Jun 16th, 2:00 PM - 3:20 PM

A Simple and Effective Approach to Global Sensitivity Analysis Based on Conditional Output Distributions

Francesca Pianosi
University of Bristol, francesca.pianosi@bristol.ac.uk

Thorsten Wagener
University of Bristol

International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, California, USA
Daniel P. Ames, Nigel W. T. Quinn, Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

A Simple and Effective Approach to Global Sensitivity Analysis Based on Conditional Output Distributions

Francesca Pianosi [a] **and Thorsten Wagener**[a]

[a] *Dept. of Civil Engineering, University of Bristol, University Walk, Bristol, BS81TR, UK, francesca.pianosi@bristol.ac.uk*

Abstract: Predictions of environmental models are affected by unavoidable and potentially large uncertainty. When models are applied to understand dominant controls of the system under study, uncertainties will reduce our ability to choose between competing hypotheses. When they are used to support decision-making, uncertainties will reduce our ability to discriminate between different management options and undermine the defensibility of the decision-making process. Global Sensitivity Analysis (GSA) provides quantitative information about the contribution to the uncertainty in the model output arising from different input factors like, for instance, model parameters, boundary conditions or forcing data. GSA thus provides insights into the model behavior and potential for simplification, indicates where further data collection and research is needed or would be beneficial, and enhances the credibility of the modelling results. In this paper, we present a novel method to GSA based on the comparison of the unconditional distribution of the model output, i.e. when all input factors vary, and the conditional distribution when one of the input factors is fixed. The main advantages of our strategy are that, in contrast to other GSA approaches, it works equally well regardless of the output distribution shape, e.g. skewed or not-skewed; it can be focused on specific regions of the output distribution, for instance extreme values; and it provides additional information about the input ranges that map into these output regions (so called Factor Mapping). We test the method using a real-world hydrological modelling application, and compare it to a well-established method of variance-based sensitivity indices (Sobol's method). We finally discuss its advantages and limitations and outline directions for further research.

Keywords: Uncertainty in environmental models; Global Sensitivity Analysis; Variance-based indices; Entropy-based indices.

1 INTRODUCTION

Global Sensitivity Analysis (GSA) is a set of mathematical techniques aimed at investigating the propagation of uncertainty through a numerical model, and specifically assessing the contribution to the model output uncertainty from different uncertain input factors. Throughout this paper, we will use the term *input factor* to denote any input that can be changed (i.e. because it might be uncertain) in the model before its execution, like for instance the input forcing data, boundary conditions, parameters, etc; and *model output* to denotes any (scalar) variable that is obtained after the model execution, for example a measure of performance metric or a summary variable of interest. Quantitative GSA approaches rely on the use of sensitivity indices. The main effect (or first-order sensitivity) measures the direct influence of an input factor on the model output, i.e. the influence it has when varying alone while all other factors are fixed. The total effect instead measures its overall effect, i.e. the main effect plus the direct influence and influence in joint variations with other input factors. Although some methods (e.g. Sobol') allow for computing sensitivity measures also for specific interactions between pairs,

triples, etc. of input factors, often the computation of the main and total effects is considered sufficient to adequately characterise the output sensitivity.

The method of Sobol is one of the most commonly used approaches to global sensitivity analysis [Saltelli et al., 2008]. Here, the main and total effects of each input factor are defined as the contribution to the total output variance from variations of that factor (alone or jointly with the other factors). These variance-based measures are widely used mainly for two reasons. First, they are absolute measures, that is, their value can be used to assess the relative importance of the input factors (factor prioritisation) but it also has a clear interpretation per se: it is the fraction of output variance due to uncertainty in the factor under study. Second, since they can be approximated by an algebraic formula [Saltelli, 2002], their computation is quite straightforward and does not require to specify any tuning parameter besides the number of random samples that should be employed. One major limitation of this approach is that the number of random samples needs to be quite high to obtain a reliable approximation of the variance-based measures in many cases, which can become a problem in case of time-consuming simulation models. Two other limitations are that: (i) the output variance is not a sensible indicator when the output distribution is highly skewed [Liu et al., 2005]; and (ii) that variance provides a summary evaluation of the whole output distribution, while sometimes specific regions, for instance the tails, are more interesting (e.g. for hazard studies or if one wants to focus on high-performing model outputs). To overcome these limitations, some authors have proposed to measure sensitivity by comparing not only variances, but the entire distributions of model outputs. Specifically, the sensitivity to an input factor is measured by the distance between the unconditional distribution of the model output that is obtained when varying all factors, and the conditional distribution that is obtained when varying that factor only. The entropy-based sensitivity measures [Liu et al., 2005], the δ-sensitivity measure [Borgonovo, 2007] and the graphical method proposed in Andres [1997] all follow this line of reasoning. However, the pick-up of these measures by other researchers has been limited, possibly because of the computational complexity in their implementation, which requires the approximation of several conditional and unconditional probability distribution functions (PDF) of the model output. Another disadvantage is that these methods require the user to specify the nominal values at which input factors should be fixed when computing the conditional PDFs, and results then hold for that choice only.

In this paper, we propose a novel method for computing the main and total effect indices that moves along the line of the above methods but introduce two main differences: first, we use cumulative CDFs rather than PDFs because the former are much easier to estimate and compare; and second, because we can efficiently estimate several conditional CDFs by considering different specific values at which to fix the input factors and therefore explore the sensitivity across the entire range of input values.

2 DESCRIPTION OF THE METHOD AND NUMERICAL IMPLEMENTATION

Here we describe the basics of our methodology, starting with the definition of the total effect since we think that it is more intuitive. The (total) sensitivity of the model output (y) to the i-th input factor (x_i) is measured by the discrepancy between the unconditional output distribution $F(y)$ and the distribution conditional to fixing that factor to a specific value, i.e. $F(y|x_i)$. Since the conditional distribution $F(y|x_i)$ accounts for what happens when all the variability due to the i-th factor is removed, both direct and jointly with other factors, the discrepancy provides a measure of the total effect of the i-th factor. As a discrepancy measure, we consider the maximum absolute difference between the two CDFs: $\max_y |F(y) - F(y|x_i)|$. Because this difference may vary depending on the value at which we fix x_i, our definition of the total effect index T_i considers a statistic (the maximum) over all possible values of x_i, i.e.

$$T_i = \max_{x_i} \left(\max_y |F(y) - F(y|x_i)| \right) \quad (1)$$

Notice that since the difference between $F(y)$ and $F(y|x_i)$ cannot be lower than 0 and higher than 1, the index T_i also has a range of variation between 0 and 1 regardless of the range of variation of the model output y. This is a very useful property especially for comparison across different definitions of the output y for the same model, or different models/case study applications with the same output

definition.

As for the main (direct) effect, we compare the unconditional output CDF with the conditional CDF $F(y|x_{\sim i})$ where $x_{\sim i}$ is the vector of all input factors but the i-th, i.e. $x_{\sim i} = |x_1, ...x_{i-1}, x_{i+1}, ..., x_M|$ (and M is the number of factors). In fact, since $F(y|x_{\sim i})$ is obtained by fixing all factors but the i-th, it accounts for the output variability induced by x_i only. In the extreme case when x_i was the only influential factor, this conditional CDF would coincide with the unconditional one and their difference would be 0; and viceversa, if x_i had no direct influence, the difference between conditional and unconditional CDF would be maximum. Therefore, as a sensitivity measure we take the complement of this difference, so that our main (direct) effect index D_i increases with the output sensitivity, specifically

$$D_i = \max_{x_{\sim i}} \left(1 - \max_y |F(y) - F(y|x_{\sim i})| \right) \tag{2}$$

The above indices do not require any assumption about the output distribution, e.g. not-skewed. Furthermore, they allow for focusing on subregions of the output distribution, for instance the tail: in fact, if one wants to focus on a specific subrange of the output values, it will be sufficient to take the maximum with respect to y only in such subrange instead of the entire range, as in Eq. (1) and (2).

Since the analytical computation of the indices T_i and D_i will be impossible in the majority of the cases, an approximate numerical procedure should be used. First of all, maximization with respect to x_i or $x_{\sim i}$ will be taken over a finite set of n evaluations of the inner argument (maximum absolute difference between CDFs), each corresponding to a randomly sampled value of x_i or $x_{\sim i}$. To select the n samples of x_i (or $x_{\sim i}$) we use maximin Latin-Hypercube [Johnson et al., 1990], so to ensure that the selected points cover the feasibility space as uniformly as possible in each direction, and at the same time that they are spread out as much as possible. Second, for each sampled value of x_i ($x_{\sim i}$), the conditional CDF $F(y|x_i)$ ($F(y|x_{\sim i})$) must be approximated. To this end, we compute m samples of the model output by varying the non-fixed input factors $x_{\sim i}$ (or x_i), derive the empirical CDF of the output samples, and finally interpolate it by a piece-wise linear function. Given the regularity properties of CDFs (continuity, monotonicity, relative smoothness) this very simple approximation strategy is quite effective also with limited number of output samples. The unconditional output CDF $F(y)$ is obtained analogously from m output samples obtained by sampling the entire feasibility space of the input factors. This approximation procedure thus has only two tuning parameters, n and m, which regulates the estimation accuracy as well as the computational complexity in terms of model simulation time. In fact, the total number of output evaluations needed for approximating all sensitivity measures T_i and D_i ($i = 1, ..., M$) is $N = m + 2 \cdot M \cdot n \cdot m$.

3 APPLICATION EXAMPLE

In order to test our method we perform a GSA of the parameters of the HyMod applied to the Leaf catchment, a 1950 km^2 catchment located north of Collins, Mississippi, USA. The HyMod is a lumped conceptual hydrological model that can be used to simulate rainfall-runoff processes at the catchment scale. It was first introduced by Boyle [2001] and is discussed extensively in Wagener et al. [2001]; a detailed description of the Leaf catchment can be found in Sorooshian et al. [1983]. HyMod is characterised by five storages, the soil moisture reservoir - represented by a Pareto distribution function to describe the rainfall excess model (see Moore [1985]), three linear reservoirs in series mimicking the fast runoff component, and one slow reservoir. HyMod has five parameters: the maximum soil moisture storage capacity (Sm); a coefficient accounting for the spatial variability of soil moisture in the catchment (β); the ratio of effective rainfall that is sent to the fast reservoirs (α), the discharge coefficient of the fast reservoirs (R_F) and that of the slow reservoir (R_S). Model equations were solved using the forward explicit Euler method with a daily resolution time series of rainfall (mm/day) and evaporation (mm/day) over a simulation horizon of two years starting from 10/10/1948.

Then, we apply GSA to investigate the propagation of parameter uncertainty when the model is used for flood frequency analysis. We thus define as a scalar model output y the frequency of exceeding a

high flow threshold (30 mm/day). Figure 1 shows that this variable varies quite a lot depending on the model parameterisation. The specific objective of GSA is to assess the contribution to such uncertainty from the five model parameters. To this end we use and compare traditional variance-based indices and the indices proposed in the previous section.

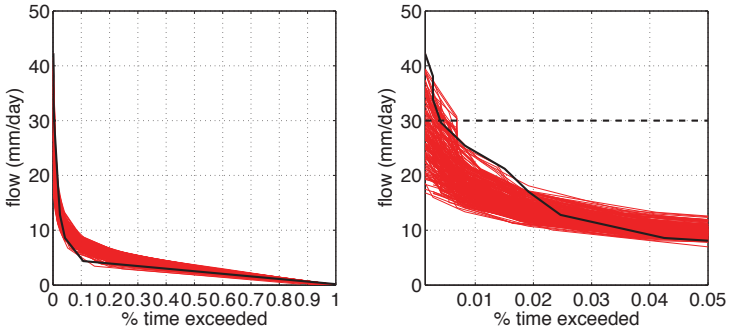

Figure 1. Left: flow duration curves based on flow observations (black) and model simulations (red) under different behavioural parameterisations. A parameterisation is considered behavioural if the corresponding model has a coefficient of determination (defined as $1-VAR(e)/VAR(q)$ where $VAR(e)$ is the variance of simulation errors and $VAR(q)$ is the variance of observed flows) higher than 0.6. Right: zoom of the left panel on low frequency values (corresponding to high flow conditions); the dashed line is the flood threshold considered in this study.

Figure 2 shows the total and main effects measured by variance-based indices (Sobol') and by our method. Variance-based indices were computed using the efficient sampling strategy and algebraic approximation formula proposed by Saltelli [2002], requiring a total number of model evaluations equal to 7000. Our method was applied using n=12 and m=55, for a total of 6655 model evaluations. Both methods sensibly indicate that the most influential parameters are α, which controls the amount of effective rainfall that goes into the fast pathway, and R_F, which controls how quickly water moves through the fast pathway. However, while our method find some direct effect of Sm, β and Rs (see bottom right panel), variance-based indices suggest that they have none (upper right). This difference is not due to numerical approximation, in fact, if we use the dataset generated by our method to approximate the variance-based main index simply applying its definition ($\mathrm{VAR}_{x_i}[\mathrm{E}_{x_{\sim i}}(y|x_i)]$), we obtain the estimates reported in Fig. 2 by grey boxes (again bottom right panel), which are consistent with the estimates by the efficient formula (upper panel). The difference between variance-based indices and ours thus reflect a deeper difference in the meaning of the index to be explored further.

One advantage of our method is that the intermediate results used to compute the sensitivity measures can be effectively visualised to gather more insights into the model behaviour. For instance the top panels in Figure 3 compare the unconditional output CDF $F(y)$ and the conditional CDFs $F(y|x_i)$ for each input factor (model parameter). In each panel, n conditional CDFs are reported corresponding to a different fixed value of the parameter (the darker the line, the higher the fixed value). It can be noticed that the unconditional CDFs for parameters α and Rf are spread out, indicating a high total effect, while those of parameters Sm, β, and especially Rs are almost overlapping with the unconditional CDF, which denotes a low total effect. The bottom panels expand the analysis for parameter Rf, providing a one-to-one comparison of the some of the unconditional (red) and conditional (black) CDFs, showing that when increasing Rf, i.e. the fast pathway becomes faster, the conditional CDFs move down, i.e. flood frequency increases.

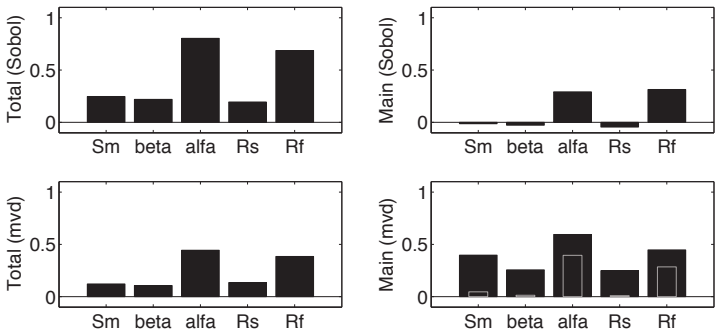

Figure 2. Total effect (left) and main effect (right) measured by variance-based indices (top) and the indices here proposed in Equations (1) and (2) respectively (bottom). Negative values of variance-based indices are due to numerical approximation errors and should be regarded as zero.

4 Discussion and conclusions

In this paper we proposed a new fast and effective method for GSA, specifically to compute sensitivity indices of main and total effects, and demonstrate it in a hydrological application example. The ranking of input factors (model parameters) provided by our method is consistent with the one generated by the widely used Sobol's method based on variance-based indices. In this example the computational complexity of the two methods is approximately the same. Further research will focus on assessing how our method scales to more complex problems, that is, involving a larger number of input factors and/or a more time-consuming simulation model which would impose a lower sample size. However, there is no indication at this stage that our method should require larger sample size than the variance-based approach. Finally, our method has the advantage that it provides many useful insights besides the ranking of the input factors, including information about the input values, ranges and thresholds that produce specific behaviour of the model output (factor mapping) and information about sensitivity in specific regions of the output distribution, for instance its tail. Future research on the proposed method will focus on (i) integrating methods for robustness and convergence analysis of the proposed indices; (ii) clarifying if and how our approach can be used to assess interactions between input factors; (iii) testing other measures of discrepancy between conditional and unconditional CDFs as alternative options to the maximum absolute difference that was proposed here; (iv) testing other statistics to aggregate local results besides the maximum here used; (v) designing a strategy to expand the applicability of the proposed method to input/output datasets not specifically generated for application of this method.

5 Acknowledgments

This work was supported by the Natural Environment Research Council [Consortium on Risk in the Environment: Diagnostics, Integration, Benchmarking, Learning and Elicitation (CREDIBLE); grant number NE/J017450/1]

References

Andres, T. (1997). Sampling methods and sensitivity analysis for large parameter sets. *Journal of Statistical Computation and Simulation*, 57(1–4):77–110.

Borgonovo, E. (2007). A new uncertainty importance measure. *Reliability Engineering and System Safety*, 92:771–784.

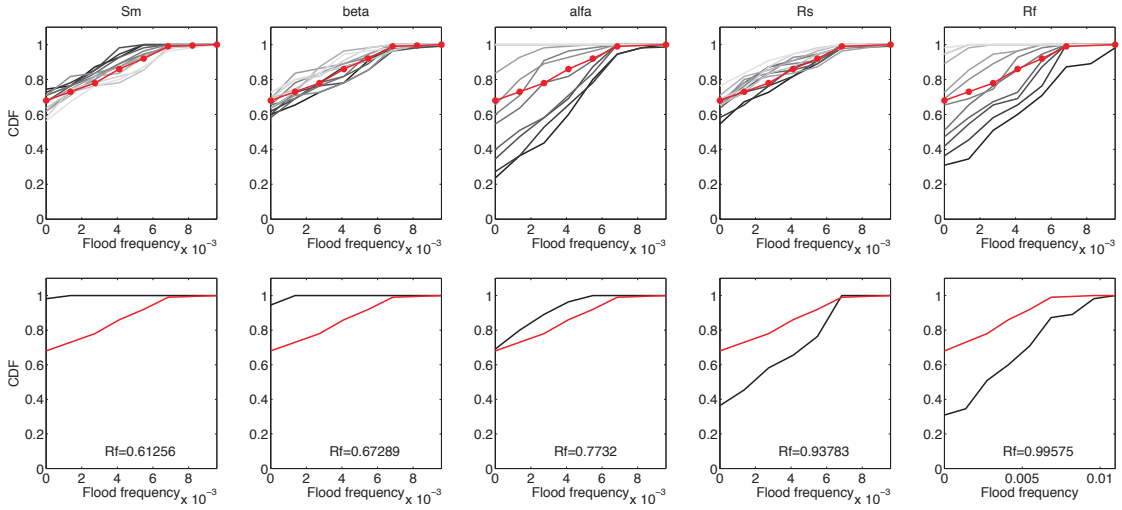

Figure 3. Top: unconditional output distribution (red circles) and conditional distribution (grey to black lines) when each input factor is fixed to a specific value (the darker the line the higher the fixed value). Bottom: one to one comparison between unconditional (red) and conditional (black) output distribution when parameter Rf is set to different fixed values.

Boyle, D. (2001). *Multicriteria calibration of hydrological models*. PhD thesis, Dep. of Hydrol. and Water Resour., Univ. of Ariz., Tucson.

Johnson, M., Moore, L., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–14.

Liu, H., Sudjianto, A., and Chen, W. (2005). Relative entropy based method for probabilistic sensitivity analysis in engineering design. *J. Mech. Des.*, 128:326–336.

Moore, R. (1985). The probability-distributed principle and runoff production at point and basin scales. *Hydrol. Sci. J.*, 30(2):273–297.

Saltelli, A. (2002). Making best use of model valuations to compute sensitivity indices. *Computer Physics Communications*, 145:280–297.

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

Sorooshian, S., Gupta, V., and Fulton, J. (1983). Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility. *Water Resour. Res.*, 19:251–259.

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