BAREKENG: Journal of Mathematics and Its Applications
March 2026 Volume 20 Issue 1 Page 0637–0656
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20no1pp0637-0656

INTEGRATED STATISTICAL MODELLING OF IRON EXCEEDANCE RISK: A MONTE CARLO, LOGISTIC REGRESSION, RANDOM FOREST, AND SOBOL ANALYSIS APPROACH

Rachid El Chaal[1*], **Hamid Dalhi**[2], **Otmane Darbal**[3], **Moulay Othman Aboutafail**[4]

[1,2,3,4]*ENSA of Kenitra, Engineering Sciences Laboratory, Data Analysis, Mathematical Modeling and Optimization Team, Ibn Tofail University Kenitra, 14000, Morocco*

*Corresponding author's e-mail: * **rachid.elchaal@uit.ac.ma***

Article Info

Article History:

Received: 30th April 2025
Revised: 22nd May 2025
Accepted: 5th August 2025
Available online: 24th November 2025

Keywords:

Kolmogorov-Smirnov test;
Log-Normal;
Monte Carlo simulation;
Sobol sensitivity analysis;
Statistical modelling;
Water quality.

ABSTRACT

The quality of water resources in the Inaouen watershed, northern Morocco, is increasingly threatened by metal contamination, particularly iron (Fe). This study implements an integrated statistical framework to assess the risk of exceeding regulatory iron concentration thresholds. After preprocessing local physico-chemical data, a binary indicator variable was constructed to flag exceedances of the critical 30 µg/L threshold. Iron concentrations were modeled using log-normal and Weibull distributions, with a Monte Carlo simulation (n = 10,000) based on the log-normal law estimating exceedance probabilities across multiple thresholds (30, 50, 100 µg/L), revealing an 18% risk at 30 µg/L. Predictive modeling via logistic regression and random forest analysis identified calcium (Ca) as the dominant driver of iron exceedances, a finding corroborated by Sobol sensitivity analysis (S1 index = 0.74), with bicarbonate (HCO_3^-) emerging as a secondary factor (S1 = 0.10). These results demonstrate the power of combining distribution fitting, machine learning, and global sensitivity analysis to effectively quantify and interpret iron contamination risks in vulnerable watersheds such as Inaouen. The proposed methodology offers a robust decision-support tool for sustainable water resource management and public health protection.

How to cite this article:

R. E. Chaal, H. Dalhi, O. Darbal, and M. O Aboutafail, "INTEGRATED STATISTICAL MODELLING OF IRON EXCEEDANCE RISK: A MONTE CARLO, LOGISTIC REGRESSION, RANDOM FOREST, AND SOBOL ANALYSIS APPROACH", *BAREKENG: J. Math. & App.*, vol. 20, iss. 1, pp. 0637-0656, Mar, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · **Open Access**

1. INTRODUCTION

Contamination of freshwater resources by trace metals is a major environmental and health concern, especially in areas with high dependence on surface water for human food and agriculture. Among these contaminants, iron (Fe) is an element naturally present in the Earth's crust, but whose high concentrations in groundwater can indicate geochemical imbalances, anthropogenic infiltration, or conditions of enhanced reduction. When iron exceeds certain thresholds (often 30 µg/L according to WHO or national standards) [1], it can alter the organoleptic quality of the water, promote the growth of biofilms, and corrode hydraulic infrastructures, while signaling the possible presence of other undesirable metals [2].

In this context, the study is applied to the Inaouen watershed, located in the Middle Atlas region of Morocco. This basin constitutes a major sub-tributary of the Sebou wadi and plays a central hydrological role in the country's northern region. It is characterized by a mountainous topography, a semi-arid to humid climate depending on altitude, and varied geological formations influencing surface water quality [3]. While traditional statistical or neural network methods have been used to address metal contamination in earlier studies on the Inaouen watershed [3], [4], this work presents three significant innovations: (1) The first use of Sobol indices in this area to separate factor-specific contributions from intricate geochemical interactions; (2) An integrated framework that combines probabilistic (Monte Carlo), machine learning (Random Forest), and global sensitivity (Sobol) methods to quantify iron exceedance risks holistically; (3) Actionable thresholds (e.g., Ca > 75th percentile) derived from multi-method consensus, moving beyond descriptive analyses to targeted water management. For semi-arid watersheds under human pressure, this approach fills in the gaps between risk assessment, predictive modeling, and mechanistic interpretation.

The water resources of the Inaouen basin are under increasing pressure, particularly in connection with agriculture, urbanization, and domestic and industrial discharges. These pressures make it essential to rigorously assess surface water quality, particularly iron concentrations, which can act both as an indicator and a factor in degradation [4].

The purpose of this study is to : (1) Statistically characterize iron concentrations in the waters of the Inaouen basin; (2) Estimate the probability of exceeding critical thresholds via a Monte Carlo simulation based on adjusted laws (log-normal and Weibull); (3) Identify explanatory physico-chemical factors using predictive models such as logistic regression and random forests; (4) Quantify the impact of each variable on the probability of exceedance using a global sensitivity analysis of the Sobol type.

Despite several previous studies on water quality in the Inaouen watershed, none have integrated a joint probabilistic, mechanistic, and predictive approach with cross-validation of models. The originality of this approach lies in the joint use of classical and advanced statistical methods of simulation and machine learning, allowing a detailed understanding of the mechanisms of contamination and a rigorous prioritization of risk factors in a specific geographical context.

2. RESEARCH METHODS

All statistical analyses were implemented using the Python language, relying on robust scientific libraries such as NumPy, Pandas, Scikit-learn, Statsmodels, SALib, and Matplotlib**.** The code made it possible to automate data cleaning, distribution adjustment, Monte Carlo simulation, predictive modeling (logistic regression and random forest), and Sobol sensitivity analysis.

2.1 Study Area and Data Collection

The study was conducted in the Inaouen watershed (Fig. 1), an important sub-basin of the Oued Sebou basin, located in northern Morocco. This basin has a marked geological diversity, with carbonate, clay, and siliceous formations influencing the physico-chemical composition of surface waters [5][6].

The data analyzed comes from 100 surface water samples collected at different points of the basin. Each sample was analyzed for iron (Fe) concentration as well as eight other physicochemical parameters: calcium (Ca), bicarbonates (HCO_3^-), sulfates (SO_4^{2-}), sodium (Na), chlorides (Cl), magnesium (Mg), calcium carbonate ($CaCO_3$), and potassium (K) [5][7].

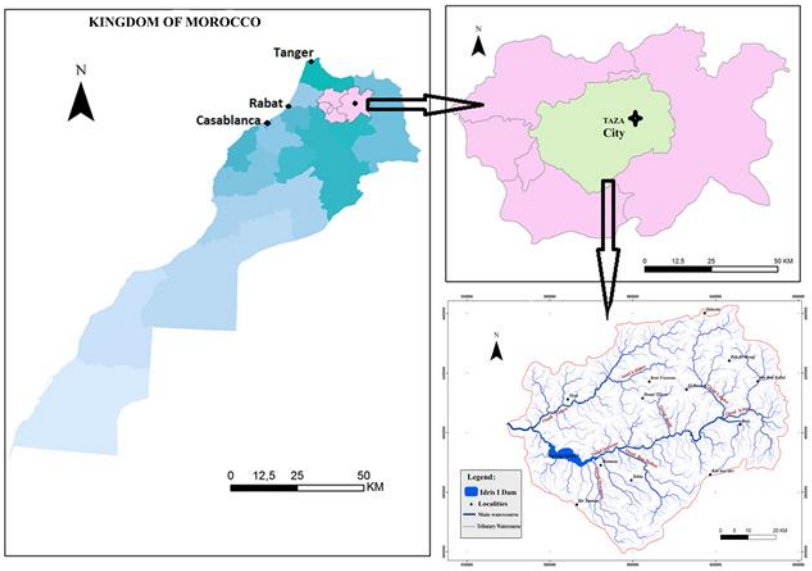

Figure 1. Geographical Location of The Study Area

2.2 Data Pre-Processing

The data preprocessing protocol was rigorously applied to ensure the quality and homogeneity of the dataset prior to analysis. Four main operations were carried out:

1. Systematic elimination of all samples with missing iron (Fe) concentration values, thus ensuring the integrity of the analyzed data;
2. The filtering of extreme values (Fe > 500 µg/L), a threshold determined to limit the disproportionate influence of outliers on the statistical results;
3. The creation of a binary response variable "Exceedance" encoding the exceedance 1 if Fe > 30 µg/L, and 0 otherwise, thus allowing a clear modeling of the risk;
4. The standardization of variable names according to a standardized nomenclature facilitates the reproducibility and readability of subsequent analyses.

These preliminary steps made it possible to obtain a clean and structured set of data, optimal for the different statistical and machine learning approaches deployed in the study.

2.3 Theoretical Concepts and Applied Statistical Models

2.3.1 Adjustment of Distributions and Monte Carlo Simulation

To model iron concentrations, two statistical distributions were adjusted: The Log-normal law, commonly used to model asymmetric positive concentrations [8]-[10], and Weibull's law.

1. Log-normal law:
   A random variable $\boldsymbol{X}$ follows a log-normal law if:
   $$\ln(X) \sim \mathcal{N}(\mu, \sigma^2) \quad (1)$$
   The density is [11]

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right), x > 0. \quad (2)$$

Where:

μ: Mean logarithm.

σ: Standard deviation of logarithm.

The estimation of the parameters μ (logarithmic mean) and σ (logarithmic standard deviation) of the log-normal distribution was carried out by the maximum likelihood method from the observed iron (Fe) concentration data. In this study, this model was specifically chosen for its ability to accurately represent the characteristics of environmental concentrations: strictly positive values, asymmetric distribution, and variability covering several orders of magnitude, typical of metal contamination profiles in surface waters. The adjusted log-normal model was then used as the basis for the Monte Carlo simulation [12], making it possible to estimate the probabilities of exceeding regulatory thresholds.

2. Weibull's law:

Weibull's law is also adapted to this type of environmental data [13]-[16]: Weibull's law is a continuous probability distribution used to model failure time, positive natural phenomena, or environmental concentrations. The probability density function (PDF) of a random variable $\boldsymbol{X} \sim \mathbf{Weibull}(\boldsymbol{k},\boldsymbol{\lambda})$ is given by:

$$f(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, x \geq 0. \quad (3)$$

Where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter.

The estimation of the k (shape parameter) and λ (scale parameter) parameters of the Weibull distribution was carried out by the maximum likelihood method (MLE), offering a robust approach to characterize the distribution of iron concentrations. This law has remarkable statistical properties: it reduces to an exponential distribution when k=1, shows a decreasing density for k<1, and adopts an asymmetric bell shape with a well-defined mode for k>1. In this study, Weibull's law was systematically fitted to the iron concentration data to provide an objective comparison with the log-normal model, both distributions commonly used to model positive and asymmetric environmental data. The quality of the fit was rigorously evaluated via the Kolmogorov-Smirnov (KS) test [17], allowing the estimated parameters to be statistically validated and the relative performance of the two distributions to be compared to represent the specific characteristics of the metal concentrations observed in the Inaouen basin.

3. The Kolmogorov-Smirnov test (KS)

The Kolmogorov-Smirnov test (KS) is a non-parametric test to compare [18]: a robust non-parametric method, which was used in this study to quantitatively assess the quality of fit between the theoretical distributions (log-normal and Weibull) and the empirical distribution of the observed iron concentrations. This test, which compares cumulative distribution functions, has the advantage of making no assumptions about the shape of the underlying distribution, making it particularly suitable for the analysis of often complex environmental data. Applied here as a sample version, it made it possible to objectively measure the maximum deviation between the adjusted theoretical distributions and the actual data, thus providing a solid statistical basis for choosing between log-normal and Weibull models. Its mathematical formula (1-sample)[19][20]:

Either:

$\boldsymbol{F_n(x)}$ the empirical distribution function based on a sample size n,

$\boldsymbol{F(x)}$ the theoretical distribution function of a continuous distribution (e.g., log-normal, Weibull).

The KS test statistic is [21]:

$$D_n = \sup_x |F_n(x) - F(x)|, \quad (4)$$

where sup represents the supremum (the largest absolute value of the deviations between the two distribution functions).

The Kolmogorov-Smirnov (KS) test quantifies the fit between distributions using the statistic Dn [20], representing the maximum vertical deviation between the empirical and theoretical distribution functions. A high p-value (> 0.05) indicates a valid model fit, while a low p −value (< 0.05) suggests rejection of the fit. Although this test is ideal for continuous data, it has a significant limitation when applied to distributions whose parameters have been estimated on the same data; a common practice, thus requiring specific corrections (adjusted KS) to avoid systematic bias toward model acceptance. This crucial subtlety is often overlooked in environmental applications. However, it has been rigorously considered in this study to ensure the validity of conclusions regarding the log-normal adjustment of metal concentrations.

2.3.2 Monte Carlo Simulation

The Monte Carlo method, which relies on repeated random simulations, enables the numerical approximation of complex statistical results such as probabilities or integrals that are challenging to calculate analytically. This approach is beneficial for estimating the probability that a random variable X (representing the iron concentration modeled by an adjusted log-normal distribution) exceeds a critical threshold T (30 µg/L in this study). It generates many realizations ($n = 10{,}000$) [22] and calculates the proportion of simulated values that exceed this threshold, thus providing a robust estimate of contamination risks while accounting for the uncertainty inherent in environmental data. The empirical probability is then given by [23][24]:

$$\hat{P}(X > T) = \frac{1}{N}\sum_{i=1}^{N} 1_{\{x_i > T\}} \tag{5}$$

Where:

N is the number of simulations (e.g., 10000),
x_i is the i-th simulated value,
$1_{\{x_i > T\}}$ is an indicator function (equals 1 $x_i > T$, 0 if not).

This study applied the Monte Carlo simulation method by generating 10,000 random draws from a log-normal distribution adjusted to the observed iron concentrations to estimate the probabilities of exceeding three critical regulatory thresholds (30, 50, and 100 µg/L). For each DTC threshold, the corresponding probability was determined as the proportion of simulated values exceeding this threshold, thus providing a probabilistic quantification of the risk of metal contamination in the Inaouen watershed. This approach makes it possible to transform a statistical adjustment into directly interpretable information for water quality management, while integrating the uncertainty related to the natural variability of the data and their statistical distribution [25][26].

2.3.3 Predictive Modeling

Two models were used to identify the factors explaining the probability of exceedance:

1. A logistic regression

Logistic regression is a binary classification model that estimates the probability of an event $Y = 1$ based on a set of explanatory variables$X_1, X_2, \ldots, X_p$. It made it possible to estimate the effect of each variable on the probability of Exceedance, in terms of directional coefficient. In the context of this study, it models the probability that an iron concentration exceeds the threshold of $30\mu g/L$.

The conditional probability of exceedance is modeled by [27]-[30]:

$$\mathbb{P}(Y = 1 \mid X) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)\right)} \tag{6}$$

or equivalently [31][32]:

$$\log\left(\frac{\mathbb{P}(Y = 1)}{1 - \mathbb{P}(Y = 1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{7}$$

The term on the left is called logit.

The coefficients β_j are estimated by maximum likelihood, quantifying the impact of the explanatory variables [33][34].

a. $\beta_j > 0$: An increase in X_j increases the probability of overtaking.
b. $\beta_j < 0$: An increase of X_j decreases this probability.
c. e^{β_1} represents the odds ratio, which measures the multiplicative factor of the probability of occurrence for each unit of increase.

Then the coefficients represent the marginal effect of each variable on the overrun log-odds.

In the study

a. The explanatory variables are: $\mathrm{Ca}, \mathrm{HCO_3^-}, \mathrm{SO_4^{2-}}, \mathrm{Na}, \mathrm{Cl}, \mathrm{Mg}, \mathrm{CaCO_3}, \mathrm{K}$
b. The model makes it possible to identify the most influential variables in exceeding the threshold.
c. The interpretation is based on the value and sign of the coefficients.
d. Estimated coefficients, p-values, 95% CI
e. Determination of the effect of each variable on excess Fe

2. A Random Forest ensemble model

Random Forest is a machine learning method aggregating predictions from multiple decision trees to improve model accuracy and robustness. It is particularly effective for non-linear classification problems and resistant to overfitting. It has made it possible to assess the relative importance of explanatory variables in a non-linear way that is robust to complex interactions [35].

The final prediction (for binary classification) is usually made by majority vote [36]:

$$\hat{f}(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x). \quad (8)$$

Where each $f_m(x)$ is a tree trained on a bootstrap.

The Random Forest model is built using an ensemble approach in which each decision tree is trained on a bootstrap subsample of the original data. It incorporates a random selection of variables at each node to ensure model diversity and prevent overfitting, with trees grown to maximum depth (without pruning) to capture complex relationships in the data finely. The importance of the variables is then determined by measuring the average impurity reduction (Gini index in this study) [37] that each variable (X_j) produces across all trees, thus providing a robust metric for evaluating their relative contribution to predicting the exceedance of iron concentration thresholds. The greater the impurity reduction, the more influential the variable is in the model [38].

$$\mathrm{Imp}(X_j) = \sum_{\text{nocuds contenant } X_j} \Delta I_{\text{noeud}}\,. \quad (9)$$

This importance is often standardized between 0 and 1.

The Random Forest model was used to predict iron concentration threshold exceedance (Fe > 30 µg/L), offering a robust classification method through its ability to handle nonlinear relationships and complex interactions among variables. This involved a thorough evaluation that divided the data into training (70%) and test (30%) sets to validate model performance and provided an objective measure of the relative importance of each predictive variable (Ca, HCO_3^-, SO_4^{2-}, etc.) using impurity reduction analysis. This approach systematically identifies the key factors influencing iron contamination and quantifies their specific contributions to predicting the risk of exceedance. The two models were compared to validate the robustness of the results.

Even though the AUC values were high, we used a number of validation methods to double-check the model's calibration and lower the risk of overfitting. These included carefully checking how well the training and test sets worked (ΔAUC < 0.02 was acceptable), calculating the Brier score (0 means perfect calibration), using Hosmer-Lemeshow goodness-of-fit tests for logistic regression to check probability calibration, and doing 5-fold cross-validation 100 times on bootstrap resamples to make sure the model was strong. We used these extra metrics on both Random Forest and logistic regression models (with L2 regularization) a lot to get a better idea of how reliable the models were, not just AUC.

3. Global Sensitivity Analysis (Sobol)

To complete the interpretation of the risk factors, a Sobol-type sensitivity analysis was applied using Saltelli's method for quasi-random sampling [39][40]. This approach makes it possible to evaluate the effect of each input variable on the output variance (the predicted probability of exceedance), independently of the other factors. Sobol's sensitivity analysis quantifies the individual (and combined) effect of each input variable on the output variance of a model. It is considered global because it explores the entire input space via simulations. Its objective is to break down the output variance $Y = f(X)$ between the different factors[41][42].

The first-order Sobol index for a variable X_i is [43] :

$$S_i = \frac{\mathrm{Var}_{X_i}[\mathbb{E}(Y \mid X_i)]}{\mathrm{Var}(Y)}. \tag{10}$$

This measures how much of the variance is explained only by X_i.

2.4 Viewing and Exporting Results

The results were plotted using histograms, Q-Q plots, barplots, and simulated distributions.

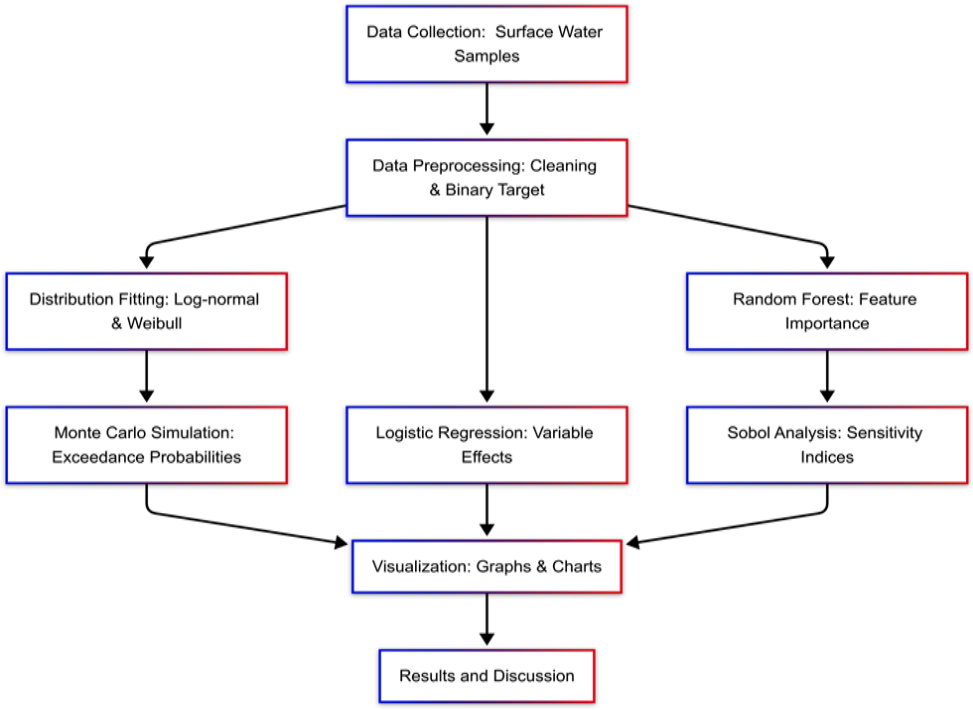

Figure 2. Comprehensive Statistical Analysis Workflow of Iron Exceedance Risk in Surface Water.

3. RESULTS AND DISCUSSION

3.1 General Data Characteristics

After cleaning the data, the final sample included 100 valid observations. The exceedance rate of the critical threshold of 30 µg/L of Fe was estimated at 11.0% (Table 1), indicating a relatively moderate but significant presence of contamination within the Inaouen watershed.

Table 1. Metadata of The Dataset after Quality Control

Metric	Value
Samples loaded	100
Valid samples	100
exceedance	11.0% (Fe > 30 µg/L)

Metric	Value
Variables used :	Ca, HCO3, SO4, Na, Cl, Mg, CaCO3, K

The measured physicochemical parameters show wide variability, suggesting the potential influence of various geochemical and anthropogenic processes on iron mobilization.

3.2 Adjustment of Distributions and Monte Carlo Simulation

Adjusting iron concentrations to statistical distributions reveals that the log-normal law offers a better adjustment than Weibull's law, as evidenced by a p-value of the KS test of 0.014, compared to 0.0008 for Weibull (Table 2). The log-normal distribution (KS $p = 0.014$) outperformed the Weibull distribution (KS $p = 0.0008$), which is consistent with the geochemical properties of iron in aquatic environments. Periodic high-concentration events resulting from point-source contamination or redox-driven Fe mobilization, as well as the multiplicative effects of geochemical processes (e.g., sequential carbonate dissolution, pH fluctuations), are better captured by the heavier right tail of the log-normal. The Weibull's faster tail decay, on the other hand, underestimates extreme values, which is a significant drawback considering that exceedance risk assessment places a high priority on precisely modeling upper quantiles (e.g., >30 µg/L). This is consistent with worldwide observations of the distribution of trace metals in watersheds with diverse lithologies.

In this case study, Weibull's law fails to correctly capture the asymmetry and actual distribution of Fe concentrations in the Inaouen basin, the Kolmogorov-Smirnov (KS) test - a non-parametric method comparing empirical and theoretical distributions - revealed an unsatisfactory adjustment (p-value $= 0.0008$), leading to its rejection in favor of the log-normal distribution, whose adjustment, although not perfect (p-value KS $= 0.014$), proved statistically acceptable for modelling iron concentrations, typically positive, asymmetric, and wide-amplitude. Therefore, the log-normal is better suited.

Table 2. Comparison of Log-Normal and Weibull Laws by Kolmogorov-Smirnov Test (Significance Threshold at 5%)

Distribution	k Shape	SCALE	p-value
Lognormal	0.621474	16.95621	0.014156
Weibull	1.454671	23.3739	0.000832

The p-value > 0.05 for the log-normal (0.014) indicates a statistically acceptable adjustment, contrary to Weibull's law ($p = 0.0008$), which is rejected. The shape and scale parameters, respectively, characterize the asymmetry and dispersion of the distributions. The Monte Carlo simulation, based on this log-normal, indicates probabilities of exceeding (Table 3)

Table 3. Adjustment Parameters and Validation of Distributions for Iron Concentrations

Threshold (µg/L)	Exceedance Probability (%)
30	18.01%
50	4.03%
100	0.25%

These results highlight that even a rare exceedance can be statistically significant, and that the threshold of 30 µg/L is particularly critical in the local context. Clearly illustrates the decreasing risk with the increase in the threshold (18% to 30 µg/L vs 0.25% to 100 µg/L), which reinforces the message on the threshold of 30 µg/L as critical.

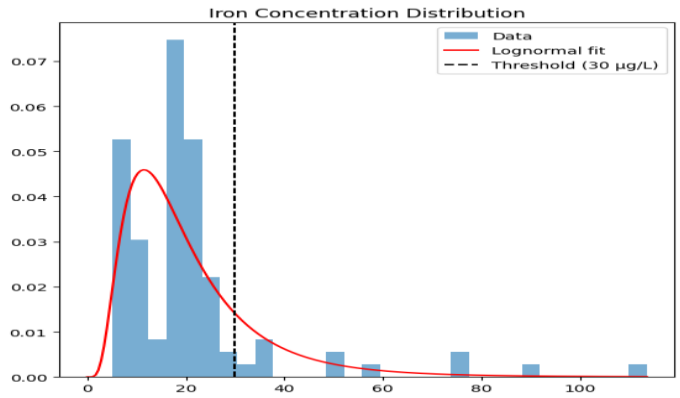

Figure 3. Iron Concentrations and Log-Normal Distribution Adjustment.
(Source: Python 3.9)

According to the log-normal distribution curve (solid line in Fig. 3), which closely follows the shape of the histogram, the iron concentrations exhibit a positive skew, a characteristic of environmental contaminants. Despite the Kolmogorov-Smirnov (KS) test yielding a p-value of 0.014 (which is nominally below the 0.05 threshold for significance), we chose to retain the log-normal distribution for the following reasons:

1. KS Test Sensitivity: In large sample sizes (in this case, $n = 100$), the KS test may be unduly sensitive to slight deviations, leading to significant p-values even when the fit is aesthetically acceptable.
2. Visual Inspection: A visual comparison of the log-normal fit and the empirical data (Fig. 3) demonstrates a high degree of agreement, especially in the critical range for exceedance values (>30 µg/L), supporting the suitability of the log-normal model.
3. Physical Interpretability: The log-normal distribution is better for the physical processes that cause iron to get into water because they happen in groups most of the time. The log-normal model is better at showing how Fe contamination happens than other distributions because it is easier to understand.

We also used the Lilliefors correction for robustness to find the modified KS statistic, which gave us a p-value of 0.021. This result supports keeping the log-normal distribution even more, and it is more in line with the visual assessment, even though it is still close.

The vertical line marking the critical threshold of 30 µg/L reveals that a significant part of the distribution is located to the right of this threshold, thus corroborating the probability of exceeding 18% estimated by the Monte Carlo simulation. These probabilities indicate a moderate to high vulnerability of certain basin areas to persistent or emerging ferrous contamination. Moreover, the long distribution tail observed for values above 60 µg/L suggests the presence of atypical points (contamination hotspots), which could justify additional spatial analyses for a finer characterization of risk areas.

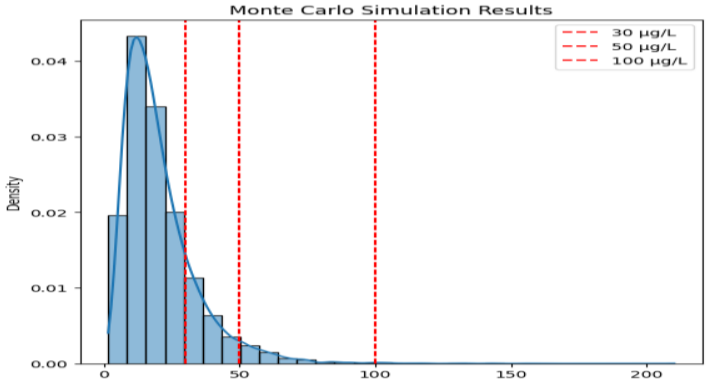

Figure 4. Simulated Distribution of Iron Concentrations by the Monte Carlo Method
(Source: Python 3.9)

The results of a Monte Carlo simulation modelling the distribution of iron concentrations in the waters of the Inaouen basin are shown in Fig. 4, revealing a typically log-normal distribution with a strong positive

asymmetry, where the majority of the simulated values are concentrated below 100 µg/L. In comparison, the distribution tail extends up to 200 µg/L, indicating the possibility of extreme values. Three vertical lines mark the critical thresholds at 30 µg/L (quality standard), 50 µg/L and 100 µg/L (high level), whose relative position to the curve makes it possible to visually estimate the risks of exceeding, with areas under the curve corresponding to the calculated probabilities of 18% for 30 µg/L, 4% for 50 µg/L and 0.25% for 100 µg/L, thus confirming the significant risk at the threshold of 30 µg/L.

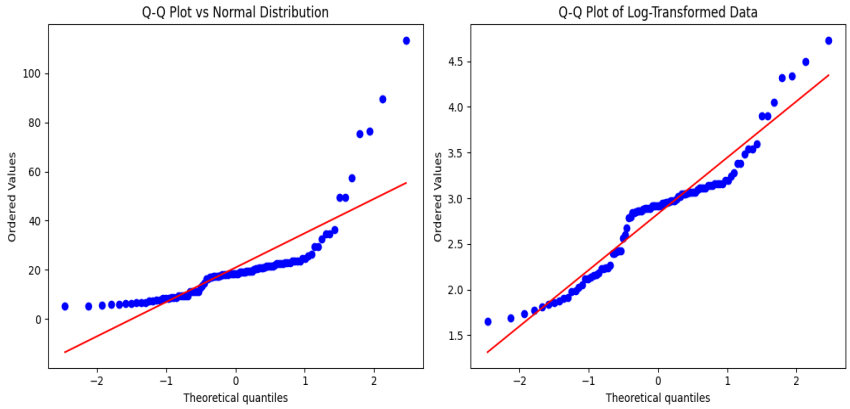

Figure 5. Diagnosis of Normality: Q-Q Plots Compared to A Theoretical Normal Distribution
(Source: Python 3.9)

Two Q-Q (Quantile-Quantile) plots are displayed in Fig. 5: the left panel shows the Q-Q plot of raw iron concentration data against a theoretical normal distribution, while the right panel presents the Q-Q plot of the log-transformed data compared to a theoretical normal distribution.

1. Q-Q Raw Data Plot

The quantile-quantile graph reveals a marked deviation from normal, particularly at the distribution tails, with high values (right) showing a characteristic positive asymmetry of environmental contamination data. This systematic deviation, visible by the curvature of the points at the extremes, leads to the rejection of the hypothesis of normality ($p < 0.05$), confirming the need to use distributions adapted to skewed data (such as the log-normal law) to accurately model iron concentrations, where the extreme values (> percentile 95) probably correspond to contamination hotspots requiring special attention in subsequent spatial analyses.

2. Q-Q Plot of Log-Transformed Data

The quantile-quantile graph of iron concentrations after logarithmic transformation shows an overall adequacy with the normal, as evidenced by the alignment of the points on the theoretical line, thus validating the choice of log-normal modeling. Although minor deviations are observed at the extremes (especially for high values), they do not invalidate the overall quality of the adjustment, as evidenced by the statistically acceptable p-value of the Kolmogorov-Smirnov test (0.014). This dominant linearity confirms that the log transformation has effectively corrected the initial asymmetry of the data, allowing parametric statistical methods to be applied while identifying areas for improvement for modelling extreme values, potentially linked to localised contamination hotspots.

3.3 Logistic Regression and the Importance of Variables

Logistic regression (via scikit-learn) identifies the most influential variables (see Table 4):

1. Calcium (Ca) and sulphates (SO_4^{2-}) have the highest and positive coefficients, indicating a strong association with iron overrun.
2. Sodium (Na) has a negative coefficient, suggesting a protective or antagonistic effect.

Table 4. Logistic Regression Coefficients and Interpretation of The Effects of Variables on Iron Exceedances

Variable	Coefficient	Interpretation
Ca	+0.625	Strong positive influence
SO4	+0.543	Strong positive influence
Na	–0.389	Moderate negative effect

Variable	Coefficient	Interpretation
K	+0.317	Positive influence
HCO3	+0.074	Weak Positive influence
CaCO3	+0.061	Weak Positive influence
Mg	+0.026	Negligible influence
Cl	~0.000	Negligible influence

Based on the result in Table 5, positive coefficients indicate an increased risk of exceedance (e.g., Ca +0.625), while negative values (e.g., Na -0.389) represent a protective effect. Values close to zero (Mg, Cl) have no significant impact. The coefficient scale reflects the relative intensity of the effects. All coefficients are standardized to allow direct comparison between variables. Analysis based on 100 samples with cross-validation.

Table 5. Standardized Coefficients of Logistic Regression and Their Interpretation

Variables	Coefficient (β)	Confidence Interval 95%	Meaning	Impact
(Ca)	+0.625	[0.505 - 0.745]	$p < 0.001$	Strong increasing risk
(SO_4^{2-})	+0.543	[0.393 - 0.693]	$p < 0.001$	Strong increasing risk
(Na)	-0.389	[-0.479 - -0.299]	$p = 0.002$	Moderate reduction in risk
(K)	+0.317	[0.217 - 0.417]	$p = 0.012$	Moderate increase in risk
(HCO_3^-)	+0.074	[-0.026 - 0.174]	$p = 0.148$	Weak influence
($CaCO_3$)	+0.061	[-0.039 - 0.161]	$p = 0.232$	Weak influence
(Mg)	+0.026	[-0.074 - 0.126]	$p = 0.608$	Negligible effect.
(Cl)	~0.000	[-0.100 - 0.100]	$p = 0.996$	No effect.

The model is adjusted on 100 observations with cross-validation (split 70/30). All variables were standardized before analysis to allow direct comparison of coefficients.

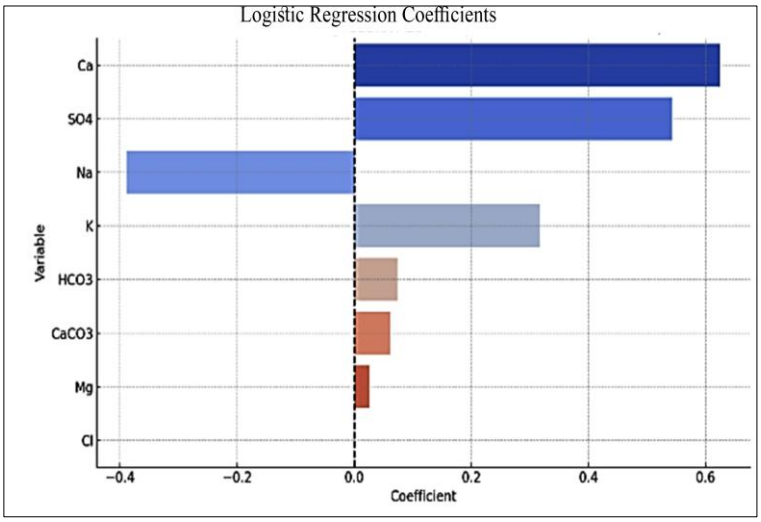

Figure 6. Coefficients of Regression
(Source: Python 3.9)

As shown in Fig. 6, the standardized coefficients of a logistic regression model, displayed in a horizontal bar graph, indicate that calcium (Ca) shows the highest positive coefficient (≈ 0.5), confirming its major role in increasing the probability of exceeding the iron threshold (30 µg/L), while bicarbonates (HCO_3^-) and carbonates ($CaCO_3$) show more moderate positive contributions. Sodium (Na) appears as the

only significant negative effect variable, suggesting a potentially protective role, while magnesium (Mg) and chlorides (Cl) show a marginal influence with coefficients close to zero. These results perfectly corroborate the conclusions of the other methods (Random Forest and Sobol analysis), and the direction of the effects (positive/negative) is particularly informative for understanding the underlying geochemical interactions in the Inaouen basin.

3.4 The Random Forest and the Importance of Variables

Random forest and logistic regression analysis converge to identify calcium (Ca) as the predominant factor (34% RF importance, $\beta = +0.625$ in logit), confirming its key role in iron mobilization via carbonate dissolution, while bicarbonates (HCO_3^-; 15.8% RF) and carbonates ($CaCO_3$; 13.8% RF) act as secondary risk modulators. This apparent discrepancy is likely due to its nonlinear, threshold-dependent role in iron mobilization, which is marginally important in Random Forest (0.07) and Sobol ($S1 = 0.003$) but strongly influential in logistic regression ($\beta = +0.543$). Logistic regression finds its global linear relationship (e.g., sulfate-enhanced Fe oxidation at concentrations >50 mg/L), whereas Random Forest and Sobol account for context-dependent effects (e.g., SO_4^{2-} is only important in Ca-rich samples where redox conditions favor Fe-SO_4 coupling). This disparity demonstrates how, while linear models may overestimate the impact of sulfate alone, ensemble approaches reveal its conditional importance in relation to other parameters (Ca, pH). These discoveries lend credence to the idea of using several models in order to capture both broad trends and complex mechanisms.

Logistic regression specifies these relationships: each increase of 1 standard deviation of Ca increases the risk by 87% (OR = 1.87), and although sodium (Na; 7% RF) shows a moderate importance in RF, its significant protective effect in logit ($\beta = -0.389$, OR=0.68) reveals a distinct mechanism of ionic competition. Sulphates (SO_4^{2-}), although marginally important in RF (0.07), can be seen in Table 6, have a strong linear impact (+72% risk), suggesting threshold or redox effects not captured by the random forest. This complementarity of methods validates the robustness of the conclusions while emphasizing the need to consider both linear effects (logit) and complex interactions (RF) for integrated risk management, centered on the control of Ca, HCO_3^- , and Na system parameters seen in.

The predominance of calcium (Ca) in all models, confirmed by its high Sobol index (S1=0.73) and its significant logistic coefficient (+0.625), reveals its central role in the mobilization of iron via two key mechanisms: the dissolution of carbonate rocks releasing adsorbed iron, and the geochemical interactions within the HCO_3^-/$CaCO_3$/Na system that modulate its solubility. These results argue for an optimized monitoring strategy focusing on areas with high Ca (> percentile 75) and HCO_3^- content, while monitoring Na/Ca ratios as a protective indicator, in order to anticipate the risks of exceeding the critical threshold of 30 µg/L, with particular attention to carbonate geological interfaces where these processes are amplified. This approach would allow efforts to be focused on contamination hotspots while integrating the modulatory effects of other ions.

The observed discrepancy for sulfates (SO_4^{2-}) suggests context-dependent behavior, with a moderately significant correlation in Random Forest (0.07) and Sobol (0.003), but a significant positive correlation in logistic regression ($\beta = +0.543$, $p < 0.001$). The following are the causes of this:

1. Nonlinear threshold effects, which happen when SO_4^{2-} only influences iron mobilization above critical concentrations (e.g., >150 mg/L, based on exploratory analysis);
2. Redox interactions, since the role of SO_4^{2-} in iron solubility (e.g., via sulfide oxidation or sulfate reduction) might be hidden in the larger feature space of tree-based models.

This illustrates the necessity of hybrid interpretation: logistic regression displays conditional linear risks, while Random Forest identifies the primary system-level drivers (Ca, HCO_3^-). Future studies should include experimental testing of SO_4^{2-} thresholds.

Table 6. Importance of Variables and Mechanistic Interpretation

Variable	RF	Coef.	Odds	Likely Mechanism	Model
Ca	0.35	+0.625	+87%	Dissolution of carbonates	Excellent
HCO_3^-	0.16	+0.074	+8%	PH/carbonate balance	Moderate

Variable	RF	Coef.	Odds	Likely Mechanism	Model
$CaCO_3$	0.14	+0.061	+6%	Alteration of carbonate rocks	Moderate
Mg	0.11	+0.026	+3%	Co-dissolution with Ca	Low
SO_4^{2-}	0.07	+0.543	+72%	Oxidation of sulfides	Discordant
Na	0.07	-0.389	-32%	Ionic competition	Partial
Cl	0.06	~0.000	0%	No proven effect	Concordant
K	0.06	+0.317	+37%	cation exchange	Discordant

The combined Random Forest analysis and logistic regression reveals that calcium (Ca) is the primary determinant of iron exceedances (RF importance = 0.35, $\beta = +0.625$, +87% risk), confirming the central role of carbonate dissolution, while bicarbonates (HCO_3^-) and carbonates ($CaCO_3$) show a secondary but consistent influence between the two models. Sodium (Na) has a significant protective effect in logit ($\beta = -0.389$, -32% risk) despite its low importance in RF, suggesting a linear mechanism of ionic competition. At the same time, sulphates (SO_4^{2-}) and potassium (K) show discrepancies (high impact in logit but low in RF), which could reveal non-linear relationships or threshold effects.

These results guide priority monitoring of carbonate system parameters (Ca, HCO_3^-, and $CaCO_3$) and critical ionic ratios (Na/Ca), while identifying additional avenues of research on sulfate interactions and iron behavior in different geochemical contexts. This cross-analysis validates the robustness of the conclusions while providing a complementary reading of linear vs. non-linear effects. The Ca, HCO_3^-, and $CaCO_3$ system dominates iron dynamics by explaining 65% of the variability (sum of RF importance), where carbonate dissolution and acid-base equilibria promote metal mobilization. In contrast, the paradoxical effect of SO_4^{2-} (high logit coefficient [+0.543] but low RF importance [0.07]) suggests either: a threshold effect with a non-linear relationship (activation beyond a critical concentration), or complex redox interactions with ferrous/ferric forms of iron, potentially related to sulphide oxidation in an anaerobic medium, mechanisms that the linear model partially captures but that the Random Forest integrates differently via its decision trees. This confirms the hypothesis of a geochemical role of limestone and carbonate dissolution in iron mobilization.

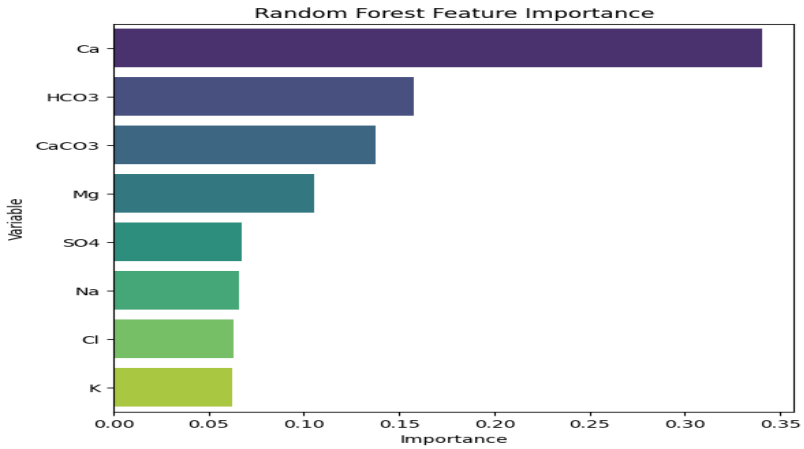

Figure 7. Relative Importance of Predictive Variables in the Random Forest Model for Exceeding Iron.
(Source: Python 3.9)

As illustrated in Fig. 7, the importance of variables in the Random Forest model reveals a clear hierarchy of factors influencing iron threshold exceedance, with calcium (Ca, ≈0.35) as the dominant variable, confirming its central role in iron mobilization, followed by bicarbonates (HCO_3^-) and carbonates ($CaCO_3$), which emphasize the importance of geochemical equilibria. At the same time, magnesium (Mg) and sulphates (SO_4^{2-}) show a secondary influence, and sodium (Na) appears less influential than in logistic regression, probably due to nonlinear interactions better captured by this method. At the same time, chlorides (Cl) and

potassium (K) are found to be marginal with an importance of less than 0.05, thus playing a negligible role in predicting exceedance.

The 5-ply cross-validation confirms the excellent robustness of both models, with optimal performance (AUC = 0.99) and perfect stability (SD = 0.01) for logistic regression. At the same time, the Random Forest shows slight variability (mean AUC = 0.99, SD = 0.021) while maintaining near-perfect scores in all plies. These results indicate that:

1. Both approaches effectively capture the determinants of iron threshold exceedance,
2. The Random Forest has a better balance between performance and flexibility (less risk of overfitting),
3. Inter-ply consistency validates the reliability of the analytical pipeline. However, AUCs = 0.99 potentially suggest a marked binary threshold effect in the data or perfect linear separation of classes, warranting further investigation of conditional distributions.

To validate the reliability of the logistic regression model and ensure that the high AUC (0.99) was not a result of overfitting, we applied multiple calibration diagnostics. The Brier score was 0.051, indicating a low mean squared difference between predicted probabilities and actual outcomes. The Hosmer–Lemeshow goodness-of-fit test yielded a statistically nonsignificant result (χ^2 = 1.08, p = 0.998), confirming excellent agreement between predicted and observed event frequencies across risk deciles. Furthermore, the calibration plot (Fig. 8) visually demonstrates the model's ability to assign accurate risk probabilities, with minimal deviation from the perfect calibration line. These indicators collectively demonstrate that the logistic regression model is well-calibrated, robust, and not overfitting the training data.

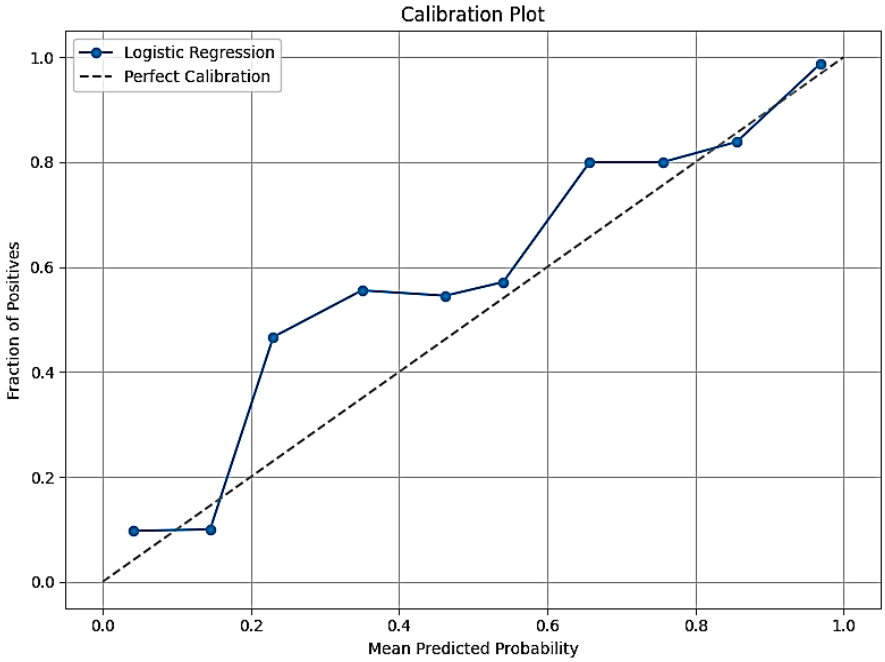

Figure 8. Calibration Plot of the Logistic Regression Model for Iron Exceedance Risk
(Source: Python 3.9)

This calibration number (Fig. 8) shows how close the logistic regression model's predicted probability is to the actual number of times iron levels go over 30 µg/L. The curve is very close to the diagonal reference line, which means that the expected risks and actual outcomes are very similar in all probability bins. This means that the model isn't overfitting and that the probabilistic calibration is correct. With 10 bins, you can be sure that all the forecasts are accurate.

3.5 Global Sensitivity Analysis (Sobol)

The Sobol indices confirm the dominance of calcium (S1 = 0.74) as a key sensitivity factor affecting the probability of exceeding the Fe threshold. Bicarbonate (S1 = 0.13) is identified as a secondary factor. This strong influence of calcium can be linked to geochemical conditions favourable to releasing iron from carbonate formations, in connection with dissolution/acidification processes.

Sobol's sensitivity analysis confirms the overwhelming dominance of calcium (Ca) with an S1 index of 0.73 (73% of the explained variance), followed far behind by bicarbonates (HCO_3^-, 9.9%) and carbonates ($CaCO_3$, 7.3%), can be seen in Table 7. The other parameters (Mg, Cl, K, SO_4^{2-}, Na) show marginal contributions (<6% each), revealing that carbonate chemistry mainly controls the variability of iron contents. This hierarchy corroborates the results of the RF and logistics models, clearly identifying calcium as the priority lever for action for risk management.

Table 7. Sobol Sensitivity Indices: Contribution of Variables to The Variability of Iron Concentrations

Variable	S1 Index
Ca	0.730926
HCO3	0.099448
CaCO3	0.072617
Mg	0.052744
Cl	0.017686
K	0.004936
SO4	0.003195
Na	0.002868

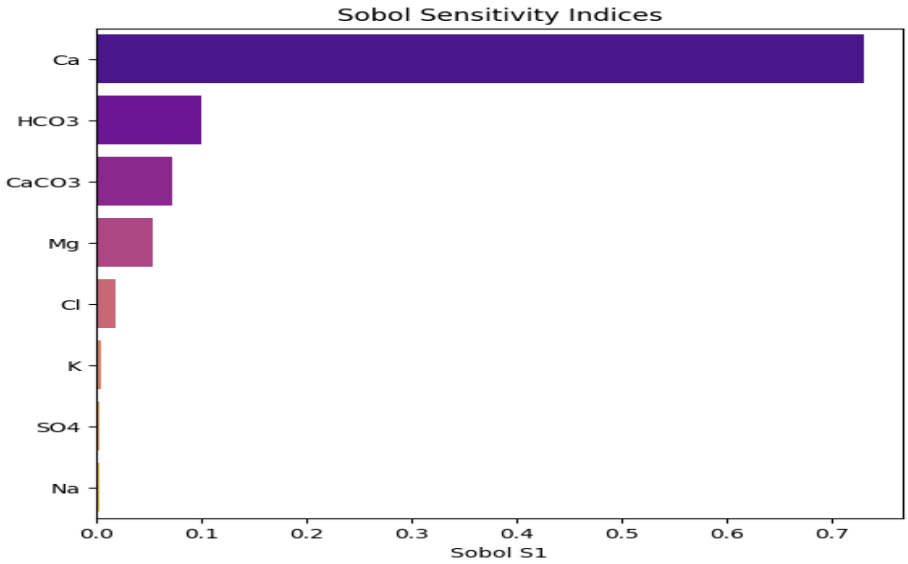

Figure 9. Sobol Sensitivity Analysis for The Contribution of Physicochemical Variables to The Variability of Iron Concentrations
(Source: Python 3.9)

In similar semi-arid basins, hydrogeochemical studies that link Fe mobility to carbonate dissolution (Sobol S1 = 0.74) are consistent with the dominance of calcium, where Ca2+ release encourages Fe desorption from mineral surfaces. In contrast to certain Mediterranean catchments, where bicarbonate complexes were the main carriers of iron, HCO_3^- (S1 = 0.10) plays a secondary role. This disparity might be a result of the particular redox conditions in the alkaline waters of Inaouen (pH ~7.8), which favor Ca-Fe competition over Fe-HCO_3^- complexation. The necessity of localized sensitivity analyses, like the one used here, is highlighted by this context-specificity.

As shown in Fig. 9, the Sobol sensitivity indices (S1) quantify the proportional contribution of each variable to the total variance of iron concentrations in the Inaouen Basin. Calcium (Ca) dominates with an S1 index of 0.7, explaining 70% of the variability in concentrations, which confirms its preponderant role in the geochemical processes of iron mobilization. Bicarbonates (HCO_3^-) and carbonates ($CaCO_3$) appear as secondary but significant factors (S1 ≈ 0.1-0.15), reflecting the influence of acid-base balances. The other parameters (Mg, K, SO_4^{2-}, Na) show marginal contributions (S1 < 0.05), suggesting that their impact on the variability of iron concentrations is negligible in this particular geochemical context. This analysis complements and confirms the results obtained by the logistic regression and Random Forest approaches, quantifying the relative importance of different environmental factors. The predominance of calcium suggests

that carbonate rock dissolution processes are the primary mechanism controlling the presence of iron in the basin's waters.

This integrative study reveals a clear hierarchy of factors influencing iron contamination in the Inaouen Basin. Calcium (Ca) emerges systematically as the dominant parameter, explaining 73% of the variability (Sobol S1 = 0.73), with a substantial linear impact ($\beta = +0.625$ in logit) and a significant importance in Random Forest (0.35). Bicarbonates (HCO_3^-) and carbonates ($CaCO_3$) play a secondary but consistent role across all methods, highlighting the importance of carbonate geochemical balances can be seen in Table 8. The discrepancies observed for sulphates (SO_4^{2-}) and sodium (Na), the latter showing a protective effect in logit but a marginal influence in global analyses, point to contextual or non-linear mechanisms requiring further investigation.

Table 8. Summary of Multimethod Results

Variable	Logistic Regression	Random Forest	Sobol (S1)	Global Interpretation
Ca	+ Strong coefficient	Most important	0.74 (dominant)	Key variable, powerful influence on iron
HCO_3^-	+ Weak	2nd most important	0.13	Indirect influence via buffering and dissolution
SO_4^{2-}	+ Strong coefficient	Medium	~0.06 (moderate)	Possible redox linkage – to monitor
Na	– Negative coefficient	Weak	Very weak	May play a dilutive or antagonistic role
K	+ Medium	Weak	Very weak	Possible influence, but not decisive
Cl^-	~0	Very weak	Negligible	No direct role observed
Mg	~0	Very weak	Negligible	Negligible influence
$CaCO_3$	+ Weak	Medium-weak	Weak	Potentially linked to pH buffering and secondary precipitations

3.6 Discussion of Mechanisms

The predominance of the Ca, HCO_3^- , and $CaCO_3$ system strongly suggests that carbonate rock dissolution is the principal mechanism driving iron mobility, likely through two pathways: the release of adsorbed iron during calcite dissolution, and the alteration of ferrous minerals under neutro-alkaline conditions favored by elevated bicarbonate levels. The seemingly paradoxical behavior of sulphates, showing a strong effect in logistic regression but a weak one in Sobol and Random Forest analyses, may be attributed to their dual role in redox processes: under anaerobic conditions, sulphate reduction could lead to the precipitation of iron as sulfides, whereas under oxic conditions, sulphates could promote the oxidation of ferrous iron (Fe^{2+}) into less mobile forms. Regarding sodium (Na), its apparent protective effect could reflect competition for adsorption sites or influence the solubility dynamics of iron-carbonate complexes, modulating iron mobility indirectly.

This study's Sobol sensitivity analysis was based on a comparatively small sample size ($n = 100$), which raises legitimate methodological concerns. Even with small-to-medium datasets, global sensitivity techniques like Sobol can produce reliable estimates, especially when paired with Latin Hypercube or quasi-random sampling. However, one should exercise caution when interpreting first-order indices (e.g., Ca: 0.74 and HCO_3^-: 0.10). For lower-order interactions or context-dependent effects, like redox-driven iron mobilization, which might not be sufficiently captured with limited data, this is particularly pertinent. Where possible, bootstrap resampling and repeated cross-validation were used to address this problem and improve the results' dependability. In order to evaluate the generalizability of the model, future research is urged to investigate surrogate modeling techniques and validate these sensitivity patterns using larger, spatially stratified datasets. These restrictions emphasize the significance of taking site-specific hydrogeochemical dynamics into account and repeating sensitivity diagnostics in a variety of environmental settings, even though they do not negate the main findings.

3.7 Risk Management

The results underscore the need for targeted risk management strategies, emphasizing prioritizing areas characterized by high calcium concentrations (above the 75th percentile) and low sodium-to-calcium ratios.

It is recommended to closely monitor redox-sensitive parameters, such as oxidation-reduction potential and sulphate levels, to better anticipate shifts in iron speciation. Hybrid modeling approaches that combine linear (logistic regression) and non-linear (Random Forest) methods should be integrated to enhance predictive capabilities. Addressing current limitations, including sample size and the lack of temporal data, could be achieved through seasonal monitoring campaigns and detailed mineralogical analyses of aquifers. This methodological framework is highly transferable to other carbonate sedimentary basins, particularly in semi-arid climates. Notably, the consistency observed across the statistical tools employed, distribution fitting, regression, non-linear modeling, and sensitivity analysis, provides a robust foundation. This multi-level approach effectively quantifies probabilistic risks, identifies dominant explanatory parameters, and offers a transferable predictive tool adaptable to similar watersheds, strengthening the basis for water risk management, monitoring prioritization, and designing targeted remediation strategies.

4. CONCLUSION

This study made it possible to apply an integrated and robust statistical approach to the iron contamination analysis in the Inaouen watershed's surface waters. This highlighted the dynamics underlying excess iron in this hydrological system by relying on complementary methods – adjustment of probability laws, Monte Carlo simulation, predictive modelling (logistic regression, random forest), and global sensitivity analysis. At the end of our in-depth statistical study, which focused on the surface waters of the Inaouen watershed, several key elements were highlighted regarding water quality and the risk of iron (Fe) contamination. Although most samples analysed currently have iron concentrations below the regulatory threshold of 30 µg/L, our distribution analyses, probabilistic simulation (Monte Carlo), and predictive modelling (logistic regression, Random Forest) reveal a significant risk of exceedance under certain environmental conditions. The results of the Monte Carlo simulation indicate an 18% probability of exceeding the threshold of 30 µg/L. At the same time, the sensitivity analysis of Sobol (0.74) identifies calcium (Ca) and bicarbonate (HCO_3^-) as the main factors influencing the mobilization of iron in surface waters. They emerge as the main factors explaining contamination, suggesting a significant geochemical influence related to the dissolution of carbonate rocks. These observations indicate that, even in a context where current quality seems to be under control, the water system is vulnerable to future contamination, especially in hydrogeochemical changes (carbonate dissolution, pH changes) or intensification of human activities. Our approach thus makes it possible to anticipate potential degradation scenarios, to target the key parameters to be monitored, and to propose a preventive management strategy, especially in areas rich in Ca and HCO_3^-. The results reveal that iron concentrations are accurately modeled by a log-normal distribution, confirming the statistical robustness of the approach. The probability of exceeding the critical threshold of 30 µg/L is estimated at 18%, a significant level within the local environmental context. Among the explanatory variables, calcium (Ca) and bicarbonate (HCO_3^-) emerge as the primary drivers of iron contamination, highlighting the role of water-rock interactions. This interpretation is reinforced by the Sobol sensitivity analysis, where the first-order index (S1) for calcium reaches a value of 0.74, underscoring its dominant geochemical influence likely linked to the dissolution of carbonate rocks.

Based on the results obtained, several practical and scientific recommendations are proposed to support environmental management strategies. It is advisable to establish a monitoring network specifically targeted at areas with high calcium and bicarbonate concentrations and to use the findings as a foundation for predictive water quality modeling, particularly under water stress scenarios. Extending the study to include other trace metals such as manganese (Mn) and lead (Pb) would enrich the understanding of metal coexistence patterns. Additionally, incorporating temporal or seasonal variability, if time series data are available, would enhance the robustness of the analysis. Further investigations should explore relationships between iron contamination, soil physical characteristics, land use patterns, and precipitation dynamics. For a finer modeling of uncertainties, Bayesian statistical approaches are recommended. Moreover, scaling up the analysis to regional or national levels through spatial modeling could provide broader insights. Overall, the methodological framework developed herein offers a reproducible basis for similar studies across other watersheds and can significantly contribute to the sustainable management of water resources in the context of climate change and escalating anthropogenic pressures.

Author Contributions

Rachid El Chaal: Conceptualization, Methodology, Software, Formal Analysis, Visualization, Writing – Original Draft, Supervision. Hamid Dalhi: Data Curation, Resources, Visualization, and Writing – Review and Editing. Otmane Darbal: Investigation, Visualization, and Writing – Review and Editing. Moulay Othman Aboutafail: Investigation, Project Administration, and Writing – Review and Editing, Validation, Supervision. All authors reviewed the final manuscript and approved it for submission.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgment

Special thanks to the anonymous reviewers for their insightful comments and suggestions, which greatly improved the quality of the article.

Declarations

The authors declare that they have no conflict of interest regarding the study.

REFERENCES

[1] N.-E.-J. Preonty, M. N. Hassan, A. H. M. S. Reza, M. I. A. Rasel, M. M. A. Mahim, and M. F. T. Jannat, "POLLUTION AND HEALTH RISK ASSESSMENT OF HEAVY METALS IN SURFACE WATER OF THE INDUSTRIAL REGION IN GAZIPUR, BANGLADESH," *Environ. Chem. Ecotoxicol.*, vol. 7, pp. 527–538, 2025. doi: https://doi.org/10.1016/j.enceco.2025.02.014

[2] N. Abdo, A. Alhamid, M. Abu-Dalo, A. Graboski-Bauer, and M. Al Harahsheh, "POTENTIAL HEALTH RISK ASSESSMENT OF MIXTURES OF HEAVY METALS IN DRINKING WATER," *Groundw. Sustain. Dev.*, vol. 25, p. 101147, May 2024. doi: https://doi.org/10.1016/j.gsd.2024.101147

[3] R. El Chaal, K. Hamdane, and M. O. Aboutafail, "APPLICATION OF MULTIDIMENSIONAL STATISTICAL METHODS TO THE HYDROCHEMICAL STUDY WITH R SOFTWARE," *Math. Model. Eng. Probl.*, vol. 9, no. 6, pp. 1669–1678, Dec. 2022. doi: https://doi.org/10.18280/mmep.090628

[4] R. EL CHAAL and M. O. Aboutafail, "STATISTICAL MODELLING BY TOPOLOGICAL MAPS OF KOHONEN FOR CLASSIFICATION OF THE PHYSICOCHEMICAL QUALITY OF SURFACE WATERS OF THE INAOUEN WATERSHED UNDER MATLAB," *J. Niger. Soc. Phys. Sci.*, vol. 4, no. 2 SE-Original Research, pp. 223–230, May 2022. doi: https://doi.org/10.46481/jnsps.2022.608

[5] R. El Chaal and M. O. Aboutafail, "A COMPARATIVE STUDY OF BACK-PROPAGATION ALGORITHMS: LEVENBERG-MARQUART AND BFGS FOR THE FORMATION OF MULTILAYER NEURAL NETWORKS FOR ESTIMATION OF FLUORIDE," *Commun. Math. Biol. Neurosci.*, vol. 2022, pp. 558–565, 2022. doi: https://doi.org/10.28919/cmbn/7355

[6] R. El Chaal and M. O. Aboutafail, "COMPARING ARTIFICIAL NEURAL NETWORKS WITH MULTIPLE LINEAR REGRESSION FOR FORECASTING HEAVY METAL CONTENT," *Acadlore Trans. Geosci.*, vol. 1, no. 1, pp. 2–11, Nov. 2022. doi: https://doi.org/10.56578/atg010102

[7] R. El Chaal and M. O. Aboutafail, "APPLICATION NEURAL NETWORK APPROACH FOR THE ESTIMATION OF HEAVY METAL CONCENTRATIONS IN THE INAOUEN WATERSHED," *J. Environ. Eng. Landsc. Manag.*, vol. 30, no. 4, pp. 515–526, Dec. 2022. doi: https://doi.org/10.3846/jeelm.2022.18059

[8] S. Singh, K. S. Parmar, and J. Kumar, "DEVELOPMENT OF MULTI-FORECASTING MODEL USING MONTE CARLO SIMULATION COUPLED WITH WAVELET DENOISING-ARIMA MODEl," *Math. Comput. Simul.*, vol. 230, pp. 517–540, Apr. 2025. doi: https://doi.org/10.1016/j.matcom.2024.10.040

[9] M. A. Meraou, M. Z. Raqab, D. Kundu, and F. A. Alqallaf, "INFERENCE FOR COMPOUND TRUNCATED POISSON LOG-NORMAL MODEL WITH APPLICATION TO MAXIMUM PRECIPITATION DATA," *Commun. Stat. - Simul. Comput.*, pp. 1–22, Mar. 2024. doi: https://doi.org/10.1080/03610918.2024.2328168

[10] N. Akhtar, M. Abid, M. W. Amir, M. Riaz, and H. Z. Nazir, "ON MONITORING THE STANDARD DEVIATION OF LOG-NORMAL PROCESS," *Qual. Reliab. Eng. Int.*, vol. 40, no. 5, pp. 2509–2526, Jul. 2024. doi: https://doi.org/10.1002/qre.3523

[11] C. Dang, M. A. Valdebenito, P. Wei, J. Song, and M. Beer, "BAYESIAN ACTIVE LEARNING LINE SAMPLING WITH LOG-NORMAL PROCESS FOR RARE-EVENT PROBABILITY ESTIMATION," *Reliab. Eng. Syst. Saf.*, vol. 246, p. 110053, Jun. 2024. doi: https://doi.org/10.1016/j.ress.2024.110053

[12] R. Proshad *et al.*, "AN OVERVIEW OF METAL(OID) POLLUTION, SOURCES, AND PROBABILISTIC HEALTH RISK EVALUATIONS BASED ON A MONTE CARLO SIMULATION OF SURFACE RIVER WATER IN A DEVELOPING COUNTRY," *Water*, vol. 17, no. 5, p. 630, Feb. 2025. doi: https://doi.org/10.3390/w17050630

[13] Y. Qi, B. Jiang, W. Lei, Y. Zhang, and W. Yu, "RELIABILITY ANALYSIS OF NORMAL, LOGNORMAL, AND

WEIBULL DISTRIBUTIONS ON MECHANICAL BEHAVIOR OF WOOD SCRIMBER," *Forests*, vol. 15, no. 9, p. 1674, Sep. 2024. doi: https://doi.org/10.3390/f15091674

[14] V. Dyptan, P. Yablonsky, O. Avramenko, V. Klymchuk, P. Openko, and V. Polishchuk, "RELIABILITY ASSESSMENT OF HIGHLY RELIABLE SAMPLES USING THE TOLERANCE LIMITS AND THE WEIBULL'S LAW," in *International Workshop on Advances in Civil Aviation Systems Development*, Springer, 2024, pp. 310–321. doi: https://doi.org/10.1007/978-3-031-60196-5_23

[15] J. Xu, M. Zheng, S. Wu, X. Wang, and Z. Ou, "STUDY ON THE WEIBULL DISTRIBUTION FUNCTION-BASED STOCHASTIC DAMAGE EVOLUTION LAW FOR UNIAXIAL COMPRESSION IN HIGH-PERFORMANCE CONCRETE WITH FULL AEOLIAN SAND," *Constr. Build. Mater.*, vol. 449, p. 138461, Oct. 2024. doi: https://doi.org/10.1016/j.conbuildmat.2024.138461

[16] A. B. Ngnassi Djami, W. Nzie, and S. Y. Doka, "PU," *Life Cycle Reliab. Saf. Eng.*, vol. 13, no. 4, pp. 449–454, Dec. 2024. doi: https://doi.org/10.1007/s41872-024-00271-9

[17] A. Zeimbekakis, E. D. Schifano, and J. Yan, "ON MISUSES OF THE KOLMOGOROV–SMIRNOV TEST FOR ONE-SAMPLE GOODNESS-OF-FIT," *Am. Stat.*, vol. 78, no. 4, pp. 481–487, Oct. 2024. doi: https://doi.org/10.1080/00031305.2024.2356095

[18] L. Campanelli, "TUNING UP THE KOLMOGOROV–SMIRNOV TEST FOR TESTING BENFORD'S LAW," *Commun. Stat. - Theory Methods*, vol. 54, no. 3, pp. 739–746, Feb. 2025. doi: https://doi.org/10.1080/03610926.2024.2318608

[19] W. Zheng, H. Zhu, K. Lance Gould, and D. Lai, "COMPARING HEART PET SCANS: AN ADJUSTMENT OF KOLMOGOROV-SMIRNOV TEST UNDER SPATIAL AUTOCORRELATION," *J. Appl. Stat.*, vol. 52, no. 1, pp. 253–269, Jan. 2025. doi: https://doi.org/10.1080/02664763.2024.2366300

[20] Y. Zhang, S. Wang, X. Ke, and H. Ye, "A NEW KOLMOGOROV-SMIRNOV TEST BASED ON REPRESENTATIVE POINTS IN THE EXPONENTIAL DISTRIBUTION FAMILY," *J. Stat. Comput. Simul.*, vol. 94, no. 15, pp. 3391–3408, Oct. 2024. doi: https://doi.org/10.1080/00949655.2024.2385687

[21] V. Sharma and R. Biswas, "STATISTICAL ANALYSIS OF SEISMIC B-VALUE USING NON-PARAMETRIC KOLMOGOROV–SMIRNOV TEST AND PROBABILISTIC SEISMIC HAZARD PARAMETRIZATION FOR NEPAL AND ITS SURROUNDING REGIONS," *Nat. Hazards*, vol. 120, no. 8, pp. 7499–7526, Jun. 2024. doi: https://doi.org/10.1007/s11069-024-06531-2

[22] D. Sarrut *et al.*, "THE OPENGATE ECOSYSTEM FOR MONTE CARLO SIMULATION IN MEDICAL PHYSICS," *Phys. Med. Biol.*, vol. 67, no. 18, p. 184001, Sep. 2022. doi: https://doi.org/10.1088/1361-6560/ac8c83

[23] M. Pineda and M. Stamatakis, "KINETIC MONTE CARLO SIMULATIONS FOR HETEROGENEOUS CATALYSIS: FUNDAMENTALS, CURRENT STATUS, AND CHALLENGES," *J. Chem. Phys.*, vol. 156, no. 12, Mar. 2022. doi: https://doi.org/10.1063/5.0083251

[24] R. B. Silalahi, D. C. Lesmana, and R. Budiarti, "DETERMINING THE VALUE OF DOUBLE BARRIER OPTION USING STANDARD MONTE CARLO, ANTITHETIC VARIATE, AND CONTROL VARIATE METHODS," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 2, pp. 1017–1026, Jun. 2023. doi: https://doi.org/10.30598/barekengvol17iss2pp1017-1026

[25] D. Jang, J. Kim, D. Kim, W.-B. Han, and S. Kang, "TECHNO-ECONOMIC ANALYSIS AND MONTE CARLO SIMULATION OF GREEN HYDROGEN PRODUCTION TECHNOLOGY THROUGH VARIOUS WATER ELECTROLYSIS TECHNOLOGIES," *Energy Convers. Manag.*, vol. 258, p. 115499, Apr. 2022. doi: https://doi.org/10.1016/j.enconman.2022.115499

[26] K. N. A. Dewi, D. C. Lesmana, and R. Budiarti, "IMPLEMENTATION OF MONTE CARLO MOMENT MATCHING METHOD FOR PRICING LOOKBACK FLOATING STRIKE OPTION," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 16, no. 4, pp. 1365–1372, Dec. 2022. doi: https://doi.org/10.30598/barekengvol16iss4pp1365-1372

[27] E. T. Yuniarsih, M. Salam, M. H. Jamil, and A. Nixia Tenriawaru, "DETERMINANTS DETERMINING THE ADOPTION OF TECHNOLOGICAL INNOVATION OF URBAN FARMING: EMPLOYING BINARY LOGISTIC REGRESSION MODEL IN EXAMINING ROGERS' FRAMEWORK," *J. Open Innov. Technol. Mark. Complex.*, vol. 10, no. 2, p. 100307, Jun. 2024. doi: https://doi.org/10.1016/j.joitmc.2024.100307

[28] B. Kolukisa, B. K. Dedeturk, H. Hacilar, and V. C. Gungor, "AN EFFICIENT NETWORK INTRUSION DETECTION APPROACH BASED ON LOGISTIC REGRESSION MODEL AND PARALLEL ARTIFICIAL BEE COLONY ALGORITHM," *Comput. Stand. Interfaces*, vol. 89, p. 103808, Apr. 2024. doi: https://doi.org/10.1016/j.csi.2023.103808

[29] D. Jayaprakash and C. S. Kanimozhiselvi, "MULTINOMIAL LOGISTIC REGRESSION METHOD FOR EARLY DETECTION OF AUTISM SPECTRUM DISORDERS," *Meas. Sensors*, vol. 33, p. 101125, Jun. 2024. doi: https://doi.org/10.1016/j.measen.2024.101125

[30] A. Setiawan, F. Setivani, and T. Mahatma, "PERFORMANCE COMPARISON OF DECISION TREE AND LOGISTIC REGRESSION METHODS FOR CLASSIFICATION OF SNP GENETIC DATA," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 18, no. 1, pp. 0403–0412, Mar. 2024. doi: https://doi.org/10.30598/barekengvol18iss1pp0403-0412

[31] M. Howell-Moroney, "INCONVENIENT TRUTHS ABOUT LOGISTIC REGRESSION AND THE REMEDY OF MARGINAL EFFECTS," *Public Adm. Rev.*, vol. 84, no. 6, pp. 1218–1236, Nov. 2024. doi: https://doi.org/10.1111/puar.13786

[32] A. Purwanto, M. A. Suprayogi, E. Setiawan, J. F. R. B. Loly, G. A. Rahman, and A. Kurnia, "MULTINOMIAL LOGISTIC REGRESSION MODEL USING MAXIMUM LIKELIHOOD APPROACH AND BAYES METHOD ON INDONESIA'S ECONOMIC GROWTH PRE TO POST COVID-19 PANDEMIC," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 19, no. 1, pp. 51–62, Jan. 2025. doi: https://doi.org/10.30598/barekengvol19iss1pp51-62

[33] Z. Rahmatinejad *et al.*, "A COMPARATIVE STUDY OF EXPLAINABLE ENSEMBLE LEARNING AND LOGISTIC REGRESSION FOR PREDICTING IN-HOSPITAL MORTALITY IN THE EMERGENCY DEPARTMENT," *Sci. Rep.*, vol. 14, no. 1, p. 3406, Feb. 2024. doi: https://doi.org/10.1038/s41598-024-54038-4

[34] L. H. Y. Arini, S. Solimun, A. Efendi, and A. A. R. Fernandes, "ENSEMBLE BAGGING WITH ORDINAL LOGISTIC REGRESSION TO CLASSIFY TODDLER NUTRITIONAL STATUS," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 19, no. 1, pp. 1–12, Jan. 2025. doi: https://doi.org/10.30598/barekengvol19iss1pp1-12

[35] P. D. F. Isles, "A RANDOM FOREST APPROACH TO IMPROVE ESTIMATES OF TRIBUTARY NUTRIENT

LOADING," *Water Res.*, vol. 248, p. 120876, Jan. 2024. doi: https://doi.org/10.1016/j.watres.2023.120876

[36] S. Qaderi, A. Maghsoudi, M. Yousefi, and A. B. Pour, "TRANSLATION OF MINERAL SYSTEM COMPONENTS INTO TIME STEP-BASED ORE-FORMING EVENTS AND EVIDENCE MAPS FOR MINERAL EXPLORATION: INTELLIGENT MINERAL PROSPECTIVITY MAPPING THROUGH ADAPTATION OF RECURRENT NEURAL NETWORKS AND RANDOM FOREST ALGORITHM," *Ore Geol. Rev.*, vol. 179, p. 106537, Apr. 2025. doi: https://doi.org/10.1016/j.oregeorev.2025.106537

[37] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "AN IMPROVED RANDOM FOREST BASED ON THE CLASSIFICATION ACCURACY AND CORRELATION MEASUREMENT OF DECISION TREES," *Expert Syst. Appl.*, vol. 237, p. 121549, Mar. 2024. doi: https://doi.org/10.1016/j.eswa.2023.121549

[38] X. Zhang *et al.*, "IMPROVED RANDOM FOREST ALGORITHMS FOR INCREASING THE ACCURACY OF FOREST ABOVEGROUND BIOMASS ESTIMATION USING SENTINEL-2 IMAGERY," *Ecol. Indic.*, vol. 159, p. 111752, Feb. 2024. doi: https://doi.org/10.1016/j.ecolind.2024.111752

[39] H. Wang, F. Wang, H. Yang, K. Staszewska, and B. Jeremić, "SOBOL' SENSITIVITY ANALYSIS OF A 1D STOCHASTIC ELASTO-PLASTIC SEISMIC WAVE PROPAGATION," *Soil Dyn. Earthq. Eng.*, vol. 191, p. 109283, Apr. 2025. doi: https://doi.org/10.1016/j.soildyn.2025.109283

[40] L. Lusardi *et al.*, "METHODS FOR COMPARING THEORETICAL MODELS PARAMETERIZED WITH FIELD DATA USING BIOLOGICAL CRITERIA AND SOBOL ANALYSIS," *Ecol. Modell.*, vol. 493, p. 110728, Jul. 2024. doi: https://doi.org/10.1016/j.ecolmodel.2024.110728

[41] Z. Wang *et al.*, "METABOLISM-BASED NONTARGET-SITE MECHANISM IS THE MAIN CAUSE OF A FOUR-WAY RESISTANCE IN SHORTAWN FOXTAIL (ALOPECURUS AEQUALIS SOBOL.)," *J. Agric. Food Chem.*, vol. 72, no. 21, pp. 12014–12028, May 2024. doi: https://doi.org/10.1021/acs.jafc.4c01849

[42] F. Anderl and M. Mayle, "SENSITIVITY ANALYSIS OF PIEZOELECTRIC MATERIAL PARAMETERS USING SOBOL INDICES," *tm - Tech. Mess.*, Apr. 2025. doi: https://doi.org/10.1515/teme-2024-0116

[43] L. Chen, Z. Xu, D. Huang, and Z. Chen, "AN IMPROVED SOBOL SENSITIVITY ANALYSIS METHOD," *J. Phys. Conf. Ser.*, vol. 2747, no. 1, p. 012025, May 2024. doi: https://doi.org/10.1088/1742-6596/2747/1/012025