What has Global Sensitivity Analysis ever done for us? A systematic review to support scientific advancement and to inform policy-making in earth system modelling

Thorsten Wagener[1,2,*] and Francesca Pianosi[1,2]
[1]Department of Civil Engineering, University of Bristol, UK
[2]Cabot Institute, University of Bristol, UK
*Corresponding author: thorsten.wagener@bristol.ac.uk

Abstract

Computer models are essential tools in the earth system sciences. They underpin our search for understanding of earth system functioning and support decision- and policy-making across spatial and temporal scales. To understand the implications of uncertainty and environmental variability on the identification of such earth system models and their predictions, we can rely on increasingly powerful Global Sensitivity Analysis (GSA) methods. Previous reviews have characterised the variability of GSA methods available and their usability for different tasks. In our paper we rather focus on reviewing what has been learned so far by applying GSA to models across the earth system sciences, independently of the specific algorithm that was applied. We identify and discuss 10 key findings with general applicability and relevance for the earth sciences. We further provide an A-B-C-D of best practise in applying GSA methods, which we have derived from analysing why some GSA applications provided more insight than others.

1. Introduction

Computer models are essential tools in the earth system sciences. They underpin our search for understanding of earth system functioning and influence decision- and policy-making at various spatial and temporal scales. For example, computer models of the atmospheric system are used to produce short-term weather forecasts, which inform operational decisions at regional or local scale, or to make long-term projections of the global climate, which forms the basis of the international debate around climate change. Global hydrologic models can now provide a coherent picture of hydrological dynamics across our planet under past, current and potential future conditions (Schewe et al., 2014); while integrated assessment models integrate our climate system with the socio-economic behaviour of society to assess the consequences of future policy scenarios (Stanton et al., 2009). Many other examples of the value of computer models can be made for a variety of earth science areas, from atmospheric circulation (Cotton et al., 1995) to biogeochemical processes in the sea (Soetaert et al., 2000), from mantle dynamics (Yoshida and Santosh, 2011) to tsunamis impacts (Gelfenbaum et al., 2011).

A key issue in the development of computer models is that they can quickly exhibit complicated behaviours because of the potentially high level of interactions between their variables, and subsequently their parameters, even when they only represent a relatively low number of physical processes. The amount of internal interactions is destined to grow as we build models that are increasingly more detailed and applied to larger domains. Two key factors are boosting this process: the increasing availability of computing resources, which enables the execution of models at unprecedented temporal and spatial resolutions (Wood et al., 2011; Washington et al., 2012), and the increasing availability of earth observations that can be used to force computer models and evaluate their predictions (O'Neill and Steenman-Clark, 2002; Ramamurthy, 2006; Nativi et al., 2015). For example, Figure 1 shows the increase in resolution and components of climate system models that was made possible by the growth of computing power over the last decades.

Increasingly detailed computer models working at ever larger scales and finer resolutions are expected to play a key role in advancing the earth system sciences (Rauser et al., 2016; Wood et al., 2011; Bierkens et al., 2015), but this growth in model complexity also comes at a price. As the level of interactions between model components increases, modellers quickly lose the ability to anticipate and interpret model behaviour and hence the ability to evaluate that a model achieves the right response for the right reason (Beven and Cloke, 2012), i.e. that the model is consistent with the underlying 'perceptual model' of system functioning (e.g. Klemes, 1986; Grayson et al., 1992; Wagener and Gupta, 2005; Kirchner, 2006; Beven, 2007; Gupta et al., 2012; Hrachowitz et al., 2014). This issue is particularly problematic in earth system modelling where incomplete knowledge of the system makes it impossible to validate models simply based on fitting model predictions to observations. Oreskes et al. (1994) therefore suggest that models should rather be evaluated in relative terms, and model validation should consist in identifying the models that are free from detectable flaws and that are internally consistent. Therefore, in the remainder of this paper, we will rather use the term model 'evaluation' to refer to any kind of model assessment or validation.

Another difficulty in the application and evaluation of earth system computer models is that, even if internally consistent, their predictions may still be erroneous as models are often forced by input variables that are only known with a significant degree of uncertainty (McMillan et al., 2012). The difficulty is even greater for models with a large number of initial and boundary conditions, for which measurements may be erroneous or simply unavailable. The problem is sometimes seemingly mitigated by the growth in data products made available by recent advances in earth monitoring (Butler, 2007) and environmental sensing (Hart and Martinez, 2006). However, the translation of raw measurements into data products usable for the modelling purpose (for example, from a satellite measurement of soil microwave radiation to an estimate of the soil water content) requires a set of pre-processing calculations that constitute a modelling activity per se. As a consequence, distinguishing between possible errors in the "main" hypothesis (the earth system computer model) and other "auxiliary" hypotheses, such as the pre-processing of input data used to force the model, can be difficult (Oreskes et al. 1994).

Uncertainty about the forcing inputs of earth system models, and consequently about their predictions, may have at least two other origins besides measurement and pre-processing errors. One is the scarcity of observations that still affects many areas of the world, either because regions are too remote or because it is impossible to establish and maintain a reliable monitoring network (Blöschl et al., 2013; Hrachowitz et al., 2013). The other is the shrinking value of historical observations in a quickly-changing world (e.g. Jain and Lall, 2001). Traditionally many modelling studies have relied on the so called 'stationarity' assumption, i.e. the assumption that "natural systems fluctuate within an unchanged envelope of variability" (Milly et al., 2008), when time periods studied were not longer than maybe a few decades. This assumption implies that observations collected in the past can inform the construction of computer models that are intended to predict future conditions. The assumption is hardly acceptable in a world where human activities are exerting an unprecedented influence on natural systems leading to unprecedented rates of environmental change (Crutzen and Stoermer, 2000). As socio-economic and technological changes are largely unpredictable, they introduce significant uncertainty about future properties of the earth system and dramatically limit our ability to make quantitative predictions about its evolution (Wagener et al., 2010)

Lack of transparency about the scope of validity, the limitations and the predictive uncertainty of earth system computer models is not just a challenge for model developers but also for the users of the model outputs, such as environmental managers and policy-makers. Inadequate description of the uncertainties that affect model predictions may lead model users to overestimate the model's predictive ability which might create the false belief that the model can adequately reproduce all the consequences of the decisions to be made. On the other hand, ineffective communication of those uncertainties may induce decision-makers to underestimate the model's predictive ability and lead to rejecting the model predictions completely (Saltelli and Funtowicz, 2013).

The discussion so far highlights the importance of investigating uncertainty propagation in computer models in earth system science for both scientific and operational purposes. This task is often performed by rather simple approaches where uncertain input factors (such as input (forcing) data, model parameters or even underlying assumptions) are changed one-at-a-time and the effect in model predictions is assessed either visually or through simple quantitative indicators such as "the amount of change in model predictions for a fixed variation of the investigated input". However, this approach quickly becomes cumbersome if one has to investigate a large number of uncertain input factors. It also does not guarantee to provide a full picture of the model's behaviour given that only a limited number of input variations can be tested manually. Therefore, there is an increasing agreement that more structured, transparent and comprehensive approaches should be used to fully explore the impacts of input uncertainties on computer model predictions. Global Sensitivity Analysis

(GSA) is a set of statistical analysis techniques that provides such a structured approach (Saltelli et al., 2008). GSA can address questions like:

- Which variable (or component) of a computer model mostly influences model predictions, when and where? Hence, is the model's behaviour consistent with our conceptual understanding of the system functioning?
- Which uncertain input (or assumption) mostly contributes to the uncertainty in the model predictions? Hence, where should we focus efforts for uncertainty reduction?
- Can we find thresholds in the input factor values that map into specific output regions (e.g. exceeding a stakeholder-relevant threshold) of particular interest? Hence, what are the tipping points that, if crossed, would bring the system to specific conditions we want to avoid or want to reach?
- How robust are model predictions to modelling assumptions? Hence, how much would model-informed decisions change if different assumptions were made?

GSA has the potential to massively advance the value of computer models in the earth system sciences, contributing to improved model development, better evaluation and more robust decision-making. However, despite such potential, the application of GSA in many areas of earth system sciences is still relatively limited. A recent literature survey by Ferretti et al. (2016) showed an increase in the share of scientific articles using the term 'sensitivity analysis' (SA) since the year 2004. They also found that the largest fraction of those papers uses a 'local' approach, whose differences with respect to the 'global' approach, on which this paper focuses, will be clarified in the next section. We therefore believe that there is a lot of potential to further expand the use of GSA and benefit from its strengths.

The goal of this paper is to demonstrate the value of GSA for the construction, evaluation and use of earth system models by showing examples of what its application has achieved so far for scientists, modellers and policy-makers. We do not cover in-depth mathematical aspects of GSA algorithms, which the interested reader may find in other recent reviews, e.g. Norton (2015) and Pianosi et al. (2016). Also, differently from recent special issues and books on GSA applications to earth system models and observations (e.g. Kettner and Syvitski (2016) and Petropoulos and Srivastava (2017)), which focus on individual methodological advances and novel applications of GSA, our aim is to provide a synthesis of some key and generic lessons that the earth science community has learnt through the application of GSA over the last 15 years. Through such review we hope to increase the appreciation of the approach in a wider community and promote its uptake by a larger number of earth system scientists.

In the next Section we introduce key definitions and concepts that are needed to understand the basic functioning of GSA and organise them into key guidelines for GSA application. Then, we present several examples from the literature where GSA was used to address the issues discussed in the

Introduction section on the topics of construction, evaluation and use of computer models for earth sciences. Again, we organise this literature review into 10 generic lessons learnt through the application of GSA to earth system models. We conclude our paper with what we think is an "A-B-C-D" for future research and applications of GSA.

2. A brief Introduction to GSA

In this section, we discuss the basics of Sensitivity Analysis (SA) in general and Global Sensitivity Analysis (GSA) in particular. We also provide key guidelines for the application of GSA to earth system models. We use the term 'model' to refer to a numerical procedure that aims at reproducing the behaviour of earth system components, typically via numerical integration of differential equations over a space and time domain. Because we assume such a numerical procedure to be implemented by a computer algorithm, we could equally use the term 'computer model' in this context. We further call 'input factor' any element that can be changed before running the model, and 'output' any variable that is obtained after the model's execution.

Figure 2(a) provides examples of input factors. They can be broadly divided into four groups:
[1] The equations implemented in the model to represent physical processes, for which our often-incomplete scientific knowledge might offer multiple options (including omissions, if a process is deemed negligible given the scope and scale of the application).
[2] Set-up choices that are needed for the execution of the model on a computer, for example the selection of temporal or spatial resolutions for numerical integration of the model equations.
[3] The numerical values to be attributed to the parameters appearing in the model equation, which are often 'effective' parameters i.e. quantities that cannot directly be measured due to a scale mismatch between model element and instrument footprint (Beven, 2002). These parameters are called 'effective' since they are typically set to values that make the model component, e.g. a soil moisture store, approximate the behaviour of the real-world system without representing the full heterogeneity of that system (Wagener and Gupta, 2005).
[4] Any input data (system forcing, initial conditions and boundary conditions), which may be uncertain due to errors in both measurement and pre-processing (Figure 2(b)). Examples of pre-processing errors include the spatial interpolation of point observations or the manipulation of raw observations (such as remote sensing data) to transform them into the actual variable needed as input to the computer model. The importance of initial and boundary conditions varies significantly with the type of model, for example the simulation results of an atmospheric model might be very sensitive to uncertainty in initial conditions, while those of a groundwater model will depend more strongly on the assumed boundary conditions. The impact of initial conditions will also grow over the simulation period for some models, e.g. numerical weather prediction models, while it will diminish with time for others, such as rainfall-runoff models,

which means it might be less relevant if a sufficiently long warm-up period is available in such cases.

The specific goal of SA is to investigate the relative influence that input factors have on one or more model outputs. If the relationship between input factors and output is nonlinear, then small variations of an input factor (e.g. x_i) may induce large variations in the output (y), while large variations of another input factor (x_j) may induce much lower variations in the output. In such cases we would say that x_i is more influential than x_j, or equivalently that y is more sensitive to x_i than to x_j. Sometimes, output sensitivities can be estimated by analysing the model equations directly (*algebraic* SA). However, when the relationships between input factors and outputs are numerous and complex, sensitivities can only be discovered 'empirically', i.e. by running the model against different combinations (samples) of the input factors and by analysing the statistical properties of the input-output sample (*sampling-based* SA). Since algebraic SA is rarely a viable option in earth system models, in this paper we focus on sampling-based SA and refer the reader to Norton (2008; 2015) for algebraic SA.

The following sections briefly outline and discuss key elements in any Global Sensitivity Analysis process. We focus mainly on the key choices a GSA user has to make in this process.

2.1 Multiple definitions of the model output are possible

The model output y can be any variable that is obtained after model execution and that is of interest for the user, for example the predicted value of the system state at a prescribed time or location, or a summary metric such as the average (or any other statistic) of time-varying and spatially-varying states (Figure 2(c)). If observations of a simulated variable are available, the output y can also be defined by an error metric that measures the distance between observed and simulated variables, e.g. the mean squared error. In this case, what is called 'output' for the purposes of SA is not the 'output' of the computer model but rather a measure of the model's predictive accuracy (or 'objective function' in the automatic calibration literature).

2.2 Global methods measure direct and joint effects of input factors across their variability space (so no baseline point needs to be defined)

The simplest and most intuitive way to perform sampling-based SA is by a so-called 'One-At-a-Time' (OAT) approach. Here, baseline values for the input factors have to be defined and the input factors are varied, one at a time, by a prescribed amount (perturbation) while all others are held at baseline values. An example of OAT sampling for the case of 3 input factors is shown in Figure 3(a). SA results can be displayed for instance using a tornado plot (Figure 3(b)), which shows the output variations from the baseline, sorted from largest to smallest. If the perturbations applied to the baseline are small, the analysis is referred to as *local* SA, and output sensitivities can be measured by the (approximate) output derivatives at the baseline point.

The OAT approach is appealing as it calculates the variation in the model output in relation to a baseline, which is easy to interpret if the baseline has a clear meaning for the model user, for example the 'default' model set-up or the 'optimal' set-up after model calibration. Local methods are widely applied in different fields of study – especially where the feasible number of model runs is a limiting factor (Hill et al., 2016). However, the OAT approach has two main disadvantages. Firstly, OAT sampling only explores a small portion of the space of variability of the input factors, especially as the number of input factors increases. Therefore, the OAT approach is mostly useful if one is interested in exploring the model behaviour in relation to the baseline rather than across the entire space of input variability. Secondly, the OAT approach cannot detect interactions between input factors, i.e. the fact that the joint perturbations of two (or more) input factors may induce larger (or smaller) output variations than the perturbation of each individual factor. The latter problem can be partially overcome in local SA, where second-order derivatives of the output can be estimated with a relatively small number of additional model runs, thus providing information about local interactions between input factors (see Norton (2015) for more details). However, such sensitivity information is only valid in the neighbourhood of the baseline point, which may be limiting if one needs to investigate the effects of larger deviations or if there is simply no 'baseline' point of particular interest.

To address these issues and investigate the effects (direct and/or through interactions) of input variations regardless of a baseline, 'global' approaches to sensitivity analysis (GSA) have been proposed. In GSA, all input factors are varied simultaneously with the objective of covering their joint variability space as evenly as possible in accordance with the distributions underlying each factor (Figure 3(c)). Different random sampling (e.g. Latin-Hypercube) or quasi-random sampling (e.g. Sobol') techniques can be applied to this end and/or combined with OAT approaches – as done for example in multiple-start OAT approaches where multiple baseline points are randomly selected within the variability space of inputs (as further discussed in Sec. 2.3). The model outputs obtained for all the sampled input factors can then be analysed qualitatively (via visualisation techniques) and/or quantitatively (via statistical techniques). Quantitative GSA methods typically provide a set of sensitivity indices (Figure 3(d)), which measure the overall effects on the output from varying each input factor, usually on a scale from 0 to 1. A simple practical example of how to visualise and interpret a set of global sensitivity indices is given in Figure 4. Examples of how global sensitivity indices can help overcome the limitations of OAT approaches and avoid missing or misclassifying key sensitivities are given for example by Saltelli and D'Hombres (2010) and Butler et al. (2014).

2.3 Method choice matters as it can result in different sensitivity estimates (so, using multiple methods is advisable)

Global sensitivity indices can be defined in several different ways. A review of available methods is given for example by Pianosi et al. (2016) where a broad classification was proposed comprising four classes: (1) multiple-start perturbation approaches, where global sensitivity is obtained by aggregation of

'OAT' sensitivities obtained at different baseline points (e.g. the Elementary Effects Test or method of Morris); (2) correlation and regression approaches, where sensitivity is measured by the correlation between input and output samples; (3) regional sensitivity analysis (or Monte Carlo filtering) methods, where sensitivity is related to variations in the distributions of input factors induced by conditioning the outputs; and (4) variance-based and density-based approaches, where sensitivity is linked to variations in the output distribution induced by conditioning the inputs. A more in-depth discussion of these approaches and their advantages and disadvantages goes beyond the scope of this review and can be found in Saltelli et al. (2008), Norton (2015) or Pianosi et al. (2016).

GSA methods are based on different assumptions and use different definitions of sensitivity, which may lead to different sensitivity values and hence differences in outcomes of ranking and screening of the input factors (e.g. Tang et al. 2007a; Gan et al., 2014). A detailed discussion of this issue would be beyond the scope of this paper, but we generally suggest comparing the outcomes of different methods to understand the impact of the assumptions made. This multi-method approach can often be achieved very cheaply (in computational terms) since the same input-output sample can be used to estimate sensitivity indices according to different methods (e.g. Pianosi et al. (2017); Borgonovo et al. (2017); or the variogram analysis by Razavi and Gupta (2016), which encompasses variance-based and derivative-based methods as special cases).

2.4 The definition of the space of variability of the input factors has potentially a great impact on GSA results

Regardless of the GSA method chosen, a critical and yet not sufficiently explored issue is the choice of the space of variability from which input factors are sampled (i.e. the box in Figure 3c and the associated probability for sampling). When the uncertain input factors are model parameters, sampling is most often based on independent uniform distributions so that only the upper and lower bounds for each parameter have to be defined. Yet this definition of boundaries is often not easy to make, given the unclear physical meaning of many of the parameters used in earth system models, i.e. their 'effective' nature as discussed above. Some might vary from 0 to 1, and some might have at least a fixed lower bound (usually 0), but often this is not the case. Several papers (e.g. Kelleher et al., 2011; Shin et al., 2013; Wang et al., 2013) have demonstrated that, when multiple choices for parameter ranges are acceptable, changing the range for uniform sampling can significantly change the estimated sensitivity indices. Paleari and Confalonieri (2016) analysed other parameter distributions (e.g. normal) and found again that sensitivity estimates were strongly affected by the chosen distribution parameters. So, a pitfall of GSA is the possibly significant impact of the chosen input distributions, which should be carefully scrutinised.

Intuitively one might opt for relatively wide ranges to ensure that any impact of a parameter is captured. However, this can lead to the problem that poorly performing parameter values are included and impact the sensitivity analysis (e.g. Kelleher et al., 2011). A key to understanding this problem is to combine the GSA with an analysis of the performance of the simulations included in the analysis so to possibly exclude poorly performing simulations and avoid that they 'dominate' the estimation of sensitivity indices. Such a performance-based screening step would identify what is sometimes referred to as the behavioural simulations, i.e. those that produce a performance metric above (or below) a certain modeller chosen threshold value (Beven and Binley, 1992; Freer et al.,1996). It is generally good advice to perform the sensitivity analysis with and without considering such performance screening to understand the potential impact of poorly performing simulations on the sensitivity analysis result.

2.5 Sample size affects GSA results (so, the robustness of sensitivity indices should be checked)

As intuitively understandable from Figure 3(c), GSA requires many more input samples, and therefore more model executions, than OAT (local) SA. Therefore, when the computing time for each model run is long and/or a large memory space is required to store the output of each run, GSA can become difficult to apply. While the number of model executions (N) typically increases proportionally to the number of input factors (M), the proportionality relationship between M and N can vary significantly from one method to another, as well as from one application to another for the same method. As a rule of thumb, we would say that the most frugal methods (e.g. multiple-starts perturbation approaches) require around 10 to 100 model runs per uncertain input factor, while more expensive methods (e.g. variance-based) may require a number as large as 10,000 or even 100,000 times the number of input factors. This said, giving a 'one-fit-for-all' rule to link M to N can be misleading because it would assume that all GSA applications with the same number of factors require the same sample size, which is not the case (see for example Figure 5 in Pianosi et al. (2016) and Sarrazin et al. (2016)).

Given that the rules of thumb mentioned above can only provide very rough guidance and the actual numbers can vary greatly with the model under study (and even with the specific system to which the model is applied) we suggest that, rather than worrying too much about the number of samples a priori, it is better practice to analyse a posteriori the robustness of the GSA results. This can for example be achieved via bootstrapping, a resampling strategy that provides confidence limits on the sensitivity indices without the need for re-running the model (e.g. Sarrazin et al., 2016). Essentially, overlapping confidence limits between factors suggest that no robust conclusion between the importance of the factors can be drawn, and that the sample size should be increased.

Also, what sample size is adequate may vary depending on the GSA purpose. In fact, while obtaining precise estimates of sensitivity indices (i.e. with narrow confidence limits) may require a very large number of model executions, several studies (e.g. the one discussed below by Baroni and Tarantola (2014) and summarised in Fig. 5) have demonstrated that a robust separation between influential and non-influential factors (referred to as 'screening' in the GSA literature) or a robust ranking of the influential factors can often be obtained at much lower sample size. Therefore, for these purposes, a relatively small number of model executions is often sufficient even when applying a supposedly expensive GSA method (Sarrazin et al., 2016).

Another critical issue arises when the objective of GSA is the screening of non-influential input factors. If sensitivity indices where calculated exactly, one would simply test which factors have sensitivity indices of zero. However, approximation errors generally mean that values will deviate from zero even for non-influential factors. Additionally, users might also want to screen out factors with very little influence on the model output. Typically, users subjectively select a threshold to cope with this problem. Any factor showing a sensitivity index value below this threshold is assumed to be non-influential (e.g. Van Werkhoven et al., 2009; or Vanrolleghem et al., 2015 for an application and methodology to set the screening threshold). Alternatively, Zadeh et al. (2017) suggested the use of a dummy factor. This dummy factor is added to the model in a way that its variability does not influence the model output by design. Therefore, the sensitivity index value obtained for this dummy factor is an estimate of the approximation error only. Hence, it provides a threshold to discriminate between factors that can be confidently considered influential, since their sensitivity index exceeds this threshold, and those that may be non-influential, because they have an index around or below the threshold.

Another option to reduce the computational burden of GSA is the use of an emulator, i.e. a computationally efficient algebraic representation of the original complex computer model, which is able to approximate the input-output relationship of the original model and can be used in its place during computationally expensive GSA applications (e.g. Borgonovo et al. 2012; Ratto et al., 2012; Girard et al., 2016; Verrelst et al., 2016).

3. Review of GSA applications in earth system modelling and lessons learnt

In this section, we present applications of GSA to earth system models or to models of earth system components. We structure our review as 10 key lessons learnt through application of GSA and their implications for the construction and use of computer models in earth system sciences. These lessons cover different stages of the model building and application process, from model calibration (lessons 1,2,3,4), to the assessment and improvement of the data used to force or calibrate the model (4,5,6), model evaluation/validation (2,7,8) and the use of models in support of decision-making (9,10). We use examples from a variety of earth science disciplines although some disciplines are relatively more represented because the use of GSA in those areas is more widespread. One example of such an area is hydrology as is visible from the extensive review by Xiaomeng et al. (2015).

3.1 Only a small number of parameters typically dominates the variability of a given model output, though which parameters are dominant might vary with the chosen error or summary metric

A key observation when performing GSA to measure the relative importance of uncertain parameters is that the number of parameters that control the variability of a specific model output, be it defined as a summary or error metric, is rather low, typically in the order of 5 or 6 parameters. Other parameters might have a small direct effect or be involved through interactions, but they are not dominant.

An example is given in the top panel of Figure 5 where Wang et al. (2013) showed that out of 47 parameters of a crop growth model, less than 10 have a dominant influence on the selected output (final yield). Other examples with similar conclusions include Ben Touhami et al. (2013) for an ecological model, Girard et al (2016) for an atmospheric dispersion model; Bastidas et al. (1999) for a land surface model, Esmaeili et al. (2014) for a water quality model, and many others for hydrological models (e.g. Wagener et al., 2001; Van Werkhoven et al., 2009; Massmann and Holzmann, 2015; Hartmann et al., 2017; Shin and Kim, 2017).

The main implication of this limited number of influential parameters is that, if a computer model is mainly used to predict a specific summary metric (like annual yield as discussed in the previous paragraph), or it needs to be calibrated according to a given error metric (like the Root Mean Squared Error), it is often possible to significantly reduce the cost of model calibration (e.g. acquisition of new data to constrain the parameter values, or use of computationally-expensive automatic calibration algorithms to determine 'optimal' parameter estimates) by focusing on the small subset of parameters that are influential for that metric. The non-influential parameters can simply be set to 'default' values (taken from literature or previous applications) without significantly affecting model predictions or their accuracy.

On the other hand, this also means that there is an opportunity to define multiple output metrics (for example high and low river flows in hydrologic models), which are controlled by different parameters, to identify all or at least most of the model parameters. And indeed, GSA examples where multiple outputs were used, consistently demonstrated that different outputs are sensitive to different subsets of parameters (e.g. Bastidas et al., 1999; Tang et al., 2007a; Rosolem et al., 2012; Gan et al., 2015). An example is given in the bottom panel of Figure 5, taken from Song et al. (2012). Importantly for our argument here, the influential parameters vary somewhat across outputs but the total number per output remains small. A consequence of this finding is that if we want to understand the level of model complexity that is supported by a given dataset, we must take great care in defining several contrasting output metrics to maximize our chances of extracting all relevant information from the data (e.g. Gupta et al., 2008).

3.2 Dominant parameters can vary with the earth system (location) modelled

Besides varying with the output metric chosen by the modeller, parameter sensitivities can also vary when the same computer model is applied to different earth system locations (e.g. different catchments or drainage basins). We typically assume that our models have a degree of generality, i.e. that they are not only build to represent a single system, such as a particular catchment or hillslope, but that they can be used to represent the behaviour of the same type of system at different locations. A single model is then tailored to different locations when its model parameters are assigned values to reflect the specific characteristics of the system under study.

For example, Rosero et al. (2010) analysed a land surface model across different meteorological monitoring sites in the southern USA. The sites are located along a precipitation gradient and they also differ in land use and soil types. The assumption in their study was that the vegetation and soil parameters of the physically-based land surface model would be controlled by the differences in land use and soil type. However, they found that the dominant control on these parameters was the variability in precipitation, thus putting the physical interpretation of the parameters into question and suggesting that they are effective parameters. The importance of climate characteristics in conditioning parameter behaviour is further demonstrated in Van Werkhoven et al. (2008a). Here, parameter sensitivities for a conceptual rainfall-runoff model were computed in 12 catchments located in increasingly drier climates. The results (shown in Figure 6) revealed that parameter sensitivity varies with the output metric and application site, and that some of this variability can be linked to climatic characteristics, since patterns of increasing or decreasing sensitivity are found when moving from drier to wetter catchments. Other GSA applications showing similar variability of parameter sensitivities with the model's application locations include Confalonieri et al. (2010); Ben Touhami et al. (2013), Shin et al. (2013), Hartmann et al. (2013) and Herman et al. (2013).

A practical implication of this finding is that when calibrating a computer model for a new site, one should avoid making assumptions based on extrapolation from GSA results obtained elsewhere. For the purpose of better understanding the model behaviour, it is also interesting to investigate how parameter sensitivities vary from site to site and to test whether these variations can be linked to the site's physical or climatic characteristics. This could be reasonably expected when parameters are assumed to correspond to physical characteristics of the modelled system. Application of formal GSA may confirm or challenge this expectation.

3.3 Parameter sensitivity often varies in space (across the simulation domain) and in time (over the simulation period)

So far, we discussed GSA applications where the model output *y* is a scalar variable obtained by aggregation of the temporally and/or spatially distributed predictions of the model – either as an aggregation of the model outputs or state variables, or as an error metric derived from the difference between simulated and observed outputs (see Fig. 2c). In both cases, it is likely that this aggregation leads to a loss of information in both space and time. For example, when calibrating a rainfall-runoff model we normally estimate any measure of model performance (i.e. an error metric) over a sufficiently long and variable time period to trigger a range of responses of the model (Yapo et al., 1999). This maximises our chances of extracting sufficient information from the data to calibrate the parameters of interest. Conversely, the temporal aggregation does not reveal when in time each parameter is controlling the model's response and when it is not.

However, we can avoid this information loss by estimating disaggregated sensitivity indices in space and time. Applications of GSA where the analysis is applied to either individual time steps or to a small moving window period have become common. One interesting application of such time varying sensitivity analysis is a comparison between active model controls and expected process controls during different response modes of the system (e.g. Wagener et al., 2003; Reusser et al., 2011; Vezzaro and Mikkelsen, 2012; Guse et al., 2014; Pfannerstill et al., 2015). We will discuss this time varying analysis of parameter sensitivity in detail in section 3.7 in the context of model validation.

An example of spatial GSA results, focused on understanding how sensitivity indices vary across a model's domain, is given in Figure 7 for a computer model of chemical transport in the atmosphere. In this study, Brewer et al. (2017) showed that parameter sensitivities can exhibit complex spatial patterns, with some parameters being very influential but only in specific portions of the simulated spatial domain. These insights are very useful to tailor the model calibration efforts to where it is most effective, a piece of information that would otherwise be lost if applying GSA to aggregate output metrics. High levels of spatial variability in parameter sensitivities were also reported in Sieber and Uhlenbrook (2005), Hall et al. (2005), Treml et al. (2015), and in Savage et al. (2017). Tang et al. (2007b) and Van Werkhoven et al. (2008b) additionally linked the spatial variability of sensitivity indices to the spatial variability of forcing inputs.

Avoiding the loss of information induced by using aggregate output metrics has consequences for a range of activities, including model calibration, model validation and evaluation, observation network design etc. GSA can be used to understand which data periods or which domain parts contain information and which do not. Such analyses also highlight opportunities for creating more detailed models without adding parameters that cannot be identified. We provide further examples of the value of disaggregation in sections 3.7 and 3.8.

3.4 Uncertainty in the observations of the system outputs can prove as influential as uncertainty in the model parameters or forcing inputs

A big challenge in earth systems modelling is that the observations of the variables simulated by the computer model are often affected by large errors. If error metrics are very sensitive to such errors, their value for evaluating model accuracy and guiding model calibration is undermined. GSA can be used to explore the issue in a formal way by including errors in observations among the uncertain input factors subject to the sensitivity analysis (several techniques to do this are discussed in Sec. 4.3.2 of Pianosi et al. (2016)) and can be used to quantify their relative influence with respect to uncertain parameters or other factors.

Figure 8 depicts an example for a computer model of soil-water-atmosphere-plant dynamics by Baroni and Tarantola (2014). Here, uncertainty in soil moisture observations was found to influence model accuracy (measured using the root mean squared error between simulated and observed soil moisture) as much as uncertainty in the soil parameters. Moreover, the analysis showed a high level of interactions between the two uncertain factors, which implies that parameters can only be properly estimated if the uncertainty in the soil moisture observations is simultaneously reduced.

Uncertainty in the observations of the system outputs are regularly ignored in modelling studies once an error metric (which typically encapsulates a set of assumptions about the statistical properties of the observational errors) has been defined. Observations of system outputs are the main data that we evaluate our model against, both when estimating parameters (calibration) and when making predictions (what is sometimes called 'validation'). However, the potentially large uncertainties in such observations are increasingly recognised (see for example Westerberg and McMillan (2015) or Coxon et al. (2005) for an assessment of uncertainty in streamflow observations). We still require a better understanding of the implications of such uncertainties, especially when it comes to predictions of extremes (such as floods or heatwaves) for which observations are sparser and more error prone. This is an under-researched area in terms of GSA applications and where GSA has the potential to help us learn much about how influential such uncertainties can be.

3.5 Uncertainty in forcing input data affects model output uncertainty, not only because of errors in the measurements but also because of uncertainties in data pre-processing

Similarly to considering uncertainty in observations of the system output, GSA can also be used to analyse the impact of uncertainty in the input data of the model simulation, such as forcing data and initial or boundary conditions. For example, in the GSA application presented in Figure 8 (Baroni and Tarantola, 2014), errors in the time series of weather forcing data (air temperature, humidity, wind, rain and global radiation) were included in the analysis, although in this particular case they proved to have a relatively negligible effect on the model output. The result is case specific and other GSA applications found that uncertainty in the such inputs can at times be as influential as parameter uncertainty (e.g. Pianosi and Wagener (2016)). Figure 9 shows another interesting example taken from Yatheendradas et al. (2008) for a distributed hydrological model. Here, the forcing input was based on rainfall estimates from radar reflectivity measurements. The GSA showed that the uncertainty in the parameters translating the reflectivity signal into rainfall estimates (the so-called Z-R relationship) dominated the uncertainty in the flow predictions and was more influential than the uncertainty in the parameters or initial conditions of the hydrological model. Hence there is little to be gained by improving the hydrological model unless this pre-processing uncertainty can first be reduced.

This is a nice example of the difficulty in distinguishing between errors in the 'main' hypothesis, i.e. the earth system computer model, and in the 'auxiliary' hypothesis, i.e. the pre-processing procedure by which the model forcing inputs are generated (Oreskes et al., 1994). The latter is subject to uncertain assumptions that may prove as important as those embedded in the model. A typical problem in this context is that there is often little additional information available to determine such uncertainties (e.g. discussion in Beven and Cloke (2012)), which are therefore poorly understood. Approaches to back-out the uncertainty in the forcing data through inverse analysis of hydrological models have shown that the result depends strongly on other assumptions made (Renard et al., 2010; 2011). It is therefore important to understand the potential impact and relevance of such data pre-processing uncertainties so that efforts to reduce the final model output uncertainty can be pointed to the right factors (forcing data, parameters, output observations, etc).

3.6 Discrete modelling choices can be as influential as the uncertainty in parameters or in data

A common issue in earth system modelling is that model developers have to make discrete modelling choices or uncertain assumptions, for instance about which equation should be used to represent a specific process, or about the appropriate temporal or spatial resolution for the numerical integration of differential equations. One might therefore want to know how much these modelling choices matter given uncertainties in the model parameters, in the input data and in other elements of the modelling chain. Although much less explored, GSA can be used to address this question because it can quantify the relative influence of discrete modelling choices on model predictions. A simple strategy to achieve this aim is to include among the uncertain input factors x_i a discrete random variable that switches between a finite number of possible values. Each of these values corresponds to one of the possible discrete choices, so that the relative importance of that choice can be compared to that of the other uncertain factors.

An example of how to implement this strategy is provided again in the hydrology field by Baroni and Tarantola (2014). In their study, the model's vertical resolution was included in the GSA and found to play a negligible role with respect to parameter and data uncertainty as can be seen in Figure 8. Savage et al. (2017) instead found – using the same strategy – that the choice of the spatial resolution grid can have a significant influence on flood inundation predictions. It can even overtake the uncertainties in parameters and boundary conditions, although the ranking of these uncertain input factors varies in time, space and with the flood metric (output *y*) used. Another example, again for flood prediction, is the study by Abily et al. (2016) shown in Figure 10. Here GSA revealed that the chosen spatial resolution grid and the level of detail in describing above ground features affected water depth predictions more than errors in high-resolution topographic data.

The cited studies demonstrate that the importance of discrete modelling choices can be quantified in a structured way just as traditionally done for uncertainty sources such as parameters and forcing data. By doing so, the authors show that these discrete choices can be as significant as the continuous uncertainties more typically considered. By revealing when such discrete choices (or uncertainties) matter relative to other uncertainty sources, GSA provides a formal criterion to assess whether simplifying choices are acceptable. The analysis might also help to prioritise efforts for model improvement.

3.7 Consistency of model behaviour with the underlying perceptual model of the system is as important as the ability to reproduce observations

Another reason for using GSA is to evaluate the consistency between the model behaviour and the modeller's expectations, i.e. their 'perceptual model' of the system. GSA can contribute to this task by providing a formal assessment of the dominant controls on the model outputs, possibly disaggregated in space and time. A minimum requirement for a computer model to be considered acceptable is that these patterns of dominance are consistent with the modeller's understanding of the system's dominant drivers. We would say this criterion reflects Oreskes et al (1994) definition of model validation as demonstration of the model's "internal consistency".

An example is given in Figure 11 for the case of a hydrological model from the study by Reusser and Zehe (2011). Here, different groups of parameters represent different flow formation processes, which means they are expected to be more or less influential as hydro-meteorological conditions vary. The authors used time-varying GSA to quantify the temporal patterns of parameter influence and to identify events where those patterns were not consistent with expectations. Further scrutiny of simulated variables and sensitivities during these events helped to identify weaknesses in the model, e.g. missing processes, and systematic errors in the data used to assess model predictions. Other examples from hydrology include Wagener et al. (2003), Sieber and Uhlenbrook (2005), Pfannerstill et al. (2015), or Kelleher et al. (2015). This type of GSA utilization is also increasing in other areas of the earth system sciences, recent examples being Treml et al. (2015) (larvae dispersal in the ocean) and Arnaud et al. (2016) (soil-landscape evolution).

The conclusions of these studies are in line with the suggestion that consistency with the underlying perception of the real-world system is equally or potentially even more important than the optimal fit to available observations (Wagener and Gupta, 2005). Moving beyond model fit-to-data as the main model quality criterion, and rather focusing on the concept of consistency, has proven highly beneficial in model assessment (Martinez and Gupta, 2011; Euser et al., 2013; Hrachowitz et al., 2014; Pfannerstill et al., 2015; Shafii and Tolson, 2015). This finding has wide reaching implications that have so far not been fully appreciated, therefore leaving much room for further exploration. The current predominant approach to model evaluation still largely relies on the comparison of modelled and observed system outputs. In this traditional approach, a model is proclaimed to have been 'validated' if predictions are reasonably close to observations, particularly if the match is achieved on a sub-sample of the available dataset that was not used during model calibration. However, such an optimal fit of predictions to observations might be a relatively fragile result, as discussed for example in Beven and Binley (1992) and many subsequent papers by Beven. It is easy to unintentionally fit the noise in the data, which is often poorly known, or to obtain biased parameter estimates because of unaccounted for errors in either forcing inputs or output observations. Biased parameters estimates can also be obtained because the calibration dataset is small and/or not representative of the entire range of system conditions (a typical example in hydrology being a dataset that predominantly includes particularly dry or wet years). The bias can also be caused because any chosen error metric is likely to only capture some aspects of the system response. A typical example is the root mean squared error, which in a hydrological model would be largely controlled by the model's ability to reproduce flow peaks and less by its ability to reproduce other aspects of the hydrological system, such as the volume error. The problem is even more relevant if the modelling objective is hypothesis testing regarding dominant processes, or if the model is expected to provide longer term projections with changing boundary (e.g. climate) or system (e.g. land use) conditions (Fowler et al., 2016). Here understanding how the model represents system controls, and how such controls in the model might change in the future, is crucial and much more important than the model's ability to reproduce historical observations.

3.8 The design of observation networks and measurement campaigns can be more effective when analysing how the data information content varies in space and time

A question regularly encountered in earth system sciences is when and/or where measurements should be taken in order to maximize uncertainty reduction in model parameters, input forcing data, and ultimately model predictions. Cost-effective data collection requires a good understanding about which measurements are informative so that a targeted field campaign or an observational network can be designed (Moss, 1979).

An example is Raleigh et al. (2015), who used GSA to explore how different error characteristics (e.g. type, magnitude and distribution) in different forcing inputs (such as air temperature, precipitation, wind speed, etc.) influenced predicted snow variables such as snow water equivalent and ablation rates. Another example is provided by Wang et al. (2017), who analysed when isotope samples from streams should be collected to reduce the uncertainty in model parameters. Using time-varying GSA, they showed that specific time periods provide more informative samples for different parameters. Furthermore, they demonstrated that taking only 2 samples during the appropriate hydrologic conditions was as effective for uncertainty reduction as using all the 100 available samples from the entire data collection period. A slightly more complex issue is where to take measurements across a spatial domain. An example where GSA is used to answer this question is described in van Werkhoven et al. (2008b) (discussed in detail in section 3.3). Here, spatially-varying sensitivities of a distributed hydrologic model revealed that at least one more streamflow gauging station was required in the catchment to ensure identifiability of the model parameters.

We believe that this issue is one of the most interesting application areas for GSA in the years to come. Growing model complexity, dramatically increasing data volumes and novel sensors continually change the problem of which data are required for model identification and hypothesis testing. Addressing this problem demands powerful frameworks for the optimal design of measurement campaigns. Advances in modelling and sensing techniques also offer new interesting questions for GSA. For example, can we achieve a similar uncertainty reduction by applying many mobile and often much cheaper sensors over a short time period compared to what is achieved by a much more expensive continuous measurement station? Surprisingly though, this has so far been one of the less active areas of GSA studies.

3.9 If model predictions are expected to support decision-making, then they have to be sensitive to decision-related input factors

As discussed in the Introduction section, earth system models are increasingly used as tools to support decision-making, often in combination with socio-economic models. In this case, input factors of a single or of several models are related to possible planning/management decisions (for example, a model's input factor may define the land use practices in agricultural areas, or the operating rules for managing a reservoir, or do we have to evacuate an area due to a high probability of flooding). The model is then used to assess and compare the effects of different decisions (input factors) on an output of interest (for example, a drought index or the biomass produced in a growing season). In this context, GSA can be used to quantify the effects of decision-related input factors in the context of other uncertain factors (such as the parameters or forcing inputs of the earth system model) that also influence the output of interest but are outside the decision-maker's control. In fact, one would hope that the decision-related input factors exert an influence on the output that is at least comparable to that of other factors – otherwise the decision-making problem would be ill-posed. While this influence might be present in the real world, one cannot take for granted that it also happens in the computer model that is used to reproduce this reality. Indeed, models built for supporting decision-making typically integrate a range of interacting and often nonlinear components, which means that their responses to variations across their many input factors are not immediately obvious.

Examples of GSA applications to assess the relative influence of decision-relevant inputs include the study by Pastres et al. (1999), who applied GSA to a model of the Venice lagoon to estimate the relative importance of controllable drivers (e.g. nitrogen load or reaeration rate) and uncontrollable ones (e.g. dispersion coefficients or initial algae density) on anoxic crises. GSA results showed that variability in the initial algae density dominates the predicted duration of anoxic conditions, while the reaeration rate and the nitrogen load play a minor role. For management purposes this implies that measures aimed at short-term reduction of nitrogen loading may not be effective if not combined with long-term actions to reduce the accumulation of algae. Another example is the study by Xie et al. (2017), who used time-varying GSA of a hydrologic and sediment transport model to identify the dominant drivers of sediment export in the Three Gorge reservoir region and hence prioritise land management practices.

While models are indisputably irreplaceable and useful components of many decision-making processes, GSA can sometimes reveal that specific models are ineffective in their role. Several studies have used GSA to assess the robustness of model-informed decisions to the uncertain assumptions and choices made throughout the modelling exercise, which typically include both natural and socio-economic components.

A famous example is given by Saltelli and D'Hombres (2010), who used GSA to re-analyse the results of the Stern review (Stern et al., 2006) of economic impacts due to climate change. They found that predicted GDP losses varied dramatically with the assumptions made regarding both socio-economic factors (e.g. discount rate) and physical factors (e.g. climate response to GHG emissions), which implies that any inference drawn from such quantitative predictions would be very fragile. Another example of GSA of an integrated assessment model is given by Butler et al. (2014). Here the authors found that decision-relevant output metrics such as climate damage and abatement costs were largely insensitive to climate-related parameters (e.g. land use change, non-CO2 greenhouse gases, the carbon cycle model, and the climate model) because they were largely controlled by the uncertainty in economic parameters (e.g. the discount rate). The implication is that the performance of different simulated policy options is more strongly controlled by the socio-economic assumptions embedded in the model, than by their policy characteristics - in other words, the model predictions tell us more about the consequences of the assumptions made than they tell us about the different policy options. A third example is given by Le Cozannet et al. (2015), who used a time-varying GSA to determine the factors that mostly controlled the vulnerability of coastal flood defences over time (Figure 12). They found that – for their question – global climate change scenarios only matter for long-term planning while local factors such as near-shore coastal bathymetry – whose uncertainty is often neglected in impact studies – dominated in the short and mid-term (say over the next 50 years).

These studies demonstrate the importance of understanding the dominant controls of a model, in the context of the uncertainties that affects it, before the model can be used for impact assessment. It is crucial to understand the actual ability of a model to discriminate between decision options to avoid unreasonably conditioning the impact assessment results on the modelling choices made. While we assume that decision support models are generally build with the best of intentions, it is important to provide the evidence that the intentions have been achieved.

3.10 Even in the presence of practically unbounded uncertainties, learning about the relationship between model controls and outputs can be relevant for decision-making

Another area where GSA has been successfully employed is the investigation of so called 'deep uncertainties' (e.g. Bankes, 2002), i.e. input factors whose ranges of variability and probability distributions are poorly known and hence practically unbounded. A typical example are future carbon emission scenarios, which can diverge massively and whose probability of occurring is totally unknown.

The propagation of practically unbounded uncertain input factors through a model is technically feasible – it will be sufficient to consider all possible input values or sample from very wide ranges. However, the resulting model predictions are typically spread over such wide ranges that they are hardly usable to directly inform decision makers. Approaches that assess the risk and consequences of selecting a particular policy have been advocated as a more useful alternative strategy (Lempert et al., 2004). In these approaches, decision-relevant insights are extracted from the model simulations by adopting a so called 'bottom-up' (e.g. Wilby and Dessai (2010)) or 'scenario-discovery' strategy (Bryant and Lempert (2010)), which in turn can be implemented through a 'factor mapping' GSA technique. The idea is to start by defining thresholds (e.g. extreme values) for output variables that are relevant for decision-making, for example because exceeding the threshold is undesirable and would require taking actions. One can then create a large number of possible scenarios (e.g. of future climate) that are propagated through the model and for which the appropriate output variables are calculated. GSA can then be used to analyse these set of simulations and identify thresholds in the input factors that, if exceeded, would cause the output to cross the undesired thresholds. Decision-makers can further complement these results with other sources of information to assess how likely those input thresholds are to be crossed in the future and hence determine whether actions may be required.

Applications of this approach have been particularly reported for planning and management of water resource systems, some examples being Brown et al.

(2012), Kasprzyk et al. (2013), Singh et al. (2014) and Herman and Giuliani (2018). Figure 13 instead reports an example for landslide risk assessment taken from Almeida et al. (2017). Here the authors analysed the dominant controls of a rainfall-triggered mechanistic landslide model and found that uncertainty related to some physical slope properties can be as important as deep uncertainties related to future changes in rainfall in determining landslide occurrence (Figure 13).

The use of GSA for mapping of potentially very large and complex input-output datasets offers great potential for detailed analyses, especially in the context of highly uncertain decision-making problems. Maybe surprisingly, powerful GSA algorithms for mapping are not yet available, especially for situations where strong interactions between input factors exist, and most of the factor mapping applications mainly rely on visual tools more than quantitative approaches. This problem offers a lot of opportunity for research advancements. One very appealing feature of this strategy is that it requires the definition of vulnerability regions in the output space (e.g. what are critical thresholds such as the bankfull discharge in flood modelling). Defining this vulnerability space is often only possible for the stakeholder or the decision maker, which therefore offers communication opportunities between them and the modeller.

Outlook

Global Sensitivity Analysis (GSA) has become a widely-applied tool to understand earth system models across processes, scales and places. Our intention in this review paper was to organize and share some of the findings that have been made using GSA across earth system model applications. We believe that understanding what we have learned so far, and how these insights have been obtained, is key to guide further model development and to achieve robust decision-making using earth system model predictions. To this end, instead of attempting a comprehensive review of a large number of papers, we selected examples that we found particularly informative and accessible and discussed them in some depth. We tried as much as possible to provide additional references of other examples on the same issue (preferably in other earth system domains) as opportunity for further reading and study.

In addition to these findings, we also attempt here to identify some common characteristics in the way GSA was implemented in the most insightful applications. We call this an “ABCD” for maximising the scientific insights produced by GSA. It contains the following considerations:

A – *Adaptability* of the model to different environmental conditions changes the relevance of its input factors. It is therefore important to compare GSA results across a representative range of environmental conditions, including different places and different time periods.

B – *Behavioural* input factor samples might produce quite different sensitivity estimates compared to the samples taken from the full factor space. One should consider whether very poor performing input factor combinations are conditioning the GSA results.

C – *Combining* different SA methods, especially visual and quantitative ones, increases insight and robustness of the analysis. Using a single GSA approach, with its specific assumptions, might provide a skewed picture of the actual model behaviour.

D – *Disaggregating* inputs and outputs in both space and time increases the amount of information extracted during the analysis. A very simple, but also very effective way, to enhance learning during GSA studies is to estimate sensitivity indices for sub-periods or sub-domains.

Much, if not all, of earth system science relies on the use of models. Even if we do not use a computer model to simulate or forecast the system response, we are still likely to use a model of sorts to translate raw observations (e.g. from a remote sensing) into a variable of interest (e.g. soil moisture). Understanding how these models' function is crucial for robust science. The complexity of these models quickly outruns our ability to analyse their behaviour without formal approaches to do so. Computational science has in recent years been challenged to ensure that its studies and their outcomes are reproducible, transparent and robust (Peng, 2011; Hutton et al., 2016). This challenge is growing quickly in size with the continuing increase in model complexity which can make GSA problematic due to computational constraints. Nonetheless, we believe that GSA offers an important way to respond to this challenge and our review hopefully provides examples of how effective GSA can be in this regard.

Acknowledgments

This work was partially supported by the Natural Environment Research Council through the Consortium on Risk in the Environment: Diagnostics, Integration, Benchmarking, Learning and Elicitation (CREDIBLE) project (grant number NE/J017450/1). TW is also supported by a Royal Society Wolfson Research Merit Award and FP is further supported by the Engineering and Physical Sciences Research Council through an Early Career "Living with Environmental Uncertainty" Fellowship (grant number EP/R007330/1). We also thank the three anonymous reviewers for their constructive criticism which helped to improve the manuscript.

References

Abily, M., Bertrand, N., Delestre, O., Gourbesville, P., Duluc, C.-M. (2016), Spatial Global Sensitivity Analysis of High Resolution classified topographic data use in 2D urban flood modelling, Environmental Modelling and Software, 77, 183-195.

Almeida, S., Holcombe, E. A., Pianosi, F., and Wagener, T. (2017), Dealing with deep uncertainties in landslide modelling for disaster risk reduction under climate change, Nat. Hazards Earth Syst. Sci., 17, 225-241.

Arnaud J.A.M. Temme, Tom Vanwalleghem, (2016), LORICA – A new model for linking landscape and soil profile evolution: Development and sensitivity analysis, Computers & Geosciences, 90, 131-143.

Bankes, S. C., (2002). Tools and techniques for developing policies for complex and uncertain systems, Proceedings of the National Academy of Sciences, 99, 7263-7266.

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.

Bastidas, L. A., H. V. Gupta, S. Sorooshian, W. J. Shuttleworth, and Z. L. Yang (1999), Sensitivity analysis of a land surface scheme using multicriteria methods, J. Geophys. Res., 104(D16), 19,481‑19,490, doi:10.1029/1999JD900155.

Ben Touhami, H., Lardy, R., Barra, V., Bellocchi, G. (2013), Screening parameters in the Pasture Simulation model using the Morris method, Ecological Modelling, 266, 42-57.

Beven, K.J. (2002). Towards a coherent philosophy for modelling the environment. Proceedings of the Royal Society A, 458(2026), DOI: 10.1098/rspa.2002.0986

Beven, K. J. (2007), Working towards integrated environmental models of everywhere: Uncertainty, data, and modelling as a learning process, Hydrol. Earth Syst. Sci., 11(1), 460‑467.

Beven, K. and Binley, A. (1992), The future of distributed models: Model calibration and uncertainty prediction. Hydrol. Process., 6: 279–298.

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

Beven, K.J., and H.L. Cloke (2012). Comment on “Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth’s terrestrial water” by Eric F. Wood et al., Water Resour. Res., 48, W01801, doi:10.1029/2011WR010982.

Bierkens, M.F.P. et al. (2015). Hyper- resolution global hydrological modelling: what is next? Hydrological Processes, 29, 310-320.

Borgonovo, E., Castaings, W., Tarantola, S. (2012) Model emulation and moment-independent sensitivity analysis: An application to environmental modelling, Environmental Modelling & Software, 34, 105-115

Borgonovo, E., Lu, X., Plischke, E., Rakovec, O., Hill, M.C. (2017) Making the most out of a hydrological model data set: Sensitivity analyses to open the model black-box, Water Resources Research, 53(9), 7933-7950

Brewer, J. F., M. Bishop, M. Kelp, C. A. Keller, A. R. Ravishankara, E. V. Fischer, (2017), A sensitivity analysis of key natural factors in the modeled global acetone budget, J. Geophys. Res. Atmos., 122, 2043–2058

Brown, C., Ghile, Y., Laverty, M. and Li, K. (2012), Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector, Water Resour. Res, 48, W09537, doi:10.1029/2011WR011212

Bryant, B.P., Lempert, R.J., 2010. Thinking inside the box: a participatory, computer- assisted approach to scenario discovery. Technological Forecasting and Social Change 77, 34-49.

Butler, D., (2007), Earth Monitoring: The planetary panopticon, Nature, 450, 778-781.

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, 10e29.

Confalonieri, R. et al. (2010), Sensitivity analysis of the rice model WARM in Europe: Exploring the effects of different locations, climates and methods of analysis on model sensitivity to crop parameters, Environmental Modelling & Software, 25(4), 479-488.

Cotton, WR, Alexander, GD, Hertenstein, R, Walko, RL, McAnelly, RL, Nicholls, M (1995) Cloud venting - A review and some new global annual estimates, EARTH-SCIENCE REVIEWS, 39(3-4), 169-206.

Coxon, G., J. Freer, I. K. Westerberg, T. Wagener, R. Woods, and P. J. Smith (2015), A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water Resour. Res., 51, 5531–5546.

Crutzen, P.J. and E.F. Stoermer (2000), The ‘Anthropocene’. Global Change Newsletter. 41, 17-18.

de' Michieli Vitturi, M., Neri, A., and Barsotti, S. (2015), PLUME-MoM 1.0: A new integral model of volcanic plumes based on the method of moments, Geosci. Model Dev., 8, 2447-2463.

de' Michieli Vitturi, M., S.L. Engwell, A. Neri and S. Barsotti (2016), Uncertainty quantification and sensitivity analysis of volcanic columns models: Results from the integral model PLUME-MoM, Journal of Volcanology and Geothermal Research, 326, 77-91.

Deshmukh, A., and R. Singh (2016), Physio-climatic controls on vulnerability of watersheds to climate and land use change across the U. S., Water Resour. Res., 52, 8775-8793.

EC (2009), Impact Assessment Guidelines. Technical Report 92, SEC.

Esmaeili, S., Thomson, N.R., Tolson, B.A., Zebarth, B.J., Kuchta, S.H. and Neilsen, D. (2014). Quantitative global sensitivity analysis of the RZWQM to warrant a robust and effective calibration. Journal of Hydrology, 511, 567-579.

EPA (2009), Guidance on the Development, Evaluation, and Application of Environmental Models. Technical Report EPA/100/K-09/003 Office of the Science Advisor, Council for Regulatory Environmental Modeling

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. (2013, A framework to assess the realism of model structures using hydrological signatures, Hydrology and Earth System Sciences*, 17*, 1893-1912.

Ferretti F., Saltelli A., and Tarantola, S. (2016), Trends in sensitivity analysis practice in the last decade, Science of The Total Environment, 568, 666-670. https://doi.org/10.1016/j.scitotenv.2016.02.133

Fowler, K.J.A. , Peel, M.C., Western, A.W., Zhang, L., and Peterson,T.J. (2016), Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall‐runoff models, Water Resources Research 52(3), 1820-1846

Funtowicz, S.O. and Ravetz, J.R. (1990), Uncertainty and Quality in Science for Policy, Springer, Dordrecht (1990) p. 54.

Freer, J., Beven, K. and Ambroise, B. (1996). Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. Water Resources Research 32 (7), 2161-2173

Gan Y., Duan Q., Gong W., Tong C., Sun Y., Chu W., Ye A., Miao C., Di Z., (2014), A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model, Environmental Modelling & Software, 51, 269-285.

Gan, Y., X.-Z. Liang, Q. Duan, H. I. Choi, Y. Dai, and H. Wu (2015), Stepwise sensitivity analysis from qualitative to quantitative: Application to the terrestrial hydrological modeling of a Conjunctive Surface-Subsurface Process (CSSP) land surface model, J. Adv. Model. Earth Syst., 7, 648–669, doi:10.1002/2014MS000406

Gelfenbaum, G, Apotsos, A. Stevens, AW, Jaffe, B (2011), Effects of fringing reefs on tsunami inundation: American Samoa, EARTH-SCIENCE REVIEWS, 107(1-2), 12-22.

Girard, S., V. Mallet, I. Korsakissok, and A. Mathieu (2016), Emulation and Sobol' sensitivity analysis of an atmospheric dispersion model applied to the Fukushima nuclear accident, J. Geophys. Res. Atmos., 121, 3484–3496

Groves, D.G. and Lempert, R.J. (2007), A new analytic method for finding policy relevant scenarios, Global Environ Chang, 17, 73-85.

Grayson, R. B., I. D. Moore, and T. A. McMahon (1992). Physically based hydrologic modeling, 2, Is the concept realistic?, Water Resour. Res., 28(10), 2659-2666.

Guse, B., Reusser, D.E. and 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. Hydrological Processes, 28(4), 2651-2670.

Gupta, H.V., Wagener, T. and Liu, Y. (2008). Reconciling theory with observations: Towards a diagnostic approach to model evaluation. Hydrological Processes, 22. doi: 10.1002/hyp.6989.

Gupta, H.V., M.P. Clark, J.A. Vrugt, G. Abramowitz, and M. Ye (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resour. Res.*, 48, W08301, doi:10.1029/2011WR011044.

Hall, J. W., Tarantola, S., Bates, P.D., & Horritt, M. S. (2005). Distributed sensitivity analysis of flood inundation model calibration. American Society of Civil Engineers, Journal of Hydraulic Engineering, 131(2), 117-126

Hart, J.K, and K. Martinez (2006), Environmental Sensor Networks: A revolution in the earth system science? EARTH-SCIENCE REVIEWS, 78(3-4), 177-191.

Hartmann, A., Barberá, J.A., Andreo, B. (2017). On the value of water quality data and informative flow states in karst modelling. Hydrol. Earth Syst. Sci. 2017, 21, 5971–5985.

Hartmann, A., Wagener, T., Rimmer, A., Lange, J., Brielmann, H., and Weiler, M. (2013): Testing the realism of model structures to identify karst system processes using water quality and quantity signatures, Water Resources Research, 49, 3345–3358, 10.1002/wrcr.20229.

Herman, J. and Giuliani, M. (2018). Policy tree optimization for threshold-based water resources management over multiple timescales. Environmental Modelling and Software, 99, 39-51.

Herman, J. D., P. M. Reed, and T. Wagener (2013), Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior, Water Resour. Res., 49, 1400–1414.

Hill, M. C., Kavetski, D. Clark, M. Ye, M. Arabi, M. Lu, D. Foglia, L. & Mehl, S. (2016). Practical use of computationally frugal model analysis methods. Groundwater. DOI:10.1111/gwat.12330

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H.H.G. and Gascuel-Odoux, C. (2014). Process Consistency in Models: the Importance of System Signatures, Expert Knowledge and Process Complexity. *Water Resources Research*, 50, 7445-7469.

Hutton, C., T. Wagener, J. Freer, D. Han, C. Duffy, and B. Arheimer (2016), Most computational hydrology is not reproducible, so is it really science?, Water Resour. Res., 52, 7548–7555, doi:10.1002/2016WR019285.

Jain, S., and U. Lall (2001), Floods in a changing climate: Does the past represent the future?, Water Resour. Res., 37(12), 3193–3205.

Kelleher, C., T. Wagener, and B. McGlynn (2015), Model-based analysis of the influence of catchment properties on hydrologic partitioning across five mountain headwater subcatchments, Water Resour. Res., 51, 4109–4136

Kettner, A.J, Syvitski, J.P.M., (2016), Uncertainty and Sensitivity in Surface Dynamics Modeling, Computers & Geosciences, 90, Part B, 1-5.

Kirchner, J.W., (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42, W03S04, doi:10.1029/2005WR004362.

Klemes, V., (1986), Dilettantism in hydrology - transition or destiny?, Water Resources Research, 22(9s), 177S-188.

Kasprzyk, J.R., Nataraj S., Reed, P.M., Lempert, R.J. (2013), Many objective robust decision making for complex environmental systems undergoing change, Environmental Modelling & Software, 42, 55-71

Leamer, E.E. (2010), Tantalus on the Road to Asymptotia. Journal of Economic Perspectives, 24(2): 31-46.

Le Cozannet, G., Jeremy Rohmer, Anny Cazenave, Déborah Idier, Roderik van de Wal, Renske de Winter, Rodrigo Pedreros, Yann Balouin, Charlotte Vinchon, Carlos Oliveros, (2015) Evaluating uncertainties of future marine flooding occurrence as sea-level rises, Environmental Modelling & Software, 73, 44-56.

Lempert, R., N. Nakicenovic, D. Sarewitz, and M. Schlesinger (2004), Characterizing climate-change uncertainties for decision-makers, Clim. Change, 65, 1‐9.

Martinez, G. F. and Gupta, H. V. (2011), Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States, Water Resources Research, 47, W12540, doi:10.1029/2011WR011229.

Massmann, C. and Holzmann, H. (2015), Analysing the sub-processes of a conceptual rainfall-runoff model using information about the parameter sensitivity and variance. Environ. Model. Assess. 20 (1), 41–53.

McMillan, H., Krueger, T. and Freer, J. (2012). Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. Hydrological Processes, 26, 4078–4111.

Milly P.C., Betancourt J., Falkenmark M., Hirsch R.M., Kundzewicz Z.W., Lettenmaier D.P., Stouffer R.J. (2008), Stationarity is dead: whither water management? Science, 319(5863):573-4.

Moreau, P., Viaud, V., Parnaudeau, V., Salmon-Monviola, J., Durand, P. (2013), An approach for global sensitivity analysis of a complex environmental model to spatial inputs and parameters: A case study of an agro-hydrological model, Environmental Modelling & Software, 47, 74-87.

Moss, M.E. (1979). Some basic considerations in the design of basic hydrologic data networks. Water Resources Research, 15(6), 1673-1676.

Nativi, S., P. Mazzetti, M. Santoro, F. Papeschi, M. Craglia, and O. Ochiai, (2015), Big data challenges in building the global earth observation system of systems, Environ. Model. Softw., 68, 1-26.

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.

O'Neill, A. and Steenman-Clark, L. (2002), The computational challenges of Earth-system science, Phil. Trans. R. Soc. Lond. A, 360, 1267-1275.

Oreskes N, Shrader-Frechette K, Belitz K. (1994) Verification, validation, and confirmation of numerical models in the Earth sciences. Science, 263(5147), 641-6.

Paleari L. and Confalonieri R. (2016), Sensitivity analysis of a sensitivity analysis: We are likely overlooking the impact of distributional assumptions, Ecological Modelling, 340, 57-63.

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.

Peng, R. D. (2011), Reproducible research in computational science, Science, 334, 1226–1227, doi:10.1126/science.1213847.

Petropoulos G. and P.K. Srivastava, (2017), Sensitivity Analysis in Earth Observation Modelling, Elsevier, 448 pages.

Pfannerstill, M., Guse, B., Reusser, D. and Fohrer, N. (2015). Process verification of a hydrological model using a temporal parameter sensitivity analysis. Hydrology and Earth System Sciences, 19, 4365–4376. doi:10.5194/hess-19-4365-2015

Pianosi, F., Beven, K., Freer, J.W. Hall, J. Rougier, J. Stephenson, D.B., Wagener, T. (2016), Sensitivity analysis of environmental models: A systematic review with practical workflow, Environmental Modelling & Software, 79, 214-232.

Pianosi, F., and Wagener, T. (2016) Understanding the time-varying importance of different uncertainty sources in hydrological modelling using global sensitivity analysis. Hydrol. Process., 30: 3991-4003.

Pianosi, F., Iwema, J., Rosolem R. and Wagner T., (2017), A Multimethod Global Sensitivity Analysis Approach to Support the Calibration and Evaluation of Land Surface Models, Chapter 7 in: Sensitivity Analysis in Earth Observation Modelling, Edited by G.P. Petropoulos and P.K. Srivastava, Elsevier, pp.125-144

Raleigh, M.S., Lundquist, J.D., and Clark, M.P. (2015) Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework, Hydrol. Earth Syst. Sci., 19, 3153-3179.

Ramamurthy, M. K., (2006). A new generation of cyberinfrastructure and data services for earth system science education and research. Adv. Geosci., 8(8), 526 69-78.

Ratto, M., Castelletti, A. and Pagano, A. (2012). Emulation techniques for the reduction and sensitivity analysis of complex environmental models. *Environmental Modelling & Software*, 34, 1-4.

Rauser, F., M. Alqadi, S. Arowolo, N. Baker, J. Bedard, E. Behrens, N. Dogulu, L. Gatti Domingues, A. Frassoni, J. Keller, S. Kirkpatrick, G.

Langendijk, M. Mirsafa, S. Mohammad, A. Naumann, M. Osman, K. Reed, M. Rothmüller, V. Schemann, A. Singh, S. Sonntag, F. Tummon, D. Victor, M. Villafuerte, J. Walawender, and M. Zaroug (2016), Earth System Science Frontiers - an ECS perspective. Bull. Amer. Meteor. Soc., doi.org/10.1175/BAMS-D-16-0025.1.

Razavi, S. and Gupta, H.V. (2016), A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory, Water Resources Research, 52(1)

Renard B, Kavetski D, Leblois E, Thyer M, Kuczera G and Franks SW (2011) Towards a reliable decomposition of predictive uncertainty in hydrologic modeling: Characterizing rainfall errors using conditional simulation, Water Resources Research, 47, W11516.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour. Res., 46, W05521, doi:10.1029/2009WR008328.

Reusser, D.E., and Zehe, E., (2011). Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity. Water Resour. Res. 47 (7).

Reusser, D., W. Buytaert, and Zehe, E. (2011): Temporal dynamics of model parameter sensitivity for computationally expensive models with FAST (Fourier Amplitude Sensitivity Test.) Water Resour. Res., doi:10.1029/2010WR009947.

Rosero, E., Z.-L. Yang, T. Wagener, L. E. Gulden, S. Yatheendradas, and G.-Y. Niu (2010), Quantifying parameter sensitivity, interaction, and transferability in hydrologically enhanced versions of the Noah land surface model over transition zones during the warm season, J. Geophys. Res., 115, D03106.

Rosolem, R., H. V. Gupta, W. J. Shuttleworth, X. Zeng, and L. G. G. de Gonçalves (2012), A fully multiple-criteria implementation of the Sobol′ method for parameter sensitivity analysis, J. Geophys. Res., 117, D07103, doi:10.1029/2011JD016355.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, (2008), Global Sensitivity Analysis. The Primer. Wiley.

Saltelli, A., B. D'Hombres, (2010), Sensitivity analysis didn't help. A practitioner's critique of the Stern review, Global Environmental Change, 20(2), 298-302.

Saltelli, A. and S. Funtowicz, (2013), When all models are wrong, Issues in Science and Technology, 79.

Sarrazin, F.J., Pianosi, F. and Wagener, T., (2016), Global Sensitivity Analysis of environmental models: Convergence and validation. Environmental Modelling and Software, 79, 135-152

Schewe J, Heinke J, Gerten D, Haddeland I, Arnell NW, Clark DB, Dankers R, Eisner S, et al. (2014). Multimodel assessment of water scarcity under climate change. Proceedings of the National Academy of Sciences 111 (9): 3245-3250. DOI:10.1073/pnas.1222460110.

Shafii, M. and Tolson, B. (2015), Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives, *Water Resour. Res., 51*.

Shin, M.J., J.H. Guillaume, B.F. Croke, A.J. Jakeman, (2013), Addressing ten questions about conceptual rainfall–runoff models with global sensitivity analyses in R, J. Hydrol., 503(0), 135-152.

Shin, M.-J. and Kim, C.-S. (2017). Assessment of the suitability of rainfall–runoff models by coupling performance statistics and sensitivity analysis. Hydrology Research, 48.5, 1192-1213.

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., T. Wagener, R. Crane, M. E. Mann, and L. Ning (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, doi:10.1002/ 2013WR014988.

Soetaert, K., Middelburg, JJ, Herman, PMJ, Buis, K (2000), On the coupling of benthic and pelagic biogeochemical models, EARTH-SCIENCE REVIEWS, 51(1-4), 173-201.

Song, X., Bryan B.A., Paul, K.I., Zhao, G. (2012) Variance-based sensitivity analysis of a forest growth model, Ecological Modelling, 247, 135-143.

Stanton, E.A., Ackerman, F., Kartha, S., 2009. Inside the integrated assessment models: four issues in climate economics. Clim. Dev. 1, 166e184.

Stern, N., S.Peters, V.Bakhshi, A.Bowen, C.Cameron, S.Catovsky, D.Crane, S.Cruickshank, S.Dietz, N.Edmonson, S.-L.Garbett, L.Hamid, G.Hoffman, D.Ingram, B.Jones, N.Patmore, H.Radcliffe, R.Sathiyarajah, M.Stock, C.Taylor, T.Vernon, H.Wanjie, and D.Zenghelis (2006), Stern Review: The Economics of Climate Change, HM Treasury, London.

Tang, Y., Reed, P., Wagener, T., van Werkhoven, K., (2007a). Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. Hydrol. Earth Syst. Sci. 11, 793-817.

Tang, Y., Reed, P., van Werkhoven, K., Wagener, T. (2007b). Advancing the identification and evaluation of distributed rainfall-runoff models using global sensitivity analysis, Water Resour. Res., 43 (6).

Tarantola S., N. Giglioli, J. Jesinghaus, A. Saltelli. (2002), Can global sensitivity analysis steer the implementation of models for environmental assessments and decision-making?, Stoch Env Res Risk A, 16, 63-76.

Treml, E.A., Ford, J.R. Black, K.P. and S.E. Swearer, (2015), Identifying the key biophysical drivers, connectivity outcomes, and metapopulation consequences of larval dispersal in the sea, Movement Ecology, 3(17).

Vanrolleghem, P., Mannina, G., Cosenza, A., Neumann, M., 2015. Global sensitivity analysis for urban water quality modelling: terminology, convergence and comparison of different methods. Journal of Hydrology, 522, 339-352.

Van Werkhoven, K. T. Wagener, P. Reed, Y. Tang (2008a). Characterization of watershed model behavior across a hydroclimatic gradient, Water Resour. Res., 44 (1).

Van Werkhoven, K., Wagener, T., Tang, Y., and Reed, P. (2008b). Rainfall characteristics define the value of streamflow observations for distributed watershed model identification. Geophysical Research Letters, 35, L11403, doi:10.1029/2008GL034162.

Van Werkhoven, K., Wagener, T., Tang, Y., and Reed, P. (2009). Complexity reduction in multiobjective watershed model calibration. Advances in Water Resources, 32(8), 1154-1169.

Verrelst, J., Sabater, N., Rivera, J.P., Muñoz-Marí, J., Vicent, J., Camps-Valls, G., Moreno, J. (2016), Emulation of Leaf, Canopy and Atmosphere Radiative Transfer Models for Fast Global Sensitivity Analysis, Remote Sensing 8(8), 673.

Vezzaro, L. and Mikkelsen, P.S. (2012). Application of global sensitivity analysis and uncertainty quantification in dynamic modelling of micropollutants in stormwater runoff. Environmental Modelling & Software, 27-28, 40-51.

Wagener, T. and Gupta, H.V. (2005). Model identification for hydrological forecasting under uncertainty. Stochastic Environmental Research and Risk Assessment. doi:10.1007/s00477-005-0006-5.

Wagener T., Boyle D.P., Lees M.J., Wheater H.S., Gupta H.V., Sorooshian S. (2001), A framework for development and application of hydrological models. Hydrol Earth Syst Sci, 5(1), 13–26.

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.

Wagener, T., M. Sivapalan, P. A. Troch, B. L. McGlynn, C. J. Harman, H. V. Gupta, P. Kumar, P. S. C. Rao, N. B. Basu, and J. S. Wilson, (2010), The future of hydrology: An evolving science for a changing world, Water Resour. Res., 46, W05301.

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.

Wang, L., van Meerveld, H.J., Seibert, J., (2017), When should stream water be sampled to be most informative for event-based, multi-criteria model calibration?, Hydrology Research, doi:10.2166/nh.2017.197

Washington, W.M., L. Buja, and A. Craig (2008). The computational future for climate and Earth system models: on the path to petaflop and beyond. Philosophical Transactions of the Royal Society A, 367(1890), 833-846.

Westerberg, I. and McMillan, H. (2015). Uncertainty in hydrological signatures. Hydrological Earth System Sciences, 12, 4233-4270.

Wilby, R.L., Dessai, S., (2010). Robust adaptation to climate change. Weather 65 (7), 180-185.

Wood, E.F. et al. (2011), Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, Water Resour. Res., 47, W05301

Xiaomeng S., Jianyun Z., Chesheng Z., Yunqing X., Ming Y., Chonggang X. (2015), Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications, Journal of Hydrology, 523, 739-757.

Xie, H., Shen, Z., Chen, L., Qiu, J., Dong, J., (2017), Time-varying sensitivity analysis of hydrologic and sediment parameters at multiple timescales: Implications for conservation practices, Science of The Total Environment, 598, 353-364.

Yapo, P.O., Gupta, H.V. and Sorooshian, S. (1999). Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. *Journal of Hydrology*, 181(1-4), 23-48.

Yatheendradas, S., T. Wagener, H. Gupta, C. Unkrich, D. Goodrich, M. Schaffner, and A. Stewart (2008), Understanding uncertainty in distributed flash flood forecasting for semiarid regions, Water Resour. Res., 44, W05S19

Yoshida, M, Santosh, M (2011). Supercontinents, mantle dynamics and plate tectonics: A perspective based on conceptual vs. numerical models, Earth System Reviews,105(1-2), 1-24.

Zadeh, F.K., Nossent, J., Sarrazin, F., Pianosi, F., van Griensven, A., Wagener, T. and Bauwens, W. (2017). Comparison of variance-based and moment-independent global sensitivity analysis approaches by application to the SWAT model. Environmental Modelling & Software, 91, 210-222.

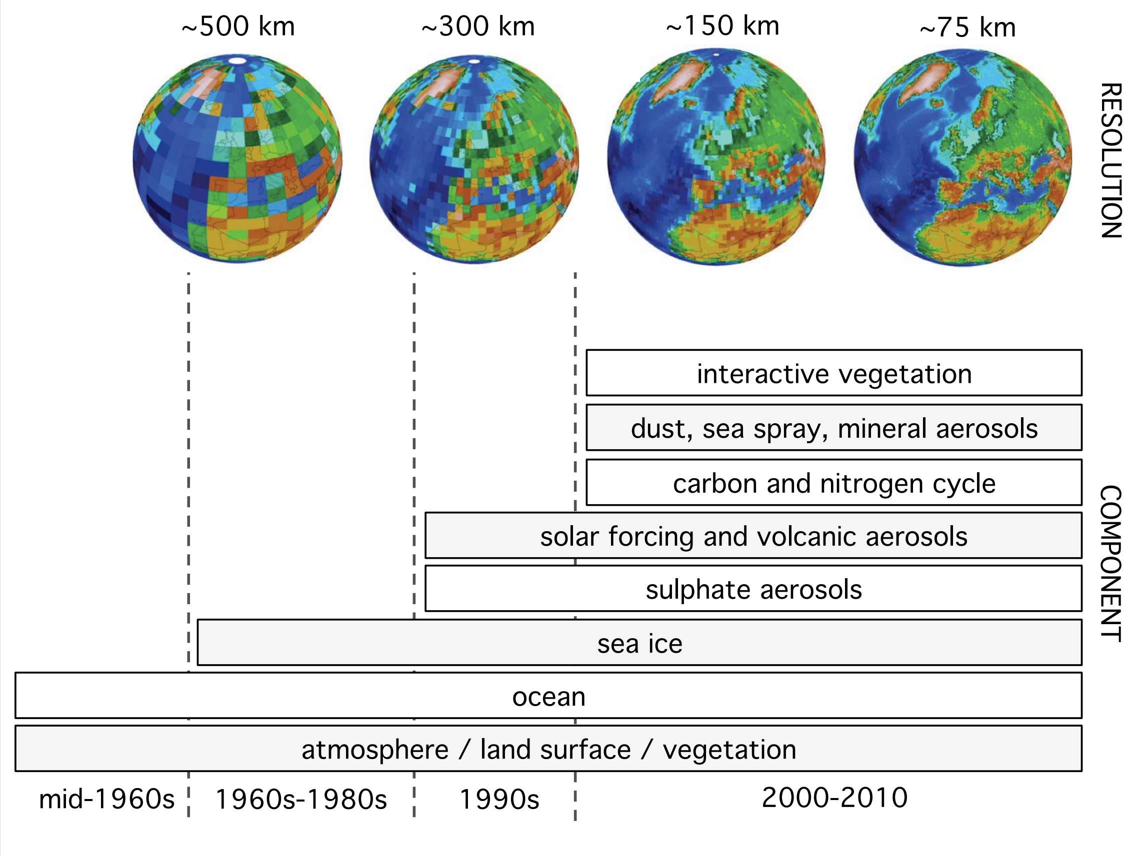

Figure 1. Increase in complexity of earth system models made possible by growing computing power: an example from atmospheric and ocean climate models. Top: growth in spatial resolution, bottom: growth in number of model components. Authors' elaboration based on Washington et al. (2012).

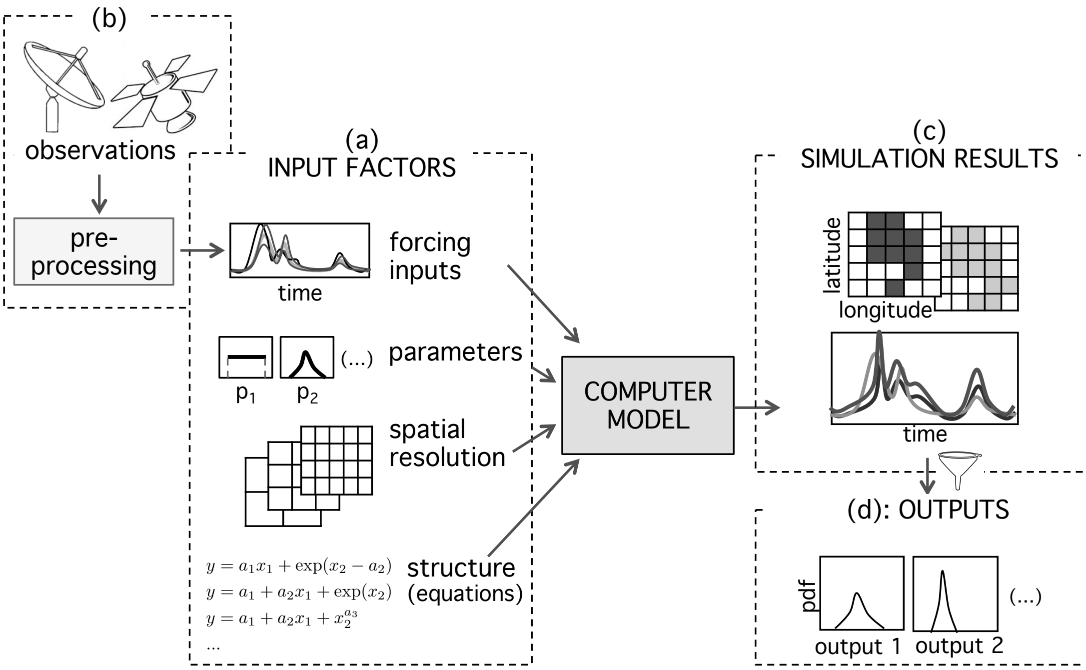

Figure 2. Schematic illustrating the (uncertain) 'input factors' and 'outputs' of a computer model, whose relationships are investigated by GSA.

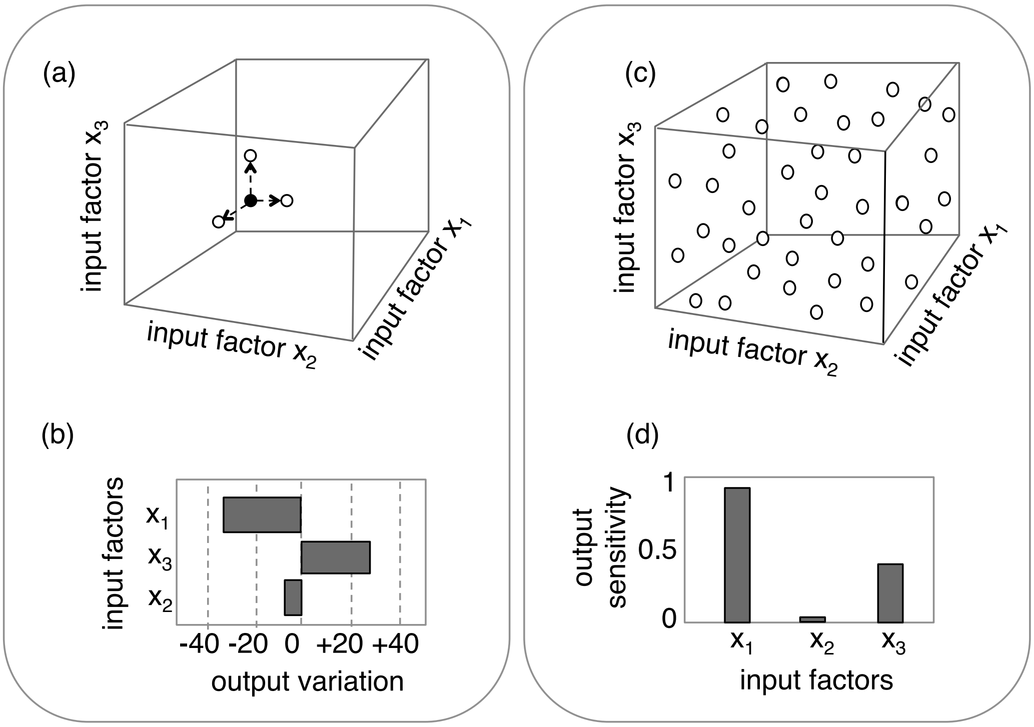

Figure 3. Schematic illustrating the difference between One-At-the-Time (OAT) sampling (a) and associated SA results (b) against All-At-the-Time (simultaneous) sampling (c) and corresponding sensitivity indices (d).

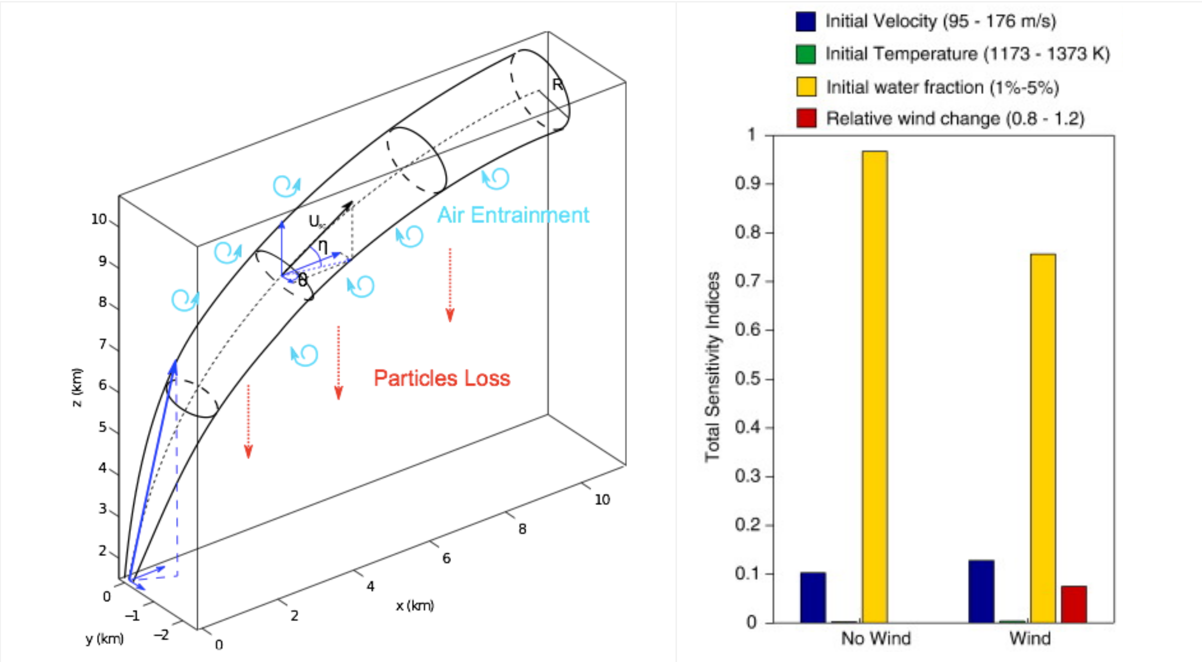

Figure 4. An example of GSA results for investigating the relative influence of four parameters on volcanic plume height predictions. Left: a schematic of the volcanic plume computer model taken from de' Michieli Vitturi et al. (2015). The model output *y* is the plume height attained at the end of the simulation period. Right: sensitivity indices (from de' Michieli Vitturi et al. (2016)) when varying the parameters in the ranges specified in the legend and under two weather scenarios ("wind" or "no wind" conditions). In both scenarios, the initial water fraction is associated with the largest sensitivity index, which means that that varying this parameter has the greatest influence on predicted plume height. Initial velocity is the second most influential input. Relative wind change has an influence only when wind is taken into account (as reasonable), and initial temperature has no influence given that the sensitivity index is close to zero in both scenarios. These results are useful for assessing the consistency of the model's behaviour and to prioritise the variables that would require targeted research in order to have the greatest reduction in output uncertainty.

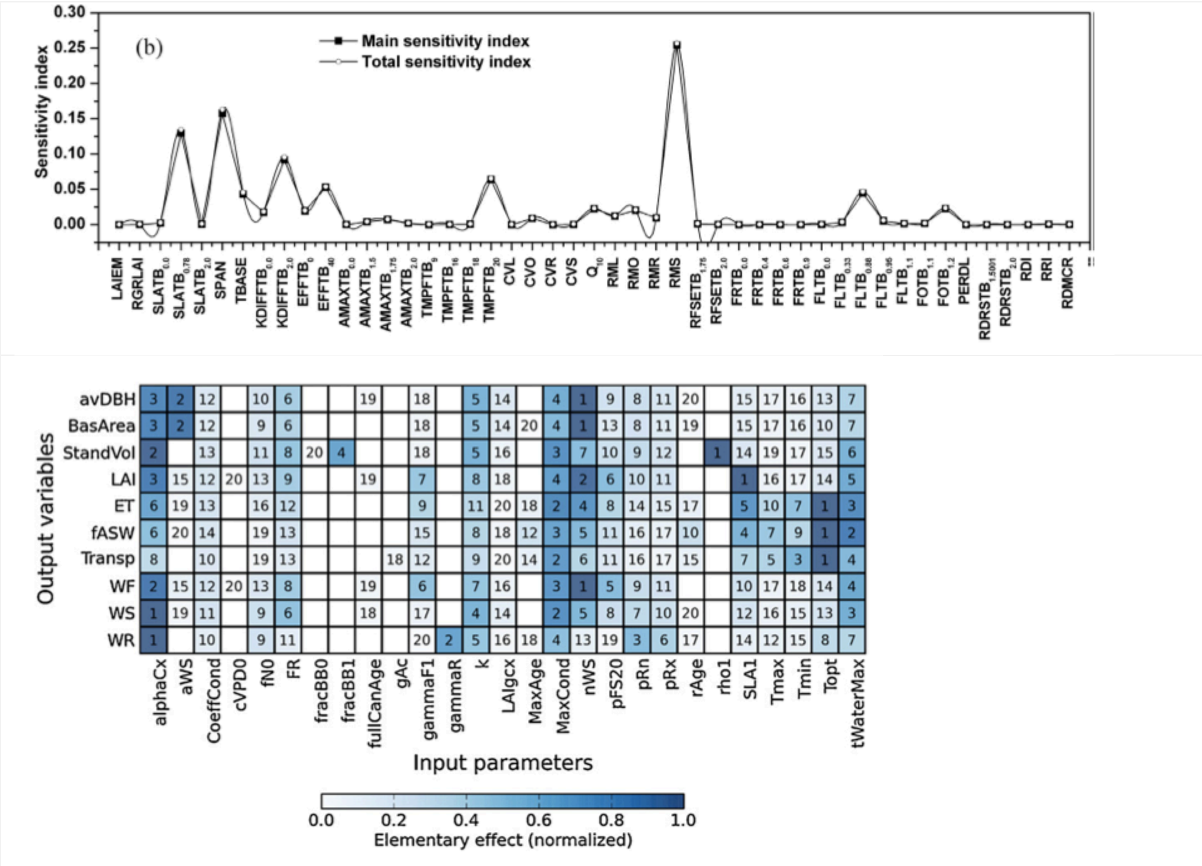

Figure 5. Examples of using GSA to analyse the relative influence of parameters on model predictions. Top: sensitivity indices of the 48 parameters of a crop growth model (taken from Wang et al., 2013). Most of the parameters have a sensitivity index close to zero, meaning that their influence on the selected output metric (the simulated final yield) is negligible. Bottom: sensitivity indices of the 27 parameters of a forest growth model for 10 different output metrics, each representing a different aspect of simulated biomass growth and water exchange between soil, plants and atmosphere (taken from Song et al. 2012). While few parameters have consistently large sensitivity indices for all output metrics, the majority of them have a significant influence only on few output metrics.

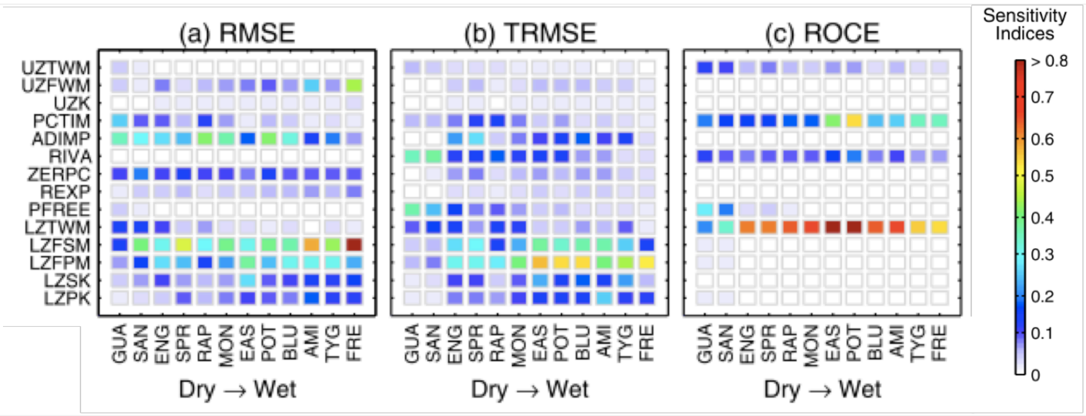

Figure 6. Example of using GSA to analyse the parameter influence of a hydrological model when applied in different sites (taken from van Werkhoven et al., 2008). Sensitivity of three different error metrics (RMSE, TRMSE, ROCE) to the 14 model parameters of a rainfall-runoff model applied to 12 catchments in the US. Catchments (on the horizontal axis) are sorted from drier to wetter climate. The plots show that sensitivity changes with the error metric but also from one catchment to another. Some patterns seem to emerge: for example, when moving from dry to wet catchments, the RMSE sensitivity to parameter UZFWM (upper zone free storage) increases and the sensitivity to PCTIM (percent of impervious area) decreases. The explanation is that in wet catchments flow peaks predictions (which control RMSE) are more often generated by saturation of the upper zone free water storage, while in dry catchments peaks are mainly controlled by direct runoff from impervious areas. Another pattern easily interpretable is that of the parameter RIVA (riparian vegetation area), which has no influence on RMSE but an increasing influence on TRMSE in dry catchments. The explanation is that riparian vegetation mainly control evapotranspiration, which in turn has little impact on high flows (which control RMSE) and a greater impact on low flows (which control TRMSE) especially in dry watersheds. Further discussion and interpretation of other sensitivity indices can be found in van Werkhoven et al. (2008).

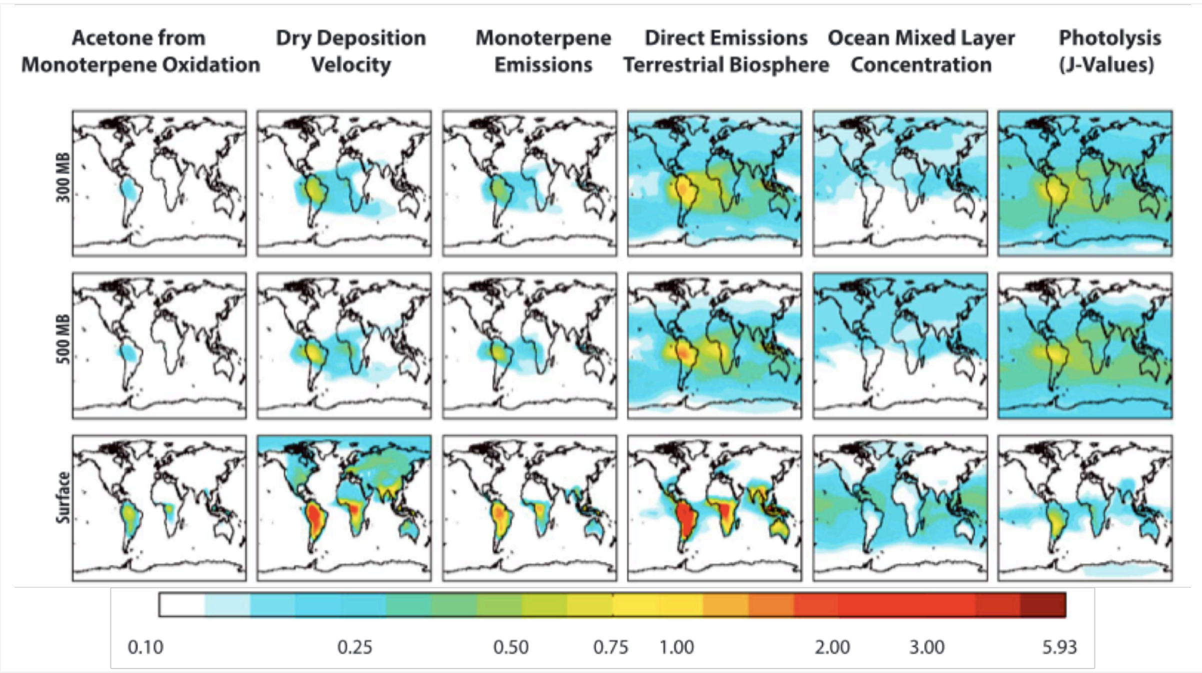

Figure 7. Example of using GSA to analyse the influence of parameters on spatially distributed output (taken from Brewer et al., 2017). Columns correspond to six input parameters of a global 3-D chemical transport model. Rows correspond to different outputs, i.e. acetone mixing ratios in three atmospheric layers. Range of variation of the sensitivity index exceed 1 because of the specific GSA method employed (Morris method, see e.g. Pianosi et al., 2016) however the interpretation is the same as in other Figures, i.e. the higher the index the more influential the input factor. The plots reveal that sensitivity changes massively across the spatial domain.

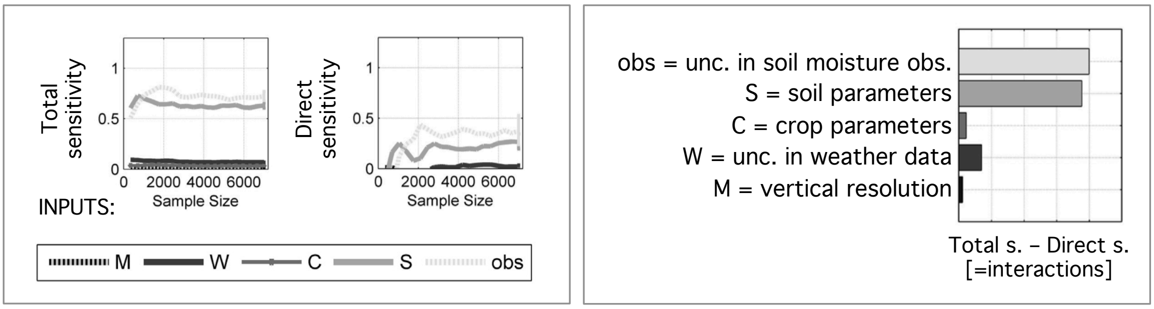

Figure 8. Example of using GSA for investigating the relative influence of uncertainty in parameters and in the observations of simulated variables of a soil-water-plan model (authors' re-elaboration of figures in Baroni and Tarantola (2014)). Left: 'total sensitivity' indices provide a measure of the overall influence of each factor on the error metric (root mean squared error between soil moisture predictions and observations) and 'direct sensitivity' indices measure the direct influence only, i.e. without considering interaction effects. Both 'direct' and 'total' sensitivity indices are evaluated using an increasing number of samples in order to assess their convergence. The plot shows that uncertainty in soil moisture observations (obs) and in soil properties (S) are dominant while other investigated input factors (crop parameters, meteorological forcing inputs, and chosen vertical resolution of the model) have a relatively negligible effect. Right: the difference between total and direct indices (evaluated at largest sample size) provides an indication of the level of interactions of each input factor with the others. Given the high difference values found for soil moisture observations and soil parameters, it can be inferred that the two must have a large amount of interactions with each other.

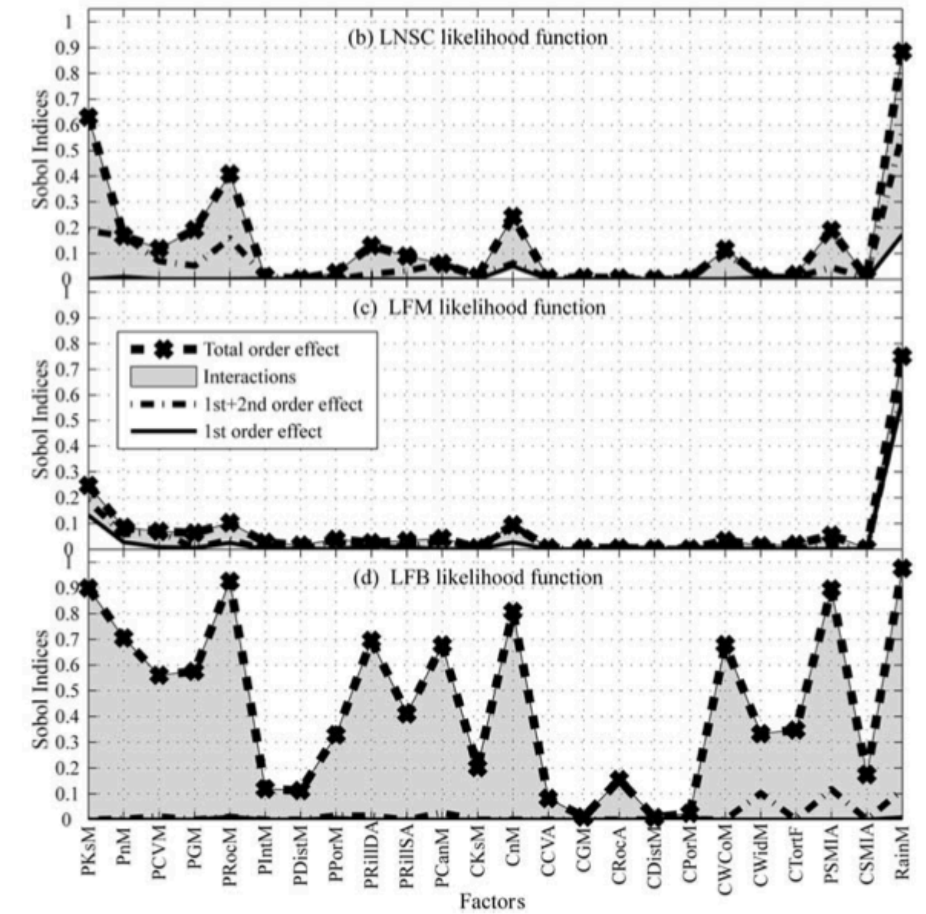

Figure 9. Example of using GSA for investigating the relative influence of uncertainty in parameters, initial conditions and input forcing data of a flow forecasting model (taken from Yatheendradas et al. (2008)). Each panel reports the sensitivity indices for a different error metric (LNSC, LFM, LFB). The input factors shown on the horizontal axis are the model parameters (acronyms starting by P), the model initial conditions (acronyms starting by C) and the rain depth bias factor (RainM) that is used to estimate rainfall rate from radar reflectivity observations. The example shows that the latter parameter has a very large influence on all error metrics and almost completely dominate the second one.

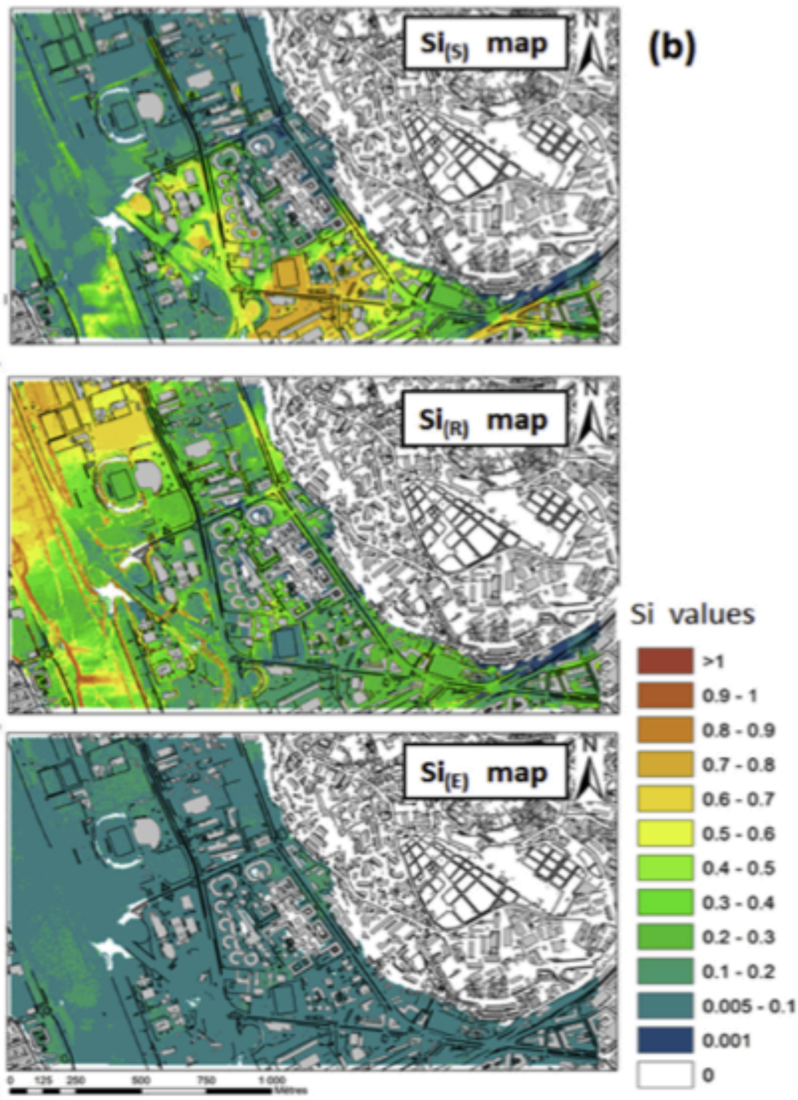

Figure 10. Example of using GSA for investigating the relative influence of measurement errors and discrete modelling choices for a flood inundation model (taken from Abily et al. (2016)). The panels show the spatial distribution of the sensitivity of water depth predictions to three uncertain input factors: chosen level of details in representing above ground features (top), resolution grid (middle), and measurement errors in high resolution topographic data (bottom). The figure highlights that the influence of different factors vary spatially but also that the modeller choices (first two panels) are overall much more important than measurement errors in this particular case.

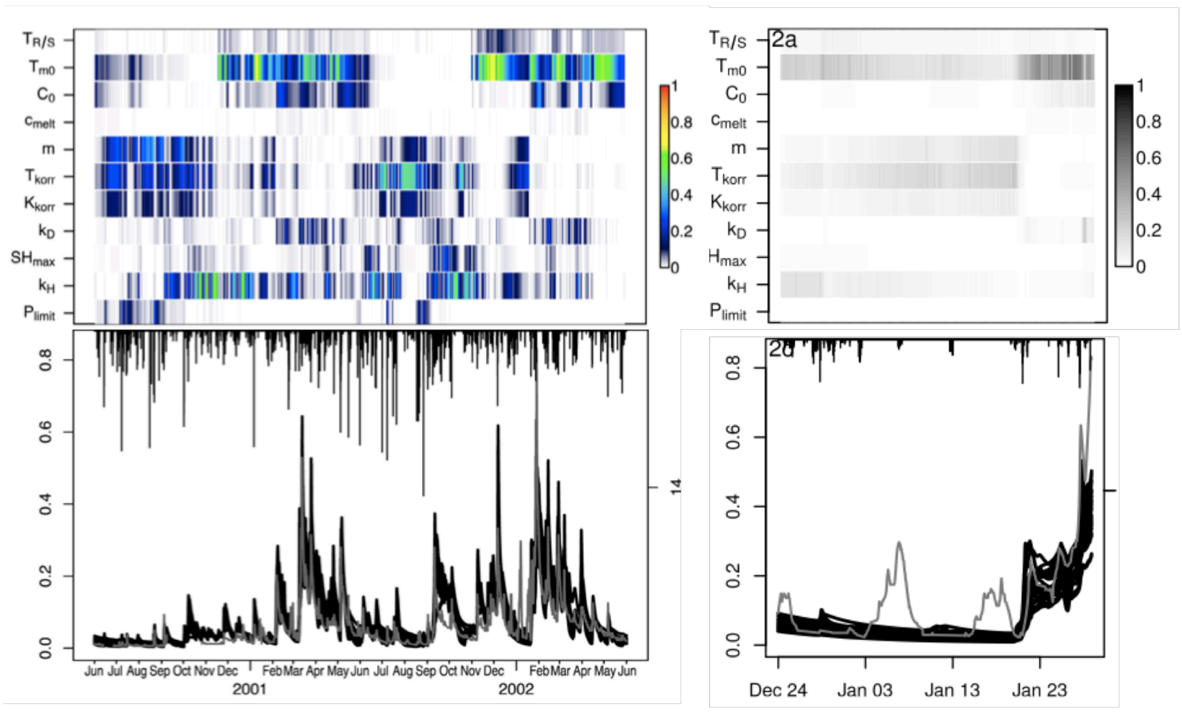

Figure 11. Example of using GSA for model validation (taken from Reusser and Zehe, 2011). The top panels show the temporal evolution of the sensitivity of flow predictions for the 11 parameters of a hydrological model (on the left the entire simulation period, on the right the zoom on selected days). To support interpretation, the bottom panel shows the time series of river flows (grey: observations; black: uncertain model predictions) and of rainfall forcing (from top) over the same periods. The left panels show an overall alignment between dominant parameters revealed by GSA and processes that are expected to dominate flow formation. For example, the top 3 parameters, which control snow accumulation and melt dynamics, are only influential in periods of the year when those processes are expected to occur. Another example is the fourth parameter from the bottom (kd), which is the recession constant for surface runoff and is only influential after large flood events. The right panels focus on a period (between January 3 and January 23) where the model fails to reproduce two observed flow peaks events. The missing sensitivity to the temperature melt index (third parameter from the top, C_0) indicates that no snowmelt can occur in the model during this period, and therefore the mismatch between predictions and observations must be attributed to a model deficiency (for example, the exclusion of radiation-induced melt processes) or a misinterpretation of flow observations (for example, rises in river flow caused by backwater effects due to ice jams).

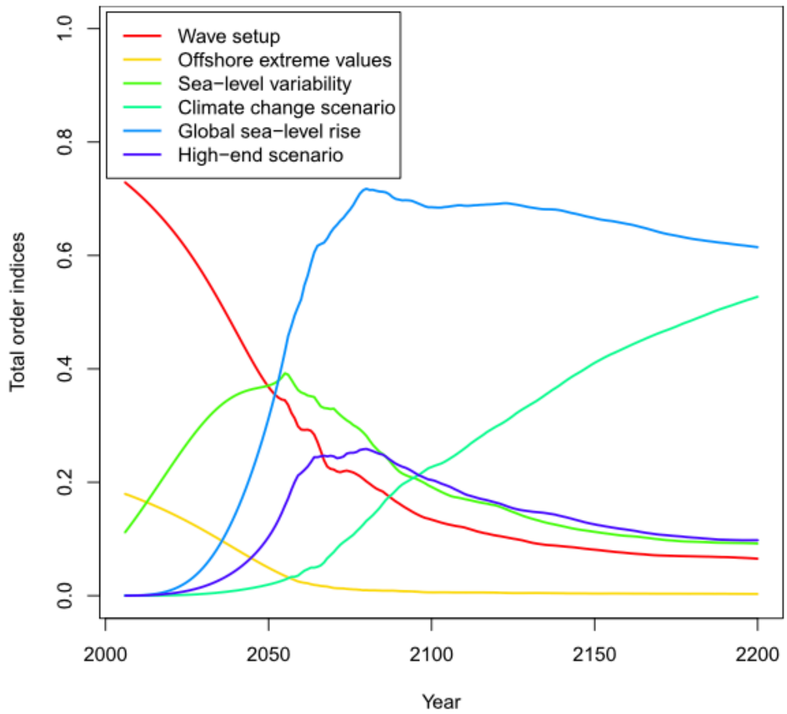

Figure 12. Example of using GSA to support long-term assessment of coastal defences (taken from Le Cozannet et al., 2015). The Figure shows the temporal sensitivity of predicted coastal defence vulnerability (specifically the output metric is the yearly probability of exceeding the threshold height of coastal defences). The figure shows that dominant drivers change significantly over time, for example global climate change scenario only matters beyond 2070 while offshore extreme values have no influence after then. Interestingly, for the time period up to 2050 the dominant factor is the ‘wave set-up’ parameter, which accounts for sea level rise induced by wave breaking. This is a local process determined by the near-shore coastal bathymetry and often neglected in coastal hazard assessments studies. GSA reveals that failing to incorporate the uncertainty in this process may invalidate conclusions and lead to an overestimation of the effects of other drivers at least on short and mid-term planning period.

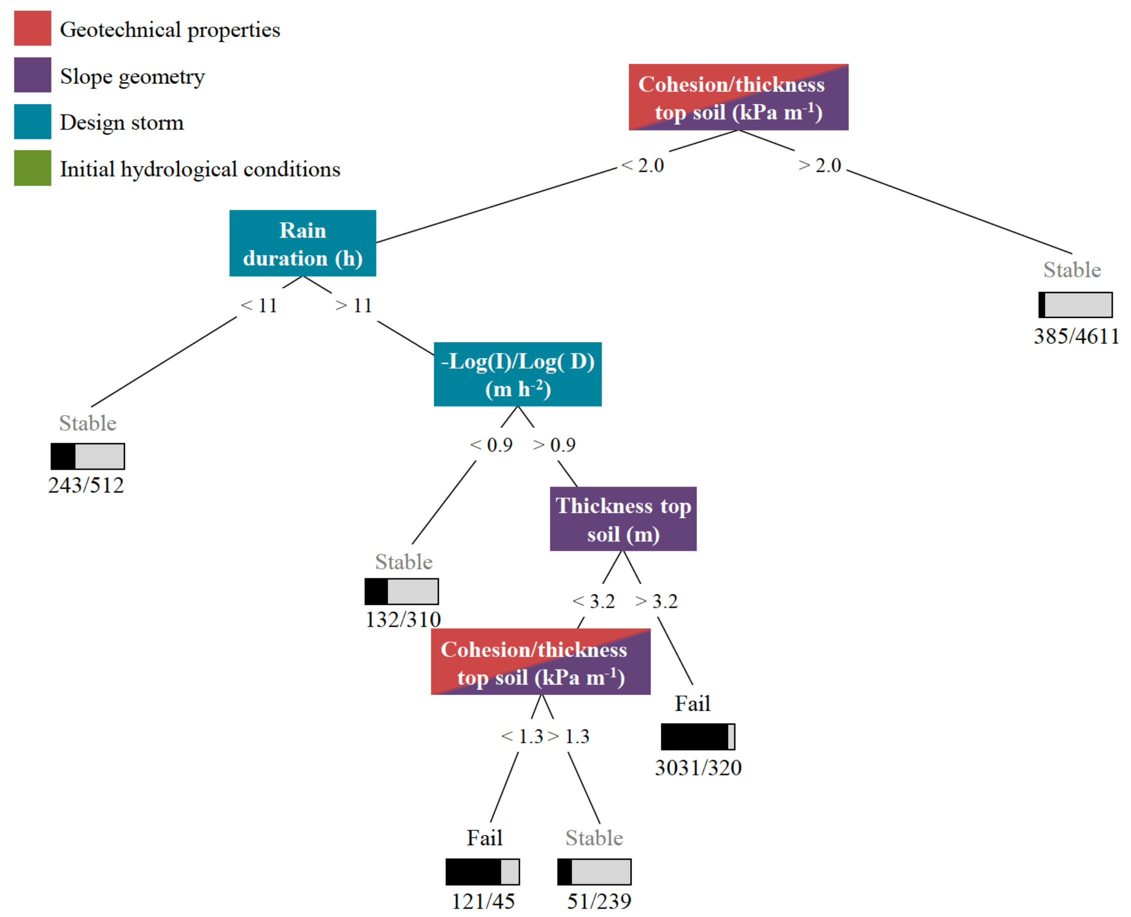

Figure 13. Example of using GSA to implement a 'bottom-up' approach to decision-making in the presence of unbounded uncertainties (taken from Almeida et al. (2017)). A Classification And Regression Tree (CART) is used to map the input factors of a hillslope scale landslide model onto model outcomes that are above (slope fails) or below (slope stable) a critical threshold of the so-called "factor of safety". Each coloured node corresponds to one of the analysed uncertain input factors, which include model parameters (geotechnical and geometrical slope properties), initial conditions and design storm characteristics (rain intensity and duration). The bars at the end of each branch show the proportion of simulations that resulted in slope failure (black) or stability (grey) for that leaf. The CART also displays the critical threshold values that cause a transition from one class to another (< >).