Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-185-RC2, 2016

Geoscientific Model Development Discussions
Open Access
EGU

Interactive comment on "Tuning without over-tuning: parametric uncertainty quantification for the NEMO ocean model" *by* Daniel Williamson et al.

R. D. Wilkinson (Referee)

r.d.wilkinson@sheffield.ac.uk

Received and published: 6 October 2016

The paper by Williamson et al. looks at an approach for estimating unknown parameters in the NEMO ocean model. History matching is the approach taken, in conjunction with the use of Gaussian process emulators. It is great to see advanced statistical methods of this kind being used in climate modelling. The paper has a pedagogical feel to it in places, but I appreciate this, and think it will benefit the community. Overall, I like the paper, and hope it is published in gmd. There are a few places where I disagree with the presentation of the statistical methodology, but that is my only complaint.

Firstly, I agree with the other referee's comments that the additional terminology introduced by the paper, for concepts which already have established names, is unneces-

Printer-friendly version

Discussion paper

sary and potentially will further muddy the waters. The computer experiment literature already insists on using specialist language which confuses some other statisticians (emulators, calibration, etc), and I feel that 'iterative refocussing' will only add to the problem, just at the stage where more people are becoming aware of history matching.

I also dislike the use of 'over-tuning'. Overfitting a model occurs when we use a model which is too complex, so that we describe the noise not the signal. That may be happening here, but isn't the point that is being made. For example on p2, line 25-26, the authors are warning against the dangers of just using a single parameter value, rather than considering parametric uncertainty, a point I agree with, but which isn't related to overfitting or over-tuning (which as far as I can tell means the same thing).

The paper is also too dismissive of 'Bayesian calibration', which is what most Bayesians would just think of as inference. I like history matching (HM), and agree that is has some strengths that make it an attractive choice in situations where we want to avoid specifying a detailed statistical model.

For example, I like the benefit of HM as expressed in the para containing line 15 on page 13, that it is a conservative method that only seeks to rule out bad regions of space, rather than find good regions (the former being much easier than the latter). For this reason, it is particularly suited to situations where the errors are not known well, as described on page 15.

But the comment on page 3 line 7, that Bayesian calibration 'can also be described as forms of optimization and suffer from some of the drawbacks stated above' makes little sense to me. Bayesian calibration can be made equivalent to history matching if we use a likelihood function which is an indicator function based on the implausibility metric, and flat priors (see for example Holden et al 2016 available as arXiv:1511.03475, where the similarity between ABC and HM is discussed). The problem with Bayesian calibration occurs if we assuming likelihoods that are inappropriate. If we instead use conservative likelihoods as in HM, then we would also avoid the problems described.

I also dislike the description of history matching as a method that inherently is based upon the use of emulators. To me, history matching is defined by the use of implausibility metrics to find not ruled out yet (NROY) regions of parameter space. How we go about doing this is a matter of implementation, where we may (indeed most likely will) find that emulators are of benefit. The advantage of viewing HM as distinct from the algorithm used to implement it, is that we can then define the 'true' HM answer (which is a NROY set). This then allows us to independently answer the two questions

1. How does the NROY set compare to answers from using other likelihoods (ie Bayesian calibration approaches) or implausibility measures?
2. How good is any given implementation/algorithm at finding the true NROY set?

If we define HM to be iterative refocussing with GP emulators, then these two questions are conflated.

If we separate these points, then I'm not sure that line 9 on page 9, that the order in which metrics are applied matters, makes sense. Unless something is happening I don't understand, the final true NROY set should be independent of the order in which the metrics are applied: if it isn't, then it says something worrying about HM as an approach. Note that I can see how the order matters in the implementation, and that some orderings will make finding the NROY set harder (something analogous to mixing in MCMC or SMC for example), but that relates to question 2 above (implementation) not question 1.

The para around line 20 on page 3 claims several benefits for history matching, and the implication is, that these would not be available if one were to use for example, calibration, which simply isn't the case. They are benefits from the careful statistical analysis (particularly the emulation) that is done, and shouldn't just be claimed for HM (or iterative refocussing).

Printer-friendly version

Discussion paper

Another case where HM is over-sold, is at the bottom of p12. In this case, the choice of a poor score (line 10) is to blame for the counter-intuitive result described. If the variance is not constant, then we should score using the log-likelihood say (or some other proper score), rather than the Mahalanobis distance, which is an improper score. If we added (or subtracted depending on your setup) $\log\det\Sigma$ to the Mahalanobis distance, then we would no longer find the optimum value occurs at an x where we are most uncertain.

In conclusion, I would like to see a more balanced discussion of HM in the paper. I think the approach makes a lot of sense for these models and can be argued for persuasively, without conflating the issues discussed above.

Finally, is there an error in equation 2 in the nugget term? Shouldn't the nugget only get added for the same model run, i.e.

$Cov(x_i, x_j) \propto I_{i=j}$

rather than

$Cov(x_i, x_j) \propto I_{x_i=x_j}$

if the nugget is there to represent internal variability?

Minor points

p1, line 21, 'and' -> 'which'?

p2, lines 15-19. Something about how this is structured confuses me. Perhaps consider rephrasing.

p4, line 13 'depracated' -> 'deprecated'

p10, line 5, change 'it's' to 'its'

p10, line 15, an emulator isn't inherently Bayesian, unless you make it so.

p11, line 33, where can the code be downloaded from?

p12, line 9, ’Mahalanobis’ rather than ’malhalanobis’ (spelling and capitalisation)

p12, line 18, I don’t think you want the inverse variance (delete the ’-1’)?

p12, line 28, comma after x*

p17, line 7, has something gone wrong with this equation? It doesn’t seem quite right.

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-185, 2016.

Printer-friendly version

Discussion paper