Geosci. Model Dev. Discuss., 6, C926–C933, 2013
www.geosci-model-dev-discuss.net/6/C926/2013/

Geoscientific Model Development Discussions

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

I. Vernon (Referee)

i.r.vernon@durham.ac.uk

Received and published: 4 July 2013

Referee Review for Geoscientific Model Development Discussion

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

General Comments

This paper by McNeil et al. is an interesting and important work that makes a significant contribution to the field. The authors address the problem of designing a possible future data collection campaign, where one has access to a complex and computationally

expensive computer model for the physical system of interest. The main challenge is to identify which real world quantities (corresponding to outputs of the computer model) would, if measured, lead to the most strict constraints on the inputs of the computer model (in this case, the Glimmer ice sheet model).

To address this, the authors use several powerful pieces of statistical methodology.

1. The computer simulator is slow, and is of high enough input dimension (5D) to make direct exploration of the input space using model runs alone infeasible (a critical problem for a broad class of computer models). The authors use Bayesian emulators, fast surrogates of the model, specifically those of Oakley and O'Hagan, which allow for substantially faster exploration.

2. To determine the extent to which input space is constrained, the authors sensibly use the history matching approach of Craig et al. (see below for more references), which provides a fast and simple method of ruling out bad parts of the input space using implausibility measures. This is a useful alternative to the full Bayesian calibration calculation, which for a decision problem such as this could involve running MCMC a vast number of times, and hence be prohibitively expensive in terms of computing time (an unfortunate problem for many Bayesian design calculations).

3. The authors then suggest and use two metrics that summarise the input space removed, for use in analysing the effect of measuring each of the three outputs, and for deciding on a possible future data collection campaign. The metrics and suitable summaries of the methods as applied to the Glimmer ice sheet model, are given in a series of well presented plots.

The above statistical techniques are appropriate and are in most cases carefully applied (see below for some slight concerns though). In particular, the emulator diagnostics were impressive and give confidence in the subsequent implausibility measure used to cut out input space. Combining these techniques together allows the authors to show real insight into the structure of the model in terms of relationships between in-

puts and outputs, and which outputs are most important to measure. The speed of the calculations also allows investigation into the effect of the size of observational errors and model discrepancy. This work clearly represents a substantial advance in this area and I would therefore fully support the publication of this manuscript in Geosciences Model Development, provided the authors can address the points below.

Specific Comments

1. Page 2379, lines 27 - 28 "we let the model output y take the place of a theoretical observational data-set z in our analysis". Using y in place of z is correct for the case of zero model discrepancy delta and zero observation errors e, but when delta and e are not assumed zero then this represents an approximation to the full calculation (a good approximation though!). The full calculation would involve simulating from delta and e to get z, and doing this many times for each ensemble value of y considered, applying the implausibility constraints for each simulated value of z, to get a distribution of space cut out for each y value. (Incidentally this process is described in the presentation: Vernon, I., Goldstein, M., Liu, J., Lindsey, K. "Emulation and Efficient History Matching of Stochastic Systems Biology Models" Presentation at UCM 2012. http://www.mucm.ac.uk/UCM2012/Talks&Posters.html)

Now the authors mention several times that their analysis will provide "a maximum bound for our ability to constrain the model inputs" e.g. in the abstract, p2370 lines 7-8. Due to the above, this claim should perhaps be tempered as it relies upon the above approximation to the full calculation. What the authors calculate is perhaps a maximum bound on the expected space cutout in the full calculation, but this depends on the distributional assumptions. For example, the calculation may say 50% of input space is the maximum that can be constrained, however measurements of the real world could lie outside the range of model outputs and hence rule out 100% of the input space.

2. Page 2372: in the description of the two approaches of Bayesian calibration and his-

tory matching, it is useful to mention the main differences between these approaches. Bayesian calibration assumes a "best input xˆ*", uses a prior for xˆ* and updates this to a posterior. History matching does not assume the existence of a best input xˆ* and instead just tests points in input space to determine if they are consistent with the specified model and associated uncertainties (observation error, model discrepancy). This is an important difference because in the Bayesian calibration case, if you have explored some parts of the input space (by doing runs of the model there) and have found those parts to be bad in terms of giving outputs far from observed data, you automatically think other as yet unexplored parts of the input space are good. This is because you have assumed one single xˆ*, and it must lie somewhere! This is avoid in history matching, where your assessment of one part of the space does not affect other parts of the space.

A full description of the benefits of history matching and its application to a large model of galaxy formation, along with a discussion comparing it to Bayesian calibration, can be found in:

Vernon, I, Goldstein, M. & Bower, R. G. 2010. "Galaxy Formation: a Bayesian Uncertainty Analysis". Bayesian Analysis 05(04): 619 - 670 (with discussion)

which it may be suitable for the authors to cite.

3. Page 2374, line 19: "where e represents systematic errors or biases in the observations". Do these errors really have to be "systematic" or be "biases". Surely e is a random variable just representing measurement error, which can be systematic or otherwise! Technically, if there were biased, this should change the implausibility measures too. Perhaps just saying "where e represents measurement errors in the observations" would be clearer?

4. Page 2376, lines 2-3 The authors use Pukelsheim's 3 sigma rule, a very powerful and general result, but it only holds true for unimodal distributions. Unimodality of the distribution underlying the implausibility measure is a reasonable assumption here, but

should be stated.

5. Page 2376, lines 4-7 the authors combine implausibilities from different outputs by maximising over the implausibilities. This is a sensible choice, however, it is worth noting that there are other more complex implausibility measures available such as the multivariate measure described in (Vernon et. al. (2010) (ref given above)).

6. Page 2376. Defining a suitable metric is no easy task. The authors introduce the marginal range of NROY space. This is a fairly sensible metric, however, the volume of NROY space is, in my opinion, a far superior metric as it is a clearly defined object and represents how much we have learned about the input space, where as the marginal range can be misleading in several situations. In this example, where 3 outputs are used to constrain a 5 dimensional input space, we expect the NROY space to be in the form of hyper-surfaces (specifically 2 dimensional) due to the remaining degrees of freedom (as the authors have noticed at a later point). These hyper-surfaces may stretch across the input space, ensuring that the marginal range is quite misleading. Could the authors, if they agree, perhaps mention that the volume metric is superior or safer in many cases?

7. Page 2376, lines 24 - 25 "We can define a volume V of “not implausible” input parameter space, or alternatively that input space “Not Ruled Out Yet” – as the region bounded by the convex hull where I < 3". No, the volume V is not the region bounded by the convex hull, it is simply the region defined by I < 3, whatever its shape or geometry. The Monte Carlo estimate for this volume V does not need any complex hull results: it is a direct estimate of V.

8. Page 2377, line 22, " or the least implausible point". When an implausibility measures gives a high value, it means we can rule out that input, but when it give a low value it simply means we are still not sure about the input at this stage (hence Not Ruled Out Yet). Further runs of the simulator (waves), or more detailed statistical modelling (using say more advanced implausibility measures) may subsequently rule out this in-

put. Hence the least implausible point has no real importance and should be treated with caution (indeed, this is why implausibility measures are much easier to use that full bayesian posteriors as the former just models where the bad inputs are, where the later tries to model the much more complex question of where all the good inputs are!). Perhaps the authors could just remove the phrase " or the least implausible point".

9. Page 2378, line 12-13 "This is unlikely to be a practical solution, given the possibly complex nature of z, and conflicting demands on expensive simulator output". I would add the difficulty of searching high dimensional spaces which in general can have large numbers of local minima.

10. Page 2384, "We fix the standard deviation of the representative observational error as 10% of the maximum simulated value for each of the outputs in the ensemble". It is good to see that the authors do explore the effects of non-zero observational errors. This value does seem quite large though as this means 3 sigma in the implausibility will correspond to greater than 30% of this value. The authors should then not be too discouraged to see that not much space is subsequently ruled out.

11. Page 2388, "case that there exists a poorly modeled discrepancy (an “unknown unknown”), the ability of data to constrain the simulator will be overconfident." This may need clarification, because as was discussed previously, the unknown unknown could lead to far more space being ruled out than the maximum bound from this calculation, or far less than the worse case scenario given. As can be seen, reasonable modelling of model discrepancy is usually unavoidable when considering any observed results from the real world.

12. The authors cite Craig et al. (2001) when discussing history matching and implausibility. It might be reasonable to cite the first two papers using this method, which also feature emulation and model discrepancy:

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1996). “Bayes linear strategies for history matching of hydrocarbon reservoirs.” In Bernardo, J. M., Berger,

J. O., Dawid, A. P., and Smith, A. F. M. (eds.), Bayesian Statistics 5, 69–95. Oxford, UK: Clarendon Press.

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1997). “Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments.” In Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R., Rossi, P., and Singpurwalla, N. D. (eds.), Case Studies in Bayesian Statistics, volume 3, 36–93. New York: Springer-Verlag.

Technical Corrections

1. Page 2374, lines 12 - 16 the authors write:

"We can represent output y as an uncertain function of input x thus:

y = g(x). (1)

The simulator is complex enough that we cannot trivially predict the output y at a given x before the simulator is run."

This section seemed a little unclear, surely the sentence after (1) should come before the authors talk about y being an uncertain function of x?

2. Page 2375, line 8 "a constraining X" should read "a constraint on X" or something similar.

3. Page 2376, line 6 "at the point", perhaps should be "at the input point x" for clarity.

4. Page 2380, line 3 "and also check that with the true value" should read "and also check that the true value"

5. Page 2380, line 17 "it it" should read "if it".

6. Page 2382, lines 5-6 "The emulator is composed of a basic linear statistical model, along with a more flexible part," I know the authors want to avoid too much statistical terminology but I think it would be reasonable to add here "know as a Gaussian

process" or some such phrase.

7. Page 2387, line 10 "This could be a powerful in the process of simulator development." should read "This could be a powerful technique in the process of simulator development." or something similar.

8. Page 2396 lines 2-3 "Implausibility is calculated the maximum of that from all three summaries" should read "Implausibility is calculated as the maximum of that from all three summaries".

9. Page 2399, Fig 7.a. 3rd panel from left: the blue representing the PDDFI input seems to be missing from this plot, where as the plot directly below includes this blue. As this input (and several other inputs) are not constrained by the maximum ice thickness output this is not so important but for consistency it would be good to have the same background blue colour in both Fig 7.a. 3rd panel and Fig 7.b. 3rd panel, if it is not too much trouble.

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