Lessons from the short history of ice sheet model intercomparison

E. Bueler

Dept. of Mathematics and Statistics, University of Alaska, Fairbanks, USA

Received: 21 April 2008 – Accepted: 21 April 2008 – Published: 29 May 2008

Correspondence to: E. Bueler (fjelb@uaf.edu)

Published by Copernicus Publications on behalf of the European Geosciences Union.
Abstract

Intercomparison should include measurement of differences, between model results and observations, among the model results themselves, or between model results and exact solutions. The processes of measuring differences and critically analyzing those differences are vital. Without such measurement as a component of intercomparison, the only expected benefits of an intercomparison project are participation, possibly the discovery of communal confusion, and the establishment of public, non-proprietary data sets.

1 On motivation

I perceive four cultural purposes of simplified geometry intercomparison projects:

(i) to get everyone together to work on a new and intimidating job of common interest,
(ii) to produce publications early in the development of new codes,
(iii) to produce benchmark experiments (or “setups”) usable by current and future modelers in understanding new models and debugging new codes, and
(iv) to produce benchmark results for the same reasons as in (iii).

My perception is based upon completed rounds of simplified geometry intercomparisons, namely EISMINT\textsuperscript{1} I (Huybrechts et al., 1996) and EISMINT II (Payne et al., 2000), and upon the yet-to-be-completed simplified geometry parts of ISMIP\textsuperscript{2}, namely HOM, HEINO, and MISMIP.\textsuperscript{3}

\textsuperscript{1}“EISMINT” = European Ice Sheet Modeling INiTiative.
\textsuperscript{2}“ISMIP” = Ice Sheet Model Intercomparison Project.
\textsuperscript{3}Regarding the simplified geometry parts of ISMIP, as of April 2008: The intercomparison work of ISMIP-HOM is completed, the collective report is under public review, and a perfor-
All of the above-mentioned simplified geometry intercomparisons have succeeded in purpose (i). Notably, ISMIP-HOM coincided with, and stimulated, explosive growth in the number of higher-order and Stokes models. Intercomparison exercises are justified by the prospect of participation and discussion. The “value” that participation represents, for instance to the agencies which fund intercomparison, far exceeds their direct creation of publications. As an additional example, participation in intercomparison exercises has stimulated to the development of public ice sheet codes (Bueler et al., 2008; Payne et al., 2008) and monographs (MacAyeal, 1997).

Purpose (ii) is equally serious. Modelers are more willing to take risks if some kind of result is likely in the shorter term. ISMIP-HOM and MISMIP clearly provide this benefit to the modeling community. The shallow ice approximation (SIA)-based intercomparisons EISMINT I/II and HEINO compared existing ice sheet codes, to a significant degree, and probably involved less risk-taking in this sense.

Purpose (iii) is always an intention of an intercomparison, not just the simplified geometry type. If the only result of an intercomparison was creation of a few standard problems, however, then it would be a dull one indeed. Individuals or informal groups can usually think up good experiments, so intercomparisons ought to add value beyond goal (iii). There ought to be comparison in intercomparison.

2 On measurement and mathematics

Benchmark results, purpose (iv) in the list of motivations above, is a stated or implied goal of all intercomparison. If the results of a simplified geometry intercomparison are tightly clustered then future readers will inevitably expect to compare new model results. Analysis of one participating result is under public review (Pattyn and Payne, 2008; Pattyn et al., 2008; Gagliardini and Zwinger, 2008). The intercomparison work of ISMIP-HEINO is completed but the collective report is not submitted, although one participating result is published (Calov and Greve, 2008; Greve et al., 2006). The intercomparison work of MISMIP is not yet completed (Schoof et al., 2008).
The EISMINT I report (Huybrechts et al., 1996) is explicit that the average result is intended to be a benchmark in this sense.

By contrast, the EISMINT II report (Payne et al., 2000) more-or-less disowns the benchmark claim. There is widespread acknowledgement that EISMINT II revealed an important feature of the nonsliding thermomechanically coupled shallow ice approximation, namely its instability as a fluid model in certain circumstances (a flat bed). There was further complication related to the use in one experiment of a pressure-melting-temperature-activated sliding law; these complications have yet to be directly addressed. The organizers of EISMINT II may have felt apologetic about the lack of benchmark results at the time, but in fact their intercomparison was successful in sparking further research (Payne and Baldwin, 2000; Hindmarsh, 2004, 2006; Saito et al., 2006; Bueler et al., 2007). This research occurred because the intercomparison addressed a problem of sufficient complexity and subtlety (at the time, and perhaps still) so that “success” by any standard was not certain.

The draft ISMIP-HOM report (Pattyn et al., 2008) is not so clear about whether the average of the results from the participants, among the full Stokes results in particular, is or is not a benchmark result. But purpose (iv) is implied by the design of ISMIP-HOM. An intercomparison report should not claim, on the one hand, that no benchmark results were sought or found, while, on the other, claim that the results show the success and agreement of the models.

In any case I am not impressed with the “benchmark results” claim implied in ISMIP-HOM (Pattyn et al., 2008) or stated for EISMINT I (Huybrechts et al., 1996). Certainly in the latter case (Bueler et al., 2005), and I suspect in the case of ISMIP-HOM, exact solutions will appear to address the same issues and actually provide benchmarks. If exact solutions only appear ten years after an intercomparison then they are irrelevant to the publishability or “success” of the intercomparison. If exact solutions are submitted for publication in the next couple of years, or if we find that they already exist in the literature, then this becomes an indication that the organizers of an intercomparison
did not look around for better options.

In the case of ISMIP-HOM an “experiment 0”, done before the others, would have been helpful. Namely an experiment based on an exact solution of the flow line, linear, constant viscosity Stokes problem for some boundary conditions like those of the simpler ISMIP-HOM experiments. An exact solution technique in that case, at least, is completely addressed by “potential fluids” (biharmonic) methods in well-known and classical sources (for example, Ladyzhenskaya, 1963, as well as many other places). The exact solutions in that case are easy enough to write down that one “benchmark result” improvement is immediate: checking future model results against a one-line formula is a lot easier and more precise than checking results against a bunch of pictures in a “supplemental” Cryosphere Discussions supplemental file. This issue alone suggests that simplified geometry intercomparisons should start with exact solutions even if they are inadequate in the modeling sense.

An “experiment 0” exact solution as described for ISMIP-HOM circumstances is not shear-thinning, but surely any Stokes solver for non-Newtonian fluids can be modified easily for constant viscosity. On the other hand there are exact solutions to shear-thinning, power law Stokes flow. The “manufactured solutions” idea (Roache, 2004; Bueler et al., 2005, 2007) can be applied to full power-law Stokes flow, by inserting a divergence-free velocity field into the full Stokes model and determining the body force necessary to make that velocity field a solution. This has been done already in the case of the (Blatter, 1995) higher order model by Glowinski and Rappaz (2003). It is also possible that the technique can be modified to get a field of flow factors, interpretable as a temperature field, and have the body force be the physical one (gravity).

If the goals are “simplified geometry experiments” or “benchmarks”, without the warts of real data, then automatically a significant part of the job is mathematical. Clear thought about the qualities of predictions from the continuum model, and about the particular initial/boundary value problem, is worthwhile even if no one knows what that prediction is. There are professionals for this job.

Organizers of an intercomparison could at least consult with mathematicians on this
task, if nothing else as insurance against possible embarrassment. Imagine the re-
versed situation: Mathematicians Alice and Bob build computer models of glaciers 
based on deeply understood first principles (working only from a well-worn copy of Eu-
clid, say). They want to know if the model results correspond to reality. Should they 
seek the advice of glaciologists on how to do field measurements? Or just wing it? 
Surely it is easy to just throw a GPS on the ice and watch it for a few days . . .

Here is an example of the mathematical attention which could be paid in future in-
tercomparisons. Suppose there is an easier, simplified experiment for which an exact 
solution is known, and a harder “serious” problem of actual interest, for which no exact 
solution is known. The intercomparison could compare the typical size of numerical 
errors for individual numerical models on the easier problem to the average pairwise 
differences between numerical results on the harder problem(s). The choice of norm 
for this comparison would be of some importance, perhaps.

If this were done then the published intercomparison could, if lucky, conclude with an 
assertion of numerical “significance”, analogous to but not as precise as statistical sig-
nificance, for the differences between numerical models based on different continuum 
assumptions. If pairwise differences between models on the harder problem exceed 
the individual model’s numerical errors on the easier problem then one has some ba-
sis for claiming the models have revealed a difference between models comes from 
differences between the underlying continuum models. Conversely, if the differences 
between models are all about the same size as the numerical errors made on the 
easy problem, then this is a useful result, too. It says you should implement whichever 
continuum model is easiest to handle, or make whichever numerical choices are most 
convenient (within the range of possibilities considered by the intercomparison).

Of course there are mathematicians and there are mathematicians. The ice flow 
modeling community is perhaps most familiar with mathematicians whose primary con-
cern is the formulation of continuum models (Hutter, 1983; Fowler, 1997; Blatter, 1995, 
among other sources). Such mathematics is essential. It may be a surprise to the 
reader to learn, however, that the main stream of the mathematics of continuum mod-
els has, in the last fifty years, been devoted to something else. Namely the study of the qualities of predictions and the quantification of behavior of particular continuum models and not the derivation of new models. For instance, far more time has been devoted to the mathematical consequences of linear Navier-Stokes than to formulating variations on Navier-Stokes which might be better models of different fluids. Physicists justifiably complain about this mathematical obsession with a particular case, but such obsessions are exactly appropriate to creating effective simplified geometry intercomparisons and “benchmarks”.

Said another way, there is a difference between the mathematics of small parameter arguments which derive new differential equations (much practiced in ice flow), and the mathematics associated to estimating the size of the solution to a particular differential equation, or the degree of approximatability of solutions of a particular equation. The latter areas of mathematics, more-or-less characterized by use of the phrase “Sobolev space”, must be accepted, even if peripherally, into the world of intercomparison for ice flow. By comparison, standard statistical measurements like “$\chi^2$” are already accepted (MacAyeal et al., 1996). Only by quantitatively relating the smoothness of inputs to the smoothness (and other qualities) of model predictions will the relative difficulty of various ice sheet modeling tasks be addressed. Only by considering problems mathematically posed in their natural spaces will the dependence of results on parameters be clearly understood (Carey et al., 2004, takes steps in this direction). Only mathematics can explain the inevitable, frustrating inconsistencies between simplified geometry intercomparison results. Mathematical clarification of such inconsistencies will allow the

---

4A space of functions with a specified amount of smoothness, along with a method for measuring size which penalizes non-smoothness (the “Sobolev norm”). Sobolev spaces has been common in the last 50 years of study of all sorts of fluids models related to ice flow, including Navier-Stokes equations (Ladyzhenskaya, 1963) and the nonlinear porous medium (Vázquez, 2007, and references therein) among others. Sobolev spaces are essential to determining when the finite element method (Braess, 2007) will succeed. Sobolev spaces were absent from the mathematical ice flow literature until quite recently (Colinge and Rappaz, 1999; Calvo et al., 2002a; Schoof, 2006, for example).
broader community to see what the models say about nature.

3 On comparisons with observations

More broadly than simplified geometry experiments alone, the short history of completed geophysical ice flow intercomparison suggests other possible outcomes:

(a) the intercomparison revealed or explored issues of communal confusion, or

(b) the intercomparison measured model results against observations of real ice flow.

Category (a) is highly desirable. Issues get dealt with out in the open, instead of festering. EISMINT II is a premier example, as described above. MISMIP probably fits in this category, more-or-less by public acknowledgement in advance. I think time will reveal that HEINO also fits in this category, for technical reasons not pursued here.

Category (b) is subject to a well-known caveat which is nonetheless worth repeating (van der Veen, 1999). Namely, if the intercomparison involves measurements against observations then some observations must be set aside for this purpose and not used in initializing models, or as boundary conditions to models, or for tuning the models.

Only one case of category (b) has been published to my knowledge: MacAyeal et al. (1996). Two additional intercomparisons of this type have been completed without publication, the EISMINT-Greenland (Ritz, 1997) and EISMINT-Antarctica (Huybrechts, 1998) intercomparisons.

As an outsider to these intercomparisons, I am impressed by their usefulness to researchers not involved in the intercomparison itself. They involved the posting of public data sets. These intercomparisons have, thereby, set a standard for the minimum quality of a model for the system under consideration. This standard is effective because the model inputs are public and (hopefully) permanent. Modelers can test whether new models reach a minimum level of quality independent of the availability of improved data, which always makes models work better.
The public-ness of the data sets is critical. It breaks the stranglehold of proprietary-in-practice data sets (though usually generated using public funding). It allows a model developer to grab real data and then be forced to deal with all its warts because, after all, others have already done so with some success. One can hardly be satisfied with new work until it is at least as good as the old intercomparison results, however messy those results were from the point of view of intercomparison. Such has been the impact of these observations-based intercomparisons on the design of the ice sheet model PISM (Bueler et al., 2008), in any case.

Ice shelf intercomparisons already have a gold standard. EISMINT-Ross (MacAyeal et al., 1996) performed quantitative, statistically-based comparison of model results and observations. That is, the model velocity results were compared to observations by measuring differences, by summing the squares of point differences in the usual way. As a result it is also qualitatively clear that ice shelf models based on the shallow shelf approximation (Weis et al., 1999) are doing the intended job, namely computing flow velocity from boundary conditions. The available surface velocity data were sufficiently numerous to leave no question that the tuning of at least one parameter (ice hardness) was appropriate. More sophisticated tuning experiments (inverse modeling) have been performed since EISMINT-Ross, with evident confidence in the effectiveness of the basic diagnostic ice shelf paradigm (Humbert et al., 2005; Rommelaere and MacAyeal, 1997; Larour et al., 2005).

Where is the grounded case for ice sheets? Could we recall Mahaffy (1976) and go back to the Barnes Ice Cap and check that the nonsliding thermomechanically coupled shallow ice approximation (TSIA) can be tuned, using few knobs, to give good agreement with observed surface velocities, or that it does not? Will there forever be the suspicion that the TSIA is good for nothing on earth, but never a test to establish, more firmly, one way or the other?

Of course, for grounded ice the critical lower mechanical boundary condition is not itself observed. Furthermore, major ice sheets have some sliding base regions. Supposing that intercomparison is a good approach at all, is intercomparison of inverse
models needed more than for forward models? If so, what do you measure, against what observations, if the observed surface velocities have already been used for the inverse modeling?

4 On language

So far I have avoided the use of the words “verify” and “validate”. Ink has been spilled, though so far no blood to my knowledge, on the meaning of these terms. Nonetheless “verification” and and “validation” are useful terms corresponding to pragmatic, achievable processes of measurement of model outputs. These processes can be parts of intercomparison, and perhaps these terms can be used without causing trouble.

The broader computational fluid dynamics (CFD) community defines “verification” to mean the comparison of results from a numerical approximation to exact, or possibly high-quality numerical, solutions of the same continuum model equations (Pierce, 2004; Roache, 2004; Wesseling, 2001, among many examples). It defines “validation” to mean the comparison of model results to trusted observations of real systems. Here “trusted” means that observations are complete enough to determine the initial and boundary values for the model, and that experimental measurement error for the observations is believed to be small. Thus, the difference between model results and observations reflects on the quality of the model not the quality of the observations.

Note that both terms may apply to a particular numerical code approximating a particular continuum model. There is no automatic association of these terms with intercomparison, but an intercomparison project can use the terms, and the processes to which they refer, as part of the “measurements of differences” advocated here.

A minor hiccup first: Reference van der Veen (1999) reverses these terms, and that reversal is propagated to van der Veen and Payne (2004). Note van der Veen (1999) cites Oreskes et al. (1994) as a source for the terms, though they are not reversed there. This is a minor point, because what is crucial is that two different measurement processes are possible. The particular names are only significant in avoiding commu-
nication errors. The CFD community is sufficiently large so that the ice sheet modeling community should conform, however.

The views in Oreskes et al. (1994) are both well known and extreme. The abstract in that source starts with the sentence “Verification and validation of numerical models of natural systems is impossible,” and the authors proceed accordingly. I see no reason to squeeze practical language, which within CFD describes particular measurement processes applied to model outputs, into a philosophical corner, but that is done in Oreskes et al. (1994). Adding a third term, “confirmation”, helps not at all.

What is perhaps worth extracting from Oreskes et al. (1994), and this point is made in van der Veen (1999) as well, is that there is no purpose in claiming that a nontrivial numerical model for a nontrivial natural system has been either “verified” or “validated”. In fact, I will try to never again say, though I have in the past, that my ice sheet model “has been verified” or “has been validated”. Though processes of verification and validation improve them, models do not tell the truth once these processes are performed. The finality of the present perfect tense leaves a needless false impression.

A model is more trustworthy if there is a record of measuring its output relative to exact solutions, and if there is a parallel record of measuring its output relative to observations of real systems. These measurement processes can be built into ice sheet models and their supporting documentation. These processes can be, and should be, repeated routinely as part of the maintenance of models. These processes of verification and validation can be components of, and complement, intercomparison itself.

References


Hutter, K.: Theoretical Glaciology, D. Reidel, 1983. 404


