Reply to reviewer #2

Overall, this manuscript builds on the work of Bury et al. (2021) by developing a binary CNN–LSTM classifier specifically designed to identify fold-bifurcation dynamics. The classifier is combined with an energy-based out-of-distribution detection method in an attempt to move beyond general early-warning signals towards the identification of bifurcation type. The research question is important, the methodological approach is promising, and the applications to palaeoclimate and palaeoceanographic records are potentially valuable.
However, the manuscript currently needs to define more clearly what can and cannot be inferred from the model output. The robustness of the empirical results with respect to classifier length, analysis-window selection, and the choice of event intervals also requires further evaluation. In addition, the construction of the synthetic training data and the extent to which patterns learned from low-dimensional synthetic systems can be transferred to complex Earth-system records should be explained in greater detail. My main comments are outlined 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.

Major comments

1. The relationship between fold bifurcations and irreversible Earth-system tipping points needs to be explained more carefully
The manuscript repeatedly associates fold bifurcations with abrupt and irreversible Earth-system tipping points. However, the mathematical meaning of a local fold bifurcation is that a stable and an unstable equilibrium collide and disappear. This may produce an abrupt transition and, under an appropriate bistable and global branch structure, may also lead to hysteresis. A local fold bifurcation alone, however, does not directly demonstrate strict irreversibility in an empirical system.
This distinction is particularly important here because the empirical analyses mainly use time series preceding the prescribed transition point. They do not examine the post-transition state, recovery trajectory, or system response when the external forcing weakens or reverses. The conclusions supported by the present analysis are therefore more accurately framed as follows: the pre-transition time series contain dynamical patterns consistent with those generated before synthetic fold bifurcations. This is not equivalent to demonstrating that the corresponding Earth-system transition was irreversible.
I recommend that the authors distinguish more clearly among fold bifurcation, hysteresis, difficult-to-reverse behaviour, and strict irreversibility. They should explain to what extent potential hysteresis or limited reversibility can be inferred from pre-transition time series alone, and discuss the limitations of that inference. Statements such as "detect irreversible tipping points" should be moderated, for example to "identify time-series patterns consistent

with an approaching fold bifurcation". Similarly, wording such as "how abrupt and irreversible" should be avoided because the model provides a class or class probability rather than a quantitative measure of the degree of abruptness or irreversibility.

→ We agree that the manuscript would benefit from a clearer explanation of the terms mentioned by the reviewer. Hysteresis is a defining feature of fold bifurcations. When a system crosses a fold bifurcation, a forward shift occurs and the system abruptly moves from the upper branch to a lower branch (see Fig. 1). Once on the lower branch, the original stable state cannot be restored by only changing the system parameters to the values at the bifurcation point (F2). The system parameters have to be changed even further (F1), to move from the new stable state back to the original stable state (backward shift). This behavior is called hysteresis. Hysteresis does not mean strict irreversibility, the system can return back to its original state, but only by 'overcompensating'. This makes fold bifurcations difficult (but not impossible) to reverse. Timescale also plays an important role in how irreversibility is defined (Buhr et al., 2025). Related to environmental and climate tipping points, the IPCC defines a perturbed state as 'irreversible', if the recovery due to natural processes takes substantially longer than the timescale of interest. We will add these points to our manuscript and will adapt it accordingly.

We also agree with the reviewer, that our classifier cannot measure the degree of abruptness or irreversibility of a transition and will change this statement.

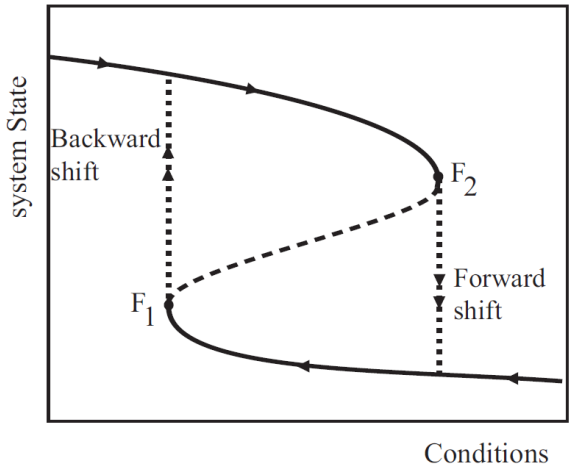

Fig. 1: Schematic diagram of a fold bifurcation (Scheffer, 2009).

2. The results are sensitive to classifier length and analysis-window selection, but no systematic sensitivity analysis is presented

One of the most important results in the manuscript is that different input lengths can lead to opposite classifications for the same event. This is not a minor shift in probability, but a change in the inferred bifurcation class. The PETM and the Oligocene–Miocene Transition provide clear examples. The discussion offers several possible explanations, but these are currently largely post hoc and are not tested systematically.

I suggest that the authors use the existing classifiers to analyse a set of nested pre-transition windows for each sufficiently long record. Fold probability could then be shown as a continuous function of both the effective number of observations and the physical duration

represented by the window. This would reveal whether the classification is stable over a range of window lengths or changes abruptly as earlier parts of the record are included.
The current comparison between length-500 and length-1500 changes both the empirical window and the classifier itself. It therefore remains unclear whether the differing results arise from the amount of empirical data included or from differences between models trained at different sequence lengths. These two effects should be separated as far as possible.
In addition, 500 data points represent very different physical durations in the PETM, CENOGRID, and sapropel records. The phrase "500 points" therefore combines sampling resolution, physical duration, and the intrinsic timescale of the system. Window length should consequently be reported in both number of observations and actual time span.
The manuscript also needs a reproducible criterion for selecting classifier length. When the length-500 and length-1500 classifiers disagree, it is not clear how a future user should decide which result is more reliable or how the two outputs should be combined. If a single selection rule cannot be provided, this length dependence should at least be treated explicitly as an important source of uncertainty rather than resolved through event-specific mechanistic interpretations alone.

→ We agree with the reviewer that a sensitivity analysis of pre-transition windows of different length would strengthen the manuscript and will perform these on the empirical records, that show different classifications of the same event. Further disentangling the effects of different empirical timeseries lengths and classifier lengths would also require training several new sets of classifiers with various lengths. We propose to train a set of new classifiers of different lengths ranging from length 500 to length 1500 (in steps of 100). However, due to high computational demand, we will limit the ensemble of these additional classifiers to two classifiers (one model 1 and one model 2). We share the full code and instructions to train additional classifiers with this manuscript, so future users of our DL method can easily train additional classifiers with lengths tailored to their system of interest. Providing a reproducible criterion for selecting classifier length is difficult and very much dependent on the analyzed system. Earth systems are unique and subject to different forcings on varying timescales. Thus, the chosen classifier should generally be long enough in order to actually see a response to the system's forcing. Identifying these forcings and their timescales requires domain specific knowledge. This 'caveat' has already been described for generic EWS, which are dependent on window length and thus require domain specific knowledge of forcings and timescales to choose the most fitting window length. In Earth system applications there is generally no 'one size fits all' solution, neither for the window length, nor for classifier length. To get a better idea of the timescales and forcings acting on a particular Earth system of interest, spectral analysis could be performed, which can help in choosing an appropriate classifier length for the analyzed system. We will discuss these points in the manuscript and emphasize this as a source of uncertainty in our analyses.

We agree with the reviewer, that different amounts of datapoints can mean very different physical durations in our empirical timeseries. Our manuscript already reports window length as both the number of datapoints and actual timespan covered in Tables 1-3, but to clarify this further we will also add it to the text in the discussion when appropriate.

3. The empirical validation lacks non-event control intervals

The empirical applications focus on predefined major transitions and analyse the intervals immediately preceding them. This is useful for comparison with earlier EWS studies, but it does not provide a realistic estimate of the background false-positive rate in geological records.

Even if high fold probabilities occur before known events, it remains necessary to determine whether similar signals also occur in ordinary background intervals, in non-stationary intervals that are not associated with a recognised tipping event, or in intervals affected mainly by proxy noise or preprocessing choices.

I recommend that the authors include several non-event control windows for each type of empirical record. These should be matched as closely as possible to the event windows in terms of duration, sampling resolution, and preprocessing. Suitable controls could include adjacent background intervals, intervals without recognised major transitions, or a small number of randomly selected ordinary intervals. The full workflow should then be applied in the same way, preferably without using event labels during the analysis.

Such controls would allow the authors to estimate the background false-positive rate and determine whether high fold probabilities are specifically concentrated before known transitions. This is especially important for the sapropel application, where S3–S9 are all classified as folds and share similar proxy characteristics, depositional settings, and processing procedures. If adjacent non-sapropel intervals also produce high fold probabilities, the event specificity of the model output would need to be reconsidered.

→ We concur that calculating the background false-positive rate would be informative in order to better determine whether fold bifurcations are specifically associated with the analyzed empirical transitions. We will carry out such an analysis on the sapropel record. The main complication is the availability of time intervals of background noise of adequate length. Only for sapropel event S3 and S5 there exists enough data outside of the time intervals already used for analysis with our DL classifier. We will focus on these intervals and will extract timeseries from this data to determine the false positive rate.

4. The origin, construction, and representativeness of the synthetic training data require fuller explanation

The description of the training data relies heavily on Bury et al. (2021). Referring readers to the earlier paper is reasonable, but the present manuscript should still be sufficiently self-contained for readers to understand what data were used, what the classifier was trained to recognise, and how the synthetic systems differ from the empirical records.

At minimum, the manuscript or Supplement should provide the general form of the randomly generated two-dimensional dynamical systems, explain how system parameters were sampled, define the fold and null classes explicitly, identify which state variable was retained, and describe how the control parameter approached the bifurcation. The distributions or ranges of the parameter-drift rate and stochastic-noise intensity should also be reported, because both can affect variance, autocorrelation, the duration of critical slowing down, and classification performance.

The hierarchy of the synthetic dataset also needs clarification. Did each underlying dynamical system generate multiple stochastic realisations? Which elements varied among those realisations? Was the train/validation/test split performed at the level of individual time series, or were all realisations from the same underlying system kept within a single subset? If different realisations of the same underlying dynamical system occur in both the training and test sets, the reported performance may overestimate generalisation to genuinely unseen systems. Where the dataset structure allows, a group-wise evaluation in which underlying dynamical systems are held out from training would provide a stronger test. If such an analysis is not feasible, the manuscript should at least describe the current splitting procedure clearly and discuss its implications for interpreting model performance.

→ We agree that the manuscript will benefit from a section explaining the training data set in more detail, as also indicated by reviewer #1. We will add information on the points brought up by the reviewer, including detailed parameters and the equations used to generate the training data, the stochastic simulation algorithm and the bifurcation identification (using AUTO-07P).

Regarding the hierarchical structure of the training data, each dynamical-system model was used to generate multiple stochastic realizations with varying parameter values. These realizations are not intended to represent different model classes; rather, they expose the classifier to variability within the two target classes: tipping and non-tipping. The class labels are assigned independently to each realization according to whether its underlying simulation approaches a bifurcation, and tipping and non-tipping realizations are not mixed. The same synthetic training dataset was used by Bury et al. (2021), whose trained classifiers successfully generalized to unseen mechanistic model simulations and empirical data. This out-of-distribution performance indicates that the classifiers learned transferable signatures associated with approaching tipping points rather than the identities of particular dynamical-system models. We therefore consider the training dataset sufficiently diverse for the binary classification task considered here.

Finally, the manuscript should discuss more fully the transferability of patterns learned from random, low-dimensional synthetic systems to complex Earth-system records. Palaeoclimate time series are influenced by multivariate coupling, non-stationary external forcing, age-model uncertainty, irregular sampling, proxy-specific processes, and sedimentary or diagenetic effects. The authors should explain which pre-bifurcation features are expected to be sufficiently universal across systems, which empirical characteristics may fall outside the training distribution, and how these differences limit the interpretation of the PETM, CENOGRID, and sapropel results.

→ The transferability of patterns learned from low-dimensional synthetic systems (as we use for training our DL classifier) to complex Earth system records is motivated by the Center Manifold Theorem (Kuznetsov, 1998). This theorem describes how the dynamics of a system close to a bifurcation can be reduced to a low-dimensional manifold, on which a corresponding normal form captures the leading order dynamics of the transition. This universality suggests that classifiers trained on diverse realizations of these normal forms

may learn dynamical features that genearlize across systems, including more complex Earth system models and observations. Thus, we use these low-dimensional models as synthetic training data for our DL model. We will add these explanations to our manuscript in the discussion.

Specific comments Abstract and Introduction

Lines 23–24:

The classifier appears to assess whether an input series is more consistent with fold or null dynamics. It does not quantify how abrupt or how irreversible a transition is. Please revise the wording accordingly.

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

Line 36:

"The Earth system" would be more conventional than "System Earth".

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

Lines 61–66:

Generic EWS are based on quantitative measures such as autocorrelation and variance. Their limitation is not that they are qualitative, but that they do not generally identify the type of bifurcation. Please revise this comparison and state the added value of the DL approach more precisely.

→ We will remove the term 'qualitative' and 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 (fold bifurcations)'.

Around Line 76:

The statement that transcritical bifurcations are relatively rare in nature should be qualified. Please clarify whether this refers to mathematical non-genericity or to their demonstrated frequency in empirical natural systems, which is difficult to establish directly from observations.

→ Transcritical bifurcations require special conditions and are thus non-generic, whereas fold bifurcations are generic and provide a canonical mechanism for tipping in climate models (Guckenheimer and Holmes 1990, Dijkstra 2019). Fold bifurcations have been

identified in different simulations of Earth system components, e.g. the AMOC (Dijkstra ad Weijer 2005; Hawkins et al. 2011, van Westen et al. 2025) and are generally associated with positive feedback loops. Such positive feedback mechanisms have also been proposed for the empirical systems analyzed in our manuscript. Examples are positive feedback loops related to the carbon cycle during the Paleocene/Eocene hyperthermals (Dickens et al., 1995; Deconto et al., 2012; Kurtz et al., 2003; Setty et al., 2023), positive feedbacks in ocean mixing during the transition into the Antarctic glaciation (Zachos and Kump, 2005) and positive feedbacks associated with sapropel formation in the Mediterranean (Hennekam et al., 2020). We will add these additional explanations to our manuscript.

Lines 82–87:

The phrase "two bifurcation types of interest" is inaccurate because the null class does not represent a bifurcation. "Two classes of interest" would be more appropriate.
Methods

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

Lines 99–109:

Please add a concise but self-contained description of the synthetic training data, including the general model form, parameter sampling, class definitions, control-parameter drift, stochastic forcing, and dataset hierarchy.

→ We agree that the manuscript will benefit from a section explaining the training data set in more detail. Please see our answer above to comment no 4.

Line 114:

Please explain the rationale for the unusual 0.95/0.04/0.01 train/validation/test split. It would also be helpful to report the variability in model performance across different random initialisations or data splits.

→ We decided for the 0.95/0.04/0.01 train/validation/test split for both the length 500 and length 1500 classifier due to the high number of synthetically generated timeseries available. A test set of a few thousand time series is adequate to provide a representative estimate of the performance metrics. In addition, we compare our newly trained binary classifier to an already existing multiclass classifier (Bury et al. 2021) with the same train/validation/test split and thus wanted to keep the parameters consistent.

We agree that examining variability across random initializations or data splits would provide an additional assessment of robustness. However, retraining the models across multiple initializations or data splits is beyond the scope of the present revision. In the revised manuscript, we will identify variability across retraining runs and data splits as a limitation and an avenue for future work.

Lines 143–145:

Please explain why nearest-neighbour interpolation was selected and discuss its potential effects on temporal structure. This method can introduce repeated values and step-like patterns, potentially altering autocorrelation and the form of the model input. A sensitivity comparison with at least one alternative interpolation method would strengthen the analysis.

→ We decided for the nearest-neighbor interpolation, as this has the least effect on autocorrelation and variability and best keeps the underlying dynamics intact. This has previously been described by Setty et al. (2023), who performed generic EWS analyses on this specific empirical dataset. We aim to keep our data preprocessing comparable to Setty et al. (2023), as we compare the results of our DL classifier with the generic EWS calculated in this study for the Paleocene/Eocene hyperthermals. Based on the reviewer's suggestion we will perform tests with other interpolation methods than the nearest-neighbor method to assess the robustness of our results.

Line 213:

Using only 100 transcritical time series to estimate the softmax misclassification rate seems limited given the larger synthetic test dataset. Please explain how these series were sampled, report an uncertainty interval, and, if feasible, repeat the estimate using a larger sample.

→ The transcritical timeseries for the softmax approach were chosen randomly from the test portion of the transcritical timeseries dataset using the numpy random.choice() function in python. 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. 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 with 1000 transcritical timeseries instead of 100 does not change the percentage of transcritical timeseries misclassified as fold (about 76%). We'll add these new results to the manuscript.

Regarding the uncertainty level we're not quite clear what the reviewer means by that specifically. If our explanations don't cover it, we're happy to receive more details on this.

Figures and Results

Figure 5:

The figure combines the synthetic OOD distributions with empirical results that are not discussed until later in the manuscript. This interrupts the sequence of the results. Please consider separating the synthetic validation and empirical applications into different figures, or revising the placement and order of the figure references.

→ This point was also addressed by reviewer #1. 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:

The caption should briefly explain how the yellow threshold was determined, or direct the reader to the relevant part of the Methods.

→ 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:

The main text refers to median noise levels, whereas the table heading reports means. Please verify the statistic used and make the terminology consistent.

→ We use median noise values in our analysis and will update the typo in our manuscript and the associated python scripts.

Section numbering:

Section 3.3.3 is followed directly by Section 3.6. Please check whether Sections 3.4 and 3.5 are missing or whether the numbering should be corrected.

→ The numbering is indeed incorrect, we will change this to 3.4 in the manuscript.

Interpretation of the empirical results:

In several places, agreement between the model output and an existing geological interpretation is described as "corroboration", for example in relation to ETM3 being the weakest event. Please provide appropriate references and distinguish between consistency with an established interpretation and independent validation of that interpretation.

→ We agree, that 'corroboration' might not be the clearest phrasing choice. We will change these instances in our manuscript and will also add supporting references. The new phrasing will reflect that our DL results provide support for, and additional information about, the abrupt changes that can be observed in the empirical records. Our results agree with existing generic EWS analysis, which have been used to identify critical transitions in the empirical datasets. While generic EWS cannot determine the type of transition, our DL approach can identify the type of the approaching bifurcation.

Line 475:

"Other additional mechanisms" is redundant. Please use either "additional mechanisms" or "other mechanisms".

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