Response to RC2:

Referee comments in **red**
Author comments in **green**

Summary

This paper presents a new ice thickness and bed map for Svalbard based on a method previously developed by the authors. Here, they use three different methods for small, land-terminating glaciers (IGM), larger and tidewater glaciers (PISM), and surging glaciers (perfect plasticity) to derive their results, having performed a considerable amount of calibration and processing to reduce errors and ensure agreement across their results (no big jumps in ice thickness at ice divides). They then compare their results to other recent bed-thickness datasets for Svalbard, showing that their work sits within the expected range, but substantially reduces errors and bias across the board.

I think this is an innovative paper that attempts to leverage recent developments to obtain the best-possible results. However, I have a few major concerns, as well as several minor ones before I would be happy to recommend the paper for publication. I wonder whether the use of three different methods, particularly when it is not clear to me the rationale for using two different ice-flow models, has not overcomplicated the paper and sacrificed internal consistency – normally an advantage of these large-scale datasets – for an unclear gain in accuracy. The authors establish using PISM where they use IGM would lead to a worse outcome, but not whether the reverse is true, which seems to me a major oversight that makes it difficult to see what advantage using PISM and complexifying the method really brings. I am also very unclear as to how the authors set up their model domains and how they dealt with the resulting boundary conditions, which makes it difficult for me to assess the quality of their modelling. Overall, I therefore think major revisions are required: it may just be a question of adding/clarifying some information that is not obvious as the paper is currently written, which I hope is the case, as I think the end outcome and method are very interesting!

I should also note that I read the paper and wrote the review before I read Reviewer 1's comments. The fact that both of us largely raise the same issues is therefore not a result of groupthink.

Line and page numbers refer to those in the manuscript.

We are very grateful for the detailed and constructive comments, which have helped us greatly to improve the manuscript!

Major Points

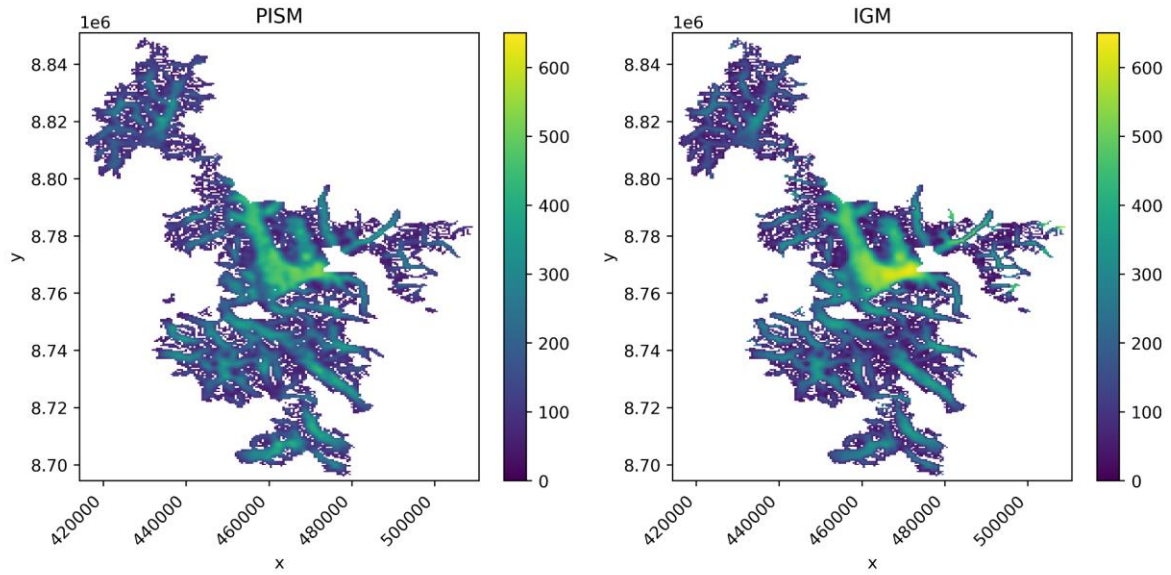
- Consistency: this is perhaps something of a more philosophical point, but a major advantage of these kinds of large-scale studies is that (usually) they apply the same processing steps to a wide area so that, even if one doesn't believe the absolute numbers very much, one can be confident that the results are internally self-consistent. Here, the use of three (very) different methods means this cannot be taken for granted in the same way. I think the authors have done a lot of work

to try to overcome this, particularly with interpolating results onto different grids, but I wonder if this was the right approach.

- Choice of methods: I do not think the authors really provide a clear justification for why they chose to use PISM for the larger (and tidewater) glaciers, as opposed to IGM. I take the point that PISM in SIA+SSA mode does a good job for larger and tidewater glaciers, but IGM is a higher-order model that would also do a good job here. Usually, this choice would be justified by saying that higher order models are too computationally expensive to make them practical at this scale, but IGM has been explicitly designed to run fast and overcome that objection. So what motivated the choice? Because, if the authors had used IGM for all the larger and tidewater glaciers, that would have allowed them to achieve a much larger degree of consistency in the results and avoided them a considerable degree of work at the calibration, method and post-processing stage, it seems to me (of course, they would have had to find a strategy to vary the sliding parameter in IGM, but that feels to me an easier thing to do than have two separate methods working on different assumptions at different resolution)

These are very good points that we have given a lot of thought before submission, but also in recent weeks. Since similar questions were raised by the other reviewer, we post a copy of that response below.

Ideally all glaciers should be modeled with the best possible physics. Following that logic it would have made sense to model also the large tidewater glaciers using IGM. There are however several reasons for us to choose PISM instead for the large glaciers. PISM is an ice flow model that has previously been successfully applied to model glaciers and ice caps that exhibit both sliding and non-sliding flow (by combining the shallow shelf and shallow ice approximations). IGM has only recently been released and has so far primarily been applied to (small) land-terminating glaciers. Only recently (after we did our experiments with IGM) progress has been made by IGM developers toward modeling tidewater glaciers, e.g. by adding a term that describes the horizontal stress at a calving front (see e.g. the preprint by Jouvet et al. 2024, doi:10.31223/X5T99C). In literature, there is so far only one example of IGM being tested on an synthetic ice shelf (Jouvet and Cordonnier, 2023; doi:10.1017/jog.2023.73), using the SSA approximation (i.e. not higher-order), and no published results exist yet for grounded tidewater glaciers. Main developer of IGM, Guillaume Jouvet, recommended us during a meeting in June this year to perform tests with IGM on tidewater glaciers but was unsure about its performance and potential bugs as no such tests have been performed yet. As we are also curious about how well IGM would do we have now performed a test with IGM for all large glaciers around Kongsfjorden (and beyond) in northwestern Svalbard. To allow for direct comparison with the PISM results, we have, as we did with PISM, set a velocity threshold (25 m a^{-1}) to distinguish areas where the sliding coefficient is locally optimized (>25 m a^{-1}) or where instead a constant viscosity is used (<25 m a^{-1}). Furthermore, exactly the same gridded input datasets (velocity, apparent mass balance, surface height and glacier outlines) are used, and simulations are done at the same model resolution (500 m). The figure below shows a comparison of the thicknesses estimated with PISM and IGM. When comparing both maps with available thickness data, we find a slightly better performance with PISM (RMSE = 86 m) than with IGM (RMSE = 92 m) for these glaciers.



Based on the above we have decided that currently it is not worth it yet to model all glaciers with IGM, although this may change in the near future with ongoing progress with IGM. We hypothesize that the reason for the slightly worse performance of IGM on large glaciers (despite the better underlying physics) is that IGM is a machine learning model which does not produce the exact output that a conventional higher order model would produce. It is notably faster, and may through internal learning come close to the results of a higher-order model, but it is not an exact copy. Jouvet and Cordonnier (2023) further noted that IGM experiences a loss of accuracy with increasing domain size, which is another confirmation that IGMs output is not a replica of regular higher-order model results. Other differences between PISM and IGM are in the numerical implementation of e.g. mass fluxes and boundary conditions, which could potentially explain some of the differences even though we are not well enough introduced into both models to give definite answers.

In general, we are not fully convinced that we should strive for more consistency between the methods for small and large glaciers. We in fact want the methods to be separately optimized for both glacier classes, and it ultimately lowers the thickness errors. What also plays a role here is that there is a physical limit to the degree of detail in the bed that still gives a surface expression (e.g. Gudmundsson et al. 2008). This typically implies that bed features smaller than the ice thickness can not be recovered through inversion. As a result of this we can expect to recover more detailed beds for small glaciers than for large glaciers, and want to modify inversion parameters accordingly.

We have added the following in Sect. 4.2: "*This confirms that the use of IGM for small glaciers leads to better agreement with thickness measurements. One reason may be the higher-order physics behind IGM, which helps to resolve small-scale ice flow and bed features better than with a model like PISM which is based on shallowness assumptions. IGM is under constant development, and to date no extensive tests have been performed yet on grounded tide-water glaciers. Using IGM and the same input datasets and model assumptions as with PISM we performed first tests on a selection of tidewater glaciers in Svalbard showing slightly worse performance (more details in Response to Reviewer 1). This may lie in the machine-learning character of IGM, which can only approximate the results of conventional ice flow models that directly solve the stress equations. It is also*

worth noting that IGM experiences a loss of accuracy with increasing domain size (Jouvet and Cordonnier, 2023), further underscoring that IGMs output is not a replica of regular higher-order model results."

The following was added in Sect. 4.3: "*Arguably, using different ice flow models, spatial resolution, and individual parameter calibration per glacier class some consistency between the methods is lost. However, advantageously we achieve a lower misfit with thickness observations. We further note that there is a limit to the degree of detail in the bed that can be recovered from inversion, which scales with the ice thickness (Gudmundsson et al. 2008). Hence, smaller-scale bed details can theoretically be recovered for smaller (thinner) glaciers than for larger (thicker) glaciers. This supports our use of different resolutions and inverse method calibration for different glacier sizes*."

- Boundary conditions and model domains: I am unclear how the authors defined their model domains. I assume, given they use RGI6.0, that they take each RGI outline and invert it individually? Or do they take all contiguous RGI6.0 outlines and invert them as one entity? In the former case, how do they then deal with ice-ice boundaries, where two different RGI entities are in contact? In both cases, what boundary condition is imposed at the front of tidewater glaciers? As both the PISM and IGM methods involve small forward timesteps, these issues need to be considered. At the very least, a few lines in the discussion about how not considering these likely introduces some local inaccuracies need to be added.

Thanks for this comment, we realize it was not clearly explained in the original manuscript. In PISM, all large glaciers (class 2), actively surging glaciers (class 3) and small glaciers (class 1) that are part of large connected ice systems were modeled in one go. The thickness of surging glaciers, from the perfect plasticity assumption, was held fixed during the simulation, but mass exchange at the ice divides was possible. In IGM, glaciers are generally modeled individually, but glaciers that are connected to other glaciers are modeled in one go. The above approach avoids thickness jumps both between glaciers in classes 2 and 3 in PISM, and between connected glaciers in class 1 modeled with IGM. Finally, to avoid that large ice systems consist of glaciers of class 1 (modeled with IGM) and 2 (modeled with PISM), which would create jumps between them, we decided to not use IGM but rather PISM output for small glaciers within those large ice systems.

At the front of tidewater glaciers, we simply assume all ice to calve off that flows out of the outline. So glaciers cannot advance beyond the outline. Given the positive apparent mass balance of tidewater glaciers and the mass balance correction term (to compensate for mass lost through side boundaries), fronts of tidewater glaciers generally have no tendency to retreat either.

The following is reformulated / added to Sect. 3 (first paragraph):
"*One nuance to the three groups above is that all (small) glaciers in class 1 that are part of / connected to larger ice caps are modeled with PISM. This is to avoid thickness jumps at the ice divides. Furthermore, to avoid thickness jumps within ice caps between PISM-modeled and surging glaciers, experiments with PISM also include the surging glaciers as static entities with thicknesses based on the perfect-plasticity assumption*."

and this is added in Sect 3.1:

"*The positive apparent mass balance for tidewater glaciers together with a positive M_{corr} commonly assure a positive mass flux (i.e. calving / frontal ablation) at the calving front. Hence, calving fronts do not retreat. They do not advance either since all mass that flows out of the outlines defined by the RGI dataset is instantly removed.*"

- Discussion: Ultimately, I think this comes back to my point on the choice of methods above, but I don't find that the discussion does a very good job of highlighting what this study brings to the table and why people should use the bed calculated in this study as opposed to those from other studies. The authors provide plenty of description for how their results compare to other datasets, but mostly do not analyse why these differences occur, making it hard for readers to assess which product is better for their particular application. The fact that it is also unclear as to why the authors made particular methodological choices (see my comment above) then further muddies the waters here. The authors do show that they substantially reduce the error on larger glaciers compared to previous studies and the bias across all glaciers (Table 2), which I would argue is the main selling point of their results in the current formulation of the paper, but this gets a bit lost in the discussion and no mention of it is made in the abstract (there is a partial reference in the conclusion, but only to the error on larger glaciers), making it very easy for readers to lose sight of it completely.

Thanks for this comment. We agree the strengths of our approach could be emphasized more.

First of all, we see the use of dedicated methods for small, large and surging glaciers as a strength of our work, and we hope the changes made in response to the first major comments on consistency and choice of methods has helped to better emphasize this.

Furthermore, in the original manuscript, we did already quantify how well our study as well as Millan et al. (2022) performed against thickness observations. This information was (is) in Table 2 and showed that our study yields comparable results for small glaciers and a marked improvement for large glaciers and surging glaciers compared to Millan et al. Such a comparison is unfortunately not possible with the results of Fürst et al. (2018) who use an approach that more or less imprints the local thickness observations in their final map. An independent thickness observation dataset would be needed to compare performance of our study with the product by Fürst et al. which to date is unfortunately not possible. It is noteworthy though that the Fürst et al. approach can be seen as an "interpolation method" as the observations are imprinted in the map and mass conservation and viscosity tuning are applied to generated thickness in between observations. Our study is less informed by the observations (only to constrain global parameters) which we argue leads to a map that may be more consistent in space (in terms of spatial detail/roughness and uncertainty) and has the advantage that it can be used as a numerically stable spin up state for prognostic modeling. We have added discussion on this to Sect. 4.2:
"*It is noteworthy though that the Fürst et al. products can be seen as an 'interpolation method' as the observations are imprinted in the map and mass conservation and viscosity tuning are applied to generated thickness in between observations. Our study is less*

informed by the observations (only to constrain global parameters) which we argue leads to a map that may be more consistent in space (in terms of spatial detail/roughness and uncertainty) and has the advantage that it can be used as a numerically stable spin up state for prognostic modeling."

Additionally, after related comments by the other reviewer, we have added scatter plots of modeled vs observed ice thickness also for Millan et al. to Figure 7 (panels c and d). These figures show that the larger MAE for glacier classes 2 and 3 in Millan et al. are a result of a general larger spread and more outliers, especially for larger thicknesses. For glaciers of class 1, i.e. the small ones, Millan et al. tends to underestimate large thickness and overestimate small thicknesses, which is a sign that their thickness distributions are too smooth. We have added sentences on this to Sect. 4.2:
"*Similar scatter plots comparing thicknesses by Millan et al. (2022) with observations (Fig. 7c-d) show that the larger errors for glaciers in classes 2 and 3 (Table 2) are a result of a general larger spread in the Millan et al. (2022) dataset, primarily for large thicknesses. For the small glaciers (class 1) Millan et al. (2022) show an underestimation of large thicknesses and an overestimation of small thicknesses, indicating that the Millan et al. (2022) thickness product is smoother than reality.*"

Minor Points

- p.1, l.10: I might venture to say that a mean ice thickness at the scale of the whole Svalbard archipelago isn't that useful or meaningful a number to include in the abstract (at least, as a headline figure for the paper, it seems some way down the list of things that readers would want to know)? The total volume, yes, but I would suggest maybe converting that into an SLR equivalent for the second number, or reporting the maximum ice thickness, which is something that makes a bit more sense at that scale. Or, possibly even more useful, say something about how the volume estimate presented here compares to other studies' estimates.

Thanks for this comment. We have now removed the mean thickness and instead added the sea level equivalent to the abstract. Furthermore, the sea level equivalent estimate has been added to Sect. 4.1.

- p.2, l.42: Reference formatting for Farinotti et al.

This is now corrected.

- Table 1: Why use the 20 m NPI DEM when it has to be downscaled to 100 or 500 m immediately? Wouldn't the COP90 DEM have been a better choice to fit with the modelling resolutions and also sit more in the middle of the range of most of the other data (2010-2019ish)? Also, the RGI6.0 outlines for Svalbard have dates of 2000-2010, so please update the table to reflect that.

Good point, although most of the outlines are from 2007-2008, there are some outlines from other years (e.g. 2001). We have changed the period in Table 1 to 2000-2010 as suggested. We have chosen the NPI DEM over global DEMs such as the COP90 DEM, since it is a

dedicated product for Svalbard that incorporates a large amount of recent and older data from aerial photography (https://doi.org/10.21334/npolar.2014.dce53a47). Furthermore, the error (standard deviation) of the NPI DEM has been quantified (2-5 m, possibly slightly larger on glaciers), whereas such information for Svalvard is, to our best knowledge, missing for e.g. the COP90 DEM. Finally, in response to the other reviewer we have extended the input data description in Sect. 2.

- Figure 1: I can see already that, on Austfonna, there are two methods being used to generate the results, despite the assertion in lines 116-117 that all the glaciers connected to larger ice caps are modelled using PISM (i.e. the same method). Can the authors confirm whether the figure or the text is correct here? If the figure is right, how are they dealing with the jumps in thickness at the ice divides on Austfonna?

See also our earlier reply. Our statement in the original manuscript that PISM is used for all glaciers that are part of larger ice systems was incorrect. Surging glaciers that are part of large ice systems were modeled with the perfect plasticity assumption. The surging glaciers were included in the PISM model runs (as static entities) to provide boundary conditions at ice divides for the modeling of the non-surging glaciers.

- p.7, l.169: ‘with a’

Corrected.

- p.9, l.200-204: Can the authors comment as to how far using uniform parameter values might introduce some error into the results?

Using uniform parameters here is foremost a practical choice as a result of 1) the observational datasets of ice velocity being of too low quality for slow-flowing mountain glaciers to deliver a reliable signal that could be used for a spatially variable sliding coefficient inversion as in our PISM approach; 2) the sample size of glaciers in class 1 with observations not being big enough to deduce any spatial/climatic patterns of A and c that could be extrapolated to unsurveyed glaciers. Not the least, the complex poly-thermal nature of many Svalbardian glaciers is a complicating factor. As such, the calibrated values for A and c are on average the best fitting ones, but naturally there may be glaciers with specific local conditions where they are not ideal. However, we do not see a feasible way of systematically classifying where this would be the case. Consequently we are left to conclude that any errors resulting from the spatially homogenous A and c values likely are included in the overall uncertainty (Table 2), given that the calibration glaciers form a fairly representative sample of all glaciers.

- p.9, l.214: I confess I’m not entirely clear on why the thickness field produced by IGM would have gaps in it that need interpolating?

We correct for the mass leaking out of the glacier domain with a spatially uniform adjustment of the specific mass balance. However, also some part of this mass addition can be leaking out, which then means that (usually small) parts of the domain remain ice free. These holes we interpolate in the end. This reasoning is also described in Frank and van Pelt (2024) which we cite. To clarify, we have rephrased line 214 as follows: “*The final thickness field is*

obtained by interpolating gaps in the modeled thicknesses which may remain in the case of persistent mass leaking and applying a thickness-dependent Gaussian filter as in Frank and van Pelt (2024).”

- p.10, l.231-232: This seems to me quite a substantial upsampling of the majority of the dataset that might introduce a considerable number of artefacts. Would not downsampling the 100 m proportion of the dataset to 500 m have been the more conservative choice?

Thanks for bringing this up. By using nearest neighbor interpolation when reprojecting the 500-m results to a 100-m resolution grid, we do not add any detail to the bed (it will look exactly the same on a 100 and 500 m grid). Furthermore, we prefer to keep the 100-m results at their original resolution so that no detail is lost there. We would also like to repeat that we believe it is justified to have finer resolution bed topography for small (thin) glaciers than for large (thick) glaciers (see our response to the first major comments).

- p.10, l.245-257: Yes, but are there observed glaciers in each of the three categories, such that all three types of inversion are bias-free? More generally, with three different inversion methods, would there not need to be three separate estimates of sigma H bar, one for each category? Because a mean error of 3.5 m on ice thickness across the whole of Svalbard seems a little too good to be true. The observations themselves would have bigger errors than that!

We would like to highlight that the uncertainty estimate (3.5 m) applies to the *mean ice thickness* for all of Svalbard (205 m). It is the mean thickness and its error that are relevant for the ice volume calculation. The fact that the mean thickness bias (i.e. volume error) is zero for 169 glaciers across Svalbard greatly reduces the volume uncertainty for all ice in Svalbard. The volume uncertainty would have been markedly higher when fewer glaciers were observed. Furthermore, the thickness error at a random location in Svalbard is much larger, e.g. in Table 2 it can be found that it is 75.5 m for glaciers in class 2) and 3) and 50.1 m for glaciers in class 1. In other words, local errors that occur at individual sites in Svalbard to a large extent balance / average out at the Svalbard-wide scale. The following has been added to Sect. 3.4:
“*Please note that the relative error of the volume and mean thickness is much smaller than the local (point) uncertainty of modeled thicknesses (the latter is quantified in Sect. 4.2).*

And in Sect. 4.3:

“*The large and well-distributed thickness observations dataset available for Svalbard used for model calibration, including data from 169 glaciers, helped to reduce the Svalbard-wide volume uncertainty (estimated at 3.5 %). Whereas the RMSE of Svalbard mean glacier thickness is only 3.5 m as a result of averaging and calibration, the local (point) thickness error is considerably larger (50.1 m for class 1 and 75.5 m for class 2 and 3, Tab. 2).*”

Furthermore, we argue that splitting the glaciers into separate categories for the uncertainty assessment would only complicate the uncertainty assessment and is unlikely to lead to a very different error estimate for the mean thickness. The error estimate would only differ significantly if the relative area fraction of observed glaciers in a certain class differ from the relative area fraction of the same class for all glaciers in Svalbard. Since observations are

well-spread over Svalbard and include glaciers of all types, we assume this effect would be small. Please also note that rather than splitting between glacier classes we could also group the glaciers in other categories with individual errors (e.g. creating categories of thin and thick glaciers, or tide-water and non-tidewater glaciers, or to split Svalbard in regions). All would give a slightly different Svalbard-wide error estimate (sometimes higher, sometimes lower) than when lumping all observed glaciers together.

To better clarify the uncertainty assessment strategy we have reformulated some sentences in Sect. 3.4:
"*The range of biases narrows if we select more than one glacier for calibrating the model, and, following the same logic as is used to calculate a standard error of a mean, it can be found that dividing by the square-root of the number of samples is required to calculate the remaining standard deviation for larger sets of glaciers used for calibration."*

- p.16, l.314-319: Have the authors performed the same comparison in the other direction? As in, what happens if IGM is used to model the larger glaciers where PISM was the preferred method? Otherwise, I think it's difficult to say that the combination of the two methods is superior to either alone. The approximations in PISM may be suitable on the larger glaciers, but it doesn't follow that that means they're more suitable than using a higher-order model.

Thanks for this comment. We refer to our response to the first major comment.

- p.16, l.321: OK, yes, 6855 is higher than 6800 and 207 is higher than 205, but I'm not sure that it's really a meaningful difference, especially when both those numbers are well within this study's own error bars. Consider rephrasing this to make it clearer that this study's integrated volume and mean thickness results are not significantly different to those from Millan et al.

We agree, this is now corrected.

- p.16, l.320-333: Could the authors provide some more analysis of why these differences exist? They posit sensible reasons for why Millan et al. likely overestimate ice thickness on larger glaciers, but I think it would make the paper much more useful for the community if they can suggest some reasons for the other differences (spatial distribution more similar to Fürst, thicker ice at lower elevations than Fürst, less pronounced jumps at ice divides than Millan, etc.), as it would help people work out which is the best bed product for them to use for their particular application

We refer to our response to the last major comment above.

- p.18, l.384: This is only true provided other people use the exact same set of final modelled bed, velocity, surface, etc. as used in this study as their initial conditions. If someone took the bed from this study and then used, say, the COP90 surface DEM and ITSLive velocities to initialise their model, they would not have a harmonious set of initial conditions. Please rephrase this to make it more clear.

Thanks for pointing this out. We have rephrased the sentence to make it clear that it is general benefit of this type of inverse methods (i.e. iterative ones) that the reconstructed beds are a starting point for potential future runs when using the same model, setup and compatible input datasets:
"*A benefit of thickness maps produced with iterative inverse methods, i.e. for all not actively surging glaciers, is that they simultaneously provide initial conditions for future simulation of the same set of glaciers. However, this does require the use of the same ice flow model, setup, and temporal consistency of input datasets.*"