Reply to reviewer #1

I found the manuscript to be an interesting read and a useful follow-on from the Bury et al., 2021 paper, with a focus that switches to important considerations for the Earth system, such as are the abrupt shifts the DL model might predict the approach to irreversible or not. I think the work presents a great step forward in the literature, but I have some concerns about the robustness of the results and I think it would benefit from stepping away from the Bury et al., paper by explaining certain aspects in more detail rather than leading the reader to read the other paper. I have detailed my comments below.

→ Thanks for reviewing our manuscript 'An Earth system deep learning classifier for tipping point detection' and for your insightful comments! We will revise our manuscript in light of your comments and provide a detailed point-by-point response to each of the issues raised below.

One of my main comments is that I do not think the training set is explained well enough. I think the time series essentially are the ones from the 'fold' and 'null' categories in the other paper, but perhaps an equation explaining the derivation would help the reader understand what is being used in this training specifically. I get the impression that the underlaying equation can be used to simulate a number of different bifurcations but this information would help the reader understand properly.

→ We agree that the manuscript will benefit from a section explaining the training data set in more detail. We will add information on the parameters and the equations used to generate the training data, the stochastic simulation algorithm and the bifurcation identification (using AUTO-07P) in the material and methods section.

To a certain degree I can understand why the transcritical time series were used as an OOD test but currently it is difficult to look past what feels like a strong focus on fold and transcritical bifurcations. An instant question would be why not include transcritical time series in the training set as null cases and I assume this is because it would be difficult to determine where to draw the line on what to include. I would be interested to know how well a DL model trained on a binary output of those two differs from the current model.

→ The focus on fold and transcritical timeseries is intentional. We do not include transcritical time series in the training set as null cases, as this would prevent us from assessing the degree of their misclassification with the OOD detection approach. To be identifiable as OOD, the classifier is not allowed to have encountered the respective type of timeseries during training. Therefore, we restrict our training to fold and null cases in order to be able to use our OOD detection approach. However, we agree that training a classifier with transcritical timeseries included as null cases is a valuable alternative approach that would be an important direction for future work comparing different methodologies.

Regarding the OOD tests, I would be interested in seeing how other time series are classified e.g. different types of bifurcations (non-catastrophic Hopf), white to red noise processes where the memory in the system is increased, or stable time series where the noise level is increased. Furthermore, including catastrophic Hopf bifurcations in an OOD test could tell you how reliant the model is on only fold bifurcations being considered catastrophic shifts.

→ We agree that these would be interesting aspects to investigate. In this manuscript we specifically focus on training a classifier for the B-tipping case, not taking into account varying noise processes; this is out of scope of the current study. However, based on the reviewer's suggestion, we've run tests including Hopf bifurcations within our OOD detection framework. All length 500 empirical timeseries previously identified as fold bifurcations by our classifier are recovered as in-distribution (fold bifurcations), meaning they are not misclassified as Hopf bifurcations. For these analyses, we treated subcritical (catastrophic) and supercritical (non-catastrophic) Hopf bifurcations as one class, as they are not distinguished in the training data. To our knowledge, there is not yet a reliable early warning signal that can distinguish between these two types of Hopf bifurcation. We judge Hopf bifurcations to be only moderately common in the Earth system, primarily describing systems with an oscillatory nature. Therefore, to keep the manuscript focused, we intend to show these additional new results in the supplement and keep our focus in the manuscript on fold bifurcations since they appear widely in models of Earth system tipping elements. Generally, we agree that this should definitely be expanded on in further studies, also including the other timeseries mentioned by the reviewer.

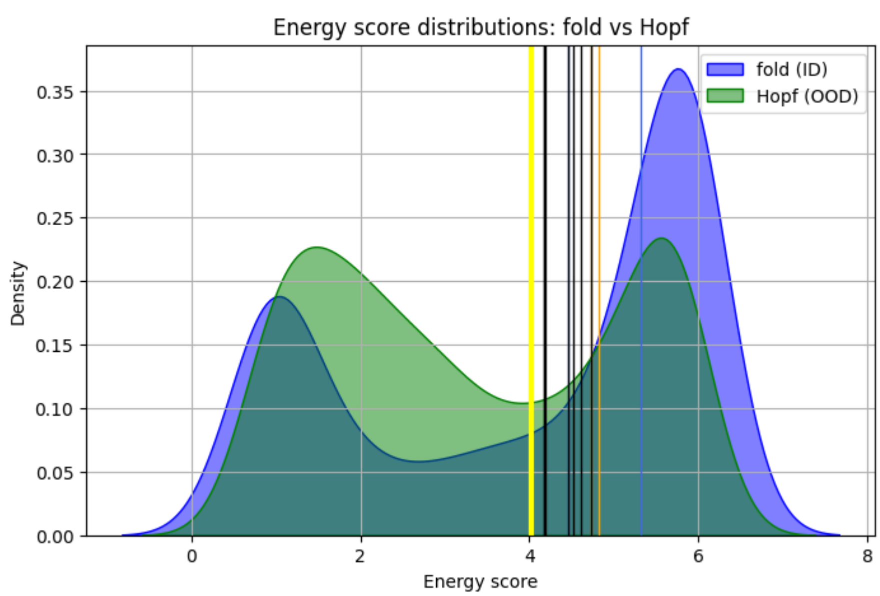

Figure 1: Energy density histograms based on the length 500 test dataset of synthetically generated Hopf timeseries (OOD) and fold timeseries (ID), the yellow line indicates the threshold between OOD and ID, other lines mark energy scores of empirical timeseries (black: sapropel timeseries, blue: ETM2 timeseries, orange: Oligocene/Miocene timeseries). The threshold between OOD and ID (yellow line) is set at the energy score, for which the number of correctly identified ID and OOD timeseries is equal.

I have a few minor comments regarding readability and presentation which I have detailed below.
Line 24 – The classifier only identifies if a transition is abrupt and irreversible, not 'how' abrupt or irreversible it is.

→ We will change this to 'if the transition is abrupt and difficult to reverse'.

Line 36 – Consider 'The Earth system' rather than 'System Earth'.

→ We will change this to 'The Earth system'.

Line 65 – I'd argue that generic EWS do not really have a qualitative nature, they use basic statistical measure and trends can be quantified. Perhaps rephrase to properly describe the added benefit of DL.

→ We agree and will rephrase this to 'we need to move away from our sole reliance on generic EWS, that indicate approaching transitions in general, to a more nuanced approach, that can confidently detect catastrophic tipping points of an abrupt and difficult to reverse nature'.

Line 85 – 'Two bifurcation types of interest' suggests that the 'null' category is a bifurcation in this instance.

→ We will change this to 'class types of interest'.

Line 213 – 100 time series seems quite low, why was this?

→ We originally decided for 100 timeseries to keep computational demand feasible. Every timeseries is evaluated with an ensemble of 10 classifiers, which already led to a runtime of several hours for the classification of 100 timeseries. Originally we used the ewstools apply_classifier_inc() function for classification, which applies our classifier to incrementally increasing timeseries lengths allowing us to see how the classifier behaves as a function of time. The ewstools apply_classifier() function on the other hand enables us to look at the whole timeseries, only returning one prediction value and is therefore less computational demanding. In order to evaluate more timeseries we chose for this function and applied it to 1000 transcritical timeseries to evaluate how many of them are misclassified as fold. The new analysis does not change the percentage of transcritical timeseries misclassified as fold (about 76%). We will update the corresponding manuscript and code sections.

Mentions of Fig. 5 – I would consider how Figure 5 is referenced as it is used a lot throughout the manuscript. Maybe do not reference it until the results section so it comes after Fig. 4. It is also confusing as it has results on it regarding the palaeo data that are not discussed in the text until much later.

→ We agree that the order of figures should be maintained in the manuscript. Thus, we will remove the references to Fig. 5 within the materials and methods section. To save figure space and avoid showing the energy density histograms twice, we decided to plot the empirical palaeo timeseries together with the OOD detection results already in Fig. 5. We will add additional explanations in Fig. 5 to clarify, that the OOD results of the empirical palaeo timeseries will be discussed later in the manuscript.

Figure 5 caption – Should briefly explain how the threshold (yellow line) was calculated or point the reader to the main text to find out.

→ We will add an additional sentence to the caption of Fig. 5, that explains that the threshold was chosen as the energy score, for which the number of correctly identified ID and OOD timeseries is equal.

Table 5 – To me (both in the table and in the main text) it reads that time series with higher noise are likely to be above the energy threshold regardless of if they are ID or OOD, meaning that it is classifying noisier time series as fold bifurcations regardless. Higher energy scores according to Fig. 5 should be associated with fold bifurcations yet the text on line 294 says they are associated with greater confidence in identification. Line 514 says something similar.

→ We acknowledge that this aspect might be confusingly worded in the manuscript and will change the respective parts in the results and discussion section accordingly. We will also update Fig. 5 and the corresponding code sections. Higher energy scores are indeed associated with identification of the timeseries as fold bifurcation. High energy scores are a measure of how confidently the classifier assigns the timeseries to a known **in distribution class**, i.e. the fold class. We find that timeseries with high noise values generally show higher energy scores than timeseries with lower noise levels and thus generally get classified as fold bifurcations. This is true for both ID and OOD timeseries. We conclude that the ‘noisiness’ of some transcritical timeseries might be the reason for their misclassification as fold bifurcations.

Section 3 – I think some specific referencing would be beneficial throughout the explaining the context of the results. Line 314 for example saying that your results corroborate ETM3 being the weakest event needs a reference.

→ We will add appropriate references throughout section 3.

Line 475 – 'other' should come after 'additional mechanisms' not before.

→ We will change this to 'additional mechanisms other than'.

I would suggest on the heatmaps (Fig. 6-8) on having one colour bar that goes from blue to red with grey in the middle. I am struggling with ~0.6 looking quite similar to me on both bars. Also be careful around these values on using the black fold text (Figure 7 for example), although solving the colour bar problem might fix this.

→ We will change the color scheme of Figs. 6-8 to a diverging one ranging from blue to red with grey in between, which also allows for better readability of the labels. We will update the corresponding manuscript and code sections.