Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-185-AC1, 2016

Geoscientific Model Development Discussions
Open Access

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

Daniel Williamson et al.

D.Williamson@exeter.ac.uk

Received and published: 24 November 2016

[a4paper,11pt]article

Author's response to Reviewer 1

November 24, 2016

We'd like to thank all 3 referees for their constructive and considered reviews of our paper. The reviews we have received fall into two categories. The first two reviewers appear generally happy with the methodology and the application, but have comments, questions and concerns about our presentation and terminology relating to the statistical ideas in the paper. We will respond to each of their points in turn and add further clarifying remarks to the paper where appropriate. However, we would like to say, in general, that we have been very careful in our choices of language and terminology and have made these choices in order to resonate with and be familiar within the community of GCM/ESM developers and tuners at modelling centres. Experience of working with modellers from some of these centres for the last 5 years, and engaging with these important communities has led us to many of the choices that our statistical reviewers take objection to. Hence much of our response to these comments will be an attempt to explain and defend our choices. We certainly welcome the chance to do so in a public discussion and are grateful to both reviewers for raising the important questions that they do.

The response of the final reviewer, a model developer/tuner and a member of our target audience contains a number of points of clarification and we will endeavour to answer

Printer-friendly version

Discussion paper

each of these and add suitable passages to the paper in response.

Referee 1

The authors propose a method for narrowing down the parameter space of a numerical model by ruling out parameter values which are not consistent with observations. The procedure is iterative. At each stage of the iteration additional observations are considered. Observational uncertainties are taken into account. The method avoids over-fitting since it does not aim at selecting a single optimal parameter combination which brings the model simulation closest to observations. The manuscript is well written and illustrates the methodology nicely using an ocean model. The main issue in my opinion is that the methodology is essentially a Bayesian parameter estimation procedure and does not deserve a new name. As far as I can see, in a Bayesian formulation, at each step the authors basically rule out parameter values which have a posterior likelihood below a particular threshold.

In fact, the whole methodology would be much clearer and more transparent if the authors would acknowledge that what they do is a Bayesian procedure, and write down the prior distributions and the likelihood function. Also in the context of a carefully conducted Bayesian parameter estimation, the modeler would not just blindly select a single optimal value, but explore a range of parameter values which are broadly consistent with observations. In case the authors disagree, they should discuss the relation to Bayesian parameter estimation, highlight what is new in their approach, and indicate the advantage of their methodology. I am very doubtful that there is an aspect to "iterative refocussing" which is not naturally (and more transparently) part of an iterative Bayesian procedure.

History matching and iterative refocussing is not a (fully) Bayesian procedure, though this reviewer is not the first to claim it as such. The reason I use the term ‘fully’ above is that it could be argued that the procedure is a partial or second order Bayesian procedure, coming as it does from a Bayes linear methodological background. The essence of the Bayes linear procedure is to avoid altogether the specification of probability distributions and to work instead with partial moments and expectation primitive. The latter convention ensures that we do not even assume probability distributions must exist given the specification of a finite collection of moments. Without a probability distribution, either explicit or implicit, there is no justification for the claim that this is a Bayesian procedure. This is argued elsewhere, specifically in Craig et al. (1997), where the term "iterative refocussing" originates, Williamson et al. (2013) which discusses the use of expectation as primitive in history matching, and, effectively, in the discussion and rejoindre to Vernon et al. (2011).

Whilst accurate, a pragmatic argument is sometimes made that even though we never need probability distributions and do not assume them, you would get the same answer if you did, given a procedure based on removing near-zero likelihood regions of the parameter space. This is my interpretation of the reviewer's argument here. However, there is a subtlety that sets the methodology apart from Bayesian parameter estimation even if you were to interpret much of what we do probabilistically as the reviewer has done, rather than viewing expectation as primitive as we have. Bayesian parameter estimation, an approach first laid out for computer simulators fully by Kennedy and O'Hagan in 2001, requires a prior distribution $p(x^*)$ that is then updated by the ensemble, F, and the observations, z, through Bayes theorem to give $p(x^*|z, F)$.

Put technically, the parallel with our method would be if we were to assume a uniform prior for x^*, then

$$p(x^*|z, F) \propto p(z, F|x^*)$$

and our procedure, roughly speaking, sets $p(z, F|x^*) = 0$ if $-log(p(z, F|x^*)) > a$ for some threshold a. But we would be doing this assuming that this implies $p(x^*|z, F) \approx 0$ due to its proportion to the likelihood in those regions of parameter space, which is not necessarily true. Within the probabilistic paradigm, $p(x^*|z, F)$ must integrate to 1, hence even when the model wildly differs from the observations and our discrepancy assessment, we will have a non-zero posterior. In fact, in this case, we will see a highly peaked posterior at the value of the model parameters where either uncertainty in the emulator is largest or at the value closest to reproducing the observations (even though it could be any number of standard deviations away under our model). The more inconsistent our model and specification with our data, the sharper the peak of this posterior. By taking the Bayesian route at all and having a prior, we assume x^* exists (with probabiliy 1), an assumption we can never coherently row back from.

What is happening here is even low near-zero likelihoods are modified by huge normalising constants to give considerable posterior probabilities to 'very far away yet closer than anywhere else' regions of parameter space. History matching has no problem here. We never assumed nor work with probability distributions and hence can coherently remove all of parameter space advising the modellers that the answer to the important question "can the model reproduce these observations/behaviours to within a given tolerance to error?" is no. How can the Bayesian proceed? The Bayesian has already ruled out the negative answer to the above in the prior. Bound (correctly) by coherence and in the face of the approximation $p(x^*|z, F) \approx 0$ for near zero likelihood breaking down, it seems that all is to be done is to revert to the posterior. Exactly what is to be reported or what inference can be drawn from here beyond pointing to the region of space where the posterior mass is concentrated as the most likely to contain good models is unclear.

Though we could imagine an iterative Bayesian analysis undertaken until convergence on a poor solution (if we have enough computing power to get there) could lead to the same conclusions (we simply get to a point where we stop and declare our statistical model invalid), we would question why should we be required to make infinitely more assumptions and judgements (in the form of full joint continuous probability distributions) in addition to new methodology required to establish and interpret convergence so that we can claim the Bayesian title for our procedure.

We say a lot more about the comparison to Bayesian parameter estimation in our response to reviewer 2 as they have raised a number of specific objections on this theme. To conclude the response to reviewer 1, we hope this answer is sufficient to convince you that History Matching is fundamentally different to Bayesian parameter estimation. A full discussion of the respective merits of these two approaches falls without the scope of this paper, and we believe would be more at home in a statistics journal. Including such detail in this paper would, we believe, risk deterring our target audience of geoscientific model developers and tuners from engaging fully with the paper.