Reviewer 1

The authors fit third order models, representing saddle-node bifurcations, to model data of three tipping elements; the Greenland Ice Sheet (GrIS), West Antarctic Ice Sheet (WAIS) and Atlantic Meridional Overturning Circulation (AMOC). The aim of this approach is to develop simple models representing tipping dynamics that allow for fast simulations, making them suitable for decision making.

The aim of the paper is worthwhile, however the current state does not go far beyond fitting a third order polynomial to model data. I have several suggestions/comments which I believe will strengthen this work and make it more valuable for the tipping community.

Thank you for the summary and seeing the value of our contribution. In the following we will answer to all of your comments (blue font).

Main comments:

- Uncertainty. Currently no uncertainty of the emulator-models is given, while uncertainty is key to know how much the results can be trusted. Hence, I think it is key to address the uncertainty question here, and not postpone it to later work. This does not need to be extensive, but a first idea of the reliability of the fit would be very valuable.

Thank you for this valuable comment. We agree that uncertainty should be addressed already in the present manuscript. In the revised version, we will therefore include a first uncertainty quantification of the emulator fits. Specifically, we will provide a first estimate of emulator uncertainty by approximating the parameter uncertainty locally from the least-squares Jacobian at the optimum and sampling feasible parameter sets around the best fit. Propagating these samples through the reduced model yields uncertainty ranges for the fitted equilibrium branches and the inferred tipping thresholds. Together with the selection of several different models for, which implicitly accounts for structural uncertainty, we believe that this gives a good first estimate of the uncertainty.

- Verification. The authors run transient simulations with their fitted models, but do not show any comparison with transient runs from the models to verify how well their framework works. As transient runs are available for at least some models (and used to determine the timescale), it would be valuable to compare the fitted models to those. This would also give insight in the reliability. Something to consider here is also testing the models' reliability out of sample, as e.g. for the ice sheets you fit it to data up to 3-4 degrees, but do transient simulations up to 6 degrees. What is the reliability of those results?

Thank you for this comment, in the revised manuscript, we will compare our fitted models with transient out-of-sample simulations, where available. For PISM and Yelmo, there are

transient (and overshoot) simulations available, that we will use for comparison and verification of our fitted model.

- Modular. In the abstract the authors state their approach is modular. However, this is not discussed in some detail till the discussion, and I am uncertain how exactly it works. As it is a main claim, it needs a improved arguments and discussion.

As mentioned by the second reviewer, we might not have made clear enough what we mean by modularity in the manuscript. By modularity, we mean the possibility to easily extend and include new model simulations or even new tipping elements. We will change it accordingly in the manuscript and, in the discussion, we also outline for which tipping elements such new simulations would be particularly beneficial, i.e. the WAIS since there is only one model that performed the necessary simulations.

Specific comments:

- I am not convinced by the statement that the method is process-informed. It is fit to physical models, but the model that is fit is not necessarily informed by any physical constraints. For example, in the transient simulations for GrIS the final volume of the ice sheet depends on the level of global warming (i.e. the warmer, the more ice loss), while I would say at a certain point all ice has disappeared, meaning it should no longer depend on the level of warming. Using the current approach this will never be the case (due to the chosen 3rd order model).

We agree that our method is not explicitly process-informed but only implicitly, i.e., by fitting to the existing comprehensive model results that include such processes. We will clarify this in the manuscript accordingly. Nevertheless, we consider this an important step forward compared to previous studies that relied on more heuristic third-order models.

However, the GrIS is indeed dependent on the global warming level. If we look at the ice-sheet model simulations (e.g. Fig.1 https://www.nature.com/articles/s41586-023-06503-9/figures/1), the final ice volume is indeed dependent on the level of global warming even in the comprehensive models. Of course, after passing long enough time at a high enough warming level all ice has disappeared, which we also find in our framework similar to the original simulations in https://www.nature.com/articles/s41586-023-06503-9. In addition, we cap the ice volume at 0m sea-level equivalent volume, because as you mentioned above, the third-order model would otherwise allow for unphysical negative ice volumes. We did not mention this clearly in the manuscript but will make it clear in the revised version.

- Introduction. There are a number of studies looking at abrupt change in climate models which are relevant, but not referenced. Drijfhout et al. (2015), Terpstra et al. (2025), Harteg et al. (preprint, TOAD), Angevaare and Drijfhout (preprint).

Thank you, we will add the according references.

> Methods. The reference simulations, timescale function and calibration all need to be explained better. Currently it is hard to know how exactly the fit is done and which choices are made and why. Similarly for the rescaling. I understand it is for computational reasons, but I would suggest to give some more detail on how it's done in the appendix and how the values to which is rescaled are chosen (order one, but somehow they differ between tipping elements). Specifically:
>
> 1. L111-114 are not clear to me.
>
> 2. Eqn. (2) – suggest making the theta dependence on the rhs explicit.
>
> 3. L121 – Shouldn't it be z_data instead of x_data? (appreciate it's the same, but it reads confusing).

We completely agree and will fix the method section to make it clearer.

In L111-114, we want to say that we model the WAIS and GrIS (SICOPOLIS) as two distinct dynamical systems due to the clear two step response in these simulations instead of modelling it as one higher order dynamical system (e.g. x^5 system). Yes, it should be z_data indeed, thank you!

- Figures. A lot of figures have a normalised state (a.u.) on the y-axis, making them hard to interpret. For example for the AMOC it is not till the end that it is mentioned that Veros and POP have very different AMOC background states, which is key for interpreting the results. This needs to be mentioned earlier, as otherwise the transient simulations are hard to interpret. Changing the figures to physical units can help with this.

We always include the physical units as well on the y-axis (right y-axis) but we agree that this could be made more explicit by putting the physical units on the primary y-axis. We will also mention in the manuscript the considerable difference in AMOC strength in the different models, which is indeed an uncertainty of the underlying Earth System Models that our approach in this work can take into account explicitly.

- By design the tipping thresholds match those of existing literature, as you also mention. So I would not mention that explicitly as a result.

Agreed.

Suggestions to consider for the discussion:

1. Is there the possibility to use observational constraints in this framework?

2. You currently have a fit for each climate model. Is there a way to combine these, reflecting the trust we have in them?

3. Do you expect your emulator to respond realistically to overshoot scenerios?

We will consider your comments in the revised version. Specifically:

1) It is indeed interesting to consider how observations could be used within this framework. In principle, observational constraints can be incorporated, although we do not yet implement this in the present manuscript. One possible approach would be to use observations to assign differential weights to the comprehensive models before combining their emulator results. For example, the weights could be based on how well each comprehensive model reproduces relevant observed features, such as the present-day mean state, recent trends, regional patterns, or other diagnostics. The resulting weighted ensemble would provide a combined emulator estimate while retaining information about structural spread across models. We note, however, that this strategy is only meaningful when several comprehensive models are available for the same tipping element, as is the case for the AMOC and the GrIS, and when observations are sufficiently informative for model evaluation.

2) Connecting to 1), the simplest way to combine the fits from the different comprehensive models would be to weight them according to our level of confidence in each model and then run simulations accordingly. However, it is difficult to quantify this confidence, or lack thereof, in a rigorous way. We will present results from such an approach in the revised manuscript.

3) Regarding overshoot scenarios, we expect the emulator to provide a useful first-order approximation of the response to overshoot forcing. At the same time, we do not want to overstate its realism. Because the emulator is low-dimensional and deterministic, it is not expected to capture all transient details of the comprehensive models, especially under strongly time-dependent forcing. To assess this more directly, we will compare the emulator with overshoot trajectories from the comprehensive models in the revised manuscript.