Technical note: Machine learning metamodelling for global sensitivity analysis

Patricio Yeste[1,2,3], Lieke A. Melsen[4], João Paulo L.F. Brêda[5], Nicolás Tacoronte[3], Andrea Saltelli[6,7], Giulia Vannucci[8], Roberta Siciliano[8], and Axel Bronstert[1]

[1]Institute of Environmental Science and Geography, University of Potsdam, Karl-Liebknecht-Straße 24–25, 14476 Potsdam, Germany
[2]GFZ Helmholtz Centre for Geosciences, Potsdam, 14473, Germany
[3]Department of Applied Physics, University of Granada, Granada, Spain
[4]Hydrology and Environmental Hydraulics Group, Wageningen University, Wageningen, The Netherlands
[5]Instituto de Pesquisas Hidráulicas, Universidade Federal do Rio Grande do Sul (IPH/UFRGS), Porto Alegre, Brazil
[6]University Pompeu Fabra, Barcelona School of Management, Carrer de Balmes, 132, 08008, Barcelona, Spain
[7]Centre for the Study of the Sciences and the Humanities, University of Bergen, Parkveien 9, PB 7805, 5020, Bergen, Norway
[8]Department of Electrical Engineering and Information Technology, Polytechnic and Basic Sciences School, University of Naples Federico II, Via Claudio, 21, 80125 Napoli (NA), Italy

Correspondence: patricio.yeste@uni-potsdam.de

Abstract. Global sensitivity analysis (GSA) plays a central role in hydrologic modelling by supporting model understanding, diagnosis, and decision-making through the identification of influential and non-influential parameters and their interactions. Variance-based methods provide a rigorous framework for GSA but are often computationally expensive, as their estimation requires a large number of model evaluations. Metamodelling has therefore been widely adopted as a strategy to alleviate this issue, with recent advances in machine learning (ML) offering new opportunities to construct accurate and flexible surrogates for complex models. This technical note examines the practical relationship between Sobol' total-effect indices (T_i) and feature importance measures derived from ML metamodels within a hydrologic modelling context. Building on theoretical results that link T_i to permutation variable importance (PVI_i) under independence assumptions, we provide systematic numerical evidence using three hydrologic models of varying complexity (HBV, HyMod, and VIC) applied to three headwater catchments in northern Germany, together with three ML metamodels: a gradient boosting decision tree (GBDT), a neural network (NN), and a linear model (LM). The three metamodels were trained on Monte Carlo samples and used to estimate sensitivities through PVI_i and SHapley Additive exPlanations (SHAP_i). The results demonstrate that GBDT and NN metamodels reliably reproduce both the ranking and relative magnitude of T_i using PVI_i across all hydrologic models, providing clear empirical support for the theoretical connection between the two measures. In contrast, the performance of LM-based estimates depends strongly on the degree of linearity in the underlying model response. Mean absolute SHAP_i values exhibit a consistent monotonic relationship with T_i and preserve parameter rankings, while sample-specific SHAP_i values enable a distributed evaluation of sensitivities across both the parameter space and the target variable space. Overall, this study highlights ML metamodelling as a computationally efficient and conceptually sound framework for GSA in hydrologic modelling and beyond.

1 Introduction

Global sensitivity analysis (GSA) provides a systematic framework to quantify how uncertainty in model outputs can be attributed to uncertainty in model inputs, both individually and through their interactions (Saltelli et al., 2008, 2010). GSA supports model understanding, diagnosis, and decision-making by revealing dominant controls, identifying influential and non-influential parameters, and highlighting the role of interactions in complex systems. These properties make GSA particularly valuable for hydrologic modelling, where models are often nonlinear, high-dimensional, and affected by substantial uncertainty. In this context, GSA enables a sensitivity analysis based on different settings, such as the prioritization of influential input factors to inform data collection and parameter estimation efforts, or for the identification of non-influential factors that can be fixed to reduce model complexity (Saltelli and Tarantola, 2002). Sensitivity analysis facilitates the diagnosis and comparison of alternative model structures, hypotheses, and assumptions (Razavi et al., 2021; Sheikholeslami et al., 2021).

Among GSA approaches, variance-based methods provide a formal way to quantify how uncertainty in model outputs can be attributed to uncertainty in individual input factors and their interactions. Within this class of methods, Sobol' indices describe how uncertainty in each input propagates through the model, with first-order indices measuring additive effects and total-effect indices accounting for the overall contribution of each factor, including higher-order interactions (Saltelli et al., 2010). Variance-based sensitivity analysis has been widely adopted across scientific disciplines and is commonly used as a reference for GSA (Lo Piano et al., 2021; Puy et al., 2022). A limitation of variance-based approaches is that they rely on the assumption that the variance of the model output is an adequate proxy for its uncertainty, which has motivated the development of moment-independent importance measures (Borgonovo et al., 2011). In hydrologic modelling, several studies have also highlighted important challenges in applying variance-based approaches, particularly when assessing sensitivity with respect to performance metrics. In such cases, global aggregation can obscure varying behaviours across different levels of model performance, and sensitivity estimates may vary substantially depending on the choice of performance metric and the defined parameter boundaries (Razavi and Gupta, 2015, 2016). These considerations motivate a careful and critical use of variance-based sensitivity analysis in hydrologic applications, while reinforcing its role as a rigorous and informative sensitivity framework. While these challenges are particularly evident in hydrologic modelling, they are not exclusive to it and extend to complex models in general.

A major practical limitation of variance-based GSA is its computational cost. Accurate estimation of total-effect indices typically requires a large number of model evaluations, scaling with both the number of uncertain inputs and the sample size required for convergence (Saltelli et al., 2010; Puy et al., 2022). For computationally expensive or high-dimensional models, these costs can become prohibitive, revealing the paradox of GSA: the models that would benefit most from GSA are typically those for which a full analysis is computationally infeasible. To alleviate this issue, metamodelling has long been proposed as a strategy to replace the original model with a computationally efficient approximation, also referred to as surrogate modelling (Saltelli et al., 2008; Razavi et al., 2012). Metamodels can enable GSA at a fraction of the original cost, but their reliability in yielding accurate sensitivity estimates critically depends on the ability of the surrogate to faithfully reproduce the input–output relationship of the original model.

Recent advances in machine learning (ML) offer new opportunities for metamodelling in the context of GSA. ML algorithms are particularly well suited to approximate complex, nonlinear relationships in high-dimensional spaces, making them attractive candidates for metamodelling. In a comprehensive review, Antoniadis et al. (2021) discussed the use of random forests for GSA and highlighted the close conceptual link between feature importance measures and sensitivity indices. In particular, for a given ML algorithm, permutation variable importance quantifies the increase in prediction error induced by breaking the association between an input and the model output, a mechanism that is closely related to the notion of variance-based total-effect sensitivity. This connection suggests that ML-based feature importance measures may provide an alternative route to estimate sensitivities through metamodelling, provided that the surrogate accurately reproduces the behaviour of the original model.

The objective of this technical note is to examine the practical relationship between variance-based total-effect sensitivity indices (Homma and Saltelli, 1996) and feature importance measures derived from ML metamodels within a hydrologic modelling context. Building on theoretical results that link total-effect indices and permutation variable importance under the assumption that the input factors have independent probability distributions, we provide systematic numerical evidence using three hydrologic models with varying structural complexity. We assess the ability of different ML metamodels to reproduce variance-based sensitivities, analyse the conditions under which this approximation is reliable, and explore alternative sensitivity measures enabled by ML surrogates. By doing so, this work aims to clarify the role of ML metamodelling as a computationally efficient and conceptually sound framework for GSA in hydrologic modelling.

Before presenting the specific objectives and details of the numerical experiment, the next section introduces the mathematical preliminaries that underpin the rationale of this work. We present the key concepts of variance-based sensitivity analysis and permutation variable importance, complemented by intuitive explanations to facilitate interpretation. While this section relies on probabilistic and statistical reasoning that may be challenging for readers less familiar with these topics, it is included to ensure clarity and transparency and to establish a solid foundation for the methodological developments and numerical results that follow.

2 Mathematical preliminaries

This section presents the mathematical preliminaries underlying this work. Sections 2.1 and 2.3 introduce the theoretical definitions of variance-based sensitivity indices and permutation variable importance, respectively, while sections 2.2 and 2.4 describe their practical estimation. This distinction between theoretical definition and practical estimation is reflected in the notation: quantities without a hat denote theoretical measures, whereas their estimators are denoted with a hat.

2.1 Variance-based sensitivity analysis

Given a deterministic model of the form $Y = f(\mathbf{X})$, where Y is a scalar model output and $\mathbf{X} = (X_1, \ldots, X_p)$ is a vector of p uncertain input factors, variance-based sensitivity analysis aims to quantify how the uncertainty in Y can be attributed to the uncertainties in the individual input factors and their interactions.

The first-order Sobol' index of the i-th factor, X_i, is defined as (e.g., Saltelli et al., 2008, 2010):

$$S_i = \frac{\mathrm{Var}_{X_i}\left(\mathbb{E}_{\mathbf{X}_{\sim i}}[Y \mid X_i]\right)}{\mathrm{Var}(Y)} \tag{1}$$

where $\mathrm{Var}(.)$ and $\mathbb{E}[.]$ are the variance and expectation operators, respectively, and $\mathbf{X}_{\sim i}$ denotes the vector of all input factors except X_i.

S_i is normalized between 0 and 1, with larger values indicating a stronger additive effect of X_i on the model output. The inner expectation in $\mathrm{Var}_{X_i}\left(\mathbb{E}_{\mathbf{X}_{\sim i}}[Y \mid X_i]\right)$ represents the conditional expectation of Y given X_i, and intuitively averages out the effect of all other factors, isolating the contribution of X_i. The outer variance measures how much this contribution varies across all possible values of X_i.

When $\sum_{i=1}^{p} S_i = 1$, the model is additive, which means that $\mathrm{Var}(Y)$ can be decomposed as the sum of all the first-order effects as there are no interactions among the input factors. However, in the presence of non-negligible interactions $\sum_{i=1}^{p} S_i \leq 1$, and the first-order indices would not be informative enough to quantify sensitivities. This is generally the case for any real-world model.

The total-effect Sobol' index of X_i measures the total contribution of X_i to the output variance, including all interactions with other factors, and is given by (e.g., Homma and Saltelli, 1996):

$$T_i = 1 - \frac{\mathrm{Var}_{\mathbf{X}_{\sim i}}\left(\mathbb{E}_{X_i}[Y \mid \mathbf{X}_{\sim i}]\right)}{\mathrm{Var}(Y)} \tag{2}$$

One intuitive interpretation is that the term $\mathrm{Var}_{\mathbf{X}_{\sim i}}(\mathbb{E}_{X_i}[Y \mid \mathbf{X}_{\sim i}])$ represents the variance explained by all factors except X_i. Subtracting this from the total variance $\mathrm{Var}(Y)$ isolates the contribution of X_i and all its interactions. T_i satisfies the condition $S_i \leq T_i \leq 1$, and for additive models $S_i = T_i$ as there are no factor interactions.

Further details and a complete derivation of the Sobol' indices can be found, for example, in Saltelli et al. (2008). It is important to note that the definitions presented above assume that the input factors have independent probability distributions. Under this assumption, $\mathrm{Var}(Y)$ can be uniquely decomposed into contributions from individual factors and their interactions. When the inputs are not independent, this decomposition is no longer unique. Extensions of Sobol' indices to dependent inputs have been proposed in the literature (e.g., Chastaing et al., 2015; Il Idrissi et al., 2021; Jacques et al., 2006; Kucherenko et al., 2017; Li et al., 2010; Mara et al., 2015), but these are beyond the scope of this work.

2.2 Estimation of Sobol' indices

The estimation of Sobol' indices requires two essential components: a sampling strategy to generate points in the space of the input factors, and an estimator to compute sensitivity (Lo Piano et al., 2021). Here, we follow the Monte Carlo–based numerical approach recommended by Saltelli et al. (2010) to estimate S_i and T_i using a single set of model evaluations. For a comprehensive overview of available estimators, we refer the reader to Puy et al. (2022).

Let $\mathbf{A}$ and $\mathbf{B}$ be two independent sampling matrices, each containing n realizations of the p input factors. For each input factor, we construct the hybrid matrix $\mathbf{A}_B^{(i)}$ by replacing the i-th column of $\mathbf{A}$ with the corresponding column of $\mathbf{B}$. S_i and T_i

can be estimated as:

$$\widehat{S}_i = \frac{\frac{1}{n}\sum_{v=1}^{n} f(\mathbf{B})_v \left[f(\mathbf{A}_B^{(i)})_v - f(\mathbf{A})_v \right]}{\mathrm{Var}(Y)} \tag{3}$$

$$\widehat{T}_i = \frac{\frac{1}{n}\sum_{v=1}^{n} \left[f(\mathbf{A})_v - f(\mathbf{A}_B^{(i)})_v \right]^2}{2\mathrm{Var}(Y)} \tag{4}$$

In the previous expressions, $\mathrm{Var}(Y)$ can be estimated as (Puy et al., 2022):

$$\widehat{\mathrm{Var}}(Y) = \frac{1}{2n-1} \sum_{v=1}^{n} \left[\left(f(\mathbf{A})_v - f_0 \right)^2 + \left(f(\mathbf{B})_v - f_0 \right)^2 \right] \tag{5}$$

where $f_0 = \frac{1}{2n}\sum_{v=1}^{n} [f(\mathbf{A})_v + f(\mathbf{B})_v]$. Here and throughout the remainder of the paper we employ the hat notation for estimators.

Equation 3 was originally proposed in Saltelli et al. (2010) as an improvement to previous approaches, although it has been outperformed by better practices, see e.g. Azzini et al. (2021). Equation 4 was introduced in Jansen (1999) and represents the best practice thus far to estimate T_i. As a result, the estimation of S_i and T_i based on the triplet of matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{A}_B^{(i)}$ has a total cost of $n(p+2)$ simulations: $2n$ simulations are required for matrices $\mathbf{A}$ and $\mathbf{B}$, and pn simulations for the hybrid matrices $\mathbf{A}_B^{(i)}$.

2.3 Permutation variable importance

Let Y be a real-valued random variable and $\mathbf{X} = (X_1, \ldots, X_p)$ a vector of real-valued random variables. We assume that $\mathbf{X}$ and Y are jointly distributed according to some joint distribution encoding a general relationship between $\mathbf{X}$ and Y, which may be deterministic or stochastic.

To assess the contribution of each variable X_i to the variability of Y, we define the modified vector $\mathbf{X}_{(i)} = (X_1, ..., X_i', ..., X_p)$, where X_i' is an independent replication of X_i that is also independent of Y and of all other variables. The permutation variable importance (PVI) of the i-th variable is formally defined as (Zhu et al., 2015; Gregorutti et al., 2017):

$$\mathrm{PVI}_i = \mathbb{E}\left[\left(Y - \mathbb{E}[Y \mid \mathbf{X}_{(i)}] \right)^2 \right] - \mathbb{E}\left[\left(Y - \mathbb{E}[Y \mid \mathbf{X}] \right)^2 \right] \tag{6}$$

where $\mathbb{E}[Y \mid \mathbf{X}]$ denotes the conditional expectation of Y given $\mathbf{X}$.

$\mathbb{E}[Y \mid \mathbf{X}]$ represents the best function of $\mathbf{X}$ for predicting Y in the least-square sense (Bertsekas and Tsitsiklis, 2008), and therefore the expression $\mathbb{E}\left[\left(Y - \mathbb{E}[Y \mid \mathbf{X}] \right)^2 \right]$ corresponds to the smallest possible mean squared error (MSE) achievable if one could use the true conditional expectation as a predictor. This quantity represents the irreducible uncertainty in Y, that is, the part of its variability that cannot be explained by $\mathbf{X}$, no matter what model is used.

On the other hand, the term $\mathbb{E}\left[\left(Y - \mathbb{E}[Y \mid \mathbf{X}_{(i)}] \right)^2 \right]$ measures the expected MSE when X_i is replaced by an independent copy X_i' and the information carried by X_i has been destroyed. Hence, PVI_i measures the increase in MSE when X_i is made independent of Y, with a large PVI_i indicating that X_i contributes substantially to predicting Y.

If we restrict our attention to the case when the dependence between $\mathbf{X}$ and Y is purely deterministic, the relationship between $\mathbf{X}$ and Y can be written as $Y = f(\mathbf{X})$. This is the case of the generic model defined in the previous section. Here, $\mathbb{E}[Y \mid \mathbf{X}] = f(\mathbf{X})$, and Equation 6 reduces to:

$$\mathrm{PVI}_i = \mathbb{E}\left[\left(f(\mathbf{X}) - f(\mathbf{X}_{(i)})\right)^2\right] \tag{7}$$

PVI_i measures the average change in $f(\mathbf{X})$ when the i-th component is permuted. Then, if the components of $\mathbf{X}$ are independent, Wei et al. (2015) established the following relationship for T_i and PVI_i:

$$T_i = \frac{\mathrm{PVI}_i}{2\mathrm{Var}(Y)} = \frac{\mathbb{E}\left[\left(f(\mathbf{X}) - f(\mathbf{X}_{(i)})\right)^2\right]}{2\mathrm{Var}(Y)} \tag{8}$$

This relation constitutes an important connection between sensitivity analysis and ML: T_i is central to GSA, while PVI_i provides a model-agnostic approach widely used in explainable artificial intelligence (XAI) to quantify feature importance for ML models. Although beyond the scope of this work, interested readers are referred to the work of Benoumechiara (2019), which proposed a modified version of PVI_i to calculate S_i and provided an in-depth analysis of the dependent case using a Rosenblatt transformation of the input variables.

Based on Equation 8, an estimator of T_i can be constructed with the matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{A}_B^{(i)}$ introduced in the previous section. The generation of $\mathbf{A}_B^{(i)}$ by replacing the i-th column of $\mathbf{A}$ with the i-th column of $\mathbf{B}$ mimics the independent replication of X_i in $\mathbf{X}_{(i)}$. Therefore, T_i can be estimated with the following expression:

$$\widehat{T}_i = \frac{\frac{1}{n}\sum_{v=1}^{n}\left[f(\mathbf{A})_v - f(\mathbf{A}_B^{(i)})_v\right]^2}{2\mathrm{Var}(Y)} \tag{9}$$

where $\mathrm{Var}(Y)$ can be estimated using Equation 5.

Equation 9 is identical to the total-order estimator proposed by Jansen (1999) (see Equation 4). Our goal now is to demonstrate how the estimation of PVI_i is carried out in practice for ML models. This provides an alternative way for estimating T_i based on ML metamodelling, which is the focus of this work.

2.4 Estimation of PVI_i

Let $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ denote a learning set consisting of n independent observations from the joint distribution of $\mathbf{X}$ and Y, where $\mathbf{X}$ and Y are defined as in the previous section. The goal in a regression setting is to find an estimator $\hat{f}$ of $\mathbb{E}[Y \mid \mathbf{X}]$ (Zhu et al., 2015; Gregorutti et al., 2017). Broadly speaking, this involves fitting a function to the learning set that minimizes a loss function of choice, thereby producing a reliable approximation of the underlying relationship between $\mathbf{X}$ and Y.

The most common approach to estimate PVI_i in practice is the importance measure proposed in Breiman (2001) for random forests (Antoniadis et al., 2021; Vannucci et al., 2024). Let $\hat{f}_k$ denote the k-th tree in an ensemble of n_{tree} trees and let $\bar{\mathcal{D}}_k$ be its out-of-bag (OOB) sample, namely the subset of observations not used to train that tree. The prediction error $\hat{R}$ of $\hat{f}_k$, expressed

as MSE, is given by the following estimator over $\bar{\mathcal{D}}_k$ (Gregorutti et al., 2017):

$$\hat{R}(\hat{f}_k, \bar{\mathcal{D}}_k) = \frac{1}{|\bar{\mathcal{D}}_k|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{\mathcal{D}}_k} \left(Y_i - \hat{f}_k(\mathbf{X}_i)\right)^2 \tag{10}$$

where $|\bar{\mathcal{D}}_k|$ is the number of samples in $\bar{\mathcal{D}}_k$.

For each input factor X_i, let us now define a perturbed version of the OOB sample, denoted $\bar{\mathcal{D}}_k^i$, by randomly permuting the values of X_i while keeping all other variables unchanged. The importance of X_i is then estimated through the degradation in prediction error induced by this perturbation averaged over all trees in the forest (Gregorutti et al., 2017):

$$\widehat{\mathrm{PVI}}_i = \frac{1}{n_{\text{tree}}} \sum_{k=1}^{n_{\text{tree}}} \left[\hat{R}(\hat{f}_k, \bar{\mathcal{D}}_k^i) - \hat{R}(\hat{f}_k, \bar{\mathcal{D}}_k)\right] \tag{11}$$

This expression is the empirical counterpart of the theoretical definition of PVI_i presented in Equation 6. Although random forests provide a natural setting for estimating PVI_i, the permutation principle is not restricted to this class of models. Given any regression algorithm, one may estimate the importance of X_i by evaluating the predictive performance of the fitted model $\hat{f}$ and comparing it with the performance obtained after permuting the values of X_i. This general procedure is a model-agnostic approach widely used across ML algorithms (Pedregosa et al., 2011). In all cases, the resulting $\widehat{\mathrm{PVI}}_i$ quantifies how much the predictive accuracy of the particular model deteriorates when the information carried by X_i is destroyed.

Finally, when the learning set is generated via a Monte Carlo-based approach for GSA—for example, using the sampling matrix $\mathbf{A}$ and the corresponding model outputs $f(\mathbf{A})$ defined in Section 2.2—the regression setting becomes a metamodelling approach. In this case, the regression algorithm provides a surrogate model $\hat{f}$ for the true model f, and $\widehat{\mathrm{PVI}}_i$ can be used to approximate T_i. The quality of this approximation depends directly on the predictive accuracy of the surrogate: only when $\hat{f}$ adequately captures the behaviour of f on the learning set can $\widehat{\mathrm{PVI}}_i$ be expected to reproduce the corresponding theoretical quantity, and consequently yield reliable estimates of T_i. This approach yields a substantial computational advantage over classical estimation of T_i, which requires $n(p+2)$ model evaluations (see Section 2.2). The surrogate is trained on the n evaluations of matrix $\mathbf{A}$ alone and subsequently used to perform the additional evaluations required for $\widehat{\mathrm{PVI}}_i$ computation across all p parameters, replacing $n(p+1)$ model runs with surrogate predictions that take advantage of highly optimized ML libraries at negligible cost.

3 Numerical experiment

3.1 Specific objectives

Specific objectives in this technical note are:

1. To provide a representative hydrologic modelling context in which to generate numerical evidence for the relationship between T_i and PVI_i expressed in Equation 8.

2. To evaluate the capabilities of different ML metamodels to accurately approximate T_i in a computationally cheaper way than a traditional variance-based sensitivity analysis.

3. To assess the potential of alternative approaches in ML metamodelling to estimate sensitivities.

It is important to note that the specific choice of hydrologic models, catchments, and model output of interest is not central to the objectives of this study.

3.2 Hydrologic modelling

In this study, three hydrologic models that are frequently used in hydrologic modelling were employed: Hydrologiska Byråns Vattenbalansavdelning (HBV, Seibert, 1997), HyMod (Wagener et al., 2001) and Variable Infiltration Capacity (VIC) version 5 (Hamman et al., 2018). The number of free parameters p for these models is set to 13, 5, and 9, respectively. It should be noted that VIC is a physically based land-surface model and therefore includes a more complex parameterization than conceptual hydrologic models such as HBV and HyMod, with numerous soil and vegetation parameters describing land–atmosphere interactions. In this study, only a subset of VIC parameters was treated as free, while the remaining soil and vegetation parameters were obtained from the VICGlobal dataset (Schaperow et al., 2021) and kept fixed. Definitions of each parameter and their feasible ranges, as informed by literature (Kollat et al., 2012, for HBV and HyMod; Schaperow et al., 2021 and Yeste et al., 2023; 2024, for VIC), are provided in the Supplementary Material. For the purpose of this experiment, all models were implemented in a lumped configuration with a daily time step.

The study focuses on three headwater catchments located in northern Germany, namely the Sude River (gauge Garlitz), the Treene River (gauge Augaard), and the Trave River (gauge Schackendorf). These catchments have semi-humid climatic conditions. Runoff generation is predominantly groundwater-driven, while during wet and high-flow conditions saturation-excess processes also play a role. The landscape is characterized by glacial deposits with deep, sandy to loamy soils, and land use is predominantly agricultural. The catchments are free of major flow regulation, allowing for the analysis of hydrologic dynamics under near-natural conditions. Key physiographic and hydroclimatic characteristics of the catchments are summarized in Table 1. Catchment areas range from 136 to 675 km^2. Mean annual precipitation ranges from 661 mm yr^{-1} (Sude) to 909 mm yr^{-1} (Treene), while snow contributions are minor, with fractions below 0.05 for all catchments. Aridity indices range from 0.59 (Treene) to 0.96 (Sude), and runoff ratios are relatively similar (0.32–0.38), although the Sude catchment is somewhat drier, receiving about 20% less precipitation and yielding lower specific annual runoff. Overall, these groundwater-dominated catchments display similar annual hydrologic variability, suggesting comparable long-term water balance characteristics.

The study catchments are part of the CAMELS-DE dataset (Loritz et al., 2024): Garlitz (gauge ID DE812200), Schackendorf (gauge ID DEF13760), and Augaard (gauge ID DEF10190). They were selected based on the performance of the HBV model reported in the original CAMELS-DE data publication, in which HBV was applied across the full dataset and the Nash–Sutcliffe Efficiency (NSE) was provided as a performance metric. The three catchments with the highest NSE values were retained for the present analysis.

For each model–catchment combination, a Monte Carlo experiment was conducted using the sampling matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{A}_B^{(i)}$ described in Section 2.2. The sampling matrices were generated using the Sobol' quasi-random sequence scheme implemented in the SALib Python package (Herman and Usher, 2017) after setting the number of samples n to $10,000$. This

Table 1. Main physiographic and hydroclimatic characteristics of the three study catchments.

River	Gauge name	Catchment size (km^2)	Catchment altitude range (m.a.s.l.)	Mean precipitation ($mm\ yr^{-1}$)	Average low/mean/high flow ($l\ s^{-1}\ km^{-2}$)	Snow fraction	Runoff ratio	Aridity index
Sude	Garlitz	675	55–8	661	0.7/5.6/20.7	0.05	0.32	0.96
Treene	Augaard	136	58–21	909	1.3/10.8/64.2	0.03	0.38	0.59
Trave	Schackendorf	349	65–21	785	1.7/8.1/32.1	0.03	0.32	0.76

design resulted in a total of $n(p+2)$ model evaluations per model and catchment combination. The model output of interest was the Kling–Gupta Efficiency (KGE) of daily streamflow over the period 1 October 2001 to 30 September 2020.

For sensitivity analysis and metamodelling, model parameters were treated as uncertain inputs and assumed to be independent and uniformly distributed within their prescribed bounds. This choice corresponds to the most common practice in hydrologic sensitivity analysis and model calibration, where parameter uncertainty is represented by specifying lower and upper bounds and assigning a uniform distribution to each parameter (Mai, 2023). Importantly, the assumption of independent inputs is required for the theoretical relationship between T_i and PVI_i expressed in Equation 8. In the VIC model, there is a dependence relationship between two parameters. This dependency was explicitly accounted for and handled so that the resulting input space satisfies the independence assumption required for the analyses conducted here. Details of this procedure are provided in Appendix A.

3.3 ML metamodelling

To emulate the behaviour of the hydrologic models, a learning set was constructed using the parameter columns in matrix $\mathbf{A}$ as input variables and the corresponding KGE values from the Monte Carlo experiment as the target variable. This setup defined a regression problem, enabling the ML models to act as surrogate models for the hydrologic simulations. The learning set constructed from the sampling matrix $\mathbf{A}$ was used exclusively for training the metamodels, without applying cross-validation or a separate test set, as would be common in a predictive ML workflow. This choice reflects the purpose of the metamodels in this study, which is not to generalize to unseen data, but to accurately emulate the behaviour of the original hydrologic models within the sampled parameter space. From a sensitivity analysis perspective, faithful reproduction of the input–output relationship on the training domain is the primary requirement for obtaining reliable sensitivity estimates. Accordingly, model configurations that would typically be regarded as overfitting in a predictive setting are acceptable here, provided that they do not distort the resulting sensitivity measures. The implications of this choice and its consistency are further examined through a dedicated sanity check, as described in Section 3.5.

Three ML algorithms were tested as metamodelling approaches. A gradient boosting decision tree (GBDT) was implemented in XGBoost (Chen and Guestrin, 2016) with three configurations of 100, 150, and 200 trees. Choosing GBDT over a conventional random forest was motivated by computational efficiency: while both share the same tree-based architecture, the

boosting implementation in XGBoost is advantageous over the bagging approach of random forests. A neural network (NN) was implemented in TensorFlow (Abadi et al., 2016) with two hidden layers, three configurations of 100, 150, and 200 neurons per layer, and ReLU activation. The choice of a conventional feed-forward network over recurrent architectures such as long short-term memory (LSTM) networks is motivated by the nature of the surrogate task: each Monte Carlo simulation produces a single scalar KGE value from a fixed parameter set, meaning the surrogate's input space consists of static parameter vectors rather than sequential data. LSTM networks are designed to capture long-range dependencies in sequential data, and their recurrent structure provides no advantage when the target is a single scalar output derived from a fixed-length input vector. A feed-forward network with ReLU activations is therefore a natural and computationally efficient choice for this regression task. Finally, a linear model (LM) was implemented in scikit-learn (Pedregosa et al., 2011). For all three ML algorithms, MSE was selected as the loss function, which is a crucial step in assessing the applicability of Equation 8. These metamodelling approaches allowed us to efficiently reproduce the hydrologic simulations, thus providing a computationally cheap framework for GSA.

3.4 Sensitivity measures

While Section 2 presents the theoretical definitions and general estimation procedures for the sensitivity measures, this section describes their specific implementation in the context of the numerical experiment.

Variance-based sensitivity $\widehat{T}_i$ was calculated using the Python library SALib (Herman and Usher, 2017), which implements Equation 4 described previously, and ML-based $\widehat{\mathrm{PVI}}_i$ was computed from the three ML metamodelling approaches using the implementation available in scikit-learn (Pedregosa et al., 2011).

Feature importance was also evaluated using SHapley Additive exPlanations (SHAP, Lundberg and Lee, 2017), a model-agnostic method rooted in cooperative game theory. SHAP decomposes each individual model prediction into additive contributions associated with each input feature, ensuring local accuracy and global consistency. The formal mathematical definition of SHAP values is provided in Appendix A. In this work we used the SHAP Python library, which provides architecture-specific explainer implementations: GPUTreeExplainer (or TreeExplainer when CUDA is not available) was used for the GBDT metamodel, DeepExplainer for the NN metamodel, and LinearExplainer for the LM metamodel.

In the present application, where the inputs correspond to hydrologic model parameters, we denote by SHAP_i the contribution of parameter X_i to the target variable. From a sensitivity analysis perspective, SHAP_i values enable two complementary levels of analysis. First, a global measure of parameter importance can be obtained by averaging the absolute SHAP_i values across all samples, yielding mean $|\mathrm{SHAP}_i|$, which provides a scalar summary of the overall importance of X_i on the model response. Second, the sample-specific SHAP_i values allow for a distributed evaluation of parameter importance across both the parameter space and the model output space. This property is particularly relevant when assessing sensitivities with respect to model performance metrics. In such cases, global averages may emphasize parameter sensitivity in regions associated with poor model performance, potentially obscuring the controls that dominate in well-performing regions of the parameter space. By retaining sample-level contributions, SHAP_i values allow parameter importance to be examined across behavioural and non-behavioural regions of the model response surface.

Finally, standardized regression coefficients were computed for the LM as an additional sensitivity measure. The standardized coefficient for parameter X_i is defined as

$$\widehat{\beta}_i = \frac{\widehat{b}_i \, \sigma_{X_i}}{\sigma_Y} \tag{12}$$

where $\widehat{b}_i$ is the estimated regression coefficient, and σ_{X_i} and σ_Y are the sample standard deviations of X_i and the model output Y, respectively. Standardized regression coefficients have been used in previous hydrologic sensitivity analyses as a computationally inexpensive alternative to variance-based methods (e.g., Yeste et al., 2020; Zhang et al., 2025), at the cost of assuming linear model behaviour.

If the relationship between $\mathbf{X}$ and Y is given by a linear model of the form $Y = b_0 + \sum_{i=1}^{p} b_i X_i$, the model is additive and S_i and T_i coincide. In this case, the following equality holds (Saltelli et al., 2008):

$$S_i = T_i = \beta_i^2 \tag{13}$$

Moreover, combining this result with the theoretical relationship between T_i and PVI_i (Equation 8) yields

$$\beta_i^2 = \frac{\mathrm{PVI}_i}{2\mathrm{Var}(Y)} \tag{14}$$

Accordingly, the ability of the LM to approximate T_i through $\widehat{\beta}_i^2$ or $\widehat{\mathrm{PVI}}_i$ depends directly on the degree to which the hydrologic model response can be represented by a linear relationship.

To assess the uncertainty of the sensitivity estimates, 95% confidence intervals were computed for all sensitivity measures. For $\widehat{T}_i$, bootstrap resampling was performed using the implementation available in SALib (1000 resamples). For $\widehat{\mathrm{PVI}}_i$ and mean $|\mathrm{SHAP}_i|$, confidence intervals were obtained by evaluating the trained metamodels on resampled versions of the training data (1000 resamples). For the GBDT metamodel, subsampling without replacement with a subsample size of 80% was used, following a similar approach to Rouzies et al. (2023) for random forests, to avoid interference with the internal subsampling mechanism of XGBoost. For the NN and LM metamodels, standard bootstrap with replacement was used.

3.5 Sanity check and convergence test

To evaluate the consistency of the ML metamodelling approach with GSA, two a priori assessments were performed: a sanity check of the bias–variance tradeoff for the important measures, and a convergence test for all sensitivity measures. Simply put, the bias–variance tradeoff establishes that increasing model flexibility produces smaller training error, but may lead to larger test error due to overfitting (e.g., James et al., 2023). In the present context, however, the objective was not to generalize to unseen data, but to accurately emulate the true model response. As discussed in Section 2.4, $\widehat{\mathrm{PVI}}_i$ can provide a reliable approximation of T_i only if the metamodel is able to adequately reproduce the behaviour of the original model. From this perspective, a low training error was a necessary condition, and configurations that would be otherwise regarded as overfitting could be acceptable provided that they do not affect the feature importance estimates. Based on this, $\widehat{\mathrm{PVI}}_i$ and mean $|\mathrm{SHAP}_i|$ were calculated for the GBDT and the NN metamodels using five overfitting configurations, as reported in Table 2. For the

subsequent analysis, the less complex configuration was selected in each case, as it provided feature importance estimates comparable to the more complex configurations while achieving similar predictive performance.

Table 2. GBDT and NN metamodel configurations used during the sanity check. Configurations differ in model complexity and random initialization.

Configuration	GBDT	NN
1	100 trees, seed 42	2 hidden layers, 100 neurons per layer, seed 42
2	150 trees, seed 42	2 hidden layers, 150 neurons per layer, seed 42
3	200 trees, seed 42	2 hidden layers, 200 neurons per layer, seed 42
4	100 trees, seed 711	2 hidden layers, 100 neurons per layer, seed 711
5	100 trees, seed 1992	2 hidden layers, 100 neurons per layer, seed 1992

The convergence test was performed to verify that the sensitivity measures stabilized as the number of Monte Carlo samples increased. The test was carried out using a random subsampling strategy. Each subsample was drawn as an independent random sample without replacement from the full set of Monte Carlo simulations, with subsample sizes incremented in steps of 100 up to the full set of 10,000 samples, resulting in 100 subsampling matrices. For each subsampling matrix, $\widehat{T}_i$, $\widehat{\mathrm{PVI}}_i$, mean $|\mathrm{SHAP}_i|$, and $\widehat{\beta}_i^2$ were computed, and convergence was assessed by comparing their values between consecutive subsample sizes. Sensitivity estimates corresponding to each subsampling matrix were normalized by dividing by the sum of all sensitivities. The Euclidean distance between the resulting sensitivity vectors at consecutive subsample sizes was then used as the convergence metric. This procedure ensured that the sensitivity results were robust with respect to the size of the training dataset. The purpose of this test was to verify that the full set of 10,000 samples was sufficient for the sensitivity estimates to stabilize, rather than to identify an exact sample size at which convergence was achieved. Since the subsamples were drawn from the same finite pool, consecutive draws of increasing size inevitably share some samples, leading to smoother variations as the subsample size approaches the full set. Despite this overlap, the test provided a reliable assessment of the stability of the sensitivity estimates across the range of subsample sizes considered.

4 Results and discussion

For clarity of presentation, the results for the Garlitz catchment are used as the guiding thread in this section. Corresponding figures for the Schackendorf and Augaard catchments are provided in the Supplementary Material. All three catchments exhibit consistent patterns, and their joint analysis is intended to ensure the robustness of the results rather than to highlight site-specific differences.

4.1 Sanity check and convergence test

Figure 1 summarizes the predictive performance of the GBDT, NN, and LM metamodels by comparing fitted and original KGE values across the three hydrologic models. For GBDT and NN, only configuration 1 is shown (see Table 2), as all five

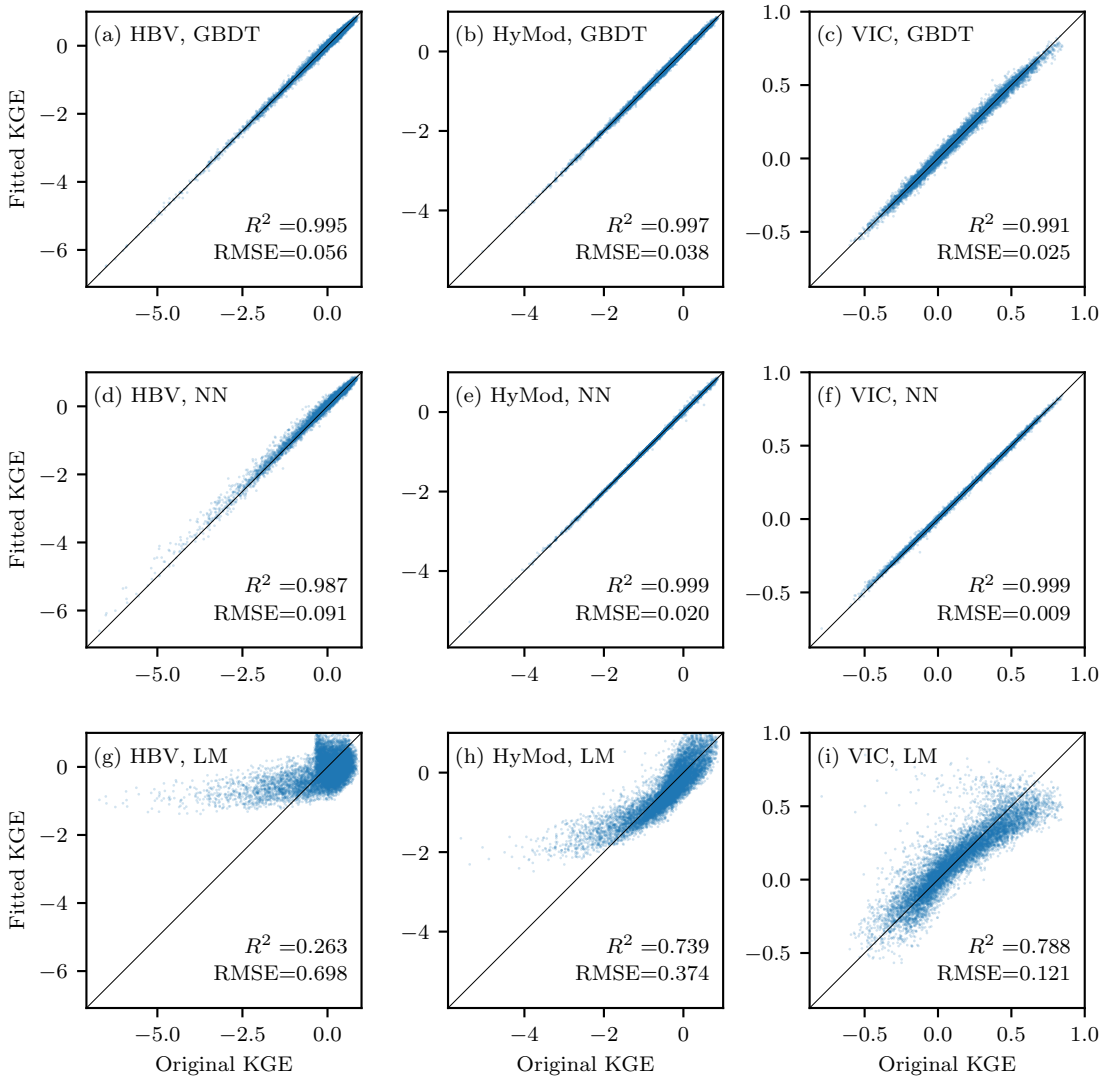

Figure 1. Predictive performance of the GBDT, NN, and LM metamodels for the three hydrologic models. GBDT and NN results are shown for configuration 1 (see Table 2). Scatter plots compare fitted and original KGE values for (a–c) GBDT, (d–f) NN, and (g–i) LM applied to HBV, HyMod, and VIC, respectively. The solid line indicates the 1:1 relationship. Coefficients of determination (R^2) and root mean square errors (RMSE) are reported in each panel.

configurations exhibit similar predictive performance. Both GBDT and NN closely match the reference simulations for all three hydrologic models, indicating that they capture the dominant structure of the parameter–output relationship. The high R^2 values are expected, as the metamodels are evaluated on the same points used for training. This confirms that the surrogates adequately capture the behaviour of the original models, which is a necessary condition for the validity of Equation 8 (see Section 2.4). In contrast, LM exhibits markedly weaker performance, particularly for HBV, with HyMod and VIC showing the highest linear response. The suitability and limitations of using an LM metamodel in contrast to GBDT and NN are discussed in the next section.

Figure 2 shows the sensitivity estimates obtained for the GBDT and NN metamodels across five flexible configurations, using both $\widehat{\mathrm{PVI}}_i$ and mean $|\mathrm{SHAP}_i|$. All five configurations correspond to deliberately overfitted metamodels (see Table 2), with coefficients of determination exceeding $R^2 > 0.98$ for all hydrologic models. The importance curves corresponding to the different configurations largely overlap, indicating that the magnitude of the importance estimates remains remarkably consistent across varying model complexity and random initialization. In addition, panel titles in Figure 2 report the mean $\pm$ standard deviation of pairwise Kendall τ and Spearman ρ between the importance vectors of the five configurations, which consistently show high values across all model–metamodel–importance combinations. This behaviour is observed for both $\widehat{\mathrm{PVI}}_i$ and mean $|\mathrm{SHAP}_i|$, suggesting that potential overfitting does not affect the resulting importance estimates. For a given importance measure, GBDT and NN exhibit nearly identical magnitudes and rankings, highlighting the robustness of the results with respect to the choice of metamodel. When comparing $\widehat{\mathrm{PVI}}_i$ and mean $|\mathrm{SHAP}_i|$, the relative ranking of influential parameters is generally preserved, while small differences in magnitude are observed. These differences are expected given the distinct theoretical foundations of the two approaches, but their close agreement indicates a consistent identification of dominant parameters across methods. Based on these results, configuration 1 was selected for GBDT and NN for the subsequent analysis, as it is the least complex while providing feature importance estimates comparable to the more complex configurations (see Section 3.5).

Finally, Figure 3 presents the convergence behaviour of all sensitivity measures as a function of the number of Monte Carlo samples during the subsampling experiment. For all three hydrologic models, $\widehat{T}_i$ exhibits stable and monotonic convergence as the subsample size increases. A similar convergence pattern is observed for $\widehat{\mathrm{PVI}}_i$ and mean $|\mathrm{SHAP}_i|$ derived from the ML metamodels, indicating that the importance estimates stabilize with increasing training data. For all sensitivity measures, stabilization occurs at subsample sizes well below the full set of 10,000 samples, confirming that the number of Monte Carlo simulations used in this study was sufficient. Convergence is generally faster and smoother for HyMod and VIC, whereas HBV shows comparatively slower convergence and higher variability across sensitivity measures. This behaviour is consistent with the higher structural complexity of HBV, which involves 13 parameters, compared to the more parsimonious HyMod (5 parameters) and VIC (9 parameters), and reflects the increased sampling requirements associated with higher-dimensional parameter spaces. In addition, mean $|\mathrm{SHAP}_i|$ values computed from the GBDT metamodel display slightly reduced stability for HBV relative to the other cases.

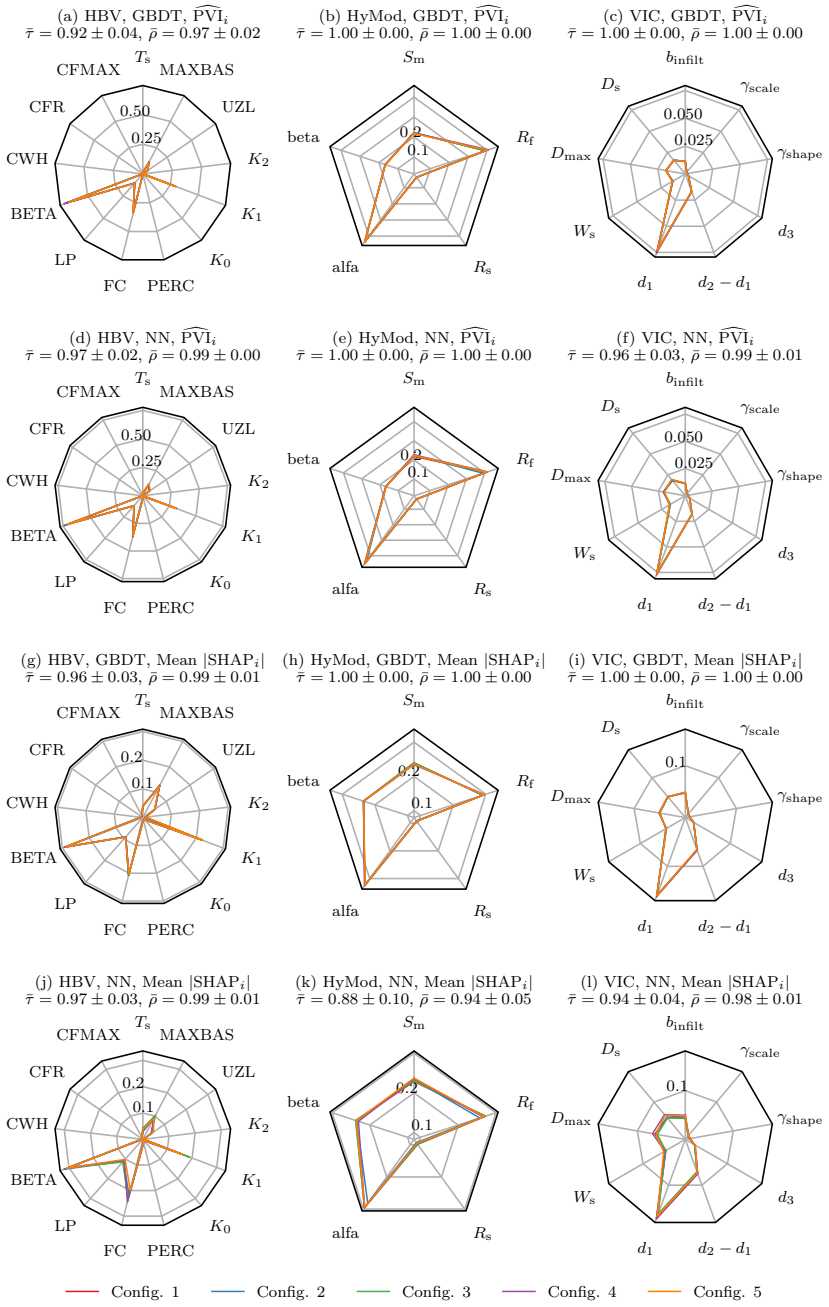

Figure 2. Sanity check of feature importance estimates for GBDT and NN metamodels across five overfitting configurations (see Table 2). Panels (a–c) show $\widehat{\mathrm{PVI}}_i$ and panels (g–i) show mean $|\mathrm{SHAP}_i|$ for the GBDT metamodel applied to HBV, HyMod, and VIC, respectively, while panels (d–f) and (j–l) show the corresponding results for the NN metamodel. Each coloured line represents a different model configuration. Panel titles report the mean $\pm$ standard deviation of pairwise Kendall τ and Spearman ρ between the importance vectors of the five configurations. The strong overlap of importance curves across configurations, supported by high $\bar{\tau}$ and $\bar{\rho}$ estimates, indicates that the feature importance estimates are robust to overfitting.

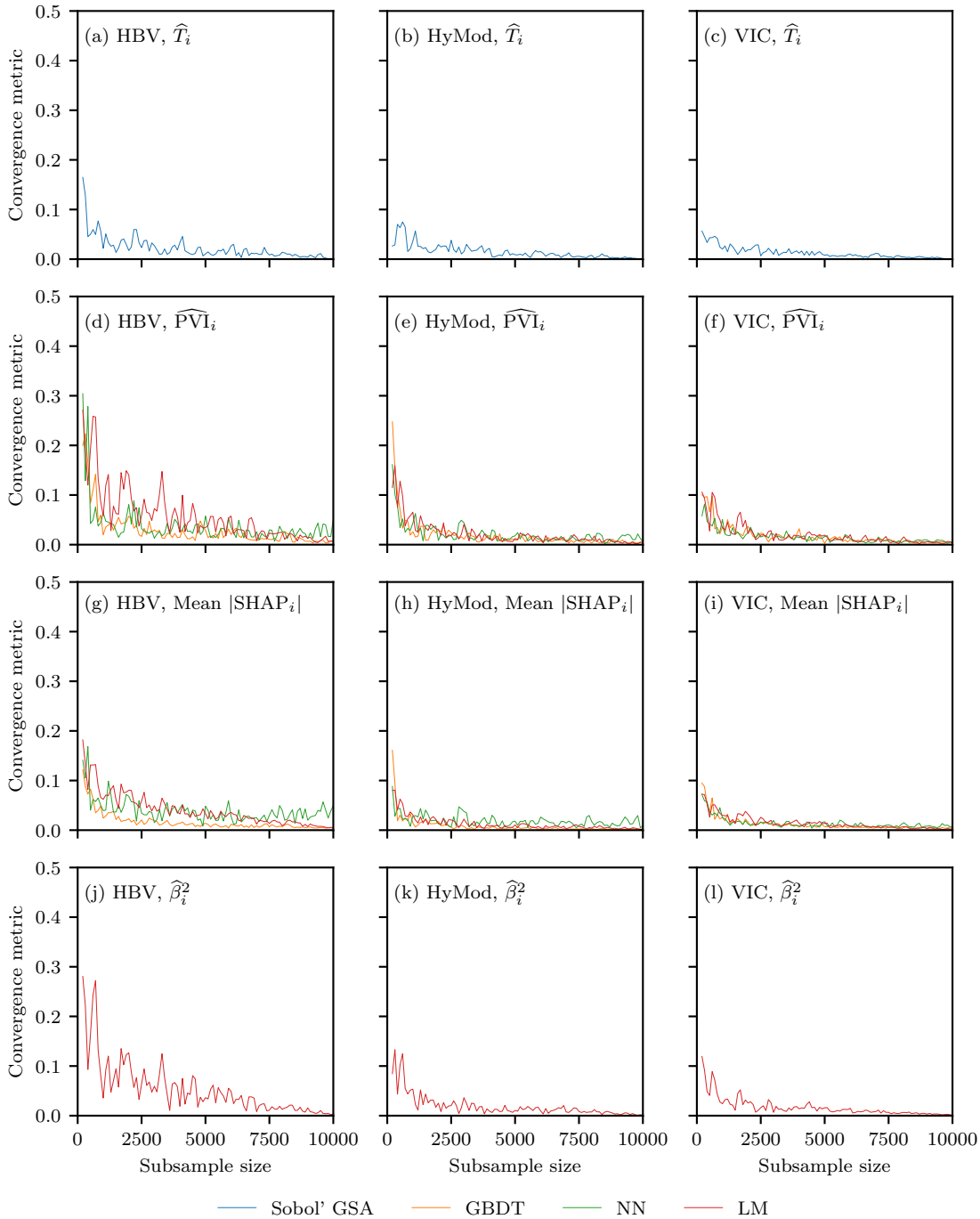

Figure 3. Convergence of sensitivity measures as a function of subsample size. Panels (a–c) show the convergence of $\widehat{T}_i$, panels (d–f) of $\widehat{\mathrm{PVI}}_i$, panels (g–i) of mean $|\mathrm{SHAP}_i|$, and panels (j–l) of $\widehat{\beta}_i^2$ for HBV, HyMod, and VIC, respectively. The convergence metric is computed as the Euclidean distance between normalized sensitivity vectors obtained from consecutive subsample sizes. Normalization was performed by dividing by the sum of all sensitivities.

4.2 Relationship between T_i and PVI_i

The theoretical relationship between T_i and PVI_i expressed in Equation 8 was examined empirically for the HBV, HyMod, and VIC models using the importance measures derived from GBDT, NN, and LM metamodels. Figure 4 shows the normalized $\widehat{T}_i$ together with normalized $\widehat{\mathrm{PVI}}_i$ estimates and normalized mean $|\mathrm{SHAP}_i|$ values obtained from the three metamodels. As in the convergence analysis (Figure 3), values were normalized by dividing by the sum of all sensitivities in order to facilitate a direct comparison and produce a consistent ranking of parameters. For the GBDT and NN metamodels, $\widehat{\mathrm{PVI}}_i$ reproduces both the ranking and the relative magnitude of $\widehat{T}_i$ across all three hydrologic models, reflecting only minor discrepancies. In contrast, LM-based $\widehat{\mathrm{PVI}}_i$ provides a good approximation of $\widehat{T}_i$ for HyMod and VIC, but exhibits larger deviations for HBV, particularly in the relative importance assigned to dominant parameters. This behaviour is consistent with the comparatively poor predictive performance of the LM for HBV (Figure 1). This expected limited fit of the LM underscores that the ability of $\widehat{\mathrm{PVI}}_i$ to approximate $\widehat{T}_i$ depends critically on the accuracy with which the metamodel captures the underlying model response.

The correspondence between $\widehat{T}_i$ and $\widehat{\mathrm{PVI}}_i$ is further illustrated in Figure 5 (panels a–c), which directly compares the two quantities for each hydrologic model parameter and metamodel. GBDT- and NN-based estimates closely follow the theoretical relationship, providing clear numerical support for the connection established in Equation 8. For the LM, agreement with the theoretical relationship is observed only for HyMod, whereas a weaker correspondence is observed for HBV and VIC – in line with how well the LM is able to capture hydrologic model behaviour. The close agreement observed between $\widehat{T}_i$ and $\widehat{\mathrm{PVI}}_i$ for GBDT is in line with previous theoretical and numerical results under independence assumptions (Wei et al., 2015; Antoniadis et al., 2021). While earlier work primarily focused on GBDT, the present results extend these findings by demonstrating that alternative ML metamodels can also provide reliable approximations of $\widehat{T}_i$, provided that they accurately reproduce the underlying input–output relationship. In particular, the strong performance of the NN metamodel, and the satisfactory behaviour of the LM for HyMod, indicate that the validity of $\widehat{\mathrm{PVI}}_i$-based sensitivity estimates is not tied to a specific algorithm, but rather to the fidelity of the regression surrogate. This observation corroborates the insight of Gregorutti et al. (2017), who emphasized that the empirical estimation of PVI_i strongly depends on the algorithm used to approximate the regression function. In contrast, Rouzies et al. (2023) reported that the relationship between $\widehat{\mathrm{PVI}}_i$ and $\widehat{T}_i$ was not well respected when using a random forest surrogate for a distributed, physically based pesticide transfer model with 145 input parameters, and after applying a screening procedure to reduce the dimension of the problem. In their study, the poor predictive performance of the surrogate was identified as the likely cause, further underscoring that the practical applicability of Equation 8 depends on surrogate fidelity. It should be noted, however, that in our work all three models were used in lumped configurations: HBV and HyMod are lumped conceptual rainfall-runoff models, while VIC, though a land-surface model with a more complex parameterization including surface and energy balance components, was likewise run in a lumped setup. This limits the degree of structural complexity relative to fully distributed applications such as that of Rouzies et al. (2023), and may partly explain why high surrogate fidelity was more readily achieved in our study.

In recent hydrologic studies, tree-based ensemble approaches and feature importance measures have been widely used as exploratory tools to analyse large-sample datasets and to identify dominant drivers of hydrologic behaviour, for example, in the

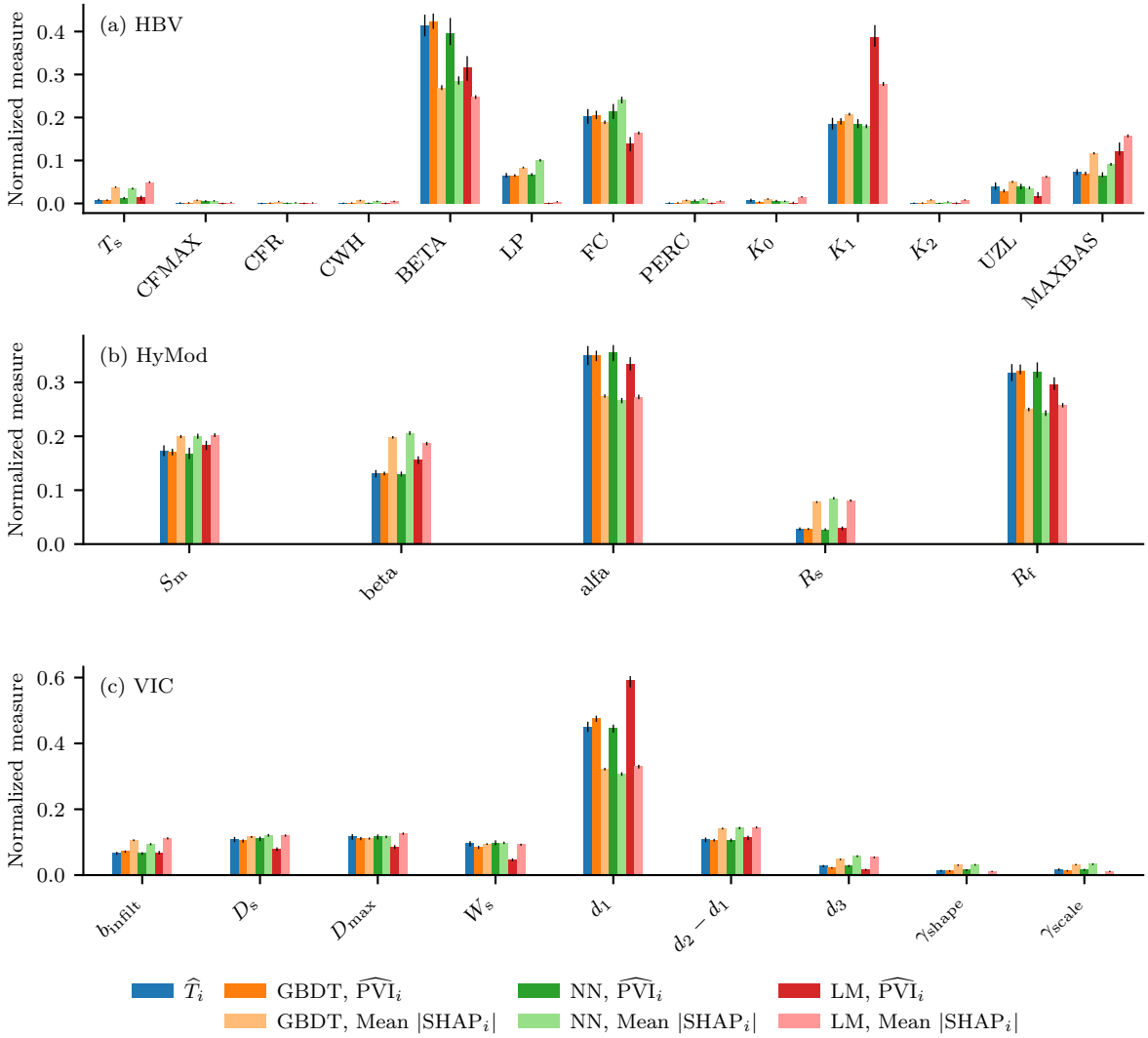

Figure 4. Parameter rankings for (a) HBV, (b) HyMod, and (c) VIC. Bars show normalized $\widehat{T}_i$ from variance-based GSA, normalized $\widehat{\mathrm{PVI}}_i$ estimates, and normalized mean $|\mathrm{SHAP}_i|$ with $\widehat{T}_i$ obtained from GBDT, NN, and LM metamodels. Sensitivity measures were normalized by dividing each by the sum of all sensitivities, as in the convergence test. Thin black lines indicate 95% confidence intervals.

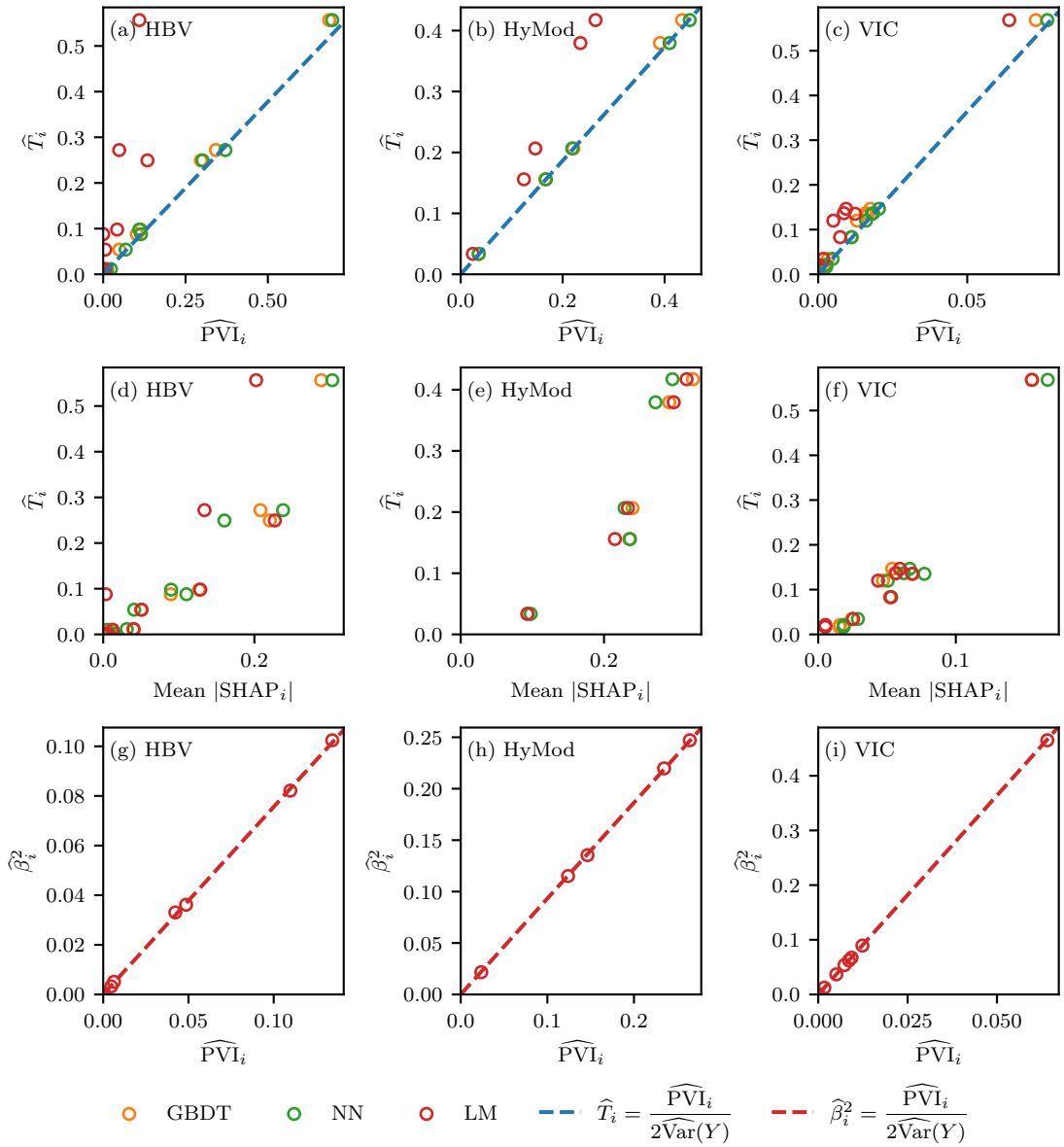

Figure 5. Comparison of different sensitivity estimates across hydrologic models. Panels (a–c) compare $\widehat{\mathrm{PVI}}_i$ with $\widehat{T}_i$ for HBV, HyMod, and VIC, respectively. Panels (d–f) compare mean $|\mathrm{SHAP}_i|$ with $\widehat{T}_i$, and panels (g–i) compare $\widehat{\mathrm{PVI}}_i$ with $\widehat{\beta}_i^2$ for the LM metamodel. Dashed lines indicate the corresponding theoretical relationships. Colours denote the ML metamodel used.

context of intercatchment groundwater flow (Liu et al., 2020), hydrologic signatures (Addor et al., 2018), and flood-generating processes (Stein et al., 2021). In these applications, different XAI approaches are used to help elucidate empirical relationships learned by the GBDT. ML metamodels are also efficient alternatives to complex hydrologic models, such as VIC (Gu et al., 2020; Sun et al., 2023) and ParFlow.CLM (Lim and Wang, 2022), and have already been applied to accelerate studies on climate impacts (Schnorbus and Cannon, 2014) as well as for regionalization and calibration of conceptual hydrologic models (Farahani et al., 2025; Tang et al., 2025). A closer connection to sensitivity analysis is found in the work of Brêda et al. (2024), who employed feature importance as a diagnostic approach for global hydrologic and land-surface models using a random forest surrogate. The present study builds directly on this emerging line of research by explicitly linking feature importance measures to variance-based total-effect sensitivity indices across multiple hydrologic models and metamodel classes.

4.3 Alternative approaches to estimate sensitivities

As an alternative to $\widehat{\mathrm{PVI}}_i$, Figure 5 also examines the possibility of using mean $|\mathrm{SHAP}_i|$ and $\widehat{\beta}_i^2$ for GSA. The normalized mean $|\mathrm{SHAP}_i|$ values included in Figure 4 further illustrate this comparison at the parameter level. Panels (d–f) of Figure 5 contrast mean $|\mathrm{SHAP}_i|$ with $\widehat{T}_i$, revealing a clear monotonic relationship across all hydrologic models, with larger dispersion than $\widehat{\mathrm{PVI}}_i$ (panels a–c). Results obtained with GBDT and NN are very similar for all hydrologic models, whereas, as observed for $\widehat{\mathrm{PVI}}_i$, LM-based results are satisfactory for HyMod and VIC, where the model response exhibits a stronger degree of linearity. Despite the increased dispersion, mean $|\mathrm{SHAP}_i|$ preserves nearly the same ranking of influential parameters as identified by $\widehat{\mathrm{PVI}}_i$ and $\widehat{T}_i$. Panels (g–i) compare $\widehat{\mathrm{PVI}}_i$ with $\widehat{\beta}_i^2$ for the LM metamodel, confirming the proportionality between the two measures established in Equation 14.

Beyond providing a global ranking, SHAP_i values enable a more detailed and informative sensitivity analysis across both the parameter space and the target variable space, as illustrated in Figure 6 for HyMod using the GBDT metamodel. Panel (a) reports a global importance ranking based on mean $|\mathrm{SHAP}_i|$, which coincides with the results shown in Figure 5 (panel e). Panel (b) shows the distribution of SHAP_i values across the parameter space, highlighting that parameters with similar global importance can exhibit markedly different local contributions depending on their values (e.g., S_m and beta). Panel (c) focuses on the most influential parameter, alfa, illustrating how its contribution to the model response varies across its range. Finally, panel (d) displays SHAP_i values for alfa as a function of KGE, showing that the region of highest model performance (i.e., larger KGE values) is associated with positive and comparatively larger SHAP_i values for alfa.

The ability of SHAP_i values to characterize how parameter influence varies across both the parameter space and the target variable space is conceptually closely related to the Distributed Evaluation of Local Sensitivity Analysis (DELSA) introduced by Rakovec et al. (2014). DELSA was developed to move beyond single summary measures of sensitivity by providing distributions of parameter sensitivities across the parameter space, thereby revealing regimes in which parameters may alternate between being influential or negligible. As emphasized by Rakovec et al. (2014), the ability to understand how parameter sensitivities vary throughout parameter space is critical for informing decisions related to data collection and model development, particularly when sensitivities are evaluated with respect to performance metrics. This perspective aligns with the broader critique of GSA articulated by Razavi and Gupta (2015, 2016), who demonstrated that global aggregation can obscure important

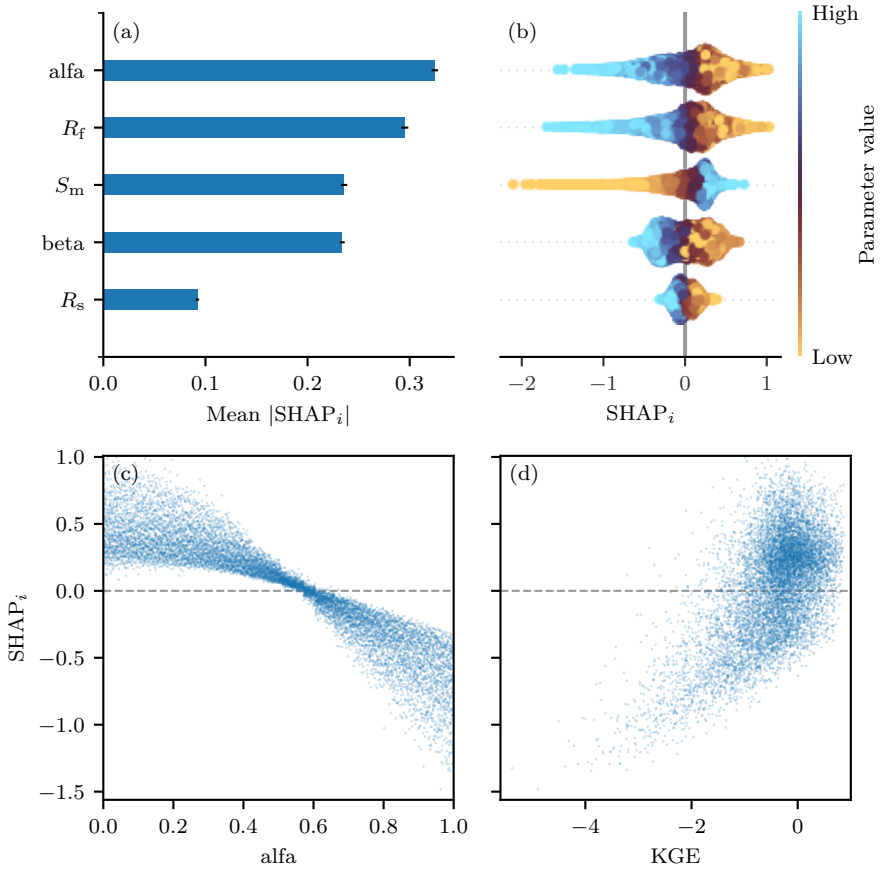

Figure 6. Detailed SHAP_i-based sensitivity analysis for HyMod using the GBDT metamodel. Panel (a) shows the global parameter ranking based on mean $|\mathrm{SHAP}_i|$, with thin black lines indicating 95% confidence intervals. Panel (b) displays the distribution of SHAP_i values across the parameter space, coloured by parameter value. Panel (c) illustrates the variation of SHAP_i for the most influential parameter, alfa, across its range. Panel (d) shows SHAP_i values for alfa as a function of KGE, highlighting how parameter contributions vary across model performance levels.

behaviour when sensitivity is assessed using performance metrics. The SHAP_i-based analyses presented here address these limitations by providing sample-specific contributions that can be examined jointly in parameter space and target variable space.

In this sense, SHAP_i values provide a complementary perspective to distributed sensitivity approaches such as DELSA, while naturally integrating with ML metamodelling frameworks. It is important to emphasize that neither SHAP_i nor DELSA should be regarded as equivalent to classical GSA measures such as $\widehat{T}_i$. DELSA is a hybrid local-global approach that provides distributions of parameter sensitivities across the parameter space, rather than the single summary measures characteristic of GSA (see also Dell'Oca et al., 2020, for a non-aggregated approach based on copula density-driven metrics). Similarly, SHAP_i operates as a local prediction-explanation method that attributes the contribution of each input to individual predictions, and the global importance derived from mean absolute SHAP_i values is an aggregation of these local explanations rather than a variance decomposition index. This complementary relationship between SHAP and SA-based approaches has been explored in recent studies. Panigrahi et al. (2025) compared an SA-based explainer, the Explainer Variogram Analysis of Response Surface (XVARS), with SHAP, showing that both approaches mostly align in their feature rankings but differ in that XVARS characterizes the entire response surface of ML models, whereas SHAP provides local explanations that are generalized by averaging over sampled data points and may miss behaviour in regions without training data. Panigrahi et al. (2026) extended this comparison, finding that XVARS and SHAP are broadly consistent in identifying dominant features but differ in scope and robustness: SA-based methods are independent of training sample size and more robust to sampling variability, whereas SHAP scales linearly with sample size and its importance estimates depend on the underlying ML architecture.

Last, despite the absence of a formal link between SHAP_i and $\widehat{T}_i$, both measures share a common theoretical root in the Shapley value framework from cooperative game theory, which provides a principled approach for allocating contributions among interacting variables. In the context of GSA, the same principle has been adopted through Shapley effects, which allocate contributions to the variance of model outputs among input variables. Owen (2014) demonstrated that, for independent inputs, Shapley effects are bounded by the first-order and total-effect Sobol indices. Although SHAP and Shapley effects address different attribution problems, both rely on the same theoretical foundation. Hence, when SHAP values are aggregated over a representative sample of the input space, as in the mean $|\mathrm{SHAP}_i|$ values used in this work, they provide a global measure of variable influence that, while not equivalent to Sobol indices or Shapley effects, can be considered a complementary global sensitivity measure.

5 Conclusions

This technical note examined the practical relationship between GSA and ML metamodelling within a hydrologic modelling context. Using three hydrologic models (HBV, HyMod, and VIC) applied over three catchments in Germany, and taking KGE as the target variable, we computed Sobol' T_i and assessed whether GBDT, NN, and LM metamodels can reproduce these sensitivities through two feature importance measures, namely PVI_i and SHAP_i values.

The numerical experiments provided clear empirical evidence supporting the theoretical relationship between T_i and PVI_i underlying Equation 8. Across all hydrologic models, the GBDT and NN metamodels reproduced both the ranking and the relative magnitude of T_i using PVI_i, with only minor discrepancies. The results further showed that, as expected, the reliability of PVI_i as a proxy for T_i depends critically on metamodel performance: when the LM failed to accurately reproduce the hydrologic model response, agreement between PVI_i and T_i deteriorated, whereas satisfactory performance and agreement were obtained for the more linear case.

Alternative sensitivity measures derived from ML metamodels were also assessed. Mean $|\mathrm{SHAP}_i|$ exhibited a consistent monotonic relationship with T_i and preserved nearly the same ranking of influential parameters observed for T_i and PVI_i. For the LM, β_i^2 was proportional to PVI_i, as expected from Equation 14. Importantly, SHAP_i values provided additional diagnostic insight by enabling a distributed evaluation of parameter sensitivities across both the parameter space and the model output space, thereby revealing important effects that are averaged out by single summary measures of sensitivity.

The results also highlight the key computational motivation for using ML metamodelling for GSA. Classical variance-based estimation of $\widehat{T_i}$ requires $n(p+2)$ model evaluations, which can be prohibitive for complex or high-dimensional models. In contrast, once a surrogate model is trained, $\widehat{\mathrm{PVI}}_i$ can be computed at negligible additional cost, while relying only on the n model evaluations used for training. This makes ML metamodels particularly attractive for efficient sensitivity analysis of computationally demanding models.

From a hydrologic modelling perspective, these findings support ML metamodelling as a practical and flexible framework for GSA, with direct relevance for parameter prioritization, model diagnosis, and the interpretation of model behaviour across model performance levels. By establishing a clear and empirically validated link between variance-based sensitivity indices and ML-based feature importance measures, this work helps bridge methodological developments in GSA and XAI with applied hydrologic modelling practice. The analysis was conducted for catchments characterized by semi-humid climatic conditions and predominantly groundwater-driven runoff generation, which may influence the generality of the results. Future research could extend this analysis to catchments with contrasting hydroclimatic and physiographic characteristics, to settings with dependent inputs, alternative model outputs of interest, and more complex hydrologic modelling configurations, as well as to applications where metamodels are used beyond sensitivity estimation, such as for exploring uncertainties, supporting calibration, or transferring across catchments.

Code and data availability. All data and code used in this study are available in a Zenodo repository (Yeste et al., 2026). All implementation details and computational times are documented in the repository description pdf.

Appendix A: Mathematical definition of SHAP values

Let $Y = f(\mathbf{X})$ be the deterministic model defined in Section 2.1, and let $x = (x_1, \ldots, x_p)$ be a specific realization of the input vector $\mathbf{X}$. The SHAP value of the i-th input variable for the prediction $f(x)$ is defined as the Shapley value of the conditional expectation function (Lundberg and Lee, 2017):

$$\phi_i(f, x) = \sum_{S \subseteq P \setminus \{i\}} \frac{|S|!(p - |S| - 1)!}{p!} \left[\mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_{S \cup \{i\}} = x_{S \cup \{i\}}] - \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = x_S] \right] \tag{A1}$$

where $P = \{1, \ldots, p\}$ is the set of input indices and the sum runs over all subsets S of P that exclude i.

The two conditional expectations represent the expected model output when only a subset of inputs is known:

- $\mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_{S \cup \{i\}} = x_{S \cup \{i\}}]$ is the expected model output when the inputs in S and X_i are set to their values in x, while the remaining inputs vary according to their prescribed distributions.
- $\mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = x_S]$ is the same expectation, but now X_i is left unknown and marginalized over its distribution.

Their difference measures the change in expected model output when X_i is made known. This difference depends on which other inputs are already known, so the Shapley value averages this difference over all possible subsets S that exclude i. The weight $|S|!(p - |S| - 1)!/p!$ gives each subset S its appropriate share according to cooperative game theory (Shapley, 1953), ensuring that the contribution of each input is fairly allocated regardless of the order in which inputs are considered.

The exact computation of SHAP values is challenging (Lundberg and Lee, 2017). In practice, SHAP values are approximated using estimators that exploit the structure of the prediction model, such as TreeExplainer for tree-based models or DeepExplainer for neural networks. This computational challenge also motivates the use of ML metamodels $\hat{f}$ as efficient surrogates of the original model f, for which SHAP values can be computed much more efficiently.

Throughout this paper, the ϕ_i values obtained from a ML metamodel $\hat{f}$ are referred to as SHAP_i, and their absolute values averaged over all samples yield the mean $|\mathrm{SHAP}_i|$, which is used here as a global measure of parameter importance.

Author contributions. Conceptualization: PY, LAM, JPLFB, AS, GV, RS. Methodology: PY, LAM, JPLFB. Software: PY, NT. Validation: NT, AB. Formal analysis: PY. Investigation: PY. Data curation: PY, NT. Writing – Original Draft: PY. Writing – Review & Editing: PY, LAM, JPLFB, NT, AS, GV, RS, AB. Visualization: PY. Supervision: AB. Funding acquisition: PY, AB

Competing interests. The contact author has declared that none of the authors has any competing interests.

Acknowledgements. The first author acknowledges the Alexander von Humboldt Foundation for a Humboldt Research Fellowship for postdoctoral researchers. The first author is currently funded by a UGR-Banco Santander postdoctoral fellowship. This work was supported

by the Italian Ministry of Research, under the complementary actions to the NRRP “Fit4MedRob - Fit for Medical Robotics” Grant (# PNC0000007). All the simulations were conducted in the HPC cluster at the University of Potsdam (https://www.uni-potsdam.de/en/zim/angebote-loesungen/hpc, last access: 05 February 2026). We thank OpenAI’s ChatGPT (version 4) for providing suggestions on wording and phrasing during manuscript preparation.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, https://doi.org/10.48550/arXiv.1603.04467, 2016.

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

Antoniadis, A., Lambert-Lacroix, S., and Poggi, J.-M.: Random forests for global sensitivity analysis: A selective review, Reliability Engineering & System Safety, 206, 107 312, https://doi.org/10.1016/j.ress.2020.107312, 2021.

Azzini, I., Mara, T. A., and Rosati, R.: Comparison of two sets of Monte Carlo estimators of Sobol' indices, Environmental Modelling & Software, 144, 105 167, https://doi.org/10.1016/j.envsoft.2021.105167, 2021.

Benoumechiara, N.: Treatment of dependency in sensitivity analysis for industrial reliability, Sorbonne Université, 2019.

Bertsekas, D. and Tsitsiklis, J. N.: Introduction to Probability, Athena Scientific, 2008.

Borgonovo, E., Castaings, W., and Tarantola, S.: Moment Independent Importance Measures: New Results and Analytical Test Cases, Risk Analysis, 31, 404–428, https://doi.org/10.1111/j.1539-6924.2010.01519.x, 2011.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Brêda, J. P. L., Melsen, L. A., Athanasiadis, I., Van Dijk, A., Siqueira, V. A., Verhoef, A., Zeng, Y., and van der Ploeg, M.: Predictor Importance for Hydrological Fluxes of Global Hydrological and Land Surface Models, Water Resources Research, 60, https://doi.org/10.1029/2023WR036418, 2024.

Chastaing, G., Gamboa, F., and Prieur, C.: Generalized Sobol sensitivity indices for dependent variables: numerical methods, Journal of Statistical Computation and Simulation, 85, 1306–1333, https://doi.org/10.1080/00949655.2014.960415, 2015.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, Association for Computing Machinery, https://doi.org/10.1145/2939672.2939785, 2016.

Dell'Oca, A., Guadagnini, A., and Riva, M.: Copula density-driven metrics for sensitivity analysis: Theory and application to flow and transport in porous media, Advances in Water Resources, 145, 103 714, https://doi.org/10.1016/j.advwatres.2020.103714, 2020.

Farahani, M. A., Wood, A. W., Tang, G., and Mizukami, N.: Calibrating a large-domain land/hydrology process model in the age of AI: the SUMMA CAMELS emulator experiments, Hydrology and Earth System Sciences, 29, 4515–4537, https://doi.org/10.5194/hess-29-4515-2025, 2025.

Gregorutti, B., Michel, B., and Saint-Pierre, P.: Correlation and variable importance in random forests, Statistics and Computing, 27, 659–678, https://doi.org/10.1007/s11222-016-9646-1, 2017.

Gu, H., Xu, Y.-P., Ma, D., Xie, J., Liu, L., and Bai, Z.: A surrogate model for the Variable Infiltration Capacity model using deep learning artificial neural network, Journal of Hydrology, 588, 125 019, https://doi.org/10.1016/j.jhydrol.2020.125019, 2020.

Hamman, J. J., Nijssen, B., Bohn, T. J., Gergel, D. R., and Mao, Y.: The Variable Infiltration Capacity model version 5 (VIC-5): infrastructure improvements for new applications and reproducibility, Geoscientific Model Development, 11, 3481–3496, https://doi.org/10.5194/gmd-11-3481-2018, 2018.

Herman, J. and Usher, W.: SALib: An open-source Python library for Sensitivity Analysis, Journal of Open Source Software, 2, 97, https://doi.org/10.21105/joss.00097, 2017.

Homma, T. and Saltelli, A.: Importance measures in global sensitivity analysis of nonlinear models, Reliability Engineering & System Safety, 52, 1–17, https://doi.org/10.1016/0951-8320(96)00002-6, 1996.

Il Idrissi, M., Chabridon, V., and Iooss, B.: Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs, Environmental Modelling & Software, 143, 105 115, https://doi.org/10.1016/j.envsoft.2021.105115, 2021.

Jacques, J., Lavergne, C., and Devictor, N.: Sensitivity analysis in presence of model uncertainty and correlated inputs, Reliability Engineering & System Safety, 91, 1126–1134, https://doi.org/10.1016/j.ress.2005.11.047, 2006.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J.: An Introduction to Statistical Learning: with Applications in Python, Springer International Publishing, https://doi.org/10.1007/978-3-031-38747-0, 2023.

Jansen, M. J. W.: Analysis of variance designs for model output, Computer Physics Communications, 117, 35–43, https://doi.org/10.1016/S0010-4655(98)00154-4, 1999.

Kollat, J. B., Reed, P. M., and Wagener, T.: When are multiobjective calibration trade-offs in hydrologic models meaningful?, Water Resources Research, 48, 1–19, https://doi.org/10.1029/2011WR011534, 2012.

Kucherenko, S., Klymenko, O. V., and Shah, N.: Sobol' indices for problems defined in non-rectangular domains, Reliability Engineering & System Safety, 167, 218–231, https://doi.org/10.1016/j.ress.2017.06.001, 2017.

Li, G., Rabitz, H., Yelvington, P. E., Oluwole, O. O., Bacon, F., Kolb, C. E., and Schoendorf, J.: Global Sensitivity Analysis for Systems with Independent and/or Correlated Inputs, The Journal of Physical Chemistry A, 114, 6022–6032, https://doi.org/10.1021/jp9096919, 2010.

Lim, T. and Wang, K.: Comparison of machine learning algorithms for emulation of a gridded hydrological model given spatially explicit inputs, Computers & Geosciences, 159, 105 025, https://doi.org/10.1016/j.cageo.2021.105025, 2022.

Liu, Y., Wagener, T., Beck, H. E., and Hartmann, A.: What is the hydrologically effective area of a catchment?, Environmental Research Letters, 15, https://doi.org/10.1088/1748-9326/aba7e5, 2020.

Lo Piano, S., Ferretti, F., Puy, A., Albrecht, D., and Saltelli, A.: Variance-based sensitivity analysis: The quest for better estimators and designs between explorativity and economy, Reliability Engineering & System Safety, 206, 107 300, https://doi.org/10.1016/j.ress.2020.107300, 2021.

Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., Hassler, S. K., Hauffe, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., and Tarasova, L.: CAMELS-DE: hydro-meteorological time series and attributes for 1582 catchments in Germany, Earth System Science Data, 16, 5625–5642, https://doi.org/10.5194/essd-16-5625-2024, 2024.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777, Curran Associates Inc., 2017.

Mai, J.: Ten strategies towards successful calibration of environmental models, Journal of Hydrology, 620, 129 414, https://doi.org/10.1016/j.jhydrol.2023.129414, 2023.

Mara, T. A., Tarantola, S., and Annoni, P.: Non-parametric methods for global sensitivity analysis of model output with dependent inputs, Environmental Modelling & Software, 72, 173–183, https://doi.org/10.1016/j.envsoft.2015.07.010, 2015.

Owen, A. B.: Sobol' Indices and Shapley Value, SIAM/ASA Journal on Uncertainty Quantification, 2, 245–251, https://doi.org/10.1137/130936233, 2014.

Panigrahi, B., Razavi, S., Doig, L. E., Cordell, B., Gupta, H. V., and Liber, K.: On Robustness of the Explanatory Power of Machine Learning Models: Insights From a New Explainable AI Approach Using Sensitivity Analysis, Water Resources Research, 61, e2024WR037 398, https://doi.org/10.1029/2024WR037398, 2025.

Panigrahi, B., Razavi, S., Doig, L. E., Gupta, H., and Liber, K.: On the complementary roles of XVARS and SHAP in explainable AI, Environmental Modelling & Software, 203, 107 007, https://doi.org/10.1016/j.envsoft.2026.107007, 2026.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Puy, A., Piano, S. L., Saltelli, A., and Levin, S. A.: sensobol: An R Package to Compute Variance-Based Sensitivity Indices, Journal of Statistical Software, 102, 1–37, https://doi.org/10.18637/jss.v102.i05, 2022.

Rakovec, O., Hill, M. C., Clark, M. P., Weerts, A. H., Teuling, A. J., and Uijlenhoet, R.: Distributed Evaluation of Local Sensitivity Analysis (DELSA), with application to hydrologic models, Water Resources Research, 50, 409–426, https://doi.org/10.1002/2013WR014063, 2014.

Razavi, S. and Gupta, H. V.: What do we mean by sensitivity analysis? The need for comprehensive characterization of "global" sensitivity in Earth and Environmental systems models, Water Resources Research, 51, 3070–3092, https://doi.org/10.1002/2014WR016527, 2015.

Razavi, S. and Gupta, H. V.: A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory, Water Resources Research, 52, 423–439, https://doi.org/10.1002/2015WR017558, 2016.

Razavi, S., Tolson, B. A., and Burn, D. H.: Review of surrogate modeling in water resources, Water Resources Research, 48, https://doi.org/https://doi.org/10.1029/2011WR011527, 2012.

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Lo Piano, S., Iwanaga, T., Becker, W., Tarantola, S., Guillaume, J. H. A., Jakeman, J., Gupta, H., Melillo, N., Rabitti, G., Chabridon, V., Duan, Q., Sun, X., Smith, S., Sheikholeslami, R., Hosseini, N., Asadzadeh, M., Puy, A., Kucherenko, S., and Maier, H. R.: The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support, Environmental Modelling & Software, 137, 104 954, https://doi.org/10.1016/j.envsoft.2020.104954, 2021.

Rouzies, E., Lauvernet, C., Sudret, B., and Vidard, A.: How is a global sensitivity analysis of a catchment-scale, distributed pesticide transfer model performed? Application to the PESHMELBA model, Geoscientific Model Development, 16, 3137–3163, https://doi.org/10.5194/gmd-16-3137-2023, 2023.

Saltelli, A. and Tarantola, S.: On the Relative Importance of Input Factors in Mathematical Models, Journal of the American Statistical Association, 97, 702–709, https://doi.org/10.1198/016214502388618447, 2002.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S.: Global Sensitivity Analysis. The Primer, Wiley, https://doi.org/10.1002/9780470725184, 2008.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S.: Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index, Computer Physics Communications, 181, 259–270, https://doi.org/10.1016/j.cpc.2009.09.018, 2010.

Schaperow, J. R., Li, D., Margulis, S. A., and Lettenmaier, D. P.: A near-global, high resolution land surface parameter dataset for the variable infiltration capacity model, Scientific Data, 8, 1–14, https://doi.org/10.1038/s41597-021-00999-4, 2021.

Schnorbus, M. A. and Cannon, A. J.: Statistical emulation of streamflow projections from a distributed hydrological model: Application to CMIP3 and CMIP5 climate projections for British Columbia, Canada, Water Resources Research, 50, 8907–8926, https://doi.org/10.1002/2014WR015279, 2014.

Seibert, J.: Estimation of Parameter Uncertainty in the HBV Model: Paper presented at the Nordic Hydrological Conference (Akureyri, Iceland - August 1996), Hydrology Research, 28, 247–262, https://doi.org/10.2166/nh.1998.15, 1997.

Shapley, L. S.: 17. A Value for n-Person Games, in: Contributions to the Theory of Games (AM-28), Volume II, pp. 307–318, Princeton University Press, https://doi.org/10.1515/9781400881970-018, 1953.

Sheikholeslami, R., Gharari, S., Papalexiou, S. M., and Clark, M. P.: VISCOUS: A Variance-Based Sensitivity Analysis Using Copulas for Efficient Identification of Dominant Hydrological Processes, Water Resources Research, 57, 1–24, https://doi.org/10.1029/2020WR028435, 2021.

Stein, L., Clark, M. P., Knoben, W. J. M., Pianosi, F., and Woods, R. A.: How Do Climate and Catchment Attributes Influence Flood Generating Processes? A Large-Sample Study for 671 Catchments Across the Contiguous USA, Water Resources Research, 57, e2020WR028 300, https://doi.org/10.1029/2020WR028300, 2021.

Sun, R., Pan, B., and Duan, Q.: A surrogate modeling method for distributed land surface hydrological models based on deep learning, Journal of Hydrology, 624, 129 944, https://doi.org/10.1016/j.jhydrol.2023.129944, 2023.

Tang, G., Wood, A. W., and Swenson, S.: On Using AI-Based Large-Sample Emulators for Land/Hydrology Model Calibration and Regionalization, Water Resources Research, 61, e2024WR039 525, https://doi.org/10.1029/2024WR039525, 2025.

Vannucci, G., Siciliano, R., and Saltelli, A.: Enhancing Variable Importance in Random Forests: A Novel Application of Global Sensitivity Analysis, https://doi.org/10.48550/arXiv.2407.14194, 2024.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrology and Earth System Sciences, 5, 13–26, https://doi.org/10.5194/hess-5-13-2001, 2001.

Wei, P., Lu, Z., and Song, J.: A comprehensive comparison of two variable importance analysis techniques in high dimensions: Application to an environmental multi-indicators system, Environmental Modelling & Software, 70, 178–190, https://doi.org/10.1016/j.envsoft.2015.04.015, 2015.

Yeste, P., García-Valdecasas Ojeda, M., Gámiz-Fortis, S. R., Castro-Díez, Y., and Esteban-Parra, M. J.: Integrated sensitivity analysis of a macroscale hydrologic model in the north of the Iberian Peninsula, Journal of Hydrology, 590, 125 230, https://doi.org/10.1016/j.jhydrol.2020.125230, 2020.

Yeste, P., Melsen, L. A., García-Valdecasas Ojeda, M., Gámiz-Fortis, S. R., Castro-Díez, Y., and Esteban-Parra, M. J.: A Pareto-Based Sensitivity Analysis and Multiobjective Calibration Approach for Integrating Streamflow and Evaporation Data, Water Resources Research, 59, e2022WR033 235, https://doi.org/10.1029/2022WR033235, 2023.

Yeste, P., García-Valdecasas Ojeda, M., Gámiz-Fortis, S. R., Castro-Díez, Y., Bronstert, A., and Esteban-Parra, M. J.: A large-sample modelling approach towards integrating streamflow and evaporation data for the Spanish catchments, Hydrology and Earth System Sciences, 28, 5331–5352, https://doi.org/10.5194/hess-28-5331-2024, 2024.

Yeste, P., Melsen, L., Lyra Fialho Brêda, J. P., Tacoronte, N., Saltelli, A., Vannucci, G., Siciliano, R., and Bronstert, A.: Data and Code Supplement to "Technical note: Machine learning metamodelling for global sensitivity analysis", https://doi.org/10.5281/ZENODO.19222162, 2026.

Zhang, Q., Zhang, K., Bárdossy, A., Li, Y., and Wu, N.: Improving representation of hydrological process heterogeneity in grid-Xin'anjiang model through a stepwise approach, Journal of Hydrology, 655, 132 897, https://doi.org/10.1016/j.jhydrol.2025.132897, 2025.

Zhu, R., , Donglin, Z., , and Kosorok, M. R.: Reinforcement Learning Trees, Journal of the American Statistical Association, 110, 1770–1784, https://doi.org/10.1080/01621459.2015.1036994, 2015.