20 October 2026

Author's response

"To tip or not to tip", egusphere-2026-3507

Please find below our response (black text) to reviewer 2's comments (blue text). Reviewer 1 had no comments to address. All changes indicated below have been implemented. Additionally changes include:

- The reference list has been updated and expanded.
- Figures 3 and 4 have been modified following the suggestion of the reviewer. The CO2 forcing timeseries of RCP2.6 in Fig. 4 was incorrect in the previous version (it was not constant after 2100). This has now been corrected. The results have not been affected by this.
- The description of Eq. (1) has been slightly adapted to be more accessible and closer to classical formulations of linear response equations. Furthermore, we added a short appendix (now Appendix A) to briefly discuss assumptions and generalizations of Eq. (1).

Response to reviewer 2

> In this manuscript, the authors propose a new, much-needed, general definition of tipping points in the Earth sciences, outline specific criteria for meeting this definition, and provide a mathematical framework for evaluating whether these criteria are met for a given physical system. In my view, the proposed definition itself is clear, robust, and fulfills the authors' stated goal of bridging mathematical intuition on tipping and considerations around practicality/impacts of identifying tipping points, and if adopted would bring a lot of clarity to the field of Earth system tipping dynamics. I have some comments that I hope will strengthen the manuscript by improving the logic of some paragraphs and definitions of some concepts, but I think that overall, this manuscript will be a very useful perspective piece.

We thank the reviewer for the positive evaluation and the very useful comments, which we respond to point-by-point below.

> Specific comments

1. Line 91: "However, it is often not useful to define a concept via its relevance." Without further justification, this seems to be a rather vague and subjective statement. I recommend the authors instead say what the risk/downside is of defining a concept by its relevance.

We agree with this and will change the sentence accordingly: "However, defining a type of dynamics based on relevance or impact risks being subjective, poorly distinctive and inconsistent with mathematical descriptions."

2. Line 100: "…arguably not all abrupt changes classify as tipping events." Here it is not clear what definition of "tipping event" the authors are referring to--- certainly by the Lenton definition, an abrupt change is classified as a tipping point (simply because that is how Lenton has defined a tipping point). From the next clause, it becomes clear that they mean that an abrupt change may not be associated with any of the

mathematical principles (bifurcation, multi-stability) that they deem important, so I recommend making the language more specific/clear in this way. In general, any time the authors refer to a "tipping event" or "tipping point" prior to introducing their own definition (e.g., anywhere in sections 1 and 2), they should be careful to say which of the other (insufficient) definitions they're using or what specific characteristics they're referring to.

We will replace "classify" as "should classify" to clarify that we are referring to a desired tipping definition, rather than a specific existing one. We will also add a sentence of clarification to the paragraph. Note that the Lenton et al. 2008 definition does not focus on abruptness but mainly on a qualitative change caused by a tiny perturbation. Their definition thus excludes forced abrupt changes due to large, abrupt perturbations.

3. Line 106—108 (and specifically, "this can result in spurious hysteresis effects…"): I understand here the point the authors are trying to make about irreversibility and spurious hysteresis, but I think the authors should expand these two sentences to explain things a bit more explicitly for a broad audience. For example, I recommend they first define hysteresis as the mathematical concept underpinning "irreversibility" (I don't believe they do this anywhere in the manuscript, and this should maybe go at the start of the paragraph to explain why it is associated with tipping points in the first place), explain how it conceptualized as a signature of multi-stability/a bifurcation, and then explain how a linear system can also exhibit this signature with the right depending on the magnitudes of inertia/forcing timescale. Then it will be clearer to a general reader why irreversibility/hysteresis is not a reliable indicator of nonlinear change.

Thank you for pointing this out; we will expand the explanation in this paragraph following the reviewer's recommendation and hope to thereby make it more accessible.

4. Line 112: Related to comment #2, what is meant by "tipping-like behavior" here? Transient/spurious hysteresis? Abruptness?

We will clarify in the revised version that by "tipping-like" we mean the phenomenology of a relatively abrupt, persistent shift.

5. Line 136: This seems like an appropriate/necessary place to add more justification for the focus on the positive-feedback aspect of tipping. How does focusing on this aspect serve the authors' goals/purpose of a new tipping definition?

The positive-feedback aspect connects the observed phenomenology to the system's governing processes and is thus important for ensuring robustness of the tipping definition. For example, purely statistical tipping indicators have been found to lack robustness because they are prone to false positives. We argue that if we would consider nonlinear response alone, without making a connection to the cause, this could lead to overuse of tipping terminology. Moreover, the positive-feedback aspect is prominent in less technical descriptions of tipping points, both in climate science and social science, and aids the intuitive understanding of tipping phenomena as internally driven ("self-propelling" or "run-away") change. It thus supports our goal of bridging between mathematical rigor and broad accessibility. We will revise the first paragraph of section 3 accordingly.

6. Given that the authors state that the qualitative focus of their definition is on the positive feedback/self-amplifying change (line 136), then the phenomenology they focus on should also be clearly tied to self-amplifying change. Is this the case? In other words, does being in a regime of dominant positive feedbacks always lead to a “persistent nonlinear transition”? If so, then with criterion 2 also being met, there would be a one-to-one relationship between a persistent nonlinear transition and the positive feedback, which would unify the two criteria of tipping around the central concept of positive feedbacks.

Criterion (1b) already implies that the change is amplifying relative to what one would expect from linear response. Assuming that the forcing F accounts for all relevant external influences, this change is then internally driven, i.e. self-amplifying. Being in a regime of dominant positive feedbacks will lead to a self-amplification of perturbations, but whether this change becomes a persistent transition depends on the balance of other feedbacks that enable the existence of an alternative linear response regime. Therefore, we argue that a “persistent nonlinear transition” always requires a positive feedback process, but the dominance of a positive feedback process is not a sufficient condition for a persistent nonlinear transition. This is why we propose that our criteria (1) and (2) are each necessary conditions for tipping and in concurrence are sufficient.

7. In criterion 1c, the condition that tau_B lasts “at least as long as tau_TH” seems like a rather arbitrary choice/definition for persistence. Can the authors justify this further? In particular, I can imagine that during a “fast” tipping transition (let’s say something that occurs over 5-10 years for the Earth system), only observing 5-10 additional years would be insufficient for establishing that a dynamically distinct and linear regime has been reached.

This is true, but any chosen value of the persistence P will be arbitrary for a general tipping system and choice of tau. Therefore, we refrain from specifying persistence further; instead, we emphasize more clearly in the revised manuscript (in the Remarks section) that tipping may be misclassified if P or tau are too small in the given context.

8. Section 3.1.1 Remark #1: Is this special case considered a tipping event by their definition? (and is this classification desirable?)

Yes, this special case is considered tipping and this is desirable in our opinion, as noise-induced transitions between competing metastable states should qualify as tipping events. We will clarify that we consider this special case a tipping event.

9. Section 3.1.1 Final remark: This scenario seems like a misclassification of a tipping point by their criteria/definition that one would want to avoid. Maybe the authors want to consider adjusting their definition to require a minimum tau that depends on the response timescale of the system, or at least suggest that best practice would be to only apply this tipping analysis over a sufficiently long tau (which again would depend on the response timescale of the system considered)?

This reflects a fundamental timescale issue. A minimum tau requirement would be arbitrary and context-dependent, which is why we do not make this specification. There is also often not a single response timescale of the system, but a spectrum of response timescales. We

will now highlight that tau must be sufficiently long (whatever that may mean in the specific context).

From a dynamical systems perspective, it would indeed be desirable to avoid this possibility for misclassification, since there is a distinction between (a) a transition between competing attractors and (b) a transition between slow phases of a fast-slow/relaxation oscillation. However, from an impact perspective on the time horizon tau, both (a) and (b) yield a very similar phenomenology and can be related to a positive feedback mechanism - so it could make sense to classify it as a tipping event even if it is a mode of variability on a longer timescale.

10. Criteria for rejecting the null model (line 210): Although it is somewhat implicit from the first two criteria that eta(t) increases between tau_A and tau_B, I feel it would add clarity to simply add this as another explicit criterion: that eta(t) surpasses some eta_c during tau_TR.

Thank you for the suggestion. The definition in section 3.1 remains qualitative, which is why we prefer not to introduce a threshold eta_c there. This threshold is later introduced when discussing the tipping test based on a Green's function.

11. The authors might want to consider adding a few sentences at the end of section 3.2.1 connecting the mathematical criteria they have outlined back to the more qualitative criteria they establish in section 3.1. This would help the reader understand that the purpose of constructing a null model and evaluating the time evolution of the residual is to be able to quantitatively assess criteria 1a, 1b, and 1c (~line 155).

We will add a short paragraph at the end of section 3.2.1.

12. Paragraph 214—221: These are very useful nuanced comments!

Thank you for the feedback.

13. While the authors spend a lot of time establishing a clear framework for evaluating criteria 1a—c (null model, etc.), they do not spend much time explaining how one would establish that criteria #2—that the persistent nonlinear change is physically caused by a positive feedback—has been met. Around line 260, they describe how one can establish the presence of a dominant positive feedback based on the concavity of the stability landscape, but for real/high-dimensional systems, the stability landscape is unknown. A bit of elaboration on practically how to identify a net positive feedback in a system where you don't a priori know the stability landscape (perhaps using the AMOC as an example) would be helpful.

Indeed, identifying the role and relative importance of the positive feedback is challenging in complex/high-dimensional systems. In fact, feedbacks are often formalized in an equilibrium setting, such that for out-of-equilibrium dynamics it is not fully clear how to quantify net positive feedback even in cases where the stability landscape is known. The NAF and BB measures proposed in the manuscript are two suggestions for this.

In practice, a dedicated investigation will be needed to convincingly identify the role of the positive feedback in the system. For AMOC tipping in CESM, where the salt-advection feedback is the relevant feedback, this has mainly been done in two ways: 1) relate the

dynamics of CESM to low-dimensional conceptual models where the salt-advection feedback is explicitly built-in as the destabilizing process and 2) perform a feedback analysis in quasi-equilibrium simulations of CESM by decomposing the Atlantic freshwater budget into different contributions. This model hierarchy approach, combined with mechanistic attribution studies, could be a general recipe for establishing criterion 2. We will comment on this in the revised manuscript.

14. Picking null models/applying the tipping test: As the authors acknowledge, accurate tipping classification with this framework requires accurate specification of the null model. For the CESM 0.5ppm/year AMOC case, the authors identify that the previously chosen null model (damped decadal oscillator) was actually a poor choice based on the fact that it led to the classification of the AMOC trajectory as a tipping event when it clearly wasn't. This led them to re-specify the null model. In essence, this meant they needed to already know whether the trajectory contained a tipping point or not in order to construct the correct null model that would give them the "correct" tipping classification. This begs the question--- can the framework the authors introduce actually be used to classify timeseries/systems as tipping/non-tipping when we don't already know what they "should" be classified as? In other words, is this just a way of formalizing two different classes of behavior, or can it be used for active prediction/classification? Perhaps a way to answer this is to formulate an example where the tipping/non-tipping classification of a trajectory is truly arbitrary based on the forcing and observable timeseries alone (or where tipping/non-tipping trajectories within the same system are qualitatively similar), but where an accurate formulation of the null model leads to a meaningful classification result.

We used the CESM simulations as a second, more complex example precisely because we have a ground truth, i.e. we know what outcome the tipping test should yield based on our existing understanding of the system. This serves to illustrate also the challenges associated with the choice of null model, especially a simplistic elementary null model like the harmonic oscillator. In principle, however, active classification of events with unknown class (tipping/not tipping) is certainly possible with our approach, provided that the null model is appropriate. To verify this, no information about tipping events is needed, since the null model must only describe the linear response dynamics of the reference regime, encoded in the Green's function that can be estimated from small perturbations. This relates to the core idea of response theory to predict the response of the system purely based on information about the unperturbed reference state.

To keep the paper concise, we will stick with the two examples and simple null models presented in the original version of the manuscript.

15. How is the interval tau_TR determined in Figures 2/3?

The intervals tau_TR in Figs. 2 and 3 are chosen by eye based on the timeseries of the observable, ensuring that tau_B is sufficiently long and that, in cases where the forcing varies in time, tau_A contains some change of F (otherwise the fit of the sensitivity parameter beta would be poorly constrained).

Technical comments

1. Line 56: Should be "In the case of a forcing threshold…"

Will be corrected, thank you.

2. Figure 3: I recommend mentioning the frequency used for the null model in the figure caption to avoid confusion with the different null models used in Figure 4.

We will add this to the figure caption.

3. Relatedly, it's a little confusing that the colors blue/orange are used to represent two different null models in Figure 4 (and black is used to represent actual timeseries), while blue and orange (although different shades) are used to represent the different forcing scenarios in Figure 3. It would help to rethink the color schemes/panel layout in Figure 3 to be either more consistent with Figure 4 (even though there's only one null model in Figure 3) or be completely different from Figure 4. For example, they could consider having a column of panels for each forcing scenario in Figure 3 showing all the desired quantities (as in Figure 4).

Thank you for this point; we will update the color scheme and bring Fig. 3 into the same layout as Fig. 4 to streamline the presentation.

Sincerely,

Reyk Börner and Henk Dijsktra