# Surrogate-based global sensitivity analysis with statistical guarantees via floodgate

Massimo Aufiero and Lucas Janson

Department of Statistics, Harvard University

**Abstract**

Computational models are utilized in many scientific domains to simulate complex systems. Sensitivity analysis is an important practice to aid our understanding of the mechanics of these models and the processes they describe, but performing a sufficient number of model evaluations to obtain accurate sensitivity estimates can often be prohibitively expensive. In order to reduce the computational burden, a common solution is to use a surrogate model that approximates the original model reasonably well but at a fraction of the cost. However, in exchange for the computational benefits of surrogate-based sensitivity analysis, this approach comes with the price of a loss in accuracy arising from the difference between the surrogate and the original model. To address this issue, we adapt the floodgate method of Zhang and Janson (2020) to provide valid surrogate-based confidence intervals rather than a point estimate, allowing for the benefit of the computational speed-up of using a surrogate that is especially pronounced for high-dimensional models, while still retaining rigorous and accurate bounds on the global sensitivity with respect to the *original* (non-surrogate) model. Our confidence interval is asymptotically valid with almost no conditions on the computational model or the surrogate. Additionally, the accuracy (width) of our confidence interval shrinks as the surrogate's accuracy increases, so when an accurate surrogate is used, the confidence interval we report will correspondingly be quite narrow, instilling appropriately high confidence in its estimate. We demonstrate the properties of our method through numerical simulations on the small Hymod hydrological model, and also apply it to the more complex CBM-Z meteorological model with a recent neural-network-based surrogate.

# 1 Introduction

## 1.1 Problem Statement

The use of computational models to describe and simulate complex phenomena is ubiquitous in many scientific fields, and such models play a critical role in both understanding natural processes and informing high-stakes decisions across many domains. Complex systems are first represented with a mathematical model—usually implicitly using coupled differential

equations—and are then implemented in computer code to exactly or approximately solve these equations. While these models are very useful for simulating processes under various conditions that may be impossible or infeasibly expensive to observe in reality, the models themselves can often be incredibly computationally expensive, with a single run taking very powerful computers on the order of hours or even days to evaluate, prohibiting the generation of large amounts of data (Razavi et al., 2012a).

Computational models typically rely on a large number of parameters, some of which may be observed data or known physical constants, while many are unobservable quantities for which only their distributions or ranges are known. The large number of uncertain inputs and the complexity of the model results in high variability in the outputs (Do and Razavi, 2020; Xu and Gertner, 2008). Thus, the question of how variation in each input factor affects the model outputs is of great interest for a number of reasons. It can be informative both to developers of the model determining how best to improve its structure or reduce uncertainty, as well as to researchers and decision-makers studying the model as a proxy for reality.

Sensitivity analysis is a set of mathematical methods for analyzing how the inputs of a system influence its output(s) by quantitatively measuring the attribution of output variability to uncertainty in the inputs (Pianosi et al., 2015). Sensitivity analysis is a critical component of understanding the intricacies of a model's behavior. However, due to the complexity of the models, it is in most cases impossible to derive an analytical description of input-output relationships, especially since the model itself is almost never given in closed-form. Thus, most methods for estimating input sensitivities rely on querying the model many times while varying the input values, often with very specific sampling schemes (see Section 1.4). However, obtaining a large enough number of samples from the model to reach a desired level of accuracy is often prohibitively expensive, especially when it needs to be done repeatedly for a large number of inputs. Much of the research on sensitivity analysis is therefore aimed at reducing the total number of model evaluations needed to get accurate estimates.

One popular solution is to use a surrogate model (also referred to as an 'emulator' or 'metamodel') to approximate the original model at a much lower computational cost. Surrogates can be either data-driven machine learning models or lower-fidelity simulation models that reduce the resolution or number of components of the original. While a computational model's surrogate can be orders of magnitude faster to run (Kelp et al., 2018), its outputs will in general not be exactly the same as those of the original model when both are given the same inputs. Therefore, simply applying sensitivity analysis to the surrogate model sacrifices many of the desirable statistical properties, such as consistency and asymptotic normality, with respect to the sensitivity in the original (non-surrogate) model. This means that even though we can generate a much larger number of samples from the surrogate for sensitivity analysis, it is unclear how well the resulting estimates will correspond to the actual sensitivity value of interest.

## 1.2 Contribution

We present a novel method for conducting sensitivity analysis that extends Zhang and Janson (2020)'s floodgate method for inferring variable importance in high-dimensional regression. Our extension of floodgate (hereafter referred to simply as "floodgate") leverages surrogate

models for computational efficiency, yet retains statistical guarantees on its estimation with respect to the sensitivity in the *original* model. In particular, floodgate outputs upper- and lower-bounds for the (original) model sensitivity with provably high asymptotic coverage no matter how accurate the surrogate is. The widths of these intervals directly improve with the accuracy of the surrogate, so that floodgate ensures appropriate uncertainty quantification always, while still providing high precision given a sufficiently high-fidelity surrogate.

Floodgate offers a significant computational advantage compared to non-surrogate-based techniques, by a factor of up to the dimension of the input space given a sufficiently fast and accurate surrogate, making it especially advantageous for high-dimensional models. In addition, given a dataset of sampled inputs and their model evaluations, floodgate can be applied with *no additional* evaluations of the original model (i.e., one only needs to evaluate the surrogate on new inputs), which is not the case for many standard sensitivity analysis techniques since they require very specific pairs of samples. Furthermore, it accounts for surrogate inaccuracy to rigorously quantify the uncertainty of estimates using intervals that are much narrower and require fewer assumptions than existing theoretical bounds on surrogate-based estimation error. Floodgate is applicable to any computational model, any surrogate model, and nearly any input distribution and sampling scheme.

## 1.3 Notation

In this paper, we assume that the computational model $f : \mathsf{R}^d \to \mathsf{R}$ is a deterministic function—which is true for most computational models—though even randomized models can be subsumed into our framework by simply including the exogenous randomness (i.e., the random seed) in $f$ as an additional input. Note that we use $f$ to denote the original computational model and $f$ to denote the corresponding surrogate, though in the sensitivity analysis literature the latter is commonly used for computational models.

We will present floodgate as being applied to a single input of interest, denoted $X$, at a time and we denote the remaining $d - 1$ of $f$'s inputs collectively as $Z$. Of course this does not mean floodgate cannot or should not be used for sensitivity analysis of many different inputs, and indeed the more inputs it is applied to, the larger its computational advantage; see Sections 2.3, 3, and 4. We use $(X_i; Z_i)$ to denote an individual sample from the input distribution, and we denote the distribution of $X_i$ conditional on $Z_i$ as $P_{X_i|Z_i}$. Finally, we use $z_\alpha$ to denote the $\alpha$th quantile of the standard normal distribution.

## 1.4 Background and Related Work

Variance-based global sensitivity analysis methods seek to quantitatively attribute output variance to uncertainty in each input or group of inputs. The estimand that we consider in this paper is the total-order sensitivity index (Homma and Saltelli, 1996). This quantity measures the proportion of the total variance of the model output that results from variation in the input $X$, through the sum of all direct effects and interactions with other inputs. A formal definition of this quantity is provided below.

De nition 1.1. (Total-order sensitivity index) For a computational model $f$, the total-order sensitivity index for input $X$ is de ned as

$$S := \frac{E[\mathrm{Var}(f(X,Z)|Z)]}{\mathrm{Var}(f(X,Z))}, \tag{1.1}$$

whenever the appropriate moments exist, with the convention that $0/0 = 0$.

Several Monte Carlo (MC) estimators for $S$ have been proposed, including by Jansen (1999), Homma and Saltelli (1996), and Sobol' (2007). Most of them utilize the Sobol' pick-freeze (SPF) scheme (Sobol', 1993, 2001; Janon et al., 2013). In MC estimation, SPF estimators use a set of $n$ i.i.d. pairs of points where within each pair, the values of $Z$ are the same, but $X$ is sampled (conditionally) independently for both points. To formalize this, assume that for a set of i.i.d. samples $\{(X_i, Z_i)\}_{i=1}^n$ we can sample $\tilde{X}_i \sim P_{X|Z_i}$ such that $\tilde{X}_i \perp\!\!\!\perp X_i \mid Z_i$ for each $i$. An example SPF estimator for $S$ introduced by Jansen (1999) is

$$\hat{S} := \frac{\frac{1}{2n}\sum_{i=1}^n \left(f(X_i,Z_i) - f(\tilde{X}_i,Z_i)\right)^2}{\frac{1}{n-1}\sum_{i=1}^n f(X_i,Z_i)\cdot\frac{1}{n}\sum_{i=1}^n f(X_i,Z_i)}. \tag{1.2}$$

$\hat{S}$ is consistent for $S$ and was proven by Sobol' (2001) and Saltelli et al. (2010) to have lower variance than other similar SPF estimators, so we only focus on this particular $\hat{S}$ as a comparison to Goodgate in Sections 3 and 4. However, for any of these SPF estimators, computation is very expensive, especially in high dimensions, since computing the full set of $n$-sample sensitivity index estimates for all $d$ inputs requires $n(d+1)$ total evaluations of $f$. Even if one had access to a dataset of i.i.d. input samples with their corresponding model evaluations, $\hat{S}$ could not be computed without nding additional evaluations of $f$ to create $d$ sets of $n$ paired points as described above. Sheikholeslami and Razavi (2020) and Plischke et al. (2013) have developed sensitivity analysis methods applicable to any given data that target di erent estimands, but to our knowledge there are no such methods for estimating $S$.

As mentioned in Section 1.1, surrogate modeling techniques are often employed to sidestep this computational obstacle, though they of course sacri ce some accuracy in the obtained estimates. Many surrogates are dynamic models that fully simulate the original model by predicting all of the model's outputs over arbitrary time scales (e.g., Castelletti et al. (2012); Kelp et al. (2020)). This includes reduced or simpli ed versions of the original model (e.g., in Arandia et al. (2019)) or even just running the same model at a lower spatio-temporal resolution. In some cases, a dynamic surrogate is built by using a machine-learning model to replace only a single particularly expensive subcomponent of the original model rather than trying to replace the entire model (e.g., in Mills et al. (2019)) or by using a machine-learning model to map from the outputs of a lower- delity simulation to those of the higher- delity model (e.g., in Zhang et al. (2019)). However, since in sensitivity analysis we are typically only concerned with a particular scalar transformation of the computational model's outputs, data-driven response surface surrogates that learn a direct mapping from the inputs to the output of interest (for some xed time scale and forcing data, if applicable) can also be

used (see Razavi et al. (2012b) for a review). For any of these types of surrogates, the naïve method of estimating $S$ using a surrogate $f$ is to simply replace $f$ with $f$ in the computation of some estimator for $S$. In the case of the SPF estimator $\hat{S}$ from (1.2), the surrogate-based estimator is

$$\hat{S}^f := \frac{\frac{1}{2n}\sum_{i=1}^{n} f(X_i;Z_i) - f(\tilde{X}_i;Z_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n} f(X_i;Z_i) - \frac{1}{n}\sum_{i=1}^{n} f(X_i;Z_i)}. \tag{1.3}$$

Another common use of surrogate models for estimating sensitivity indices that is somewhat less similar to our approach is to fit a special type of response surface surrogate from which estimates of the sensitivity indices can be directly computed from the model coefficients rather than by MC estimation. These methods rely on the fact that $f$ can be exactly represented by an expansion onto some infinite basis, and good approximations of $f$ can often be obtained by using only a finite number of terms in the basis for $f$. The coefficients of the terms in this reduced basis are then used to compute the truncated-sum estimates $\hat{S}^f$. Two of the most popular such methods are the Fourier amplitude sensitivity test (FAST) (Cukier et al., 1973; Saltelli et al., 2010) and polynomial chaos expansion (PCE) (Kalinina et al., 2020; Cheng and Lu, 2018), which use Fourier and polynomial bases, as the names suggest. Tsokanas et al. (2021) use PCE, kriging, and polynomial chaos kriging surrogates for estimating sensitivity indices for a virtual hybrid model representing a prototype motorcycle. Stephens et al. (2011) compare radial basis functions, neural networks, and least squares support vector machines for surrogate-based sensitivity analysis of a computational fluid dynamics model. Le Gratiet et al. (2017) demonstrate the use of Gaussian processes to obtain confidence intervals on sensitivity indices for a truss structure engineering model, as well as PCE for point estimates.

Both approaches to surrogate-based sensitivity analysis face the same issue: the estimates that use $f$ in place of $f$ do not truly estimate $S$, but rather the quantity $S^f := \mathrm{E}[\mathrm{Var}(f(X_i;Z_i) \mid Z)]/\mathrm{Var}(f(X_i;Z_i))$; where the difference $S^f - S$ is unknown, making it hard to rigorously quantify the error of the estimates $\hat{S}^f$ with respect to $S$. This is because there is now a surrogate error (i.e., $S^f - S$) in addition to the sampling error ($\hat{S}^f - S^f$) in the estimates. Janon et al. (2014) present a method that uses the MC estimates $\hat{S}^f$ to obtain upper and lower bounds on $S$, but it requires a computationally intensive optimization procedure repeated over bootstrap samples as well as knowledge of the pointwise error bound on $|f(X;Z) - f(X;Z)|$, which is only computable for a few specific types of surrogates and computational models. Janon et al. (2013) establish conditions on the rate of convergence of a sequence of surrogates $f_n$ to $f$ in order for $\hat{S}^{f_n}$ to be consistent for $S$ and asymptotically normal in the double limit as $n \to \infty$ and $f_n \to f$, where the surrogate $f_n$ depends on the sample size (e.g., if some fixed proportion of the data is used for training the surrogate and the rest for computing $\hat{S}^{f_n}$). However, these conditions require convergence rates of the model error to zero that are only satisfied by a limited class of computational model–surrogate pairs, and thus the results are not applicable in general. Most recently, Panin (2021) provides a bound on the surrogate error $|S^f - S|$ that depends on the mean squared error (MSE) of $f$ and is estimable from data. We compare goodgate to their bounds both theoretically (Section 2.4) and empirically (Section 3.4) by extending their methodology to

obtain confidence intervals as well, demonstrating that floodgate produces intervals that are consistently and substantially narrower than theirs.

As mentioned in Section 1.2, our method is an extension of the original floodgate method for high-dimensional inference from Zhang and Janson (2020). Their algorithm outputs an asymptotically valid lower confidence bound for the numerator of $S$ in the general regression setting. Instead of the lower-bound function used in Zhang and Janson (2020), we use the function introduced in (Zhang, 2022), a later work focused on other estimands. Zhang and Janson (2020) notes that the numerator of $S$ can be upper-bounded as well, but that the bound cannot be made tight except in the noiseless setting—an edge case in that paper but the case of primary interest in this paper—so they do not pursue the idea further. In contrast to these works, we derive and give full treatment to an upper confidence bound (which is tight, as we work in the noiseless setting) so that we can provide a (two-sided) confidence interval for $S$. Indeed, for sensitivity analysis, upper confidence bounds are often of even more value than lower confidence bounds, as they allow dimensionality reduction via dropping inputs with low sensitivities. We also present novel results on the asymptotic width of our confidence interval and on the computational speedups floodgate offers compared to non-surrogate-based sensitivity analysis methods, as well as numerical demonstration of floodgate's value for sensitivity analysis and efficient code[1] in python to ease the use of floodgate by others.

# 2 Methods

## 2.1 Bounds for the total-order sensitivity index

As outlined in Section 1.4, there are a number of existing estimators that are consistent and asymptotically normal for $S$ when using samples from the computational model $f$, but when a surrogate model $\hat{f}$ is substituted for $f$, these properties are no longer guaranteed for $\hat{S}^{\hat{f}}$. Indeed, as detailed in Section 1.4, for a fixed surrogate $\hat{f}$ (i.e., $\hat{f}$ does not change with $n$), $\hat{S}^{\hat{f}}$ will converge to a value $S^{\hat{f}}$ that is not equal to $S$ in general. Thus, we would like to have a way to leverage a surrogate model $\hat{f}$ to estimate $S$ with a computable bound on the error, ideally as small as possible.

We introduce a bias-aware surrogate-based method for inference on $S$ that allows us to quantify uncertainty in the form of confidence intervals. Floodgate uses surrogate examples to estimate upper and lower bounds on $S$. In particular, we

(a) define functions $\ell(\hat{f})$ and $u(\hat{f})$ such that $\ell(\hat{f}) \le S \le u(\hat{f})$ for any surrogate $\hat{f}$,

(b) construct estimators $\hat{\ell}_n(\hat{f})$, $\hat{u}_n(\hat{f})$ that converge to $\ell(\hat{f})$ and $u(\hat{f})$, respectively,

(c) derive a confidence interval $[L_n, U_n]$ with provable coverage for $S$.

The intuition behind this process is that if $L_n$ is a lower confidence bound for $\ell(\hat{f})$, then it is by construction also a lower confidence bound for $S$, and similarly for the upper confidence bound $U_n$.

[1] Code available at https://github.com/aeroma12/floodgate

We use the following upper- and lower-bound functions, whose numerators were originally derived by Zhang and Janson (2020) and Zhang (2022), respectively. For any surrogate $f : \mathbb{R}^d \to \mathbb{R}$, define

$$u(f) = \frac{E\left(f(X,Z) - E[f(X,Z)|Z]\right)^2}{\mathrm{Var}\left(f(X,Z)\right)} \tag{2.1}$$

$$\ell(f) = u(f) - \frac{E[(f(X,Z) - f(X,Z))^2]}{\mathrm{Var}\left(f(X,Z)\right)}; \tag{2.2}$$

again with the convention that 0/0 = 0.[1] The numerator of $u(f)$ is simply the MSE of $f_z(Z) := E[f(X,Z)|Z]$, which is a function of only $Z$, and the numerator of $\ell(f)$ is the difference between this quantity and the MSE of $f$ itself. Lemma 2.1 is a key result for proving the accuracy and validity of GOODGATE. The proof of this lemma is provided in Appendix A.2.

Lemma 2.1. For any $f$ and $f'$ such that $\ell(f)$ and $u(f')$ exist,

$$\ell(f) - \ell(f') = S = u(f') - u(f). \tag{2.3}$$

## 2.2 Estimators and confidence intervals

Given samples $\{(X_i, Z_i)\}_{i=1}^n$, we can construct simple MC estimators $\hat{\ell}_n(f)$ and $\hat{u}_n(f)$. We start by defining the quantities

$$\mathrm{MSE}(f) = E\left(f(X,Z) - f(X,Z)\right)^2 \tag{2.4}$$

$$\mathrm{MSE}(f_z) = E\left(f(X,Z) - f_z(Z)\right)^2. \tag{2.5}$$

Thus, we can express $u(f) = \frac{\mathrm{MSE}(f_z)}{\mathrm{Var}(f(X,Z))}$ and $\ell(f) = \frac{\mathrm{MSE}(f_z) - \mathrm{MSE}(f)}{\mathrm{Var}(f(X,Z))}$. If the samples $(X_i, Z_i)$ are i.i.d., then we easily obtain unbiased MC estimators for each term in the numerators and denominators by simply replacing the outer expectations with sample means. For $\mathrm{MSE}(f)$, this is straightforward. We can compute the sample mean of

$$M_i := \left(f(X_i, Z_i) - f(X_i, Z_i)\right)^2, \tag{2.6}$$

for $i \in \{1, \ldots, n\}$ and use this as our estimator. Note that $M_i$ is a (random) function of $f$, but we drop the dependence on $f$ in the notation for simplicity. $M_i$ is trivially unbiased for $\mathrm{MSE}(f)$. However, unbiased estimation of $\mathrm{MSE}(f_z)$ and $\mathrm{Var}(f(X,Z))$ is somewhat less straightforward as both are expected squared errors with respect to expectations which are themselves generally intractable to compute analytically. For $\mathrm{Var}(f(X,Z))$, we can use the standard unbiased variance estimator, which is the sample mean of

$$V_i := \frac{n}{n-1}\left(f(X_i, Z_i) - \frac{1}{n}\sum_{i'=1}^{n} f(X_{i'}, Z_{i'})\right)^2. \tag{2.7}$$

[1] Note that when $\mathrm{Var}(f(X,Z)) = 0$, the numerator of $u(f)$ is always non-negative (so $u(f)$ may take the value $\infty$) and the numerator of $\ell(f)$ is always non-positive (so $\ell(f)$ may take the value $-\infty$). This is clearly true for $u(f)$ from (2.1) and follows for $\ell(f)$ from the proof in Appendix A.1.

For MSE($f_z$), the challenge is dealing with the analytically intractable $f_z(Z) = E[f(X,Z)|Z]$. Assuming that we can sample $K \geq 1$ copies $\tilde{X}_i^{(k)}$ of $X_i$ from the conditional distribution $P_{X_i|Z_i}$, where each $\tilde{X}_i^{(k)}$ is conditionally independent of $X_i$, then for each $i \in \{1,\ldots,n\}$ we can estimate $f_z(Z_i)$ with

$$\frac{1}{K}\sum_{k=1}^{K} f(\tilde{X}_i^{(k)}, Z_i),$$

which we will use in our estimator for MSE($f_z$). In particular, the quantity

$$M_i^z := \left(f(X_i,Z_i) - \frac{1}{K}\sum_{k=1}^{K} f(\tilde{X}_i^{(k)},Z_i)\right)^2 - \frac{1}{K+1}\left(f(X_i,Z_i) - \frac{1}{K}\sum_{k=1}^{K} f(\tilde{X}_i^{(k)},Z_i)\right)^2 \tag{2.8}$$

is unbiased for MSE($f_z$). This is formalized in Lemma 2.2, the proof of which is provided in Appendix A.3. Again, we drop the dependence of $M_i^z$ on $f$ in the notation for simplicity.

Lemma 2.2. For $K \geq 1$, given a set of i.i.d. original samples $\{(X_i,Z_i)\}_{i=1}^n$ and modified samples $\tilde{X}_i^{(1)},\ldots,\tilde{X}_i^{(K)} \sim P_{X_i|Z_i}$ such that $\tilde{X}_i^{(k)} \perp\!\!\!\perp X_i | Z_i$ for each $i,k$, then for any $f$, $f : \mathbb{R}^d \to \mathbb{R}$, the quantity $M_i^z$ defined in (2.8) satisfies:

$$E[M_i^z] = \mathrm{MSE}(f_z).$$

Since $\{(M_i^z, M_i, V_i)\}_{i=1}^n$ are i.i.d. and unbiased for $(\mathrm{MSE}(f_z), \mathrm{MSE}(f), \mathrm{Var}(f(X,Z)))$, their sample means $M^z$, $M$, and $V$ are asymptotically normal estimators for their respective estimands. We can then combine these to get estimators for $v(f)$ and $u(f)$ that are also asymptotically normal by the delta method. In particular, the estimators are

$$\hat{v}_n(f) = \frac{M^z - M}{V} \tag{2.9}$$

$$\hat{u}_n(f) = \frac{M^z}{V}. \tag{2.10}$$

Due to the asymptotic normality of these estimators, it is straightforward to obtain asymptotically valid upper and lower confidence bounds, as described in Algorithm 1.

**Algorithm 1** Floodgate for surrogate-based sensitivity analysis

Input: Samples $\{(X_i, Z_i)\}_{i=1}^n$, surrogate $f : \mathbb{R}^d \to \mathbb{R}$, $K \geq 2 \in \mathbb{N}$, and confidence level $\alpha \in (0, 1)$.

For each $i \in \{1, \ldots, n\}$, compute $M_i^z$, $M_i$, $V_i$ according to (2.8), (2.6), and (2.7), their sample means $(\bar{M}^z, \bar{M}, \bar{V})$, and their $3 \times 3$ sample covariance matrix $\hat{\Sigma}$.

If $\bar{V} = 0$, set $L_n = 0$ and $U_n = 1$. Else, compute

$$s_u^2 = \frac{1}{\bar{V}^2}\left(\hat{\Sigma}_{11} - 2\frac{\bar{M}^z}{\bar{V}}\hat{\Sigma}_{13} + \left(\frac{\bar{M}^z}{\bar{V}}\right)^2\hat{\Sigma}_{33}\right)$$

$$\text{and } s_\ell^2 = \frac{1}{\bar{V}^2}\left(\hat{\Sigma}_{11} + \hat{\Sigma}_{22} + \left(\frac{\bar{M}^z}{\bar{V}}\bar{M}\right)^2\hat{\Sigma}_{33} - 2\hat{\Sigma}_{12} + 2\frac{\bar{M}^z}{\bar{V}}\bar{M}\hat{\Sigma}_{23} - \hat{\Sigma}_{13}\right);$$

$$\text{and set } L_n = \max\left(0, \frac{\bar{M}^z}{\bar{V}}\bar{M} - z_{1-\alpha/2}\frac{s_\ell}{\sqrt{n}}\right) \text{ and } U_n = \min\left(1, \frac{\bar{M}^z}{\bar{V}} + z_{1-\alpha/2}\frac{s_u}{\sqrt{n}}\right).$$

Output: Confidence interval $[L_n, U_n]$.

Theorem 2.3 establishes the asymptotic coverage of the interval $[L_n, U_n]$, and the proof is given in Appendix A.4. It requires only very mild moment assumptions that are standard for the central limit theorem (CLT). We need $f(X, Z)$ and $f(X, Z)$ to have finite fourth moments rather than the standard second moments because of the squared terms in the estimators and estimands.

Theorem 2.3. (Asymptotic coverage). For i.i.d. samples $\{(X_i, Z_i)\}_{i=1}^n$, any computational model $f$ and surrogate $f : \mathbb{R}^d \to \mathbb{R}$, and any $\alpha \in (0, 1)$, if $E[f^4(X, Z)], E[f^4(X, Z)] < \infty$, then the bounds $L_n$ and $U_n$ output by Algorithm 1 satisfy

$$\liminf_{n\to\infty} P(L_n \leq S \leq U_n) \geq 1 - \alpha.$$

Note that the validity of the confidence interval derived from $f$ does not depend on $f$ itself. We do not require any conditions on the quality of $f$—in fact, a less accurate $f$ would tend to have higher coverage, since $\ell(f)$ and $u(f)$ would span $S$ by a wider margin, whereas the confidence bounds derived from a naïve surrogate-based estimator such as $\hat{S}_f$ in (1.3) are not guaranteed to cover the true value at all when $f$ differs from $f$.

Of course, in practice we hope that $f$ is an accurate surrogate for $f$. Since there are very few assumptions on the properties of $f$, we can leverage state-of-the-art machine learning algorithms and arbitrary domain knowledge to construct a surrogate that is as accurate as possible. For example, $f$ can be any of the various types of machine learning, physically-based, or hybrid surrogate models described in Section 1.4, with the only restriction being that it is independent of the data used for constructing the floodgate bounds (e.g., if $f$ is fitted via machine learning, its training data should be independent of the data used to compute $L_n$ and $U_n$). Theorem 2.4 establishes that the width of the confidence interval $[L_n, U_n]$ converges to a value depending on the accuracy of $f$, and it does so at an $O_p(n^{-1/2})$ rate. Thus, when $f$ is very close to $f$, we can asymptotically achieve very tight bounds while still guaranteeing coverage. The proof of Theorem 2.4 is provided in Appendix A.5.

Theorem 2.4. (Width of confidence intervals). Under the same assumptions as in Theorem 2.3 and the additional assumption that $\mathrm{Var}(f(X,Z)) > 0$, the bounds $L_n$ and $U_n$ output by Algorithm 1 satisfy

$$U_n - L_n \leq \frac{\mathrm{MSE}(\hat{f})}{\mathrm{Var}(f(X,Z))} + O_p\left(n^{-1/2}\right).$$

## 2.3 Computational savings

We consider here the computational expense of goodgate when applied to every input of $f$. Thus, for the remainder of this section (and in Sections 3 and 4), we will explicitly label the total-order sensitivity index for the $j$th input using $S_j$ (i.e., $S_j$ corresponds to labelling the $j$th input as $X$).

For an individual sensitivity index, Algorithm 1 requires a set of only $n$ points $(X_i, Z_i)$ evaluated by $f$, plus an additional $K$ samples $(\tilde{X}_i^{(k)}, Z_i)$ evaluated by $f$ for each $i \in \{1, \ldots, n\}$, for a total of $nK$ evaluations of $f$. While these $nK$ evaluations of $f$ on the resampled inputs are distinct for each of the $d$ sensitivity indices, the same set of $n$ original points evaluated by $f$ are used every time. Since we generally assume that the surrogate is much less expensive to evaluate than $f$, we expect the cost of the $nK$ total evaluations of $f$ necessary for inference on the full set of $\{S_j\}_{j=1}^d$ is negligible compared to the $n$ total evaluations of $f$ necessary. If we need to train the surrogate ourselves first, this will require an additional $n_{\text{train}}$ evaluations of $f$.

For comparison, $n$-sample estimation of all $d$ sensitivity indices by $\hat{S}$ requires $n(d+1)$ total evaluations of $f$. Note that when $K = 1$ and $\hat{f} = f$, $\hat{\tau}_n^2(\hat{f}) = \hat{\tau}_n^2(f) = \hat{S}$, so in the ideal case where we have a surrogate that is perfectly accurate and takes zero computation time, the output of goodgate with $K = 1$ is identical to estimation/inference using $\hat{S}$ but with computation time lower by a factor of $1/(d+1)$ (and goodgate will only be more accurate with larger $K$). In general (continuing to assume $K = 1$ for simplicity), if we let $c$ denote how many times higher $f$'s computational expense is than $\hat{f}$'s, the same speedup approximately holds as long as $c \ll d$ and as long as $\mathrm{MSE}(\hat{f})/\mathrm{Var}(f(X,Z)) \ll n^{-1/2}$. If one or both of these conditions do not hold, the computational advantage of goodgate will be less than a factor of $1/(d+1)$, though the speedup may still be significant. Goodgate's advantage disappears roughly when $\mathrm{MSE}(\hat{f})/\mathrm{Var}(f(X,Z)) \gtrsim n^{-1/2}\left(\frac{1}{c} + \frac{1}{d}\right)^{-1/2}$, which makes intuitive sense since it means the surrogate is either not particularly fast or not particularly accurate (or both), and naturally we should not expect to be able to leverage such a low-quality surrogate to beneficial effect.

Another advantage of goodgate compared to most non-surrogate methods is that it can be applied to almost any pre-existing dataset—which may have been collected for a purpose unrelated to sensitivity analysis or obtained from another source—without the need for additional evaluations of $f$. While we present our results for the case of i.i.d. data, they generalize to any sampling scheme that is compatible with asymptotically normal estimation of expectations of functions of the samples. For example, CLTs exist for Latin hypercube sampling (Owen, 1992) and general randomized quasi-Monte Carlo sampling techniques (Owen, 2013), which are frequently used in sensitivity analysis because they achieve faster rates of convergence than standard MC estimation. Indeed, we demonstrate the application of

oodgate to a dataset of non-i.i.d. samples in Section 4.

## 2.4 Comparison to existing bounds on error of $\hat{S}^f$

As mentioned in Section 1.4, Panin (2021) proved bounds on the surrogate error term $|S^f - S|$ and provide a method for estimating those bounds. In particular, Corollary 1 of their paper establishes the bound

$$|S^f - S| \leq \min\left\{1, E + 2\sqrt{\bar{S}}, E + 2\sqrt{1 - \bar{S}}\right\} E, \tag{2.11}$$

where $E = \sqrt{\text{MSE}(f)/\text{Var}(f(X,Z))}$. Therefore, adding and subtracting the bound on the right-hand side of (2.11) to the surrogate sensitivity $S^f$ yields an interval guaranteed to contain the true sensitivity $S$. Note that if $S$ is not exactly 0 or 1, then as $f$ converges to $f$, the width of this interval is $O(E)$.

The interval $[\ell(f), u(f)]$ that we provide in Section 2.1 is also guaranteed to contain the true sensitivity $S$, but its width is only $O(E^2)$, so it shrinks at a much faster rate as the quality of the surrogate improves (i.e., $\text{MSE}(f) \to 0$). In fact, since its width is exactly $E^2$ (see (2.2)), by comparing $E^2$ to each of the three terms in the bound in (2.11), it is immediate that our interval $[\ell(f), u(f)]$ is strictly narrower than that derived from (2.11) whenever $S \notin \{0, 1\}$ and when $E < 1$, the latter of which we would expect to be true for any reasonable surrogate.

Panin (2021) also propose a natural plug-in estimator for the bound in (2.11) in Section 3.1.2., which we note can be extended to obtain con dence intervals as well via a similar CLT argument to what we use. While computing $\hat{S}^f$ requires no evaluations of $f$ (or $n_{\text{train}}$ evaluations if $f$ must be trained from scratch), one still needs $n > 1$ samples from $f$ in order to estimate $E$. Thus, computing con dence intervals for a set of $d$ sensitivity indices requires $n$ evaluations of $f$ (and $nd$ evaluations of $f$), similar to oodgate.

We perform empirical comparisons of our bound to Panin (2021)'s in Sections 3.4 and 4.

# 3 Simulations

## 3.1 Overview of computational model

We conducted numerical experiments using the Hymod (Boyle, 2001; Wagener et al., 2001) hydrological model to demonstrate oodgate's guarantees on coverage and the relative widths of its con dence intervals compared to standard SPF estimators. Hymod is relatively simple, low-dimensional, and inexpensive to evaluate (though still slower than most common surrogates). Thus, it would not likely be necessary to use a surrogate for it and it would not be the ideal target of oodgate in practice; however, it was a good candidate for these simulations because it is so inexpensive to evaluate, meaning we were able to run many independent trials and obtain precise approximations of the sensitivity indices for reference. In addition, Hymod has been used frequently within the sensitivity analysis literature as a test case for new methods (e.g., Herman et al. (2013); Razavi and Gupta (2016); Cheng and Lu (2018); Sheikholeslami and Razavi (2020)).

There are five uncertain parameters in the model governing the mechanics of the system, and we treat these as the model inputs whose sensitivities we study. The names, descriptions, units, and ranges of these inputs are given in Table 1 in Appendix A.6. We used forcing data and observed outputs obtained from the Leaf Catchment in Mississippi (Pianosi and Wagener, 2016) in these experiments. An implementation of Hymod in Python was obtained from the SAFE Toolbox (Pianosi and Wagener, 2016).

## 3.2 Simulation setup

The response variable we considered for the Hymod simulations was the Nash–Sutcliffe efficiency criterion (NSE) (Nash and Sutcliffe, 1970). The NSE is a very commonly used metric in hydrology to assess how well model predictions agree with observed data. It is often used in sensitivity analysis studies, since it helps understand what inputs significantly impact model accuracy and thus should be calibrated more carefully, or what inputs have a very small effect and could thus be fixed or dropped for model simplification.

Since we need to be able to draw i.i.d. samples $(X_i, Z_i)$ and conditionally i.i.d. samples $\tilde{X}_i^{(k)}$ for each sensitivity index, we must define the joint input distribution. As is standard in other global sensitivity analysis studies of this model, we assumed that the inputs were independent and followed uniform distributions. The ranges we used for each variable's distribution are provided in Table 1 in Appendix A.6.

We applied `goodgate` to compute confidence intervals for the full set of $d = 5$ sensitivity indices using different computational budgets $N$ ranging from 100 to 50000 (to be defined shortly). We compare our intervals to asymptotically normal confidence intervals derived from $\hat{S}$ (1.2) and $\hat{S}^f$ (1.3), clipping the bounds to be within $[0, 1]$. In Section 3.4 we also provide a more detailed comparison to the error bounds from Panin (2021). We define the computational budget as the total number of evaluations of $f$ required in the full process of estimating all $d$ sensitivity indices.

- For `goodgate`, the full set of $n = N$ evaluations of $f$ is used for each input. Recall that `goodgate` uses the same set of $\{f(X_i, Z_i)\}_{i=1}^n$ and distinct sets of $n$ $K$ evaluations of $f$ for each input.
- For the non-surrogate SPF estimator $\hat{S}$, only $n = N/(d+1)$ pairs of evaluations of $f$ are used for each input. Recall that computation of each $\hat{S}$ uses the same set $\{f(X_i, Z_i)\}_{i=1}^n$ and a distinct set $\{f(\tilde{X}_i, Z_i)\}_{i=1}^n$ for each input.

This relationship between $n$ and $N$ puts `goodgate` and $\hat{S}$ on an equal computational footing, assuming that evaluation of $f$ is negligible compared to $f$. However, since we use pretrained surrogates, it is not possible to do the same for $\hat{S}^f$ as it uses no evaluations from $f$. We choose to use $n = N$ pairs of evaluations of $f$ for each input. While $n$ could have been chosen to be arbitrarily large for $\hat{S}^f$, the main conclusion about $\hat{S}^f$ from these experiments is its lack of coverage, which is only more evident at larger sample sizes.

Note that if one did not have a pretrained surrogate and had to train one from scratch, then for `goodgate`, some fraction of the $N$ samples (e.g., $N/2$) would have to be reserved for training, while the remaining $n = N/2$ are used for computing the confidence intervals. In this case, the surrogate-based SPF estimator $\hat{S}^f$ would use all $N$ samples to train and 0

samples from $f$ plus an arbitrary number of samples from $f$ for inference. The relationship between $n$ and $N$ remains unchanged for $\hat{S}$ since it does not use a surrogate.

To simulate having surrogates of various qualities, we conducted all experiments using kernel ridge regression (KRR) with radial basis function kernel pretrained on different amounts of data. In particular, we train low-quality ($\text{MSE}(f)/\text{Var}(f(X,Z)) \approx 0.07$; dashed lines) and high-quality ($\text{MSE}(f)/\text{Var}(f(X,Z)) \approx 0.01$; solid lines) surrogates. In practice, we would expect the type of surrogates used to incorporate specific domain knowledge and potentially to have been trained offline and made publicly available, rather than be fitted by an out-of-the-box machine learning model as done here. However, Hymod is simple and low-dimensional enough that there are no existing high-fidelity surrogates, so fitting a non-parametric model on a large separate dataset provided a reasonable alternative for obtaining surrogates to test our method.

## 3.3 Results

As mentioned in Section 2.3, we will index $S_j$ (and similarly for its estimators) here to explicitly denote the total-order sensitivity index for the $j$th input, since we apply `oodgate` and other methods to every input. Figure 1 shows the confidence intervals obtained by `oodgate` and the two other methods, along with their empirical coverage. The curves in the confidence bound plots are the averages over 1000 independent trials. All standard errors are less than 0.008 on the left-hand sides of the plots and less than 0.001 on the right-hand sides. The horizontal dotted red lines in the confidence bound plots represents the "ground truth" $S_j$ values, which were estimated using a consistent estimator with $10^8$ samples. These ground truth values were used for computing the coverage. The horizontal dotted red lines in the coverage plots represent the nominal level of 95%.

These plots demonstrate that `oodgate` can output narrow (and valid) confidence intervals when using an accurate surrogate. Our high-quality intervals (solid green curves) are almost always tighter than those using the non-surrogate SPF estimator (blue curves) for smaller sample sizes. For example, for the input Rf, our estimated interval is $[0.667, 0.757]$ for 100 model evaluations, whereas the non-surrogate-based method gives the much wider interval $[0.116, 0.758]$. For the input Rs, we provide an upper confidence bound of 0.102 for $N = 100$, whereas the non-surrogate-based method's is more than double that (0.228).

In addition, our coverage is almost always valid with only a few small violations at the smallest sample sizes, and it is consistently higher than the other methods' even when our intervals are narrower. Some of the reasons contributing to this is that each of our estimators gets to use $d + 1 = 6$ times as many samples evaluated by $f$ as $\hat{S}_j$ due to the computational budget constraints (putting `oodgate` closer to "asymptopia") and our bounds are conservative against the bias of $\hat{S}$. While the surrogate-based SPF estimator (orange curves) also outputs narrow confidence intervals, they fail to account for surrogate inaccuracy and thus have no guarantees on validity, as demonstrated in the coverage plots. Thus, this estimator can output very high confidence for an incorrect estimate. As an example, on the right-hand side of the plot for the input Sm, the low-quality surrogate-based SPF method outputs a very narrow interval around a value roughly double the true value—the coverage here is of course zero.

As mentioned in Section 3.1, these results are all for a low-dimensional model where the

Figure 1: 95% con dence intervals and empirical coverage for the Hymod model sensitivity indices using oodgate and the non-surrogate- and surrogate-based SPF estimators with di erent quality surrogates.

advantage of floodgate should be particularly unpronounced. We were able to achieve much narrower intervals (while maintaining valid coverage) than the non-surrogate-based method with just 6 times the number of samples here, but for a higher-dimensional model we would expect to gain a larger advantage—we will demonstrate an application of floodgate to a roughly 100-dimensional model in Section 4.

### 3.4 Comparison to Existing Bounds on the Error of $\hat{S}^f$

We also compare floodgate to the bounds derived by Panin (2021), discussed in Section 2.4. In particular, we compare our confidence intervals for each $S_j$ to those computed by adding and subtracting their estimated bound on surrogate error to $\hat{S}_j^f$ and applying the CLT with the multivariate delta method, clipping the bounds to be within $[0, 1]$. Figure 2 shows the width of each method's confidence intervals averaged over 1000 independent trials along with their empirical coverage for different values of the computational budget $N$ and using low-quality and high-quality surrogates. All standard errors are less than 0.0083 on the left-hand sides of the plots and less than 0.0004 on the right-hand sides. The dependence on $N$ for floodgate (green curves) is the same as in the previous section. For the confidence intervals based on Panin (2021)'s bounds (purple curves), all $N$ evaluations of $f$ are used to estimate $E$, and we again use $n = N$ pairs of samples from $f$ for computing $\hat{S}_j^f$. Thus, both methods use $N$ total evaluations from $f$. We chose to plot the widths of the intervals rather than the intervals themselves on the same set of axes as in Figure 1 because both methods give the same guarantees on coverage of $S_j$, and thus the positions of the intervals are less interesting than their relative widths. Indeed, both methods' empirical coverage is valid almost everywhere, reaching practically 100% on the right-hand side of the plots since both methods' bounds are conservative.

We see that the width of our confidence intervals are consistently substantially smaller than those of Panin (2021) for each sensitivity index and surrogate. The empirical coverage for both methods is above the nominal level in nearly every setting. Floodgate's coverage does dip below most notably for the input Sm for smaller $N$ values, though even in this case it never falls below 80%. The intervals derived using Panin (2021)'s bounds tend to have higher coverage than the floodgate interval, which is expected given that their bounds are looser in the accurate-surrogate regime as shown in Section 2.4. Coverage for both methods becomes consistently 95% as $N$ increases, as they both have the same asymptotic guarantees.

We thus demonstrate that floodgate achieves rigorous quantification of the uncertainty of surrogate-based estimates using tighter bounds than those provided by Panin (2021).

## 4 Application

### 4.1 Overview of computational model and surrogate

To demonstrate a more realistic application of floodgate, we used the Carbon Bond Mechanism Z (CBM-Z) meteorological model (Zaveri and Peters, 1999) for simulating tropospheric gas-phase chemistry. It is both higher-dimensional and more computationally expensive than Hymod, and there are existing surrogates that have been built for it. CBM-Z models the evo-

Figure 2: Widths of 95% con dence intervals and empirical coverage for the Hymod model sensitivity indices using oodgate and Panin (2021)'s bounds with di erent quality surrogates.

lution of 101 gaseous and aerosol species over a given time period by numerically integrating a system of partial di erential equations. In addition to the initial concentrations of each species, the model has four additional meteorological inputs: temperature, pressure, relative humidity, and the cosine of the solar zenith angle. While we use a stand-alone version of CBM-Z for our experiments, it is commonly employed as the gas-phase chemistry module within larger chemistry transport models used for air quality forecasting, and it is typically the most time-consuming component (Wang et al., 2019).

For our experiments, we used a multitarget regression neural network surrogate from Kelp et al. (2020). The network consists of an encoder that reduces the input concentrations to a lower-dimensional latent representation, an operator that approximates the integration step in the latent space, and a decoder that maps the integrated system back to the original space. The operator can be applied recurrently to make predictions over arbitrary time scales. For a 24-hour simulation, Kelp et al. (2020) report a speedup by a factor of roughly 3700 compared to the true CBM-Z model on the same hardware, while maintaining high accuracy for various outputs of interest on an independent test set of randomly sampled inputs.

## 4.2 Experimental setup

We consider the predicted concentration of ozone (O3) over a 2-hour interval as the response variable of interest. Ozone is a common subject of sensitivity analysis studies with atmospheric chemistry models, as in Constantin and Barrett (2014) and Christian et al. (2018),

since it has a high impact on air quality. We chose 2 hours as the time interval because we found that the surrogate from Kelp et al. (2020) was most accurate at this time scale, with $\mathrm{MSE}(f) = \mathrm{Var}(f(X, Z)) \approx 0.06$.

Since we did not have access to a working implementation of the CBM-Z model, we used a dataset of 80,000 samples provided by Kelp et al. (2020) that was independent of the data used for training and validating their surrogate. The initial concentrations and meteorological inputs were sampled from independent uniform distributions with ranges outlined in Kelp et al. (2020). In particular, the dataset consists of 625 i.i.d. batches of samples, where each batch is a full Latin hypercube with 128 points. Thus, we used sample sizes that were multiples of 128 and applied the CLT to the means within each batch for deriving confidence intervals.

We again compare goodgate to the confidence intervals given by the surrogate-based SPF estimator $\hat{S}_j^f$ and using the error bounds from Panin (2021) for every input. Since the implementation of the CBM-Z model used by Kelp et al. (2020) to train their surrogate was proprietary and not available to us, we were not able to evaluate $f$ ourselves and thus could not implement the non-surrogate baseline for these experiments. As discussed in Section 2.3, the fact that goodgate can be applied to a pre-existing dataset is itself an advantage over the estimator $\hat{S}_j$.

As in Sections 3.2 and 3.4, for a given computational budget $N$, goodgate uses the full set of $N$ samples from $f$ to estimate each sensitivity index. Panin (2021)'s bounds use all $N$ samples to estimate $E$. Again, since $\hat{S}_j^f$ does not use any evaluations of $f$, it has no dependence on $N$, so we choose to evaluate it with $m = N$ terms in the summation for each sensitivity index as well.

## 4.3 Results

Figure 3 shows the confidence intervals obtained by goodgate, the surrogate-based SPF estimator, and Panin (2021)'s bounds for a representative subset of the inputs (the plots for the full set of inputs can be found in Appendix A.7). In this case, since we have only a small dataset that does not contain samples in the paired form required for an SPF estimator, we cannot obtain sufficiently precise estimates of the ground truth sensitivities to add the horizontal red lines to the plots or to calculate coverage.

For most of the inputs, the estimated sensitivities were very close to 0, and their plots look almost identical to that for HCl. This is because the green and purple curves are essentially the same when $\hat{S}_j^f \approx 0$, since both give lower bounds of 0, and the error bound in (2.11) will be simply $E^2$, which is the same as the width of our intervals. However, when $\hat{S}_j^f$ is even slightly above 0, the gap between the green and purple upper bounds becomes quite noticeable, as in the plot for OLEI. For O3, which naturally has the highest sensitivity of all the inputs, Panin (2021)'s confidence interval is roughly twice as wide as ours for $N = 512$, and it is nearly five times wider than ours for $N = 80000$. The intervals derived using the surrogate-only method (orange curves) converges to a value outside our interval entirely, which again demonstrates that this naïve method can give high confidence for an incorrect result.

Figure 3: 95% con dence intervals for a representative subset of sensitivity indices for the CBM-Z model using oodgate, the surrogate-based SPF estimator, and Panin (2021)'s bounds.

# 5 Conclusion

We present a novel method, oodgate, for conducting statistically rigorous sensitivity analysis using surrogate models to achieve substantial computational speedups relative to existing methods. Floodgate provides asymptotically valid con dence intervals whose accuracy directly improves with the quality of the surrogate used. Since all of our theoretical results are very general, the method can be applied to any computational model with any surrogate, allowing users to take full advantage of arbitrary domain knowledge and state-of-the-art machine learning models, regardless of their complexity, to achieve results that are as accurate as possible. Furthermore, the con dence intervals provide rigorous quanti cation of the uncertainty of surrogate-based estimates, informing the user as to when they can and cannot be trusted.

We highlight some of the advantages of oodgate compared to existing work empirically through simulations with the relatively simple Hymod hydrological model in Section 3 and in a more realistic application to the CBM-Z meteorological model in Section 4. These results validate the fact that having a high-quality surrogate allows us to obtain very tight bounds, and when this is not the case, our con dence intervals account for the surrogate's inaccuracy and still provide valid coverage. We compare oodgate both theoretically and empirically to similar results for statistically valid surrogate-based inference from Panin (2021), and we show that the con dence intervals we provide are signi cantly narrower while still maintaining coverage.

More broadly, we see the future implications of our work in this paper as aiding in the general goal of being able to study complex, expensive models through their less-expensive surrogates without sacri cing statistical guarantees. Computational models are used in

several high-stakes settings, aiding in important scientific discoveries, shaping engineering designs, and making forecasts or simulations that inform high-impact decisions; thus it is crucial to have both an accurate understanding of the input-output relationships in these models, as well as to be conscious of the uncertainty in the studies we perform. This principle extends beyond just sensitivity analysis, applying to all forms of inference on the models' features and outputs in which one might be interested.

# 6 Acknowledgements

The authors would like to thank Lu Zhang for sharing early drafts of her PhD thesis that our work built upon and Makoto Kelp for generously providing code and data necessary for the CBM-Z application. M. A. and L.J. were partially supported by a CAREER grant from the National Science Foundation (Grant #DMS2045981).

# References

Ernesto Arandia, Fearghal O'Donncha, Sean McKenna, Seshu Tirupathi, and Emanuele Ragnoli. Surrogate modeling and risk-based analysis for solute transport simulations. Stochastic Environmental Research and Risk Assessment, 33:1907–1921, 12 2019. ISSN 14363259. doi: 10.1007/s00477-018-1549-6.

Douglas Patrick Boyle. Multicriteria calibration of hydrologic models. PhD thesis, The University of Arizona, 2001.

A. Castelletti, S. Galelli, M. Restelli, and R. Soncini-Sessa. Data-driven dynamic emulation modelling for the optimal management of environmental systems. Environmental Modelling and Software, 34:30–43, 6 2012. ISSN 13648152. doi: 10.1016/j.envsoft.2011.09.003.

Kai Cheng and Zhenzhou Lu. Adaptive sparse polynomial chaos expansions for global sensitivity analysis based on support vector regression. Computers and Structures, 194:86–96, 1 2018. ISSN 00457949. doi: 10.1016/j.compstruc.2017.09.002.

K. E. Christian, W. H. Brune, J. Mao, and X. Ren. Global sensitivity analysis of geos-chem modeled ozone and hydrogen oxides during the intex campaign. Atmospheric Chemistry and Physics, 18(4):2443–2460, 2018. doi: 10.5194/acp-18-2443-2018. URL https://acp.copernicus.org/articles/18/2443/2018/.

B V Constantin and S R H Barrett. Application of the complex step method to chemistry-transport modeling. Atmospheric Environment, 99:457–465, 2014.

R. I. Cukier, C. M. Fortuin, K. E. Shuler, A. G. Petschek, and J. H. Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory. Journal of Chemical Physics, 59:3873–3878, 10 1973. ISSN 10897690. doi: 10.1063/1.1680571.

Nhu Cuong Do and Saman Razavi. Correlation effects? a major but often neglected component in sensitivity and uncertainty analysis. Water Resources Research, 56, 3 2020. ISSN 19447973. doi: 10.1029/2019WR025436.

J. D. Herman, P. M. Reed, and T. Wagener. Time-varying sensitivity analysis clari es the e ects of watershed model formulation on model behavior. Water Resources Research, 49: 1400{1414, 3 2013. ISSN 19447973. doi: 10.1002/wrcr.20124.

Toshimitsu Homma and Andrea Saltelli. Importance measures in global sensitivity analysis of nonlinear models, 1996.

Alexandre Janon, Thierry Klein, Agnes Lagnoux-Renaudie, Ma elle Nodet, and Clementine Prieur. Asymptotic normality and e ciency of two sobol index estimators. 2013. doi: 10.48550/ARXIV.1303.6451. URL https://arxiv.org/abs/1303.6451 .

Alexandre Janon, Ma elle Nodet, and Clementine Prieur. Uncertainties assessment in global sensitivity indices estimation from metamodels. International Journal for Uncertainty Quanti cation , 4(1):21{36, 2014. URL https://hal.inria.fr/inria-00567977 .

Michiel J W Jansen. Analysis of variance designs for model output, 1999.

Anna Kalinina, Matteo Spada, David F. Vetsch, Stefano Marelli, Calvin Whealton, Peter Burgherr, and Bruno Sudret. Metamodeling for uncertainty quanti cation of a ood wave model for concrete dam breaks. Energies, 13, 7 2020. ISSN 19961073. doi: 10.3390/en13143685.

Makoto M Kelp, Christopher W Tessum, and Julian D Marshall. Orders-of-magnitude speedup in atmospheric chemistry modeling through neural network-based emulation. 2018.

Makoto M. Kelp, Daniel J. Jacob, J. Nathan Kutz, Julian D. Marshall, and Christopher W. Tessum. Toward stable, general machine-learned models of the atmospheric chemical system. Journal of Geophysical Research: Atmospheres, 125, 12 2020. ISSN 21698996. doi: 10.1029/2020JD032759.

Lo c Le Gratiet, Stefano Marelli, and Bruno Sudret. Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes, pages 1289{1325. Springer International Publishing, Cham, 2017. ISBN 978-3-319-12385-1. doi: 10.1007/978-3-319-12385-1_38. URL https://doi.org/10.1007/978-3-319-12385-1_38 .

Kyle Mills, Kevin Ryczko, Iryna Luchak, Adam Domurad, Chris Beeler, and Isaac Tamblyn. Extensive deep neural networks for transferring small scale learning to large scale systems. Chem. Sci, 10:4129{4140, 2019. doi: 10.1039/C8SC04578J.

J.E. Nash and J.V. Sutcli e. River ow forecasting through conceptual models part i | a discussion of principles. Journal of Hydrology, 10(3):282{290, 1970. ISSN 0022-1694. doi: https://doi.org/10.1016/0022-1694(70)90255-6. URL https://www.sciencedirect.com/science/article/pii/0022169470902556 .

Art B. Owen. A central limit theorem for latin hypercube sampling. Journal of the Royal Statistical Society. Series B (Methodological), 54(2):541{551, 1992. ISSN 00359246. URL http://www.jstor.org/stable/2346140 .

Art B. Owen. Monte Carlo theory, methods and examples. 2013.

Ivan Panin. Risk of estimators for sobol' sensitivity indices based on metamodels. Electronic Journal of Statistics, 15(1), 1 2021. doi: 10.1214/20-ejs1793. URL https://doi.org/10.12142F20-ejs1793.

Francesca Pianosi and Thorsten Wagener. Understanding the time-varying importance of different uncertainty sources in hydrological modelling using global sensitivity analysis. Hydrological Processes, 30(22):3991{4003, 2016. doi: https://doi.org/10.1002/hyp.10968. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.10968 .

Francesca Pianosi, Fanny Sarrazin, and Thorsten Wagener. A matlab toolbox for global sensitivity analysis. Environmental Modelling & Software, 70:80{85, 2015. ISSN 1364-8152. doi: https://doi.org/10.1016/j.envsoft.2015.04.009. URL https://www.sciencedirect.com/science/article/pii/S1364815215001188 .

Elmar Plischke, Emanuele Borgonovo, and Curtis L Smith. Global sensitivity measures from given data. European Journal of Operational Research, 226:536{550, 2013. doi: 10.1016/j.ejor.2012.11.047.

Saman Razavi and Hoshin V. Gupta. A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. application. Water Resources Research, 52:440{455, 1 2016. ISSN 19447973. doi: 10.1002/2015WR017559.

Saman Razavi, Bryan A. Tolson, and Donald H. Burn. Numerical assessment of metamodelling strategies in computationally intensive optimization. Environmental Modelling and Software, 34:67{86, 6 2012a. ISSN 13648152. doi: 10.1016/j.envsoft.2011.09.010.

Saman Razavi, Bryan A. Tolson, and Donald H. Burn. Review of surrogate modeling in water resources. Water Resources Research, 48, 2012b. ISSN 00431397. doi: 10.1029/2011WR011527.

Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. Computer Physics Communications, 181:259{270, 2 2010. ISSN 00104655. doi: 10.1016/j.cpc.2009.09.018.

Razi Sheikholeslami and Saman Razavi. A fresh look at variography: Measuring dependence and possible sensitivities across geophysical systems from any given data. Geophysical Research Letters, 47, 10 2020. ISSN 19448007. doi: 10.1029/2020GL089829.

I. M. Sobol'. Sensitivity estimates for nonlinear mathematical models. MMCE, 1:407{414, 1993.

I. M. Sobol'. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates, 2001.

I. M. Sobol'. Global sensitivity analysis indices for the investigation of nonlinear mathematical models. Matematicheskoe Modelirovanie, 19:23{24, 2007.

D. W. Stephens, D. Gorissen, K. Crombecq, and T. Dhaene. Surrogate based sensitivity analysis of process equipment. Applied Mathematical Modelling, 35:1676–1687, 4 2011. ISSN 0307904X. doi: 10.1016/j.apm.2010.09.044.

Nikolaos Tsokanas, Roland Pastorino, and Bozidar Stojadinovic. A comparison of surrogate modeling techniques for global sensitivity analysis in hybrid simulation. Machine Learning and Knowledge Extraction, 4:1–21, 12 2021. doi: 10.3390/make4010001.

T. Wagener, D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian. A framework for development and application of hydrological models. Hydrology and Earth System Sciences, 5(1):13–26, 2001. doi: 10.5194/hess-5-13-2001. URL https://hess.copernicus.org/articles/5/13/2001/.

Hui Wang, Junmin Lin, Qizhong Wu, Huansheng Chen, Xiao Tang, Zifa Wang, Xueshun Chen, Huaqiong Cheng, and Lanning Wang. Mp cbm-z v1.0: Design for a new carbon bond mechanism z (cbm-z) gas-phase chemical mechanism architecture for next-generation processors. Geoscientific Model Development, 12:749–764, 2 2019. ISSN 19919603. doi: 10.5194/gmd-12-749-2019.

Chonggang Xu and George Zdzislaw Gertner. Uncertainty and sensitivity analysis for models with correlated parameters. Reliability Engineering and System Safety, 93:1563–1573, 10 2008. ISSN 09518320. doi: 10.1016/j.ress.2007.06.003.

Rahul A. Zaveri and Leonard K. Peters. A new lumped structure photochemical mechanism for large-scale applications. Journal of Geophysical Research: Atmospheres, 104(D23): 30387–30415, 1999. doi: https://doi.org/10.1029/1999JD900876. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999JD900876.

Lu Zhang. Inference on nonparametric targets and discrete structures. PhD thesis, Harvard University, 2022.

Lu Zhang and Lucas Janson. Floodgate: inference for model-free variable importance. arXiv preprint arXiv:2007.01283, 2020.

Xinyue Zhang, Yanfang Wang, Wei Zhang, Yueqiu Sun, Siyu He, Gabriella Contardo, Francisco Villaescusa-Navarro, and Shirley Ho. From dark matter to galaxies with convolutional networks. 2 2019. URL http://arxiv.org/abs/1902.05965.

# A Appendix

## A.1 Numerator of $\grave{}(f)$

From (2.1) and (2.2), we can write

$$\grave{}(f) = \frac{E[(f(X,Z) - f_z(Z))^2] - E[(f(X,Z) - f(X,Z))^2]}{Var(f(X,Z))}.$$

When $Var(f(X,Z)) = 0$, we have $f(X,Z) \overset{a.s.}{=} E[f(X,Z)] =: a$. In this case, the numerator of $\grave{}(f)$ simplifies to

$$\begin{aligned}
&E(a - f_z(Z))^2 - E(a - f(X,Z))^2 \\
&= a^2 - 2aE[f_z(Z)] + E f_z^2(Z) - a^2 + 2aE[f(X,Z)] - E f^2(X,Z) \\
&= E f_z^2(Z) - E f^2(X,Z) - 2a(E[f_z(Z)] - E[f(X,Z)]) \\
&= E f_z^2(Z) + 2E[f_z(Z)f(X,Z)] - E f^2(X,Z) - 2a(E[f(X,Z)] - E[f(X,Z)]) \\
&= E(f(X,Z) - f_z(Z))^2 ; \qquad \text{(A.1)}
\end{aligned}$$

where the third equality follows from the law of iterated expectations:

$$E[f_z(Z)] = E[E[f(X,Z)|Z]] = E[f(X,Z)]$$

and

$$E[f_z(Z)f(X,Z)] = E[E[f_z(Z)f(X,Z)|Z]] = E[f_z(Z)E[f(X,Z)|Z]] = E f_z^2(Z) .$$

Note that the term in (A.1) is always non-positive. In particular, it will be 0 when $Var(f(X,Z)|Z) \overset{a.s.}{=} 0$, meaning that $\grave{}(f) = 0$ by definition, and it will be negative when $E[Var(f(X,Z)|Z)] > 0$, meaning that $\grave{}(f)$ can take the value $-\infty$.

## A.2 Proof of Lemma 2.1

Lemma. For any $f$ and $f'$ such that $\grave{}(f)$ and $u(f)$ exist,

$$\grave{}(f) - \grave{}(f') = S = u(f') - u(f).$$

Proof. Recall the definition of the function $f_z(Z) = E[f(X,Z)|Z]$ from Section 2.1, and similarly define

$$f'_z(Z) = E[f'(X,Z)|Z].$$

We can rewrite the expressions for $u(f)$ and $S$ given in (2.1) and (1.1) as

$$u(f) = \frac{E(f'(X,Z) - f_z(Z))^2}{Var(f'(X,Z))} \quad \text{and} \quad S = \frac{E(f'(X,Z) - f_z(Z))^2}{Var(f'(X,Z))}.$$

When $Var(f'(X,Z)) = 0$, $S = 0$ by definition, $u(f)$ can take the values 0 or $-\infty$, and $\grave{}(f)$ can take the values 0 or $-\infty$, so the inequality holds.

When $\mathrm{Var}(f_\theta(X,Z)) > 0$, in order to prove that $u(f_\theta) \in S$, it suffices to show $D_1 := E(f_\theta(X,Z) - f_z(Z))^2 - E(f_{\theta_0}(X,Z) - f_z(Z))^2 \geq 0$. With some algebraic manipulation, we have

$$\begin{aligned} D_1 &= E(f_\theta(X,Z) - f_z(Z) + f_z(Z) - f_z(Z))^2 - E(f_{\theta_0}(X,Z) - f_z(Z))^2 \\ &= E(f_\theta(X,Z) - f_z(Z))^2 + 2E[(f_\theta(X,Z) - f_z(Z))(f_z(Z) - f_z(Z))] \\ &\quad + E(f_z(Z) - f_z(Z))^2 - E(f_{\theta_0}(X,Z) - f_z(Z))^2 \\ &= 2E[(f_\theta(X,Z) - f_z(Z))(f_z(Z) - f_z(Z))] + E(f_z(Z) - f_z(Z))^2 \end{aligned} \tag{A.2}$$

We can expand the expectation in the first term in (A.2) to get

$$\begin{aligned} & E(f_\theta(X,Z) - f_z(Z))(f_z(Z) - f_z(Z)) \\ &= E[f_\theta(X,Z)(f_z(Z) - f_z(Z))] - E[f_z(Z)(f_z(Z) - f_z(Z))] \\ &= E[E[f_\theta(X,Z)(f_z(Z) - f_z(Z)) \mid Z]] - E[f_z(Z)(f_z(Z) - f_z(Z))] \\ &= E[E[f_\theta(X,Z) \mid Z](f_z(Z) - f_z(Z))] - E[f_z(Z)(f_z(Z) - f_z(Z))] \\ &= E[f_z(Z)(f_z(Z) - f_z(Z))] - E[f_z(Z)(f_z(Z) - f_z(Z))] \\ &= 0. \end{aligned}$$

Therefore, by plugging this into (A.2), we have

$$D_1 = E(f_z(Z) - f_z(Z))^2 \geq 0. \tag{A.3}$$

Now, we consider the lower bound $\ell(f_\theta)$ from (2.2). Again, to prove $\ell(f_\theta) \in S$, it suffices to show $D_2 := E(f_\theta(X,Z) - f_z(Z))^2 - E(f_\theta(X,Z) - f(X,Z))^2 - E(f(X,Z) - f_z(Z))^2 \geq 0$. Note that

$$D_2 = D_1 - E(f_\theta(X,Z) - f(X,Z))^2. \tag{A.4}$$

Since we found $D_1$ in (A.3), we just need to deal with this second term. Define the functions

$$\begin{aligned} h(X,Z) &= f(X,Z) - f_z(Z) \\ h_\theta(X,Z) &= f_\theta(X,Z) - f_z(Z). \end{aligned}$$

Thus, we can rewrite

$$\begin{aligned} & E(f_\theta(X,Z) - f(X,Z))^2 \\ &= E(h_\theta(X,Z) + f_z(Z) - h(X,Z) - f_z(Z))^2 \\ &= E(h_\theta(X,Z) - h(X,Z))^2 + 2E[(h_\theta(X,Z) - h(X,Z))(f_z(Z) - f_z(Z))] \\ &\quad + E(f_z(Z) - f_z(Z))^2. \end{aligned} \tag{A.5}$$

From (A.3), (A.4), and (A.5), we have

$$D_2 = -E(h_\theta(X,Z) - h(X,Z))^2 - 2E[(h_\theta(X,Z) - h(X,Z))(f_z(Z) - f_z(Z))]. \tag{A.6}$$

Simplifying the expectation in the second term in (A.6), we have

$$\begin{aligned} E[(h_\gamma(X,Z) - h(X,Z))(f_z(Z) - f_z(Z))] &= E[E[(h_\gamma(X,Z) - h(X,Z))(f_z(Z) - f_z(Z)) \mid Z]] \\ &= E[E[(h_\gamma(X,Z) - h(X,Z)) \mid Z](f_z(Z) - f_z(Z))] \\ &= 0, \end{aligned}$$

where the last equality holds because $E[h_\gamma(X,Z) \mid Z] \overset{a.s.}{=} E[h(X,Z) \mid Z] \overset{a.s.}{=} 0$ by definition. Finally, plugging this into (A.6), we have

$$D_2 = -E\left[(h_\gamma(X,Z) - h(X,Z))^2\right] \leq 0. \tag{A.7}$$

Finally, it is clear from (A.3) and (A.7) that

$$\ell(f_\gamma) = S = u(f_\gamma). \qquad \square$$

## A.3 Proof of Lemma 2.2

**Lemma.** For $K \geq 1$, given a set of i.i.d. original samples $\{(X_i, Z_i)\}_{i=1}^n$ and modified samples $\tilde{X}_i^{(1)}, \ldots, \tilde{X}_i^{(K)} \overset{iid}{\sim} P_{X_i|Z_i}$ such that $\tilde{X}_i^{(k)} \perp\!\!\!\perp X_i \mid Z_i$ for each $i, k$, then for any $f_\gamma, f : \mathbb{R}^d \to \mathbb{R}$, the quantity $M_i^z$ defined in (2.8) satisfies:

$$E[M_i^z] = \mathrm{MSE}(f_z).$$

**Proof.** We start by defining the term

$$F_i^z := \frac{1}{K} \sum_{k=1}^K f_\gamma(\tilde{X}_i^{(k)}, Z_i),$$

so we can rewrite the expression for $M_i^z$ from (2.8) as

$$M_i^z = (f_\gamma(X_i, Z_i) - F_i^z)^2 - \frac{1}{K+1}(f_\gamma(X_i, Z_i) - F_i^z)^2,$$

Note that

$$E[M_i^z] = E\left[(f_\gamma(X,Z) - F_i^z)^2\right] - \frac{1}{K+1} E\left[(f_\gamma(X,Z) - F_i^z)^2\right], \tag{A.8}$$

by linearity of expectation, so we can examine each term on the right hand side of (A.8) separately. From here on, we drop the index $i$ on random variables within expectations since samples are i.i.d. Expanding the first term, we have

$$E\left[(f_\gamma(X,Z) - F^z)^2\right] = E\left[f_\gamma^2(X,Z)\right] - 2E[f_\gamma(X,Z)F^z] + E\left[F^{z2}\right]. \tag{A.9}$$

Simplifying the expectation in the second term of (A.9), we get

$$
\begin{aligned}
E[f(X;Z)F_z] &= \frac{1}{K}E\left[f(X;Z)\sum_{k=1}^{K} f(\tilde{X}^{(k)};Z)\right] \\
&= \frac{1}{K}\sum_{k=1}^{K} E\left[f(X;Z)f(\tilde{X}^{(k)};Z)\right] \\
&= E\left[f(X;Z)f(\tilde{X}^{(1)};Z)\right] \\
&= E\left[E\left[f(X;Z)f(\tilde{X}^{(1)};Z)\,|\,Z\right]\right] \\
&= E\left[E[f(X;Z)\,|\,Z]\,E\left[f(\tilde{X}^{(1)};Z)\,|\,Z\right]\right] \\
&= E[f(X;Z)E[f(X;Z)\,|\,Z]] \\
&= E[f(X;Z)f_z(Z)]; \qquad \text{(A.10)}
\end{aligned}
$$

where the third and sixth equalities use the fact that $\{X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(K)}\}$ are all exchangeable, the fourth and sixth lines apply the law of iterated expectations, and the fifth line follows from the conditional independence of $X$ and $\tilde{X}^{(1)}$.

Simplifying the expectation in the third term of (A.9), we get

$$
\begin{aligned}
E\left[F_z^2\right] &= \frac{1}{K^2}E\left[\left(\sum_{k=1}^{K} f(\tilde{X}^{(k)};Z)\right)^2\right] \\
&= \frac{1}{K^2}E\left[\sum_{k=1}^{K} f^2(\tilde{X}^{(k)};Z) + 2\sum_{k=1}^{K}\sum_{t=k+1}^{K} f(\tilde{X}^{(k)};Z)f(\tilde{X}^{(t)};Z)\right] \\
&= \frac{1}{K^2}\sum_{k=1}^{K} E\left[f^2(\tilde{X}^{(k)};Z)\right] + \frac{2}{K^2}\sum_{k=1}^{K}\sum_{t=k+1}^{K} E\left[f(\tilde{X}^{(k)};Z)f(\tilde{X}^{(t)};Z)\right] \\
&= \frac{1}{K}E\left[f^2(X;Z)\right] + \frac{K-1}{K}E\left[f(X;Z)f(\tilde{X}^{(1)};Z)\right] \\
&= \frac{1}{K}E\left[f^2(X;Z)\right] + \frac{K-1}{K}E\left[E\left[f(X;Z)f(\tilde{X}^{(1)};Z)\,|\,Z\right]\right] \\
&= \frac{1}{K}E\left[f^2(X;Z)\right] + \frac{K-1}{K}E\left[E[f(X;Z)\,|\,Z]\,E\left[f(\tilde{X}^{(1)};Z)\,|\,Z\right]\right] \\
&= \frac{1}{K}E\left[f^2(X;Z)\right] + \frac{K-1}{K}E\left[(E[f(X;Z)\,|\,Z])^2\right] \\
&= \frac{1}{K}E\left[f^2(X;Z)\right] + \frac{K-1}{K}E\left[f_z^2(Z)\right]; \qquad \text{(A.11)}
\end{aligned}
$$

where the fourth and seventh equalities follow from the exchangeability of $\{X, \tilde{X}^{(1)}, \ldots, \tilde{X}^{(K)}\}$, the fifth line follows from the law of iterated expectations, and the sixth line follows from the conditional independence of $X$ and $\tilde{X}^{(1)}$.

Substituting (A.10) and (A.11) into (A.9), we get

$$\begin{aligned} & E\ (f\ (X;Z)\quad F^{z})^{2} \\ &= E\ f\ ^{2}(X;Z)\quad 2E[f\ (X;Z)f_{z}(Z)] + \frac{1}{K}E\ f^{2}(X;Z) + \frac{K\quad 1}{K}E\ f_{z}^{2}(Z) \\ &= E\ f\ ^{2}(X;Z)\quad 2E[f\ (X;Z)f_{z}(Z)] + E\ f_{z}^{2}(Z) + \frac{1}{K}\ E\ f^{2}(X;Z)\quad E\ f_{z}^{2}(Z) \\ &= E\ (f\ (X;Z)\quad f_{z}(Z))^{2} + \frac{1}{K}E\ (f\ (X;Z)\quad f_{z}(Z))^{2}\ ; \end{aligned} \tag{A.12}$$

where in the second term in the last line, we use the fact that the cross term $E[f\ (X;Z)f_{z}(Z)] = E[f_{z}^{2}(Z)]$ by the law of iterated expectations.

Now we can simplify the expectation in the second term of (A.8). Expanding, we have

$$\begin{aligned} & E\left(f\ (X;Z)\quad F^{z}\right)^{2} \\ &= E\left[f\ (X;Z)\quad \frac{1}{K}\sum_{k=1}^{K} f\ (\tilde{X}^{(k)};Z)\right]^{2} \\ &= E\ f^{2}(X;Z)\quad \frac{2}{K}\sum_{k=1}^{K} E\left[f\ (X;Z)f\ (\tilde{X}^{(k)};Z)\right] + E\left[\frac{1}{K}\sum_{k=1}^{K} f\ (\tilde{X}^{(k)};Z)\right]^{2} \\ &= E\ f^{2}(X;Z)\quad 2E\left[f\ (X;Z)f\ (\tilde{X}^{(1)};Z)\right] + \frac{1}{K^{2}}E\left[\sum_{k=1}^{K} f\ (\tilde{X}^{(k)};Z)\right]^{2}: \end{aligned} \tag{A.13}$$

We have already found the expectation of the last term here in (A.11). The middle term is very similar to (A.10):

$$\begin{aligned} E\left[f\ (X;Z)f\ (\tilde{X}^{(1)};Z)\right] &= E\left[E\left[f\ (X;Z)f\ (\tilde{X}^{(1)};Z)\right] j\ Z\right] \\ &= E\left[E[f\ (X;Z)\ j\ Z]E\left[f\ (\tilde{X}^{(1)};Z)\right] j\ Z\right] \\ &= E\ f_{z}^{2}(Z)\ : \end{aligned} \tag{A.14}$$

Substituting (A.11) and (A.14) into (A.13), we get

$$\begin{aligned} E\ (f\ (X;Z)\quad F^{z})^{2} &= E\ f\ ^{2}(X;Z)\quad 2E\ f_{z}^{2}(Z) + \frac{1}{K}E\ f^{2}(X;Z) + \frac{K\quad 1}{K}E\ f_{z}^{2}(Z) \\ &= \frac{K+1}{K}\ E\ f^{2}(X;Z)\quad E\ f_{z}^{2}(Z) \\ &= \frac{K+1}{K}E\ (f\ (X;Z)\quad f_{z}(Z))^{2}\ ; \end{aligned} \tag{A.15}$$

where the last line follows for the same reason as in (A.12).

Finally, plugging (A.12) and (A.15) into (A.8), we have

$$\begin{aligned} E[M_i^z] &= E\,(f\ (X;Z) \quad f_z(Z))^2 + \frac{1}{K}E\,(f\,(X;Z) \quad f_z(Z))^2 \\ &\quad \frac{1}{K+1} \quad \frac{K+1}{K} \quad E\,(f\,(X;Z) \quad f_z(Z))^2 \\ &= E\,(f\ (X;Z) \quad f_z(Z))^2 \\ &= \mathrm{MSE}(f_z): \end{aligned}$$

□

## A.4 Proof of Theorem 2.3

Theorem. For i.i.d. samples $f(X_i;Z_i)g_{i=1}^n$, any computational model$f$ and surrogate $f : R^d ! R$, and any $2 (0;1)$, if $E\ f^4(X;Z)\ ;E[f^4(X;Z)] < 1$, then the bounds$L_n$ and $U_n$ output by Algorithm 1 satisfy

$$\liminf_{n!1} P(L_n \quad S \quad U_n) \quad 1 \quad :$$

Proof. We rst handle the case where Var$f(\ (X;Z)) = 0$. Since this implies that $V \overset{a.s.}{=} 0$, Algorithm 1 will return $[0;1]$ with probability 1. Since $S = 0$ by de nition when $\mathrm{Var}(f\ (X;Z)) = 0$, the oodgate interval has 100% coverage.

Now we consider the case where Var$f((X;Z)) > 0$. The rst step in this proof is to show that $M^z$, $M$ and $V$ are all asymptotically normal and consistent for MSE$f(_z)$, MSE$(f)$, and $\mathrm{Var}(f\ (X;Z))$, respectively. Since these estimators are sample means of i.i.d. terms that we have already established are unbiased for their corresponding estimands, it su ces to show that $\mathrm{Var}(M_i^z)\ ;\mathrm{Var}(M_i)\ ;\mathrm{Var}(V_i) < 1$, or equivalently that their second moments are nite, in order to apply the CLT. As in the previous proof, we will drop the index$i$ on random variables withing expectations, since samples are i.i.d.

Starting with $M$, we have

$$\begin{aligned} E\ M^2 &= E\,(f\ (X;Z) \quad f\,(X;Z))^4 \\ &= E\ f^4(X;Z) \quad 4E\ f^3(X;Z)f\,(X;Z) + 6E\ f^2(X;Z)f^2(X;Z) \\ &\quad 4E\ f\ (X;Z)f^3(X;Z) + E\ f^4(X;Z) \\ &< 1\ ; \end{aligned}$$

where the nal inequality holds because the rst and last terms are nite by assumption,

and the middle three terms can be bounded by Hölder's inequality:

$$
\begin{aligned}
E\left[f^3(X,Z)f(X,Z)\right] &\le E\left[f^3(X,Z)f(X,Z)\right] \\
&\le \left[E\left[f^3(X,Z)\right]^{\frac{4}{3}}\right]^{\frac{3}{4}} \left[E\left|f(X,Z)\right|^4\right]^{\frac{1}{4}} \\
&= \left[E\left[f^4(X,Z)\right]\right]^{\frac{3}{4}} \left[E\left[f^4(X,Z)\right]\right]^{\frac{1}{4}} \\
&< 1 \\
0 \le E\left[f^2(X,Z)f^2(X,Z)\right] &\le \left[E\left[f^2(X,Z)\right]^2\right]^{\frac{1}{2}} \left[E\left|f^2(X,Z)\right|^2\right]^{\frac{1}{2}} \\
&= \left[E\left[f^4(X,Z)\right]\right]^{\frac{1}{2}} \left[E\left[f^4(X,Z)\right]\right]^{\frac{1}{2}} \\
&< 1 \\
E\left[f(X,Z)f^3(X,Z)\right] &\le E\left[f(X,Z)f^3(X,Z)\right] \\
&\le \left[E\left|f(X,Z)\right|^4\right]^{\frac{1}{4}} \left[E\left|f^3(X,Z)\right|^{\frac{4}{3}}\right]^{\frac{3}{4}} \\
&= \left[E\left[f^4(X,Z)\right]\right]^{\frac{1}{4}} \left[E\left[f^4(X,Z)\right]\right]^{\frac{3}{4}} \\
&< 1 :
\end{aligned}
$$

Next, we consider $M^z$. Define

$$T_1 := \left(f(X,Z) - F^z\right)^2 \tag{A.16}$$

$$\text{and } T_2 := \left(f(X,Z) - F^z\right)^2; \tag{A.17}$$

so that $M^z = T_1 + \frac{1}{K+1}T_2$. Then,

$$
\begin{aligned}
E\left[M^{z2}\right] &= E\left[T_1^2\right] - \frac{2}{K+1}E\left[T_1T_2\right] + \frac{1}{(K+1)^2}E\left[T_2^2\right] \\
&\le E\left[T_1^2\right] + \frac{1}{(K+1)^2}E\left[T_2^2\right];
\end{aligned} \tag{A.18}
$$

since $T_1, T_2 \ge 0$ almost surely. We first consider $T_1$:

$$
\begin{aligned}
&E\left[T_1^2\right] \\
&= E\left[\left(f(X,Z) - F^z\right)^4\right] \\
&= E\left[f^4(X,Z)\right] - 4E\left[f^3(X,Z)F^z\right] + 6E\left[f^2(X,Z)F^{z2}\right] - 4E\left[f(X,Z)F^{z3}\right] + E\left[F^{z4}\right];
\end{aligned} \tag{A.19}
$$

The first term is finite by assumption, so $E\left[F^{z4}\right] < 1$ is sufficient to show $E\left[T_1^2\right] < 1$, since the middle three terms can all be bounded by Hölder's inequality. Substituting in the expression for $F^z$, we have

$$E\left[F^{z4}\right] = \frac{1}{K^4}E\left[\left(\sum_{k=1}^{K} f(X^{-(k)},Z)\right)^4\right]; \tag{A.20}$$

While we spare the messy algebra here, expanding the right-hand side of (A.20) yields a sum of the terms $E[f^4(X,Z)]$, $E\left[f^3(X^{(1)},Z)f(X^{(2)},Z)\right]$, $E\left[f^2(X^{(1)},Z)f^2(X^{(2)},Z)\right]$, $E\left[f^2(X^{(1)},Z)f(X^{(2)},Z)f(X^{(3)},Z)\right]$, and $E\left[f(X^{(1)},Z)f(X^{(2)},Z)f(X^{(3)},Z)f(X^{(4)},Z)\right]$ with finite coefficients. The first of these expectations is finite by assumption, which also implies the existence of the second expectation by applying Hölder's inequality with $p = \frac{4}{3}$ and of the last three expectations by repeated application of Cauchy–Schwarz. Now we consider $T_2$:

$$\begin{aligned} & E\left[T_2^2\right] \\ &= E\left[(f(X,Z) - F_Z)^4\right] \\ &= E\left[f^4(X,Z)\right] - 4E\left[f^3(X,Z)F_Z\right] + 6E\left[f(X,Z)^2F_Z^2\right] - E\left[f(X,Z)F_Z^3\right] + E\left[F_Z^4\right]. \end{aligned} \tag{A.21}$$

As in (A.19), the first term is finite by assumption, and since we already already established $E\left[F_Z^4\right] < \infty$, we can bound the middle three terms using Hölder's inequality, and thus $E[T_2^2] < \infty$. By (A.18), (A.19), and (A.21), we have $E\left[M_Z^2\right] < \infty$.

Finally, $E[V^2] < \infty$ is immediate given $E\left[f^4(X,Z)\right] < \infty$.

Now, by the multivariate CLT,

$$\sqrt{n}\left(\begin{bmatrix} M_Z \\ M \\ V \end{bmatrix} - \begin{bmatrix} \mathrm{MSE}(f_Z) \\ \mathrm{MSE}(f) \\ \mathrm{Var}(f - f^*(X,Z)) \end{bmatrix}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{12} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{13} & \Sigma_{23} & \Sigma_{33} \end{bmatrix} := \begin{bmatrix} \mathrm{Var}(M_Z) & \mathrm{Cov}(M_Z, M) & \mathrm{Cov}(M_Z, V) \\ \mathrm{Cov}(M_Z, M) & \mathrm{Var}(M) & \mathrm{Cov}(M, V) \\ \mathrm{Cov}(M_Z, V) & \mathrm{Cov}(M, V) & \mathrm{Var}(V) \end{bmatrix}.$$

Since we assume that $\mathrm{Var}(f - f^*(X,Z)) > 0$, we can apply the multivariate delta method to get

$$\sqrt{n}\,\frac{M_Z - M}{V}\,\frac{\mathrm{MSE}(f_Z) - \mathrm{MSE}(f)}{\mathrm{Var}(f - f^*(X,Z))} = \sqrt{n}\left(\hat{\ell}_n(f) - \ell(f)\right) \xrightarrow{d} \mathcal{N}(0, \sigma_\ell^2); \tag{A.22}$$

$$\sqrt{n}\,\frac{M_Z}{V}\,\frac{\mathrm{MSE}(f_Z)}{\mathrm{Var}(f - f^*(X,Z))} = \sqrt{n}\left(\hat{u}_n(f) - u(f)\right) \xrightarrow{d} \mathcal{N}(0, \sigma_u^2) \tag{A.23}$$

where

$$\begin{aligned} \sigma_\ell^2 := \frac{1}{\left(\mathrm{Var}(f - f^*(X,Z))\right)^2} \Bigg( & \Sigma_{11} + \Sigma_{22} + \left(\frac{\mathrm{MSE}(f_Z) - \mathrm{MSE}(f)}{\mathrm{Var}(f - f^*(X,Z))}\right)^2 \Sigma_{33} - 2\Sigma_{12} \\ & - 2\frac{\mathrm{MSE}(f_Z) - \mathrm{MSE}(f)}{\mathrm{Var}(f - f^*(X,Z))}(\Sigma_{23} - \Sigma_{13}) \Bigg) \end{aligned}$$

$$\sigma_u^2 := \frac{1}{\left(\mathrm{Var}(f - f^*(X,Z))\right)^2}\left(\Sigma_{11} - 2\frac{\mathrm{MSE}(f_Z)}{\mathrm{Var}(f - f^*(X,Z))}\Sigma_{13} + \frac{\mathrm{MSE}(f_Z)}{\left(\mathrm{Var}(f - f^*(X,Z))\right)^2}\Sigma_{33}\right).$$

By Slutsky's Theorem, we can replace $\ell$ and $\sigma_u$ in (A.22) and (A.23) with their consistent estimators $s_\ell$ and $s_u$ defined in Algorithm 1, and the same results hold.

Note that

$$L_n = \begin{cases} 0 & \text{if } V = 0; \\ 0 & \text{if } \hat{\ell}_n(f) - \frac{z_{1-\alpha/2} s_\ell}{\sqrt{n}} < 0; \\ \hat{\ell}_n(f) - \frac{z_{1-\alpha/2} s_\ell}{\sqrt{n}} & \text{else,} \end{cases}$$

and since $L_n \leq \ell$ in either of the first two cases (since $\ell \geq 0$ by definition), we have that

$$P(L_n \leq \ell) \geq P\left(\hat{\ell}_n(f) - \frac{z_{1-\alpha/2} s_\ell}{\sqrt{n}} \leq \ell\right), \tag{A.24}$$

where we say that the event $\{\hat{\ell}_n(f) - \frac{z_{1-\alpha/2} s_\ell}{\sqrt{n}} \leq \ell\}$ does not occur when $\hat{\ell}_n(f) - \frac{z_{1-\alpha/2} s_\ell}{\sqrt{n}}$ is undefined (due to $V = 0$). By the same argument,

$$P(U_n \geq u) \geq P\left(\hat{u}_n(f) + \frac{z_{1-\alpha/2} s_u}{\sqrt{n}} \geq u\right). \tag{A.25}$$

Thus, by (A.24) and (A.22),

$$\begin{aligned} \liminf_{n\to\infty} P(L_n \leq \ell) &\geq \liminf_{n\to\infty} P\left(\hat{\ell}_n(f) - \frac{z_{1-\alpha/2} s_\ell}{\sqrt{n}} \leq \ell\right) \\ &\geq \liminf_{n\to\infty} P\left(\hat{\ell}_n(f) - \frac{z_{1-\alpha/2} s_\ell}{\sqrt{n}} \leq \ell(f)\right) = 1 - \frac{\alpha}{2}; \end{aligned}$$

and by (A.25) and (A.23),

$$\begin{aligned} \liminf_{n\to\infty} P(U_n \geq u) &\geq \liminf_{n\to\infty} P\left(\hat{u}_n(f) + \frac{z_{1-\alpha/2} s_u}{\sqrt{n}} \geq u\right) \\ &\geq \liminf_{n\to\infty} P\left(\hat{u}_n(f) + \frac{z_{1-\alpha/2} s_u}{\sqrt{n}} \geq u(f)\right) = 1 - \frac{\alpha}{2}. \end{aligned}$$

A simple union bound gives us the final result:

$$\liminf_{n\to\infty} P(L_n \leq \ell \leq U_n) \geq 1 - \alpha. \qquad \square$$

## A.5 Proof of Theorem 2.4

**Theorem.** Under the same assumptions as in Theorem 2.3 and the additional assumption that $\mathrm{Var}(f_\ell(X, Z)) > 0$, the bounds $L_n$ and $U_n$ output by Algorithm 1 satisfy

$$U_n - L_n \leq \frac{\mathrm{MSE}(f)}{\mathrm{Var}(f_\ell(X, Z))} + O_p\left(n^{-1/2}\right).$$

*Proof.* Note that since $V \xrightarrow{p} \mathrm{Var}(f_\ell(X, Z)) > 0$, $P(V = 0) \to 0$. When $V > 0$,

$$\begin{aligned} (U_n - L_n) - \frac{\mathrm{MSE}(f)}{\mathrm{Var}(f_\ell(X, Z))} &\leq \frac{M z}{\sqrt{V}} + \frac{z_{1-\alpha/2} s_u}{\sqrt{n}} - \frac{M z}{\sqrt{V}} - \frac{z_{1-\alpha/2} s_\ell}{\sqrt{n}} - \frac{\mathrm{MSE}(f)}{\mathrm{Var}(f_\ell(X, Z))} \\ &= \left(\frac{M}{\sqrt{V}} - \frac{\mathrm{MSE}(f)}{\mathrm{Var}(f_\ell(X, Z))}\right) + \frac{z_{1-\alpha/2}(s_u + s_\ell)}{\sqrt{n}}. \end{aligned} \tag{A.26}$$

The reason for the inequality in the first line is that we could have $L_n = 0$ or $U_n = 1$, since we clip the bounds. Using the intermediate results from the proof of Theorem 2.3 and our assumption that $\text{Var}(f(X, Z)) > 0$, we can apply the multivariate delta method to establish that

$$\sqrt{n}\left(\frac{M}{V} - \frac{\text{MSE}(f)}{\text{Var}(f(X, Z))}\right) \xrightarrow{d} N(0, \sigma^2),$$

for some $\sigma^2 < 1$. This implies that the first term in (A.26) is $O_p(n^{-1/2})$. Since $s_u \xrightarrow{p} \sigma_u$ and $s_{\cdot} \xrightarrow{p} \sigma_{\cdot}$, $s_u$ and $s_{\cdot}$ are $O_p(1)$, meaning that the second term in (A.26) is also $O_p(n^{-1/2})$. This establishes the desired result.

□

## A.6 Description of Hymod Model

Hymod is a conceptual rainfall-runoff model based on the probability-distributed soil storage capacity principle (Sheikholeslami and Razavi, 2020). It simulates daily streamflow given precipitation and potential evapotranspiration as forcing data, where both the inputs and outputs are time series. The model consists of a soil moisture module and a routing module composed of three linear quickflow reservoirs and one linear slowflow reservoir. The outputted streamflow is given by the sum of quick and slowflow generation (Herman et al., 2013).

| Input | Description | Units | Min | Max |
|---|---|---|---|---|
| Sm | maximum soil moisture | mm | 0 | 400 |
| beta | exponent in the soil moisture routine | - | 0 | 2 |
| alfa | partition coefficient | - | 0 | 1 |
| Rs | slow reservoir coefficient | day$^{-1}$ | 0 | 0.1 |
| Rf | fast reservoir coefficient | day$^{-1}$ | 0.1 | 1 |

Table 1: Hymod inputs (Pianosi and Wagener, 2016)