Earth Syst. Dynam. Discuss.,
doi:10.5194/esd-2016-17-RC2, 2016

Earth System Dynamics Discussions

Interactive comment on "The impact of structural error on parameter constraint in a climate model" *by* D. McNeall et al.

R. D. Wilkinson (Referee)

r.d.wilkinson@sheffield.ac.uk

Received and published: 22 June 2016

This paper describes a thorough and detailed investigation into the ability of FAMOUS to predict forest fraction. The paper starts from the pretext of being given an ensemble of pre-run simulator evaluations and observation data corresponding to some of the outputs, and being asked to estimate some of the parameters. The work applies the latest statistical thinking/methodology in a largely clear and careful manner. To my non-climate trained eye, the authors seem to learn things about FAMOUS that were possibly unknown before, and likely to be of interest to the community of climate modellers. In my opinion the work deserves to be published subject to a few minor changes.

I have two main criticisms of the paper. The first is that it is slightly repetitive in places. Several of the plots show very similar information, and make the same point albeit in

different ways (which may be the intention). I felt the main point of the paper could be made in less space, and that this would improve the paper.

My second criticism is that the paper is philosophically confused in places. This isn't necessarily a criticism of the paper, as most of the computer experiments community is somewhat confused about model discrepancy (as am I), but I felt the discussion lacked depth and nuance in places. Note that many of the following points are discussion rather than suggested changes to the manuscript.

Simulator discrepancy

As discussed, estimating simulator discrepancy is hard, as it is difficult to disentangle the effect of simulator discrepancy from the problem of estimating unknown parameters. I don't like the definition of discrepancy quoted from Williamson et al 2014, that discrepancy is an error that cannot be removed by changing the parameters without introducing more serious biases to the model. The problem is that what constitutes an acceptable discrepancy function depends upon your goal. If you aim to do prediction, then something like the above would work, as we just want to characterize the simulator error for a given parameter value. However, if the aim is to infer the parameters, and for that inference to relate to the 'true' value of those parameters, then you have to aim to model the true simulator discrepancy, which is much much harder. The problem that is hard to overcome, is that we may find the smallest simulator error occurs at parameters that are far from their 'true' values if the simulator is poor. Brynjarsdottir and O'Hagan make the point that strong prior information is needed on the true parameter values if you wish to have any hope of disentangling the parametric uncertainty from the discrepancy. I think the aim of this paper is to estimate parameters, but the approach taken is one that is perhaps better suited to prediction problems.

A discrepancy emerges in the paper, and is argued for by showing that there is an

irresolvable error. The argument used is a kind of minimum error argument: we can't simulate all four forests simultaneously, but we can do three, so let's have a discrepancy just on the Amazon, and assume the simulator is fine for the others. This sounds sensible, but it could be that the Amazon is correct and the others wrong, or that there is simulator discrepancy for all four when we use the true parameter values. I could imagine that the errors are highly correlated for the forests, so that this kind of weight of evidence approach may be flawed. This also highlights for me the weakness of this approach compared to a more traditional statistical approach. If we had statistically modelled the discrepancy, described priors, and inferred posteriors, I suspect a similar conclusion may have been reached, but the weighting would have been done using the rules of probability, and the argument would instead be over the choice of model. Here, although it is unclear to me quite how the conclusion was reached, it seems that the authors avoid the need for modelling assumptions, but instead use an informal and heuristic weighting arguments to decide where to place the discrepancy. Although they have a mechanistic explanation of why their approach makes sense, the danger is that this is done post-hoc to fit the results.

A final point on the discrepancy concerns the sentence 'We do not have enough information to create a more detailed discrepancy function: for example, one that varies across parameter space'. Why would the discrepancy vary across parameter space? I thought it was the difference between the simulator and reality when the simulator is run at the 'true' or 'best' input?

History matching

In the statistical part of the computer experiment community, there is an ongoing debate about whether we should do calibration or history matching (HM). I sometimes feel that HM advocates are too critical of calibration, criticising implementation problems

as if they were fundamental flaws in the framework, and conversely that the calibration crowd simply don't consider doing anything different. I like the idea of history matching, and have used it in my own work, but my understanding is that it was developed for situations where you have a huge input space, most of which is implausible, which can then mean that it is hard to accurately emulate the simulator across the entire input space. If this is the case, conservatively ruling out parts of space in a sequence of HM waves, can make emulation much easier. I have heard HM advocates then say that they might finish the analysis with a calibration, which again makes sense to me, as this can provide more nuanced information along the lines of 'we can't rule out $\theta = 2$, but it is much less likely than $\theta = 3$', which are statements that cannot be made within a HM approach. For the situation considered in this paper, there is no need to do waves of HM, as the emulator is adequate, and the data are such that only a small proportion of space can be ruled out (43% ruled out in the end). I can't help but feel that statistical calibration would have been the better approach in this case (although this is a matter of taste). Indeed, although the authors provides a brief explanation of why they prefer HM, in several places, the authors treat the output of their inference as if it were the result of a probabilistic calibration.

For example, Figure 16 is misleading. The histogram is suggestive of this being a distribution over the parameters. But as history matching was used, not calibration, there is no relevant information about the relative weighting of the parameters. This error is compounded in the sentence 'The relative frequency of NROY points is higher in some locations than others [...] suggesting a higher probability that the best estimates of the parameters is in these regions'. No statement can be made about probability here, as no probabilities were used and so this is misleading.

Line 6-8 on page 9 puzzled me, and also made me think that probabilistic calibration was perhaps what the authors had in mind. The claim is that finding the NROY region is near the edge of parameter space suggests a discrepancy function. I didn't really understand why this should be so, unless there is a secret/undeclared prior distribution

that the authors have in mind, and that they believe the parameters really lie near the middle of the a priori plausible region. Of course, in a HM approach these consideration are not taken into account.

On page 7, line 10, the authors say that the 'key' difference between calibration and HM is that points are not-ruled-out-yet (NROY) rather than 'accepted'. I find this point to be rather pedantic, as it is just a matter of labelling. I would say the key difference is that HM classifies points, but calibration describes a probability distribution over them. If we did calibration with uniform priors and thresholded the likelihood (using a pseudo-likelihood of either 0 or 1), then the two approaches can be made algorithmically equivalent (the interpretation remains different).

Finally, HM uses the implausibility given by equation 2 to score points, and then rejects points with a high score. We know from the theory of scoring rules that it is important to use a proper score, yet we can show that this score is improper (e.g. Gneiting and Raftery, JASA, 2007). Why doesn't this matter? We could use other scores in HM, and cut-offs other than the 3 sigma rule, and indeed on page 11, line 26-30, variations on how to threshold the plausibility are discussed. I support the authors' call for more research on the behaviour of the measures of implausibility, and perhaps suggest that links to scoring rules are investigated.

Other points

- Page 8, line 27. Where does the 0.05 observation error come from? And the sentence 'This corresponds to an expectation that the true 95% CI of $\pm$ 0.15' is incorrect I think. Pukelsheim's rule says that the 95% CI is contained within $\pm$ 0.15, not that it is equal to it. For a Gaussian rv, this would be a 99% CI for example.
- There is some confusion over the projections of points in the plots. In figure 8

for example, error is shown as a function of two parameters, where the effect of the other parameters has been averaged out. Is this useful? Just because the average error is zero, doesn't mean the error is zero anywhere. I appreciate this probably isn't what is happening, but the plots aren't necessarily a good idea.

- Page 11, line 14. I don't understand the final sentence here? According to equation 2, it makes no difference whether we assign the uncertainty to the observation or the model discrepancy. And why would we want to do this? We were told observation error was known (and fixed).
- Another point that is more discussion then criticism, as I believe it is probably common practice, is the issue of treating the climate as a static system, by spinning up the climate model to reach equilibrium. Again, I'm not a climate scientist, but as the climate is dynamic, does this practice cause a bias? Suppose we had the true simulator, with zero discrepancy, would spinning-up to equilibrium induce an error in our predictions? I appreciate there is probably no way around this.
- The language needs editing in places, with errors becoming increasingly common in later sections.

Minor points

- Page 1, line 10, 'find the parameters that have most impact on simulator error'. To be slightly nit-picky, I don't know what this means. Perhaps 'find the parameters that have most impact on simulator output', as simulator error, probably means the error when run at the best input.
- Page 2, line 8-11. This description is slightly confusing. Calibration, tuning, and history matching are all solving the inverse problem in some sense. Needs rephrasing, and perhaps a reference or two.

- Page 5, line 7-8. I don't believe the claim that LHC designs are better than others. I read Urban and Fricker a long time ago, but I think they just compared LHC to grid designs, and then only in empirical experiments. I'm pretty sure it is not the case that the question of the best design is settled in general (see Zhu and Stein 2006, and Zimmerman 2006 etc). I think it would be better to say that LHC designs are 'good designs'.
- Page 6, line 25, 'Gaussian' not 'gaussian'
- Page 6, line 29, 'The emulator is a nonlinear regression model' perhaps 'non-parametric' would be better than 'nonlinear', given the potential for confusion with what is normally meant by 'nonlinear regression model', i.e., non-linear in the parameters.
- Page 6, line 31. Given it is quite a long paper, there are remarkably few details about the emulator, covariance function, mean function, estimation approach etc. The review guidelines ask me to check that the paper is reproduceable, which without these details, it would not be.
- Page 8, line 31, 'We sample from the emulator uniformly across input parameter space' - this is unclear. Presumably you sampled uniformly from the input parameter space, and then from the emulator. Same again on page 11, line 2.
- Page 10, line 7, 'total effect'
- Page 12, line 29. 'that that model discrepancy uncertainty is zero'.
- Page 16, line 29/30. Rephrase sentence 'First, is there...' A dodgy emulator would lead us to think a bias exists, not cause it.

Interactive comment on Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-17, 2016.