Geosci. Model Dev. Discuss., 7, C581–C588, 2014
www.geosci-model-dev-discuss.net/7/C581/2014/

Geoscientific Model Development Discussions

Interactive comment on "Probabilistic calibration of a Greenland Ice Sheet model using spatially-resolved synthetic observations: toward projections of ice mass loss with uncertainties" *by* W. Chang et al.

T.L. Edwards (Referee)

tamsin.edwards@bristol.ac.uk

Received and published: 20 May 2014

This paper describes a proof-of-concept study, emulating the SICOPOLIS Greenland ice sheet model and performing test Bayesian calibrations with synthetic observations. It compares this with a previous, simpler assessment of the ensemble and model parameters. This is a welcome addition to the literature and lays useful groundwork in a research area I think has exciting potential.

The paper uses standard methods of uncertainty quantification (UQ) for computation-

ally expensive computer models, and aims to explain the statistical aspects so they are understandable by the non-expert. It is somewhat similar to McNeall et al. (2013, GMD) but makes advances both in the complexity of the emulated output (showing the way to more sophisticated use of observations in future studies) and the use of Bayesian parameter calibration, rather than parameter range exclusion (aiming for probabilistic, and therefore more meaningful, projections of ice sheet contributions to sea level). These two aspects are not novel methods but are novel in their application to a Greenland ice sheet model. They therefore begin (in common with some of the literature cited below) to address the particular challenges of Bayesian calibration of ice sheet model projections. The main paper is mostly very well explained. It is therefore a useful addition to the literature not only scientifically but also for those new to emulation and Bayesian calibration generally.

My main scientific criticisms are: (a) some imprecise and unsupported statements, and missing discussion points, which I would like to see addressed; (b) unnecessarily poor justification of using synthetic rather than real observations (by unnecessarily I mean it is both valid and useful to use synthetic observations for such a proof-of-concept). The paper's conclusions would also be more robust and substantial if the leave-one-out testing were repeated for all ensemble members, as is (reasonably) standard, or at least for more than three (10 or 20 might seem a reasonable minimum number to me).

Another weakness is awareness of the relevant literature, including results from IPCC AR5 and the ice2sea project (See http://www.ice2sea.eu/programme/published-papers for references) and the UK UQ community.

Finally, it may be because I read the SI a while after the main paper but I found much of Sections 4 and 5 in the SI difficult to understand. I hope the brain dump of points I found confusing and suggested improvements given at the end of this review are useful.

==========

1. Scientific points

1906/12 I would argue you can make probabilistic projections without calibration (i.e. present prior density, if no observations available).

1906/13 You don't use observational data but synthetic observations...

1907/15 "primarily" - No, see IPCC (2013) sea level chapter: projections are primarily based on regional climate models, ice sheet models and glacier models

1907/15 "spatial distribution" - and flow

1913/16 "identify" -> present / describe / test (not sufficiently novel to warrant identify)

1915/24 This is not a good justification for not using observed geometry, and doesn't match the scope of the paper. For me the justification is that you want to test the ability to retrieve the original parameter values - which you do - rather than make calibrated projections of the future. Whether the model gets the observed geometry right or not is not relevant to the stated aims of the paper. Your statement that model limitations will "cause problems" is both ill-defined (do you mean discrepancies too large to have any effect on the posterior? or more difficult to construct a statistical model of the discrepancy?) and not supported (by showing the ensemble thicknesses against the observed to demonstrate that they disagree substantially, or performing the calibration with the observations to show it has little effect). See also comment below on p1921.

1917/2-8 These lines do not describe passing of check #1: lines 9-13 do. (Suggest new paragraph at line 13 and reordering previous lines to reflect this.)

1917/27 There *is* clustering around the best estimate for ice PDD: this should be mentioned.

1918/6 and Fig, 5 How do you obtain 95% intervals for the Applegate windowing method? By retaining those within $\pm$9.5%?

1918/7 It doesn't only reflect the utility of spatial information: it also reflects the choice of window size (it could have been $\pm$5%), and your Bayesian statistical modelling

choices.

1920/3 "incorrect" -> "non-optimal" or similar; also, some modes look like they have similar density, so worth mentioning there may not be a unique best estimate theta*.

1920/10 "true" -> "best", as you use in the SI, because you are not talking about retrieving the synthetic observation parameters here.

1920/12 There isn't always wide variation: see e.g. ice2sea's Shannon et al. (2013), Edwards et al. (2014b)...

1920/14 And potentially also differences in spin-up method, if by "reproduced the modern ice sheet equally well" you mean only the topography (rather than dH/dt, velocities, ice temperature etc).

1920/25 "generally too thick" - show ensemble and obs in Figure? (as per comment above on p1915)

1920/27 "difficulties" - see ice2sea results for improvements in spin-up methods (holding to modern geometry, relaxation, using SMB corrections through the model simulations to account for remaining errors) - as per comment on p1908 in Section 2 below.

1921/2 Here is a bit more detail on why you are not using observations. Again I don't agree with this justification - couldn't you test the effects of a large discrepancy term with the synthetic observations? But if you do include this justification, add it earlier (p1915) too.

Fig. 5 Are your kde bandwidths a bit small or are those bumps due to real physics?

SI/55 "our experiences" - reference? leave-one-out validation? What is "very accurate"?

SI/170 Does the error rate in the cross validation indicate some choices could be improved?

SI-187 Do you use the old Bamber et al. geometry to pick the ensemble members because this was done in Applegate et al., or could you use the updated (2013) geometry instead? Why not do leave-one-out for each ensemble member, not just 3? (N.B. I may not be sympathetic to the argument "it takes 8 hours"... ;))

SI-187 "essentially the same" - no, the effect of the calibration is much stronger! (maximum posterior density more than 2x greater) - why is this?

==========

2. Literature

1906 I think (of course I do...) it is relevant to cite Edwards et al. (2014a) and (2014b) The Cryosphere - probabilistic Greenland projections, calibrated in a Bayesian framework with spatial information from a regional climate model.

1906 IPCC (2013) replaces Meehl et al. (2007).

1907 If citing SeaRISE, should cite ice2sea project (had much, if not more, model development) too.

1907 Why cite a palaeo ref for melting vs elevation?

1908 Projections from the ice2sea project use the observed geometry in one stage of the spin-up to reduce these errors. For example, in Edwards et al. (2014b) and Shannon et al. (2013) multi-model papers; see also ice2sea Greenland papers led by Goelzer, Gillet-Chaulet, Quiquet.

1908 Cite Little et al. (2013, Nature Climate Change): a probabilistic study of Antarctica that uses a flow line model of the PIG (and extrapolation of observations elsewhere), calibrated with observations.

1908 As mentioned above, Edwards et al. (2014b) is a probabilistic projection for Greenland that uses a multi-model ice sheet ensemble and parameter perturbations calibrated using regional climate model data.

1909 Update McNeall et al. reference - now accepted.

1909 Edwards et al. (2014b) contains a $\sim$100 member perturbed parameter ensemble for one Greenland model, (technically PPEs for multiple models, albeit with N=3 per model for the others. . .).

SI / p3 Cite Kennedy and O'Hagan for emulation; ideally Goldstein / Rougier ref(s) too. If citing specific emulation applications, also include refs to the UK community (e.g. Sexton, Rougier, Williamson, McNeall, Lee) - or else remove Drignei et al. onwards and just cite emulation methods papers.

SI / p4 For theta* and discrepancy, again Goldstein and/or Rougier refs would be appropriate here.

==========

Clarity and other suggested improvements

It would be useful to have somewhere a clear summary of the main differences/improvements c.f. McNeall et al.: i.e. a different model, plausible magnitude present and future projections rather than idealised/palaeo-simulations, a smaller ensemble, the differences in parameters and ranges, emulation of higher dimensional output (latitudinal thickness PCs instead of scalar summaries), a Bayesian probabilistic calibration instead of a history matching approach. And anything else worth mentioning.

1909/9 May be worth saying explicitly that McNeall et al. study is not probabilistic.

1915/9-10 The observational discrepancy part is a little hard to understand: is it possible to illustrate it in a figure?

1918/5 Presumably you use the same windowing approach as Applegate et al. but with the synthetic (not observed) volume - probably a good idea to state this explicitly.

Supplementary Information

The SI would benefit from more explanation, and an assumption that the reader has not read the authors' previous work...

It also needs a more thorough proof-reading.

I am quite confused by the discrepancy modelling. Are the end of p4 and middle of p6 talking about the same thing? If so, I would move the latter forward and give the notation for the numerical choices (e.g. is phi_d 2100km? is kappa_d 2500m? how about the nugget of 1km?)

Some further suggestions if you would like to make all of the SI as clear as the main manuscript and SI introduction:

- a plain English outline before/after the maths in each paragraph / section;

- an illustrative example of one emulator (either an idealised single model output, or the future projections emulator) before showing the emulation of PCs;

- explanation of statistical terms for the benefit of numerical modellers: partial sills, nuggets, knots, identifiability, block updating, well-mixed chain (and what they mean in terms of assumptions e.g. is one of the nuggets the *emulator* discrepancy?);

- explanation of the subscripts y and d for the basis vectors, phi etc;

- a table summarising all parameters;

- explanation why kappa_y are re-estimated.

- explanation how J is chosen / how much of the variation these PCs account for

- ideally, explain how GP is different to linear regression emulation

- an illustrative figure would be useful

- translations of what the statistical model choices mean in terms of assumptions about the model

- could mention kriging, which may be familiar to the readers

- remove some unnecessary jargon, e.g. dispersed posterior density -> "has little effect" or similar

- many mentions of perfect model experiment as if it was one part of the paper, but I think it is all of it? Also I find synthetic observations more clear than perfect model (I use perfect model for studies that do not include a discrepancy variance. . .)

- you talk about joint and marginal estimation - do you use both? (e.g. for the different 1D and 2D figures?)

- explain definition of "error rate" in cross validation (p8)

Interactive comment on Geosci. Model Dev. Discuss., 7, 1905, 2014.