Nat. Hazards Earth Syst. Sci. Discuss.,
doi:10.5194/nhess-2016-288-RC2, 2016

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Multi-level emulation of a volcanic ash transport and dispersion model to quantify sensitivity to uncertain parameters" *by* Natalie J. Harvey et al.

F. Pianosi (Referee)

francesca.pianosi@bristol.ac.uk

Received and published: 18 October 2016

The manuscript presents a complex and potentially very interesting application of emulation modelling to quantify the relative impact of several uncertain parameters on the predictions of a volcanic ash transport and dispersion model (the NAME model). The main novelty of this study is that it presents one of the first applications of a formal sensitivity analysis (i.e. beyond one-at-the-time approaches) to a volcanic ash transport and dispersion model. Moreover, some of the techniques and lessons learnt in this study (for example on how to link fast and slow simulators) might be of interest to a broader community who deals with uncertainty in computationally expensive transport and dispersion models, not only volcanic ashes.

However, the manuscript in its present form is quite unclear and the structure unbalanced, which makes it difficult to fully evaluate and appreciate the findings of the study. Below my main issues and some suggestions for revision.

[1] The choice of contents in the methodology sections is confusing. Too much space is given to topics that are generic and covered in many other papers or even textbooks, which divert attention from and leave too little space to specific information about the NAME application. For example, on P. 14 L. 16-29 almost the same space is devoted to describing the difference between fast and slow simulators, which is a very important setting of the model under study (see also comment [3] below), as to describing the well-established and totally generic Latin Hypercube sampling technique. The description of emulators and Bayes linear methods covers 4 pages (P. 16-19) although much of the content is very generic, easily accessible elsewhere, and - most importantly - not strictly related to the application here. In fact, from P. 20 L. 10-12 ("a successful method has been to use a standard (non-Bayesian) least-squares regression") I understand that the non-linear/linear Bayesian approaches discussed on previous pages are not actually used here because linear least-squares are sufficient. As a reader I was confused by these long descriptions and diverted from focusing on the specific features of this application, for example the link between emulators of the fast and slow simulators. Another example is Section 5.4: this is all generic and standard methods that do not need to be discussed in an application-oriented paper, especially if not used then in the Results section (see for example the comment on "examining the residuals"...). In summary, I would recommend to deeply revise these sections, to make them shorter and include only the information relevant to the specific application and thus needed to understand the results.

[2] The results section seems to focus a lot on the validation of the emulators - i.e. how good they are in representing the fast and slow simulators - and their similarity, rather than the ultimate goal of the analysis, which is to use the emulators "as a research tool to better understand the simulator, the role of the parameters, the interactions between

them..." (P. 4 L. 17). For example, while Fig. 5 and 6 report validation results for the emulators, there are no Figures reporting sensitivity results, parameter mapping or interactions. The only results related to sensitivity analysis are in Table 4, which however does not even use the word "sensitivity"! This is odd especially considering the title of the manuscript.

[3] The difference between fast and slow simulators should be better clarified. In particular, on P. 14, L. 16-24: What does the "increase in particle-sampling noise" mean in simpler terms? How does it impact the simulation results? Not so much - at least for predicting average column loadings - according to what the authors say later on P. 15 (L. 7-12). On the other hand, how fast is the fast simulator? These aspects need to be clarified because they are at the basis of the motivation of the study: if the fast simulator is fast enough, and its predictions of the output of interest (average column loading) are reasonably close to those of the original (slow) simulator, then why building an emulator? One could directly apply a Global Sensitivity Analysis technique (for instance, the Morris method, or Regional Sensitivity Analysis, which are both reasonably "low-cost") to the fast simulator. I am not saying that this is necessarily the case (maybe the fast simulator is not so fast, or there is some other reason I am missing to avoid using it) but more details should be provided to clarify this crucial point.

[4] Choices underpinning the analysis and their impact on sensitivity results needs further discussion. P. 6, L. 20 onwards: there are three set of choices that are made here: - the choice of the uncertain parameters to be included in the sensitivity analysis (out of a larger set of parameters appearing in NAME, which are set to default values - two of them are further described in Sec. 3.2.7 and 3.2.8, together with the reasons for not varying them - what about the others?) - the choice of plausible ranges for the uncertain parameters - the choice of a particular event, and hence a particular set of forcing data, for the model simulation (and the choice of ignoring the uncertainty in those data) I presume that these choices may have a very strong impact on the results and thus on the generality/transferability of the findings. Assessing such impact might

be beyond the scope of this manuscript, but the point should be at least mentioned and discussed.

OTHER SPECIFIC POINTS:

P. 3 L. 28: "Finally, the analysis cannot provide ...". A bit vague, please clarify what is an "overall assessment of uncertainty"?

P. 8 L. 3-4: Difference between "full depth" and "thin layer" is not clear to me. Also, very unclear how the 1700 + 1700 runs here mentioned are connected with the 1500 + 200 runs mentioned on P. 14 (the same experiment of P. 14 is repeated twice, once per each source type?). Maybe the confusion could be avoided by simply not giving all these details on the thin layer source case, since its results are not shown (as commented on P. 23 L. 15-16)?

P. 10 L. 18: " are varied by the same proportion". Unclear.

P. 12 L. 23: "between 0 and 2": or between 0.5 and 2 (assuming this section refers to x_{17} in Table 1).

P. 14, L. 10-11: "the average ash column loading predicted ..." not completely clear. Are column loadings per each region averaged over the simulation period, thus defining 75 "outputs" (and 75 emulators), or are predictions for each hour analysed separately (thus defining 75xT "outputs", where T is the number of hours in the simulation)? Please clarify, maybe also inserting an equation here.

P. 14, L. 14: Again unclear: "regions used for the first hour are marked..." So, the definition of regions changes from one hour to another? And also their number then? And so how does it connect to my previous question?

P. 15, L. 7-12 and Figure 3. Again on the difference between fast and slow simulators. I understand that the main conclusion here is that simulations from fast and slow simulators are similar, however this paragraph and the Figure are rather unclear. "agreement between the two simulators" means that they provide similar predictions of average

column loading? What does "to be related but not in agreement" mean? The "correlations" of 0.99 and 0.7 are the correlation between simulated column loadings (in different regions? at different time in the simulation period?). Please be more specific. Figure 3: units of measurements on the axes are missing!!! (I guess they are "Log ash" as in Fig. 5?).

P. 15, L. 7: "the goal" ... This is a bit misleading: the ultimate reason for building the emulator is not making inference, but rather understanding the role of parameters (see discussion on P. 4 L. 14-17). Maybe good to remind it here.

P. 28, L. 9-10: "for some parameters... it is not plausible": so you included in the analysis some parameter combinations that are implausible? This sounds contradictory (who would be interested in sensitivity to parameter variations that are not plausible?). Please clarify

P. 29, L. 18-23: This is a generic comment that was already made in the Introduction (P. 3) and is not part of the results. I would remove it.

Table 1: - term "default" on first row is a bit misleading - does it refer to the "data from Keflavik radar" (as explained on P. 7, L. 19)? If so, clarify in the Table - term "default" on third row is a bit misleading - does it refer to the MER value from Eq. (1) (as explained on P. 8, L. 20)? If so, clarify in the Table - connect better to the text, for example the mathematical symbols used in the text could be included in the description of "Parameter name" (for example R_a on row 17...) - rows 7-10: "varied in proportion..." is unclear (here and in the text) - row 18: so the default value coincides with the maximum value? This looks strange.

Figure 3: Caption mention "(a)" and "(b)" but letters are not reported in the panels.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., doi:10.5194/nhess-2016-288, 2016.