



Comparison of feature importance measures and variance-based indices for sensitivity analysis: case study of radioactive waste disposal flow and transport model

Valentina Svitelman¹ · Elena Saveleva¹ · Georgy Neuvazhaev¹

Accepted: 14 November 2024 / Published online: 11 December 2024

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Sensitivity analysis is a crucial step in the development of computational models for any complex system, as it allows for comparison of the relative influence of model parameters on the simulation results. Its role is even more critical in the context of numerical safety assessment for future geological repositories of radioactive waste since it provides insights into understanding relevant to safety processes, establishes grounds for prioritizing additional research, enhances confidence in the safety assessment results. Unfortunately, groundwater flow and radionuclide transport models for radioactive waste repositories often become quite detailed and computationally expensive during the iterative process of the safety assessment. Moreover, any meaningful statistical analysis of the simulation results requires hundreds, if not thousands, of model realizations at different parameter combinations. This necessity has led to the increasing popularity of sensitivity analysis methods coupled with metamodeling approaches where a portion of the points in the parametric space (model realizations) is obtained through numerical simulation of processes, and another portion is approximated from available realizations using less computationally demanding data-driven algorithms. Modern machine learning methods also frequently address the problem of predicting a response function at new points using data from known points. Feature importance measures for machine learning models serve a role akin to sensitivity analysis: they assist in quantifying the effect of model inputs (features) on outputs (predictions). In this paper, we compare the modern first-choice variant of sensitivity analysis, variance-based Sobol' indices, obtained using polynomial chaos expansion metamodel, with well-known importance measures from the machine learning world.

Keywords Feature importance measures · Variance-based sensitivity analysis · Machine learning · Metamodels · Gradient boosting · Numerical safety assessment

1 Introduction

Numerical simulations play a key role in the safety assessment of geological repositories for high-level radioactive waste due to the complexity of the system, which includes the waste itself, and both engineered and natural barriers that are designed and selected to minimize exposure to the environment over the long periods of time (NEA 2013). Even full-scale experiments that may last for several decades

cannot provide a complete understanding of the variety of processes relevant to safety and running at different spatial and temporal scales. The “as best as you can get” solution for this problem is a wide spectrum of various numerical models that simulate the evolution of the repository system using available site-specific data.

Various numerical tools for uncertainty management and, in particular, sensitivity analysis instruments are a significant part of this sophisticated simulation framework: they help to test the adequacy and the robustness of models, identify the parameters that need additional studies, and finally—provide confidence in the obtained assessment results (Vigfusson et al. 2007; Swiler et al. 2021).

In general, sensitivity analysis always answers the question how uncertainty in input variables (parameters) may

✉ Valentina Svitelman
svitelman@ibrae.ac.ru

¹ Nuclear Safety Institute, Russian Academy of Sciences, Moscow, Russia

influence the uncertainty of the simulation results (Saltelli et al. 2008; Pianosi et al. 2016). Numerical simulations typically generate complex outputs, such as multiple variables at various temporal and spatial points, along with additional technical information (for example, software execution logs). A formal approach to sensitivity assessment is to calculate a set of indices that measure the influence of model parameters on the variability in a specific simulated variable (Pianosi et al. 2016). These specific simulated results may be represented by various summary variables or aggregate statistics that characterize the model's performance or the simulation results' compliance with certain criteria. They are usually referred to as the 'quantities of interest' (QoI); however, the term 'outputs' is also used if the context is sufficient to distinguish between the whole simulation output and a specific output variable. In the context of safety assessments for repositories of radioactive waste, the quantities of interest are often represented by various metrics of contaminant transport from the repository to the environment, such as the annual dose rate to an individual, peak concentration of a contaminant (or a conservative tracer as a substitute for the contaminant), activity flux, or sometimes characteristics of a subsystem or group of processes, for example, pH or hydraulic heads (Swiler et al. 2021; Samper et al. 2016; Saveleva et al. 2021). The corresponding parameters may include various flow and transport characteristics of the host rock and engineered barriers, such as porosities, permeabilities, hydraulic conductivities, sorption and diffusion coefficients, as well as the properties of the contaminant source (e.g., rates of waste matrix dissolution or container corrosion).

The simplest type of sensitivity analysis relies on the variation of the input parameters of the model from their nominal values, one at a time (OAT), and assessing the impacts on the simulation outputs. Historically, this approach was widely used due to its simplicity and computational inexpensiveness. However, OAT methods cannot effectively explore a multidimensional space (Saltelli et al. 2019), and meaningful sensitivity analysis for complex non-linear models requires global sensitivity analysis methods. In these methods output variations are induced by varying all the input factors simultaneously, and therefore, the sensitivity to each factor considers not only the direct influence of that factor but also the joint influence due to parameter interactions.

The variety of global sensitivity analysis method originates from the fact that influence of a parameter on the quantity of interest can be expressed by various mathematical measures, such as the proportion of variance, correlation, regression coefficient, conditional probability density function, etc. Consequently, over the last couple of decades, sensitivity analysis has evolved into a distinct discipline

(Razavi et al. 2021) encompassing a vast array of methods and approaches. There are also several variants of sensitivity analysis problem setting (Saltelli and Tarantola 2002; Borgonovo and Plischke 2016; Pianosi et al. 2016; Swiler et al. 2021). For instance, typical sensitivity analysis tasks include prioritization of significant parameters, screening of insignificant ones, the assessment of model additivity and linearity, identification of trends and exploration of input parameter space regions that produce extreme output values. Different mathematical methods may be beneficial for analyzing a particular system (and corresponding model), answering specific questions, or overcoming the limitations of a particular case study. Reviews of methods in publications (Pianosi et al. 2016) and (Swiler et al. 2021) offer a condensed view on matching of methods to sensitivity analysis tasks.

The first choice from this variety is usually variance-based methods, and Sobol' indices in particular, because their calculation requires no assumptions about the linearity of the model, they account for interactions between the model parameters, and easy to interpret (Puy et al. 2022; Swiler et al. 2021). The core idea of variance-based sensitivity analysis is the assumption that input parameters can be treated as stochastic variables, the corresponding simulation results form a distribution in the output space, and the variance is an adequate measure of output uncertainty. Under this assumption, the variance decomposition of an output, obtained based on a large sample of input parameters, can be used to estimate the effect of each parameter (or group of parameters) on the considered output. This sample is usually constructed in a specific way, using Sobol' low-discrepancy quasi-random sequence (Saltelli et al. 2010). The Sobol' sequence demonstrates significantly higher performance for the even exploration of the parametric space than random sampling and also allows for the sequential addition of new points, which is beneficial when the sufficient size of the sample is initially unknown (Razavi et al. 2021; Kucherenko et al. 2015).

One of the persistent issues associated with variance-based sensitivity analysis in the scope of numerical simulations of almost any complex natural and industrial systems is the computational burden (Iooss and Lemaître 2015). To overcome the requirements on the size and structure of an input sample (which are not always feasible due to the corresponding computational costs), Sobol' method is very often coupled with the polynomial chaos expansion (PCE) metamodel (Sudret 2008; Crestaux et al. 2009; Blatman and Sudret 2010). The popularity of the PCE in particular among other metamodels (i.e., approximations that mimic the input–output relationship of the original computational model) results from the fact that sensitivity indices can be computed analytically from the PCE coefficients.

However, this is by far not the only metamodeling method. The general approach was originally proposed under the name ‘response surface methodology’ as a statistical tool for finding the operating conditions of a process that would produce optimal responses (Box and Draper 2007). Subsequent generalizations have led to this approach being used to obtain approximating functions for computationally demanding numerical simulations across various scientific and engineering problems. The mathematical methods suitable for constructing metamodels (also referred to as surrogate models) include linear and polynomial regression, multivariate adaptive regression splines, kriging, and many others (Iooss and Lemaître 2015; Alizadeh et al. 2020). Modern machine learning (ML) algorithms are also widely used for metamodeling purposes, as they demonstrate much better accuracy in predicting outcomes of complex and non-linear phenomena (Iooss and Lemaître 2015).

At this point, sensitivity analysis research very closely approaches the field of machine learning (ML) (Iooss et al. 2022), where models are from the outset data-driven and purposely constructed in such a way that enables recognizing patterns in available data points and making predictions for new ones.

The key distinction between numerical and ML models lies in how they establish the linkage between input and output variables. Numerical models form this connection through hypothesized causal relationships, whereas ML models extract information about input–output relationships solely from the data itself and initially appear as a black box, where the reasoning behind predictions is unknown. The need to reduce this non-transparency, crucial for practical use and acceptance by decision-makers, has naturally evoked the development of diagnostic tools to explain the expected behavior of ML models (Molnar 2022). This is broad class of methods called model-agnostic interpretation/explanation tools, with purposes similar to those of sensitivity analysis. Some of them, usually referred to as feature importance methods, rank model inputs (features) based on their relevance to the predicted output. Other techniques, called feature effects, express how changes in a feature affect the predicted outcome.

Despite developing from different origins and often having different purposes, these two research areas—sensitivity analysis and ML interpretability tools—appear to benefit from influencing and complementing each other. For example, “classical” tools from the sensitivity analysis field such as variance-based methods are known and used for the explanation of machine learning-based predictions as well (Molnar 2022; El Moçayd and Seaid 2021). Some approaches such as the Shapley effects (Owen and Prieur 2017) have been developing and progressing independently both in sensitivity analysis and machine learning areas.

At the same time, the distinctive aspect of ML-based feature importance methods, their independence of usually unknown properties of the model and the relationship between input parameters, makes them quite attractive for analysis of input–output relations of numerical models too. The application of feature importance measures as a sensitivity analysis tool both to benchmark analytical functions and models of real-world systems has shown promising results (Wei et al. 2015; Jaxa-Rozen and Kwakkel 2018; Song et al. 2021; Torres 2021; Cappelli et al. 2023; Cappelli and Grimaldi 2023). In particular, these works have demonstrated the applicability of various feature importance measures, such as permutation-based, Shapley-based, graphical-based (Accumulated Local Effects, ALE), and mean decrease impurity (Gini), for the sensitivity analysis of numerical models.

Results obtained using feature importance approaches were compared with those from a wide range of classical global sensitivity analysis (GSA) methods, for example, variance-based including but not limited to the Sobol’ indices (Jaxa-Rozen and Kwakkel 2018; Song et al. 2021; Cappelli et al. 2023), Morris elementary effects method (Wei et al. 2015; Jaxa-Rozen and Kwakkel 2018), and correlation and regression approaches (Torres 2021; Cappelli and Grimaldi 2023), as well as distribution-based approaches (Cappelli et al. 2023; Cappelli and Grimaldi 2023).

The test cases for sensitivity analyses ranged from analytical functions with known behaviors to models of more complex systems. Specifically, a special case with a known ranking was investigated in (Cappelli and Grimaldi 2023). The Ishigami function (Ishigami and Homma 1990), a well-known analytical benchmark model was considered in (Jaxa-Rozen and Kwakkel 2018; Song et al. 2021), and other benchmarks were used in (Song et al. 2021; Torres 2021). Comparisons between different feature importance measures and GSA methods were performed on simplified versions of models describing real ecological (Wei et al. 2015), biological (Jaxa-Rozen and Kwakkel 2018; Torres 2021), and economic (Jaxa-Rozen and Kwakkel 2018) systems.

In this paper, we present another attempt to explore the practical applicability of feature importance measures for the purposes of sensitivity analysis of numerical models. Some considerations important for global sensitivity analysis were beyond the scope of the aforementioned publications. In particular, we could point out the following gaps.

Firstly, only the classical version of the Sobol’ method was used in the comparisons of ML importance measures and sensitivity indices. This is sufficient for estimating the number of samples required for the analysis, since the Sobol’ method, when coupled with the PCE metamodel, allows for a reduction in the number of required samples (Blatman and

Sudret 2010). Since minimal requirements for the number of realizations are sufficient when selecting a method for dealing with computationally expensive models, it is reasonable to compare importance measure approaches against the widely used metamodel-assisted variant of the method.

A related issue is the potential influence of the type of sampling on the sensitivity analysis results. Since using low-discrepancy pseudo-random sequences significantly improves the approximation required for estimating Sobol' indices (Kucherenko et al. 2015), we also investigated the possible influence of the type of sampling on feature importance performance.

Another consideration was related to the choice of the importance measures for the comparison. Within this selection of methods, the Gini impurity measure, associated with the structure of a tree-based metamodel, appears rather appealing according to the analyses performed for benchmarks and simplified models of real processes (Jaxa-Rozen and Kwakkel 2018; Song et al. 2021; Torres 2021). However, in the sensitivity analysis of the real hydrological model (Cappelli et al. 2023), the Gini measure was not included in the rather wide set of compared methods: two variants of permutation feature importance, the total index and the feature importance measure based on the accumulated local effects plot, along with three different sensitivity indices (variance-based, density-based, and distribution-based). It seems worthwhile to explore the applicability of the Gini importance measure for sensitivity analysis of a complex, nonlinear, and computationally expensive model describing a real system, and compare its performance against the permutation-based and Shapley-based approaches, which are widely used as well.

We believe that our research could complement these investigations regarding the listed above aspects. To be practical, we selected the three most widely-used methods within the ML community from the entire range of importance measures. These methods are implemented in open-source libraries, making them readily available to practitioners from any field. As a baseline for comparison, we used the Sobol' indices with the PCE metamodel, which is currently considered the “gold standard” of global sensitivity analysis.

As a case study for our exercise, we used a three-dimensional groundwater flow and radionuclide transport model of the prospective site for the geological repository of radioactive waste. This model represents an interim stage in the iterative safety assessment process: on the one hand, it is more advanced than the previous cross-sectional model (Valetoev et al. 2019; Saveleva et al. 2021), on the other hand, the site characterization activities are still undergoing, and the final model will incorporate more site-specific data. The current model and its sensitivity analysis though

could be instrumental in providing useful insights about the system behavior (e.g., its inherent three-dimensionality) and even bringing appropriate corrections into the site investigation activities.

The final purpose of this exercise is not to decide which approach is universally better but to understand how they could complement each other for the sake of a better understanding of the model, especially in cases where a limited number of simulations is available.

2 Material and methods

Let $X = (X_1, \dots, X_M)$ be the vector of input parameters, M be the number of parameters, $y = f(x)$ be the output of the numerical model selected as a quantity of interest and also a subject of metamodeling, and N be the number of realizations of vector X in the stochastic sample.

2.1 Variance-based sensitivity analysis and polynomial chaos expansion metamodel

Nowadays the most commonly used measure to quantify the sensitivity of simulated outputs to the parameters of the numerical model is Sobol' indices (Saltelli et al. 2008). This approach belongs to the class of variance-based global sensitivity analysis (GSA) methods. “Global” in this context means that all uncertain parameters are varied simultaneously and therefore parametric space is studied evenly. “Variance-based” refers to the interpretation of the analysis results: it considers the variance as an adequate statistical measure of the uncertainty of the estimated quantity of interest. The attractiveness of variance-based sensitivity analysis methods is associated with their ability to deal with nonlinear models, as they can measure the effect of parameter interactions in non-additive systems.

Considering input parameters $X_i, i = 1, \dots, M$ as independent stochastic variables resulting in an output variable y selected as quantity of interest, the variance of y can be decomposed into the sum of the direct effects of the M parameters on y and the joint effects of combinations of parameters (Sobol' 2001; Saltelli et al. 2008):

$$V = \sum_i V_i + \sum_{i < j} V_{ij} + \sum_{i < j < m} V_{ijm} + \dots + V_{123\dots M} \quad (1)$$

where $V_i = V(E(y|X_i))$ is the part of the variance corresponding to the direct effect of the input parameter X_i , V_{ij} is the part corresponding to the joint influence of input parameters X_i and X_j , and so on. The joint influence of two or more input parameters could be greater than zero if the cumulative effect on the variance of y is not equal to the sum of their individual direct effects. Sobol' main (or

first-order) indices represent the direct effect of each parameter on the variance of the selected output of the model y and are estimated by dividing the corresponding variance V_i by the variance V of the output variable (QoI) (Sobol' 2001; Saltelli et al. 2008):

$$S_{1i} = \frac{V_i}{V} = \frac{V_{X_i}[E_{X_{\sim i}}\{y|X_i\}]}{V}, \quad (2)$$

where $E_{X_{\sim i}}\{y|X_i\}$ is the expectation of the output variable while varying all input parameters except X_i , and V_{X_i} is the variance of output variable due to the variation of X_i . The overall contribution of a parameter X_i to the variance of output variable y , which includes both its direct effect and its joint effects with the other input parameters (so-called interactions), is described by Sobol' total indices (Sobol' 2001; Saltelli et al. 2008):

$$S_{Ti} = \frac{V_{Ti}}{V} = \frac{E_{x_{\sim i}}[V_{x_i}(y|x_{\sim i})]}{V}, \quad (3)$$

The standard approach to the numerical estimation of Sobol' indices is Monte-Carlo integration (Sobol' 2001; Saltelli et al. 2010). The practical inconveniences of this method are that it requires a specific strategy for generating a sample of input parameters and that the sample size should be quite significant (at least $N \sim 500 \times M$, where M is the number of parameters) (Pianosi et al. 2016; Swiler et al. 2021) to obtain reliable results. An alternative for cases where a sufficient number of model runs might be computationally demanding or where the required sample design is unobtainable is the estimation of Sobol' indices using metamodel. In particular, metamodels based on Polynomial Chaos Expansion (PCE) provide such an approximation by casting the model outputs onto orthogonal polynomials (Sudret 2008; Crestaux et al. 2009; Blatman and Sudret 2010).

The idea of the PCE method is representing a random variable as a polynomial function of other random variables, where the polynomials are chosen to be orthogonal with respect to the joint probability distribution of these random variables. Particularly, the method is based on the expansion $g_{PC}(X)$ of the random response (model-based estimate of QoI) $y = g(X)$ into a set of orthogonal polynomials Ψ_i (Xiu and Karniadakis 2002).

$$y = g(X) \approx g_{PC}(X) = \sum_{i=1}^{P-1} g_i \Psi_i(X), \quad (4)$$

where P is the limit of the indefinite sum which gives a sufficient approximation of the response. The choice of the polynomial basis is associated with the distribution of the input parameters. For example, when input parameters follow a normal distribution, the family of multivariate Hermite polynomials can be used as an orthogonal basis

(Baudin et al. 2015). As for a uniform distribution of input parameters, which is the case in this paper, the multivariate Legendre polynomials are employed.

The coefficients are obtained by minimization of the approximation error (Sudret 2008; Blatman and Sudret 2010). After that Sobol' indices (2) could be calculated as:

$$S_i = \frac{\sum_{j \neq i} g_j^2 E[\Psi_j^2]}{V(y)}. \quad (5)$$

And Sobol' total indices (3) then could be calculated as:

$$S_{Ti} = \frac{\sum_{j \in \mathfrak{S}_i} g_j^2 E[\Psi_j^2]}{V(y)}, \quad (6)$$

where $\mathfrak{S}_i$ is the set of all indices with a non-zero i th component.

2.2 Machine learning metamodels and feature importance measures

The specificity of feature importance assessment lies in the obligatory presence of a surrogate model (metamodel) trained on the input–output samples of the numerical model to make predictions. Modern machine learning algorithms show great ability to build accurate metamodels with strong predictive capabilities, even for systems with complex behavior and nonlinear phenomena (Iooss and Lemaître 2015; Molnar 2022). There is a wide range of available ML techniques suitable for these purposes, each with mathematically distinct approaches: tree-based, such as random forest (Breiman 2001) or gradient boosting (Mason et al. 1999), kernel-based (e.g., support vector regression, SVR) (Vapnik 1995), Gaussian regression (Rasmussen and Williams 2006), etc. An important requirement while applying ML techniques is that the resulting metamodel should reproduce the results of the numerical model adequately well. This can be verified, for instance, by the R^2 metric, coefficient of determination (Devore and Carlton 2016):

$$R^2 = 1 - \frac{\sum_{i=1}^L (\hat{f}_i - y_i)^2}{L \cdot \sigma^2}, \quad (7)$$

where $\hat{f}_i$ and y_i are, respectively, the metamodel approximation and the result of the numerical model at the same point within parametric space, L is the number of points compared, and σ^2 is the dispersion of the numerical model results. The R^2 metric provides a simple and intuitive measure of how well the numerical model is replicated by the metamodel, based on the proportion of the total variation of the model results predicted by the metamodel. It normally ranges from 0 to 1, where a value of 0 indicates that

the trained metamodel does not explain any variability in the resulting variable, and a value of 1 means a perfect fit. Negative values may arise when the metamodel predictions being compared to the corresponding numerical model outcomes have not been derived using those outcomes.

In our case, the gradient boosting approach appeared to produce the most suitable surrogate model among several tested methods (comparison results are provided in the “Results and Discussion” section). The gradient boosting, along with the decision tree and the random forest approaches, belongs to the class of tree-based algorithms. Their main idea is as follows (Fig. 1a): a decision (classification or regression) tree comprises the root, internal nodes, and leaves (terminal nodes). Each internal node has a decision rule on how to split the dataset into two subsets with similar responses within each. These rules are inferred from prior training data based on the criterion of entropy maximization.

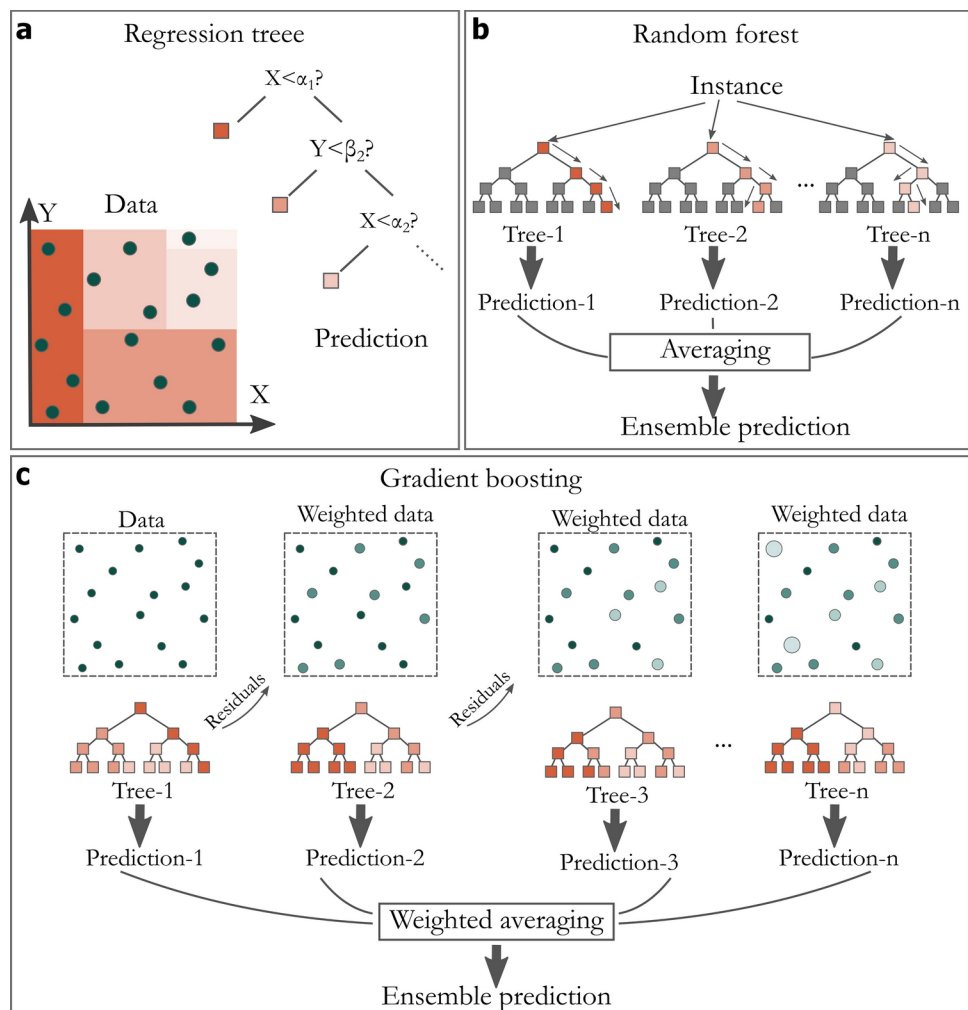
A commonly known problem with decision tree metamodels is their tendency to become overly complex, leading to

overfitting. Metamodels based on an aggregation of multiple trees were introduced to address this issue and improve classification or regression accuracy. The simplest way to aggregate the predictions of multiple decision trees is the random forest method (Breiman 2001). It employs multiple classification or regression trees trained on various subsets of the given dataset and averages the results to improve predictive accuracy (Fig. 1b). Gradient boosting (Mason et al. 1999) constructs an ensemble of weak prediction models (simple classification or regression trees) in another way; it is trained iteratively: the subsequent model focuses on correcting the errors of the previous one (Fig. 1c).

2.2.1 Gini importance

The Gini importance measure was specifically designed for tree-based algorithms and is computed directly from the tree structure (Breiman 2001). Gini impurity is a measure that indicates the prediction error at a given terminal node (t). For decision trees, it is described by the likelihood of misclassifying (appointing to the wrong leaf) an output:

Fig. 1 Schematic diagram for tree-based algorithms for regression problem: **a** regression tree; **b** random forest; **c** gradient boosting



$\sum_{i=1}^n p_i \cdot (1 - p_i)$, where n is the number of leaves and p_i is the probability of reaching leaf i . For regression trees, the impurity is described by the mean square error:

$$i_R = \sum_{i=1}^M \left(y_i - \hat{f}_{(i)} \right)^2, \quad (8)$$

where y_i is an output from the training dataset, and $\hat{f}_{(i)}$ is the mean of the values in the leaf that case i belongs to.

To calculate the Gini importance of a particular input parameter (feature) j , one can assess how it influences the impurity of each individual tree and the average over all trees in the ensemble:

$$G(j) = \frac{1}{T} \sum_{\tau=1}^T \sum_{t \in \varphi_\tau} 1 \cdot (s_t == j) [\Delta i_R(s_t, t)], \quad (9)$$

where s_t is the variable used for splitting the node (t), $\Delta i_R(s_t, t)$ is the difference in impurity (8) due to the corresponding split, T is the number of trees φ_τ in the ensemble.

The drawback of this method is its tendency to favor numerical features and categorical features with high cardinality. Moreover, in the case of correlated features, it may detect one of them and neglect the importance of the other, which can lead to wrong conclusions.

2.2.2 Permutation importance

Permutation feature importance overcomes the limitations of impurity-based feature importance: it does not exhibit bias toward high-cardinality features and can be computed on a test set that was not used for the model training. The method is applied to an already trained model $\hat{f}$ with an available error measure (mean square error), $MSE(y, \hat{f}(X))$, where y is the known correct output and X is the vector of inputs.

To estimate the importance of a parameter (feature) j , the values corresponding to this feature in the input vector are randomly shuffled to produce a different vector of inputs $X_{(j)}$ which corresponds to unknown from training data output.

The importance measure of a feature is defined as the absolute value of the difference between the mean square errors of permuted and initial predictions:

$$P(j) = \left| L(y, \hat{f}(X_{(j)})) - MSE(y, \hat{f}(X)) \right|. \quad (10)$$

The input that most significantly impacts prediction performance is considered the most important. This technique benefits from being model-agnostic and can be calculated many times with different permutations of the feature. It is noteworthy that under condition of input independence,

permutation feature importance combined with a quadratic loss function is equivalent to variance-based total sensitivity (Hooker et al. 2021).

2.2.3 SHAP

SHAP (SHapley Additive exPlanations) is a method to explain individual predictions of different machine learning models (or, in our case, metamodels for a numerical model), inspired by game theory. It is based on so-called Shapley values, a solution concept in cooperative games (Roth 1988).

The setup of a problem describing a cooperative game is as follows: a coalition of M players cooperate, and individual players expect a fair share of the total gain. The Shapley value for each player (or a group of players) is the quantity that indicates the worth of coalitions with them. One possible mathematical definition of Shapley value for player j is (Lundberg and Lee 2017):

$$\varphi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{j\}}(x_{S \cup \{j\}}) - f_S(x_S)], \quad (11)$$

where F is the whole set of players and S is its subset, $|F|$ and $|S|$ are their sizes, and f is the resulting gain of the game for the corresponding group.

In sensitivity analysis, the input factors are considered as the players, and Shapley values indicate how input parameters (features) affect the output (prediction). The calculation of Shapley values using Eq. (11) is computationally expensive, it requires at least $2^{|F|}$ model launches (Plischke et al. 2021).

SHAP allows simplifying the calculations by constructing an additive representation (linear explanation model) of Shapley values for a prediction $y = \hat{f}(x)$ based on x , individual stochastic realization of parameter vector X (Lundberg and Lee 2017):

$$g(z') = \phi_0 + \sum_{j=1}^M \phi_j z'_j, \quad (12)$$

where g is the explanation model, $z' \in \{0, 1\}^M$ is a coalition vector of inputs (1 if the input is present in the coalition, and 0 — if it is absent), z'_j is the coalition containing input j , ϕ_0 usually refers to the mean value over metamodel $\hat{f}(x)$ outputs, $\phi_j(i) \in R$ is a Shapley value of feature j for i th metamodel output $\hat{f}(x_i)$, M is the number of features. The explanation model attributes the effect of each feature so that their sum approximates the output of the original numerical model. There are several special variations of the SHAP approach depending on the internal properties (structure) of

the original metamodel such as Tree SHAP, Kernel SHAP, Max SHAP, and Deep SHAP (Molnar 2022).

The SHAP feature importance is calculated as the average of absolute Shapley values per feature across the whole dataset:

$$Sh(j) = \frac{1}{N} \sum_{i=1}^N |\phi_j(i)|, \quad (13)$$

where N is the number of samples in the dataset.

2.3 Computational implementation

All analyses were carried out in the Python language using well-known open-source libraries. In particular, SALib (Iwanaga et al. 2022) was used for generating Sobol' sequences, and openTURNS (Baudin et al. 2015) for evaluating of Sobol' indices based on the PCE metamodel. Random Forest Regression, Support Vector Regression, and Gaussian Process Regression metamodels, as well as the calculation of Gini and permutation importance measures, were implemented using the scikit-learn library (Pedregosa et al. 2011). The XGBoost library was used for the gradient boosting metamodel (Chen and Guestrin 2016), and the SHAP library (Lundberg and Lee 2017) for SHAP importance measures. Regression tree visualization was implemented using the dtreeviz library (Parr et al. 2023).

The interaction with the investigated model was implemented in the MOUSE software package developed as a part of the numerical safety assessment framework (Linge et al. 2020). It provides an interface for applying widely-known statistical methods to models developed within this software framework, with knowledge of the file formats and data structures of specific simulation software packages. For instance, in the case of the GeRa flow and transport simulation code (Konshin and Kapyrin 2017) employed in this work, model data are stored in a binary file with a specific complex structure due to the need to work with very detailed computational meshes. During the model development process, the user defines model parameters and selects observation points via the graphical interface, which is not an option for the stochastic variation of parameters during sensitivity analysis or model calibration. Information about monitoring points is stored in .sqlite database files, and simulated outputs include multiple groups of files in various formats (e.g., VTK files for fields of variables, text, binary, or JSON files for summary information by monitoring points, etc.). MOUSE facilitates interaction with GeRa through templates of interaction files introduced specifically for this purpose, along with instructions for extracting quantities of interest. The instruction can be organized in multiple ways: either by defining keywords in the output

file and then specifying row and column numbers relative to the keywords or by a sequence of keys that locate the value within the JSON structure.

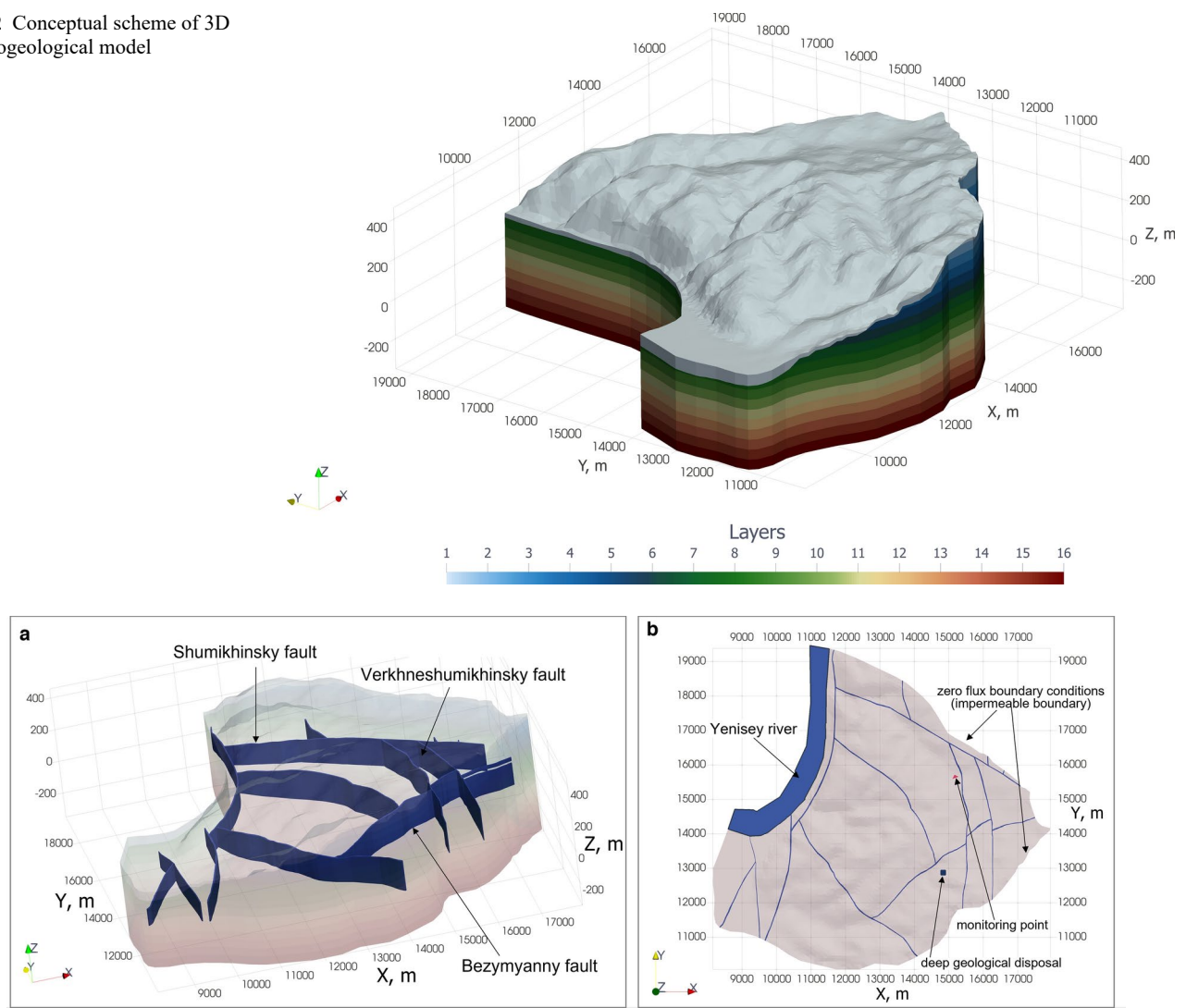
3 Case study description: groundwater flow and radionuclide transport model

To illustrate the applicability of feature importance measures as sensitivity analysis tools, in this paper, we use the three-dimensional groundwater flow and radionuclide transport model of the Yeniseysky site. This site has been selected as a potential location for a future geological disposal facility for high-level radioactive waste in Russia. Implementing such a facility, which meets safety requirements and can be practically constructed, operated, and closed, is a long and multidisciplinary process. Currently, a license has been obtained for the construction of the underground research facility (URF). The planned URF experiments, along with additional, more detailed site surveys in the next stages of the investigations, are expected to confirm the site's suitability for the construction of the disposal facility.

The development of computational models and corresponding numerical safety assessment results are also essential parts of the deep geological disposal development strategy (Dorofeev et al. 2017; Rosatom 2018). The model described below represents the current state of understanding of groundwater flow and transport processes in the far field of the disposal facility. The model will evolve significantly as new data is provided by the studies in the URF or site surveys. On the other hand, uncertainty and sensitivity analysis at the current stage of the models could provide insights into the simulated system that, in turn, could influence further studies, e.g., by providing feasible arguments for prioritizing specific experimental or field research activities.

The current flow and transport model is based on the geological structure described in (Morozov et al. 2019) and developed using the continuum porous medium approach implemented in the GeRa software (Konshin and Kapyrin 2017; Kapyrin et al. 2018; Bagaev et al. 2019).

The model conceptualization relies on data from the previous stage of site surveying, which included coring in more than 20 deep (~700 m) boreholes, geophysical surveys of the territory, and hydrogeological tests (Morozov et al. 2019; Rastorguev et al. 2021). Groundwater flow conditions are considered steady-state. The direction of the average groundwater flow is predominantly in the north and northwest. The main volume of water inflow (about 82%) is intercepted by small watercourses, with only about 18% discharged directly into the Yenisey River.

Fig. 2 Conceptual scheme of 3D hydrogeological model**Fig. 3** Model boundaries and considered fault zones: **a** side view; **b** top view

The domain comprises 16 geological layers (Fig. 2), this discretization was introduced in accordance with test interval lengths in order to calibrate the model against the observed head data. Regional faults within the model's boundaries are considered. Quantifying their influence is significant for long-term safety assessment since they could serve as preferential paths for groundwater flow and contaminant transport. The faults were considered vertical because survey data indicate they are steep-dipping ($\sim 70\text{--}80^\circ$ dip angle). The boundaries of the model and the fault lines are shown in Fig. 3. Zero flux boundary conditions (impermeable boundary) were set at all side boundaries except the northern one. A Dirichlet boundary condition was set on the Yenisey River (constant heads equal to its absolute mark), and a Neumann boundary condition (recharge by infiltration) was set on the top of the modeled area.

The computational grid was constructed using a triangular-prismatic grid generator. The final mesh contains 312,487 cells across the entire computational domain.

The transport model output provides the 3D field of contaminant concentrations. A typical performance metric at the final stages of safety assessment for the radioactive waste disposal system is the dose to a member of the public. However, since the exposure to the public occurs via the aquifer, the commonly used substitute performance metric for the earlier stages, when a lot of site-specific information still needs to be obtained, is the concentration of a neutral tracer. For the purposes of this paper, the peak concentration of the neutral tracer (^{129}I) at the monitoring point situated in layer 2 in the area of groundwater discharge (in the upstream reaches of the Shumikha River) was investigated as the quantity of interest.

The monitoring point location is shown in Fig. 3. The variants of model conceptualization and parameterization are not discussed here in detail as they are the subject of different ongoing research (Neuvazhaev et al. 2020). As part of the iterative process of safety assessment, new versions of the site model are developed and calibrated as new site-specific data becomes available, and sensitivity analysis is carried out at each stage for each model version.

The parameters characterizing the main processes in the model of groundwater flow under unconfined conditions (mathematical details of the model are provided in the Supplementary Materials) are the hydraulic conductivities and porosities of all matrix layers and faults, as well as the infiltration rate. The same parameter values were assigned to layers with similar origins and properties: layer 1 is Quaternary deposits, layers 2–8 belong to the bedrock, and layers 9–16 belong to the confined aquifer. The parameters of the three nearest to the URF faults (Bezmyanny, Shumikhinsky, Verkheshumikhinsky) were defined separately, while the hydraulic conductivity and porosity for the other faults were set as joint parameters for all of them.

For the sensitivity analysis studies, all 14 parameters were varied, the corresponding ranges are provided in Table 1. The reasoning for these ranges is as follows. Infiltration (recharge) corresponds to the average groundwater runoff which is equal to 103.6 mm/year or $2.8 \cdot 10^{-4}$ m/day, the value set in the model. In the region under consideration, the value conditioning groundwater runoff varies from 19 to 25% of the total annual precipitation. The sum of annual precipitation ranges from 450 to 600 mm per year. The range for sensitivity analysis is chosen with a reasonable margin because currently there are no objective data for estimating

the balance and discharge to surface runoff for the whole modeled area.

The ranges for the hydraulic conductivity of the Bezmyanny, Shumikhinsky, Verkheshumikhinsky, and other faults, and the hydraulic conductivity of Quaternary deposits are selected based on the results of 3D modeling of hydrogeological pumping tests and their analytical interpretation (Rastorguev et al. 2021). Since fault zones potentially could act as fast transport paths from the designed facility, the upper limits of the range of hydraulic conductivity for them were set higher.

The ranges for the hydraulic conductivity for the gneiss matrix of the upper (from the 2nd to the 8th) layers and the lower (from the 9th to the 16th) layers of the model were selected based on slug-test studies conducted at the Yeniseysky site for absolute elevations from 0 to 410 m (Rastorguev et al. 2021).

The ranges for the porosity of the matrix for the upper (2nd to 8th) and the lower (9th to 16th) layers were selected based on coring data sampled from various depths (Morozov et al. 2019).

The ranges of porosity values for the faults (Bezmyanny, Shumikhinsky, and other faults) represent higher values than the matrix because some microfractures were observed for corresponding rocks (dolerite dykes), and microfractures contribute to an increase in overall porosity (Morozov et al. 2019).

Figure 4 shows the time series of the mean concentration at the aforementioned monitoring point, accompanied by 95% and 50% confidence intervals. The min–max spread is too wide for a meaningful plot, minimal and maximum values at the peak are $2 \cdot 10^{-28}$ and $6.58 \cdot 10^7$, respectively. Statistics were calculated over a sample of 6000 stochastic realizations (quasi-random Sobol' sequence) further used in the sensitivity analysis. The sample size was bound by the computational costs of one simulation run and the formula $N = n \cdot (2M + 2)$ for the sampling design, Saltelli's extension of the Sobol' sequence (Saltelli 2002), which allows the estimation of main, total, and second-order indices. N here is the full sample size, M is the number of parameters, and n is the size of the subsample used to estimate an individual effect of one parameter.

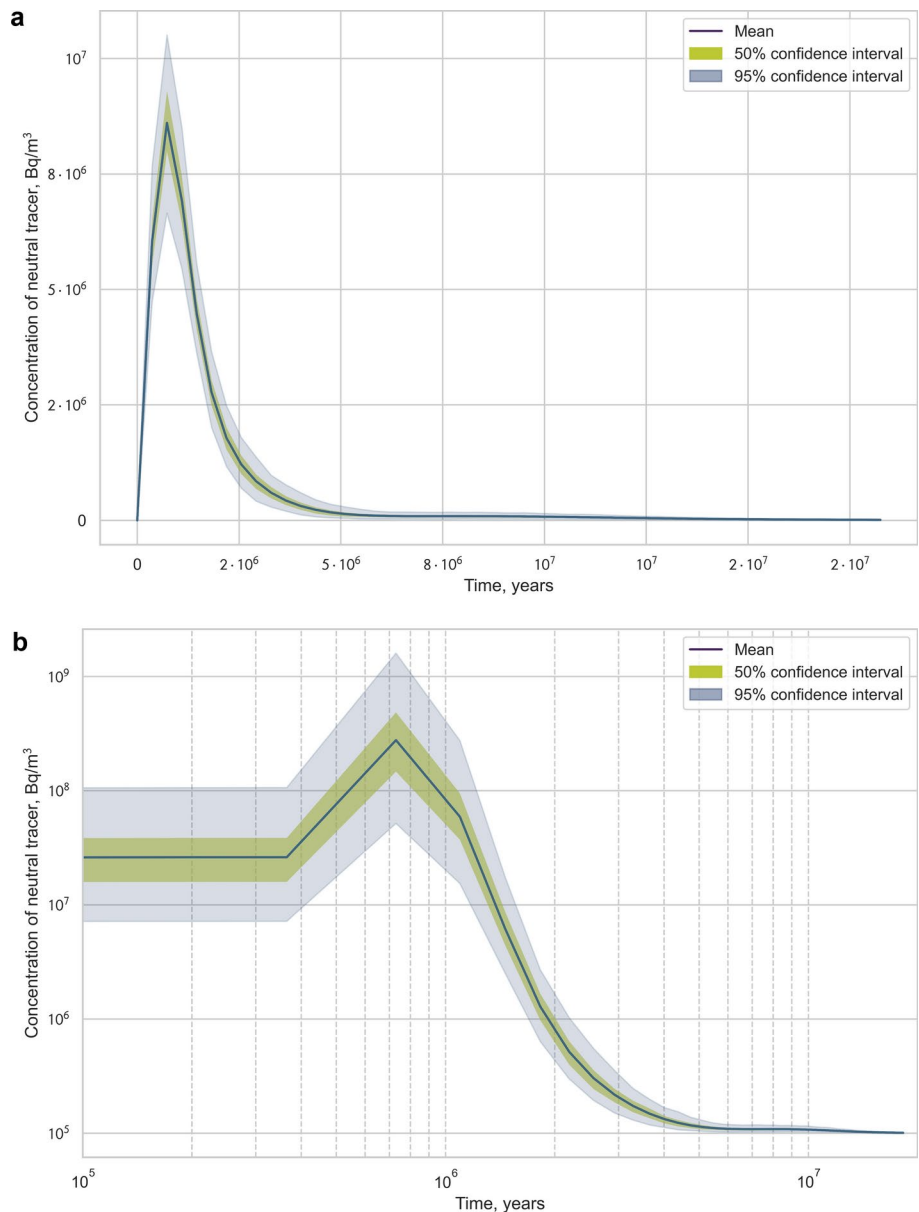
4 Results and discussion

For the purposes of this study, two sets of 6000 realizations of stochastic parameter combinations were generated within their respective variation ranges: the first sample represented a Sobol' quasi-random sequence (Saltelli et al. 2010), which is considered to fill parametric space more evenly and provides better convergence of sensitivity analysis results

Table 1 Parameter ranges used in sensitivity analysis

Parameter	Variation range
Infiltration rate [m/day]	[0.0001; 0.0004]
Hydraulic conductivity: Bezmyanny fault [m/day]	[0.001; 1]
Porosity: Bezmyanny fault [-]	[0.05; 0.25]
Hydraulic conductivity: Shumikhinsky fault [m/day]	[0.001; 1]
Porosity: Shumikhinsky fault [-]	[0.05; 0.2]
Hydraulic conductivity: Verkheshumikhinsky fault [m/day]	[0.001; 1]
Porosity: Verkheshumikhinsky fault [-]	[0.05; 0.25]
Hydraulic conductivity: other faults [m/day]	[0.001; 1]
Porosity: other faults [-]	[0.05; 0.25]
Hydraulic conductivity: layer 2–8 matrix [m/day]	[0.00005; 0.001]
Porosity: layers 2–8 matrix [-]	[0.001; 0.01]
Hydraulic conductivity: layers 9–16 matrix [m/day]	[0.00001; 0.001]
Porosity: layers 9–16 matrix [-]	[0.001; 0.01]
Hydraulic conductivity: Quaternary deposits (layer 1) [m/day]	[0.05; 1]

Fig. 4 Neutral tracer concentration at the monitoring point: mean, 95%, and 50% confidence intervals: **a** linear scale; **b** logarithmic scale



(Kucherenko et al. 2015), while the second sample was constructed by independent generation of 6000 uniformly distributed values for each parameter (further referred to as ‘uniformly distributed’). Since quasi-random sampling is known to provide better results for the estimation of variance-based sensitivity indices (Lo Piano et al. 2021), it was interesting to observe the effect of the sampling strategy on ML importance measures.

Due to the considerations regarding the sample size and presumable high non-linearity of the model under investigation, we did not include direct (without a metamodel) estimations of Sobol’ indices in the comparison. Initially, as mentioned in the previous section, we generated a dataset of the shape that allows direct calculation of main, total and 2nd-order indices with the size at the lower boundary of

acceptable (Pianosi et al. 2016; Swiler et al. 2021). Although the sample size was quite large from the computational cost perspective, we found the results of the direct calculations of Sobol’ indices inadequate (e.g., a lot of slightly negative indices, which is a known issue for not large enough samples for catching highly non-linear model behavior). Just for reference, we included these results into the Supplementary materials.

Unfortunately, the model is constructed on a quite detailed computational mesh, which is not uncommon for models in the context of safety assessment of radioactive waste disposal systems. The sensitivity analysis of almost any kind for such models is predictably computationally expensive. If one model run takes about an hour on personal computer (as it did in this case), even with parallelization possibilities,

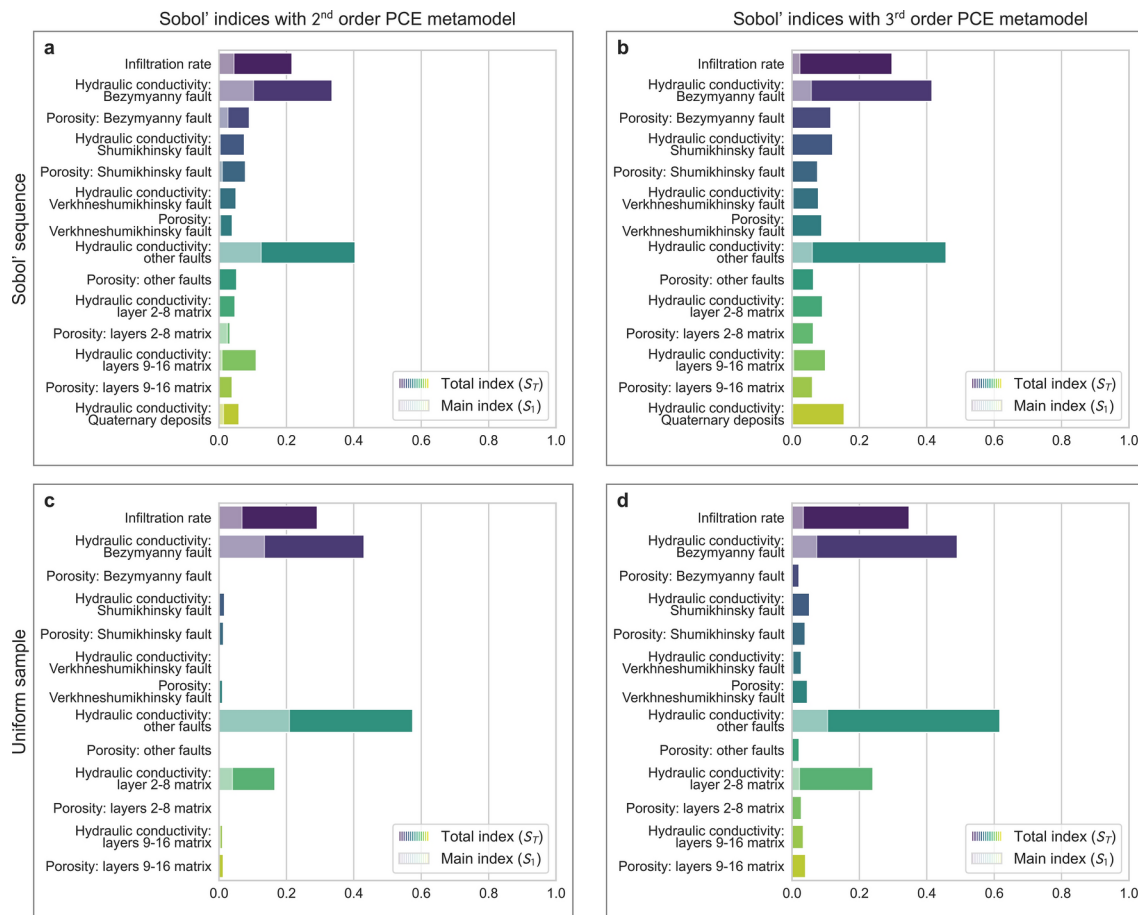


Fig. 5 Sobol' indices with 2nd-order PCE versus 3rd-order PCE for two sampling strategies

Table 2 Quality of prediction (R^2) for the PCE metamodel on a test dataset comprising 10% of the data

Type of sampling	Polynomial order						
	1	2	3	4	5	6	7
Sobol' quasi-random sequence	0.08	0.294	0.673	0.13	0.275	0.5	0.628
Uniformly distributed sample	0.076	0.218	0.347	0.306	0	0	0

additional several thousands of realizations take weeks to obtain. That is why we arrived at a decision to use this case study to explore the metamodeling possibilities.

As the first step of the analysis, Sobol' indices were estimated using the PCE metamodel with 2nd-order and 3rd-order polynomials (Fig. 5). The prediction quality (R^2 score) is higher for the 3rd-order metamodel (Table 2), although both surrogates show mostly similar parameter rankings. The main difference is that the results for the metamodel with the 3rd-order polynomial (Fig. 5) show a larger difference between the total indices and the main indices (i.e., the share of joint effects of multiple parameters) and a slightly increased influence of the least significant parameters. It is interesting that despite the R^2 score being significantly lower for the metamodels obtained based on the dataset with uniform sampling, these metamodels give almost the same parameter rankings with few discrepancies concerning the

parameters with low significance. In particular, the results for the two different sampling strategies contradict with regard to the ranking significance of the hydraulic conductivities of bedrock (layers 2–8) and confined aquifer (layers 9–16). Given the potentially insufficient number of stochastic realizations for accurately describing the “real” numerical model and the sensitivity indices, it is safe to assume that this effect (low but noticeable influence of the hydraulic conductivity of layers 2–8) might be introduced by the metamodel itself. In further comparison of overall ranking by different approaches, we use the results obtained using 3rd-order PCE on the Sobol' quasi-random sample.

Another technical point worth mentioning: in our case, we used the full PCE variant where the basis size is computed from the predefined polynomial order (Table 2), and for the approximation of the coefficients, the least squares strategy is employed, which minimizes the quadratic error

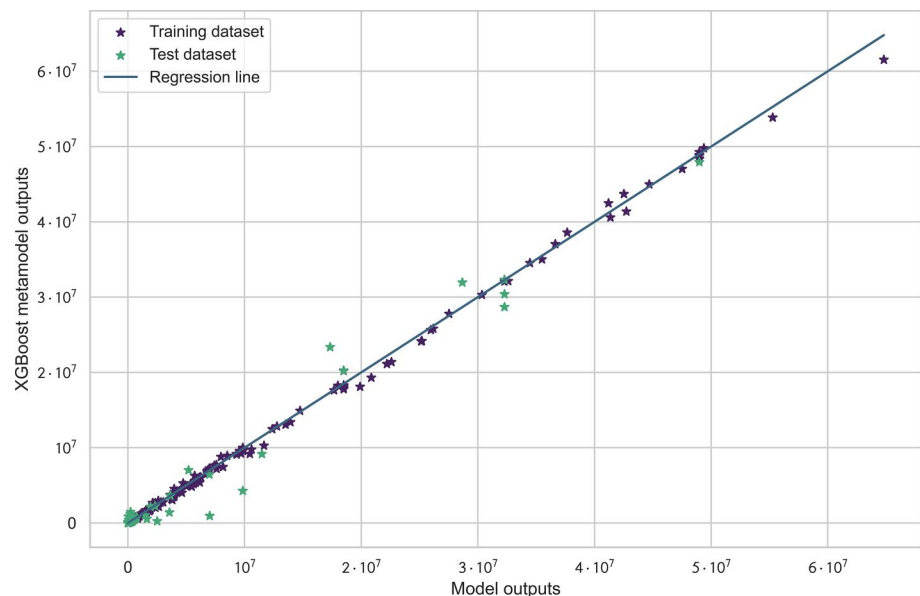
Table 3 Hyperparameters of the ML metamodels and quality of their prediction (R^2) on a test dataset comprising of 10% of the data

Type of sampling	Gradient boosting (XGBoost)	Random forest	Support vector regression (SVR)	Gaussian regression
Sobol' quasi-random sequence	0.99	0.94	0.6	0.63
Uniformly distributed sample	0.98	0.93	0.59	0.33
Hyperparameters of the metamodels	Number of trees in ensemble: 100 Max depth of the tree: 3	Number of trees in ensemble: 100 Max depth of the tree: 5	Kernel: RBF (radial basis function) (Rasmussen and Williams 2006)	Kernel: RBF (radial basis function), length scale of the kernel = 1.0 (means isotropic kernel) (Rasmussen and Williams 2006)

between the initial model response and the polynomial approximation (Björck 1996). It is worth mentioning that there is evidence that in similar cases involving high-dimensional nonlinear problems, the sparse form of PCE, i.e. the expansion with the truncated functional (Abraham et al. 2017; Lüthen et al. 2021), often performs better than classical full PCE. Yet this did not appear to be the case for our model. The corresponding test results obtained with sparse PCE based on least angle regression (Blatman and Sudret 2011) are provided in the Supplementary Materials. However, comparing different modifications of the PCE metamodel could be an interesting avenue for further investigations with this model when a larger sample is available.

As an alternative to Sobol' indices with the PCE metamodel, we decided to explore the potential of feature importance measures in combination with gradient boosting (XGBoost) metamodel. The choice of XGBoost was guided by several considerations. First, tree-based metamodels were attractive for our purposes from the start since they provide more explainability compared to many other metamodeling approaches. And second, XGBoost metamodel demonstrated higher approximation quality compared to several other methods, with R^2 (Eq. 7) values of 0.99 and 0.98 for quasi-random and uniform samples, respectively (Table 3, Fig. 6). It is evident that R^2 values for all methods applied to a uniformly distributed sample are lower than those applied to a Sobol' quasi-random sequence. This discrepancy may be due to the main difference in their generation: the Sobol' quasi-random sequence has a unique feature: it covers the entire parameter space, increasing the density of samples with each step, while the other sampling approach lacks such design features. Consequently, the test data (the first 10% of the total set of points) from the uniformly distributed sample could be more distant from the training set. It appears that the PCE and the Gaussian regression methods are particularly sensitive to this factor.

The application of all metamodels listed in Table 3 was preceded by the transforming the output to achieve a distribution close to normal. This transformation was performed to eliminate differences in the scales of the output values. No transformation was made for the input parameters, as they are considered uniform within the ranges presented in Table 1. It should be mentioned that no significant efforts were made to fit the metamodels; it is possible that some of them could perform much better. The significant hyperparameters of the metamodels are also provided in Table 3,

Fig. 6 Quality of the XGBoost metamodel prediction

other parameters were set to the default values recommended in the documentation of the corresponding libraries listed in Subsect. 2.3. The good performance of XGBoost was considered sufficient, as the focus of the study was on feature importance measures and not on a detailed comparison of metamodeling approaches.

Figure 7 displays the results of importance measure estimation using different approaches for the same two samples. All three approaches agree on the ranking of the four most significant parameters, with some differences in the identification of the least significant ones. Both the variance-based Sobol' and ML importance measures approaches rate the same parameter as the most influential. It is the hydraulic conductivity of other faults (all faults excluding Bezymyanny, Shumikhinsky, and Verkhnesumikhinsky). From a hydrogeological perspective, this is explained by the consideration that the main portion of contaminant migration occurs along fault zones, which possess higher hydraulic conductivity characteristics and lower porosity, resulting in high average velocities.

Another aspect that should be noted is that from the beginning we were particularly interested in metamodeling techniques that could be internally linked with the estimation of parameter significance, such as the combination of Sobol' indices and PCE metamodel or importance measures and tree-based metamodels. In our opinion, these combinations of approaches could have a better chance of utilizing all available information. However, we tested the primitive solution of simply expanding the dataset using a suitable

metamodel (XGBoost) and then directly estimating sensitivity indices. Predictably, this did not yield significantly more satisfactory results than the direct estimation of Sobol' indices from initial dataset since the XGBoost reproduces the same data structure. The details can also be found in the Supplementary materials.

Since from the practical point of view the relative ranking of the parameters is more valuable than the absolute values of sensitivity measures, the summary of the parameter rankings by different approaches (Sobol' indices obtained based on 3rd order PCE and the importance measures obtained based on XGBoost) is provided in Table 4. The smaller number means higher importance (1 equals the most significant parameter), and the ranks lower than 5 are not displayed. The ranking of the three parameters that appeared most important is fully consistent with the expectations of an expert hydrogeologist (model developer). The faults could affect the resulting concentration of the contaminant, and the Bezymyanny fault, being closer to the URF and the future disposal facility than other large ones, is significant, as some fraction of the contaminant could travel through it. The two main dissimilarities between the importance measures and the Sobol' indices are as follows. Firstly, the second and third most significant parameters (the infiltration rate and the hydraulic conductivity of the Bezymyanny fault) are ranked in reverse order. Secondly, importance measures distinctly highlight the last parameter—the hydraulic conductivity of Quaternary deposits—among the other parameters with minor influence. This was not expected; however,

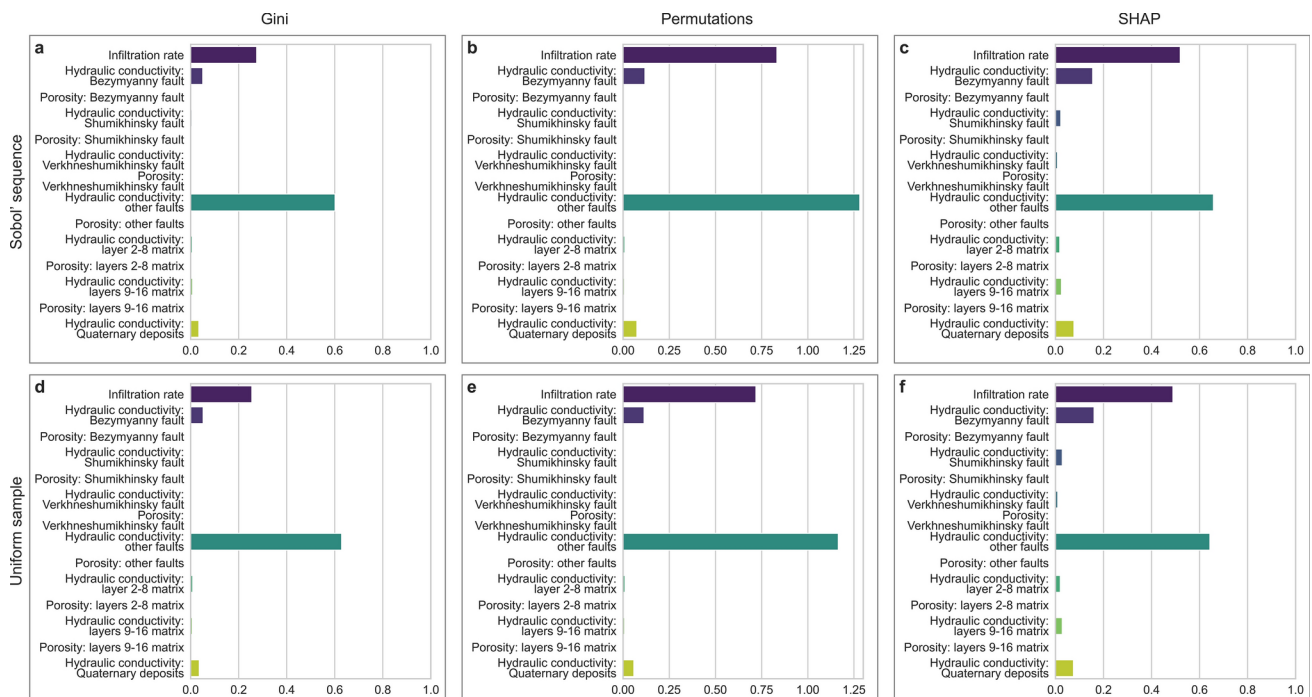


Fig. 7 Different importance measures for two sampling strategies

Table 4 Summary of parameter rankings by different approaches

Parameter	Ranking by Sobol' indices (obtained with 3rd order PCE)		Ranking by importance measures (obtained with XGBoost metamodel)		
	Main (S_1)	Total (S_T)	Gini impurity	Permutations	SHAP
Infiltration rate	3	3	2	2	2
Hydraulic conductivity: Bezmyanny fault	2	2	3	3	3
Porosity: Bezmyanny fault		5*			
Hydraulic conductivity: Shumikhinsky fault		5			5*
Porosity: Shumikhinsky fault					
Hydraulic conductivity: Verkheshumikhinsky fault	5				
Porosity: Verkheshumikhinsky fault					
Hydraulic conductivity: other faults	1	1	1	1	1
Porosity: other faults					
Hydraulic conductivity: layer 2–8 matrix			5*	5	5*
Porosity: layers 2–8 matrix					
Hydraulic conductivity: layers 9–16 matrix	4		5		5
Porosity: layers 9–16 matrix					
Hydraulic conductivity: Quaternary deposits		4	4	4	4

*Indicates that the sensitivity index value for this parameter is slightly lower than that of the previous one, and their rankings can be considered the same

it can be explained from a hydrogeological perspective: the hydraulic conductivity of Quaternary deposits and infiltration jointly affect the volume of water reaching the far-field model domain, resulting in the dilution of the contaminant over time. The ranking of this parameter is unclear from the Sobol' indices. It slightly stands out with the PCE metamodel of 3rd-order but is indistinguishable from other parameters with minor significance in other variants of Sobol' indices.

In general, differences of this kind are not uncommon when comparing sensitivity analysis approaches based on different conceptualizations of the “influence of a particular parameter on the QoI”, and they could be observed even for simple models with known parameter ranking (Cappelli and Grimaldi 2023). The visualization of the regression tree for the investigated datasets (Fig. 8) sheds some light on the reasons for the differences in findings by alternative approaches. It can be seen that the infiltration rate, while having a smaller share in the variance of the output (as established via Sobol' indices), conditions more partitions of the input–output dataset, and therefore receives a larger importance measure value. This means that the range of this parameter has more subranges (in Fig. 8, we can see at least four) with dissimilar output–input dependencies. Similarly, the hydraulic conductivity of the Quaternary deposits, a parameter ranked 4th but almost indistinguishable from other less significant parameters via variance-based analysis, causes multiple partitions of the dataset and unequivocally obtains the 4th rank according to importance measures.

Another disagreement between variance-based and importance-based approaches concerns the least influential parameters, such as the porosity of the Bezmyanny fault or the hydraulic conductivities of the Shumikhinsky fault, the layers 2–8 matrix, and the layers 9–16 matrix. The Gini and permutation importance measures practically do not observe any influence from these parameters, and SHAP ranks them very low, while they are clearly recognizable by the Sobol' method. This discrepancy can probably be explained in different ways. On one hand, we are not dealing with the “true” variance of the nonlinear output but rather with its approximation based on a limited number of realizations, and the corresponding numerical noise could be responsible for the appearance of falsely influential parameters. Due to these considerations, practitioners often treat sensitivity indices around 0.1 as a threshold for deeming a parameter insignificant. Conversely, the importance measure methods might not detect these parameters because the metamodel has smoothed higher-order effects. A definitive answer about the parameters of low significance could be obtained only with the help of additional data.

We also would like to outline some considerations for selecting a particular importance measure method from the discussed variety. In our case, the results in Table 4

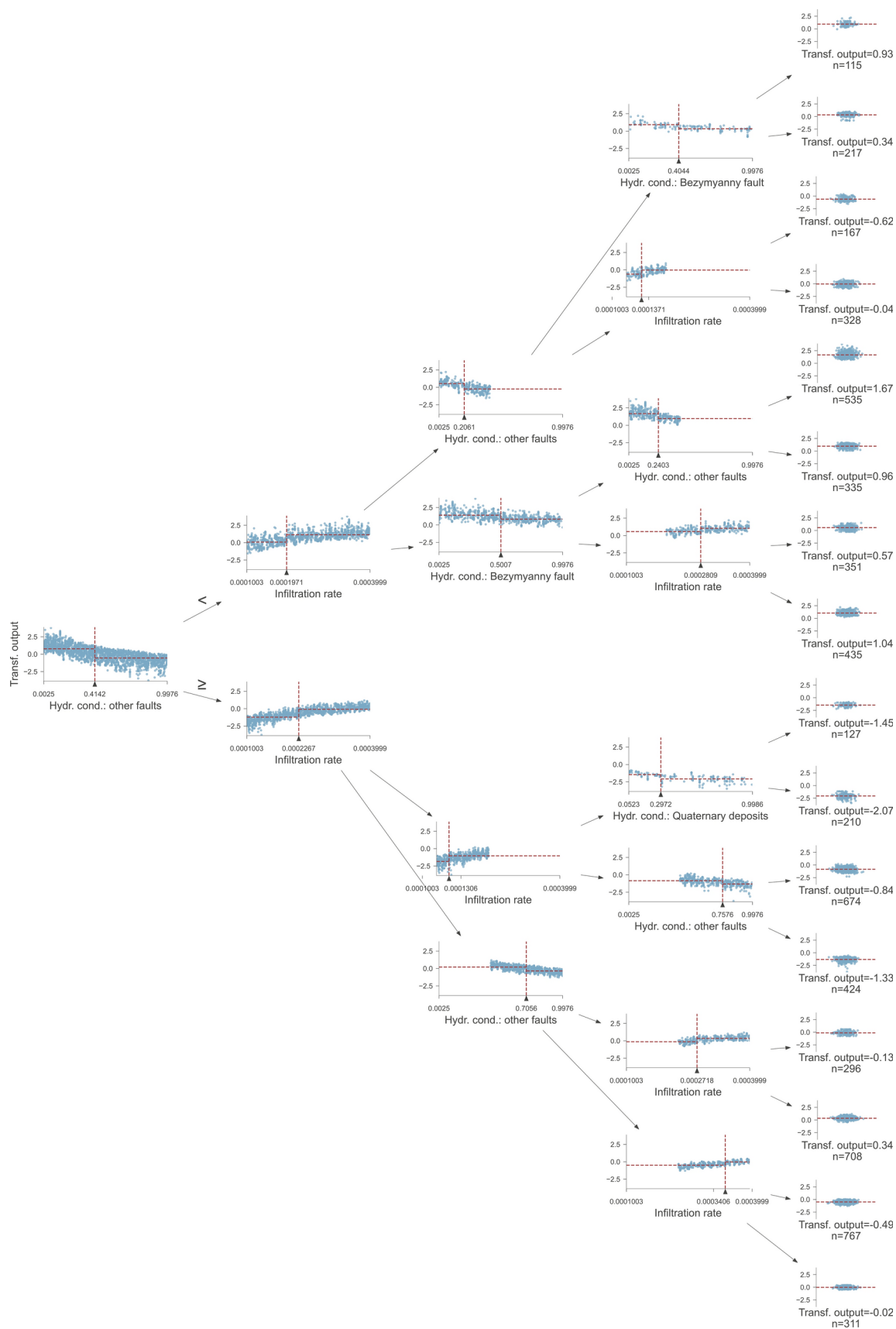


Fig. 8 Regression tree visualization

were almost identical for the Gini, permutation, and SHAP importance measures. From an explainability perspective, using Gini impurity seems beneficial if the corresponding tree-based metamodel provides adequate approximation quality, as this combination of methods reveals the reasoning behind the obtained parameter ranking. On the other hand, the permutation and SHAP measures are independent of the metamodeling technique choice and thus could represent a viable alternative in cases where tree-based metamodels seem ill-suited for the particular dataset or the overall structure of the numerical model under consideration.

An additional point that we wanted to explore was the performance of the importance measures on subsamples of decreasing sizes. The values of R^2 (Eq. 7), which assesses the quality of the XGBoost metamodel, for the subsets consisting of 1200, 600, and 200 realizations were 0.978, 0.956, and 0.757, respectively. Table 5 shows corresponding parameter rankings and Fig. 9 displays the plots of the Sobol' indices and Gini importance measures for three subsamples of one of the initial sets of realizations (the one generated via the Sobol' sequence). For this part of analysis, we used Sobol' indices obtained using 2nd-order PCE surrogate model, since the number of stochastic realizations in the subsets was evidently insufficient for capturing the higher-order effects (comparison of 2nd-order and 3rd-order PCE is given in the Supplementary materials).

Both Fig. 9 and Table 5 indicate agreement of Gini importance values obtained from the subsamples with a decreasing number of realizations. The Sobol' total indices, on the other hand, show slight discordance. For example, according to the ranking obtained from smaller-sized samples (Table 5), parameters unexpected from the largest sample and the expert's point of view, such as the porosities of the

Shumikhinsky and Verkhneshumikhinsky faults, appear in the top five significant parameters. Also, the most significant parameter according to the largest sample (hydraulic conductivity of other faults) goes unnoticed in the results obtained for the smallest subsample (200 realizations).

The absolute values of Sobol' total indices obtained from smaller subsets of realizations (Fig. 9) are lower than those obtained from the full sample (Fig. 5). There is also a visible tendency to "equalize" the values of indices for all parameters as sample size decreases.

The values of Gini importance measures for the smaller subsamples (Fig. 9) are also lower than the corresponding values obtained from the full sample (Fig. 5). However, there are no significant changes within both the most significant and insignificant parameters.

To wrap up the discussion, we would like to reiterate that the challenge with sensitivity analysis very often lies in the computational burden associated with the high number of required numerical model simulations. The comparison of the results obtained on samples of decreasing sizes demonstrated the superiority of the importance measure approach in this respect. For the case study under investigation, the ranking of the most significant parameters by the importance measure approach remained the same even for 200 realizations, which is a very small sample size in terms of sensitivity analysis. Thus, the application of metamodeling and feature importance measures allows us to tackle the main challenge of sensitivity analysis—reducing the number of numerical simulation runs, at least for the purposes of preliminary analysis.

Table 5 Summary of parameter rankings for decreasing subsamples

Sensitivity analysis approach		Ranking by Sobol' total indices				Ranking by Gini importance measure			
Sample size, realizations	6000	1200	600	200	6000	1200	600	200	
Parameter	Infiltration rate	3	2	2	1	2	2	2	2
	Hydraulic conductivity: Bezymyanny fault	2			2	3	3	3	3
	Porosity: Bezymyanny fault	5*	3	3	4*				
	Hydraulic conductivity: Shumikhinsky fault	5			5				
	Porosity: Shumikhinsky fault		4*	4*					
	Hydraulic conductivity: Verkhneshumikhinsky fault								
	Porosity: Verkhneshumikhinsky fault				4				
	Hydraulic conductivity: other faults	1	1	1		1	1	1	1
	Porosity: other faults								
	Hydraulic conductivity: layer 2–8 matrix		5	5	3	5*	4	5	5
	Porosity: layers 2–8 matrix								
	Hydraulic conductivity: layers 9–16 matrix		4	4		5			
	Porosity: layers 9–16 matrix						5		
	Hydraulic conductivity: Quaternary deposits	4				4	5*	4	4

*Indicates that the sensitivity index value for this parameter is slightly lower than that of the previous one, and their rankings can be considered the same

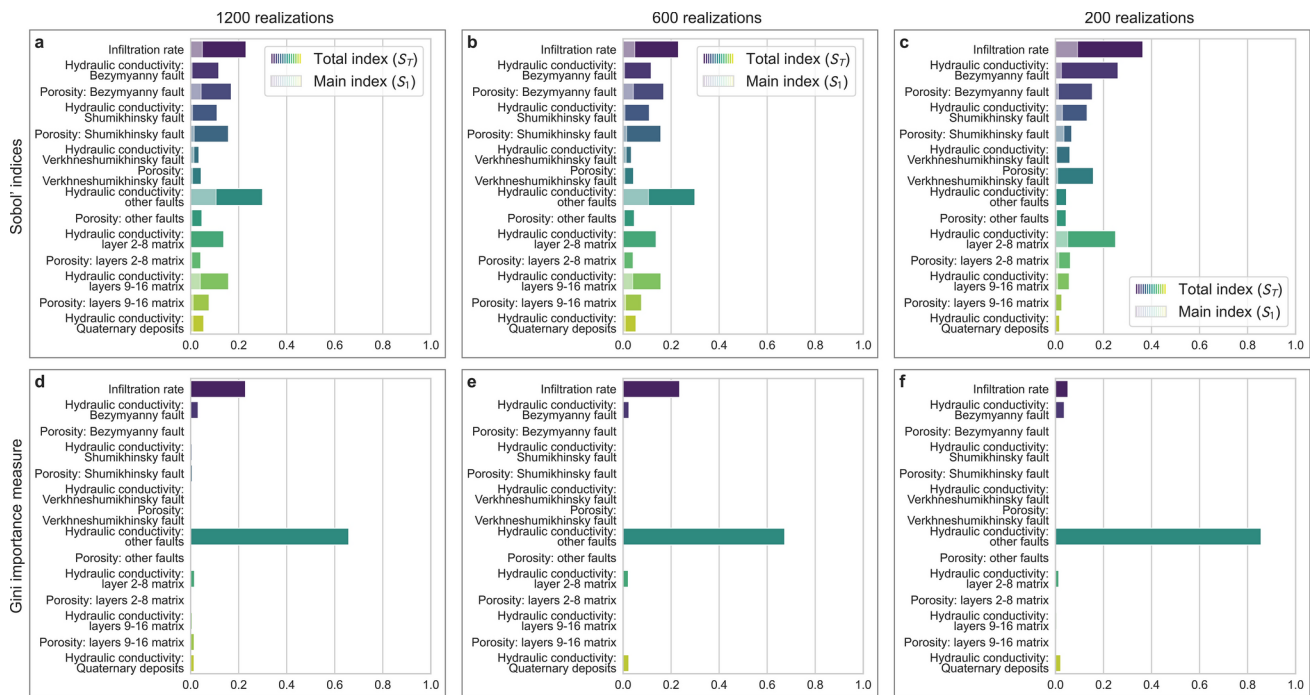


Fig. 9 Sobol' indices with PCE metamodel versus Gini importance for the subsamples of decreasing sizes

5 Conclusion

Feature importance measures, initially developed as a tool for investigating and explaining machine-learning models—which are data-driven by their origin and essence—could also be applied as a sensitivity analysis approach for process-driven numerical models.

The identification of significant parameters based on importance measures is mostly consistent with standard variance-based methods, and dissimilarities could be explained by differences in the metrics used in corresponding approaches for parameter ranking. Importance measures reveal the parameters of secondary importance, the influence of which may not be obvious against the background of the most significant parameters for the analyzed model, but is still non-negligible and can be adequately interpreted.

The importance measures can't be estimated from the dataset of inputs and outputs as is; building a metamodel is an inherent step of the approach. In this paper, we applied XGBoost, a tree-based approach, as a metamodeling technique. It was selected for its higher approximation quality for the analyzed numerical model compared to several other methods. Generally, other ML methods, such as Gaussian processes, support vector regression, and many others, could also be employed effectively with two of the three investigated importance measures: permutations and SHAP. However, the benefit of tree-based approaches and Gini importance measure is that they allow visualization of

the explicit “rules” that the metamodel uses to divide the parametric space and predict the model outputs. This option enhances the explainability of the results and provides some insights into model behavior.

The overall conclusion of the presented exercise is that importance measures could at least accompany the standard variance-based methods because the different approaches provide an understanding of different aspects of model behavior and underlying processes. At the same time, for cases where the application of variance-based indices is not recommended, data-driven importance measures are still viable and could help to carry out preliminary analysis even if only a small sample is currently available. It should be noted, though, that this conclusion about the sufficiency of a small sample should not instill a false sense of confidence. When investigating the behavior of a complex natural system with the help of a numerical model—which is already a simplification—and then attempting to understand this model using information about only values within a small portion of the parametric space and an approximation based on them, we, first, undoubtedly miss some potentially significant information, and second, mask this lack of information with some noise while building the metamodel. Thus, one should bear in mind that sufficient sample sizes are still desirable for more thorough sensitivity investigations.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00477-024-02869-y>.

Author contributions ES and VS conceived the original idea of the paper, carried out sensitivity analyses, and wrote the original draft of the manuscript. GN developed the groundwater flow and transport model. VS and GN prepared figures and tables. All authors read and approved the final manuscript.

Funding Numerical studies in the scope of “Comprehensive research program of radioactive waste disposal long-term safety” are supported by the State Atomic Energy Corporation “Rosatom” (contract N.4dsh.244.20.22.1040).

Data availability Datasets used for sensitivity analysis and the corresponding scripts are available on GitHub: https://github.com/lab13-ibr/ae/importances_vs_sobol_paper.

Conflict of interest The authors declare no competing interests.

References

- Abraham S, Raisee M, Ghorbaniasl G et al (2017) A robust and efficient stepwise regression method for building sparse polynomial chaos expansions. *J Comput Phys* 332:461–474. <https://doi.org/10.1016/j.jcp.2016.12.015>
- Alizadeh R, Allen JK, Mistree F (2020) Managing computational complexity using surrogate models: a critical review. *Res Eng Design* 31(3):275–298. <https://doi.org/10.1007/s00163-020-00336-7>
- Bagaev D, Grigoriev F, Kapyrin I, et al (2019) Improving parallel efficiency of a complex hydrogeological problem simulation in GeRa. In: Voevodin V, Sobolev S (eds) *Supercomputing: 5th Russian Supercomputing Days, RuSCDays 2019, Moscow, Russia, Sep 23–24, 2019, revised selected papers* (Communications in computer and information science). Springer, Cham, Switzerland, pp 265–277. https://doi.org/10.1007/978-3-030-36592-9_22
- Baudin M, Dutfoy A, Iooss B, Popelin A-L (2015) OpenTURNS: an industrial software for uncertainty quantification in simulation. In: Ghanem R, Higdon D, Owhadi H (eds) *Handbook of uncertainty quantification*. Springer, Cham, pp 1–38. https://doi.org/10.1007/978-3-319-11259-6_64-1
- Björck Å (1996) *Numerical methods for least squares problems*. SIAM Press, Philadelphia. <https://doi.org/10.1137/1.9781611971484>
- Blatman G, Sudret B (2010) Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. *Reliab Eng Syst Saf* 95(11):1216–1229. <https://doi.org/10.1016/j.ress.2010.06.015>
- Blatman G, Sudret B (2011) Adaptive sparse polynomial chaos expansion based on least angle regression. *J Comput Phys* 230(6):2345–2367. <https://doi.org/10.1016/j.jcp.2010.12.021>
- Borgonovo E, Plischke E (2016) Sensitivity analysis: a review of recent advances. *Eur J Oper Res* 248(3):869–887. <https://doi.org/10.1016/j.ejor.2015.06.032>
- Box GEP, Draper NR (2007) *Response surfaces, mixtures, and ridge analyses*, 2nd edn. Wiley, Hoboken
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
- Cappelli F, Grimaldi S (2023) Feature importance measures for hydrological applications: insights from a virtual experiment. *Stoch Env Res Risk Assess* 37(12):4921–4939. <https://doi.org/10.1007/s00477-023-02545-7>
- Cappelli F, Tauro F, Apollonio C et al (2023) Feature importance measures to dissect the role of sub-basins in shaping the catchment hydrological response: a proof of concept. *Stoch Env Res Risk Assess* 37(4):1247–1264. <https://doi.org/10.1007/s00477-022-02332-w>
- Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, California, USA, pp 785–794. <https://doi.org/10.1145/2939672.2939785>
- Crestaux T, Le Maître O, Martinez J-M (2009) Polynomial chaos expansion for sensitivity analysis. *Reliab Eng Syst Saf* 94(7):1161–1172. <https://doi.org/10.1016/j.ress.2008.10.008>
- Devore JL, Carlton M (2016) *Probability and statistics for engineering and the sciences*, 9th edn. Cengage Learning, Boston
- Dorofeev AN, Bolshov LA, Linge II et al (2017) Strategic master plan for R&D demonstrating the safety of construction, operation and closure of a deep geological disposal facility for radioactive waste. *Radioact Waste* 1:33–42
- El Moçayd N, Seaid M (2021) Data-driven polynomial chaos expansions for characterization of complex fluid rheology: case study of phosphate slurry. *Reliab Eng Syst Saf* 216:107923. <https://doi.org/10.1016/j.ress.2021.107923>
- Hooker G, Mentch L, Zhou S (2021) Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Stat Comput* 31(6):82. <https://doi.org/10.1007/s11222-021-10057-z>
- Iooss B, Lemaître P (2015) A review on global sensitivity analysis methods. In: Dellino G, Meloni C (eds) *Uncertainty management in simulation-optimization of complex systems: algorithms and applications*. Springer, US, Boston, pp 101–122. https://doi.org/10.1007/978-1-4899-7547-8_5
- Iooss B, Kenett R, Secchi P (2022) Different views of interpretability. In: Lepore Antonio, Poggi Jean-Michel, Palumbo Biagio (eds) *Interpretability for Industry 4.0: statistical and machine learning approaches*. Springer, Switzerland, pp 1–20. <https://doi.org/10.1007/978-3-031-12402-0>
- Ishigami T, Homma T (1990) An importance quantification technique in uncertainty analysis for computer models. In: *Proceedings. First International Symposium on Uncertainty Modeling and Analysis*. pp 398–403. <https://doi.org/10.1109/ISUMA.1990.151285>
- Iwanaga T, Usher W, Herman J (2022) Toward SALib 2.0: advancing the accessibility and interpretability of global sensitivity analyses. *Soc Environ Syst Model* 4:18155. <https://doi.org/10.18174/sesmo.18155>
- Jaxa-Rozen M, Kwakkel J (2018) Tree-based ensemble methods for sensitivity analysis of environmental models: a performance comparison with Sobol and Morris techniques. *Environ Model Softw* 107:245–266. <https://doi.org/10.1016/j.envsoft.2018.06.011>
- Kapyrin I, Konshin I, Kramarenko V, Grigoriev F (2018) Modeling groundwater flow in unconfined conditions of variable density solutions in dual-porosity media using the GeRa code. In: *Supercomputing: 4th Russian Supercomputing Days, RuSCDays 2018, Moscow, Russia, September 24–25, 2018, Revised Selected Papers* (Communications in computer and information science). Springer, Cham, Switzerland, pp 266–278. https://doi.org/10.1007/978-3-030-05807-4_23
- Konshin I, Kapyrin I (2017) Scalable computations of GeRa code on the base of software platform INMOST. In: Malyshkin V (ed) *Parallel Computing Technologies: 14th International Conference, PaCT 2017, Nizhny Novgorod, Russia, September 4–8, 2017* (Lecture notes in computer science). Springer, Cham, Switzerland, pp 433–445. https://doi.org/10.1007/978-3-319-62932-2_42
- Kucherenko S, Albrecht D, Saltelli A (2015) Exploring multi-dimensional spaces: a comparison of Latin hypercube and quasi Monte Carlo sampling techniques. *arXiv preprint arXiv:1505.02350*. <https://doi.org/10.48550/arXiv.1505.02350>
- Linge II, Utkin SS, Svitelman VS, Deryabin SA (2020) Numerical safety assessment and optimization of decisions on radioactive waste disposal and nuclear decommissioning: trends, needs,

- opportunities. *Radioact Waste* 2(11):85–98. <https://doi.org/10.25283/2587-9707-2020-2-85-98>
- Lo Piano S, Ferretti F, Puy A et al (2021) Variance-based sensitivity analysis: the quest for better estimators and designs between explorativity and economy. *Reliab Eng Syst Saf* 206:107300. <https://doi.org/10.1016/j.res.2020.107300>
- Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S et al (eds) *Advances in neural information processing systems* 30 (NIPS 2017). Curran Associates, Red Hook, pp 4765–4774
- Lüthen N, Marelli S, Sudret B (2021) Sparse polynomial chaos expansions: literature survey and benchmark. *SIAM/ASA J Uncertain Quantif* 9(2):593–649. <https://doi.org/10.1137/20M1315774>
- Mason L, Baxter J, Bartlett P, Fream M (1999) Boosting algorithms as gradient descent. In: Solla S, Leen T, Müller K (eds) *Advances in neural information processing systems* 12 (NIPS 1999). MIT Press, pp 512–518
- Molnar C (2022) *Interpretable machine learning: a guide for making black box models explainable*, 2nd edn. Mucbook Clubhouse, Munich
- Morozov OA, Rastorguev AV, Neuvazhaev GD (2019) Assessing the state of the geological environment at the Yeniseyskiy site (Krasnoyarsk region). *Radioact Waste* 4(9):46–62. <https://doi.org/10.25283/2587-9707-2019-4-46-62>
- NEA (2013) *The nature and purpose of the post-closure safety cases for geological repositories* (Radioactive Waste Management, NEA/RWM/R(2013)1). OECD Nuclear Energy Agency, Paris, France
- Neuvazhaev G, Rastorguev A, Morozov O, et al (2020) 3D hydrogeological modeling of deep geological disposal in the Nizhnekansky rock massif. In: EGU General Assembly Conference Abstracts. p 21509. <https://doi.org/10.5194/egusphere-egu2020-21509>
- Owen AB, Prieur C (2017) On Shapley value for measuring importance of dependent inputs. *SIAM/ASA J Uncertain Quantif* 5(1):986–1002. <https://doi.org/10.1137/16M1097717>
- Parr T, Lapusan T, Grover P (2023) dtreeviz: A python library for decision tree visualization and model interpretation. <https://github.com/parrt/dtreeviz>. Accessed 5 Apr 2024
- Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
- Pianosi F, Beven K, Freer J et al (2016) Sensitivity analysis of environmental models: a systematic review with practical workflow. *Environ Model Softw* 79:214–232. <https://doi.org/10.1016/j.envsoft.2016.02.008>
- Plischke E, Rabitti G, Borgonovo E (2021) Computing Shapley effects for sensitivity analysis. *SIAM/ASA J Uncertain Quantif* 9(4):1411–1437. <https://doi.org/10.1137/19M1304738>
- Puy A, Becker W, Piano SL, Saltelli A (2022) A comprehensive comparison of total-order estimators for global sensitivity analysis. *Int J Uncertain Quantif* 12(2):1–18. <https://doi.org/10.1615/Int.J.UncertaintyQuantification.2021038133>
- Rasmussen CE, Williams CKI (2006) *Gaussian processes for machine learning* (Adaptive computation and machine learning). MIT Press, Cambridge
- Rastorguev AV, Neuvazhaev GD, Smirnov KD (2021) Interpretation of groundwater inflow testing from imperfect wells in low-permeable rock mass considering a flow in the wellbore of a perfect observation well based on the Yeniseyskiy site case study. *Radioact Waste* 3(16):61–71. <https://doi.org/10.25283/2587-9707-2021-3-61-71>
- Razavi S, Jakeman A, Saltelli A et al (2021) The future of sensitivity analysis: an essential discipline for systems modeling and policy support. *Environ Model Softw* 137:104954. <https://doi.org/10.1016/j.envsoft.2020.104954>
- Rosatom (2018) *Strategy for the development of RW deep disposal facility*. *Radioact Waste* 2(3):114–120
- Roth AE (ed) (1988) *The Shapley value: essays in Honor of Lloyd S.* Cambridge University Press, Cambridge, UK, Shapley. <https://doi.org/10.1017/CBO9780511528446>
- Saltelli A (2002) Making best use of model evaluations to compute sensitivity indices. *Comput Phys Commun* 145(2):280–297. [https://doi.org/10.1016/S0010-4655\(02\)00280-1](https://doi.org/10.1016/S0010-4655(02)00280-1)
- Saltelli A, Tarantola S (2002) On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. *J Am Stat Assoc* 97(459):702–709. <https://doi.org/10.1198/016214502388618447>
- Saltelli A, Ratto M, Andres T et al (2008) *Global sensitivity analysis: the primer*. Wiley, Chichester. <https://doi.org/10.1002/9780470725184>
- Saltelli A, Annoni P, Azzini I et al (2010) Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comput Phys Commun* 181(2):259–270. <https://doi.org/10.1016/j.cpc.2009.09.018>
- Saltelli A, Aleksankina K, Becker W et al (2019) Why so many published sensitivity analyses are false: a systematic review of sensitivity analysis practices. *Environ Model Softw* 114:29–39. <https://doi.org/10.1016/j.envsoft.2019.01.012>
- Samper J, Naves A, Montenegro L, Mon A (2016) Reactive transport modelling of the long-term interactions of corrosion products and compacted bentonite in a HLW repository in granite: uncertainties and relevance for performance assessment. *Appl Geochem* 67:42–51. <https://doi.org/10.1016/j.apgeochem.2016.02.001>
- Saveleva E, Svitelman V, Blinov P, Valetov D (2021) Sensitivity analysis and model calibration as a part of the model development process in radioactive waste disposal safety assessment. *Reliab Eng Syst Saf* 210:107521. <https://doi.org/10.1016/j.res.2021.107521>
- Sobol' IM (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math Comput Simul* 55(1):271–280. [https://doi.org/10.1016/S0378-4754\(00\)00270-6](https://doi.org/10.1016/S0378-4754(00)00270-6)
- Song S, He R, Shi Z, Zhang W (2021) Variable importance measure system based on advanced random forest. *Comput Model Eng Sci* 128(1):65–85. <https://doi.org/10.32604/cmesci.2021.015378>
- Sudret B (2008) Global sensitivity analysis using polynomial chaos expansions. *Reliab Eng Syst Saf* 93(7):964–979. <https://doi.org/10.1016/j.res.2007.04.002>
- Swiler LP, Becker D-A, Brooks D, et al (2021) Sensitivity analysis comparisons on geologic case studies: an international collaboration (Technical Report, SAND2021-11053). Sandia National Laboratories, Albuquerque, New Mexico, USA. <https://doi.org/10.2172/1822591>
- Torres M (2021) A machine learning method for parameter estimation and sensitivity analysis. In: Paszynski M, Kranzlmüller D, Krzhizhanovskaya VV et al (eds) *Computational science—ICCS 2021* (lecture notes in computer science). Springer, Cham, pp 330–343. https://doi.org/10.1007/978-3-030-77977-1_26
- Valetov D, Neuvazhaev G, Svitelman V, Saveleva E (2019) Hybrid cuckoo search and harmony search algorithm and its modifications for the validation of groundwater flow models. In: *Proceedings of the 11th International Joint Conference on Computational Intelligence*. SCITEPRESS - Science and Technology Publications, Vienna, Austria, pp 221–228. <https://doi.org/10.5220/0008345502210228>
- Vapnik VN (1995) *The nature of statistical learning theory*, 1st edn. Springer, New York. <https://doi.org/10.1007/978-1-4757-2440-0>
- Vigfusson J, Maudoux J, Raimbault P, et al (2007) European pilot study on the regulatory review of the safety case for geological disposal of radioactive waste. Case study: uncertainties and their management. <https://fanc.fgov.be/nl/system/files/case-study-european-pilot-group.pdf>. Accessed 5 Apr 2024
- Wei P, Lu Z, Song J (2015) A comprehensive comparison of two variable importance analysis techniques in high dimensions:

Application to an environmental multi-indicators system. *Environ Model Softw* 70:178–190. <https://doi.org/10.1016/j.envsoft.2015.04.015>

- Xiu D, Karniadakis GE (2002) The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM J Sci Comput* 24(2):619–644. <https://doi.org/10.1137/S1064827501387826>

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.