Geosci. Model Dev. Discuss., 6, C712–C716, 2013
www.geosci-model-dev-discuss.net/6/C712/2013/

Geoscientific Model Development
Discussions
Open Access

Interactive comment on "The potential of an observational data set for calibration of a computationally expensive computer model" *by* D. J. McNeall et al.

M. Crucifix (Referee)

michel.crucifix@uclouvain.be

Received and published: 3 June 2013

This important manuscript by McNeil et al. is an application of global sensitivity analysis theory. More specifically, the authors examine systematically the relationships between a set of inputs and a set of output summaries of a climate-ice-sheet simulator (here: GLIMMER), with the objective of delivering a simulator-based estimate of which observations may best constrain model inputs. The problem would be trivial if the model was linear (the result would be given by the inverse of the Jacobian operator), but it is not in a complex simulator. Hence, the authors had to overcome a number of challenges :

- establish a cheap statistical surrogate for the simulator, i.e., an emulator. The

  authors have chosen the now-standard Gaussian process emulator proposed by Oakley and O'Hagan;
- a decision framework. The choise here relies on the history matching of Craig et al., inherited from the Bayesian Linear paradigm which asserts that decision may be taken based on variance estimators without the need of considering the higher moments of the posterior distributions
- define a set of quantitative metrics: the authors propose here two metrics: the marginal ranvge of not implausible space, and the volume of not-implausible space
- and finally a set of visual diagnostics, particularly critical for the exploration and analysis of complex systems

In the tradition of *Geoscience Model Development* the main output of the paper is methodological. In this respect the authors have definitely achieved. They propose a methodology that follows good practice rules (design emulator, assess it following leave-one-out approach, explore the input-output relationship and discuss the value of quantitative metrics). The visual diagnostics are neat and carefully designed, and they enabled the authors to reach convincing conclusions about the simulator itself. This said, fully appreciating their usefulness will require user training and more experience with other models, hence the importance of making the relevant R code publicly available. In addition, the authors deliver a more general and far-reaching message about the simulator itself. The potential of observations for constraining the model input is disappointingly low as soon as reasonable amounts of uncertainty on observations or model discrepancy are accounted for. The implication is that most of the information about the model is effictively encoded through the paramater prior distributions, obtained by expert elicitation. Whether this result should be seen as positive or negative is a matter of judgement, and I would consider it as potentially very significant.

As a conclusion, I fully support the publication of this manuscript in Geosciences Model Development, pending a number of minor comments given below.

General comment: First a remark on the form: the authors have chosen to make a use of the active form (e.g.: "we use this", "our metric") that is slightly more assertive than standard in the scientific literature. Whether this should be corrected is left as a decision of the editor.

p. 2371: There might me some semantic argument to be had about the definition of the words *calibration* and *tuning*. Here the authors define *tuning* as a form of point-estimate (best-matching) calibration. In other contexts, *calibration* and *tuning* refer to different processes, *calibration* implying the existence of a formal quantitative statistical framework while *tuning* being informal or qualitative. A reference to the definitions given by the authors might therefore be welcomed, especially if they are standard in the statistical literature.

p. 2377: Section 2.2.3 turns to be distracting and in fact not really helpful.

p. 2379, l. 21: It is read : “If the entire a priori input space is truly plausible (. . .) *given an observation of the true system*" (emphasis is mine). What is meant by this latter expression : “given an observation of the true system"?

The first part of section 2.4 seems to come at the wrong place. Considerations about the cost of observations or the interest and limit of the simulator in guiding observations are best left for the introductory material and the discussion, where they are already present. At this stage the text should be exempt of general considerations to help the reader to focus on the methodological and mathematical details. Furthermore, given that we are already p.11 of the manuscript (in web form), use of the future tense (“our metric *will* take") or the conditional (“we *might* use is making the reading impatient. Some editorial work is probably needed to present

the results a bit earlier, and possibly use them to support some methodological choices.

p. 2380, l. 4 : extra “with".

p. 2380, l. 19 : “Our metric will take into account not only the uncertainty in the emulator, but also the inherent problem of inverting the mapping for $\mathcal{Y}$ to $\mathcal{X}$" : this is a point that might require clarification. It would be useful to be more explicit in stating which choices are critical in the quality of this mapping.

p. 2382 Regarding the latter comment: little is said about the choice of the roughness lengths and emulator nuggets, known to be important. It is said that the BACCO package is used, which is fine, but the generic "These parameters are estimated empirically from the ensemble data, via an optimisation routine" is unsatisfactory: optimised on what? (probably leave-one-out criteria)? How ? Are their any visual diagnostics being involved? The leave-one-out approach implies the calibration of $N - 1$ emulators: do they all have the same roughness lengths and nuggets?

p. 2383, ll. 17 : What is meant here as the accuracy seems in fact to be the well-calibrated character.

p. 2388, ll. 3-6 : Visualisations techniques are definitely important and the authors have delivered on this in the present article. This said I failed to make sense of this paragraph. Why speak of “projecting a set of lower dimensions in high dimensional space"? What is the point?

p. 2388, l. 14 and 18 : Shouldn’t “at worst” read “at best" ?

Finally, the Editorial board of Geophysical Model Development drawing attention on the importance of *scientific reproducibility*, the authors can only but be encouraged to provide code for their nice visual diagnostics.

Interactive comment on Geosci. Model Dev. Discuss., 6, 2369, 2013.