Reviewer 2

This paper presents a method that fits 1D dynamical systems (double-well potentials with variable control parameter) to data previously obtained from ESM simulations of tipping points in various sub-systems of the climate. Using the obtained models can be useful for conceptual studies of tipping elements, given that the emulated dynamics are reliable when extrapolated beyond the ESM data. But I have major reservations with respect to the motivation and soundness of the mathematical form of the dynamical system used, as well as the overall presentation of the paper.

We thank the reviewer for their comments and, in response to their feedback, have majorly reworked our manuscript as we outline below. We hope that our revised version will help to resolve the identified shortcomings. In the following we will answer to all of your comments (blue font).

Major comments

1.

When this was introduced as a “framework”, I was expecting to

a) get uncertainty estimates on the model parameters

b) be able to synthesize information from different ESMs of the same tipping element into one model emulator

This is unfortunately not possible, and instead the method is a particular way of fitting one dynamical system (limited to a 1D fold-fold bifurcation third-order polynomial ODE) for each set of simulations in a given ESM. I think the limitations of the framework need to be highlighted better, and it needs to be explained better how the collection of obtained models can then nevertheless be used to make “decision-relevant” studies of climate tipping risks.

Thank you for this valuable comment. We will highlight the limitations of our framework better in the revised version.

In the revised version, we will also include a first uncertainty quantification of the emulator fits, which we found was possible after considering both, the comments from reviewer 2 and reviewer 1.

We also agree that the current manuscript does not yet synthesize information from several comprehensive models of the same tipping element into a single emulator. As a first step in this direction, we combine the model-specific emulator fits in a weighted ensemble. The simplest approach is to assign weights to the individual comprehensive-

model calibrations and propagate these weights through the reduced-model simulations. In practice, however, quantifying and determining such weights is challenging, since it requires some measure of model performance or credibility, particularly as the confidence of the underlying Earth system models has not yet been quantified. Instead, what our approach can offer is to explicitly propagate the uncertainty of different tipping element realizations in different models.

In addition, we will present an initial version of such a multi-model analysis in the revised manuscript and discuss its interpretation and limitations explicitly.

2.

Section 3.1.1, where as part of the method the "timescale" tau is estimated as function of the control parameter (before the actual fitting, if I understand correctly), is in my view misinformed. The purpose seems to be to account for critical slowing down, which gives a power-law divergence of the relaxation time towards equilibrium as a saddle-node bifurcation is approached.

But the form of Eq. 1 using a constant tau already gives rise to a saddle-node bifurcation, and thus will feature critical slowing down in itself. So I don't see a basis for introducing this additional, duplicate critical slowing down into the equations. Maybe the fitting procedure still works (at least introducing a monotonic tau(p) does not induce additional equilibria), but conceptually it is hard to justify. I find this a major issue, since this timescale function is the only major enhancement with regards to previous work, which simply fits Eq. 1 to data while having a constant tau.

I guess that by using such a simple 1D system, the time scale that could be estimated from fluctuations around equilibrium may yield an emulated "timescale" of the eventual collapse (which is then far from equilibrium) that is far off the actual timescale in the ESM. So I understand that there is a motivation to estimate a constant characteristic time scale using other, transient simulations. That being said, it seems that the actual time scale of the model dynamics (i.e. scaling of parameters [a, c, e]) is afterwards modified anyhow in an unknown way by fitting the polynomial parameters from the equilibrium simulations.

We thank the reviewer to pointing this important issue out. First, we want to clarify that in a first step, we always fit Eq. 1 to the equilibrium runs of the comprehensive models without the parameter τ. Currently, this is not clear from the text and we will change it accordingly in the revised version. In a subsequent step, we determined the function $\tau(p, t)$ by using the transient simulations. So, the parameters a,c,e are *not* modified by the fit of the time scale.

The motivation of using this function of τ was not to account for the critical slowing down per-se, since as you say, the saddle-node bifurcation gives already rise to critical slowing down even with a constant τ. Rather, we intended to account for the transient behavior of the system after crossing the tipping point. Specifically, Fig. A1 shows the effect we

wanted to account for, that is not captured by a constant τ across the whole forcing regime. There, we fit a scalar τ to the different transient time series of the GrIS (Yelmo) after crossing the threshold. A clear dependency of τ on the distance to the critical warming threshold is visible (Fig. A1 i). By assuming a constant τ, the transient response of the Yelmo emulator does not match the comprehensive model's response as well, in most scenarios we compared. However, your concern about accounting for the critical slowing down twice is valid, since the functional form of τ is diverging as p->p_c. To see if the complex time scale relationship holds up, we tested this relationship against the transient PISM runs (past the threshold) and, while there seems to be a dependency of τ on the temperature, it is not as clear as for the Yelmo runs. In addition, we find that transient runs, where p≈p_c for longer times lead to unphysical responses of the system, compared to the comprehensive models. Therefore, we decided to go forward with a constant τ and not with the more complex function τ(p,t).

In summary, we agree that we had mixed up these two different concepts too much in our original methods section. Accordingly, we will revise the section to our comments above.

3.

In many places, it is mentioned that the framework is "modular", "naturally designed for efficient updating", "modular update protocol" etc. But the only justification for why this is the case is that the parameter e (constant offset of the bifurcation parameter) can be changed without refitting the model, in the case where new ESM simulations come along that reveal a different critical threshold. First, I don't see why this is even worth mentioning. Anyone who chooses to do simulations with conceptual models can decide to redefine their control parameter as (p-p0) if they think the critical threshold is in one place or the other, so in what way is this unique to this method? Besides, one would not even need to make new simulations with changed value of e, but can simply relabel any plots with dependence on p → p + e. Second, are there really realistic scenarios where new data from an ESM becomes available and everything about the system stays the same (the exact entire shape of the bifurcation diagram), but the critical threshold is shifted? In my view it would be much better to do actually do a refitting, rather then to postulate the dynamics are the same while only the scale of the bifurcation parameter changed. And furthermore, why would it be a big deal to refit the model with new data? Is the fitting computationally intensive? Finally, how does a shifting of p by e affect the choice of the function tau(p-p_c)? Should Eq. 3 (or the parameter p_c) not also include a dependence on e?

Thank you for this thoughtful comment. We agree that our wording might have overstated the novelty and importance of the "updatable/modular" aspect. By modularity, we mean the possibility to easily extend and include new model simulations or even new tipping elements by either 'redefining' the control parameter or refitting. Even in the case of refitting, it is easy and straightforward to include new simulations or even tipping elements since the fitting procedure is general.

We agree that the possibility of shifting the effective control parameter by adjusting the offset parameter e in Eq. (1) is not unique to our framework. As you note, this is mathematically equivalent to a relabeling of the control parameter. We also agree that, in most realistic situations, new simulations may change not only the threshold location but also the branch geometry, hysteresis width, or transient behavior. In such cases, a full refit is the appropriate procedure and should be preferred (which is computationally not expensive).

Regarding the redefinition of the parameter e, then of course the inferred p_c changes so there is no need to change anything in τ(p-p_c). So yes, there is a dependence of p_c on e (and all other parameters).

We will tone down the language as well as make clearer what we mean.

4.

I am not convinced the authors have demonstrated that new scientific results can be achieved with the method, even though it is claimed that it is “decision-relevant” and “physically grounded”. The proposed method simply fits 1-d dynamical systems to ESM data and then hints to using the collection of models (although it remains unclear how to synthesize different models of the same tipping element) to estimate the thresholds (e.g. Sec 4.1) and the time scales of partial or full collapse (e.g. WAIS, Sec. 4.2). This is problematic, since

a) the fit is often simply not good (e.g. SICOPOLIS model), and so it seems to be more robust to simply look at the ESM data to argue for the location of the tipping point, rather then considering the extrapolation of the 1D model, which has visibly clear discrepancies with the ESM data.

b) the estimate of timescales relies on the rationale for the functional dependence of the time scale, which is unclear, see comment #2.

As hinted at in the discussion, I guess that they might use the models in the future for more conceptual studies of coupled tipping elements. But here it would be unclear which of the several 1d models for a given tipping element one should choose. Overall, I would have been more convinced if the authors simply use the method and demonstrate that new scientific results can be obtained.

We thank the reviewer for this substantial comment. Originally, this manuscript was intended to introduce the framework rather than its application. However, we agree with the reviewer that such applications would be very helpful to be able to inspect the usefulness of our approach. Therefore, we will include applications in the revised manuscript such as overshoot scenarios of the GrIS and compare them to the comprehensive model results.

We will also show how the different fits for different models can be combined. Specifically, how one can get a single estimate of the response of the tipping elements for different forcing levels by combining the model-specific emulators.

However, we do not agree with the statement that *the fit is often not good*. For SICPOLIS, we agree that the fit is not perfect, especially the reverse branches. This is especially due to the fact that the reverse branch in SICOPOLIS starting from the intermediate state shows a relative linear regrowth in response to the reduction of the temperature (yellow line Fig. 1 in Höning et al. 2023), which cannot be captured very well by our model. However, in the revised manuscript, we now include the reverse branch in the fitting directly, which was not the case in the previous version.

But for the other models we would argue that the fits are generally reasonably good (Fig. 2, 5, 7, 8). Regarding the time scale, we agree and refer to the other answers (e.g., your comment 2).

5.

I find the formulation of the abstract a bit misleading:

They contrast their approach with ones that “rely on idealized system dynamics” and which “do not take into account Earth system model processes”. At the same time it is said that the method fits “saddle-node bifurcations” to ESM model output, stating this makes it process-informed. I find this a bit misleading, since their approach indeed heavily relies on idealized dynamics by shifting and stretching the predefined bifurcation diagram of a one-dimensional system. The functional form of this system is given and not derived from any underlying processes, at least if with processes one refers to physical processes related to the climate system.

Similarly, I would not call the models “low-order” (as already in the title), but “conceptual”. At least in my understanding, low-order is usually used for models of several ODEs that have been formally derived by approximation of the underlying PDEs of, for instance, the fluid dynamics. But I might be wrong here.

They go on to say that the “emulators reproduce multistability of the GrIS and WAIS, …”, but this is the case by definition of the choice of their one-dimensional model.

In general, the language in the paper sounds at times not scientifically precise enough, and rather too much like a sales pitch. Examples are: “process-informed”, “decision-relevant”, “element-specific”, “reduced-complexity”, “domain-specific simulations”, “high-fidelity experiments”, “transparent low-order dynamics”, “traceable”.

Some of these descriptors can of course simply be rephrased and spelled out explicitly in their meaning. For others it can perhaps be motivated more explicitly why they are apt. But

many of these descriptors do not really hold up to scrutiny in my opinion, see my other comments.

Thank you for this helpful comment. We agree that the original wording overstated the extent to which the emulator is "process-informed" and may have suggested that the reduced model is derived from underlying physical processes. We also agree that the language was not scientifically sound enough in some parts. We will revise the language in the manuscript according to your suggestions. We also agree that the term “lower-order” models might have been confusing here and will stick to the term “conceptual” models instead. Regarding the statement that the emulators "reproduce multistability", we agree that this should be phrased more carefully since the chosen system is multistable by construction.

Minor comments:

2 of 3 example systems have the double two-fold, i.e., tri-stability. Why not attempt to fit a higher-degree polynomial model to it? Then one would not need to postulate that each system with tri-stability consists of two completely independent sub-systems. This postulate is not justified in general, and limits the emulated dynamics as well as the method on the whole.

We actually did try to fit a higher-order polynomial first to the WAIS and SICPOLIS simulations, however decided against it ultimately. The reason for our decision was that the uncertainty of the fit of all three stable branches with higher order polynomials (up to 7th degree) increases drastically. The separated lower order polynomial fits were always advantageous for the fit. Therefore, we decided to use two subsystems for the WAIS and SICOPOLIS.

The discussion does not relate enough to the actual content of the work, i.e., the underlying assumptions as well as benefits with respect to prior work. It is mostly a general discussion on AMOC, WAIS, etc. which are not really informed by this work (e.g. p12 first paragraph).

We agree and will change that in the revised version.

The fitting routine needs to be explained better. From the notation it does not become clear whether all data, including at different control parameter values, is fitted/optimized at the same time. It is also not clear to me from the text whether the time scale function is determined before or after doing the main fit (I am assuming before).

We fully agree and we will revise the whole method section. As mentioned above, the time scale function is fitted after doing the main fit, not before. We revised the fitting procedure in the manuscript. Specifically, we minimize the residuals between the simulated

equilibrium points and the corresponding equilibrium branches of the reduced model. We describe the fitting procedure in the revised manuscript in more detail now.

L43:

What are “highly idealized tipping frameworks”, and in what way does the present paper go beyond that? It seems the authors argue their method falls into the class of “tipping-element emulators”, and thereby bridges the gap toward fully coupled ESMs. If the authors believe that fitting parameters of the double-well potential (or other conceptual systems) to ESM data makes it an emulator, it would be good if they explicitly state this, and at the same time state what type of approach would fall even lower on the modeling hierarchy. I guess this would be using the same type of model, but without any calibration of the parameters to an ESM or observations.

We agree that the distinction between “highly idealized tipping frameworks”, “tipping-element emulators”, and the present approach was not stated clearly enough. By “highly idealized tipping frameworks” we mean conceptual or statistical representations of tipping behavior that are used without direct calibration to higher-complexity process-based models. In this category we include, for example, conceptual double-well or saddle-node systems with parameters chosen from theory, illustrative assumptions, expert judgement, or broad literature ranges, as well as more abstract Markov-chain approaches. In contrast, by “tipping-element emulator” we mean a conceptual model with explicit threshold dynamics whose parameters are calibrated against output from a higher-complexity model or observational constraints. We count our approach to this class. We will revise the manuscript to state this explicitly.

L45:

Similarly, what are “reduced-complexity climate frameworks”?

By this term we meant simplified climate or climate-carbon-cycle modelling frameworks that are computationally inexpensive compared with fully coupled ESMs and are therefore often used for scenario ensembles, sensitivity analyses, or coupling to impact modules. An example is the SURFER model, which is mentioned in the next sentence. We agree, that this term is too broad and we will be more explicit in the revised version.

Eq 1: f(z,p,t)

Thank you, we fixed it.

Regarding Eq. 1 and Eq. 2:

Does the model not include a noise-term? Eq. 1 as it stands is deterministic and only allows for monotonic relaxation towards a fixed point. The ESM simulations on the other

hand feature oscillations and quasi-stochastic variability. Comparing (in squared difference sense) time increments of the deterministic 1D model (where the time increments go to zero as it equilibrates) to those of the ESMs (time increments should reach some stationary distribution with zero average to zero, but only if perfectly equilibrated) seems not very reasonable. Or are you only predicting with Eq. 1 one step ahead, i.e. you take the ESM value at time i and predict one time step ahead to get your model increment? This is not clear from the provided text.

We thank the reviewer for this helpful comment. Eq. (1) is indeed deterministic and does not include an explicit noise term. Our aim is therefore not to reproduce the oscillatory or quasi-stochastic variability of the comprehensive model simulations, but to represent their underlying equilibrium structure with our reduced one-dimensional model.

We agree that fitting the deterministic model to time increments of the comprehensive-model output is not the most appropriate formulation in the presence of internal variability. To avoid this ambiguity, we will revise both the manuscript text and the fitting procedure. In the revised version, we do not fit transient increments or freely integrated trajectories. Instead, for each forcing level, we compare the comprehensive-model state to the corresponding equilibrium state of the reduced model and minimize the squared distance to the closest equilibrium branch. The fit is obtained in two steps, (i) a global differential-evolution search provides an initial parameter estimate; (ii) this estimate is refined with a local least-squares optimization. Where needed, additional information on the threshold positions is imposed through soft constraints in the objective function.

L128-135:

This paragraph mixes up the response of the system to perturbations before a bifurcation (critical slowing down) with the relaxation time scale after a bifurcation (as in the example of faster GrIS melt for strong warming above a threshold). It needs to be made clear that there are two different things. Also, I think it is a bit fuzzy to say the “timescale” of a system depends on its state. What is meant by timescale? The relaxation towards an equilibrium would be characterized (in the linear regime) by some e-folding time scale, but this does not depend on the state. Of course the momentary rate of a system state’s movement is given by the right-hand side of the underlying differential equation and does depends on the position in phase space (the state). In case this rate of change varies arbitrarily in phase space, in my opinion this implies that there is no characteristic timescale, as opposed to a “state-dependent” timescale. It would be perhaps better to only say – and this is in line with how the state-dependent timescale is used later - that there can be different characteristic time scales of local relaxation for different attractors.

We agree that the original wording conflated two different notions of timescale (i) the local relaxation timescale of small perturbations while the system is still on a stable branch,

which diverges near a fold bifurcation due to critical slowing down, and (ii) the transient adjustment time after the bifurcation has been crossed and the system evolves toward a different attractor. In the manuscript these two concepts were discussed too closely together and not properly distinguished. We will revise the manuscript to make the distinction clearer. The fitted τ in our manuscript should be interpreted as the latter, an effective transient timescale governing the evolution of the model after crossing the respective critical threshold. As this is important since it determines when a certain impact would be realized as opposed to the time when a certain tipping point is crossed. Also see Fig. A1 and the other comments regarding the time scale.

We will make the phrasing clearer and avoid the term *state-dependent timescale* and rather call it *effective transient timescale* or similar.

L189:

"we utilise the results from three independent ice sheet models to fit a dynamical system to the GrIS" → To me this makes it seem like the method allows to derive a single dynamical system from three independent ESMs, whereas in reality one independent dynamical system is obtained for each ESM. Similarly, and even more strongly, this is suggested in line 307 too.

Our phrasing was not clear here. We will change it accordingly.

L281-282:

I am not sure about the justification for filtering the high-frequency variability. If the authors are interested in the quasi-equilibrium state, they could just fit the fixed points of Eq. 1 to the average states in the ESM equilibria (as a function of p). What is the effect of filtering, and why is it done?

The "filtering" via a moving average gives an estimate of the quasi-equilibrium as often done in the literature. So, it removes the short-term variability from the ESM output, so the fit actually targets the quasi-equilibrium branch. In the quasi-equilibrium regime, fitting the fixed points of Eq. 1 to the averaged ESM states is exactly what we do. The model is not being calibrated to transient dynamics here, but to the equilibrium branch as a function of p. Filtering is only a preprocessing step to extract that equilibrium from noisy ESM output.

L289-301:

This relates to the previous point and to the major comment #2:

Why can the characteristic timescales not simply be estimated from the fluctuations around the equilibria from the quasi-equilibrium runs? Why is it necessary to first specify an arbitrary timescale k, and then afterwards fit the parameters [a,c,e] based on the ESM

timeseries, which can override the choice of k? Perhaps it would be possible for the authors to demonstrate that choosing a value for k is really necessary in order to obtain a fit.

Thanks for the comment. As mentioned before (e.g. major comment 2), we always fitted a,c,e first and then afterwards determined the timescale (k and alpha). We agree, that the intrinsic time scale could also be determined from the fluctuations around the equilibrium, if such simulations where available for all elements. However, we are mostly interested in the transient response after crossing the threshold, meaning the time it takes for the system for near-full disintegration. This measure is important as it determines when a certain impact would be realized as opposed to the time when a certain tipping point is crossed. Therefore, in the sense of our model we decided to take the timescale for the transient response after crossing the threshold.