Authors' response to comments on "ATAT 1.0, an Automated Timing Accordance Tool for comparing ice-sheet model output with geochronological data"

Responses in italics.

We thank both reviewers for their comments which have helped focus and clarify the manuscript. We have made changes to both the manuscript and the code having considered these comments. On the code, we have incorporated aspects of model uncertainty (margin position, elevation) into the code, and programmed the code in such a way that it can be called from the command-line and is therefore more suitable for batch processing (e.g. of a large ensemble). Changes were substantial enough that we now call this version of the code 1.1.

Line numbers below refer to the document which includes track changes.

Reviewer 1: Evan Gowan

General comment

Ely et al. present a tool that can be used to evaluate an ice sheet reconstruction or the output of an ice sheet model simulation to chronological data relating to the minimum timing of retreat or maximum timing of advance. I think such a tool is very valuable, and I can see it being useful in my own studies. As stated in the paper, there have been few attempts to directly incorporate individual dates into ice sheet reconstruction evaluation, instead using margin reconstructions such as those by Dyke (2004) and Hughes et al. (2016) for visual comparison. Ely et al. use a statistical approach to evaluate whether or not the area covered by the ice sheet reconstruction is consistent with the chronological information that indicates ice free conditions. As stated in the manuscript, all dates suffer from the "minimum age problem", which is to say that there is an unknown period of time between the retreat of the ice sheet and the age of the material dated (although cosmogenic dates might be close). However, there are few other options to directly evaluate how close the reconstruction is to reality. ATAT is a valuable addition for assessing ice sheet reconstructions that should be used alongside other evaulation methods such as fit to glacial isotatic adjustment indicators.

We are glad that the reviewer sees the value in our tool. On their final point, we have now made it explicit in the text that ATAT should be used in conjunction with other evalulation methods, including GIA (lines 47-48 and 553-554).

1. ATAT software

Unfortunately, I was unable to get the software to work. I tried to follow the instructions for the format of the NetCDF file as per Table 3, but the program would not accept them, specifically with the geochronological data file. I would suggest adding scripts to build this file, or at least give some example NetCDF files so that it is possible to put things in the right format.

We now have an example NetCDF file in the supplementary material for guidance on how to build the geochronological file.

It should be noted that the unit “Years before present" is not a valid CF compliant time unit, and will cause command line NetCDF tools like CDO to complain and not work. I would use CF compliant units to make the NetCDF files compatible with other programs.

The provided example netcdf now uses valid CF compliant units, with a calendar the same as that of the ice-sheet model (in our case years before 1-1-1). We have addressed the calendar issue in the text (Lines 366-367).

There is no recommendation on what to do if there are multiple dates in one grid cell. When I was attempting to use the program, I just took the oldest date without regard of the error, but maybe it is better to make a combined probability using a tool like OxCal (Bronk Ramsey, 2009).

The selection of dates for each cell should be left to expert judgement. The issue of data quality is paramount when choosing a geochronological constraint and requires expert judgement – as explored in Small et al. (2017). We have made it more explicit that such expert judgement is needed for individual cells (lines 369-372), and that future attempts should incorporate Bayesian modelling (as per Chiverrell et al., 2013, lines 379-382). In reality, with a high-resolution ice-sheet model (say 5 km) it is unlikely that two equally reliable dates will be contained within a cell – radiocarbon in a core for example should just use the date that is oldest, closest to the glacial contact.

Are the errors supposed to be 1-sigma or 2-sigma? The paper does not indicate which should be used.

This is up to the user, and will vary for experimental design. 1 or 2 sigma could also depend upon the source of data for a cell – radiocarbon is typically reported as 2 sigma, OSL as 1. We have reported this in the text (372-375).

Also, calibrated radiocarbon dates are not normally distributed, what is the recommendation for usage in this program?

We are using minimum (maximum) constraints for deglaciation (advance), we only look at one side of the distribution (one-tailed constraints as in our Fig. 3). We therefore have an agree/disagree metric that is not dependent upon distribution shape, but rather a user defined acceptable level of error. This is now mentioned in the text (373-375).

Another recommendation I would have is to allow the program to read the required variables (i.e. DEGLACIAL/ADVANCE, filenames, THK/MASK) from a file rather than requiring interactive input. This would greatly streamline usage in scripts where many ice sheet reconstructions are evaluated and plotted automatically.

We now enable users to specify all options at the command line, rather than interactively. Scripts could then be developed to batch process several files. The text has been changed throughout to accommodate this change.

2. Paper

In general, the paper is well written, though I think at times the authors go overboard on detail that is not directly relevant to the tool they are introducing. I think section 2 (background) should be shortened considerably. In the current form, it is almost half of the text. In particular, section2.2 is a two and a half page review of the inadequacies of ice sheet

models. I don't think Geoscience Model Development is really an appropriate venue for such a review, especially since ATAT is not really about fixing these problems. Bringing up these issues here really gives the impression that the authors don't trust ice sheet models at all, which I doubt is the intention. I think anyone doing ice sheet modeling is well aware that it may not be possible to exactly reproduce a configuration that replicates geological observations given the limitations of the models, but they may want to know how close they are!

Though we have reduced the length of this section, we think it is important to review these inadequacies of ice sheet models here. We note that whenever ice-sheet models are demonstrated to non-modellers interested in palaeo-ice sheets, they often question why a specific site or geologically recorded event is not accurately replicated. Though people who conduct ice-sheet modelling are aware of the limitations of models, those in the palaeo-community who are do not conduct ice sheet model experiments (half the audience for this paper) are often unaware of model limitations. We note that we also have a lengthy review of the inadequacies of dating, useful for modellers who may not be so close to this discipline. We also disagree that ATAT, and tools like this, won't fix model these problems. Albeit in an indirect way, such comparisons can help. This rationale was stated in the manuscript (299-307), and has been reiterated in the introduction (lines 79-84).

I think rather than going into such detail on the inadequacies of ice sheet models, it would be more appropriate to detail how ice sheet reconstructions are numerically evaluated at present, such as the extensive Monte Carlo sampling technique used by Lev Tarasov (e.g.Tarasov et al., 2012) and evaluations based purely on glacial isostatic adjustment (e.g Auriac et al., 2016).

We mention the use of GIA modelling in the introduction and have added the Auriac reference (line 59). Tarasov et al. 2012 run ice sheet models that are not independent of the dated chronology (there is a margin raster, their Fig 2, which "nudges" the ice sheet into place based upon Dykes reconstruction). This calibration is different to model evaluation. We have reduced the uncertainty section, but note that without pointing out the inadequacies of ice-sheet models, we think it would be difficult to make a valid comparison.

Section 3 does a nice job of explaining the usage of ATAT.

In section 4 and 5, there is a lot of emphasis that this tool be used with large ensemble of model runs. I don't know if this is a realistic outlook if you want to consider realistic climate scenarios. Computing a specific climate state (e.g LGM) can take weeks, and a fully coupled ice sheet-climate model is along the lines of months. While ice sheet modelling by itself takes a lot less time to run, I question how valid it is to run a large ensemble of model runs using a linear scaling of modern day climate. During glacial periods, the ocean and atmospheric patterns were substantially perturbed, and this has follow-on impacts on the growth and retreat of ice sheets. Maybe such an exercise is useful to get a general feel for the kind of climatic conditions are necessary for glaciation, but I don't think it is diagnostic. The discrepancies between the three model runs presented in this section and the chronological data could very well be due to this issue. It could also be related to using a scaling based on the GRIP record, which may not be representative of the climatic variability in the British Iles during the Weichselian Glaciation. None of these points detract from the utility of ATAT, and I think the focus should be more on evaluating the model results. Perhaps one way to do

this is to run ATAT using the DATED reconstruction and compare it with one of the model runs. This would illustrate what a good fit looks like.

We agree that perturbing modern climate by a distal climate record will not capture all of the necessary climate changes. However, this is still done by some palaeo-ice sheet modellers (e.g. Patton et al. 2016 and 2017, Seguinot et al. 2016) to reconstruct these ice masses. ATAT could be used to decipher how well these models simulate the glacial history of an area.

However, coupled earth-system models are also being developed which will capture oceanic and atmosphere changes. It will be important to evaluate how close to the data these runs achieve, and where improvement is needed.

We have now made it explicit that these 3 model runs are for demonstration purposes only, and our intention is to highlight the utility of ATAT, not to accurately capture climatic conditions over Britain and Ireland through the last deglacial (lines 492-494).

3. Minor comments

Line 57: I would include Auriac et al. (2016) here.

Auriac et al (2016) now included.

Line 301: The sentence here is not complete.

Now fixed.

Line 441: Any reason for using the SPECMAP sea level curve rather than more up to date reconstructions?

No. We just had SPECMAP available. As noted, the aim of these simulations are just to provide a bank of 3 simple experiments to compare to geochronological data.

Figure 7: There is no frame of reference in these maps. I'd suggest putting on modern shorelines to make it easier to see what is going on with the model output.

Coastlines have been added to this figure.

Figure 8: It is very hard to see the location of the geochronological data on these plots. Maybe it would be better to just plot the raw data as points, rather than plotting them as a grid. I also find it a bit confusing to put both the timing of advance to the maximum extent and the Younger Dryas readvances on the same plot. I suggest splitting it up into two panes.

We have plotted the data as points rather than cells for visual clarity. The younger dryas is included to highlight that ATAT only includes the last advance of ice (model could be stopped before younger dryas for a different experiment).

Reviewer 2: Lev Tarasov

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).

We hope to have addressed the flaws in the algorithm, many of which we believe to be miscommunication on our part in the paper and addressed by some rewriting of the model-code. These are outlined below. As stated above, we outline model uncertainty to clarify for non-ice sheet modellers (i.e. those who collect geochronological data). We also think it is important to outline this uncertainty when making a comparison tool.

ATAT now runs from the command line, to better accommodate batch processing. ATAT outputs all metrics, as they are quick to calculate. This output is shown in the updated Figure 5, which now documents the new metrics designed to account for margin position and vertical uncertainty.

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.

We have shortened the length of this discussion, but retain the section as we think that understanding the uncertainty of the model is important when comparing to data. Uncertainty handling will come with ensemble design, the tool asks which ensemble member fits the data best.

We have also changed the code to deal with some downscaling uncertainty, in margin position and ice sheet elevation. Our method for dealing with this is now stated on lines 411-422.

Lines 346-347: 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).

We addressed the issue of which date to choose for a cell in our response to Reviewer 1, and have strengthened our point, that not all dates are equal and this requires expert judgement, in lines 368-371.

It is true that non-gaussian dates occur. Our metrics are based upon whether the model hits a minimum (maximum) constraint in deglaciation (advance), meaning that all geochronological constraints are essentially one tailed depending upon stratigraphic context (see Figure 3). Therefore, the input error is a threshold beyond which model-data agreement does not occur (this is now clarified on lines 371-375. Therefore, if considering a skewed distribution a larger (or smaller) threshold should be defined by the user.

Future adaptations may account for more complex treatments of age probability.

Lines 367-368: what range? one or two or 3 sigma? What if non-Gaussian?

Our response to the issue of non-Gaussian distributions is stated above.

We now state that it is up to the user to define the level of sigma they wish to test (this may be different for radiocarbon, OSL or TCN ages) (lines 372-373).

Lines 370-371: Does this take into account age uncertainty? If so, again to what sigma before rejection?

There may be some miscommunication here, as we use the word error to refer to the uncertainty attached to a date (deliberately done to distinguish from model uncertainty). This is specifically input as a variable into ATAT, and we have clarified how to do this.

Lines 373 and 376 and Eq 1: 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.

We have changed the spatial weighting scheme to apply a search window which defines a local density of dated cells rather than a nearest neighbour distance. This is now outlined in the manuscript on lines 400-401.

Lines 387-388: 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.

ATAT outputs several metrics (these are listed at the bottom of Figure 5, and demonstrated at the bottom of Table 4 which seemed to be missing from our original submission). These include both continuous and non-continuous (agree/disagree) metrics. All metrics are output into a .csv file at the end of the comparison, which is named after the simulation name and whether deglacial or advance dates are being tested. Different users of the tool may want to use different metrics in different combinations. For example, to get rid of extremely poor simulations, it might be worth checking the percentage of sites covered. With better simulations, it may be worth checking the wRMSE of sites within dated error. It is also important keep the agree/disagree metric for the following reason: you may do 100 simulations of a palaeo ice sheet and keep getting the same sites that disagree. This may warrant investigation of the erroneous sites and re-evaluation of the data. This logic is stated in the text lines 299-307 and restated is now restated in the introduction (lines 80-84).

Lines 389-406: Equations 2 and 3 don't take into account dating uncertainty, and are therefore inappropriate.

We apply the RMSE to all dates to indicate how close to the observations the model is i.e. to develop a continuous metric. We also produce a metric which limits to only those data which have passed the original agree/disagree criteria. This is now clarified in the manuscript (lines 457-459).

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.

We agree, and have now included an elevation consideration in the code and in the text. This helps resolve thinning issues for dates on trimlines or possible nunataks. Thank you for suggestion. This is now documented in the manuscript on lines (414-422).

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.

This is a great suggestion and something we had overlooked, thank you. We now include a separate metric that accounts for this uncertainty by applying a perimeter surrounding the originally idenfied margin. This is stated in the manuscript on lines 411-414.

Lines 421-422: 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 ????

Apologies, this is a miscommunication on our part. By "first filter" we meant to identify the worst model runs (e.g. those that do not glaciate over say 50% of the dated sites). We have clarified this in the text (lines 462-466).

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.

We have redesigned the script to be run from the command line. An example of how to execute the script is included in the script header and in the instructions contained in section 6.

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.

A summary output file is produced every time ATAT is run and we have adapted the script to be run from a command line in order that batch processing can be done (e.g. from a shell script).

Small corrections

Line 201: 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)

This is now noted (lines 207-208).

Line 490: English is broken.

Now corrected.

Lines 508-509: 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".

We have rephrased accordingly (lines 553-554).