Review of 'The future of Upernavik Isstrøm through ISMIP6 framework: Sensitivity analysis and Bayesian calibration of ensemble prediction' by Jager et al.

Doug Brinkerhoff

May 22, 2024

1 Summary and Main Points

This paper presents a probabilistic forecast for mass change at Upernavik Ice Stream in West Greenland. In very broad terms the forecast is produced by running an ensemble of predictions over different model parameters using the ice sheet model Elmer/Ice, and then weighting those ensemble members based on their agreement with a variety of observations. The paper is remarkable in that performs a very detailed methodological exploration of different schemes for weighting ensemble members. The paper also presents an index based sensitivity analysis, allowing for an interesting temporal discussion of the influence of variance in model parameters on the variance in predictions.

Overall, I think that this paper is exceptional and represents an important advance in data-constrained forecasting for Greenland. In particular, the paper makes (and justifies) a rather significant claim, which is that at centennial scales, the dominant source of uncertainty for projection remains elements of climate forcing and that contemporary observational constraint on ice sheet models exhibits diminishing returns. I have no major methodological issues with this work and I find the science to be sound. I do think that the paper could be made more accessible to a broad audience – in particular one that isn't complete versed in the language of probabilistic forecasting, and for whom this paper would still be a very useful read – by some clarification in exposition. In particular, I think that the use of acronyms could be reduced, some sections could be shortened, and the lanauage of Bayesian inference modified to be in closer accordance with standard usage.

2 Line-by-line comments

L23 'limited' used twice.

L24 What is the difference between 'limited understanding of ice dynamics' and 'uncertainties in ISMs'?

L30 It's worth noting that in Aschwanden (2021) the authors state that ISMIP6 actually already does a really good job with quantifying uncertainty in model structure! For the other considerations, yes, those must still be better accounted for.

L34 This would be an appropriate reference for sensitivity studies in Greenland as well: https://doi.org/10.5194/tc-13-1349-2019

L47 For item (i), I think it makes sense to characterize this as 'establishing prior distributions' over uncertaint model parameters.

L50 It might be worth noting that this lack of cross-validation is often done because there's just not that much data to work with usually, but the authors' point generally a very fair one. I hope that future studies incorporate the authors' suggested cross-validation framework.

L111 The specification of the calving front position rather than it being a prediction from the model was and remains one of the most contentious aspects of the ISMIP6 experiments. While the position of the calving front may be specified with precision by the parameterization, that doesn't necessarily mean that the position will be accurate in the future, and successfully simulating calving rates and front positions remains one of the largest open challenges in glaciology. It is worth noting here that by ignoring this source of uncertainty in model projections, the authors' are making a large and potentially critical assumption. I don't think it's a problem, but it really does need to be discussed.

2.1.1–2.1.3 I struggle a bit with the semantic separation of the historic ensemble (hpr) from the two future ensembles (cpr and ppr), given that it is the initial period for both. I don't think it's too important but the authors for clarity may wish to refer to the historic ensemble as the 'shared hindcast period' or similar for the two prognostic ensembles.

L141–150 I find this section to be a little bit confusing. It might help to have a more complete discription of the parameterization and specifically the role of κ.

L195 'Ice Discharge' should be lower-cased.

L197 Given that this producet is based on a model result (RACMO), is there the potential for this product to contain systematic bias?

L203 'ensemblist' $\rightarrow$ ensemble

L225 I think that $\mathbf{M}$ needs to be understood as a random vector of model parameters, with $P(\mathbf{M})$ its prior distribution. It will be the case that a sample will be drawn from this distribution, which will be used to create the ensemble, but $\mathbf{M}$ is not the ensemble itself.

Eq. 6 I think that some of these components are mislabelled. In particular, Eq. 6 is not Bayes' theorem, so I'm not sure it makes sense to refer to a posterior and prior as such. Rather, Eq. 6 is the definition of the 'posterior predictive distribution' (which is what the left-hand side should be labelled), which is the distribution of future sea level outcomes conditioned on data. This is decomposed into the two terms on the right side,

$$P(\mathbf{M}|\mathcal{B}),$$

which is what would usually be called the 'posterior distribution', 'parametric posterior distribution', or 'calibration' (as it already is in the paper) to disambiguate, and

$$P(SLR|\mathbf{M}, \mathcal{B}),$$

which is the distribution of model predictions given a particular parameter value (perhaps called 'model prediction' or 'projection').

Eq. 7 An important condition here is that

$$\mathbf{M}_i \sim P(\mathbf{M}),$$

which is to say that the realizations of the particles need to be drawn from the prior distribution. If that's not the case, then the prior needs to appear in the numerator and denominator of the term in L. 237.

L240 'gaussian' should be capitalized.

Eq. 8 This intersection notation is weird – I think it would be better to just start with Eq. 9.

Eq.11 A should be $A^{n_{obs}}$ or a new constant should be used.

L248 Be explicit as to what this measurement operator is, e.g. the evaluation of the Elmer/Ice FEM basis representation of the velocity field at the desired locations in space and/or time.

L255 It is also worth noting that even if observational uncertainty were IID, model error *definitely* is not, which is what ultimately leads to the egregiously peaked distributions over ensemble members and heavily weighting only a single one. Ultimately the problem is that – priors aside – we don't know an appropriate likelihood function for comparing models to data! As such, the more *ad hoc* methodology described in this work is justified.

L265–282 This section is really great. It has significant similarities to lots of previous work on Bayesian inference in the presence of model misspecification, and it might be worthwhile to frame the discussion in terms of that. This is a good reference: https://doi.org/10.1146/annurev-statistics-040522-015915

L280 –282] I'm not sure I understand this statement.

L304 It would be worth describing whether this weighting scheme is more or less permissive than full-period weighting - I don't have a good intuition. It might also be helpful to introduce an equation for each of the weighting schemes to be concrete.

L308 I don't really understand the introduction of f_{param} weighting. This is very much tied to the particular parameterization and behavior of the authors' existing model setup (thoroughly described in a separate paper) and it's challenging for someone not so involved with that work to understand what this specific experiment is trying to capture. Can this be expanded to provide more substantial justification?

L232 This sentence changes from passive to active voice in the middle. Probably best to stick with active voice.

Sec.2.3.3 While I appreciate the desire to include SSP as a random variable, I also think that doing so sort of obscures the influence of all the other aspects of the model since this is expected (and turns out to be) a dominant factor in determining predictive variance. Is it possible in later plots to also present ensemble ranges conditioned on SSP (i.e. the sub-ensembles of particles using just SSP2.6, SSP4.5, SSP8.5)? That would be helpful for comparison against existing similar work and would also facilitate a 'best-case versus worst-case' analysis for climate change impacts.

L343 It would be super helpful to reiterate here what the difference is between the Ppr and the Cpr. I had initially thought to suggest more instructive names, but I can't think of any, so at the very least a brief reminder of the assumptions of each would be great.

L356 Is the agreement between the Hpr median and mass loss observations by design or a happy accident?

Fig. 2 and 3 Perhaps consider changing the symbology scheme to something friendly for greyscale/colorblindness, e.g. cross-hatching one of the two shaded regions.

L361 It might be worth contextualizing this with respect to Robel, 2019: https://doi.org/10.1073/pnas.190482211. The skew in the distribution is perhaps not surprising.

Sec. 3.3.1 It's a little cumbersome to start a section with a reference to another section. I understand shunting the methodological details to the appendix, but a recapitulation of the methodological approach would be helpful here.

Sec. 3.3.1 More generally, all four points introduced here seem a bit *ad hoc*. Do there exist references that could help place the current procedure on more sound probabilistic footings? Seems like this problem has to have been studied before.

L446 Where is factor mapping previously established?

L448 The parameter f_{law} sometimes appears throughout the manuscript as just law. Please revise for consistency.

Fig. 5 The overlying transparent bars aren't really readable.

Sec 3.4.1 I think that this section would benefit from a bit of extra subdivision. I think it would be helpful to break this into individual subheadings describing the historical period and the forecasts. I think it would also be helpful to separate the principal conclusions about the relative insensitivity of long term forecasts to ISM parameters from the details of weighting. I also don't think that it makes sense to refer to the changes in ranges described around L506 as 'notable' – the more notable thing is that they're almost exactly the same!

Fig. 6 The font in this figure is too small.

Sec 3.4.2 Again, I would like to reiterate that presenting ensemble results which each of the SSPs held fixed would be useful here, and would help to ameliorate some of the challenges associated with trying to guess the probabilities of future human carbon emissions (which is why previous works have treated these as fixed hypotheses rather than as random variables).

Sec. 4.1 I am not quite sure I understand the reference to ISMIP6 here. How is that relevant to the present model being able to reproduce observations?

Sec. 4.2 Again, I think that this section would be clarified by adding some more sub-headings. e.g. at L589, this paragraph could be called 'reducing uncertainty through ISM calibration), whereas at L598, this could be called 'reducing uncertainty through climate forcing calibration', or something like that.

L576 The referenced compensatory effect is not clear to me from Fig. 4 or elsewhere. Could this please be clarified?

L585 If the front parameterization has such as significant influence, then perhaps this calls into question the validity of imposing the front at all. Would it be possible to make a statement about how the predictions might be influenced if the front were allowed to evolve freely or based on a different parameterization?

L602–604 This is a very significant assertion that would have major implications for how ice sheet modeling proceeds in the future! What should the community be doing if improving ice dynamics isn't likely to help? (note that I don't disagree with the assertion – I am genuinely curious where effort should be allocated instead).

L618 Its foundation in observational data is sort of the problem – no data available in the future.

L703 There are other possibilities than the Gaussian or T.

Sec. 4.4 There is a lot of good in this section, but there is also a lot of material that is only applicable to the authors' own modeling setup (which has already been covered) it might be worthwhile to take a critical read to assess which of these insights are going to be generally applicable, and which are more like notes to guide the authors' own continuing work.

Sec. C2 The student-t distribution has an additional degree of freedom, namely the number of degrees of freedom. What was used for this, or how was it estimated?