Environmental Modelling & Software 79 (2016) 214–232

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Sensitivity analysis of environmental models: A systematic review with practical workflow

Francesca Pianosi [a, *], Keith Beven [f], Jim Freer [c], Jim W. Hall [d], Jonathan Rougier [b], David B. Stephenson [e], Thorsten Wagener [a, g]

[a] Department of Civil Engineering, University of Bristol, UK
[b] Department of Mathematics, University of Bristol, UK
[c] School of Geographical Sciences, University of Bristol, UK
[d] Environmental Change Institute, University of Oxford, UK
[e] Department of Mathematics and Computer Science, University of Exeter, UK
[f] Lancaster Environment Centre, Lancaster University, UK
[g] Cabot Institute, University of Bristol, UK

ARTICLE INFO

Article history:
Received 31 December 2014
Received in revised form
15 January 2016
Accepted 2 February 2016
Available online 18 February 2016

Keywords:
Sensitivity Analysis
Uncertainty Analysis
Calibration
Evaluation
Robust decision-making

ABSTRACT

Sensitivity Analysis (SA) investigates how the variation in the output of a numerical model can be attributed to variations of its input factors. SA is increasingly being used in environmental modelling for a variety of purposes, including uncertainty assessment, model calibration and diagnostic evaluation, dominant control analysis and robust decision-making. In this paper we review the SA literature with the goal of providing: (i) a comprehensive view of SA approaches also in relation to other methodologies for model identification and application; (ii) a systematic classification of the most commonly used SA methods; (iii) practical guidelines for the application of SA. The paper aims at delivering an introduction to SA for non-specialist readers, as well as practical advice with best practice examples from the literature; and at stimulating the discussion within the community of SA developers and users regarding the setting of good practices and on defining priorities for future research.

© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Sensitivity Analysis (SA) investigates how the variation in the output of a numerical model can be attributed to variations of its input factors. Within this broad definition, the type of approach, level of complexity and purposes of SA vary quite significantly depending on the modelling domain and the specific application aims.

In contexts where very complex simulation models are used, for instance climate or atmospheric sciences, the term SA often refers to a 'what-if' analysis where the input factors of the simulation procedure, e.g. the model parameterization or the forcing scenario, are varied one at a time. Typically, the induced variations are assessed by visual comparison of model predictions. The goal is to verify the consistency of the model behaviour (e.g. Devenish et al., 2012) or to assess the robustness of the simulation results to uncertain inputs or model assumptions (e.g. Paton et al., 2013). The increasingly common practice in weather and climate science of producing sets (ensembles) of forecasts and simulations (e.g. Stephenson and Doblas-Reyes, 2000; Collins et al., 2012 and references therein) can be regarded as a type of SA exercise. Here, forecast uncertainty due to the imperfect knowledge of initial conditions is addressed via ensembles of weather forecasts starting from perturbed initial model states, while the sensitivity of climate simulations to model parameters is addressed using *perturbed physics ensembles* where simulations are made with different choices of model parameter values.

When simulation results can be associated with a summary scalar variable, for instance a measure of model performance like the sum of squared errors or some aggregate statistic of simulated variables, e.g. the mean streamflow, a more formal approach is to measure sensitivity as the variability induced in such a scalar variable via a set of quantitative sensitivity indices. Depending on whether output variability is obtained by varying the inputs around a reference (nominal) value, or across their entire feasible space, SA is either referred to as *local* or *global*. Local SA applications typically

* Corresponding author.
E-mail address: francesca.pianosi@bristol.ac.uk (F. Pianosi).

http://dx.doi.org/10.1016/j.envsoft.2016.02.008
1364-8152/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

consider model parameters as varying inputs, and aim at assessing how their uncertainty impacts model performance, i.e. how model performance changes when moving away from some optimal or reference parameter set. Partial derivatives or finite differences are used as sensitivity indices in the context of local approaches (e.g. Hill and Tiedeman, 2007). The spatio-temporal evolution of local sensitivity can also be investigated by adjoint methods (e.g. Vautard et al., 2000) or algebraic SA (Norton, 2008).

Global SA applications may consider model parameters but also other input factors of the simulation procedure, for instance the model's forcing data (e.g. Hamm et al., 2006) or its spatial resolution (e.g. Baroni and Tarantola, 2014) simultaneously. Different types of sensitivity indices can be used, ranging from correlation measures between inputs and output to statistical properties of the output distribution, e.g. its variance, and many others. Since analytical computation of these indices is impossible for most models, sensitivity indices are usually approximated from a sample of inputs and output evaluations. Global SA is used for a range of very diverse purposes, including: to support model calibration, verification, diagnostic evaluation or simplification (e.g. Sieber and Uhlenbrook, 2005; Harper et al., 2011; Nossent et al., 2011; Kelleher et al., 2013; Shin et al., 2013; Butler et al., 2014); to prioritize efforts for uncertainty reduction (e.g. Hamm et al., 2006); to analyse the dominant controls of a system (e.g. Pastres et al., 1999); to support robust decision-making (e.g. Nguyen and de Kok, 2007; Singh et al., 2014; Anderson et al., 2014).

In this paper we provide a systematic review and structuring of the SA literature across different environmental modelling domains with three specific objectives:

1. To provide a comprehensive view of SA purposes and approaches by clarifying terminology (e.g. quantitative versus qualitative, local versus global, one-at-a-time versus all-at-a-time) and by discussing the connections between SA and other methodologies for model identification and application (e.g. uncertainty analysis, model calibration and diagnostic evaluation, model-based decision-making, emulation modelling). The goal is to illustrate the broad spectrum of aims for which SA can be used, and thus stimulate its effective use in the environmental modelling community.
2. To provide a systematic review of the SA approaches most widely used in environmental modelling. The goal here is twofold: to provide non-expert readers with a broad enough background to engage with the SA literature while suggesting references for further reading; and to propose a classification system to support SA users in the choice of the most appropriate SA method depending on the characteristics of their case study.
3. To provide practical guidelines for the application of SA. To this end, we develop a workflow for the application of SA and discuss the key choices that SA users must make at each step of this workflow. We also provide practical suggestions on how to make these choices, how to assess their impacts on SA results and how to revise them, with good practice examples from the literature.

The paper is intended for a broad audience including researchers and practitioners who want to gain a general introduction to SA purposes and approaches, and to obtain practical advice on SA applications with best practice examples from the literature. The paper also aims at stimulating the discussion within the community of SA developers and users on good practice in SA application and on setting priorities for future research.

The paper is divided into three main sections that reflect the three objectives discussed above. Section 2 introduces common definitions and concepts used in the SA literature and clarifies the link between SA and related topics. Section 3 illustrates our classification system of SA methods with a short description of the underlying key assumptions, scope of application, advantages and limitations of each class of methods. Finally, Section 4 illustrates and discusses our proposed workflow for the application of SA. Section 3 and 4 build on some initial thoughts presented in the conference paper by Pianosi et al. (2014), however, both the classification system and the workflow have been significantly expanded and improved with respect to the earlier version discussed in that conference paper.

2. Conceptualization

2.1. Definition of model, input factors and outputs

In this paper we use the term *model* to refer to a numerical procedure (often implemented in a computer program) that simulates the behaviour of an environmental system, for instance by solving a set of algebraic equations (static model) or integrating differential equations over a spatial-temporal domain (dynamic model). We call *input factor* any element that can be changed before model execution, and *output* a variable that is obtained after the model execution. Examples of input factors are the parameters appearing in the model equations, the initial states, the boundary conditions or the input forcing data of a dynamic model; as well as non-numerical factors like the model equations themselves or, in the case of dynamic models, the time/spatial grid resolution for numerical integration. For dynamic models, the term 'output' usually does not refer to the entire range of temporal and spatial variables produced by the model simulation, but to a summary variable that is obtained by a scalar function of the simulated time series. Using the terminology proposed by Shin et al. (2013), we can distinguish two types of scalar functions:

- *objective functions* (also called *loss* or *cost functions*), which are measures of model performance calculated by comparison of modelled and observed variables (for instance, the Root Mean Squared Error);
- *prediction functions*, which are scalar values that are provided to the model-user for their practical use (for instance, the value of a variable at given time in given location, or its average over a spatial and temporal domain), and that can be computed even in the absence of observations.

Fig. 1 gives a practical example of possible inputs and outputs of SA in the case of a dynamic simulation model. While the aggregation of temporally and/or spatially distributed variables into a scalar output can induce a significant loss of information, such a loss can be recovered by considering multiple definitions of the summary output or analysing the temporal or spatial patterns of the output sensitivity. This issue will be further discussed in Section 4.1.

Given the above definitions, we can assume for the purposes of this paper that one can always resort to the general formulation:

$$y = g(\mathbf{x}) = g(x_1, x_2, \ldots, x_M) \quad (1)$$

where y is the output, $\mathbf{x} = [x_1, x_2, \ldots, x_M]$ is the vector of input factors, which belongs to the input variability space $\mathscr{X}$, and g is the function that maps the input factors into the output. This input–output relation is sometimes referred to as *response surface* or *model's response*, rather than 'model', to avoid confusion with the underlying simulation model which, as stated earlier, might have more inputs and outputs than $\mathbf{x}$ and y (see again Fig. 1). As the model's response function g is hardly ever available in analytic form, we will assume hereafter that a numerical procedure is

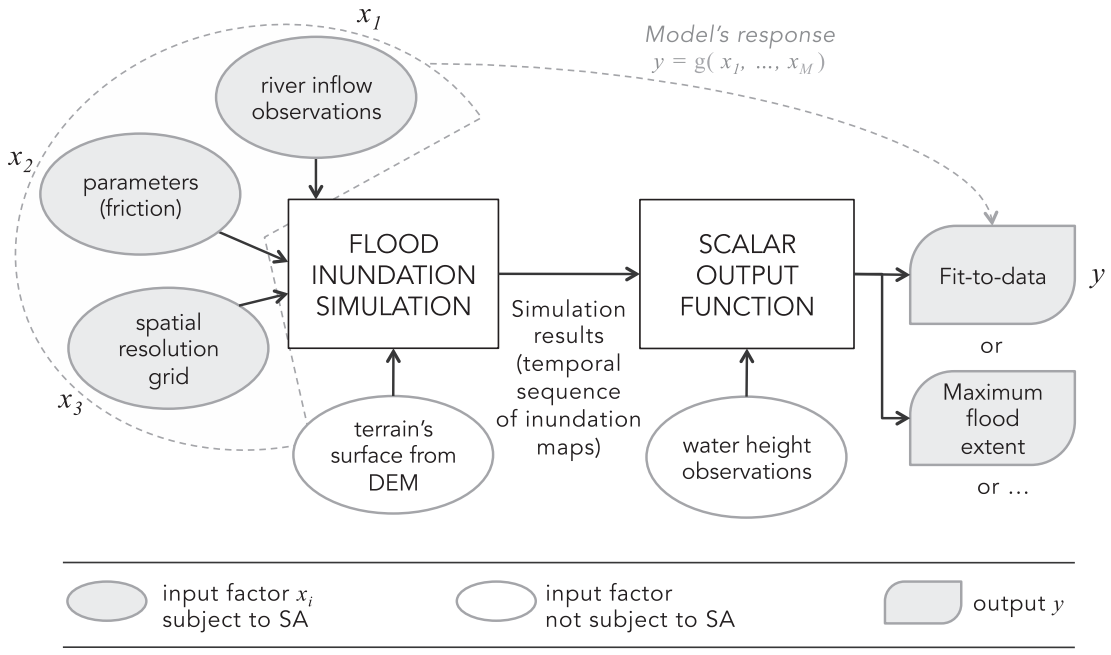


Fig. 1. Example of input factors and output definition for the SA of a (dynamic) flood inundation model.

available to evaluate it for any given combination of input factor values.

2.2. Types of Sensitivity Analysis

Sensitivity analysis investigates how the variation in the output y can be attributed to variations in the different input factors $x_1, x_2, \ldots, x_M$. Typical questions addressed by SA are: What input factors cause the largest variation in the output? Is there any factor whose variability has a negligible effect on the output? Are there interactions that amplify or dampen the variability induced by individual factors? We can distinguish different types of sensitivity analysis depending on how these questions are formulated and addressed.

2.2.1. Local and Global SA

Local sensitivity analysis considers the output variability against variations of an input factors around a specific value $\bar{x}$, while global sensitivity analysis (or GSA) considers variations within the entire space of variability of the input factors. The application of local SA obviously requires the user to specify a nominal value $\bar{x}$ for the input factors. While GSA overcomes this possible limitation, it still requires specifying the input variability space $\mathscr{X}$. When the latter is poorly known, the conclusions drawn from GSA should be taken with care.

2.2.2. Quantitative and Qualitative SA

We use the term *quantitative* SA to refer to methods where each input factor is associated with a quantitative and reproducible evaluation of its relative influence, normally through a set of sensitivity indices (or 'importance measures'). In *qualitative* SA, instead, sensitivity is assessed qualitatively by visual inspection of model predictions or by specific visualization tools like, for instance, tornado plots (e.g. Howard, 1988; Powell and Baker, 1992), scatter (or dotty) plots (e.g. Beven, 1993; Kleijnen and Helton, 1999a) or representations of the posterior distributions of the input factors (e.g. Freer et al., 1996, see also Section 3.4 and Appendix A). Often such visual tools are used complementary to a more quantitative analysis.

2.2.3. One-At-a-Time (OAT) and All-At-a-Time (AAT)

Another distinction often made is between 'One-[factor]-At-a-Time' (OAT) methods and what we propose to call 'All-[factors]-At-a-Time' (AAT) methods. This distinction refers to the sampling strategy used to estimate the sensitivity indices. In fact, in general, sensitivity indices cannot be computed analytically due to the complexity of the input–output relationship of Eq. (1) and thus they are numerically approximated from a sample of input factors and associated output evaluations (*sampling-based* SA from now on, see also Fig. 2). The distinction between OAT and AAT methods is based on the approach adopted to select input samples. Specifically:

- In OAT methods, output variations are induced by varying one input factor at a time, while keeping all others fixed.
- In AAT methods, output variations are induced by varying all the input factors simultaneously, and therefore the sensitivity to each factor considers the direct influence of that factor as well as the joint influence due to interactions.

While local SA typically uses OAT sampling, global SA can use either OAT or AAT strategies. In general, AAT methods provide a better characterization of interactions between input factors, and some of them (for instance, the variance-based methods described in Section 3.5) allow the user to analyse interactions between specific combinations (pairs, triples, etc.) of factors. OAT methods do not provide such detailed insights although some methods, for instance the EET described in Section 3.2, can give an indication on whether interactions matter or not. The drawback of AAT methods is that they typically require more extensive sampling and therefore a higher number of model evaluations (see further discussion in Sections 4.5 and 4.6).

2.2.4. Purposes (settings) of SA

Following Saltelli et al. (2008), we distinguish the following three purposes (or 'settings' in their terminology):

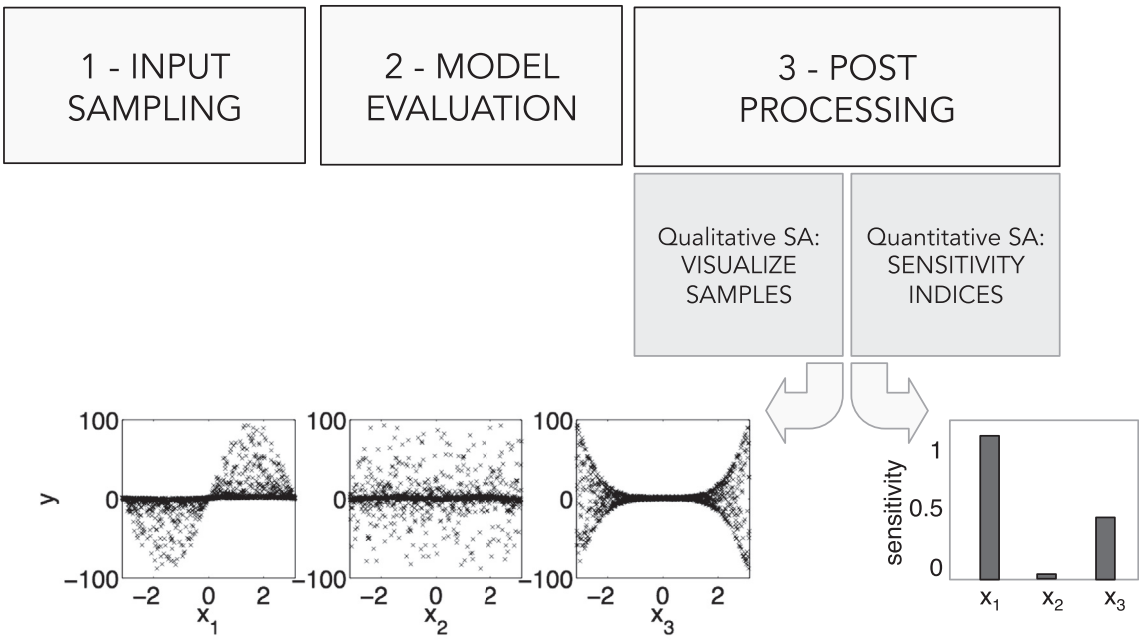


Fig. 2. The three basic steps in sampling-based Sensitivity Analysis, with an example of qualitative or quantitative results produced by the post-processing step.

- *Ranking* (or *Factor Priorization*) aims at generating the ranking of the input factors x_1, x_2,…, x_M according to their relative contribution to the output variability.
- *Screening* (or *Factor Fixing*) aims at identifying the input factors, if any, which have a negligible influence on the output variability.
- *Mapping* aims at determining the region of the input variability space that produces significant, e.g. extreme, output values.

The purpose of SA defines the ultimate goal of the analysis. It therefore guides the choice of the appropriate SA method since different methods are better suited to address different questions. Although SA is most commonly used for the three purposes above, our list is not exhaustive and other SA settings have been proposed. For instance the *direction* (or *sign*) *of change* is a question that can be addressed by SA (e.g. Anderson et al. (2014)). Another question is the presence of interactions between input factors. These aspects will be further discussed in Section 3. In the remainder of this Section, instead, we will discuss the links between SA and other related methods that can support the identification and assessment of environmental models.

2.3. SA and uncertainty analysis

When used for uncertainty assessment of numerical models, Sensitivity Analysis, and in particular global SA (GSA), is closely related to Uncertainty Analysis. Some authors (e.g. Saltelli et al., 2008), suggest that the discrimination is that UA focuses on *quantifying* the uncertainty in the output of the model, while GSA focuses on *apportioning* output uncertainty to the different sources of uncertainty (input factors). While different in focus and objectives, UA and GSA often use similar mathematical techniques. The 'forward' propagation of uncertainty by Monte Carlo simulation, which is commonly employed in many UA methodologies (e.g. Vrugt et al., 2009 or Beven and Freer, 2001) is also used to perform the initial steps of sampling-based GSA (Fig. 2). Some UA and GSA methods have been developed in close relation to each other: for instance the GLUE strategy for uncertainty analysis (Beven and Freer, 2001) was derived from the basic idea of Regional Sensitivity Analysis (see Section 3.4). In practice, GSA and UA often offer a valuable complement to each other: when performing GSA, UA should be used to verify that the output variability captured by sensitivity indices falls within the range of 'acceptable' model behaviour (see further discussion in Section 4.3); conversely, during UA, the estimation of sensitivity indices adds little computing effort while offering potentially valuable extra insights.

2.4. SA and model calibration

Sensitivity Analysis is also closely connected to the process of model calibration. By 'model calibration' we mean here the process of estimating the model parameters by maximizing the model fit to (or at least consistency with) observations. SA can be used to support and complement a model calibration exercise by providing insights on how variations in the uncertain parameters (the input factors $\mathbf{x}$) map onto variations of the performance metric (the output y) that measures the model fit. When an 'optimal' parameter estimate $\bar{\mathbf{x}}$ has been found, local SA can be used to investigate the uncertainty of such a parameterization: high local sensitivity to a parameter indicates high accuracy of its optimal estimate, while low sensitivity suggests that the parameter is poorly identified and uncertainty is large (an example is given by Sorooshian and Farid, 1982). A rigorous mathematical interpretation is available for the case when the output y is the mean squared error and gradient-based local sensitivity (see Section 3.1) is an approximation of the curvature (Hessian matrix) of y evaluated at $\bar{x}$ (for practical examples see for instance Sorooshian and Gupta (1985) or the PEST approach by Moore and Doherty (2005)). Most established analytical parameter-estimation methods for linear-in-parameters models (e.g. prediction-error method or generalized least squares and its variations) provide such local sensitivity information jointly with optimal parameter estimates (Ljung, 1999). SA is closely related to Identifiability Analysis (IA), which asks if parameters of a given model can be (uniquely or adequately) estimated from the available set of inputs and outputs.

While local SA usually follows the model calibration exercise,

global SA and model calibration are interlaced in a more complex, often iterative way. A model calibration based on the equifinality principle (Beven and Freer, 2001) can be used prior to GSA in order to constraint the input variability space $\mathscr{X}$, e.g. by finding parameter ranges that produce an acceptable level of model performance (e.g. Freer et al., 1996). On the other hand, GSA can be used before calibration of a computationally intensive model in order to: (i) identify parameters that have no influence on the model fit to observations and therefore can be ignored during refined calibration (e.g. van Werkhoven et al., 2009); (ii) investigate the parameters' influence and interactions in the regions of the parameter space associated with higher model performance, and thus provide the knowledge base for a more efficient local-search calibration in those regions (e.g. Spear et al., 1994); (iii) assess the potential for and limitations of model calibration given other uncertainty sources besides parameters, e.g. measurement errors in the observations or in the model forcing data (e.g. Baroni and Tarantola, 2014). In the latter case, the insights provided by GSA can help to set priorities for future efforts, for instance by investing in more sophisticated and computationally demanding calibration techniques or by first improving the quality of the data.

2.5. SA and model diagnostic evaluation

In cases where observations are affected by large uncertainties, due to observational errors, pre-processing errors, spatial averaging, etc., it might be hard to corroborate or reject a model based on some performance metric alone. Then, the modeller may also want to verify the consistency of the model behaviour with his/her perception of the real-world system (Wagener and Gupta, 2005). A model would be considered consistent if, for example, the parameters that control its response at a particular time or place are representative for physical processes that are also expected to dominate in reality. Being confident that the modelled controls are in line with our perceptions is particularly important if the model will be applied outside the range of variability of the calibration data (e.g. at different sites or for long term projections under nonstationary conditions). It is often difficult to predict when and where a specific parameter will have a significant influence on the simulation results when dealing with complex environmental models with many interacting components. Modified SA techniques have been used to formally address the question in what has in recent years been referred to as 'diagnostic model evaluation' (Gupta et al., 2008). For instance, Sieber and Uhlenbrook (2005), Reusser and Zehe (2011) and Herman et al. (2013) used time-varying and spatially-varying SA (see also Section 4.1) to quantify the temporal or spatial patterns of the output sensitivity to model parameters and therefore verify the model structure, i.e. assuming that different model components should be active during different system states. Similarly, the parameter screening provided by SA indicates whether there are 'unnecessarily' represented processes in the model and thus identify potential for model simplification, i.e. processes that are never activated in the model (e.g. Demaria et al., 2007). The modeller has to decide though whether this problem could be caused by limited calibration data variability and whether there is a potential for future, maybe more extreme, conditions to still trigger these processes (Gupta et al., 2008; Yilmaz et al., 2008).

2.6. SA, dominant controls analysis and robust decision-making

So far, we have discussed SA as a tool to investigate the propagation of uncertainty through a numerical model and to understand the model's intrinsic behaviour. Along the same lines, when simulation models are applied to anticipate the effects of management actions and thus support decision-making, SA is a recommended practice to assess the robustness of the assessment (and thus of the final decision) with respect to uncertain model inputs or assumptions (e.g. EC, 2009; EPA, 2009). Meaning that we can "ascertain if the inference of a model-based study is robust or fragile in light of the uncertainty in the underlying assumptions" Saltelli and D'Hombres (2010). However, SA can be applied to learn not only about models but also about systems. If the model reasonably reflects real-world processes, the application of SA to the model can provide insights into the dominant controls of the system. These insights can be used in turn to support decision-making by addressing questions like: what is the relative influence of different drivers – those that can be altered by the system managers and those that cannot - on the system response? What are critical values of the system drivers that induce threshold effects in the decision objectives? An early application of this type is reported in Pastres et al. (1999), who apply SA to a shallow water system to estimate the interactions between controllable system drivers (e.g. nitrogen load) and uncontrollable ones (e.g. dispersion or reaeration coefficients) in determining dramatic events such as anoxic crisis. More recently SA has been proposed as a tool for 'bottom-up' or 'vulnerability-based' approaches for dealing with decision-making problems under large (and often unknown levels of) uncertainties (Wilby and Dessai, 2010) like for example climate projections uncertainties. In such instances, Sensitivity Analysis, and in particular mapping methods of input factors, can be used to explore the space of possible variability of the system drivers, for instance climate or socio-economic drivers like land use, demand for natural resource, etc., and isolate combinations that would exceed vulnerability thresholds (Lempert et al., 2003); or to quantify links between the vulnerability of a system (e.g. a catchment) and its properties (e.g. climate, hydrology, see for instance Prudhomme et al. (2013)). More widely employed mapping methods include the Patient Rule Induction Method (PRIM) and Classification And Regression Trees (CART) (Lempert et al., 2008). Applications of SA for this purpose are far less numerous than those for uncertainty investigation and model calibration. However, they are increasingly investigated, see for example Brown et al. (2011), Singh et al. (2014) and references therein.

2.7. SA and emulators

An emulator (or emulation model, or surrogate model) is a computationally efficient model, e.g. a polynomial or some other algebraic relation, that is calibrated over a (small) dataset obtained by simulation of a computationally demanding model, and that can be used in its place during computationally expensive tasks. In the context of SA, emulators can be used to obtain faster evaluations of the model's response (Eq. (1)) and therefore allow for applying computationally demanding SA methods to complex simulation models. For specific choices of the emulator structure and the SA method, emulators can provide analytical solutions to compute sensitivity indices. For example Sudret (2008) presents an approach where generalized polynomial chaos expansions (PCE) are used as emulators and variance-based sensitivity indices (see Section 3.5) are computed analytically as a post-processing of the PCE coefficients. On the other hand, the use of emulators poses a number of numerical challenges related to their calibration and validation. In fact, the validity of an emulator relies on the assumption that the samples used for its identification are sufficiently representative of the behaviour of the original simulation model and for the intended model application, an assumption that is difficult if not impossible to verify. The identification and use of emulators for SA is the topic of a wide literature, whose review falls outside the scope of this paper. The interested reader is referred to Forrester et al. (2008) for a general introduction to emulation modelling, and Ratto et al.

(2012) for a review of its application in SA.

3. Systematic review of SA methods

In this section we propose a systematic classification of SA methods. This review does not aim at providing an exhaustive list of all the available SA methods, which would be hardly feasible and likely become obsolete in a short while. Rather, we group the methods most widely used in the environmental modelling domain into 5 broad classes, based on their underlying concept, which reflect different assumptions, working principles and objectives. In this sense our review is 'systematic' and hopefully open to encompass methods that we do not cite here explicitly as well as for future developments within each class. The reviewed SA methods are then placed within this classification system (shown in Fig. 3) that can be used as an operational tool to guide the choice of the most appropriate SA method for a problem at hand, depending on:

- the specific SA purpose (screening, ranking or mapping, as described in Section 2.2) that each method can address;
- the method's computational complexity, measured by the number of model evaluations required in its application.

We emphasize the role of computational complexity because sampling-based methods requiring large sample sizes can be impossible to apply to models with long run time and/or those producing large input/output data files. In Fig. 3, we provide a rough idea of the number of model evaluations required by each class of methods. More discussion of the computational complexity issue is given in Section 4.5. The remainder of this Section is dedicated to a short description of the five classes of methods, their working principles, and their advantages and limitations. The mathematical notation used throughout the Section is summarised in Table 1. We intentionally do not provide excessive mathematical details on the mechanics of the various SA methods, and refer the reader to the cited literature. A good complement of this review in this regard are the introduction to sensitivity assessment of simulation models by Norton (2015), the literature reviews (with a focus on the chemical modelling literature) by Saltelli et al. (2005, 2012) and the review of recent methodological advances by Borgonovo and Plischke (2016).

3.1. Perturbation and derivatives methods

The simplest type of SA varies (perturbs) the input factors of the simulation model from their nominal values one at a time (OAT) and assesses the impacts on the simulation results via visual inspection, for instance by pair-comparison of the time series (or spatial patterns) of simulated variables under nominal and perturbed inputs (e.g. Devenish et al., 2012 and Paton et al., 2013). If a scalar output variable y can be defined, a more formal approach is to measure the output sensitivity to the i-th input factor by the partial derivative $\partial g/\partial x_i$ evaluated at the nominal value of the factors $\bar{x}$, or by the finite-difference gradient if the input–output

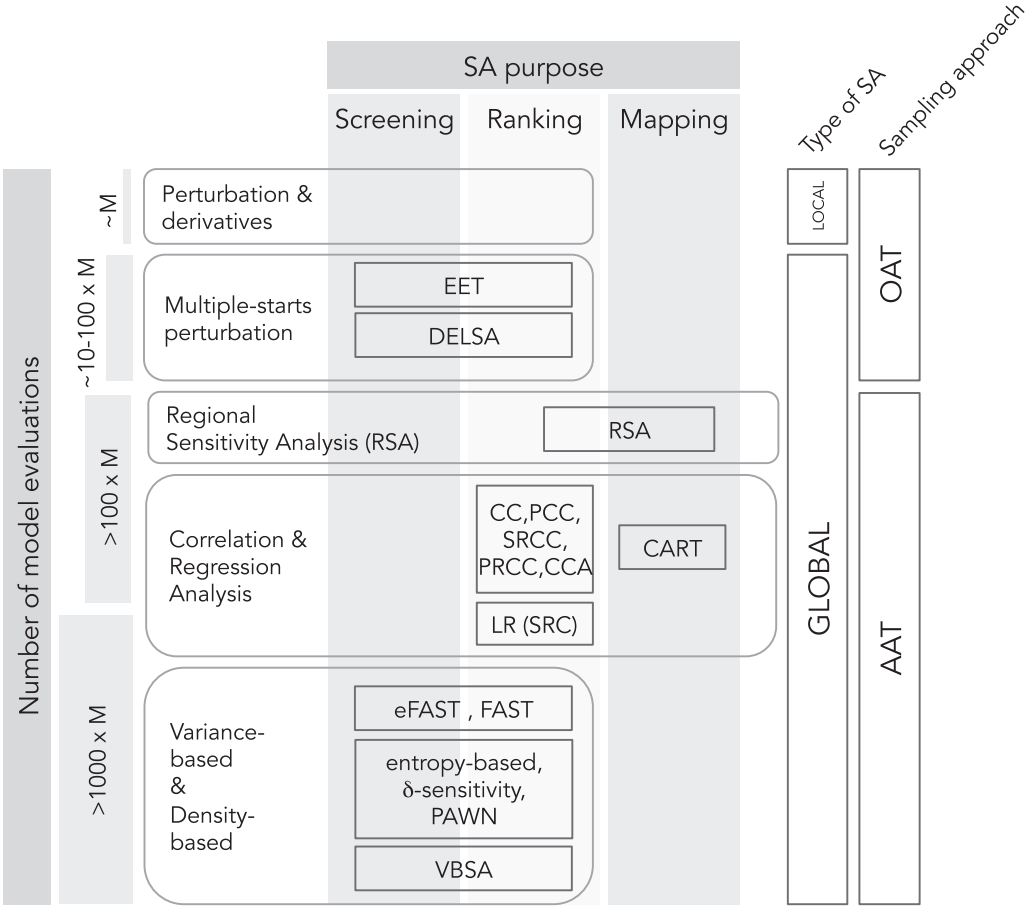


Fig. 3. Classification system of Sensitivity Analysis methods based on computational complexity (vertical axis; M is the number of input factors subject to SA) and purposes of the analysis. Some of the most widely used methods are reported (acronyms are defined in corresponding paragraphs of Sec. 3). Types of SA and sampling approaches are defined in Sec. 2.2. Figures about computational complexity are indicative, for a further discussion see Sec. 4.5.

Table 1
List of symbols used and their meaning.

E	Expected value
f	Probability Density Function (PDF)
F	Empirical Cumulative Distribution Function (CDF)
g	Relationship between the model's inputs and output investigated by SA (or model's response), as defined by Eq. (1)
M	Number of input factors subject to SA
N	Sample size (and thus number of model evaluations) in sampling-based SA
n	Base sample size for variance-based sensitivity estimators (Saltelli et al., 2010)
r	Number of local derivatives in multiple-starts perturbation methods
SD	Standard Deviation
S_i	Sensitivity index of the i-th input factor
V	Variance
$\mathbf{x}$	Vector of M input factors subject to SA
$\overline{\mathbf{x}}$	Nominal value of $\mathbf{x}$ for local SA
$\mathscr{X}$	Variability space of $\mathbf{x}$ for global SA
x_i	i-th input factor subject to SA
y	(Scalar) Model output
$Y_b(Y_{nb})$	Set of behavioural (non-behavioural) output samples in Regional Sensitivity Analysis

function g of Eq. (1) is not differentiable at $\overline{x}$. Derivative-based SA finds its rationale in the Taylor series expansion. This is well explained in Helton (1993) and generalized later on in Borgonovo (2008). In order to facilitate a comparison of sensitivities across input factors that may have different units of measurements, the partial derivatives are usually rescaled (e.g. Hill and Tiedeman, 2007). The sensitivity measure for the i-th input factor thus takes the form

$$S_i(\overline{x}) = \left.\frac{\partial g}{\partial x_i}\right|_{\overline{x}} c_i \tag{2}$$

where c_i is the scaling factor. Given that the functional relation of Eq. (1) is rarely known in analytic form, partial derivatives are usually approximated by finite differences, i.e.

$$\widehat{S}_i(\overline{x}) = \frac{g(\overline{x}_1, \ldots, \overline{x}_i + \Delta_i, \ldots, \overline{x}_M) - g(\overline{x}_1, \ldots, \overline{x}_i, \ldots, \overline{x}_M)}{\Delta_i} c_i \tag{3}$$

Using an approximation of Eq. (3), the computation of the sensitivity measures for M factors requires $M+1$ model evaluations. Derivative-based sensitivity measures are therefore computationally very cheap, with the drawback that they provide information about local sensitivity only. Second derivatives can be estimated with a relatively small number of additional model evaluations, thus providing information about local interactions between input factors. For more details on this issue see Norton (2015).

3.2. Multiple-starts perturbation methods

A global extension of the perturbation approach is to compute output perturbations from multiple points $\overline{x}^j$ within the feasible input space, and to measure the global sensitivity by aggregating these individual sensitivities. Methods falling under this category differ from each other in one or more of the following aspects: (i) whether they use finite differences directly, or some transformation such as their absolute or squared values; (ii) how they select the fixed points $\overline{x}^j$ and the length of the finite variation Δ_i to perturb the i-th input factor (design strategy); (iii) how they aggregate individual sensitivities.

The most established method of this type is the method of Morris (Morris, 1991), also called the Elementary Effect Test (EET (Saltelli et al., 2008)). Here, the mean of r finite differences (also called 'Elementary Effects' or EEs) is taken as a measure of global sensitivity, i.e.

$$\begin{aligned} S_i &= \frac{1}{r}\sum_{j=1}^{r} EE^j \\ &= \frac{1}{r}\sum_{j=1}^{r} \frac{g\left(\overline{x}_1^j, \ldots, \overline{x}_i^j + \Delta_i^j, \ldots, \overline{x}_M^j\right) - g\left(\overline{x}_1^j, \ldots, \overline{x}_i^j, \ldots, \overline{x}_M^j\right)}{\Delta_i^j} c_i \end{aligned} \tag{4}$$

Besides the above sensitivity measure, it is common practice to also compute the standard deviation of the EEs, which provides information on the degree of interaction of the i-th input factor with the others. A high standard deviation indicates that a factor is interacting with others because its sensitivity changes across the variability space. An alternative measure proposed by Campolongo and Saltelli (1997) takes the absolute value of the finite differences to avoid that differences of different signs would cancel out. Borgonovo (2010) present a method where, at the additional cost of $M+1$ model evaluations per EE, one can estimate whether the response of the model is predominantly additive or governed by interactions.

As for the sampling strategy to select the points $\overline{x}^j$ $(j = 1, \ldots, r)$ and the input variations Δ_i, different approaches have been proposed. The sampling strategy originally proposed by Morris (1991) builds r trajectories in the input space, each composed of $M+1$ points. The starting point of each trajectory is randomly selected over a uniform grid and the subsequent M points are obtained by moving one factor at a time of a fixed amount Δ, so that each trajectory allows for evaluating one EE per factor. The user has to specify the "number of levels" L, which determines the grid size (equal to $1/(L-1)$ of the range of variability of the input factor) and the size of the variation Δ (equal to $L/(2*(L-1))$). Typical values for L ranges from 4 to 8, which means that Δ ranges from $4/6 = 0.76$ to $8/14 = 0.57$ of the range of variability. Therefore, with this setup the EEs capture finite and rather large perturbations. On the one hand this avoids the risk of focussing only on very local behaviours of the model's response g (Eq. (1)). On the other hand it can produce misleading results if g is highly non-smooth and the characteristic length of its variations is much smaller than Δ.

Several variants of the sampling strategy by Morris have been proposed, including the LH-OAT approach proposed by van Griensven et al. (2006), where the starting points of each trajectory are generated by Latin-Hypercube sampling rather than random sampling over a grid; and the approach by Campolongo et al. (2007), where a high number of trajectories are generated and a subset of r trajectories is selected so to maximise the overall spread over the input space. A different approach to OAT sampling is the radial-based design, where the variations Δ_i are all taken

starting from the same (randomly selected) point in the input space. Campolongo et al. (2011) show that radial-based design provides several advantages in terms of efficiency and integration with subsequent AAT sensitivity analysis. The interested reader is referred to that paper and references therein for a discussion of different OAT sampling strategies.

For all of these sampling strategies, the computation of the mean (and standard deviation) of the EEs of M input factors requires $r(M+1)$ model evaluations, a requirement that is far lower than the majority of AAT global approaches. Therefore, the EET is often used when the computing time of a single model run is high, or when the number of factors is very large. The EET is particularly suitable for screening, i.e. to detect non-influential factors that can be discarded from a subsequent, more time-consuming global SA (see for instance Nguyen and de Kok, 2007), and for ranking.

Other multiple-start perturbation approaches use squared finite differences, which allow a link to be established with the variance-based SA approach discussed in Section 3.5. For instance, Sobol' and Kucherenko (2009) suggest use of the mean of the squared finite differences and demonstrate that it provides an upper bound on the total-order variance-based sensitivity index (see Section 3.5). This sensitivity measure is especially suitable for screening since a small value of the measure implies that the input factor is non-influential, while the same authors show that it may give false conclusions if used for ranking. Along a similar line of reasoning is the DELSA approach (Distributed Evaluation of Local Sensitivity Analysis) by Rakovec et al. (2014), which also uses the squared finite differences as a measure of sensitivity (scaled by the ratio between the a priori input variance and the total output variance). Here, local sensitivities computed at different sampling points are not aggregated but their full frequency distribution is analysed, and if aggregated, the median value is used rather than the mean. Another difference that is worth mentioning with respect to the EET is that in the DELSA approach the input variation Δ is set to 0.01 of the fixed value $\bar{x}_i^j$, so that finite differences can be regarded as approximating local derivatives.

3.3. Correlation and Regression analysis methods

The underlying idea of these methods is to derive information about output sensitivity from the statistical analysis of the input/output dataset generated by Monte Carlo simulation. Early works in the field are Iman and Helton (1988) (mainly on regression analysis) and Saltelli and Marivoet (1990) (on correlation methods). An introduction and review of these approaches are given e.g. in Kleijnen and Helton (1999a), Helton and Davis (2002) and Storlie et al. (2009).

Correlation methods use the correlation coefficient between the input factor x_i and the output y as a sensitivity measure, i.e.

$$S_i = \text{correlation}(x_i, y) \tag{5}$$

Several different definitions of correlation can be used, including Pearson correlation coefficient (CC) and partial correlation coefficient (PCC), which apply when a linear relationship exists between the input factors $\mathbf{x}$ and the output y, and the Spearman rank correlation coefficient (SRCC) or partial rank correlation coefficient (PRCC), which can be used for nonlinear but monotonic relationships (e.g. Pastres et al., 1999). The choice among these different alternatives depends on the degree of acceptability of the linearity and/or monotonicity assumption between inputs and output. An informal though effective way to assess this is through visual inspection of the input/output sample, for instance using scatter plots. More sophisticated correlation methods can be used to address specific needs. For example, Minunno et al. (2013) demonstrate the use of Canonical Correlation Analysis (CCA) for GSA in an application where multiple model outputs need to be accounted for simultaneously.

Regression analysis methods instead derive the sensitivity measure as a 'byproduct' of regression analysis applied to the input/output sample. The simplest and most widely used method is linear regression. Here, a linear relationship $y = a_i + b_i x_i$ is assumed and the linear least-squares estimate of the regression coefficient b_i is the sensitivity measure. The Standardised Regression Coefficients (SRC) are used when input factors have different units of measurement, i.e.

$$S_i = b_i \frac{SD(x_i)}{SD(y)} \tag{6}$$

where SD stands for standard deviation. Multiple linear regression can be used to obtain the sensitivities to all the individual input factors at once. The advantage of linear regression is that it can be easily applied to small datasets, however it can be inadequate if the input–output relationship is non-monotonic or strongly nonlinear (e.g. Hall et al., 2009).

A particularly interesting class of nonlinear regression methods in the context of Sensitivity Analysis is that of Classification And Regression Trees (CART, for application examples see e.g. Harper et al. (2011) and Singh et al. (2014)). CART provides several advantages, including that they can easily handle non-numerical inputs and outputs, and that they can be used for both ranking and mapping.

3.4. Regional Sensitivity Analysis (or Monte-Carlo filtering)

Regional Sensitivity Analysis (RSA), also called Monte Carlo filtering, is a family of methods mainly aimed at identifying regions in the inputs space corresponding to particular values (e.g. high or low) of the output, and that can be used for mapping and for dominant controls analysis. The idea was first proposed and investigated in Young et al. (1978) and Spear and Hornberger (1980). Here, the input samples (typically parameters) are divided into two binary sets, 'behavioural' and 'non-behavioural', depending on whether the associated model simulation exhibits the expected pattern of state variable response or not. Another way to apply RSA is by splitting input samples depending on whether the associated output is above or below a prescribed threshold. Then, the two input sets are compared to gain insight on the model behaviour and mapping. For example, Q–Q plots can be used to compare behavioural versus non-behavioural samples. Another common analysis is to over-plot the marginal empirical cumulative distribution functions (CDF) of the behavioural and non-behavioural sets. Visual inspection of these distributions provides information on factor mapping, for instance by highlighting a reduction in the variability range for behavioural inputs. The divergence between the two distributions, for example measured by the Kolmogorov–Smirnov statistic, can be used as a sensitivity index, i.e.

$$S_i = \max_{x_i} \left| F_{x_i|y_b}(x_i|y \in Y_b) - F_{x_i|y_{nb}}(x_i|y \in Y_{nb}) \right| \tag{7}$$

where $F_{x_i|y_b}$ and $F_{x_i|y_{nb}}$ are the empirical cumulative distribution functions of x_i when considering input samples associated with behavioural and non-behavioural outputs respectively (i.e. falling in the subsample Y_b/Y_{nb} of behavioural/non-behavioural outputs). The advantage of using empirical distribution functions is that they usually provide a robust approximation of the underlying distribution even if computed over small samples. This may happen for instance with overparameterised models where behavioural

parameterisations are confined to small sub-regions of the parameter space and therefore the size of the behavioural set might be very small even if starting from a large number of model simulations (Norton, 2015). However, while useful for ranking, the sensitivity measure of Eq. (7) cannot be used for screening. In fact, a value of zero of the above index is a necessary but not sufficient condition for insensitivity because input factors contributing to output variability only through interactions may have the same behavioural and non-behavioural distribution functions (see for instance the example given in Section 5.2.3 of Saltelli et al. (2008)).

One advantage of this approach is that it can be applied to any type of model output, including non-numerical ones, as long as a splitting condition can be defined and verified, possibly also by qualitative evaluation. On the other hand, the use of a splitting criterion can be a limitation whenever the discrimination between behavioural and non-behavioural outputs is not clear-cut. For instance, RSA has been widely used in applications where the model output is an objective function (i.e. a measure of the model accuracy against observations) and the splitting criterion reflects the achievement of a minimum requirement of model performance (e.g. Freer et al., 1996; Sieber and Uhlenbrook, 2005). The definition of the threshold value at which the model performance is deemed acceptable is usually a subjective choice by the modeller. The problem can be especially difficult when the scalar model output is a predictive function, unless there exists a threshold value that has a specific meaning for the model users (for instance a regulatory threshold value for an environmental variable). To overcome the issue and apply RSA without specifying thresholds, one option is to group the ranked output samples into a prescribed number, e.g. 10, of equally spaced intervals, and compare the 10 resulting distribution functions of the input factors (Freer et al., 1996; Wagener et al., 2001). For an application example and discussion see also Tang et al. (2007b).

3.5. Variance-based methods

Variance-based SA relies on three basic principles: (i) input factors are regarded as stochastic variables so that the model induces a distribution in the output space; (ii) the variance of the output distribution is a good proxy of output uncertainty; (iii) the contribution to the output variance from a given input factor is a measure of sensitivity.

Several variance-based indices can be defined. *First-order indices* (or 'main effects') measure the direct contribution to the output variance from individual input factors or, equivalently, the expected reduction in output variance that can be obtained when fixing a specific input, i.e.

$$S_i^F = \frac{V_{x_i}[E_{x_{\sim i}}(y|x_i)]}{V(y)} = \frac{V(y) - E_{x_i}[V_{x_{\sim i}}(y|x_i)]}{V(y)} \quad (8)$$

where E denotes expected value, V denotes the variance, and $x_{\sim i}$ denotes "all input factors but the i-th". The *total-order indices* (or 'total effects') introduced by Homma and Saltelli (1996) measure the overall contribution from an input factor considering its direct effect and its interactions with all the other factors, which might amplify the individual effects, i.e.

$$S_i^T = \frac{E_{x_{\sim i}}[V_{x_i}(y|x_{\sim i})]}{V(y)} = 1 - \frac{V_{x_{\sim i}}[E_{x_i}(y|x_{\sim i})]}{V(y)} \quad (9)$$

Total-order indices are particularly suitable for screening because a value of zero of the total-order index is a necessary and sufficient condition for a factor to be non-influential. First-order indices are often used for ranking, especially if interactions are not significantly contributors to output variance. Variance-based sensitivity indices of intermediate order can also be defined: for instance, second-order indices measure the contribution to output variance from pairs of factors; third-order indices from factor triples; etc. These indices can be used to analyse interactions between specific groups of input factors. An effective account of the development of variance-based indices and their connections to earlier works on 'importance measures' (e.g. Iman and Hora (1990)) can be found in Borgonovo (2007).

An interesting property of first-order and higher-order indices is that they are related with the terms in the variance decomposition of the model output (Sobol', 1993), which "reflects the structure of the model itself" (Oakley and O'Hagan, 2004) and holds under relatively broad assumptions, the strongest one being that input factors are independent. In the presence of correlations among the input factors, instead, the tidy correspondence between variance-based indices and model structure is lost (see e.g. discussion in Oakley and O'Hagan (2004)) and counterintuitive results may be obtained. For example one might observe total-order indices smaller than first-order ones for negative correlations or total-order indices tending to zero as correlation grows to unity (Kucherenko et al., 2012). The mechanism of output variance decomposition and the link to variance-based sensitivity indices are also discussed in Norton (2015).

Another reason for the popularity of the first-order and total-order indices is that they are relatively easy to implement since several closed-form algebraic equations exist for their approximation. For a review of these estimators in the case of independent input factors, see Saltelli et al. (2010); for an extension to the case of dependent inputs, see Kucherenko et al. (2012). However, the sample size required to achieve reasonably accurate approximations can be rather large (as further discussed in Section 4.5), which severely affects the applicability of this approach to time-consuming models. Several methods were proposed to reduce the number of model evaluations in the approximation of variance-based indices. These include: (i) methods using the Fourier series expansion of the model output y, like the Fourier Amplitude Sensitivity Test (FAST (Cukier et al., 1973)) for the approximation of the first-order indices, and the extended FAST (Saltelli et al., 1999) for the total-order indices (for an introduction to these techniques, see Norton (2015)); and (ii) methods using an emulator like the approach by Oakley and O'Hagan (2004).

Besides computational aspects, another limitation of variance-based indices is that, by relying on the implicit assumption that variance can fully capture uncertainty, they can be inappropriate when the output distribution is multi-modal or highly-skewed and the variance is therefore not a meaningful indicator. This issue is discussed in the next section.

3.6. Density-based methods

The limitations of the variance-based approach have stimulated a number of studies on 'moment-independent' sensitivity indices that do not use a specific moment of the output distribution to characterize uncertainty and therefore are applicable independently of the shape of the output distribution. These methods are sometimes referred to as 'density-based' methods because they look at the Probability Density Function (PDF) of the model output, rather than its variance alone.

The key idea is to measure sensitivity through the variations in the output PDF that are induced when removing the uncertainty in one input factor. In practice this is done by computing the divergence between the unconditional output PDF, which is generated by varying all factors, and the conditional PDFs that are obtained when fixing individual input factor in turn to a prescribed value. If multiple conditioning points are considered, some type of statistic

is applied to aggregate individual results. The general form of a density-based sensitivity index is

$$S_i = \underset{x_i}{\text{stat}}\ \text{divergence}\left[f_y, f_{y|x_i}(\cdot|x_i)\right] \quad (10)$$

where f_y and $f_{y|x_i}$ are the unconditional and conditional output PDFs, and 'stat' and 'divergence' denote some statistic and divergence measure. For example, in entropy-based methods the divergence between conditional and unconditional PDF is measured by the Shannon entropy (Krykacz-Hausmann, 2001) or by the Kullback-Leibler entropy (Park and Ahn, 1994; Liu et al., 2006), while the δ-sensitivity approach (Borgonovo, 2007) uses the area enclosed between the two PDFs. In the δ-sensitivity approach, different conditioning values are used for x_i and individual results are averaged, i.e. 'stat' in Eq. (10) is the mean, while in entropy-based methods only one conditioning value is typically used. Other density-based approaches, e.g. Borgonovo (2014) and the novel density-based PAWN method by Pianosi and Wagener (2015), use cumulative distribution functions in place of PDFs. The advantage is that unconditional and conditional CDFs can be efficiently approximated by the empirical CDFs of output samples, which makes the density-based sensitivity indices very simple to compute.

One advantage of density-based sensitivity indices is that they can easily be tailored to measure sensitivity over the entire range of output variability as well as a specific sub-range, for instance extreme values (the so called *Regional Response Probabilistic Sensitivity Analysis* discussed in Liu et al. (2006)). This may be very interesting in those applications, e.g. hazard assessment, where the tail of the output distribution is of particular interest. Other interesting properties of density-based methods are that they allow for using statistics that are monotonic transformation invariant, and that they can be estimated from a given sample, i.e. without requiring a tailored sampling strategy (Borgonovo (2014) and references therein).

Application examples in the environmental domains are Pappenberger et al. (2008) for the entropy-based indices, Castaings et al. (2012); Anderson et al. (2014) and Peeters et al. (2014) for the δ-sensitivity measure, and Pianosi and Wagener (2015) for PAWN.

4. Workflow for the application of SA

Despite the differences between the individual SA methods described in the previous section, their application requires performing a sequence of steps that, to some extent, can be discussed in general terms. Here we refer to these steps as 'workflow'. The workflow for the application of SA is illustrated in Fig. 4. In this section, we discuss this workflow and the choices that users have to make at each step, with the goal of providing practical guidelines to support users in their SA application.

4.1. Experimental set-up: define input factors and output

Any SA exercise starts from three basic choices that together form what we call the 'experimental setup': (i) choosing which input factors will be subject to SA; (ii) setting the values of other input factors that will be kept constant throughout the SA; and (iii) defining the model output.

When the model execution produces a temporally or spatially varying set of outputs, the application of SA typically requires aggregating the outputs into a scalar function, as described in Section 2.1. An exception is the case when the input factors are the model parameters and the mathematical form of the model allows the derivation of algebraic solutions of the model state's sensitivity in time (Norton (2015) and references therein). When a scalar output function must be used, its definition obviously affects the SA results because different scalar outputs may have different sensitivities to the input factors. For instance, Pappenberger et al. (2008) shows how the ranking of the input factors (the parameters of a flood inundation model) vary when considering the mean of the squared errors or the mean of the absolute errors as scalar outputs.

Often, it is convenient to define multiple scalar outputs that summarise different aspects of the model behaviour. Their sensitivity can then be analysed separately (e.g. Baroni and Tarantola, 2014) or jointly (e.g. Minunno et al., 2013), or reframed as a multi-criteria analysis using for example Pareto ranking (e.g. Rosolem et al., 2012).

Another option that is becoming more and more accessible with growing computing power is that of reducing the level of aggregation so to preserve more of the temporal or spatial variability of the model. Sensitivity indices can be computed at different temporal resolutions, therefore obtaining their temporal evolution over the simulation horizon (Wagener and Harvey, 1997; Wagener et al., 2003; Cloke et al., 2008; Reusser and Zehe, 2011; Kelleher et al., 2013; Guse et al., 2014). A similar approach can be applied to the aggregation of spatial patterns into a single output function, whose resolution can be varied in order to capture the space-variability of sensitivities across the model domain (Tang et al., 2007a; van Werkhoven et al., 2008). Time-varying or spatially-varying SA is especially useful to provide new insights about the dynamics of the model (e.g. when and where a given parameter is more influential) and/or the underlying system (e.g. what processes are mostly influential, when and where). However, its application poses a number of practical difficulties, for example regarding the choice of the averaging window size and of appropriate methods for complex models (Massmann et al., 2014), which constitute an opportunity for further research.

4.2. Choose the SA method

As discussed in Section 3, the choice of the most appropriate SA method for a given problem is largely driven by the purpose of the analysis (screening, ranking or mapping: see horizontal axis in Fig. 3) and by the available computing resource (and therefore the maximum number of model evaluations that can be used to approximate sensitivity indices: see vertical axis in Fig. 3). Typically, the number of model evaluations N increases with the number of input factors M subject to SA. However, the ratio between N and M significantly varies from one method to another and often also from one application to another. This is illustrated in Fig. 5, which reports some examples of combinations (M,N) taken from the literature. The choice of the appropriate sample size will be further discussed in Section 4.5.

The choice of the method can be also driven by other specific features of the problem at hand, like the linearity of the input–output relationship, the statistical characteristics of the output distribution (e.g. its skew), etc., which are handled more or less effectively by different methods, as discussed in Section 3.

When multiple options are available, it may be advisable to apply more than one method and to compare individual results so to reinforce the general conclusions drawn from SA. Often, this can be done at almost no extra computing cost because different methods can be applied to the same input–output sample without re-running the model. This topic will be further discussed in Section 4.8. When the number of input factors is high, another option is to apply methods sequentially, beginning with computationally efficient screening methods like the EET and then applying more computer-intensive methods to a reduced number of input factors. In such a case, a careful design of the OAT sampling strategy applied during the screening step could help to reduce the number of new

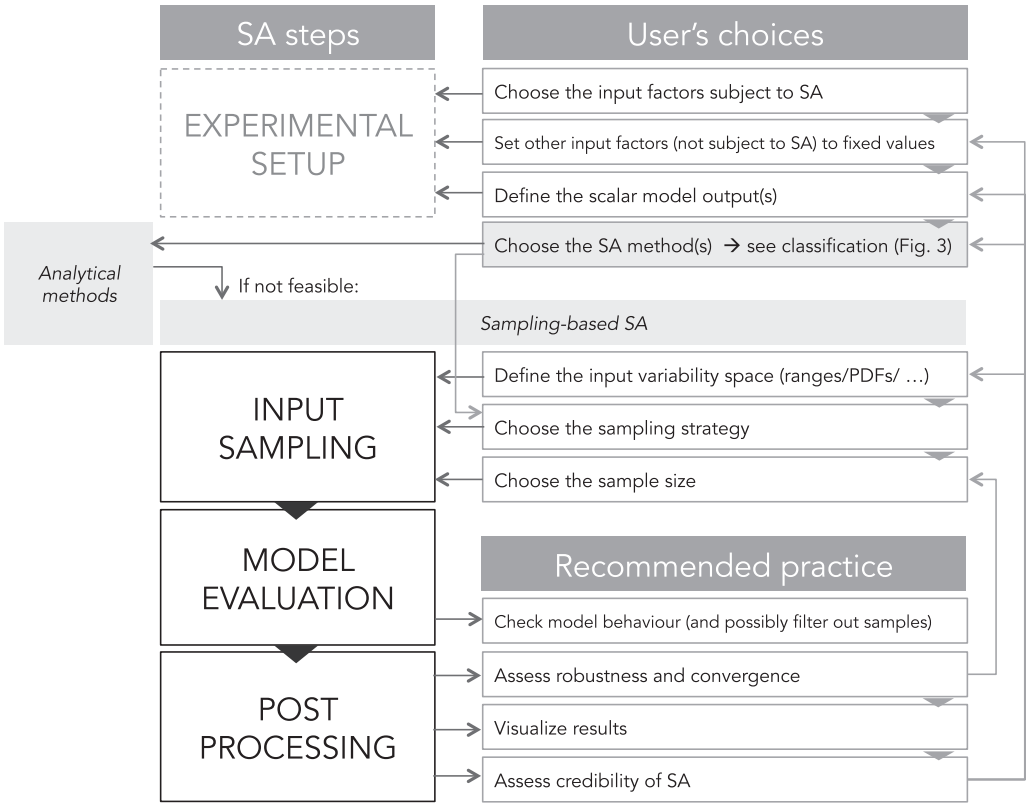


Fig. 4. Workflow for the application of Sensitivity Analysis, choices to be made and recommended practice for their revision.

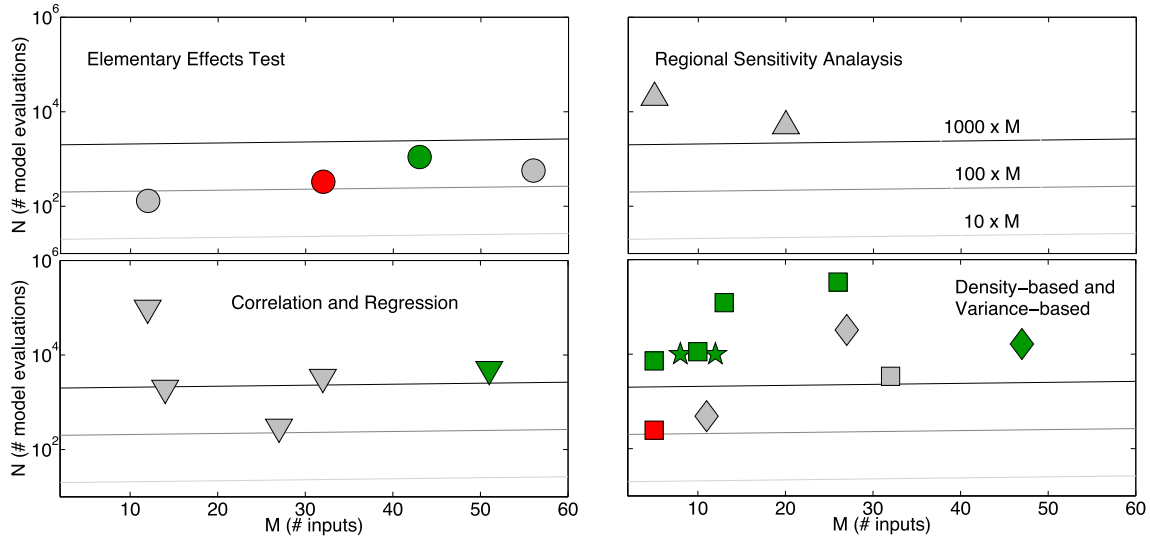


Fig. 5. Number of model evaluations (N) used in SA against the number of input factors (M) from the applications referenced in this paper. Green markers denote that the convergence of the sensitivity indices was reached, red markers that it was not reached, grey markers that convergence assessment was not reported in the paper. For density-based and variance-based (bottom right panel), squares refer first-order and total-order estimators via resampling technique (Saltelli et al., 2010), diamond denote applications of FAST/eFAST, and stars are application of the density-based δ-sensitivity method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

model evaluations required in the second step, as discussed for instance by Campolongo et al. (2011).

4.3. Define the input variability space

Whatever the chosen SA method, the first step of SA is the definition of the variability space of the input factors, i.e. the 'neighbourhood' of the nominal value $\bar{\mathbf{x}}$ in local SA, and the input variability space $\mathscr{X}$ in global SA. When using global methods where inputs are regarded as stochastic variables, like variance-based and density-based methods (Sections 3.5 and 3.6), their PDFs over the support $\mathscr{X}$ must also be defined. In the absence of specific

information regarding this choice a common approach is to assume independent, uniformly distributed inputs so that the problem reverts to the definition of $\mathscr{X}$ only.

When the input factors **x** are the model parameters, feasible ranges can often be defined based on their physical meaning or from existing literature, and further constrained using a priori information about the specific characteristics of the case study site (e.g. Bai et al., 2009). If observations of the simulated variables are available, another option is to first apply a preliminary Regional Sensitivity Analysis to assess whether literature ranges can be narrowed down by excluding sub-ranges producing a model performance below a prescribed acceptability threshold (e.g. Freer et al., 1996).

When the input factors **x** are the model forcing inputs, feasible ranges should account for the observational errors that can be expected from measuring devices, data pre-processing, interpolation, etc. Approaches to quantify data uncertainty vary depending on the type of variable under study and are gaining increasing consideration in the environmental modelling community. For an example of meteorological and water quality and quantity variables and their uncertainties see for instance McMillan et al. (2012). When suitable data are either unavailable or sparse, ranges or probability distributions can be elicited from experts. Several techniques and practical tools are discussed e.g. in O'Hagan et al. (2006) and in Morris et al. (2014).

While a review of the available data-based or expert-based methods to define the input variability space falls outside the scope of this paper, here we mainly want to point out that the definition of $\mathscr{X}$ (and possibly the associated probability distribution) is often one of the most delicate steps in the application of GSA. A number of studies demonstrate how different definitions of $\mathscr{X}$, each considered equally plausible by the analyst, can dramatically change the values of sensitivity measures and therefore the conclusions drawn from GSA (e.g. Shin et al., 2013). This is especially true for variance-based and density-based methods where the sensitivity measures are directly related to the output probability distribution, which is induced by the combination of the model structure (Eq. (1)) and the assumed input distributions.

In the following paragraphs we discuss two other specific issues that in our opinion deserve special attention when applying GSA to environmental models.

4.3.1. Handling unacceptable model behaviour

When dealing with complex environmental models, it may happen that a combination of input factors that a priori may seem feasible, generates a model's response that the analyst would reject as unacceptable (for instance, unacceptably large deviations from observations), or even causes the simulation to fail (for instance due to numerical instability). These simulations may be excluded from further analysis by adding a 'filtering' step before the post-processing step (see Fig. 2). An example is given in Pappenberger et al. (2008), where output samples associated with model performance below a prescribed threshold are discarded before the computation of the sensitivity indices. Kelleher et al. (2013) also compare the sensitivity estimates that are obtained before and after applying a performance criterion to screen out unacceptable parameter sets. Such critical look at the results of individual samples, or subsets of samples, is a practice we recommend since it may yield useful insights into the model behaviour and gives directions to revise the experimental setup of the SA exercise (for instance, to reduce or enlarge the input variability space).

4.3.2. Handling non-scalar or non-numerical input factors

GSA methods described in Section 3 are usually illustrated assuming that all the input factors are numerical scalar quantities (for instance the model parameters), so that a given combination of inputs can be represented by a vector $\mathbf{x} = [x_1, x_2, \ldots, x_M]$. However, in environmental modelling applications, candidate input factors may include entities that are not immediately represented by a scalar number, like for example the time series of forcing inputs (e.g. the input hydrograph in Fig. 1) or the model's spatial resolution. In order to include such input factors in SA, a link must be established between possible realizations of the non-numerical input factor and the values of a numerical quantity x_i. Ad hoc procedures can be used for specific types of factors: for instance, a time series of forcing inputs can be associated with a scalar characteristic used to design it (e.g. the intensity or the duration of a design storm event as in Hamm et al. (2006)) or with the scalar multiplier used to obtain it by perturbation of a reference time series (e.g. Singh et al., 2014). A more flexible procedure is the one described in Baroni and Tarantola (2014) (and references therein). Here, the variability space of each input factor is represented by a list of its possible realizations. Then, the index of each element in the list is the desired scalar quantity x_i, which is associated with a discrete uniform probability distribution. Following these definitions, sampling is performed with respect to the scalar indices $x_1,\ldots, x_M$, while the model is evaluated against the original input factors defined by the sampled indices. This procedure can be applied to any type of input, including non-numerical. However, it requires that post-processing uses output samples only, like for instance in variance-based or density-based methods, while it cannot be applied within the Elementary Effects Test or Regional Sensitivity Analysis, which by construction requires that the input variability space be a metric space (see Eqs. (4) and (7)).

4.4. Choose the sampling strategy

When sensitivity indices cannot be computed analytically, sampling-based sensitivity analysis (Fig. 2) must be used.

For OAT methods like the EET, several alternative strategies are available for sampling (see discussion in Section 3.2). For AAT methods like Correlation and Regression Analysis, Regional Sensitivity Analysis and density-based methods, in principle any random or quasi-random sampling technique can be used. Among these, the most commonly used in the GSA literature are Latin-Hypercube sampling and Sobol' quasi-random sampling. A practice-oriented introduction to these techniques can be found for example in Forrester et al. (2008) (Section 1.4) and Press et al. (1992) (Section 7.7).

Some other GSA methods may require a tailored sampling strategy. For example, the approximation of the first-order and total-order variance-based indices by the estimators discussed in Saltelli et al. (2010) (see Section 3.5) is based on a tailored two-stage procedure. First, $2n$ random samples are generated (so called *base* sample) using Sobol' quasi-random or Latin-Hypercube sampling; then, other Mn input samples are built by recombination of the vectors in the base sample. The FAST and eFAST approaches also require a tailored sampling strategy. In fact, the use of an efficient sampling strategy is what differentiates them from other estimators of variance-based indices, as described in Section 3.5.

We suggest that the implications of the sampling choice should be tested similar to the other choices made in the application of GSA. If computationally feasible, different strategies can be compared. However, it is likely that the definition of the input variability space or the output definition have a larger impact on the GSA outcome. Furthermore, independently of the chosen sampling strategy, the robustness of the sensitivity indices can be checked through confidence intervals, as discussed in the following sections.

4.5. Choose the sample size

The second choice to be made in sampling-based GSA is that of the sample size N. This choice has a dramatic impact on the overall computational burden, given that the execution of the model is usually far more computationally expensive than the post-processing step of estimating sensitivity indices. Therefore GSA users are typically confronted with the problem of finding a compromise between the need for keeping the sample size small and that of obtaining reliable estimates of the sensitivity indices. The solution to this problem is not unique and may significantly vary depending on the complexity of the model's response (Eq. (1)), which, however, is generally difficult to know before running the model.

Some suggestions for the choice of the sample size for the most widely used methods are reported in the literature. For instance, for the Elementary Effect Test, a common indication is to use $r = 10$ EEs, which results in a total number of $N = r(M + 1)$ model evaluations. However, to the authors' knowledge this choice seems to be motivated mainly by the need of keeping the total number of model evaluations limited rather than by a formal assessment of the reliability of the results. For example, Campolongo and Saltelli (1997) show that, with $r = 10$, the confidence bounds of the sensitivity indices obtained by bootstrapping are so large that factor ranking is essentially meaningless; Vanuytrecht et al. (2014) compute the EET sensitivity indices using an increasing number of samples and conclude that $r = 25$ is sufficient to discriminate between influential and non-influential factors (screening) while it is still not sufficient to stabilize factor ranking.

For variance-based indices computed using the efficient estimators discussed in Saltelli et al. (2010), the application of the resampling strategy to a base random sample of size n leads to a total of $N = n(M + 2)$ model evaluations. Common indications for n range from 500 to 1000 (Saltelli et al., 2008). However, application examples reported in the literature seem to suggest that the base sample size may significantly vary from one application to the other and that a much larger base sample might be needed to achieve reliable results (see datapoints in the bottom right panel of Fig. 5). Furthermore, the number of samples needed to reach stable sensitivity estimates can vary from one input factor to another, with low sensitivity inputs usually converging faster than high sensitivity ones (e.g. Nossent et al. (2011)).

The use of distribution functions in RSA usually provides quite robust sensitivity estimates even for relatively small sample sizes (see discussion in Section 3.4), a feature that made RSA particularly attractive when it was introduced in the early 1980s given that the computing resource for Monte Carlo sampling was very limited at the time. Correlation and regression methods are also generally applied to relatively limited datasets, typically around or less than $1000M$ model evaluations (see again Fig. 5 for some examples). However, it is difficult to provide general rules for these classes of methods especially because applications of RSA and Correlation and Regression methods rarely report a discussion of the appropriateness of the selected sample size (an exception is Kleijnen and Helton (1999b)).

To summarise, we can conclude that, roughly speaking, the number of model evaluations N increases with the number of inputs M by a factor in between 10 and 100 for multiple-starts derivatives, between 100 and 1000 for Regional Sensitivity Analysis and Correlation and Regression methods, and around 1000 or even more for density-based and variance-based methods (though significant reductions are obtained when using FAST or eFAST). However, these proportionality coefficients are expected to increase with M, and they can vary greatly from one application to another. Therefore, rather than providing specific indications on how to properly choose the sample size a priori, in the next subsection we discuss some techniques to verify *a posteriori* the appropriateness of the choice made.

4.6. Assess robustness and convergence

When applying sampling-based SA, sensitivity indices are not computed exactly but they are approximated from the available samples. The robustness and convergence of such sensitivity estimates should therefore be assessed, especially when obtained from samples of small/medium size.

Convergence analysis assesses whether sensitivity estimates are independent of the size of the input–output sample, i.e. if they would take similar values when using an (independent) sample of larger size. A simple and generic technique to address this question is to re-compute the sensitivity indices using sub-samples of decreasing size extracted from the original sample. The advantage of this approach is that it does not require running new model simulations, however it might overestimate the convergence rate because the sub-samples are not independent. Results of convergence analysis can be displayed in a 'convergence plot' like the one in Appendix A. Examples are given in Nossent et al. (2011) and Wang et al. (2013).

Robustness analysis assesses whether sensitivity indices are independent of the specific input–output sample, i.e. if they would take similar values if estimated over a different sample of the same size. Technique to address the question without running new model evaluations are subsampling and bootstrapping (Efron and Tibshirani, 1993; Romano and Shaikh, 2012). A discussion of the quality of bootstrapping-based confidence limits of some widely used sensitivity indices can be found in Yang (2011).

If convergence has not been reached and/or the confidence bounds are large, additional model simulations may be run and the sensitivity indices re-estimated over the increased sample. If this is not possible because of limited computing resources, some conclusions may still be drawn from the available results. In fact, even if the estimates of the sensitivity indices have not reached convergence, the screening of the non-influential input factors or the factor ranking might have stabilised (see for instance the discussion in Ziliani et al. (2013)).

While the evaluation of convergence and/or robustness is increasingly common in applications of variance-based methods, it is not equally common for other methods, for instance the Elementary Effects Test, although there is no technical reason not to extend the above described techniques to this approach (see for example the visualization of the EET results with bootstrapping in Appendix A). We suggest that the assessment of convergence and robustness of the estimated indices and the associated screening, ranking and mapping should be standard practice in any sampling-based SA exercise.

4.7. Visualize results

When dealing with large sets of sensitivity indices, the interpretation of SA results can be significantly enhanced by effective visualization tools that: (i) facilitate the identification of outliers and counterintuitive behaviours; (ii) help comparing results obtained by varying some of the underlying choices, e.g. different definitions of the input variability space or different sampling strategy; (iii) support the identification of temporal or spatial patterns in the output sensitivity; etc. Furthermore, effective visualization is key to improve the communication of SA results and conclusions.

General suggestions for visualizing scientific data effectively are presented in Kelleher and Wagener (2011). In Appendix A of this

paper we provide several examples of plots that have been employed in SA applications and that we found helpful. Some of these plots have been proposed for specific SA methods (e.g. the Elementary Effects Test or Regional Sensitivity Analysis) while others are meant to handle specific challenges. For instance, patterns plots (e.g. van Werkhoven et al., 2008) can be very effective to visualize large sets of sensitivity indices, e.g. when the number of input factors is large or when analysing the variations of output sensitivity across a wide temporal or spatial domain. They help highlighting patterns and trends although they do not allow for a detailed comparison between the exact index values.

Another challenge is to visualize multiple sensitivity attributes simultaneously, for instance first-order sensitivity, total-order sensitivity and interactions, in such a way that much information is conveyed without overloading the reader. Two types of plots that have been recently suggested to this end are Circos (Kelleher et al., 2013) and radial convergence diagrams (Butler et al., 2014). Our (subjective) experience is that viewers find radial convergence diagrams somewhat easier to interpret though both contain the same information.

Besides visualising sensitivity indices, it is often convenient to visualise the input and output samples for additional insights. For example, variance-based methods do not provide any mapping of the results into the input factors space, however some information about this mapping can be obtained by applying RSA or other visualization tools (e.g. scatter plots or parallel coordinate plots, see Appendix A) to the base sample generated for VBSA, at no additional computing cost.

4.8. Assess credibility

The robustness and convergence analyses discussed in Section 4.6 aim at assessing the uncertainty in the results of a specific SA method. Therefore they tell us about the reliability of the results within the context of that method. A different and equally relevant question is how much the method itself can be trusted, i.e. how suitable it is to address the questions it is expected to answer when applied to the problem at hand. For instance, variance-based methods rely on the assumption that variance is a sensible proxy for uncertainty, which may not be true for a highly-skewed output distribution. In this case, even if one were able to derive almost exact estimates of the variance-based sensitivity indices, they would not provide the correct ranking (a numerical example is given in Liu et al. (2006)). In other words, SA results may be very robust and yet not credible, and vice versa.

A way to assess credibility is by verifying that the underlying assumptions of a method are satisfied, for instance checking the linearity, monotonicity or smoothness of the input–output relationship of Eq (1) or the characteristics of the output distribution.

Another way is to compare SA results produced by different methods. As discussed in Section 4.2, the application of different GSA methods does not necessarily increase the computational burden since multiple approaches can be applied to the same input–output sample. If the screening/ranking results remain the same across different methods, the comparison reinforces the conclusions of SA. If instead there are contradictory results, it stimulates further investigations that may lead to understanding different aspects of the model's behaviour that are captured by different SA methods (see for instance the discussion in Pappenberger et al. (2008)). Also, specific techniques can be applied to validate SA conclusions, e.g. the visual test proposed by Andres (1997) to validate factor screening or the quantitative test based on the Kolmogorov–Smirnov statistic presented in Pianosi and Wagener (2015). Here conditional (on either sensitive or insensitive parameters) and unconditional output distributions are compared to check whether all insensitive factors have been identified. The limitation of these validation tests is that they require additional model runs.

Credibility assessment also involves the interpretation and explanation of the SA results. If unexpected results are obtained, for instance the output is highly sensitive to an input factor that was supposed hardly influential, the interpretation of the result could lead to either learning new aspects of the model behaviour or revising some of the choices made in the experimental setup, for example the definition of the output definition or of the input variability space.

5. Conclusions

In this paper we have provided a systematic classification of Sensitivity Analysis (SA) methods and discussed a workflow for its application, with the aim of providing the reader with the background needed:

- to further engage with the SA literature;
- to recognise the type of questions that could be addressed through SA;
- to choose the most suitable SA approach depending on the questions to be addressed, the available computing resources, and the characteristics of the problem at hand;
- to be aware of the key assumptions underlying each approach, its scope and limitations;
- to understand the typical workflow for applying SA;
- to be aware of the most sensitive choices that are made in the workflow and how to assess their impacts.

In doing so, we also highlighted some emerging trends in the SA literature that we consider of particular interest to the environmental modelling community. In particular:

(i) the application of SA to analyse the impact of non-numerical uncertain factors like model resolution or structure;
(ii) the application of time-varying and space-varying SA, which is made possible by increasing computing power and storage space, and which is a means to overcome the limitations of defining an 'aggregated' scalar output when dealing with dynamic models;
(iii) the application of SA for dominant-control analysis and robust decision-making, i.e. as a means to learn about the behaviour of models or systems.

We think that, among the topics for further research in the field, the following are of particular relevance for environmental modellers:

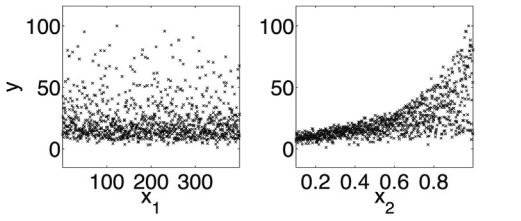
- developing multi-method approaches to overcome the limitations of individual SA methods;
- providing guidance and advice on convergence and robustness of different SA approaches;
- integrating the evaluation of model behaviour/performance in the estimation of sensitivity indices;
- improving techniques to analyse interactions between input factors: in fact, while information about factor interactions can be gathered as a byproduct of several SA techniques (for

instance, by looking at the standard deviation of the EEs or at the difference between total-order and first-order indices in variance-based SA), to our knowledge there is no SA method that has been specifically proposed to effectively investigate factor interactions;
- improving tools for visualisation and effective communication of SA results;
- reducing computing requirements for applications to complex environmental models, including the use of emulators.

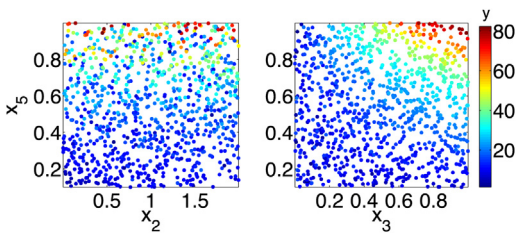
The authors thank three anonymous referees for very useful comments and suggestions that have greatly contributed to improving the manuscript. This work was supported by the Natural Environment Research Council (NERC) [Consortium on Risk in the Environment: Diagnostics, Integration, Benchmarking, Learning and Elicitation (CREDIBLE); grant number NE/J017450/1].

Appendix A. Examples of helpful visualization tools for global SA

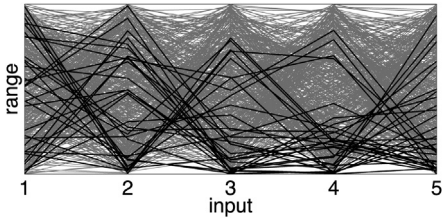
Visualize input/output samples



1. **Scatter plots** (or dotty plots): output samples against samples of the i-th input factor. One point (x_i^j, y^j) per input/output sample ($j = 1, \ldots, N$). Uniformly scattered points like in the left panel indicate low sensitivity (to x_1 here); emergence of patterns like in the right panel denotes high sensitivity. Useful for screening and ranking.

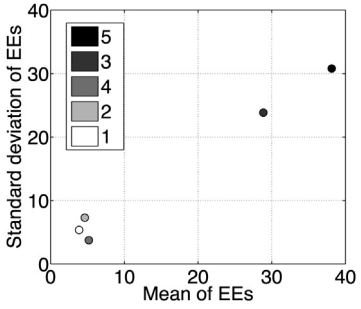


2. **Coloured scatter plots**: samples of i-th input factor against samples of the k-th, with marker colour proportional to the associated output value. One point (x_i^j, x_k^j) per input/output sample ($j = 1, \ldots, N$). Useful to detect interactions, which are highlighted by the emergence of colour patterns (as for instance in the right panel).

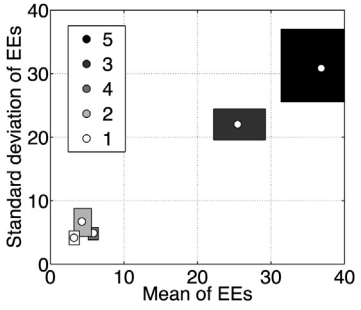


3. **Parallel coordinate plots**: distribution of input factors within their variability ranges. One line per sample $\mathbf{x}^j$ of input factors ($j = 1, \ldots, N$). Ranges are standardised to allow for comparison across factors. Lines highlighted in different colours correspond to 'particular' output values, for instance above a threshold. If highlighted lines cover the entire range of a factor (for instance black lines on factor number 2) sensitivity is low. If they concentrate in a subrange (as for instance for factor 5) then sensitivity is high. Useful for mapping.

Elementary Effects Test



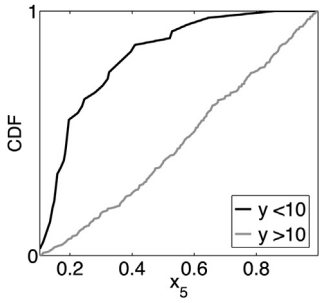
4. Average of Elementary Effects (EEs) versus their standard deviation. One point per input factor. The more to the right a point along the horizontal axis, the more influential the factor. The higher up a point along the vertical axis, the larger its degree of interactions with other factors. Useful for screening and ranking.



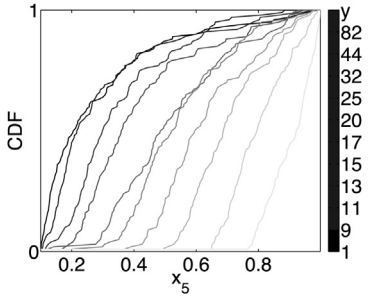
5. Same as before but with confidence bounds derived via bootstrapping around the mean and standard deviation of the EEs.

(*continued*)

Regional Sensitivity Analysis

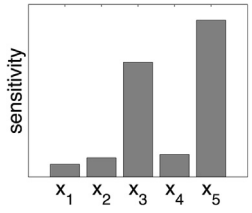


6. Empirical cumulative distribution function of the input samples associated with output values above/below a given threshold. One plot per input factor. The larger the distance between the two distribution functions, the more influential the factor. This plot can be also used to determine sub-ranges of the input factor that have no influence on the output above/below the threshold: these are the sub-ranges where the distribution functions are either zero or one (e.g. $x_5 > 0.8$ for $y > 10$ in this example). Useful for ranking and mapping.

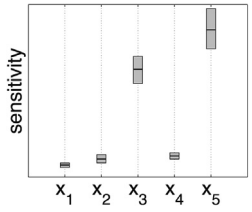


7. Empirical cumulative distribution function of input factors associated to output values within ten different ranges. Same as before without the need for specifying a threshold value.

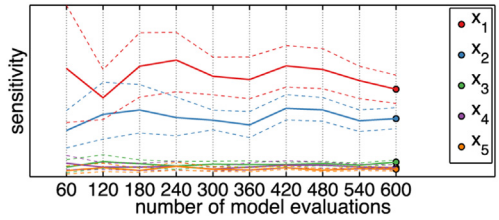
Visualize sensitivity indices



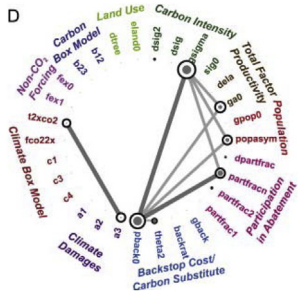
8. **Bar plot.** Value of sensitivity index for different input factors.



9. **Box plot.** Average value of sensitivity index over bootstrap resamples for different input factors (black line) and 90% confidence intervals.



10. **Convergence plot.** Sensitivity indices estimated using an increasing sample size (one line per factor). Dashed lines represent confidence bounds obtained at each sample size, for instance by bootstrapping.

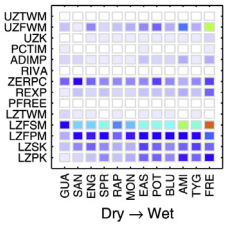


11. **Radial convergence diagrams.** For each input factor, the diagram shows: its direct (first-order) influence (proportional to the size of the inner circle); the total influence including interactions (size of the outer circle); the existence and extent of interactions between pairs of factors (lines and their width). Taken from Butler et al. (2014).

(*continued on next page*)

(*continued*)

12. **Circos.** For each input factor, the diagram shows: the total influence including interactions (proportional to the size of the pie slice on the outside of the circle); existence and extent of interactions between pairs of factors (inner connecting lines and their width). The direct (first-order) influence of each factor can be inferred as the portion of the total influence that is not connected to any other pie slice (i.e. the white space highlighted by the black arrows). Taken from Kelleher et al. (2013).



13. **Pattern plots.** For each study site (i.e. the watersheds listed on the horizontal axis) and each input factor (i.e. the model parameters on the vertical axis), the picture shows the output sensitivity via a colour scale (white denotes no sensitivity, red is maximum sensitivity). Study sites are ordered along the horizontal axis depending on their climate conditions, which facilitates visual investigation of trends and patterns linking parameter sensitivity to climate properties. Taken from van Werkhoven et al. (2008).

References

Anderson, B., Borgonovo, E., Galeotti, M., Roson, R., 2014. Uncertainty in climate change modeling: can global sensitivity analysis be of help? Risk Anal. 34 (2), 271–293.

Andres, T., 1997. Sampling methods and sensitivity analysis for large parameter sets. J. Stat. Comput. Simul. 57 (1–4), 77–110.

Bai, Y., Wagener, T., Reed, P., 2009. A top-down framework for watershed model evaluation and selection under uncertainty. Environ. Model. Softw. 24 (8), 901–916.

Baroni, G., Tarantola, S., 2014. A general probabilistic framework for uncertainty and global sensitivity analysis of deterministic models: a hydrological case study. Environ. Model. Softw. 51, 26–34.

Beven, K., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Adv. Water Resour. 16, 41–51.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249 (1–4), 11–29.

Borgonovo, E., 2007. A new uncertainty importance measure. Reliab. Eng. Syst. Saf. 92, 771–784.

Borgonovo, E., 2008. Sensitivity analysis of model output with input constraints: a generalized rationale for local methods. Risk Anal. 28, 667–680.

Borgonovo, E., 2010. A methodology for determining interactions in probabilistic safety assessment models by varying one parameter at a time. Risk Anal. 30 (3), 385–399.

Borgonovo, E., 2014. Transformation and invariance in the sensitivity analysis of computer experiments. J. R. Stat. Soc. Ser. B 76 (5), 925–947.

Borgonovo, E., Plischke, E., 2016. Sensitivity analysis: a review of recent advances. Eur. J. Oper. Res. 248 (3), 869–887.

Brown, C., Werick, W., Leger, W., Fay, D., 2011. A decision-analytic approach to managing climate risks: application to the upper great lakes. J. Am. Water Resour. Assoc. 47 (3), 524–534.

Butler, M.P., Reed, P.M., Fisher-Vanden, K., Keller, K., Wagener, T., 2014. Identifying parametric controls and dependencies in integrated assessment models using global sensitivity analysis. Environ. Model. Softw. 59, 10–29.

Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environ. Model. Softw. 22 (10), 1509–1518.

Campolongo, F., Saltelli, A., 1997. Sensitivity analysis of an environmental model: an application of different analysis methods. Reliab. Eng. Syst. Saf. 57 (1), 49–69.

Campolongo, F., Saltelli, A., Cariboni, J., 2011. From screening to quantitative sensitivity analysis. A unified approach. Comput. Phys. Commun. 182 (4), 978–988.

Castaings, W., Borgonovo, E., Morris, M., Tarantola, S., 2012. Sampling strategies in density-based sensitivity analysis. Environ. Model. Softw. 38 (0), 13–26.

Cloke, H., Pappenberger, F., Renaud, J., 2008. Multi-method global sensitivity analysis (mmgsa) for modelling floodplain hydrological processes. Hydrol. Process. 22 (11), 1660–1674.

Collins, M., Chandler, R., Cox, P., Huthnance, J., Rougier, J., Stephenson, D., 2012. Quantifying future climate change. Nat. Clim. Change 2, 403–409.

Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., Schaibly, J.H., 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. J. Chem. Phys. 59 (8), 3873–3878.

Demaria, E.M., Nijssen, B., Wagener, T., 2007. Monte Carlo sensitivity analysis of land surface parameters using the Variable Infiltration Capacity model. J. Geophys. Res. Atmos. 112 (D11).

Devenish, B., Francis, P.N., Johnson, B.T., Sparks, R.S.J., Thomson, D.J., 2012. Sensitivity analysis of dispersion modeling of volcanic ash from Eyjafjallajokull in May 2010. J. Geophys. Res. 117.

EC, 2009. Impact Assessment Guidelines. European Commission. Technical Report 92. http://ec.europa.eu/governance/impact/docs/key_docs/iag_2009_en.pdf.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC.

EPA, 2009. Guidance on the Development, Evaluation, and Application of Environmental Models. Technical Report EPA/100/K-09/003. Environmental Protection Agency. http://www.epa.gov/crem/library/cred_guidance_0309.pdf.

Forrester, A., Sobester, A., Keane, A., 2008. Engineering Design via Surrogate Modelling: a Practical Guide. John Wiley & Sons.

Freer, J., Beven, K., Ambroise, B., 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. Water Resour. Res. 32 (7), 2161–2173.

Gupta, H.V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol. Process. 22 (18), 3802–3813.

Guse, B., Reusser, D.E., Fohrer, N., 2014. How to improve the representation of hydrological processes in SWAT for a lowland catchment – temporal analysis of parameter sensitivity and model performance. Hydrol. Process. 28 (4), 2651–2670.

Hall, J., Boyce, S., Wang, Y., Dawson, R., Tarantola, S., Saltelli, A., 2009. Sensitivity analysis of hydraulic models. ASCE J. Hydraul. Eng. 135 (11), 959–969.

Hamm, N., Hall, J., Anderson, M., 2006. Variance-based sensitivity analysis of the probability of hydrologically induced slope instability. Comput. Geosci. 32 (6), 803–817.

Harper, E., Stella, J.C., Fremier, A., 2011. Global sensitivity analysis for complex ecological models: a case study of riparian cottonwood population dynamics. Ecol. Appl. 21 (4), 1225–1240.

Helton, J., 1993. Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. Reliab. Eng. Syst. Saf. 42 (2–3), 327–367.

Helton, J., Davis, F., 2002. Illustration of sampling-based methods for uncertainty and sensitivity analysis. Risk Anal. 22 (3), 591–622.

Herman, J.D., Kollat, J.B., Reed, P.M., Wagener, T., 2013. From maps to movies: high-resolution time-varying sensitivity analysis for spatially distributed watershed models. Hydrol. Earth Syst. Sci. 17 (12), 5109–5125.

Hill, M., Tiedeman, C., 2007. Effective Groundwater Model Calibration: with Analysis of Data, Sensitivities, Predictions, and Uncertainty. John Wiley & Sons.

Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis of nonlinear models. Reliab. Eng. Syst. Saf. 52 (1), 1–17.

Howard, R., 1988. Decision analysis: practice and promise. Manag. Sci. 34 (6), 679–695.

Iman, R., Helton, J., 1988. An investigation of uncertainty and sensitivity analysis techniques for computer models. Risk Anal. 8, 71–90.

Iman, R., Hora, S., 1990. A robust measure of uncertainty importance for use in fault tree system analysis. Risk Anal. 10, 401–406.

Kelleher, C., Wagener, T., 2011. Ten guidelines for effective data visualization in scientific publications. Environ. Model. Softw. 26 (6), 822–827.

Kelleher, C., Wagener, T., McGlynn, B., Ward, A.S., Gooseff, M.N., Payn, R.A., 2013. Identifiability of transient storage model parameters along a mountain stream. Water Resour. Res. 49 (9), 5290–5306.

Kleijnen, J., Helton, J., 1999a. Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1: review and comparison of techniques. Reliab. Eng. Syst. Saf. 65 (2), 147–185.

Kleijnen, J., Helton, J., 1999b. Statistical analyses of scatterplots to identify important factors in large-scale simulations, 2: robustness of techniques. Reliab. Eng. Syst.

Saf. 65 (2), 187–197.
Krykacz-Hausmann, B., 2001. Epistemic sensitivity analysis based on the concept of entropy. In: Proceedings of SAMO2001, CIEMAT, Madrid, pp. 31–35.
Kucherenko, S., Tarantola, S., Annoni, P., 2012. Estimation of global sensitivity indices for models with dependent variables. Comput. Phys. Commun. 183 (4), 937–946.
Lempert, R., Bryant, B., Bankes, S., 2008. Comparing Algorithms for Scenario Discovery. Working Papers WR-557-NSF. RAND Corp., Santa Monica, USA.
Lempert, R., Popper, S., Bankes, S., 2003. Shaping the Next One Hundred Years: New Methods for Quantitative, Long-term Policy Analysis. Monograph Reports MR-1626-RPC. RAND Corp., Santa Monica, USA.
Liu, H., Sudjianto, A., Chen, W., 2006. Relative entropy based method for probabilistic sensitivity analysis in engineering design. J. Mech. Des. 128, 326–336.
Ljung, L., 1999. System Identification: Theory for the User, second ed. PTR Prentice Hall, Upper Saddle River, NJ.
Massmann, C., Wagener, T., Holzmann, H., 2014. A new approach to visualizing time-varying sensitivity indices for environmental model diagnostics across evaluation time-scales. Environ. Model. Softw. 51, 190–194.
McMillan, H., Krueger, T., Freer, J., 2012. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. Hydrol. Process. 26 (26), 4078–4111.
Minunno, F., van Oijen, M., Cameron, D., Pereira, J., 2013. Selecting parameters for bayesian calibration of a process-based model: a methodology based on canonical correlation analysis. SIAM/ASA J. Uncertain. Quantif. 1 (1), 370–385.
Moore, C., Doherty, J., 2005. Role of the calibration process in reducing model predictive error. Water Resour. Res. 41, W05020.
Morris, D.E., Oakley, J.E., Crowe, J.A., 2014. A web-based tool for eliciting probability distributions from experts. Environ. Model. Softw. 52, 1–4.
Morris, M., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33 (2), 161–174.
Nguyen, T., de Kok, J., 2007. Systematic testing of an integrated systems model for coastal zone management using sensitivity and uncertainty analyses. Environ. Model. Softw. 22 (11), 1572–1587.
Norton, J., 2008. Algebraic sensitivity analysis of environmental models. Environ. Model. Softw. 23 (8), 963–972.
Norton, J., 2015. An introduction to sensitivity assessment of simulation models. Environ. Model. Softw. 69, 166–174.
Nossent, J., Elsen, P., Bauwens, W., 2011. Sobol sensitivity analysis of a complex environmental model. Environ. Model. Softw. 26 (12), 1515–1525.
Oakley, J., O'Hagan, A., 2004. Probabilistic sensitivity analysis of complex models: a bayesian approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 66, 751–769.
O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T., 2006. Uncertain Judgements: Eliciting Experts' Probabilities. Wiley.
Pappenberger, F., Beven, K., Ratto, M., Matgen, P., 2008. Multi-method global sensitivity analysis of flood inundation models. Adv. Water Resour. 31 (1), 1–14.
Park, C., Ahn, K., 1994. A new approach for measuring uncertainty importance and distributional sensitivity in probabilistic safety assessment. Reliab. Eng. Syst. Saf. 46, 253–261.
Pastres, R., Chan, K., Solidoro, C., Dejak, C., 1999. Global sensitivity analysis of a shallow-water 3D eutrophication model. Comput. Phys. Commun. 117, 62–74.
Paton, F.L., Maier, H.R., Dandy, G.C., 2013. Relative magnitudes of sources of uncertainty in assessing climate change impacts on water supply security for the southern Adelaide water supply system. Water Resour. Res. 49 (3), 1643–1667.
Peeters, L., Podger, G., Smith, T., Pickett, T., Bark, R., Cuddy, S., 2014. Robust global sensitivity analysis of a river management model to assess nonlinear and interaction effects. Hydrol. Earth Syst. Sci. 18 (9), 3777–3785.
Pianosi, F., Wagener, T., 2015. A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environ. Model. Softw. 67, 1–11.
Pianosi, F., Wagener, T., Rougier, J., Freer, J., Hall, J., 2014. Sensitivity analysis of environmental models: a systematic review with practical workflow. In: Vulnerability, Uncertainty, and Risk, pp. 290–299.
Powell, S., Baker, K., 1992. Management Science, the Art of Modeling with Spreadsheets, fourth ed. John Wiley & Sons, Hoboken, NJ, USA.
Press, W., Teukolsky, S., Vetterling, W., Flannery, B., 1992. Numerical Recipes in C, second ed. Cambridge University Press.
Prudhomme, C., Kay, A., Crooks, S., Reynard, N., 2013. Climate change and river flooding: part 2 sensitivity characterisation for British catchments and example vulnerability assessments. Clim. Change 119 (3–4), 949–964.
Rakovec, O., Hill, M.C., Clark, M.P., Weerts, A.H., Teuling, A.J., Uijlenhoet, R., 2014. Distributed Evaluation of Local Sensitivity analysis (DELSA), with application to hydrologic models. Water Resour. Res. 50 (1), 409–426.
Ratto, M., Castelletti, A., Pagano, A., 2012. Emulation techniques for the reduction and sensitivity analysis of complex environmental models. Environ. Model. Softw. 34, 1–4.
Reusser, D.E., Zehe, E., 2011. Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity. Water Resour. Res. 47 (7).
Romano, J., Shaikh, A., 2012. On the uniform asymptotic validity of subsampling and the bootstrap. Ann. Stat. 40 (6), 2789–2822.
Rosolem, R., Gupta, H., Shuttleworth, W., Zeng, X., deGoncalves, L., 2012. A fully multiple-criteria implementation of the Sobol' method for parameter sensitivity analysis. J. Geophys. Res. 117 (D07103).
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. Comput. Phys. Commun. 181 (2), 259–270.
Saltelli, A., D'Hombres, B., 2010. Sensitivity analysis didn't help. A practitioner's critique of the stern review. Glob. Environ. Change 20 (2), 298–302.
Saltelli, A., Marivoet, J., 1990. Non-parametric statistics in sensitivity analysis for model output: a comparison of selected techniques. Reliab. Eng. Syst. Saf. 28 (2), 229–253.
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis. The Primer. Wiley.
Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F., 2005. Sensitivity analysis for chemical models. Chem. Rev. 105, 2811–2828.
Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F., 2012. Update 1 of: sensitivity analysis for chemical models. Chem. Rev. 112, 1–21.
Saltelli, A., Tarantola, S., Chan, K.P.-S., 1999. A quantitative model-independent method for global sensitivity analysis of model output. Technometrics 41 (1), 39–56.
Shin, M.-J., Guillaume, J.H., Croke, B.F., Jakeman, A.J., 2013. Addressing ten questions about conceptual rainfall–runoff models with global sensitivity analyses in R. J. Hydrol. 503 (0), 135–152.
Sieber, A., Uhlenbrook, S., 2005. Sensitivity analyses of a distributed catchment model to verify the model structure. J. Hydrol. 310 (1–4), 216–235.
Singh, R., Wagener, T., Crane, R., Mann, M.E., Ning, L., 2014. A vulnerability driven approach to identify adverse climate and land use change combinations for critical hydrologic indicator thresholds: application to a watershed in Pennsylvania, USA. Water Resour. Res. 50, 3409–3427.
Sobol', I., 1993. Sensitivity analysis for non-linear mathematical models. Mathematical Modelling and Computational Experiment 1, 407–414, translated from Russian: I.M. Sobol', Sensitivity estimates for nonlinear mathematical models Mat. Model. 2 (1990), 112–118.
Sobol', I., Kucherenko, S., 2009. Derivative based global sensitivity measures and their link with global sensitivity indices. Math. Comput. Simul. 79 (10), 3009–3017.
Sorooshian, S., Farid, A., 1982. Response surface parameter sensitivity analysis methods for postcalibration studies. Water Resour. Res. 18 (5), 1531–1538.
Sorooshian, S., Gupta, V.K., 1985. The analysis of structural identifiability: theory and application to conceptual rainfall-runoff models. Water Resour. Res. 21 (4), 487–495.
Spear, R., Hornberger, G., 1980. Eutrophication in peel inlet. II. Identification of critical uncertainties via generalized sensitivity analysis. Water Res. 14 (1), 43–49.
Spear, R.C., Grieb, T.M., Shang, N., 1994. Parameter uncertainty and interaction in complex environmental models. Water Resour. Res. 30 (11), 3159–3169.
Stephenson, D., Doblas-Reyes, F., 2000. Statistical methods for interpreting Monte Carlo ensemble forecasts. Tellus A 52 (3), 300–322.
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. Reliab. Eng. Syst. Saf. 94 (11), 1735–1763.
Sudret, B., 2008. Global sensitivity analysis using polynomial chaos expansions. Reliab. Eng. Syst. Saf. 93 (7), 964–979.
Tang, Y., Reed, P., van Werkhoven, K., Wagener, T., 2007a. Advancing the identification and evaluation of distributed rainfall-runoff models using global sensitivity analysis. Water Resour. Res. 43 (6).
Tang, Y., Reed, P., Wagener, T., van Werkhoven, K., 2007b. Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. Hydrol. Earth Syst. Sci. 11, 793–817.
van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M., Srinivasan, R., 2006. A global sensitivity analysis tool for the parameters of multi-variable catchment models. J. Hydrol. 324 (1–4), 10–23.
van Werkhoven, K., Wagener, T., Reed, P., Tang, Y., 2008. Characterization of watershed model behavior across a hydroclimatic gradient. Water Resour. Res. 44 (1).
van Werkhoven, K., Wagener, T., Reed, P., Tang, Y., 2009. Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. Adv. Water Resour. 32 (8), 1154–1169.
Vanuytrecht, E., Raes, D., Willems, P., 2014. Global sensitivity analysis of yield output from the water productivity model. Environ. Model. Softw. 51, 323–332.
Vautard, R., Beekmann, M., Menut, L., 2000. Applications of adjoint modelling in atmospheric chemistry: sensitivity and inverse modelling. Environ. Model. Softw. 15 (6–7), 703–709.
Vrugt, J.A., ter Braak, C., Diks, C., Robinson, B., Hyman, J., Higdon, D., 2009. Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int. J. Nonlinear Sci. Numer. Simul. 10, 273–290.
Wagener, T., Boyle, D., Lees, M., Wheater, H., Gupta, H., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrol. Earth Syst. Sci. 5, 13–26.
Wagener, T., Gupta, H.V., 2005. Model identification for hydrological forecasting under uncertainty. Stoch. Environ. Res. Risk Assess. 19 (6), 378–387.
Wagener, T., McIntyre, N., Lees, M.J., Wheater, H.S., Gupta, H.V., 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis. Hydrol. Process. 17 (2), 455–476.
Wagner, B.J., Harvey, J.W., 1997. Experimental design for estimating parameters of rate-limited mass transfer: analysis of stream tracer studies. Water Resour. Res. 33 (7), 1731–1741.
Wang, J., Li, X., Lu, L., Fang, F., 2013. Parameter sensitivity analysis of crop growth

models based on the extended Fourier Amplitude Sensitivity Test method. Environ. Model. Softw. 48, 171–182.
Wilby, R.L., Dessai, S., 2010. Robust adaptation to climate change. Weather 65 (7), 180–185.
Yang, J., 2011. Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis. Environ. Model. Softw. 26 (4), 444–457.
Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. Water Resour. Res. 44 (9).
Young, P.C., Spear, R.C., Hornberger, G.M., 1978. Modeling badly defined systems: some further thoughts. In: Proceedings SIMSIG Conference, Canberra, pp. 24–32.
Ziliani, L., Surian, N., Coulthard, T., Tarantola, S., 2013. Reduced-complexity modeling of braided rivers: assessing model performance by sensitivity analysis, calibration, and validation. J. Geophys. Res. Earth Surf. 18, 1–20.