After a long description of data and model uncertainties, the authors present a description and example application of a data-model comparison software tool. As detailed below, I find several flaws in the proposed data-model comparison algorithm that need to be addressed. Furthermore, I do not understand why the authors place so much attention on model uncertainty in the text and then fully ignore it in the design of their metrics. If this tool is meant to be used by those doing model calibration against paleo observations, then model uncertainty and downscaling error needs to be explicitly accounted for in the metric. A few of my issues can be addressed by making the package more flexible to handle user choices of metrics (eg implemented by documentation of how to change the metric).

major comments

I also agree with Evan Gowan's suggestion to significantly shorten
the model uncertainty section. When I first read the paper, I
expected all the detailed uncertainty discussion to lead to a detailed
approach to handling model and data uncertainty. Given that these challenging
aspects of model-data comparison are ignored in part (or in whole for model and downscaling uncertainty),
I see no rational for such detailed attention in this paper.

ATAT requires that geochronological data
346 (advance or deglacial) are interpolated onto the same grid projection and resolution as the ice-sheet model
347 before use. Though

Samples for cosmogenic dating in glacial geology contexts are
generally gathered in non-singular quantities for a given local (given
all the uncertainties with inheritance). Consider a grid cell with
two 10Be samples that have very little overlap in their age PDFs. This
cannot be represented by a single Gaussian PDF, so I don't see how
this interpolated single age approach can work for this context
unless non-Gaussian PDF's are permitted (for which there is no indication).

Preparation of the geochronological data to be the same format and grid resolution as the ice sheet model output
352 requires use of a GIS software package such as ESRI ArcMap or QGIS. Users must define deglacial/advance
353 ages based either upon the availability of geochronological data in a cell, or based upon an empirical
354 reconstruction (Figure 4). Where there are no data (i.e. outside the ice-sheet limit), the grid value must be kept at
355 0. Given that this may involve many expert decisions (e.g. which date has the relevant stratigraphic setting,
356 which date(s) are most reliable?), this part of the process is not yet automated within ATAT.

For a deglacial
367 scenario, a model prediction will be unacceptable if the cell is ice-covered after the range of the date error is
368 accounted for, but the cell may become deglaciated any time before this.

what range? one or two or 3 sigma? What if non-Gaussian?

The opposite is true for advance ages; ice can cover a cell any time after the date

and
371 associated error, but cannot cover the cell before the date of the advance.

Does this take into account age uncertainty? If so, again to what
sigma before rejection?

To
373 account for the uneven spatial distribution of dates, a weighting for each date is then calculated based upon their
374 spatial proximity. This weighting is used later when comparing the data to the model output. To calculate this
375 weighting, the Euclidian distance from each dated cell to its nearest dated cell (d_i) is calculated. The mean
376 distance between dated cells (dbar) is then calculated, and the weight of each location (w_i) defined using Eq. (1):

wi=sqrt(di/dbar)

I don't follow the logic of the weighting scheme. Why should the
weight be proportional d_i? What if you have 2 equidistant adjacent
cells with dates? Your weighting scheme assigns the same weight to a
dated grid cell with one adjacent dated grid cell and a dated grid
cell surrounded by equi-distant dated grid cells.

387 geochronological age and modelled age at each location (Figure 4). Firstly, the grid cells which have data are
388 categorised as to whether there is model-data agreement, based on the criteria shown in Figure 3. Since all

If I follow this correctly, the algorithm is outputting a binary agree-disagree result.
If so, this should be changed to give a continuous metric (that can saturate to a large disagree
result when disagreement is well beyond 3 sigma data + downscaling + some model uncertainty.
Continuous metrics are required for efficient sampling/calibration algorithms. You
will likely start with a bunch of "bad" models, and you need to be able to decipher
which are less bad. If my interpretation of the metric is incorrect, then the
description needs to be improved.

Since all
389 dating techniques only record the absence of ice, geochronological data provides only a one-way constraint on
390 palaeo-ice sheet activity. For deglacial ages, deglaciation could occur any time before the geochronological data
391 provided and within the error of the date, but deglaciation must not occur after the error of the date is considered
392 (Figure 3). For advance ages, advance must have happened after the date or within error beforehand, but palaeo393
ice sheet advance cannot occur in the time period before that dated error (Figure 3).

Therefore, ATAT also determines the temporal proximity of the geochronological data and the model
406 prediction. Firstly, a map of the difference between modelled and empirical ages is created (Figure 5).

Equations 2 and 3 don't take into account dating uncertainty, and
are therefore inappropriate.

The comparison should also take into account elevation. If the
modelled contemporaneous ice surface is below the elevation of the
dated sample, then there is datapoint-model consistency even though
ice is present contrary to what the presented algorithm would
indicate. Given the coarse topography near the present-day margin of
Greenland, for instance, elevation needs to be accounted for.

The algorithm also lacks consideration of subgrid/downscaling issues. Eg, for an
ice marginal gridcell on a 25km grid, one would infer the actual
subgrid margin to be somewhere within the gridcell since the next
beyond margin gridcell has 0 mean ice thickness, and therefore 0 ice
throughout. The easiest way to address this is to have a metric that
takes into account proximity of the ice margin as well. This spatial
proximity accounting is also important for model calibration to
extract a continuous measure that can differentiate between two
"bad" models.

different uses. For instance, the percentage of covered dates may prove useful as a first 421 filter of model runs,
422 whilst the wRMSE of dates within error may be more convenient for choosing between filtered model runs

So by this logic, a model that was within 1 sigma of all but 2 data points and in the rejection
region for those 2 datapoints (lets say out of 1000 datapoints) would be worse than a model
that was only within 2 sigma everywhere with no data-points in the temporal rejection region.
????

6.1. General Instructions

I would recommend inclusion of an in-line documented sample run
script (ie that could be executed with a single command). Model
data comparison in generally involve large ensembles, so a script
that could be run in a loop would make this more accessible to
users.

The design of the comparison output needs more thought for use
in ensemble comparisons. A summary file should be generated that
for each line starts with a model run ID and then includes the
summary metric values for that run. The tool should come with
an looping script to cycle over model runs from some file list.

##################
small corrections

A very large source of uncertainty for modelling palaeo-ice sheets is the climate used .. few palaeo-ice sheet models are coupled with 202 climate models
should mention even state-of-the-art GCMs still have relatively
large uncertainties for this context (just need to consider the spread across PMIP 3 submissions)

490 agreement occurs, the RMSE produced are much higher when for the model is compared to the DATED
491 reconstruction.
English is broken

Note that a fuller ensemble model of hundreds
509 to thousands of runs is required for full model evaluation (e.g. Hubbard et al., 2009).
-> of thousands to potentially much more is required...
There is no way the uncertainties in a paleo cycle ice sheet model can be
honestly represented by even a thousand model runs if one is claiming "full model evaluation"