JJ (2010) Hillslope threshold response to rainfall:(2) development and use of a macroscale model. J Hydrol 393(1–2):77–93

Graham CB, Woods RA, McDonnell JJ (2010) Hillslope threshold response to rainfall:(1) a field based forensic approach. J Hydrol 393(1–2):65–76

Greenwell BM, Boehmke BC, McCarthy AJ (2018) A simple and effective model-based variable importance measure. arXiv preprint arXiv:1805.04755

Greenwell BM, Boehmke BC, Gray B (2020) Variable importance plots-an introduction to the vip package. R J 12(1):343

Gruber MH (2017) Improving efficiency by shrinkage: the James-Stein and ridge regression estimators. Routledge, London

Guastini E, Zuecco G, Errico A et al (2019) How does streamflow response vary with spatial scale? Analysis of controls in three nested alpine catchments. J Hydrol 570:705–718

Hastie T, Tibshirani R, Friedman JH et al (2009) The elements of statistical learning: data mining, inference, and prediction, vol 2. Springer, Heidelberg

Hewlett J (1974) Comments on letters relating to role of subsurface flow in generating surface runoff: 2, upstream source areas by r. allan freeze. Water Resour Res 10(3):605–607

Homma T, Saltelli A (1996) Importance measures in global sensitivity analysis of nonlinear models. Reliab Eng Syst Saf 52(1):1–17

Hooker G, Mentch L (2019) Please stop permuting features: an explanation and alternatives. arXiv e-prints pp arXiv–1905

Hopp L, McDonnell JJ (2009) Connectivity at the hillslope scale: identifying interactions between storm size, bedrock permeability, slope angle and soil depth. J Hydrol 376(3–4):378–391

Iman RL, Hora SC (1990) A robust measure of uncertainty importance for use in fault tree system analysis. Risk Anal 10(3):401–406

Iwasaki K, Katsuyama M, Tani M (2020) Factors affecting dominant peak-flow runoff-generation mechanisms among five neighbouring granitic headwater catchments. Hydrol Process 34(5):1154–1166

Jencso KG, McGlynn BL (2011) Hierarchical controls on runoff generation: topographically driven hydrologic connectivity, geology, and vegetation. Water Resour Res 47(11) Article Number: W11527. https://doi.org/10.1029/2011WR010666

Jencso KG, McGlynn BL, Gooseff MN, Wondzell, SM, Bencala, KE, Marshall LA (2009) Hydrologic connectivity between landscapes and streams: Transferring reach-and plot-scale understanding to the catchment scale. Water Resour Res 45(4). https://doi.org/10.1029/2008wr007225

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26

Kuhn M (2009) The caret package. J Stat Softw 28(5):1–26

Kuncheva LI (2014) Combining pattern classifiers: methods and algorithms. Wiley, Hoboken

Lehmann P, Hinz C, McGrath G et al (2007) Rainfall threshold for hillslope outflow: an emergent property of flow pathway connectivity. Hydrol Earth Syst Sci 11(2):1047–1063

Liaw A, Wiener M et al (2002) Classification and regression by randomforest. R News 2(3):18–22

Liu J, Engel BA, Wang Y et al (2019) Runoff response to soil moisture and micro-topographic structure on the plot scale. Sci Rep 9(1):1–13

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In NeurIPS, pages 4768–4777

McGuire KJ, McDonnell JJ (2010) Hydrological connectivity of hillslopes and streams: characteristic time scales and nonlinearities. Water Resour Res 46:W10543. https://doi.org/10.1029/2010WR009341

Mockus V (1964) Letter from victor mockus to orrin ferris. US Department of Agriculture Soil Conservation Service, Lanham, MD, USA

Molnar C (2022) Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Christophm.github.io/interpretable-ml-book/

Mourato S, Fernandez P, Marques F et al (2021) An interactive web-gis fluvial flood forecast and alert system in operation in portugal. Int J Disaster Risk Reduct 58(102):201

Papacharalampous G, Tyralis H, Papalexiou SM et al (2021) Global-scale massive feature extraction from monthly hydroclimatic time series: statistical characterizations, spatial patterns and hydrological similarity. Sci Total Environ 767(144):612

Pearson K (1905) On the general theory of skew correlation and non-linear regression, vol 14. Dulau and Company, London

Plischke E, Borgonovo E, Smith CL (2013) Global sensitivity measures from given data. Eur J Oper Res 226(3):536–550

Rajaee T, Khani S, Ravansalar M (2020) Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: a review. Chemom Intell Lab Syst 200(103):978

Ramly S, Tahir W (2016) Application of HEC-GeoHMS and HEC-HMS as rainfall–runoff model for flood simulation. In: ISFRAM 2015. Springer, Singapore, pp 181–192

Ramly S, Tahir W, Abdullah J et al (2020) Flood estimation for smart control operation using integrated radar rainfall input with the HEC-HMS model. Water Resour Manage 34(10):3113–3127

Renyi A (1959) On measures of statistical dependence. Acta Math Acad Sci Hungarica 10:441–451

Ribeiro MT, Singh S, Guestrin C (2016) "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144

Ridgeway G (2007) Generalized boosted models: a guide to the GBM package. Update 1(1):2007

Ripley B, Venables W, Ripley MB (2016) Package nnet. R package version 7(3–12):700

Saltelli A (2002) Making best use of model evaluations to compute sensitivity indices. Comput Phys Commun 145(2):280–297

Saltelli A, Tarantola S, Campolongo F et al (2004) Sensitivity analysis in practice: a guide to assessing scientific models, vol 1. Wiley Online Library, New York

Saltelli A, Ratto M, Andres T et al (2008) Global sensitivity analysis - the primer. Wiley, Chichester

Scaife CI, Band LE (2017) Nonstationarity in threshold response of stormflow in southern appalachian headwater catchments. Water Resour Res 53(8):6579–6596

Schmidt L, Heße F, Attinger S et al (2020) Challenges in applying machine learning models for hydrological inference: a case study for flooding events across germany. Water Resour Res 56(5):e2019WR025,924

Shapley LS (1953) A value for n-Person Games. Study 28. Princeton University Press, Annals of Mathematics Studies, pp 307–317

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J Roy Stat Soc Ser B (Methodol) 36(2):111–133

Strobl C, Boulesteix AL, Zeileis A et al (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform 8(1):1–21

Subagyono K, Tanaka T, Hamada Y et al (2005) Defining hydrochemical evolution of streamflow through flowpath dynamics in Kawakami headwater catchment, central Japan. Hydrol Process Int J 19(10):1939–1965

Sun AY, Scanlon BR (2019) How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environ Res Lett 14(7):073,001

Tauro F, Selker J, Van De Giesen N et al (2018) Measurements and observations in the xxi century (moxxi): innovation and multi-disciplinarity to sense the hydrological cycle. Hydrol Sci J 63(2):169–196

Teweldebrhan AT, Schuler TV, Burkhart JF et al (2020) Coupled machine learning and the limits of acceptability approach applied in parameter identification for a distributed hydrological model. Hydrol Earth Syst Sci 24(9):4641–4658

Thorslund J, Bierkens MF, Oude Essink GH et al (2021) Common irrigation drivers of freshwater salinisation in river basins worldwide. Nat Commun 12(1):1–13

Tyralis H, Papacharalampous G, Langousis A (2021) Super ensemble learning for daily streamflow forecasting: large-scale demonstration and comparison with multiple machine learning algorithms. Neural Comput Appl 33(8):3053–3068

Uchida T, Tromp-van Meerveld I, McDonnell JJ (2005) The role of lateral pipe flow in hillslope runoff response: an intercomparison of non-linear hillslope response. J Hydrol 311(1–4):117–133

Zehe E, Becker R, Bárdossy A et al (2005) Uncertainty of simulated catchment runoff response in the presence of threshold processes: role of initial soil moisture and precipitation. J Hydrol 315(1–4):183–202

Zounemat-Kermani M, Batelaan O, Fadaee M et al (2021) Ensemble machine learning paradigms in hydrology: a review. J Hydrol 598(126):266