Global sensitivity analysis using Monte Carlo estimation under fat-tailed distributions

Sin, Gürkan

Published in:
Chemical Engineering Science

Link to article, DOI:
10.1016/j.ces.2024.120124

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Chemical Engineering Science 294 (2024) 120124

Contents lists available at ScienceDirect

Chemical Engineering Science

journal homepage: www.elsevier.com/locate/ces

Global sensitivity analysis using Monte Carlo estimation under fat-tailed distributions

Gürkan Sin

PROSYS, Department of Chemical and Biochemical Engineering, Technical University of Denmark, B228A Soltofts Plads, 2800 Lyngby DK, Denmark

ARTICLE INFO

Keywords:
Global sensitivity analysis
Monte Carlo simulation
Pareto distribution
Convergence
Derivative-based sensitivity
Greenhouse gas emission
Process engineering

ABSTRACT

Global sensitivity analysis found a widespread use in the modeling community to study the input and output relationship of typically complex numerical models. Many sensitivity analysis studies have been published over the years across different domains, from considering simple test problems to more complex case studies involving large numerical models. However, very few studies have addressed the issue of the presence of fat-tailed distributions and its implication for the sensitivity analysis. First, we recall how the law of large numbers slowly convergences depending on the extent of tails in the distributions. Then, we present some methods to study Paretianity in the data and estimate the tail index. We then apply these concepts to a real- world global sensitivity problem using a case study of long- term measurements of N_2O emissions dataset from WWTPs. We then propose a robust sensitivity metric based on mean absolute deviation for parameter importance ranking under fat- tailed distributions.

1. Introduction

Consider a model function $f(\boldsymbol{x})$, with an input vector, $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ that specify an n dimensional input space H^n , which is used to predict a model output, $y = f(\boldsymbol{x})$. Global sensitivity analysis (GSA) is concerned with the question of how to decompose model output variation with respect to model inputs. This is closely related to uncertainty analysis, which aims to quantify model output uncertainty given uncertainty in the inputs of a model. Here the model inputs refer to any parameters, constants, assumptions, or modeling resolution used for constructing the model in question.

An important distinction should be made here between the local versus global sensitivity analysis. Indeed the local sensitivity analysis is defined typically as a partial derivative of a function, $f(x)$ with respect to a particular input, x, $\frac{\partial f(x)}{\partial x}$. In global sensitivity analysis emphasis is made to analyzing the model outputs response to the entire input space defined by H^n, as opposed to one particular point,x^0 , in the input space, which is used in local sensitivity analysis $\left(\frac{\partial f(x)}{\partial x}|x^0\right)$.

This class of global sensitivity analysis techniques, known as variance-based methods, is closely associated with applied statistics and employs typically a Monte Carlo approach for analysis of the entire input space. A comprehensive study of the topic is given in the seminal book of (Saltelli et al., 2007). In recent decades, GSA has received widespread applications among diverse academic fields including engineering, economics, operation research, and environmental science among others as reviewed comprehensively (Saltelli et al., 2019; Borgonovo and Plischke, 2016). It has also become part of international guidelines when using models for better policy making such as (European Commission, 2023; USA EPA, 2023). In process systems engineering, GSA is also widely employed to generate insights into the contributions of individual model inputs, or sub-groups of inputs, to the variations in the output of numerical models. Such insights are used for model calibration and quality assessment to design space and robustness analysis (Kucherenko et al., 2020; Al et al., 2019; Sin et al., 2009).

Over the decades, many sensitivity analysis techniques, from derivative-based (Morris screening, derivative-based global sensitivity analysis) to variance-based methods (such as Sobol's method), moment-dependent methods etc, have been developed together with some software tools catering to different needs in many application domains (Razavi et al., 2021).

Many of these techniques focused on input distributions typically

Abbreviations: Notation, definition; MS plot, Maximum-to-Sum plot; GSA, Global sensitivity analysis; DNN, Deep neural network; LLN, Law of large numbers; CLT, Central limit theorem; QQplot, Quantile-quantile plot; N_2O, Nitrous oxide; NH_4, Ammonium; NO_3, Nitrate; DO, Dissolved oxygen; Q_{air}, Air flowrate; Q_{inf}, Influent flowrate; T, Temperature.

E-mail address: gsi@kt.dtu.dk.

https://doi.org/10.1016/j.ces.2024.120124
Received 18 October 2023; Received in revised form 14 March 2024; Accepted 12 April 2024
Available online 14 April 2024

characterized by uniform or Gaussian distribution, which belongs to the thin-tailed distribution class. Convergence of the central limit theorem under such distributions is known to be fast (Taleb, 2020). This is essential to ensure that the Monte Carlo estimation of global sensitivity measures is reliable. However, surprisingly no studies have addressed the issue of fat tails in the distribution of model outputs and its implications for the convergence of Monte Carlo estimation of the sensitivity measures. Fat tails belong to the power law class of distributions, such as Pareto distribution, typically used to model extreme events in hydrology (rainfall intensity/flooding), insurance and finance among others. This class of distributions is known to have significant implications for the quality of the analysis and decision-making among others due to the slow convergence of the central limit theorem (Taleb, 2020; Embrechts et al., 2013).

In this contribution, we first recall the basic theory illustrating the convergence of the central limit theorem for increasingly fat tails and graphical methods to check the presence of fat tails in observations (Cirillo, 2013). Next, we will test and evaluate these methods on an industrially-relevant case study focusing on greenhouse gas emissions (N_2O) from a municipal wastewater treatment plant. The question for GSA in this case study is to study how the measured process input parameters in the plant explain the variation in the greenhouse gas. We will briefly recall the dataset and the model developed in our previous study (Hwangbo et al., 2021), which will be used as a testbed.

Next, we will introduce the derivative-based global sensitivity analysis (DGSM) technique, which uses Monte Carlo sampling to estimate sensitivity measures. This will be followed by presenting the results and discussion before concluding with a summary of points and key findings of the study.

2. Theory

The Monte Carlo technique, whether used for solving global sensitivity analysis problems or uncertainty analysis, essentially requires an understanding of the limiting behavior of a sum of random variables as a function of sampling number, n. Both law of large numbers (LLN) and central limit theorems (CLT) are concerned with addressing the limiting behavior of the sum of independent random variables, X_n with arbitrary distribution: $S_n = X_1 + X_2 \cdots + X_n$. It is important to note that the sum, S_n, is a random variable and will fluctuate in every repetition of this experiment. Hence, it is necessary to study its convergence.

The law of large numbers states the following (Feller, 1991): Let $\{X_1, X_2 \cdots X_n\}$ be a sequence of mutually independent random variables with a common distribution. If the expectation $\mu = \boldsymbol{E}(X_n)$ exists, then for every $\varepsilon > 0$ as $n \to \infty$,

$$P(|\overline{X_n} - \mu| > \varepsilon) \to 0 \tag{1}$$

In other words, the probability that the sample mean $\overline{X_n}$ will differ from the true mean,μ, of the distribution of X_n by more than an arbitrarily small number,ε tends to 0. Therefore the sample mean converges to true mean in probability: $\overline{X_n} \xrightarrow{P} \mu for n \to \infty$.

On the other hand, the central limit theorem (CLT) addresses the limiting behavior of the sum, S_n, when normalized by the $\sqrt{n}$ and centered around the mean μ (Embrechts et al., 2013). The standard version of CLT is recalled as follows: Consider a sequence of independent and identically distributed random variables $\{X_1, X_2 \cdots\}$ with a mean $= \boldsymbol{E}(X_n)$, variance $\sigma^2 = \boldsymbol{V}(X_n)$ and a sample mean, $\overline{X_n} = n^{-1} S_n$. Then the sum of centered and normalized random variables, $\sqrt{n}(\overline{X_n} - \mu)$ converges in distribution to Gaussian, as n approaches infinity (Taleb, 2020):

$$\sqrt{n}(\overline{X_n} - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{2}$$

$$vi = \int_{Hn} \left(\frac{\partial f}{\partial xi} \right)^2 dx \, for \, i = 1, \cdots n \tag{3}$$

Table 1 summarizes an experiment that analyze the behavior of the summation of random variables drawn from different distribution functions.

In particular, the sum of random variables drawn from a uniform distribution converges rapidly to the Gaussian distribution – it takes only three summands. The convergence of the sum of the exponentially distributed random variables to the Gaussian is slower, which takes about six summands. Interestingly, the sum, the convergence of independent random variables drawn from a Pareto distribution converges much slower than the Gaussian. The tail in the x-axis is still present in the distribution of the sum, S_n (see the figure at the bottom row of Table 1). These simulated observations agree with the theoretical derivation of the probability density function of the sums shown in (Taleb, 2020). In power-law distributions, most of the information is in the tails despite low probability of occurrence. Therefore, to accurately represent such distributions, a large number of samples must be taken, which is not always feasible. Thus, verifying the Paretianity in data and performing diagnostic convergence tests when using Monte Carlo or sampling-based analysis in general is important.

3. Material and methods

3.1. Derivative-based global sensitivity analysis (DGSM)

The derivative-based global analysis (DGSM) method proposed by (Sobol and Kucherenko, 2009) expands the Morris method of elementary effects, EEi. It presents an alternative function to estimate the importance of model inputs. The method is defined as follows: Consider again a model function $f(\boldsymbol{x})$, with an input vector, $\boldsymbol{x} = \{x_1, \cdots, x_i \cdots x_n\}$ that specify an n dimensional input space H^n, which is used to compute a model output, $y = f(\boldsymbol{x})$. If the partial derivative, $\frac{\partial f}{\partial x}$ exists, that is $f(\boldsymbol{x})$ is differentiable, then the following sensitivity measure,vi, is proposed:

$$vi = \int_{Hn} \left(\frac{\partial f}{\partial xi} \right)^2 dx \, for \, i = 1, \cdots n \tag{4}$$

Here the sensitivity measure,vi, corresponds to the absolute mean of the partial derivatives. This measure is used to rank the relative importance of the model inputs: a small value will indicate less influential inputs, and vice versa, a high value will indicate highly significant inputs.

The intuition behind this derivative-based approach is as follows: instead of evaluating the partial derivative of a function at one single point in the input space $\left(\frac{\partial f(x)}{\partial x} | x^0 \right)$ as typically done in local sensitivity analysis, it is suggested to evaluate the partial derivative at many randomly selected points in the input space of the model. This procedure results in randomly drawn samples from the distribution of partial derivatives. This aligns with the global analysis mindset as the input space is covered more comprehensively. The property of this distribution is then analyzed to reveal a global sensitivity measure.

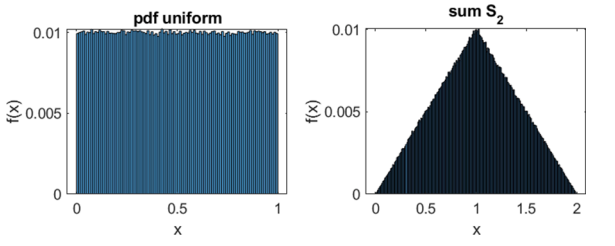
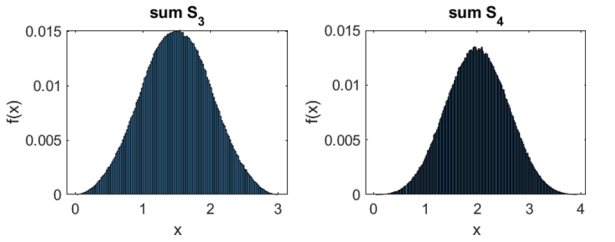
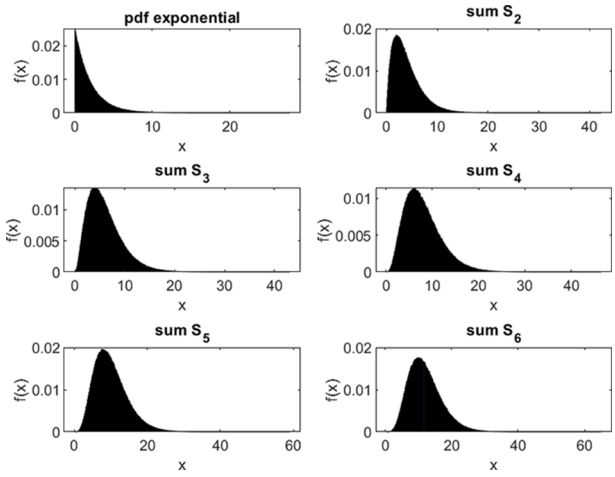
It is noted that extension of global sensitivity analysis methods to account for dependent inputs is presented elsewhere (Lamboni and Kucherenko, 2021; Mara et al., 2015). This is beyond the scope of this study.

In this contribution, we focus on the convergence of the moments of the distribution of the partial derivatives obtained via Monte Carlo sampling, which we denote as Fi, to check for the consistency of the results.

3.2. MS plots

Maximum-to-Sum (MS) plot is a graphical method used to analyze the contribution of maximum values in a sample to a given moment. This plot is useful to graphically see how the LLN converges (or not) as the number of samples, n, grows for different moments. The MS plot makes

Table 1
Convergence of central limit theorem for different distributions with increasing tail index: uniform, exponential and Pareto.

Distributions	
Uniform $P(X > x) f(x) = \frac{1}{b-a}$ $for a < x < b$ For the experiment, $a = 0 and b = 1$	
Exponential $f(x) = \lambda e^{-\lambda x} for x > 0$ For the experiment, $\lambda = 2$	
Pareto $f(x) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}} for 0 < x_0 \le x$ For the experiment, $\alpha = 3 and x_0 = 2$	

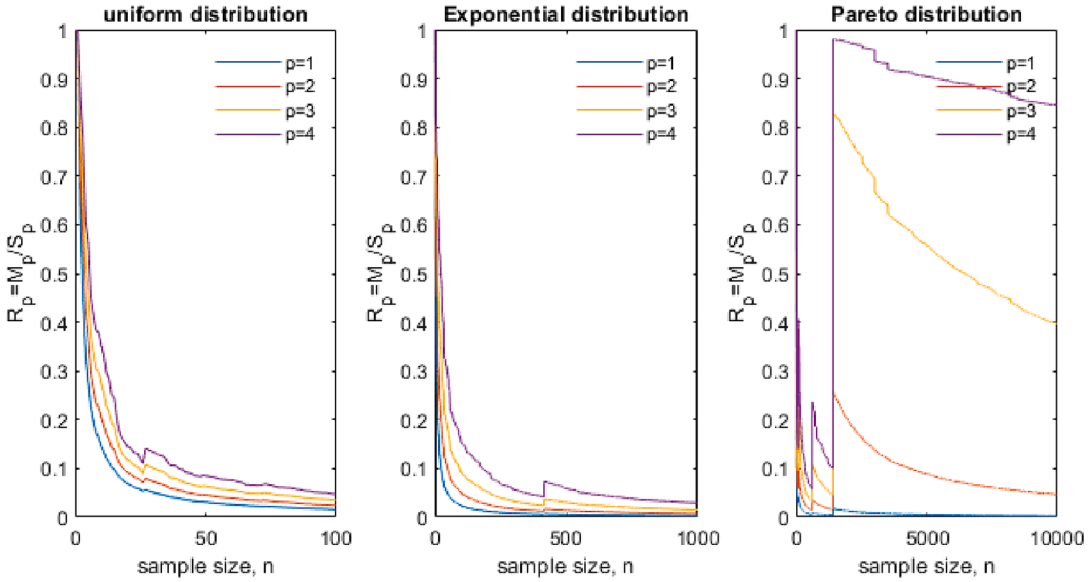


Fig. 1. The MS plot showing behavior of LLN for convergence of different moments: uniform (left), exponential (middle) and Pareto (right) distributions.

use of LLN and is defined as follows (Taleb, 2020):

Consider again a sequence of random variables, $\{X_1, X_2 \cdots X_n\}$ which are non-negative and i.i.d random variables. If finite moments exist for p = 1,2,.. with $(X^p) < \infty$, then the ratio of the maximum in a sample to the sum of the random variables in the sample will as surely converge:

$$R_n^p = \frac{M_n^p}{S_n^p} \xrightarrow{a.s.} 0 \, for \, n \to \infty \tag{5}$$

Here $S_n^p = \sum_{i=1}^n X_i^p$ is the partial sum and $M_n^p = \max(X_1^p \cdots X_n^p)$ the partial maximum for a given p moment. This ratio is useful to determine how much a single observation from the tail of a fat-tailed distribution can contribute to the total information. Therefore, it helps identify the presence or absence of Paretianity in the data. Fig. 1 plots the MS ratio for the samples presented in Table 1 and shows the Law of Large Numbers in action. We can observe rapid convergence of the first moment of the distribution (which is the mean of the distribution) for uniform random variables, slower convergence for exponential random variables, and much slower convergence for Pareto random variables. Moreover, for Pareto random variables, the convergence rate of the higher moments is even slower, requiring significantly more samples, n,. This phenomenon is known as the slow convergence rate of LLN under fat-tailed distributions (Taleb, 2020).

In addition to the MS plot, other graphical methods, such as the zipf plot, mean-excess threshold, and peak-over-threshold (POT) plots can also be used to infer Paretianity in the data (Cirillo, 2013).

3.3. The Hill estimator for the shape parameter

The Hill method is one of the several methods to estimate shape parameter,$\xi = \frac{1}{\alpha}$. The tail exponent, α indicates the heaviness of tails in observations (Embrechts et al., 2013). In particular, the lower the α, the heavier the tail. The method works by using order statistics and log transformation of the data as follows: Consider a sequence of random variables $\{X_1, \cdots X_n\}$ with a probability distribution function F as follows:

$$P(X > x) = \overline{F}(x) = Cx^{-\alpha} \, for \, x \geq u > 0 \tag{6}$$

Assume threshold u is known such that $C = u^{\alpha}$ is fully defined, in that case the maximum likelihood estimate of the tail exponent, $\widehat{\alpha}$, is obtained as follows (Embrechts et al., 2013):

$$\widehat{\alpha_n} = \left(\frac{1}{n} \sum_j^n lnX_{j,n} - lnu \right)^{-1} \tag{7}$$

In practice, as the knowledge of threshold u is unknown, the estimator, $\widehat{\alpha}$ will be obtained for a range of u and analyzed graphically. Here K order statistics is the event defined as follows (Embrechts et al., 2013):

$$K = card\{j : X_{j,n} > u, j = 1 \cdots n\} \tag{8}$$

The Hill plot, which is the plot of log of [j] the order statistics versus log of [Xj] is used to visually inspect and generate proposal for threshold values u.

3.4. Case study: Deep learning model of greenhouse gas emissions and input data

The model structure considered here has the following form: let **y** be a vector of model outputs of interest, **X** is a matrix of inputs, and f is a model that relates the inputs to outputs, $y = f(\mathbf{X})$. The case study analyzed here refers to the study of the off-gas N_2O emissions from the Avedøre wastewater treatment plant presented in (Hwangbo et al., 2021). In this case study, the model output, y, is the time series measurements off-gas nitrous oxide (N_2O) emissions from the plant. The input dataset, **X**, consist of time series measurements of six process variables namely influent flow rate (Q_{inf}), temperature (T), airflow rate (Q_{air}) and liquid-phase measurement of ammonium (NH_4), dissolved oxygen (DO), nitrate (NO_3) respectively from the period March to June 2018.

The model structure, f that relates the inputs to the output is a deep neural network (DNN) model, which is trained against the measured data using mean squared error as the objective function. The DNN model consists of 6 hidden layers with 64, 32, 16, 8 and 4 hidden neurons, respectively. Each hidden neuron uses the hyperbolic tangent sigmoid function. The methodology used for hyper-parameter tuning of the model is described in detail elsewhere (Hwangbo et al., 2021). The particular DNN model and the dataset used for training the model are provided on DTU Data repository (Sin, 2023). A time series plot of the dataset is provided in the Supplementary Material, while in Fig. 2, the summary of the model predictions versus measurements for model output is given.

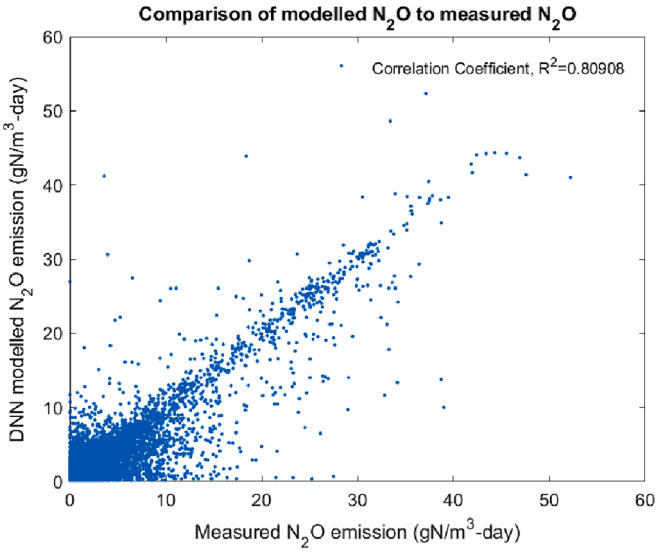


Fig. 2. The DNN model used in this study for sensitivity analysis: Comparison of the predictions versus measured emissions data.

4. Results

4.1. Monte Carlo simulations for derivative-based global sensitivity analysis

To obtain a sufficiently high number of samples from the distribution of the derivative-based sensitivity analysis, *30,000* Monte Carlo evaluations of the derivative function in Eq.3. To this end, the input domain is sampled using the Latin Hypercube Sampling (LHS) technique, which is shown in Fig. 3 as a matrix plot. The histogram of the inputs is shown along the diagonal, while the pairwise plot of the inputs is shown along off-diagonal elements of the matrix plot. One thing toobserve is that. All the inputs have visible tails in their histogram (frequency versus random variate plots) except for temperature.

The distribution of the derivative values calculated at each sample point is shown in Fig. 4. For the inputs shown, while the majority of the samples have values concentrated around zero, one can see the presence of large maximum values of the derivative both at the left and right-hand side tails of the histogram. There is now an important question related to these sensitivity results: given the presence of tails in the distribution, is the Monte Carlo analysis converged and consistent, or is this a fluctuation resulting from a chance-based sampling that is used in the analysis? This is addressed in the next section.

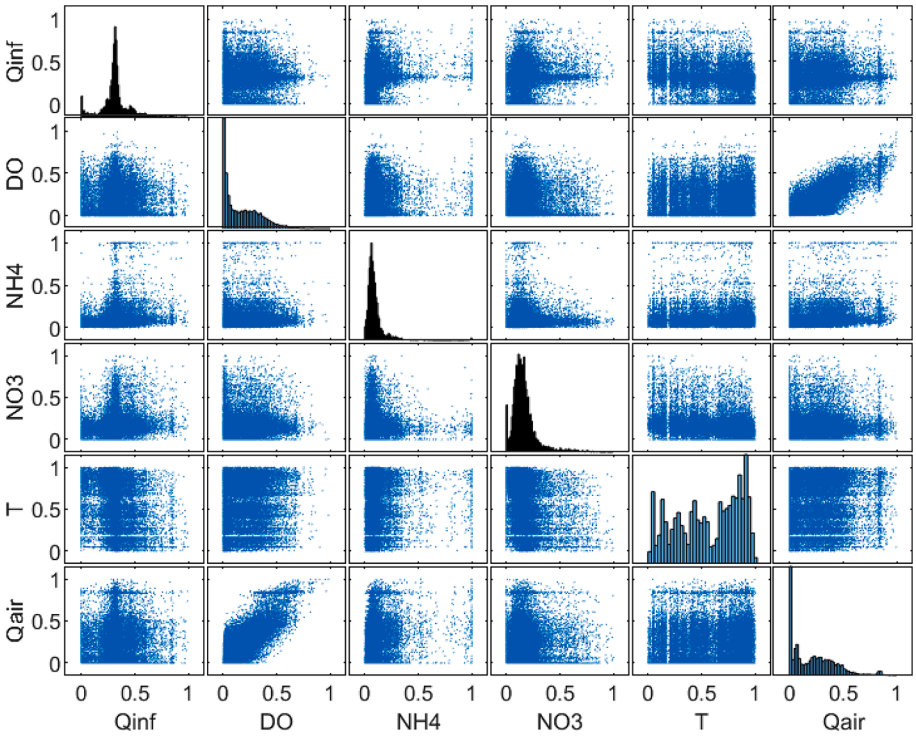


Fig. 3. Sampling from the normalized input domain using Latin Hypercube Sampling with a Gaussian copula for DO and Qair to preserve the correlation between them (the correlation matrix between inputs is shown in the Supplementary material).

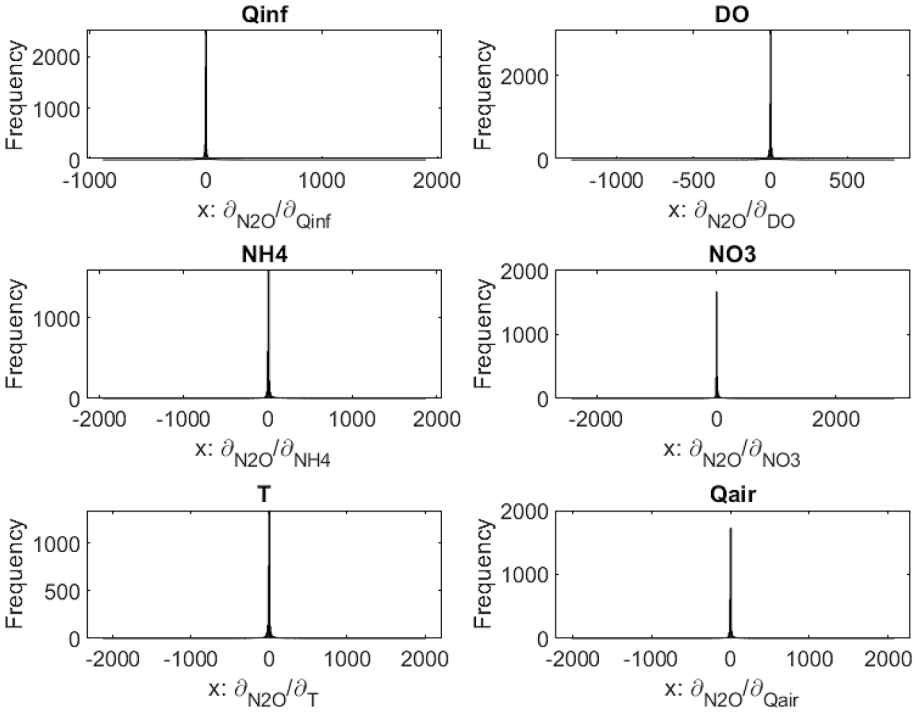


Fig. 4. Distribution of the partial derivate of inputs with respect to model output, namely the nitrous oxide emissions from the plant. Notice the scale of the x-axis indicating the range of values of the partial derivative.

4.2. Estimation of the tail of the distribution

To better understand the shape of the distribution, the tail index on the left and right-hand sides of the distributions are estimated. The main purpose here is to check for the presence of Paretianity in the data, which would indicate a convergence issue and quality of the Monte Carlo solutions, as discussed in the background above. First, we look at the QQplot, which compares the sample quantiles to the theoretical quantiles of a known distribution. This plot helps visualize the fit of data given on the y-axis to a specific distribution given on the x-axis.. In Fig. 5 we compare the quantiles of the data against the normal distribution and we see disagreement especially in the tails. Hence the data does not

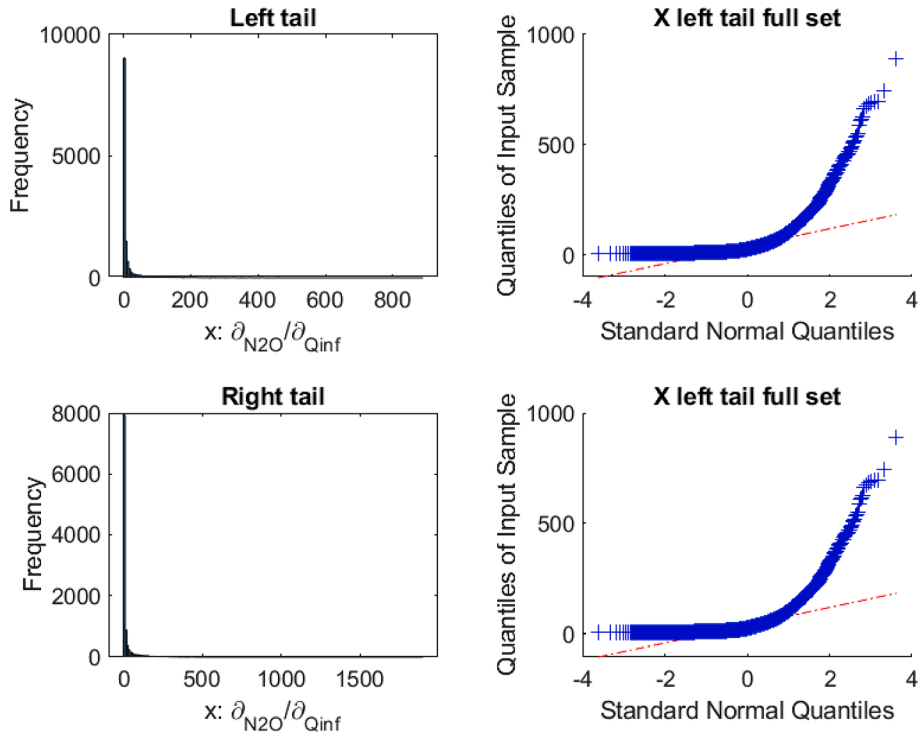


Fig. 5. The left and right tails of the distribution of partial derivative of influent flowrate (Qinf) on Nitrous Oxide (N_2O) emissions are shown. The QQplot indicates the tails of the distributions are diverging significantly from normal distribution.

follow the quantiles that that would be expected if the data were normally distributed.

Following the visual inspection, we will next look for part of the region in the tails of the data that may be explained by a power law type distribution such as Pareto. To estimate the tail exponent,α, of the power law type distribution, we use the Hill estimator and generate a plot of estimates for different thresholds u following k order statistics. The results are shown in Fig. 6. Usually, this plot is used to find the first stable region and infer proper order statistics. For the left side of the distribution, the estimation of the tail is based on k = 151 (u = 228.8) and α = 2.44 (with ± 0.4 95 % confidence interval). This means that the data in the tail larger than 228.8 follows a power law type Pareto distribution. For the right side of the tail, the estimation of the tail exponent is based on k = 151 (u = 305.4) and α = 2.29 (±0.37 95 % confidence interval).

For thin tailed distributions, the rare events does not change the calculation of the moments of the distribution (e.g. mean or variance) hence the convergence of the analysis is fast. However, for fat-tailed distributions, a single sample from the tail can alter the calculations of the moment drastically. In particular, the higher order moments converges very slowly and in fact does not exist for $\alpha < 2$ ((Embrechts et al., 2013). Based on these estimates, we can conclude that the lower bound of the 95 % confidence level is very close to 2, which indicates potential convergence issues for estimating the higher order moments of this distribution except the first order moment.

The estimation of the tail exponents for other partial derivatives for all the inputs are summarized in Table 2. The corresponding Hill estimator plots are provided in the Supplementary Material. It is noted that the tail exponents for the temperature is the highest among others (with

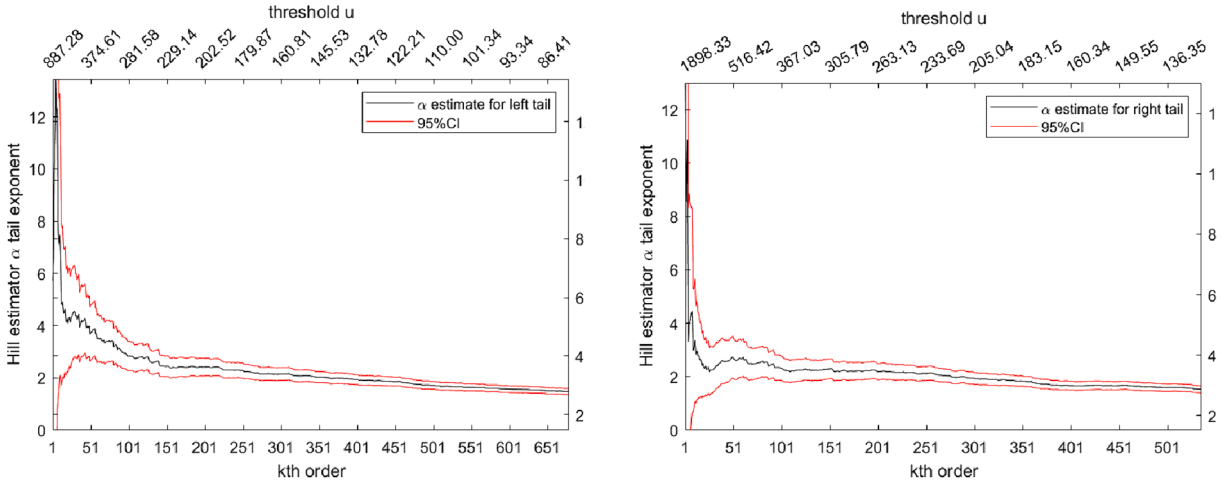


Fig. 6. The Hill estimator for the tail exponent α for the left and right tails of the distribution of the partial derivative of influent flow (Qinf) with respect to Nitrous Oxide (shown in Fig. 5). The results are shown together with their 95% confidence interval. The kth order statistics of the ranked observation of the partial derivatives shown in bottom x-axis and the corresponding threshold value, u is shown in the top x-axis.

Table 2
Estimation of the tail exponents for the left and right tails of the distributions.

Partial derivative	Left tail (k, u & α ± 95 %CI)	Right tail (k, u & α ± 95 %CI)
Q_{inf}	k = 151(u = 228.8) & α = 2.44 ± 0.4	k = 151 (u = 305.4) & α = 2.29 ± 0.37
DO	k = 101(u = 221.8), & α = 3.15 ± 0.63	k = 101 (u = 240.5) & α = 2.65 ± 0.52
NH_4	k = 76(u = 443.1), & α = 2.65 ± 0.61	k = 76 (u = 742.2), & α = 3.43 ± 0.79
NO_3	k = 76(u = 464.9), & α = 2.74 ± 0.63	k = 71(u = 799.2), & α = 2.57 ± 0.61
T	k = 25(u = 1053.5), & α = 4.28 ± 1.78	k = 25(u = 1057.8), & α = 3.40 ± 1.42
Q_{air}	k = 76(u = 435.9), & α = 2.61 ± 0.6	k = 26(u = 831.2), & α = 3.0 ± 1.23

lowest kth order statistics). This means that the distribution of temperature effect on nitrous oxide emissions does not follow a power law type behavior. On the other hand, for many of the inputs we observe a tail exponent estimate that lies close to 2 indicating presence of Paretianity in the data. This means that convergence test for the given Monte Carlo sampling number is mandatory to understand which, *if any*, inference based on the Monte Carlo analysis is reliable.

4.3. Convergence of Monte Carlo solutions

The convergence of Monte Carlo simulations are checked using the MS plot. In Fig. 7, we plot the results of convergence for the right tail of the distributions given for the first, second, third, and fourth moments, respectively. This plot shows that as the sample size goes to infinity, the MS ratio tends to converge to zero. For a given finite sampling size, *n*, the MS plots will indicate if there is asymptotic convergence or not (in case there are jumps, for example, this would indicate non-convergence).

Overall, the results are consistent with the range of tail exponent estimated in the data, where we observe that for distributions with a lower tail exponent, the convergence of the higher-order moments is not observed. While the first moment of the distribution converges for all the inputs, this cannot be observed for higher-order moments, especially 3 and 4. Regarding the influent flowrate in Fig. 7, we clearly observe that the first moment is converged, therefore the mean estimate of this sample is consistent with the true mean of the distribution.

However, there are sudden jumps seen in this ratio, indicating that one large observation in a given sample will affect this ratio. This is consistent with sampling from fat-tailed distributions (see Fig. 1). Convergence of this ratio for all the moments clearly will require a much higher number of samples.

In case the sampling size is limited and additional sampling (i.e. performing more Monte Carlo simulations) is costly, then one is advised to use only consistently estimated moments of the distribution. In this case, we can reliably use the mean estimate of the sample for all of the inputs, which have converged consistently. In some cases, we see second moments converged, but certainly for not all inputs especially the case for Q_{inf}, NO_3, Q_{air} and NH_4. Therefore, to make a reliable and consistent inference from this global sensitivity analysis for all the inputs, one needs to use the first moment of the distribution as a sensitivity measure.

The analysis was crosschecked using Sobol sequence as a low discrepancy sampling method, which is shown elsewhere to be efficient (Kucherenko et al., 2015). The MS plots obtained using Sobol sequence, shown in the Supplementary Information, confirm similar convergence issues for estimating the higher-order moments under fat-tailed distributions. This demonstrates that the convergence is not related to the sampling technique but more to the shape of the distribution.

4.4. A consistent metric to summarize global sensitivity analysis

As the first moment, which is the mean, μ of the distribution, is consistently converged for all the inputs in our analysis, we propose to use the mean as well as the mean absolute deviation (MAD) as metrics to make inferences from the distributions. The MAD makes use of the first moment as follows: $MAD = \frac{1}{n}\sum_i^n |(x_i - \mu)|$ where to describe the dispersion of the data around the mean.

Table 3 summarizes the results of the global sensitivity analysis for

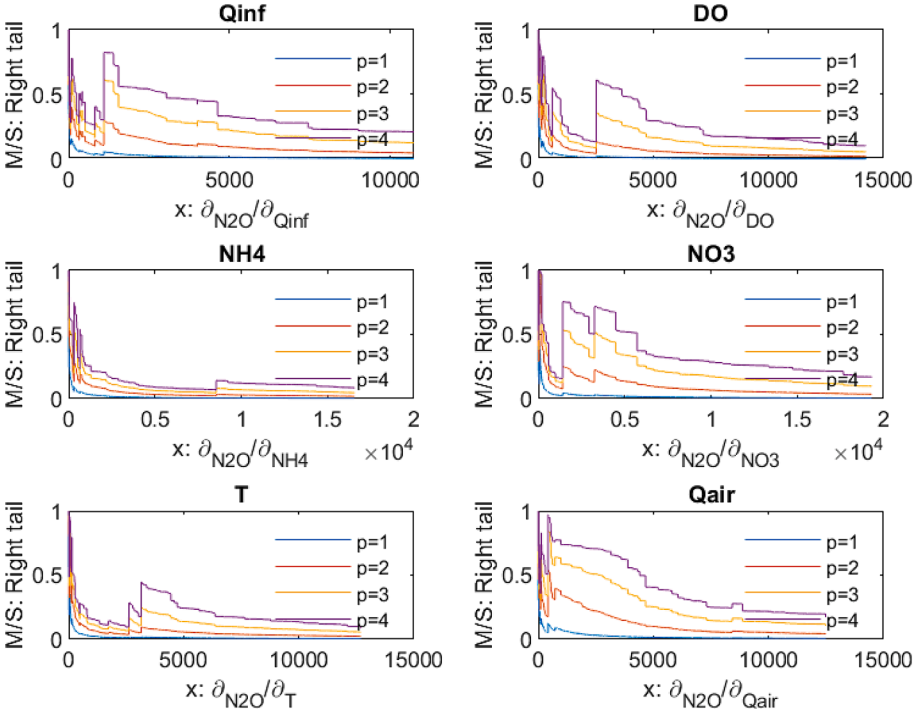


Fig. 7. The MS plots for the distribution of the partial derivatives with respect to the inputs on the model output nitrous oxide. The results are shown for the right side of the distribution and for convergence of the first fourth moments.

Table 3
Summary of derivative-based global sensitivity analysis of N_2O emissions: three sensitivity measures for importance ranking are shown.

OUTPUT→	N_2O Emissions		
InputS	Mean of the left tail	Mean of the right tail	Mean Absolute Deviation (MAD)
Q_{inf}	−16.40	24.09	16.90
DO	−15.95	14.37	12.80
NH_4	−26.37	*34.03*	**30.94**
NO_3	*−42.16*	*33.60*	**33.18**
T	*−37.71*	*36.38*	**30.74**
Q_{air}	−29.4	26.93	24.74

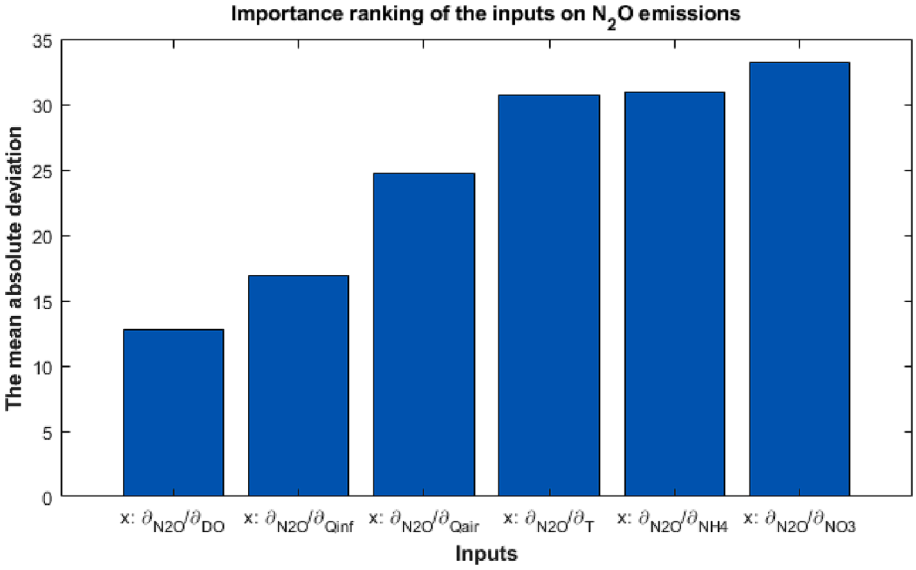


Fig. 8. Importance ranking of the inputs on the N_2O emissions based on the mean absolute deviation as a sensitivity measure.

the N_2O emissions using these three sensitivity measures: the mean of the left side, the mean of the right side, and the mean absolute deviation of the entire distribution.

The importance ranking of the inputs on the nitrous oxide emissions is shown in Fig. 8. As regards the sign of the impact of the inputs, we compare the mean of the left (which contains negative derivatives in the sample) and the mean of the right side (which contains the positive derivatives in the sample) shown in Table 3. A significance test at a 1 % confidence level (*t*-test) is performed, which indicates that the means of the two sides are not significantly different. This shows that none of the input factors analyzed here drive the variation in the N_2O emissions in one direction alone. This is a typical feature of nonlinear systems, in which an input can have both a positive and negative contribution to the output depending on the initial state of the system.

The mean absolute deviation of the distribution of the partial derivatives as a metric for the importance ranking (Fig. 8). The ranking showed that NO_3, NH_4 and T are the three most important factors driving the randomness in the nitrous oxide emissions. This finding is consistent with the process engineering knowledge. As discussed by Chen et al (2019), the nitrous oxide emissions result from a complex interplay of microbiological activity, the plant disturbances and the operation dynamics (Chen et al., 2019).

For example, the impact of DO and Q_{air} is related to the closed-loop control of the plant. The plant uses a cascade control system in which the duration of the alternating aerated and unaerated (anoxic) periods and the setpoint for DO during the aerated period are regulated. The controller manipulates the Q_{air} to ensure a certain removal efficiency for NH_4 and NO_3 (see (Chen et al., 2019). for a detailed discussion of the plant operation and its impact on the emissions).

5. Discussion

The primary purpose of global sensitivity analysis is to identify inputs that contribute to the variability in output behavior. The analysis is then used either for factor prioritization based on their importance ranking or factor fixing in case their sensitivity measure is negligibly low) (Saltelli et al., 2007).

In the given case study, the focus is to identify the inputs that are causing the randomness in nitrous oxide emissions from the plant. Some of these inputs are disturbances, such as influent flowrate (Q_{inf}), Temperature (T) and ammonium (NH_4), while others are process parameters controlled by the plant operation system, such as dissolved oxygen (DO), air flowrate (Q_{air}) and nitrate (NO_3). By understanding which inputs are the main contributors to the emissions, the existing control/operation system of the plant can be revised to reduce them. More details on this topic are discussed elsewhere detail (Chen et al., 2019)).

We used the derivative-based global sensitivity analysis method to answer the question concerning which inputs are important. The solution of this method requires the use of Monte Carlo analysis of the input domain. The solution method provides a sampling from the distribution of the partial derivatives of the inputs. We addressed the convergence issue of our analysis due to the pronounced tail phenomena in the distribution using extreme value statistics methods. The QQplot showed a clear deviation from normal distribution, especially for the tails. We used the Hill estimator to estimate the tail exponent and found that it ranges from 2.3 to 4.2, indicating that the distribution of partial derivatives is fat-tailed (Embrechts et al., 2013; Taleb, 2020). This has consequences on which tools should be used to properly interpret the results, as shown in the theory section.

We checked the convergence of the Monte Carlo solutions given a finite sampling size, *n* using the MS plot. The MS plot provided a visual

confirmation and showed that the higher-order moments of the distribution is not converged. Therefore, the estimate of the kurtosis or skewness of the distribution of the partial derivatives is inconsistent and cannot be used. A higher sample size may not help, especially since the tail exponent is below 2 (which was our case). However, we can use mean absolute deviation (MAD) as a sensitivity measure, especially since the MS plot showed that the first moment of the distribution is converged. This metric can be consistently used for analysis of the partial derivative-based sensitivity method.

As regards the results, one might ask why we observe such significant fat tail phenomena in the distribution of the partial derivatives. What generates these fat-tailed phenomena? Recall that the model structure in this case study has this form: $y = f(X)$. This has two dimensions: one is the distribution of the inputs, X, and the other is the model structure, f, which is a continuous and nonlinear function defined in the input domain **X**. In principle, both can be responsible for this.

It is interesting to consider the influent flowrate received by the plant as an example. The influent flowrate is composed of rainwater runoff and dry weather flows. The dry weather flow tends to follow a rather stable pattern, in which extreme values are rarely observed. On the other hand, the rainwater runoff is heavily influenced by weather conditions and hydrological conditions of the urban area (Tchobanoglus et al., 2003). As a result, extreme events statistics such as peak over threshold (POT) is typically used to model this behavior (Embrechts et al., 2013). This likely explains the fat tail distribution of the influent flowrate rate received by the plant (see Fig. 3). On other hand, the temperature of the wastewater follows a typical seasonal behavior with a predictable pattern that is not subject to random and sudden increases or decreases. Therefore, the distribution of the temperature belongs to thin-tailed probability distributions (see Fig. 3). It is interesting to note that the tail exponents for the partial derivatives of influent flowrate ($\alpha = 2.3$) is much lower than the tail exponent for temperature ($\alpha = 4.2$). This observation is consistent with the shape of the input distributions.

Surprisingly, in many global sensitivity analysis studies, algebraic and well-defined inputs (typically uniform to normal distributions) and their theoretical convergence using the central limit theorem are studied (Borgonovo and Plischke, 2016). However, as we have demonstrated in this case study, real-world datasets can have a region that follows a fat-tailed distribution. Hence, the theoretical convergence of the global sensitivity analysis methods, especially those employing Monte Carlo numerical solution, is needed.

For process systems engineering studies, the methods presented in this study for checking the convergence of the Monte Carlo analysis will, in general, be necessary, especially when dealing with inputs that may have a Paretianity in their data. In many techno-economic analysis studies, such as net present value, discounted cash flow models, etc., one has to deal with market price data such as projected sale prices for the products in plant useful lifetime. The shape of the distribution of the product prices may well follow a fat-tailed distribution (Taleb, 2020). Another potential application area for the convergence tests is uncertainty quantification. In the field of machine learning, sampling-based techniques such as ensemble or bootstrap sampling are becoming widely used. Therefore, verifying that a certain ensemble size is sufficient will be needed to ensure a consistent estimation of the mean or variance of the model predictions (Aouichaoui et al., 2023).

6. Conclusions

The present study performed a derivative-based global sensitivity analysis on a real-world dataset. The study underscores the significance of conducting global sensitivity analysis with a proper assessment of input distribution. The convergence and consistency of Monte Carlo analysis are strongly dependent on the shape of the input distribution. The main finding is that for a robust and consistent interpretation of the sensitivity analysis results, it is recommended to use mean absolute deviation (MAD) as a sensitivity measure. This is particularly needed since the sensitivity metrics defined based on higher-order moments converge very slowly under fat-tailed distributions.

Overall, this study emphasizes the careful consideration of the input distribution and proper convergence testing in real-world applications of global sensitivity analysis. This is to ensure that the results are statistically significant and not just a fluke resulting from the fluctuations of coin-tossing experiments.

7. Data availability and reproducibility statement

Fig. 2 can be reproduced using the dataset and the DNN model which is provided on DTU data repository (Sin, 2023). A time series plot of the dataset is provided in the Supplementary Material Of this paper, as well as the correlation matrix used in the sampling shown in Fig. 3. The original dataset comes from the more extensive data collected in (Chen et al., 2019) and the deep learning model is one of the models published and developed by (Hwangbo et al., 2021). The dataset and model used in this study are uploaded in the Supplementary Material for review.

CRediT authorship contribution statement

Gürkan Sin: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ces.2024.120124.

References

Al, R., et al., 2019. Meta-modeling based efficient global sensitivity analysis for wastewater treatment plants–An application to the BSM2 model. Comput. Chem. Eng. pp. 127, 233–246.
Aouichaoui, A., et al., 2023. Combining Group-Contribution Concept and Graph Neural Networks Toward Interpretable Molecular Property Models. J. Chem. Inf. Model. 63 (3), 725–744.
Borgonovo, E., Plischke, E., 2016. Sensitivity analysis: A review of recent advances. Eur. J. Oper. Res. 248 (3), 869–887.
Chen, X., et al., 2019. Assessment of full-scale N2O emission characteristics and testing of control concepts in an activated sludge wastewater treatment plant with alternating aerobic and anoxic phases. Environ. Sci. Technol. 53 (21), 12485–12494.
Cirillo, P., 2013. Are your data really Pareto distributed? Phys. A: Stat. Mech. its Applications 392 (23), 5947–5962.
Embrechts, P., Klüppelberg, C., Mikosch, T., 2013. Modelling extremal events: for insurance and finance. Springer Science & Business Media, s.l.
European Commission, 2023. Better regulation: guidelines and toolbox. [Online]. Available at: https://commission.europa.eu/law/law-making-process/planning-and-proposing-law/better-regulation/better-regulation-guidelines-and-toolbox_en.
Feller, W., 1991. An Introduction to Probability Theory and Its Applications. Volume 1, 3rd Edition. s.l.:Wiley .
Hwangbo, S., Al, R., Chen, X., Sin, G., 2021. Integrated model for understanding N2O emissions from wastewater treatment plants: a deep learning approach. Environ. Sci. Technol. 55 (3), 2143–2151.
Hwangbo, S.R.A.X.C.a.G.S., 2021. Integrated model for understanding N2O emissions from wastewater treatment plants: a deep learning approach. Environ. Sci. Technol., 55(3), 2143–2151.
Kucherenko, S., Albrecht, D., Saltelli, A., Exploring multi-dimensional spaces: a Comparison of Latin Hypercube and Quasi Monte Carlo Sampling Techniques, 2015 arXiv:1505.02350.
Kucherenko, S., Giamalakis, D., Shah, N., García-Muñoz, S., 2020. Computationally efficient identification of probabilistic design spaces through application of metamodeling and adaptive sampling. Comput. Chem. Eng. 132, 106608.
Lamboni, M., Kucherenko, S., 2021. Multivariate sensitivity analysis and derivative-based global sensitivity measures with dependent variables. Reliab. Eng. Syst. Safe. 212, 107519.

Mara, T.A., Tarantola, S., Annoni, P., 2015. Non-parametric methods for global sensitivity analysis of model output with dependent inputs. Environ. Model. Softw. 72, 173–183.
Razavi, S., et al., 2021. The future of sensitivity analysis: An essential discipline for systems modeling and policy support. Environ. Model. Softw. pp. 137, 104954.
Saltelli, A., et al., 2007. Global Sensitivity Analysis. The Primer. John Wiley & Sons, s.l.
Saltelli, A., et al., 2019. Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. Environ. Model. Softw. 114, 29–39.
Sin, G., Gernaey, K.V., Lantz, A.E., 2009. Good modeling practice for PAT applications: Propagation of input uncertainty and sensitivity analysis. Biotechnol. Progr. 25 (4), 1043–1053.
Sin, G., 2023. Longterm N2O emission data and deep learning model. Technical University of Denmark., Årgang https://doi.org/10.11583/DTU.24106860.v1, p. Dataset..
Sobol, I., Kucherenko, S., 2009. Derivative based global sensitivity measures and their link with global sensitivity indices. Math. Comput. Simul. 79, 3009–3017.
Taleb, N.N., 2020. Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. Stem Academic Press, s.l.
Tchobanoglus, G., Burton, F., Stensel, H., 2003. Wastewater engineering: treatment and reuse. McGraw-Hill Education, s.l.
USA EPA, 2023. Guidance for Quality Assurance Project Plans for Modeling EPA QA/G-5M. [Online] Available at: https://www.epa.gov/quality/guidance-quality-assurance-project-plans-modeling-epa-qag-5m.