A Fresh Look at Variography: Measuring Dependence and Possible Sensitivities Across Geophysical Systems from Any Given Data

R. Sheikholeslami[1,2,3] and S. Razavi[2,3,4]

[1]School of Geography and the Environment, University of Oxford, Oxford, UK
[2]School of Environment and Sustainability, University of Saskatchewan, Saskatoon, Canada
[3]Global Institute for Water Security, University of Saskatchewan, Saskatoon, Canada
[4]Department of Civil, Geological, and Environmental Engineering, University of Saskatchewan, Saskatoon, Canada

Key Points:

- Develops an efficient "data-driven" method for global sensitivity analysis (GSA) based on principles of variography.
- Enables assessing relationship strength, either causal or correlational, in geophysical systems based on any given data.
- Shows theoretical links with previous GSA methods and demonstrates robust performance even when the data size is very small.

[1]Corresponding author: Razi Sheikholeslami (razi.sheikholeslami@ouce.ox.ac.uk)

Abstract

Sensitivity analysis in Earth and environmental systems modelling typically demands an onerous computational cost. This issue coexists with the reliance of these algorithms on ad-hoc designs of experiments, which hampers making the most out of the existing datasets. We tackle this problem by introducing a method for sensitivity analysis, based on the theory of variogram analysis of response surfaces (VARS), that works on any sample of input-output data or pre-computed model evaluations. Called data-driven VARS(D-VARS), this method characterizes the relationship strength between inputs and outputs by investigating their covariograms. We also propose a method to assess 'robustness' of the results against sampling variability and numerical methods' imperfectness. Using two hydrologic modelling case studies, we show that D-VARS is highly efficient and statistically robust, even when the sample size is small. Therefore, D-VARS can provide unique opportunities to investigate geophysical systems whose models are computationally expensive or available data is scarce.

Plain Language Summary

Sensitivity analysis (SA) is about assessing how the properties of a system are influenced by different factors. It can also help us better understand the behavior of a mathematical model and the underlying real-world system that it mimics. Almost always, classic SA estimates the sensitivities by sampling the entire problem space in a specific manner. They are incapable of using a pre-existing set of input-output data or pre-computed model evaluations. Hence, classic SA becomes useless in cases where a sample of input-output data, obtained from physical experiments or computationally-expensive simulations, is already available. We propose a purely data-driven method which can effectively be used in such situations. Based on the principles of variography, our method measures dependence and possible sensitivities across a system from any given data. Here, we illustrate our method in the context of hydrologic modelling, but it can potentially be applied to study other geophysical systems models.

1 Introduction

Understanding how various uncertain factors influence the behavior of Earth and environmental systems (EES) models has greatly raised the need for continued development of the efficient methodologies for global sensitivity analysis (GSA). GSA methods identify dominant factors that exert a considerable impact on the model responses, and accordingly provide invaluable information for model simplification, risk assessment, and uncertainty analysis (Castelleti et al., 2012; Safta et al., 2015; Markstrom et al., 2016; Bhalachandran et al., 2019; Janetti et al., 2019; Li et al., 2019; Puy et al., 2020).

Almost all GSA methods are "sampling-based", starting by sampling from a d-dimensional factor/input space using design of experiments approaches. Next, the corresponding response variable of interest at all sample points are determined from an output space through an input-output relationship, i.e., a *model*. A statistical estimator can then be employed to compute the sensitivity indices, which are essentially a representation of *relationship strength* between the output and individual inputs. However, two major issues preclude efficient application of the sampling-based GSA:

- They are bound to their own sampling strategies that follow specific spatial arrangement (the "ad hoc designs" as termed by Ratto et al. (2007)).
- They are not applicable when the underlying input-output functional relationship is unavailable (e.g., we may only have a sample of input-output data obtained from field observations or previous modelling experiments, and nothing more).

The former issue is widespread among GSA methods (e.g., GSA methods of Saltelli et al. (2010); Pianosi and Wagener (2015); Razavi and Gupta (2016a,b)), and therefore, a sample taken by one method cannot be utilized by another method. Furthermore, this complicates comparison of the different methods. And due to the latter, GSA has remained limited to cases where a computational model is available, and its run-time is tractable to generate a large enough input-output sample. Hence, the mainstream methods are not useful when the distributions, correlations, and interactions between different factors need to be characterized based on the existing dataset, without model (re-)evaluations. These challenges necessitate a "given-data approach" to GSA (alternatively termed as "data-driven approach" or "post-processing GSA", see Borgonovo et al., (2016); Li and Mahadevan (2016); Plischke (2010)) that can extract sensitivity-related information from pre-existent datasets. In a burgeoning era of big data, this becomes more crucial than ever because data collection can be much easier and faster than building models.

A limited number of GSA methods can potentially be used under the given-data approach. These are mainly emulator-based techniques, limited to estimating variance-based sensitivity indices using either Monte Carlo-based method (see, e.g., Marseguerra et al., 2003; Iooss et al., 2006; Volkova et al., 2009; Storlie et al., 2009) or analytical approaches (see, e.g., Sudret, 2008; Marrel et al., 2009; Jourdan, 2012; Jia and Taflanidis, 2016; Shi et al., 2019; Sargsyan et al., 2019). Additionally, moment-independent methods can also be used in the given-data setting (Borgonovo et al. 2012; Plischke et al., 2013; Borgonovo et al. 2017; Yun et al., 2018). Each of these methods characterizes only a specific feature of the response surface and ignores the rest (Razavi and Gupta, 2015). Recognizing this fact, some researchers have recently advocated the use of multiple GSA measures together, which may be based in different theories, at the cost of augmenting computational burden (Wang and Solomatine, 2019; Borgonovo et al., 2017; Pianosi et al., 2016). We should note that, in theory, the given-data paradigm provides a unique platform to estimate various GSA measures simultaneously from the same dataset, without much additional computational effort.

Here, we contribute to the given-data paradigm by developing a highly efficient and robust data-driven GSA, based on the theory of variogram analysis of response surfaces (VARS) (Razavi and Gupta, 2016a,b). Despite the increasing popularity of VARS (e.g., Yassin et al., 2017; Krogh et al., 2017; Sheikholeslami et al., 2017; Akomeah et al., 2019; Schürz et al., 2019; Leroux and Pomeroy, 2019; Lilhare et al., 2020; Becker, 2020; Korgaonkar et al., 2020), its current version cannot be applied in the given-data setting. Conversely, our new version of VARS works on any given data, by approximating the anisotropic variogram structure of the underlying (but unknown) response surface, when only a (small) sample of the input-output space is available. We also address a crucial, but *often neglected*, component of any GSA practice, which is assessing the "robustness" of its results. Note that GSA results are typically prone to statistical uncertainty due to sampling variability and to imperfect nature of numerical methods. However, a comprehensive robustness assessment can be computationally costly or infeasible in the given-data context. Therefore, we develop a new robustness index that works well within the given

data setting. We note that typical approaches to assess robustness based on bootstrap (Efron, 1982) have limited applicability in some given-data GSA methods, because of ill-conditioning of covariance matrix caused by samples with non-unique members (i.e., when the two or more sample points are identical).

2 Method

2.1 Background

VARS is a new GSA framework that builds on anisotropic variograms to quantify the influence of input factors on the variability of response variables. Directional variograms are a rich source of information to attribute the structure and spatial variability of a response variable to the distributional properties of different factors. VARS recognizes that the variability of any continuous response surface can be better expressed by the variance of change in the response variable when a factor or group of factors are perturbed with varying perturbation sizes, across the factor space. In other words, for any pairs of sample points in the factor space, e.g., X^u and X^w, the variance of the difference between the corresponding response variables, $y(X^u)$ and $y(X^w)$, depends on their distance $\|X^u - X^w\| = h$ in the *d*-dimensional input space R^d, i.e.:

$$var[y(X^u) - y(X^w)] = E[(y(X^u) - y(X^w))^2] = 2\gamma(h) u, w = 1,2,\ldots,m \quad (1)$$

where $h = \{h_1, h_2, \ldots, h_d\}$ denotes the vector that separates two sample points with respect to distance and direction, and $\gamma(h)$ is one half of the expected squared difference between $y(X^u)$ and $y(X^w)$.

If the stationarity assumption holds (Matheron, 1971), the function that relates γ to h, known as (semi-)variogram, can be approximated by:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{l=1}^{N(h)} \{y(X^l) - y(X^l + h)\}^2 \quad (2)$$

where $N(h)$ is the number of all pairs of the sample points in the input space separated by the distance vector h. The vector h is also referred to as "perturbation scale" in VARS terminology. An example of a response surface and the estimated variogram surface is shown in **Figure S1**.

Using **Eq. (2)**, directional variograms along the j-th input factor $(j = 1,2,\ldots,d)$ can be estimated by calculating $\gamma(h_j)$ for any given set of h_j, for example $\{0, \Delta h, 2\Delta h, \ldots\}$ with perturbation resolution of Δh (Razavi and Gupta, 2016b). The directional variograms, therefore, show how the variability of the response variable is changing with respect to the direction and perturbation scale. Accordingly, VARS-based sensitivity analysis links the rate of variability to both direction and perturbation scale (for detail see Razavi and Gupta (2016a)).

Finally, to measure factor sensitivities, given perturbation scales ranging from zero to H_j, VARS defines a set of sensitivity indices for the j-th factor, called integrated variograms across a range of scales (IVARS), as follows:

$$\Gamma(H_j)=\int_0^{H_j}\gamma(h_j)\,dh_j \quad (3)$$

2.2 New extension of VARS: A given-data estimator

We extend the theory of VARS and propose a data-driven estimator, called D-VARS. The stationarity assumption (Matheron, 1971) implies that the variogram can be related to spatial covariance of the response variable according to the following equation (see **Text S1** in the **Supporting Information**):

$$\gamma(h)=var[y]-cov\left[y(X^u),y(X^w)\right]=\sigma^2-C(h) \quad (4)$$

where σ^2 and $C(h)$ are the variance and spatial covariance function of the $y(X)$, respectively.

Hence, for any given covariance function, the variogram of the response variable is uniquely constructible from the covariance function, which must be a symmetric, positive definite function of h (Rasmussen and Williams, 2006). To characterize the covariance functions in D-VARS, we assume a zero mean Gaussian Process (GP) throughout this study. In the case of GP, the covariance function can be written as (Sacks et al., 1989; Jones 2001):

$$C(h)=\sigma^2 R(h) \quad (5)$$

where $R(.)$ is the correlation function that provides spatial correlation properties.

We focus on correlation functions that can be defined as a product of one-dimensional kernel functions, i.e., r_j:

$$R(X^u,X^w)=R(h)=\prod_{j=1}^{d} r_j(h_j),\ j=1,2,\ldots,d \quad (6)$$

Following equations (**4-6**), the variogram can be obtained by:

$$\gamma(h)=\sigma^2(1-R(h))=\sigma^2\left(1-\prod_{j=1}^{d} r_j(h_j)\right) \quad (7)$$

By substituting **Eq. (7)** in **Eq. (3)**, D-VARS estimates the IVARS sensitivity indices for any perturbation scale from zero to H_j (for the j-th factor), as follows:

$$\Gamma(H_j)=\int_0^{H_j}\sigma^2\left(1-r_j(h_j)\right)dh_j \quad (8)$$

Theoretically, the one-dimensional correlation functions, r_j, can be learned purely from input-output data. With the covariance functions distilled from data, D-VARS directly estimates

several types of sensitivity indices using **Eq. (8)**, and thus can be used to perform GSA without the need to resort to a particular sampling strategy or re-running the model. We discussed numerical implementation of D-VARS in **Text S2** of the **Supporting Information**. Moreover, inspired by Sheikholeslami et al. (2019), we defined a robustness index (see **Text S2**) for evaluating how robust D-VARS is while being tested on the given data setting.

3 Numerical Experiments

3.1 Computer experiments versus physical experiments

We argue that D-VARS is applicable to both computer experiments and physical experiments. For the former, D-VARS contributes to more efficient and sampling-free GSA, while for the latter, it provides new opportunities to assess the strength of relationships, either causal or correlational, between different variables measured in an experiment. There are two major differences between these two types of experiments: computer experiments are usually deterministic, while data collected from physical experiments are prone to noise or errors, often with unknown properties; in computer experiments, one may have full control on the experimental design for collecting samples and their distributional properties, while it is typically not the case in physical experiments.

In this paper, we chose to test D-VARS on computer experiments for the following reasons. First, the full properties of the underlying relationships were available through computer models and therefore we could benchmark our results against the 'true' dependencies between different variables in question. Second, we could replicate our experiment many times, by resampling from the models with increasing sample sizes, to assess the convergence and robustness of the new method. While the outcome of our tests here can, in part, be generalized to physical experiments, a rigorous study is required to test D-VARS to cases where datasets are polluted with noise and variables follow a variety of distributional forms. This will be the topic of our follow-up paper.

3.2 Hydrologic models used

We designed two case studies with two well-established hydrologic models of increasing complexity (HYMOD and HBV-SASK) to illustrate the utility of D-VARS. The HYMOD model with five parameters was configured for the Leaf River catchment, a 1950 km^2 catchment located north of Collins, Mississippi, USA. The HBV-SASK model with 12 parameters was configured for the Oldman River basin, a catchment of 1435 km^2 located in western Canada. The structure of the models, forcings, and parameters are described in **Text S3** of the **Supporting Information**. We chose these classic models as our testbed, because of their high computational efficiency and that they have already been extensively used in GSA studies (Song et al., 2015). In future work, we will test how the performance of D-VARS will scale to models with varying complexity when the dimensionally of the model increases (see Sheikholeslami et al., 2019; Liu et al., 2020).

3.3 Synthetic input-output datasets for D-VARS

In the first case study, we chose the five parameters of HYMOD as the inputs and a goodness-of-fit metric to observations, Nash-Sutcliffe Efficiency (Nash and Sutcliffe, 197), as the output. This case study was designed to represent a widely adopted GSA setting for parameter screening, which can inform the process of *calibration* (Gupta and Razavi, 2018). The second case study, instead, was made as an example of more modern applications of GSA, where the purpose is *learning* about the system behavior under uncertainty and non-stationarity (Razavi et al., 2020), regardless of the quality of fit to observations (Razavi and Gupta, 2019; Do and Razavi, 2020). Therefore, the 12 parameters of HBV-SASK were chosen as the inputs, while the output was the model's estimate of flood peak with the 10-year return period. To compute the 10-year flood peak for each model run (under a different parameter set), we fitted a generalized extreme value (GEV) distribution to annual maximum peak discharges extracted from the simulated streamflow times series over the historical period. This case study might also be viewed as an example for flood frequency analysis in ungauged, where one seeks to know which uncertain parameters control the uncertainty in flood estimates the most. We note that these case studies are for demonstration only, and therefore, decisions like choosing GEV were rather arbitrary. More details of these case studies are available in **Supporting Information**.

3.4 D-VARS runs

We first ran the original VARS for both HYMOD and HBV-SASK case studies using the sampling-based STAR-VARS algorithm (Razavi and Gupta, 2016b). A large sample size was chosen, resulting in ~70,000 model evaluations for each case study, to ensure convergence to robust and accurate results. These results were deemed to be the '*true*' sensitivities and used as the comparison benchmark for the performance assessment of D-VARS.

Second, we generated synthetic datasets for the assessment of D-VARS by randomly sampling from the input space, with progressively increasing sample size (i.e., 20, 50, 100, 200, and 400 sample points) via progressive Latin hypercube sampling (Sheikholeslami and Razavi, 2017). We then ran all the sample sets through the models to obtain the respective outputs. Each input-output sample set was considered to be a set of 'given-data'. We replicated this data generation process 100 times with different initial random seeds to assess both average and variability of the D-VARS behavior.

Furthermore, IVARS-50, called 'total-variogram effect' (Razavi et al. 2019), was applied to assess the factor sensitivities, which means that **Eq. (8)** was computed for H_j=¿ '50% of the parameter range'. IVARS-50 indicates the overall contribution of a parameter, including its interaction with other parameters, to variability of the output. The 100 independent replicates of D-VARS on independent sample sets also helped us comprehensively assess the robustness of D-VARS against sampling variability. A GSA algorithm is perfectly robust if independent replicates of the algorithm with different samples converge to identical results.

4 Results and Discussion

4.1 Factor sensitivities and actual robustness of D-VARS

Figures 1 and **2** present the histograms of the HYMOD and HBV-SASK parameter rankings for different input-output sample sizes. In these figures, Rank1 stands for the most influential parameter on the output. In other words, the dependency of the output to the Rank1 parameter is the highest. The true ranking of the HYMOD and HBV-SASK parameters are as follows {$\boldsymbol{R_q}$, $\boldsymbol{C_{max}}$, ***alpha***, $\boldsymbol{b_{exp}}$, $\boldsymbol{R_s}$} (see **Table S1**) and {***PM***, ***C0***, ***FRAC***, ***TT***, ***FC***, ***a***, ***UBAS***, ***K1***, ***EFT***, ***LP***, ***K2***, ***beta***} (see **Table S2**) from the most influential to the least influential one.

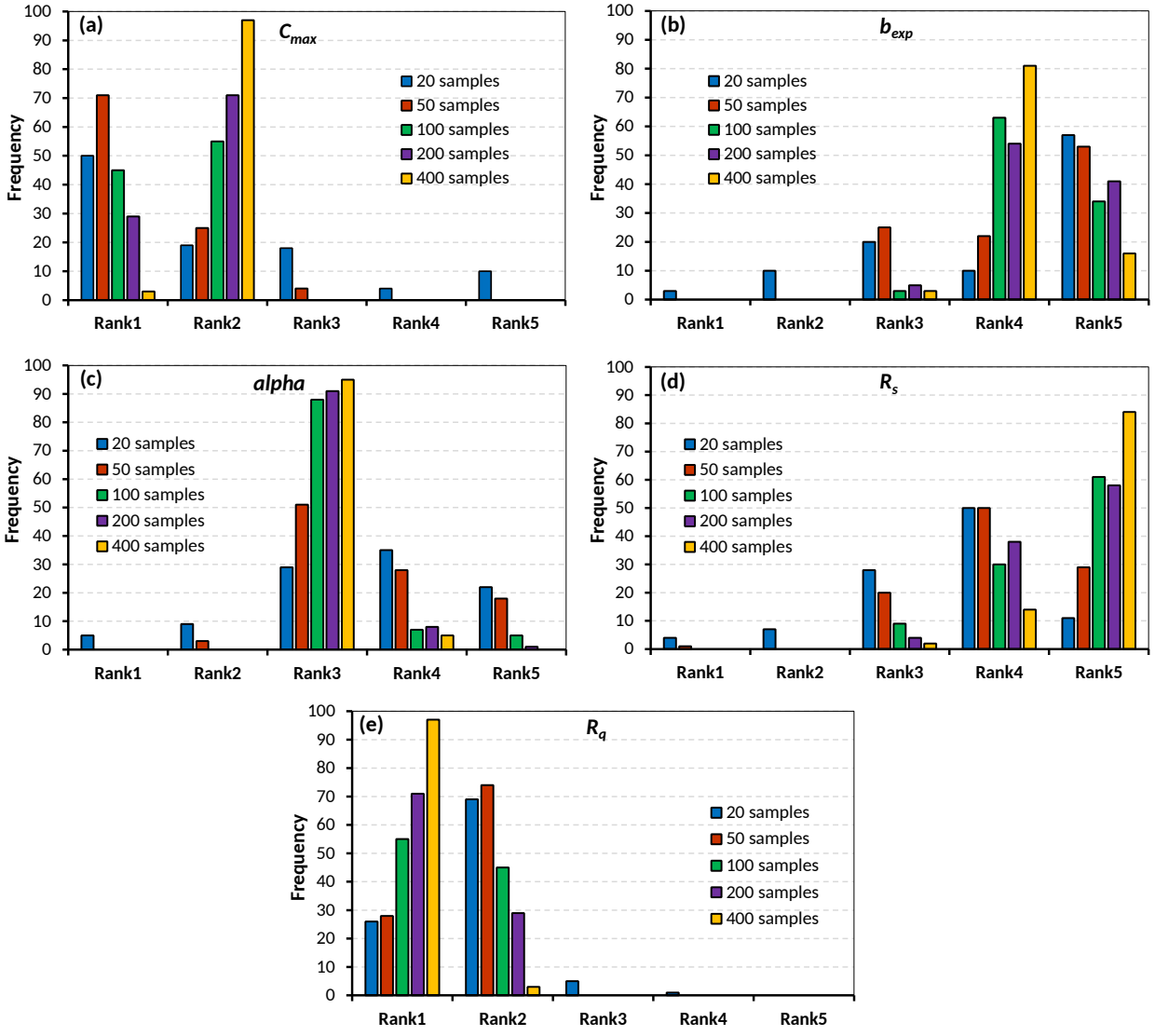

Figure 1. Parameter importance rankings calculated by D-VARS for the HYMOD model. Each subplot shows histograms of the parameter rankings obtained by an increasing sample size of given data, each with 100 replicates with different initial random seeds. For example, in subplot

(a), when the sample size is 400, 98 of 100 replicates indicate C_{max} is Rank2 while the remaining two replicates indicate it is Rank1.

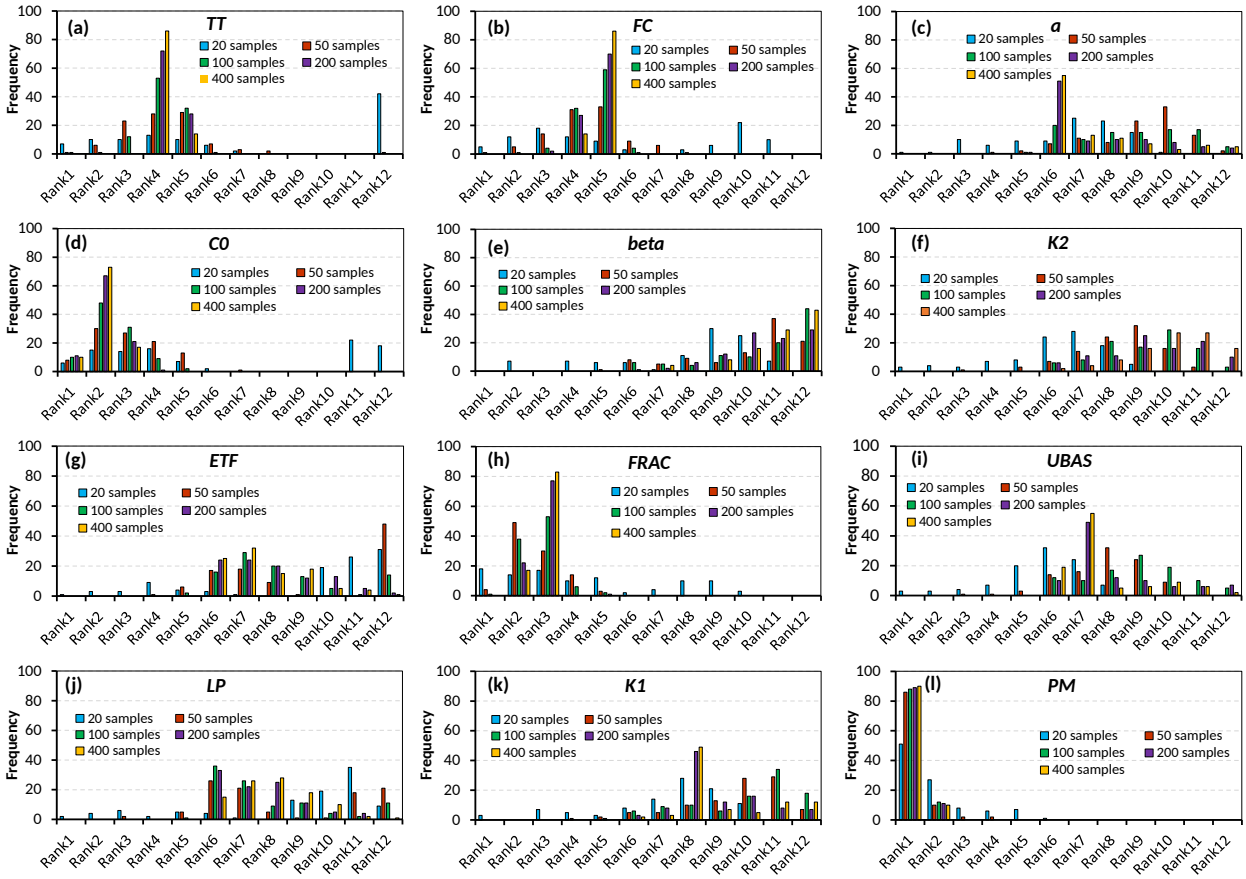

Figure 2. Parameter importance rankings calculated by D-VARS for the HBV-SASK model. Each subplot shows histograms of the parameter rankings obtained by an increasing sample size of given data, each with 100 replicates with different initial random seeds.

As shown in **Figures 1** and **2**, rankings of the most influential parameters for HYMOD, {R_q, C_{max}, ***alpha***}, and HBV-SASK, {***PM***, ***C0***, ***FRAC***, ***TT***, ***FC***}, have been well established even when the size of given data was 50. This confirms that D-VARS is extremely efficient in identifying the most influential parameters when the sample size is very small. Furthermore, **Figure 2** shows that the ranking of parameters with moderate importance in HBV-SASK, {α, ***UBAS***, ***K1***}, stabilized when the sample size was larger than 200. A close examination of the results reveals that in more than 50 replicates, the ranking of these parameters converged to the true ranking when the sample size is only 200. It is noteworthy that parameters with the least influence on HBV-SASK, {***EFT***, ***LP***, ***K2***, ***beta***}, have the slowest convergence rate in terms of ranking.

Now let us directly look at actual sensitivity indices, IVARS-50, derived by D-VARS. **Figures 3** and **4** show IVARS-50 values (scaled between zero and one) for the HYMOD and HBV-SASK

parameters, obtained from the 100 replicates of each experiment. For an extremely small sample size (i.e., 20), D-VARS showed highly variable performance. However, by increasing sample size, all the replicates quickly converged to a single set of IVARS-50 (i.e., true values), particularly for the most influential parameters. This confirms the robustness of D-VARS against sampling variability. For the least influential parameters, the IVARS-50 values may not be distinguishable across all the replicates, even for the larger sample sizes. This is mainly because these parameters are almost non-influential on the output of interest and their associated IVARS-50 values are close to zero.

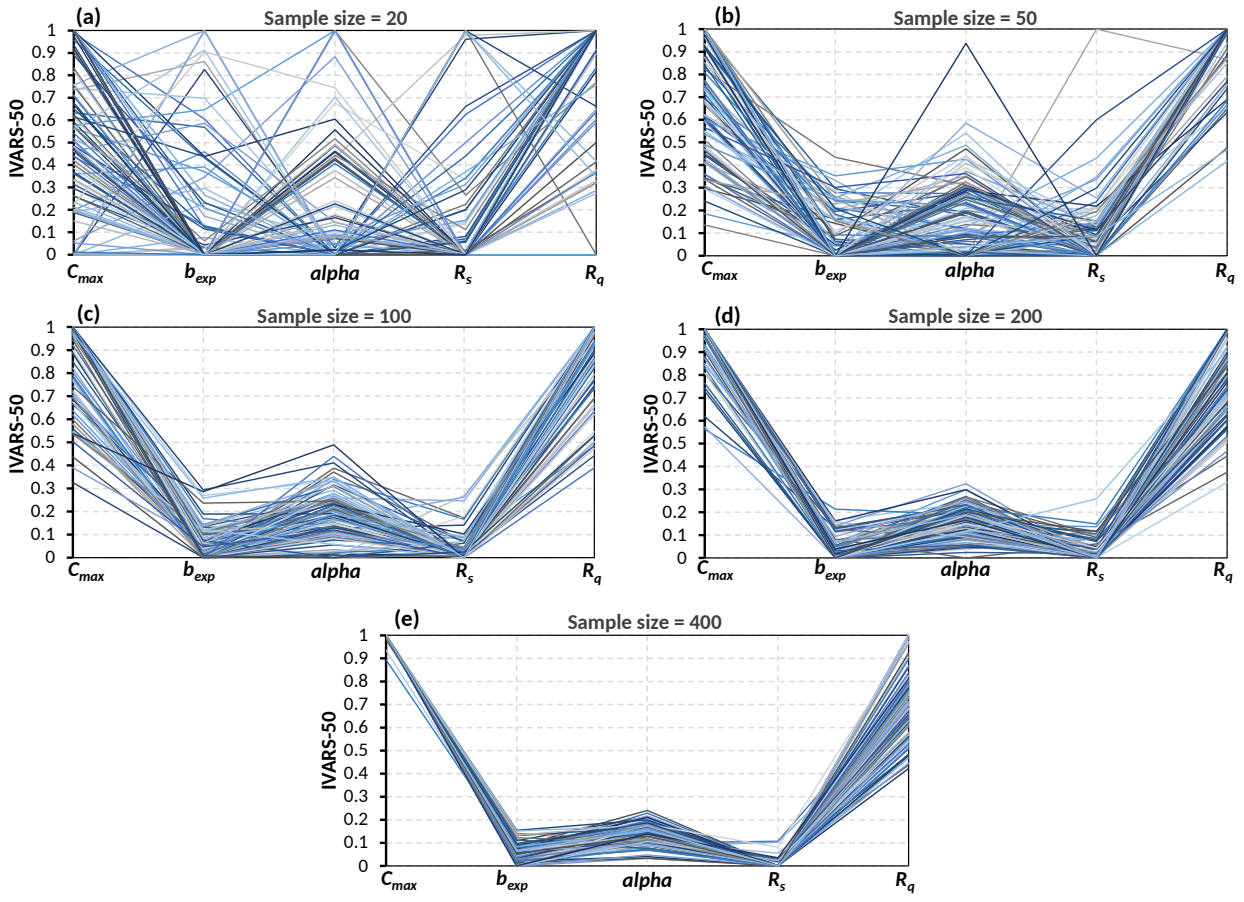

Figure 3. Sensitivity indices for the HYMOD parameters obtained by D-VARS. Subplots show the IVARS-50 values from given data with different sample sizes for 100 replicates of the experiment.

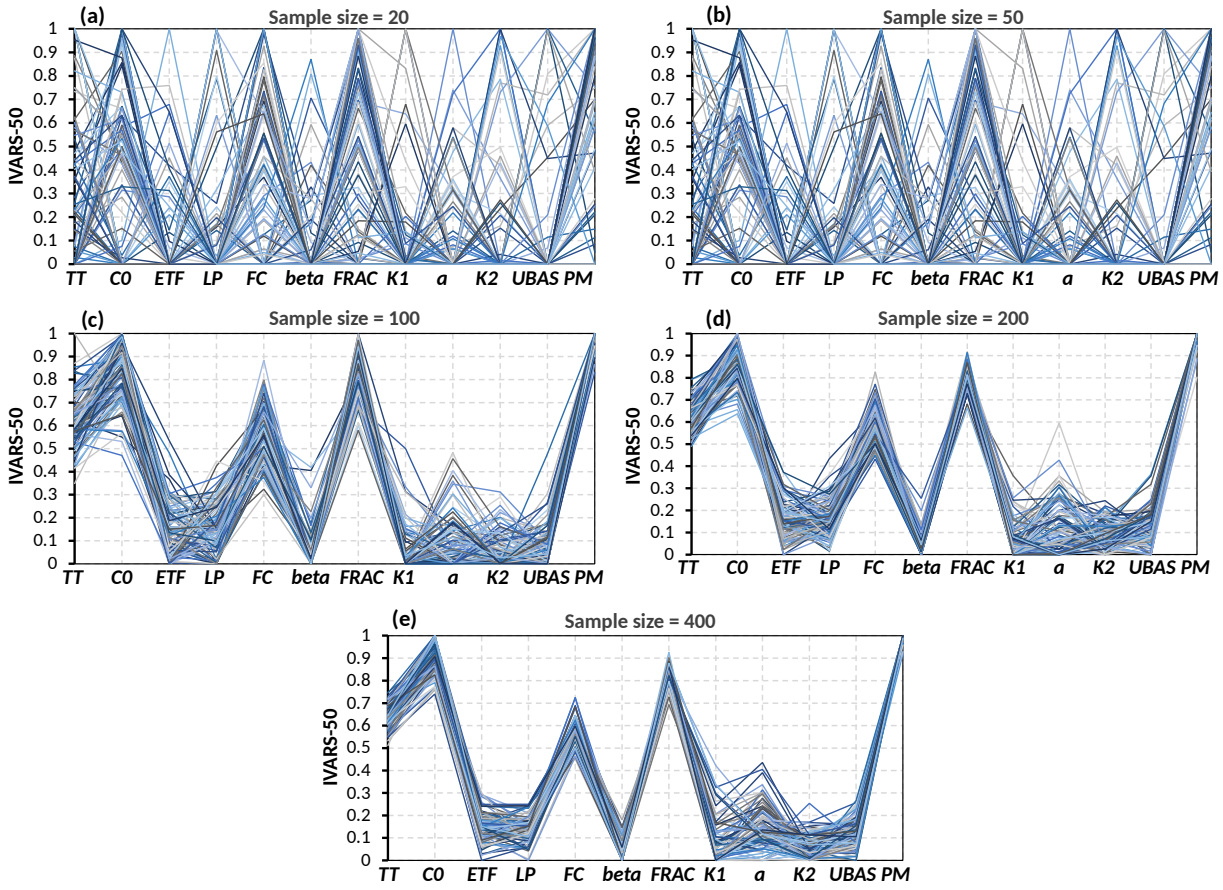

Figure 4. Sensitivity indices for the HBV-SASK parameters obtained by D-VARS. Subplots show the IVARS-50 values from given data with different sample sizes for 100 replicates of the experiment.

4.2 Physical justifiability of sensitivity assessments

A critical question that one might ask after running GSA is whether the results are physically meaningful. Based on our results, we can (rather subjectively) categorize the parameters of HYMOD into two groups, i.e., influential: {$\boldsymbol{R_q}$, $\boldsymbol{C_{max}}$, ***alpha***}, and non-influential: {$\boldsymbol{b_{exp}}$, $\boldsymbol{R_s}$}, and those of HBV-SASK into three groups, (i) strongly influential: {***PM***, ***C0***, ***FRAC***, ***TT***, ***FC***}, (ii) moderately influential {***a***, ***UBAS***, ***K1***}, and (ii) non-influential: {***EFT***, ***LP***, ***K2***, ***beta***}. Recall that for HYMOD, these assessments are based on choosing NS and 10-year flood estimates as the outputs for the HYMOD and HBV-SASK case studies, respectively. We know from hydrology domain knowledge that both of the outputs should, in principle, be dominantly controlled by high flows in the hydrograph.

In case of HYMOD, D-VARS ranked $\boldsymbol{R_q}$, a parameter controlling the quick flow generation, as the most influential parameter. This assessment is physically justifiable, as the Leaf River basin is a rainfall-dominated basin with a history of torrential storms. Most influential parameters of HBV-SASK, however, are those mainly responsible for the snowmelt (***C0*** and ***TT***) and soil

(***FRAC*** and ***FC***) processes. The high influence of ***C0*** and ***TT*** can be justified because in the Oldman River basin major floods are governed by snow accumulation and melt in early spring. From the structure of both models, it is evident that ***alpha*** (in HYMOD) and ***FRAC*** (in HBV-SASK) determine the fraction of soil release entering fast reservoir, and accordingly play a significant role on the high flow values. Moreover, C_{max} (in HYMOD) and ***FC*** (in HBV-SASK) account for partitioning of the precipitation into runoff and soil moisture, and thus can significantly impact the simulated high flows. Additionally, our method recognized ***α*** and ***UBAS*** among the influential parameters for peak flow generation in HBV-SASK, since they control timing and attenuation of the release from the fast reservoir. Finally, D-VARS identified R_s (in HYMOD) and ***K2*** (in HBV-SASK) as the non-influential parameters. These parameters represent the release pace of slow reservoir in the structure of these models, and as such, are only responsible for base flows (minimally contributing to peak flows).

4.2 The proposed robustness index

The assessment of actual robustness, as presented in Section 4.1, is typically infeasible in practice, because only a single data set is often available or the model under investigation is computationally intensive to generate multiple independent sets of input-output data. This study, therefore, proposed a novel robustness index that can estimate the robustness of D-VARS for any given data. The performance of this new index for the synthetic samples taken from the HYMOD and HBV-SASK models across 100 replicates are depicted in **Figure 5**.

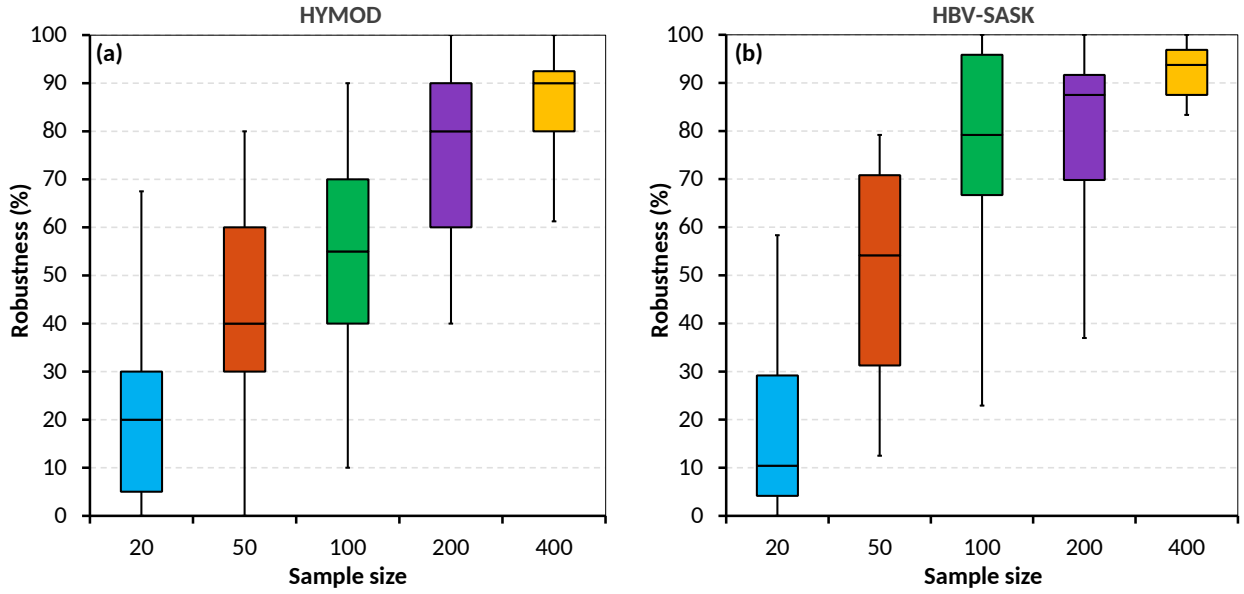

Figure 5. Robustness assessment results for the **(a)** HYMOD and **(b)** HBV-SASK models. Boxplots represent the distribution of robustness index across the 100 replicates. Robustness index varies between 0 to 100%, with the latter corresponding to perfectly robust results.

As shown in **Figure 5**, by increasing the sample size, the variability of the robustness index obtained over all replicates of the algorithm became lower, and consequently the robustness of D-VARS became higher. Also, for both models, the robustness indices quickly converged towards 100% (i.e., perfect robustness). When the sample size was larger than 100, almost all the robustness indices were higher than 50%. For clarity, see the medians of the estimated robustness indices (the horizontal black lines in **Figure 5**). Notice that the medians of the robustness indices were already quite high when the sample size was 100, i.e., 55% for HYMOD and 79% for HBV-SASK, and increased rapidly thereafter to be 90% for HYMOD and 93.7% for HBV-SASK at 400 samples.

5 Conclusions

GSA has almost always been tied to *ad-hoc* experimental designs and defined in the context of mathematical models. This study tried to take GSA to a next level by proposing one of the first methods to conduct GSA for any given data, independent of experimental designs and mathematical models. Our proposed method, called D-VARS, was based on variography via Gaussian process modelling to characterize the spatial correlation properties of the underlying response surface for estimating sensitivity/dependence indices. D-VARS is not only a computationally cheap method that works with any given data, but also makes, in principle, less confining assumptions than most of the existing sampling-based methods. For example, many GSA methods assume that the input data themselves are uncorrelated (see Do and Razavi, 2020), while D-VARS handles correlated inputs as well (not shown in this paper). We examined the performance of the method using two benchmark hydrologic case studies. Overall, our results demonstrated that D-VARS accurately estimates the true sensitivity measures with very small sample sizes.

The effectiveness and high-efficiency of D-VARS make it uniquely positioned to advance GSA paradigm on two fronts: (i) D-VARS can open up a new area of research, where GSA is applied to any data set, even when the underlying relationship and mechanisms are not known; (ii) D-VARS can enable GSA for computationally expensive models, wherein conventional GSA is handicapped, as D-VARS requires minimal computational effort to produce robust results. Further, we argued that any GSA practice must be accompanied by the assessment of the robustness, which is typically neglected in the literature. To this end, we proposed a robustness index and showed that it can consistently provide an accurate evaluation of robustness. Our proposed index can be easily used in practice since it works when only a single (albeit small) dataset is available. Future work may include testing our proposed method on real-world data obtained from field observations or remote sensing to better support model development and understanding.

Data Availability

Datasets for this research is available through Razavi et al. (2019).

References

Abrahamsen, P. (1994), A review of Gaussian random fields and correlation functions. Tech. Rep. 878, Norsk Regnesentral.

Akomeah, E., Lindenschmidt, K.E. & Chapra, S.C. (2019), Comparison of aquatic ecosystem functioning between eutrophic and hypereutrophic cold-region river-lake systems. *Ecological Modelling*, *393*, 25–36. https://doi.org/10.1016/j.ecolmodel.2018.12.004

Bhalachandran, S., Haddad, Z. S., Hristova−Veleva, S. M., & Marks, F. D. Jr. (2019), The relative importance of factors influencing tropical cyclone rapid intensity changes. *Geophysical Research Letters*, *46*, 2282–2292. https://doi.org/10.1029/2018GL079997

Becker, W., 2020. Metafunctions for benchmarking in sensitivity analysis. *Reliability Engineering & System Safety*, *204*, 107189. https://doi.org/10.1016/j.ress.2020.107189

Bergström, S. (1995), The HBV model. Computer models of watershed hydrology, edited by V.P. Sing. Water. Resour. Publ. 443–476.

Boggs, P.T. & Tolle, J.W. (2000), Sequential quadratic programming for large-scale nonlinear optimization. *Journal of Computational and Applied Mathematics*, *124(1–2)*, 123–137. https://doi.org/10.1016/S0377-0427(00)00429-5

Borgonovo, E., Castaings, W. & Tarantola, S. (2012), Model emulation and moment-independent sensitivity analysis: An application to environmental modelling. *Environmental Modelling & Software*, *34*, 105–115. https://doi.org/10.1016/j.envsoft.2011.06.006

Borgonovo, E., Hazen, G.B. & Plischke, E., (2016), A common rationale for global sensitivity measures and their estimation. *Risk Analysis*, *36(10)*, 1871–1895. https://doi.org/10.1111/risa.12555

Borgonovo, E., Lu, X., Plischke, E., Rakovec, O. & Hill, M.C. (2017), Making the most out of a hydrological model data set: Sensitivity analyses to open the model black−box. *Water Resources Research*, *53(9)*, 7933–7950. https://doi.org/10.1002/2017WR020767

Boyle, D.P., Gupta, H.V. & Sorooshian, S. (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research*, *36 (12)*, 3663–3674. http://dx.doi.org/10.1029/2000WR900207

Butler, A., Haynes, R.D., Humphries, T.D. & Ranjan, P. (2014), Efficient optimization of the likelihood function in Gaussian process modelling. *Computational Statistics & Data Analysis*, *73*, 40–52. https://doi.org/10.1016/j.csda.2013.11.017

Castelletti, A., Galelli, S., Ratto, M., Soncini-Sessa, R. & Young, P.C. (2012), A general framework for dynamic emulation modelling in environmental problems. *Environmental Modelling & Software*, *34*, 5–18. https://doi.org/10.1016/j.envsoft.2012.01.002

Chilès, J.-P. & Delfiner, P. (1999), Geostatistics: Modeling Spatial Uncertainty. Wiley, New-York.

Do, N. C. & Razavi, S. (2020), Correlation effects? A major but often neglected component in sensitivity and uncertainty analysis. *Water Resources Research*, *56*, e2019WR025436. https://doi.org/10.1029/2019WR025436

Efron, B. (1982). *The Jackknife, the bootstrap and other resampling plans*, SIAM, Philadelphia.

Gupta, H.V. & Razavi, S. (2018), Revisiting the basis of sensitivity analysis for dynamical earth system models. *Water Resources Research*, *54(11)*, 8692–8717. https://doi.org/10.1029/2018WR022668

Iooss, B., Van Dorpe, F. & Devictor, N. (2006), Response surfaces and sensitivity analyses for an environmental model of dose calculations. *Reliability Engineering & System Safety*, *91(10-11)*, 1241–1251. https://doi.org/10.1016/j.ress.2005.11.021

Janetti, E.B., Guadagnini, L., Riva, M. & Guadagnini, A. (2019), Global sensitivity analyses of multiple conceptual models with uncertain parameters driving groundwater flow in a regional-scale sedimentary aquifer. *Journal of Hydrology*, *574*, 544–556. https://doi.org/10.1016/j.jhydrol.2019.04.035

Jia, G. & Taflanidis, A.A. (2016), Efficient evaluation of Sobol' indices utilizing samples from an auxiliary probability density function. *Journal of Engineering Mechanics*, *142(5)*, 04016012. https://doi.org/10.1061/(ASCE)EM.1943-7889.0001061

Jones, D.R. (2001), A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, *21(4)*, 345–383. https://doi.org/10.1023/A:1012771025575

Jourdan, A. (2012), Global sensitivity analysis using complex linear models. *Statistics and Computing*, *22(3)*, 823–831. https://doi.org/10.1007/s11222-011-9239-y

Korgaonkar, Y., Meles, M. B., Guertin, D. P., Goodrich, D. C., & Unkrich, C. (2020). Global sensitivity analysis of KINEROS2 parameters representing green infrastructure using the STAR-VARS framework. *Environmental Modelling & Software*, *132*, 104814. https://doi.org/10.1016/j.envsoft.2020.104814

Krogh, S.A., Pomeroy, J.W. & Marsh, P. (2017), Diagnosis of the hydrology of a small Arctic basin at the tundra-taiga transition using a physically based hydrological model. *Journal of hydrology*, *550*, 685–703. https://doi.org/10.1016/j.jhydrol.2017.05.042

Leroux, N.R. & Pomeroy, J.W. (2019), Simulation of capillary pressure overshoot in snow combining trapping of the wetting phase with a nonequilibrium Richards equation model. *Water Resources Research*, *55(1)*, 236–48. https://doi.org/10.1029/2018WR022969

Li, C. & Mahadevan, S. (2016), An efficient modularized sample-based method to estimate the first-order Sobol′ index. *Reliability Engineering & System Safety*, *153*, 110–121. https://doi.org/10.1016/j.ress.2016.04.012

Li, W., Lin, K., Zhao, T., Lan, T., Chen, X., Du, H. & Chen, H. (2019), Risk assessment and sensitivity analysis of flash floods in ungauged basins using coupled hydrologic and hydrodynamic models. *Journal of Hydrology*, *572*, 108–120. https://doi.org/10.1016/j.jhydrol.2019.03.002

Lilhare, R., Pokorny, S., Déry, S.J., Stadnyk, T.A. & Koenig, K.A. (2020), Sensitivity analysis and uncertainty assessment in water budgets simulated by the variable infiltration capacity model for Canadian subarctic watersheds. *Hydrological Processes*, *34(9)*, 2057–2075. https://doi.org/10.1002/hyp.13711

Liu, H., Ong, Y.S., Shen, X. & Cai, J. (2020). When Gaussian process meets big data: A review of scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems*, 1–19. https://doi.org/10.1109/TNNLS.2019.2957109

Lophaven, S.N., Nielsen, H.B. & Søndergaard, J. (2002), Aspects of the Matlab toolbox DACE. Technical report IMM-REP-2002-13, Technical University of Denmark.

MacDonald, B., Ranjan, P. & Chipman, H. (2015), GPfit: An R package for fitting a Gaussian process model to deterministic simulator outputs. *Journal of Statistical Software*, *64(i12)*. http://hdl.handle.net/10.18637/jss.v064.i12

Markstrom, S.L., Hay, L.E. & Clark, M.P. (2016), Towards simplification of hydrologic modeling: identification of dominant processes. *Hydrology & Earth System Sciences*, *20(11)*, 4655–4671. https://doi.org/10.5194/hess-20-4655-2016

Marrel, A., Iooss, B., Laurent, B. & Roustant, O. (2009), Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering & System Safety*, *94(3)*, 742–751. https://doi.org/10.1016/j.ress.2008.07.008

Marseguerra, M., Masini, R., Zio, E. & Cojazzi, G. (2003), Variance decomposition-based sensitivity analysis via neural networks. *Reliability Engineering & System Safety*, *79(2)*, 229–238. https://doi.org/10.1016/S0951-8320(02)00234-X

Matheron, G. (1971), The Theory of Regionalized Variables and its Application. Paris School of Mines, Cah. Cent. Morphologie Math., 5. Fontainebleau.

Nash, J.E. & Sutcliffe, J.V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology*, *10(3)*, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Pianosi, F. & Wagener, T. (2015), A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. *Environmental Modelling & Software*, *67*, 1–11. https://doi.org/10.1016/j.envsoft.2015.01.004

Pianosi, F., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B. & Wagener, T. (2016), Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, 79, 214–232. https://doi.org/10.1016/j.envsoft.2016.02.008

Plischke, E. (2010), An effective algorithm for computing global sensitivity indices (EASI). *Reliability Engineering & System Safety*, *95(4)*, 354–360. https://doi.org/10.1016/j.ress.2009.11.005

Plischke, E., Borgonovo, E. & Smith, C.L. (2013), Global sensitivity measures from given data. *European Journal of Operational Research*, *226(3)*, 536–550. https://doi.org/10.1016/j.ejor.2012.11.047

Powell, M.J.D. (1978), A fast algorithms for nonlinearly constrained optimization calculations, in Watson, G.A. (ed.), Numerical Analysis, Springer-Verlag, Berlin, 144–157.

Puy, A., Lo Piano, S. & Saltelli, A. (2020), Current models underestimate future irrigated areas. *Geophysical Research Letters*, *47(8)*, e2020GL087360. https://doi.org/10.1029/2020GL087360

Rasmussen, C.E. & Williams, C.K. (2006), Gaussian Processes for Machine Learning. Cambridge, MA: MIT Press.

Ratto, M., Pagano, A. & Young, P. (2007), State dependent parameter metamodelling and sensitivity analysis. *Computer Physics Communications*, *177(11)*, 863–876. https://doi.org/10.1016/j.cpc.2007.07.011

Razavi, S. & Gupta, H.V. (2015), What do we mean by sensitivity analysis? The need for comprehensive characterization of "global" sensitivity in Earth and Environmental systems models. *Water Resources Research*, *51(5)*, 3070–3092. https://doi.org/10.1002/2014WR016527

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

Razavi, S. & Gupta, H.V. (2016b), A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application. *Water Resources Research*, *52(1)*, 440–455. https://doi.org/10.1002/2015WR017559

Razavi, S., Sheikholeslami, R., Gupta, H.V. & Haghnegahdar, A. (2019), VARS-TOOL: A toolbox for comprehensive, efficient, and robust sensitivity and uncertainty analysis. *Environmental Modelling & Software*, *112*, 95–107. https://doi.org/10.1016/j.envsoft.2018.10.005

Razavi, S., Gober, P., Maier, H.R., Brouwer, R. & Wheater, H. (2020). Anthropocene flooding: Challenges for science and society. *Hydrological Processes*, *34(8)*, 1996–2000. https://doi.org/10.1002/hyp.13723

Regis, R.G. & Shoemaker, C.A. (2013), A quasi-multistart framework for global optimization of expensive functions using response surface models. *Journal of Global Optimization*, *56(4)*, 1719–1753. https://doi.org/10.1007/s10898-012-9940-1

Regis, R.G. (2016), Trust regions in Kriging-based optimization with expected improvement. *Engineering Optimization*, *48(6)*, 1037–1059. https://doi.org/10.1080/0305215X.2015.1082350

Sacks, J., Welch, W.J., Mitchell, T.J. & Wynn, H.P. (1989), Design and analysis of computer experiments. *Statistical Science*, *4(4)*, 409–423.

Safta, C., Ricciuto, D.M., Sargsyan, K., Debusschere, B., Najm, H.N., Williams, M. & Thornton, P.E. (2015). Global sensitivity analysis, probabilistic calibration, and predictive assessment for the data assimilation linked ecosystem carbon model. *Geoscientific Model Development*, *8(7)*, 1899–1918. https://doi.org/10.5194/gmd-8-1899-2015

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M. & Tarantola, S. (2010), Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, *181(2)*, 259–270. https://doi.org/10.1016/j.cpc.2009.09.018

Sargsyan, K. and Safta, C., 2019. Surrogate-enabled sensitivity analysis and parameter inference of high-dimensional models (Report No. SAND2019-8257C). Sandia National Lab (SNL-CA), https://www.osti.gov/biblio/1641209

Schittkowski, K. (1983). On the convergence of a sequential quadratic programming method with an augmented Lagrangian line search function. *Mathematische Operationsforschung und Statistik. Series Optimization*, *14(2)*, 197-216.

Schürz, C., Hollosi, B., Matulla, C., Pressl, A., Ertl, T., Schulz, K. & Mehdi, B. (2019). A comprehensive sensitivity and uncertainty analysis for discharge and nitrate-nitrogen loads

involving multiple discrete model inputs under future changing conditions. *Hydrology & Earth System Sciences*, *23(3)*, 1211–1244. https://doi.org/10.5194/hess-23-1211-2019

Sheikholeslami, R. & Razavi, S. (2017), Progressive Latin Hypercube Sampling: An efficient approach for robust sampling-based analysis of environmental models. *Environmental Modelling & Software*, *93*, 109–126. https://doi.org/10.1016/j.envsoft.2017.03.010

Sheikholeslami, R., Razavi, S., Gupta, H.V., Becker, W. & Haghnegahdar, A. (2019), Global sensitivity analysis for high-dimensional problems: How to objectively group factors and measure robustness and convergence while reducing computational cost. *Environmental Modelling & Software*, *111*, 282–299. https://doi.org/10.1016/j.envsoft.2018.09.002

Sheikholeslami, R., Yassin, F., Lindenschmidt, K.E. & Razavi, S. (2017), Improved understanding of river ice processes using global sensitivity analysis approaches. *Journal of Hydrologic Engineering*, *22(11)*, 04017048. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001574

Shi, Y., Gong, W., Duan, Q., Charles, J., Xiao, C. & Wang, H. (2019), How parameter specification of an Earth system model of intermediate complexity influences its climate simulations. *Progress in Earth and Planetary Science*, *6(1)*. 46. https://doi.org/10.1186/s40645-019-0294-x

Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M. & Xu, C. (2015), Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. *Journal of Hydrology*, 523, 739–757. https://doi.org/10.1016/j.jhydrol.2015.02.013

Storlie, C.B., Swiler, L.P., Helton, J.C. & Sallaberry, C.J. (2009), Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliability Engineering & System Safety*, *94(11)*, 1735-1763. https://doi.org/10.1016/j.ress.2009.05.007

Sudret, B. (2008), Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, *93(7)*, 964–979. https://doi.org/10.1016/j.ress.2007.04.002

Volkova, E., Iooss, B. & Van Dorpe, F. (2008), Global sensitivity analysis for a numerical model of radionuclide migration from the RRC "Kurchatov Institute" radwaste disposal site. *Stochastic Environmental Research and Risk Assessment*, *22(1)*, 17–31. https://doi.org/10.1007/s00477-006-0093-y

Vrugt, J.A., Gupta, H.V., Bouten, E. & Sorooshian, S. (2003), A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, *39(8)*. http://dx.doi.org/10.1029/2002WR001642

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

Wang, A. & Solomatine, D.P. (2019), Practical experience of sensitivity analysis: Comparing six methods, on three hydrological models, with three performance criteria. *Water*, *11(5)*, p.1062. https://doi.org/10.3390/w11051062

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J. & Morris, M.D. (1992), Screening, predicting, and computer experiments. *Technometrics*, *34(1)*, 15–25.

Yassin, F., Razavi, S., Wheater, H., Sapriza−Azuri, G., Davison, B. & Pietroniro, A. (2017), Enhanced identification of a hydrologic model using streamflow and satellite water storage data: A multicriteria sensitivity analysis and optimization approach. *Hydrological Processes*, *31(19)*, 3320–3333. https://doi.org/10.1002/hyp.11267

Yun, W., Lu, Z., Jiang, X. & Zhang, L. (2018), Borgonovo moment independent global sensitivity analysis by Gaussian radial basis function meta-model. *Applied Mathematical Modelling*, *54*, 378–392. https://doi.org/10.1016/j.apm.2017.09.048