https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

An Earth system deep learning classifier for tipping point detection

Madleen Grohganz[1], Thomas M. Bury[2], Bregje van der Bolt[3], Gert-Jan Reichart[1,4], Rick Hennekam[1]

[1]Department of Ocean Systems, NIOZ Royal Netherlands Institute for Sea Research, Den Burg, 1790 AB, The Netherlands
2 Department of Mathematics, University of California, Riverside, Riverside, CA 92521, United States
[3]Environmental Sciences Group, Water Systems and Global Change, Wageningen University, Wageningen, 6708 PB, The Netherlands
[4]Department of Earth Sciences, Faculty of Geosciences, Utrecht University, Utrecht, 3584 CB, The Netherlands

Correspondence to: Madleen Grohganz (madleen.grohganz@nioz.nl)

Abstract. Tipping points are thresholds at which a system, often abruptly and irreversibly, transitions from a stable state to a contrasting one. Crossing such critical boundaries poses a risk to Earth system stability and may have catastrophic consequences. This is especially relevant, as current climate change is destabilizing Earth subsystems, potentially bringing them closer to tipping points. Thus, it is important to be able to detect approaching tipping points in the Earth's system, which can be achieved through calibration on palaeo-records. Recently, new deep learning (DL) methods have been established that are able to confidently and quantitatively identify different types of critical transitions characterised by their abruptness and (ir)reversibility. Based on this, we develop a new (simplified) DL classifier focusing on the quantitative detection of catastrophic tipping points (fold bifurcations) in the Earth system. Our approach reduces computational demand and improves performance, especially for short timeseries. We first test the new classifier's performance on synthetic data and subsequently on different existing Cenozoic proxy records. Our DL results are compared to the results from previous studies applying generic early warning signals (EWS), which can detect approaching transitions qualitatively but cannot distinguish bifurcation types (abruptness and (ir)reversibility of the transition). Our DL classifier enables us to identify how abrupt and (ir)reversible an approaching transition is, which is important for tipping point risk assessment and mitigation. Results are generally consistent between generic EWS from previous studies and our DL approach and fit with what is known from the geological context. We note that some results are dependent on the length of the classifier used and the time interval investigated before the bifurcation. We implement an out-of-distribution (OOD) detection method to reduce the misclassification of non-catastrophic bifurcations as catastrophic tipping points. Combined with the binary DL classifier, this approach enables reliable, quantitative detection of catastrophic tipping points in Earth system records.

Keywords: tipping point, nonlinear dynamics, Earth system, machine learning, early warning signals

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

1 Introduction

System Earth is vulnerable to tipping points, critical thresholds at which gradual changes in external forcings destabilize key components, pushing them towards often abrupt and irreversible transitions (Scheffer et al., 2009; Van Nes et al., 2016; Lenton et al., 2008). Prominent examples include the Atlantic Meridional Overturning Circulation (AMOC) (Rahmstorf, 1995; Hawkins et al., 2011; Van Westen and Dijkstra, 2023; Van Westen et al., 2024), the Greenland and Antarctic ice sheets (Van Breedam et al., 2023; Robinson et al., 2012; Garbe et al., 2020), and other Earth subsystems that exhibit catastrophic tipping behavior under sustained anthropogenic climate forcing (Lenton et al., 2025; Lenton et al., 2008). Crossing thresholds in these systems can trigger drastic global impacts affecting millions of people worldwide, such as abrupt shifts in temperature and precipitation regimes, accelerated sea level rise (Drijfhout, 2015; Jackson et al., 2015; Armstrong Mckay et al., 2022), and even tipping cascades, i.e. a domino-like collapse of several interconnected Earth systems (Wunderling et al., 2021; Kriegler et al., 2009).

Critical transitions are defined by a loss of stability. Changing external forcings, such as greenhouse gas increases or ocean temperature changes, destabilize the system and push it towards a critical boundary, a so-called bifurcation point. The resulting shift may unfold gradually or discontinuously, with the latter defining a catastrophic tipping point: rapid, self-perpetuating, and often irreversible due to hysteresis (Milkoreit et al., 2018; Scheffer et al., 2009; Lenton et al., 2008). Hysteresis means that, even if forcings are removed or reversed to the previous level after passing a tipping point, the system remains locked in its new state until conditions are reduced even further (overcompensated), thereby complicating recovery efforts (Crawford, 1991).

Current approaches to tipping point detection rely on so-called early warning signals (EWS) (Scheffer et al., 2009). EWS are based on the observation that a system approaching a tipping point loses resilience and becomes slower at recovering from small perturbations, a behaviour called critical slowing down (CSD) (Wissel, 1984; Scheffer et al., 2009). Different statistical techniques can be used to infer CSD from timeseries leading up to tipping points, e.g. increasing lag-1 autocorrelation and variance. Using these CSD-based EWS, several studies have identified approaching critical transitions in palaeoclimatic timeseries as well as observational climate timeseries, e.g. related to marine (de)oxygenation (Hennekam et al., 2020a), the Greenland ice sheet (Boers and Rypdal, 2021), the AMOC (Boers, 2021), Dansgaard-Oeschger events (Boers, 2018) and major Cenozoic climate transitions (Boettner et al., 2021). However, these generic EWS are limited to qualitatively flagging approaching critical transitions in general. They cannot distinguish between smooth, easily reversible and discontinuous, difficult to reverse transitions. However, this aspect is particularly important for Earth Systems, where reversibility of transitions have major implications. To better assess the potential impacts of a critical transition on the Earth system, we need to move away from the qualitative nature of generic EWS to a more quantitative approach, that can confidently detect catastrophic tipping points of an abrupt and irreversible nature.

Recent studies have applied machine learning to quantitatively forecast tipping points. These studies trained deep learning classifiers to recognize time series leading up to different types of critical transitions/bifurcations (Bury et al., 2021; Bury et al., 2023; Deb et al., 2022). This approach outperformed generic EWS, i.e. traditional statistical

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

indicators (lag-1 autocorrelation and variance) in signaling an upcoming tipping point. Importantly, DL classifiers provide the ability to specifically identify systems with rapid and irreversible transitions, which conventional CSD-based indicators alone cannot achieve. The most critical bifurcation type identified in Earth systems is the fold bifurcation, which is associated with abrupt, irreversible tipping points (Scheffer et al., 2009; Lenton et al., 2025; Lenton et al., 2008). There are other bifurcation types that are also subject to critical slowing down (Kéfi et al., 2014), but involve a smooth transition, like the transcritical bifurcation. Transcritical bifurcations are non-generic and generally represented by simple models with strict bounds and are relatively rare in nature (an exception being e.g. the onset of epidemics) (Arnold, 2012; Seydel, 2009; Kermack and Mckendrick, 1927). But they are particularly difficult to distinguish from fold bifurcations due to their mathematical similarity (both have quadratic, one-dimensional normal forms (Kuznetsov, 1998)) and thus are regularly misclassified as fold bifurcations by current DL algorithms. We want to limit the amount of misclassified transcritical bifurcations and ensure, that the classifier confidently detects bifurcations that are abrupt and irreversible (fold bifurcation).

Here, we build on existing multi-class DL classifiers (Bury et al., 2023; Bury et al., 2021) and train a new binary classifier which focuses exclusively on the detection of fold bifurcations (abrupt, catastrophic tipping points). Multi-class classifiers are trained on multiple different types of bifurcations, while binary classifiers specifically focus on two bifurcation types of interest. We opted for a binary classifier over a multi-class classifier for two reasons: (i) it is computationally less expensive to train, and (ii) we expect it to have better performance on the specific task of detecting a fold bifurcation. Additionally, we implement out-of-distribution (OOD) detection to flag misclassified transcritical bifurcations, enhancing reliability. We test the performance of our binary classifier with OOD on synthetic data and compare its performance metrics to the multi-classifier approach (Bury et al., 2021). We then apply our approach to different Earth system proxy records, that represent instances of past transitions in temperature, carbon cycling and marine (de)oxygenation. These records reflect processes acting on different timescales, from several thousand to millions of years and have previously been shown to exhibit generic CSD-based EWS (lag-1 autocorrelation, variance) (Setty et al., 2023; Boettner et al., 2021; Hennekam et al., 2020a). Applying our new DL classifier and OOD detection we assess abrupt tipping behaviour associated with catastrophic fold bifurcations for these records in order to evaluate the potential added value of the approach proposed here.

2 Material and methods

2.1 Training data for the DL classifier

We trained two types of binary classifiers suitable to analyse timeseries of up to 500 datapoints (length 500 classifier) and timeseries of up to 1500 datapoints (length 1500 classifier). For training the DL classifiers we used the synthetically generated training datasets of length 500 and length 1500 respectively from Bury et al. (2021). The length 500 training dataset consists of 125,000 null time series and fold time series each, the length 1500 training dataset of 50,000 null and fold timeseries each. These timeseries are derived from simulations of randomly generated, two dimensional systems, that exhibit local bifurcations. Null timeseries are obtained from simulations with fixed parameters, fold

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

bifurcations from simulation with varying parameter towards the bifurcation. To obtain the residual time series, they were detrended using Lowess smoothing with a span of 0.2 (smoothing parameter indicating the width of the moving window used to calculate the residuals; in our case 20% of the original timeseries datapoints). Each residual time series was normalized by dividing each time series data point by the average absolute value of the residuals across the entire time series in order to remove scale dependence and allow comparison across different timeseries

2.2 DL algorithm architecture and training

For our DL classifier training we adapted the CNN-LSTM DL algorithm from Bury et al. (2021) (see their paper for more details) with one convolutional (CNN) layer and two recurrent (LSTM) layers. The algorithm was trained for 500 epochs with a learning rate of 0.0005. Due to the high number of synthetically generated timeseries, we used a train/validation/test split of 0.95/0.04/0.01. Two variants (model 1 and model 2) of the length 500/length 1500 algorithm were trained on censored versions of the training time series. Model 1 was trained on time series, that were censored by padding them with a variable number of zeros at both the left and right (beginning and end of the time series). This results in time series of various lengths (as short as 50), that can represent any part of the time series; not necessarily the time phase before the transition, but also the middle or beginning sections. Model 2 was trained on a censored version of the training set, where the time series were padded with zeros only on the left (at the beginning of the timeseries). This results in time series of various lengths (as short as 50); but all of them directly leading up to the bifurcation. Due to their different censoring, model 1 excels at picking up early warning signals from the early parts of the timeseries and model 2 rather at detecting the specific type of bifurcation close to the actual transition. To leverage and integrate the capabilities of both models, we take the average prediction of five model 1 and five model 2 classifiers at each point to generate the reported results (ensemble) for each classifier (length 500 and length 1500). Prior to application on empirical systems we evaluate the performance of our binary classifier on the test portion of the synthetically generated dataset (see results section).

For each of the length 500 and length 1500 classifiers we calculated different performance metrics: precision (how many positive predictions are true positives), recall (how many of the true positives are detected), and f1 score (a combined measure of precision and recall). The final performance values are the means of these metrics for an ensemble of five model 1 and five model 2 classifiers for length 500 and length 1500 respectively.

2.3 Empirical systems for application of the binary DL classifier

We then apply the binary DL classifier to three different empirical systems from the Cenozoic and compare its results with previous studies focussing on generic EWS (lag-1 autocorrelation and variance):

1) The stable isotope record of the Paleocene and Early Eocene, in which several palaeoclimatic events associated with abrupt perturbations of the carbon cycle ($\delta^{13}C$) and changes in temperature ($\delta^{18}O$) can be identified (PETM, ETM2 and ETM3 events) (Fig. 1). The PETM is the most intense warming event (hyperthermal) of the Cenozoic accompanied by a rapid release of isotopically light carbon into the atmosphere-ocean system, global warming and significant ocean acidification; it was followed by the similar, but less intense hyperthermal events ETM2 and ETM3 (Westerhold et al., 2020; Zachos et al., 2005; Sexton

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

et al., 2011; Lauretano et al., 2018b). To assess the abruptness of these transitions we combine the isotope records from ODP Site 208-1262 (Barnet et al., 2019; Barnet et al., 2018) (PETM and ETM2) and ODP Site 208-1263 (Lauretano et al., 2018b; Lauretano et al., 2018a) (ETM3). As our DL classifier requires a time series with equally spaced data points, the $\delta^{13}C$ and $\delta^{18}O$ data are interpolated to regular sampling steps (using the nearest neighbor method). We apply the same analysis timeframes (see Table 1) and data pre-processing methods as Setty et al. (2023) and compare the results of our analyses to the generic EWS from their study. For timeseries longer than 500 datapoints (PETM) we apply both the length 1500 classifier (on the complete timeseries) and the length 500 classifier (on the last 500 points before the bifurcation).

Table 1: Beginning and end dates of analysed events (based on Setty et al., 2023) and number of data points in the respective time series.

Event	Transition [Ma]	Analysis timeframe [Ma]	Number of data points (and applied classifier)
Paleocene/Eocene Thermal Maximum (PETM)	55.935	61.28–55.935	1397 (length 1500 classifier) and 500 (length 500 classifier)
Eocene Thermal Maximum 2 (ETM2)	54.06	55.141–54.06	416 (length 500 classifier)
Eocene Thermal Maximum 3 (ETM3)	52.87	53.936–52.87	411 (length 500 classifier)

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

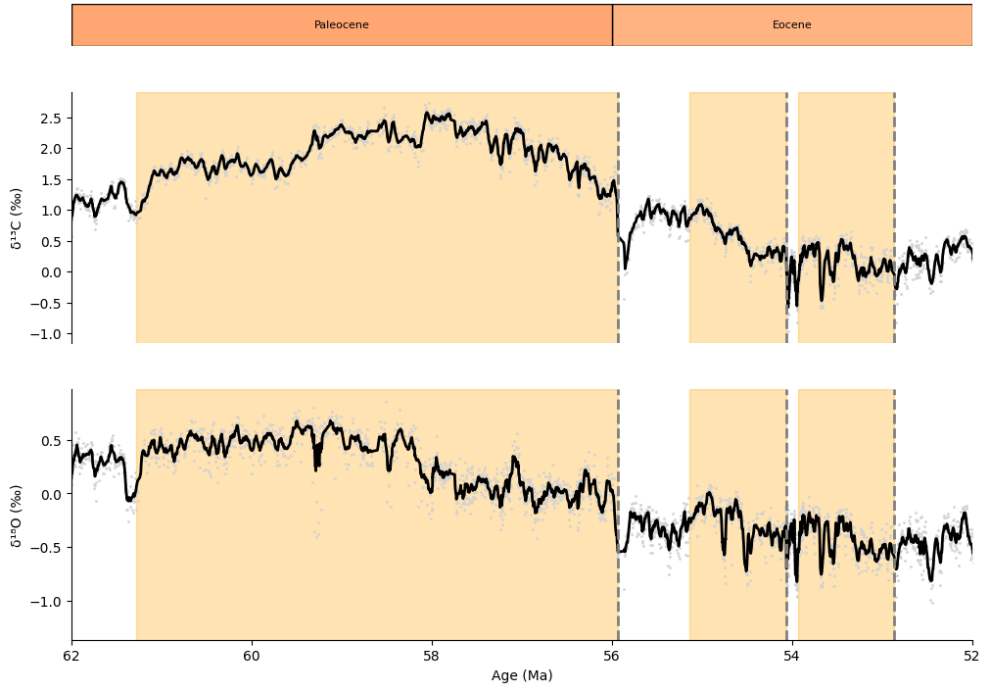


Figure 1: $\delta^{13}C$ and $\delta^{18}O$ isotope record of the Paleocene and Early Eocene with PETM, ETM2 and ETM3 events marked with dashed lines, areas shaded in yellow indicate the analysed time intervals prior to the events.

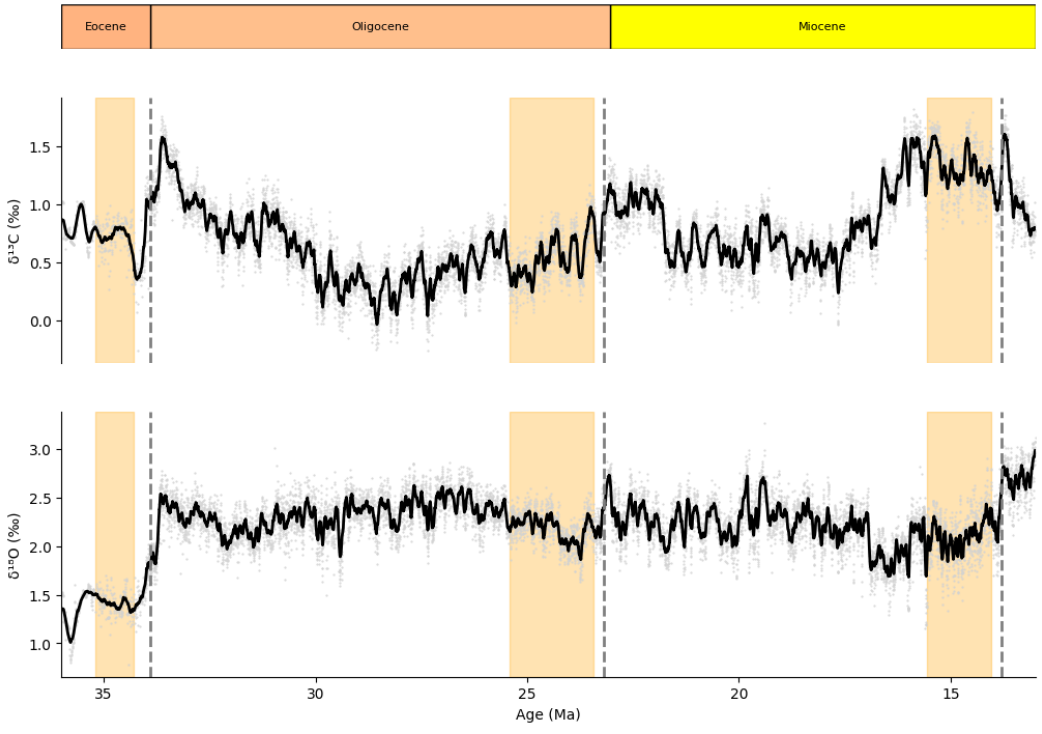
2) The CENOGRID dataset, a high-resolution $\delta^{13}C$ and $\delta^{18}O$ isotope record covering the Cenozoic (Westerhold et al., 2020). The CENOGRID dataset is a continuous composite of benthic foraminifer isotope records, derived from 14 Atlantic and Pacific ocean drilling cores mainly from low to mid latitudes (Westerhold, 2020). Several Cenozoic palaeoclimatic events associated with abrupt changes in temperature ($\delta^{18}O$) and abrupt perturbations of the carbon cycle ($\delta^{13}C$) can be identified in this record. Here, we focus on the Eocene/Oligocene Transition, Oligocene/Miocene Transition and Middle Miocene Climate Transition, events associated with significant positive $\delta^{18}O$ isotope excursions indicating shifts from a warm greenhouse climate towards a colder icehouse climate (Fig. 2). These transitions are related to different steps in the build-up and expansion of the Antarctic ice sheet (Lear et al., 2008; Wilson et al., 2008; Zachos et al., 2001; Shevenell et al., 2004), potentially representing tipping points of the ice sheet. To assess the abruptness of these events we use the binned and interpolated isotope records (Westerhold (2020), S34; see also SI of the original paper Westerhold et al. (2020)). We adapt the analysis timeframes (see Table 2) from Boettner et al. (2021), which are optimized to cover the highest data point density leading up to the transition. The results of our analyses are compared to the generic EWS from Boettner et al. (2021), who use non-equidistant time series from the binned, but non-interpolated version of the CENOGRID record. For timeseries longer than 500 datapoints

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

(Oligocene/Miocene Transition and Middle Miocene Transition) we apply both the length 1500 classifier (on the complete timeseries) and the length 500 classifier (on the last 500 points before the bifurcation).

Table 2: Beginning and end dates of analysed events (based on Boettner et al., 2021) and number of data points in the respective time series.

Event	Transition [Ma]	Analysis timeframe [Ma]	Number of data points (and applied classifier)
Eocene/Oligocene Transition	33.9	35.2-34.305	180 (length 500 classifier)
Oligocene/Miocene Transition	23.2	25.424-23.44	993 (length 1500 classifier) and 500 (length 500 classifier)
Middle Miocene Climate Transition	13.8	15.56-14.04	761 (length 1500 classifier) and 500 (length 500 classifier)



https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

Figure 2: CENOGRID $\delta^{13}C$ and $\delta^{18}O$ isotope record with Eocene/Oligocene Transition, Oligocene/Miocene Transition and Middle Miocene Climate Transition events marked with dashed lines, areas shaded in yellow indicate the analysed time intervals prior to the events.

3) The eastern Mediterranean sapropel record. Sapropels are organic-rich sediment layers occurring regularly in eastern Mediterranean sediments and represent changes from an oxic to anoxic state (and back) as a result of relatively gradual increasing freshwater inputs (Hennekam et al., 2020a). To assess the abruptness of these oxic-to-anoxic transitions we use the anoxia proxy Mo (calibrated XRF-scanning data) from core 64PE406E1 (Hennekam et al., 2020b) (Fig. 3). We use the original dataset provided; interpolation is not required, as the spacing of datapoints is almost regular. We adapt the following analysis timeframes (see Table 3) and compare our results to the generic EWS from Hennekam et al. (2020b). As all sapropel timeseries are shorter than 500 datapoints we exclusively apply the length 500 classifier.

Table 3: Beginning and end dates of analysed sapropel events (based on Hennekam et al., 2020b) and number of data points in the respective time series.

Event	Transition [ka]	Analysis timeframe [ka]	Number of data points (and applied classifier)
S9	240.378	250.318-240.378	201 (length 500 classifier)
S8	225.039	234.923-225.039	232 (length 500 classifier)
S7	198.533	208.494-198.533	289 (length 500 classifier)
S6	177.308	187.225-177.308	211 (length 500 classifier)
S5	128.393	138.361-128.393	324 (length 500 classifier)
S4	107.823	117.818-107.823	435 (length 500 classifier)
S3	85.826	95.823-85.826	376 (length 500 classifier)

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

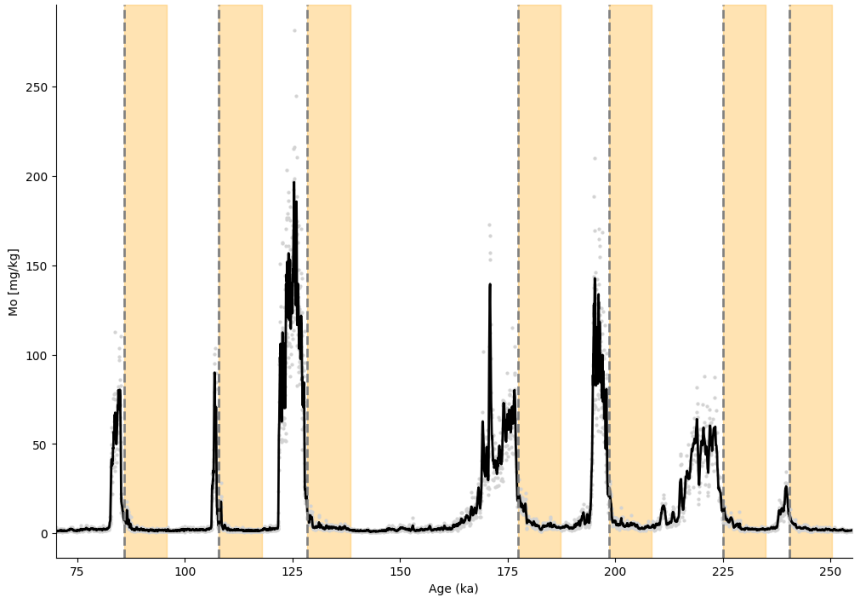


Figure 3: Eastern Mediterranean sapropel record of the anoxia proxy Mo with analysed S3 to S9 sapropel events marked with dashed lines, areas shaded in yellow indicate the analysed time intervals prior to the events.

For analysis of the results we adapted a version of ewstools (Bury, 2023; Bury et al., 2021) for our newly developed binary classifier. The script takes timeseries as inputs, which are detrended with the same filter (Lowess filter of span 0.2) as the synthetic timeseries the classifier was trained on (Dablander and Bury, 2022). We then apply our binary classifiers using the apply_classifier_inc() function defined in ewstools with an inc value of 10. This calculates DL probabilities for the input timeseries as they are revealed stepwise in time intervals of 10 until they reach the critical transition. Results are plotted as DL probability over time up to the transition. Individual DL probability graphs are calculated for the five classifiers of model 1 and the five classifiers of model 2. Ensembled results are plotted as the average DL probabilities of all models at a specific timestep. A fold or null transition is assigned if the DL probability output of the ensembled models is >0.5 respectively. We visualise these results as heatmaps, in which we plot the DL probabilities of the dominant transition type (fold or null) at the last timestep before the transition. Results are shown for variables ($\delta^{13}C$ and $\delta^{18}O$ isotope values, Mo concentrations) across the investigated events.

2.4 OOD detection methods

To evaluate how confident our classifier is in detecting catastrophic fold bifurcations and how many transcritical bifurcations are misclassified as fold, we apply OOD detection methods. We use two approaches, (i) the classical softmax confidence score approach (Hendrycks and Gimpel, 2016) and (ii) a new energy-based approach (Liu et al., 2020).

The softmax confidence scores are the same as the aforementioned DL probability outputs, that are directly derived from the application of the classifier. These values represent the model's prediction confidence across the fold and null

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

classes for an individual timeseries. We ran tests with the softmax probability to investigate how overconfident our classifier is in its results. We applied our ensembled length 500 binary classifiers (five model 1 and five model 2 classifiers) on 100 synthetically generated transcritical timeseries from the test dataset (see Bury et al., 2021). Our classifiers were only trained on null and fold bifurcations and had no previous contact with transcritical timeseries during training. We investigated how many of these transcritical timeseries get erroneously classified as fold bifurcations by our binary classifier.

We also applied an energy-based out-of-distribution (OOD) detection approach (Liu et al., 2020) as a posthoc method to distinguish if timeseries, that were originally classified as fold by our model are true fold bifurcations or misclassified transcritical bifurcations. Energy-based OOD detection has been shown to be less susceptible to overconfidence than the softmax confidence score approach and better at identifying OOD samples in image detection applications (Liu et al., 2020). We adapted this energy-based OOD detection method originally developed for image networks and implement it for timeseries. Energy-based OOD detection relies on energy scores calculated from an energy function. The energy function takes the raw logits outputs from a classifier and combines them into a scalar, that measures how strongly the model recognizes the input timeseries as belonging to any known class (the classifier was trained on). The higher the calculated energy score, the more confident the classifier is about this timeseries being in-distribution (ID) (representing a bifurcation class known to our classifier, in this case fold bifurcation). The lower the calculated energy score, the more likely it is out-of-distribution (OOD) (representing a bifurcation class unknown to our classifier, in this case transcritical). The resulting energy scores are plotted in an energy density histogram, showing the distribution of energy score values of ID and OOD timeseries (see Fig. 5).

We calculate such energy density histograms for our length 500 and length 1500 binary classifiers. We use an ensemble of five binary classifiers (model 2) for length 500 as well as length 1500. Our ensembles only contain model 2, as here we focus on determining the type of bifurcation (fold vs transcritical) and not generally detecting early warning (which model 1 works best for). In addition, the empirical data we later apply the OOD detection to, is also in the form of the input data for model 2, i.e. timeseries going right up to the bifurcation. As input for our energy density histograms we take fold and transcritical timeseries from the test dataset of length 500 (1250 timeseries each) and length 1500 (500 timeseries each) (see Bury et al., 2021). For these test dataset timeseries their bifurcation type is known, so the resulting energy density histograms represent the energy score distributions of ID (fold) and OOD (transcritical) timeseries for the length 500 and length 1500 classifier respectively (see Fig. 5). An energy threshold is set, that divides ID and OOD energy scores. We chose a threshold value of 2.9 (length 500) and 3.85 (length 1500) respectively (yellow lines in Fig. 5), as the number of correctly identified ID and OOD timeseries was equal for these energy scores.

We then compare energy scores of our empirical timeseries (identified as fold by our binary classifier) to our energy density histograms. For empirical timeseries shorter than 500 points and 1500 datapoints respectively, the timeseries were padded with zeros at the front. All empirical timeseries were detrended using a Lowess filter of span 0.2 before analysis to match the preprocessing of the test dataset timeseries (used as a basis for the energy density histograms). If the energy score values calculated for the empirical timeseries fall above the energy threshold, i.e. within the ID (fold)

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

part of the energy density histogram, they constitute true fold bifurcations; otherwise they are flagged as misclassified transcritical bifurcations.

To test for noise as the underlying cause of the bimodal nature of the ID and OOD energy density histograms (see Fig. 5), we calculated the standard deviation of the underlying test dataset timeseries. We applied Kruskal Wallis tests to determine statistically significant differences in the standard deviations (noise levels) of ID and OOD timeseries above and below the energy score threshold.

3 Results

3.1 Performance of the binary DL classifier on the synthetic test data

We evaluated the performance of our binary classifiers of length 500 (Fig. 4A and B) and length 1500 (Fig. 4C and D). The length 500 classifier achieved 86.3% precision, 85.6% recall, and an F1 score of 85.5% (Table 4). The length 1500 classifier performed slightly better (87.8% precision, 87.5% recall, F1 score 87.4%, see Table 4). Both our binary length 500 and length 1500 classifiers are able to correctly classify most timeseries, with a relatively low and balanced amount of false negatives and false positives. We compare the performance metrics of our binary classifiers with multi-class classifiers (Bury et al., 2021) (see Table 4). For shorter timeseries (length 500), our binary classifier outperformed the multiclass alternative. However, for length 1500 timeseries the multiclass classifier achieved slightly better performance values.

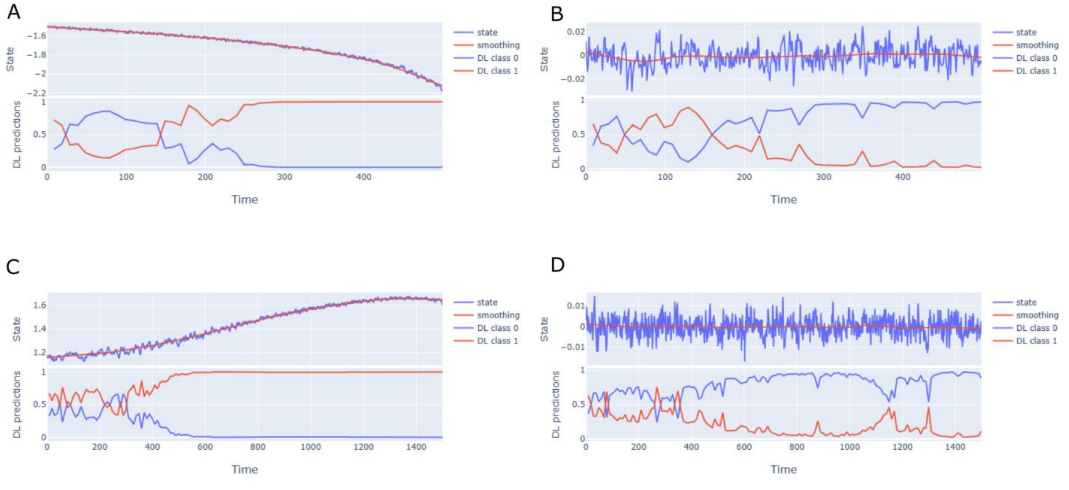


Figure 4: Examples of analyses with the binary classifier ensemble of length 500 on a synthetic fold timeseries (A) and synthetic null timeseries (B), and with the binary classifier ensemble of length 1500 on a synthetic fold timeseries (C) and a synthetic null timeseries (D); with state as the original timeseries, DL class 0 as average of ensembled model results for null bifurcation and DL class 1 as average of ensembled model results for fold bifurcation.

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

Table 4: Performance metrics of our binary classifier (length 500 and length 1500) compared to performance metrics multiclass classifiers (length 500 and length 1500) (Bury et al., 2021)

	Binary classifier (length 500)	Binary classifier (length 1500)	Multiclass classifier (length 500) (Bury et al., 2021)	Multiclass classifier (length 1500) (Bury et al., 2021)
Precision	86.3%	87.8%	84.4%	88.3%
Recall	85.6%	87.5%	84.2%	88.3%
F1 score	85.5%	87.4%	84.2%	88.2%

3.2 OOD detection on synthetically generated test data

We test different types of OOD detection methods in order to identify transcritical bifurcations, that are misclassified as fold bifurcations. The softmax confidence score method reveals that 76% of synthetically generated transcritical timeseries are misclassified as fold by our binary classifier (see supplementary data). To account for this high number of misclassification we apply an energy-based OOD detection approach. We calculate energy density histograms for our length 500 and length 1500 binary classifier (Fig. 5) (for more details see the material and methods section). Energy-based OOD detection allows for a clear distinction between the energy scores of ID (fold) and OOD (transcritical) timeseries for length 500 (Fig. 5A), and to a lesser degree also for length 1500 (Fig. 5B). This leads to about 68% (length 500) and 66% (length 1500) of fold and transcritical timeseries being correctly identified based on their energy scores, thus reducing misclassification to 32% (length 500) and 34% (length 1500) with our energy-based OOD detection method. However, there remains a certain overlap between ID and OOD in the energy density histogram due to the bimodal nature of the curves, more pronounced in length 1500 (Fig. 5B) than length 500 (Fig. 5A).

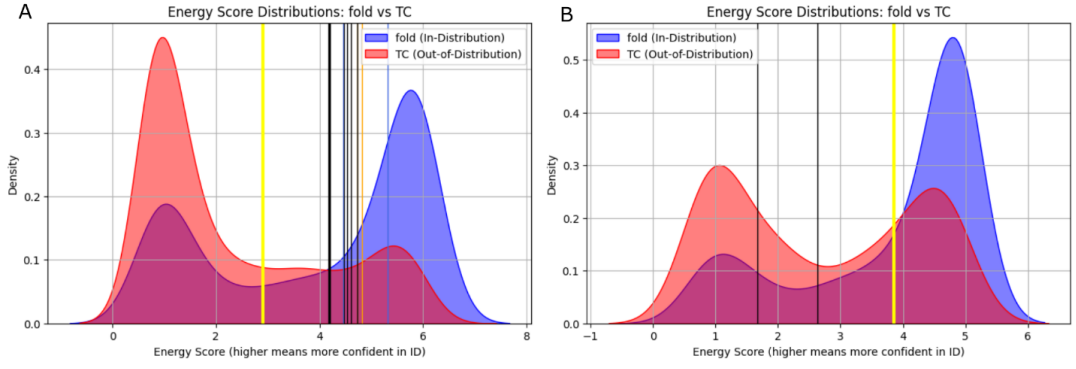


Figure 5: Energy density histograms (A) based on the length 500 test dataset of synthetically generated transcritical (TC) timeseries (OOD) and fold timeseries (ID), the yellow line indicates the threshold between OOD and ID, black lines mark energy scores of empirical sapropel timeseries, blue lines empirical ETM2

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

timeseries and orange lines empirical Oligocene/Miocene timeseries (see below); (B) based on the length 1500 test dataset of synthetically generated timeseries, the yellow line indicates the threshold between OOD and ID, black lines mark energy scores of empirical PETM timeseries (see below).

We tested for noise as the underlying cause of this bimodality, using standard deviation as a measure of noise. For both length 500 and length 1500, timeseries with higher energy scores (above the energy score threshold) are associated with higher median noise values compared to the ones with lower energy scores (below the energy threshold) (Table 5). High energy scores and thus confidence in identification generally seem to be associated with timeseries with higher noise for both ID (fold) and OOD (transcritical) timeseries.

Table 5: Mean noise levels of ID and OOD timeseries of different lengths above and below the energy threshold and results of Kruskal-Wallis test to determine significance of noise differences between the groups.

	Mean noise (std) above energy threshold	Mean noise (std) below energy threshold	p-value Kruskal-Wallis test
ID (length 500)	0.014	0.011	p-value = 5.6×10^{-25}
OOD (length 500)	0.016	0.012	p-value = 3.6×10^{-17}
ID (length 1500)	0.015	0.011	p-value = 3.6×10^{-09}
OOD (length 1500)	0.018	0.013	p-value = 7.5×10^{-12}

3.3 Application of the binary classifier to Cenozoic proxy record

3.3.1 Paleocene/Eocene isotope record (PETM, ETM2, ETM3)

The PETM, ETM2 and ETM3 events all show pronounced negative $\delta^{13}C$ as well as $\delta^{18}O$ excursions indicating strong carbon cycle perturbations and warming (see Fig. 1). The PETM is considered the strongest of the three events, followed by ETM2 and ETM3. We applied our binary DL classifier to these Paleocene/Eocene isotope events. Fig. 6 summarises the DL predictions and probabilities for the $\delta^{13}C$ and $\delta^{18}O$ isotope records of the PETM, ETM2 and ETM3 events. For the PETM our length 1500 classifier predicts fold bifurcations in both the $\delta^{13}C$ (0.99) and $\delta^{18}O$ (0.63) isotope records, indicating an approaching tipping point in the carbon cycle, as well as in temperature (albeit with a weaker signal). This supports the view of the PETM as an event mainly driven by a catastrophic tipping point in the carbon cycle. However, applying our length 500 binary classifier returns null predictions in both isotope records (0.87 for $\delta^{13}C$ and 0.62 for $\delta^{18}O$), indicating that there is no fold bifurcation approaching. For the ETM2 (analysed with the length 500 classifier) our classifier predicts fold bifurcations for both isotope records (0.67 for $\delta^{13}C$ and 0.75 for $\delta^{18}O$), indicating approaching tipping points. Our results support the ETM2 as a tipping point event driven by rapid changes in both the carbon cycle and temperature. The ETM3 event (analysed with the length 500 classifier) returns null predictions (0.88 for both $\delta^{13}C$ and $\delta^{18}O$), both with high DL probabilities, indicating no approaching bifurcations. This corroborates ETM3 being the weakest of the Early Eocene events, which didn't go through a catastrophic tipping point.

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

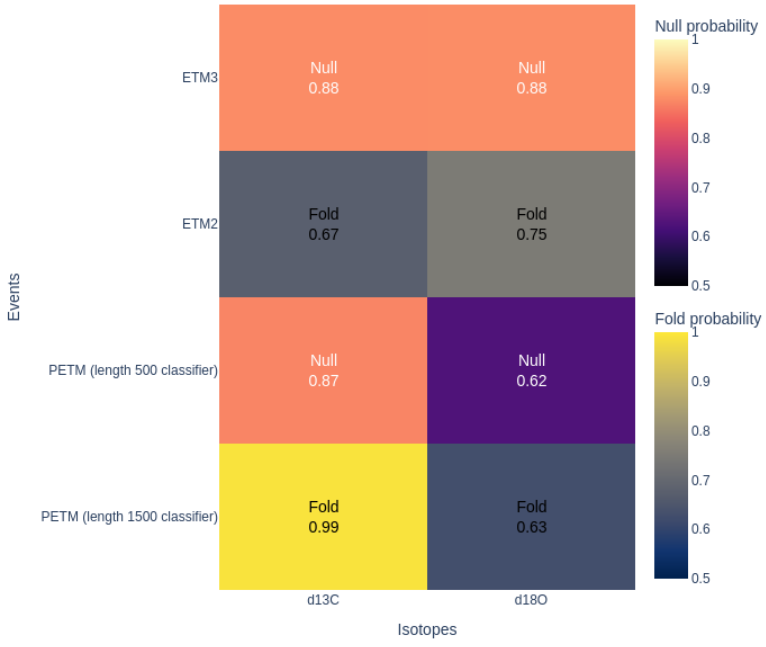


Figure 6: Heatmap of DL classifier predictions and probabilities for $\delta^{13}C$ and $\delta^{18}O$ isotope records of the PETM, ETM2 and ETM3 events, including results for different classifier lengths.

3.3.2 CENOGRID isotope record

The Eocene/Oligocene Transition, the Oligocene/Miocene Transition and the Middle Miocene Transition all show positive $\delta^{13}C$ as well as $\delta^{18}O$ excursions indicating general cooling and ice volume increase (see Fig. 2). Fig. 7 summarises our classifier's DL predictions and probabilities for the $\delta^{13}C$ and $\delta^{18}O$ isotope records of the analysed events. For the Eocene/Oligocene Transition the 500 length classifier returns null predictions for both the $\delta^{13}C$ (0.87) and $\delta^{18}O$ records (0.83), with high DL probabilities. The Eocene/Oligocene Transition is considered the first onset of the Antarctic glaciation, however no tipping point was crossed, the system did not yet transition directly into a fully glacial state (Galeotti et al., 2016). For the Oligocene/Miocene Transition our 500 length classifier returns fold predictions for both isotope records (0.75 for $\delta^{18}O$ and 0.51 for $\delta^{13}C$), pointing to substantial perturbations of the temperature record (and to a lesser degree also the carbon cycle). We link this to a tipping point being crossed in the Antarctic ice sheet subsystem, locking the system into a cool state with a stable Antarctic icesheet. However, when

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

applying our 1500 length classifier, null predictions are returned, albeit with low DL probabilities (0.56 for $\delta^{13}C$ and 0.64 for $\delta^{18}O$). For the Middle Miocene Transition both our 500 length and 1500 length classifier return null predictions for both isotope records, all with high DL probabilities (500 length classifier: 0.90 for $\delta^{13}C$ and 0.92 for $\delta^{18}O$, 1500 length classifier: 0.84 for $\delta^{13}C$ and 0.92 for $\delta^{18}O$), indicating no tipping points associated with the Middle Miocene records. During the Middle Miocene Climate Transition temperature continued to decrease, the ice sheets stabilized and the system got fully locked in an icehouse mode, but without crossing another tipping point.

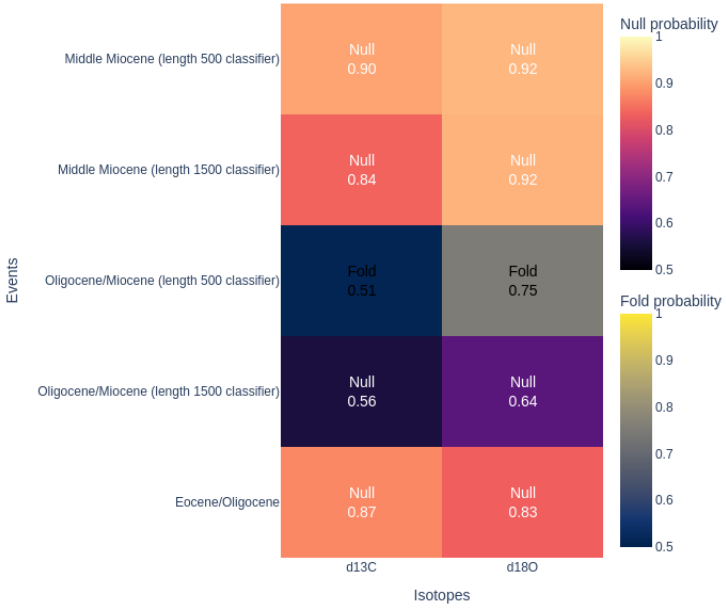


Figure 7: Heatmap of DL classifier predictions and probabilities for δ13C and δ18O isotope records of the Eocene/Oligocene Transition, Oligocene/Miocene Transition and Middle Miocene Climate Transition, including results for different classifier lengths.

3.3.3 Eastern Mediterranean sapropel data

Our binary fold classifier was applied to the Eastern Mediterranean sapropel record of S3 to S9 from core 64PE406E1. Fig. 8 summarises the classifier's DL predictions and probabilities for the Mo proxy records of the analysed events. For all sapropel timeseries (analysed with the length 500 classifier) fold bifurcations are predicted. DL probabilities are generally high (>0.95), except for the weakest sapropel events S3 (0.89) and S9 (0.87). This cleary indicates a tipping point from oxic to anoxic conditions associated with all investigated Eastern Mediterranean sapropel layers.

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

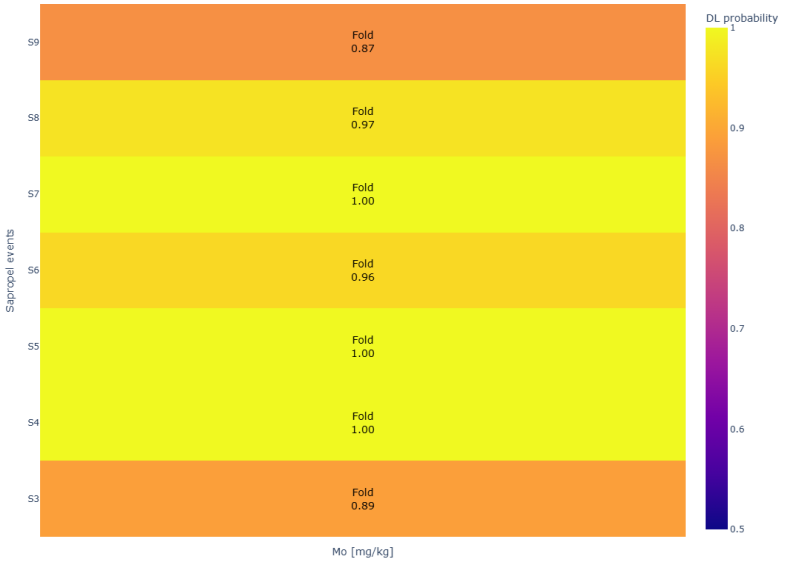


Figure 8: Heatmap of DL classifier predictions and probabilities for the Mo anoxia proxy record of various Eastern Mediterranean sapropel events.

3.6 OOD detection in empirical timeseries data

We subsequently applied our energy based OOD detection approach to the empirical timeseries identified as fold bifurcations by our binary classifier to flag any potentially misclassified transcritical timeseries. We evaluate empirical timeseries with length 500 or shorter, i.e. the sapropel timeseries, ETM2 timeseries ($\delta^{13}C$ and $\delta^{18}O$ isotope record) and Oligocene/Miocene timeseries ($\delta^{13}C$ and $\delta^{18}O$ isotope record), against the length 500 energy density histogram. The calculated energy scores of all the investigated empirical timeseries fall above the energy score threshold and thus within the ID part of the energy score distribution (black, orange and blue vertical lines in Fig. 5A). The sapropel Mo timeseries (S3 to S9), the ETM2 timeseries ($\delta^{13}C$ and $\delta^{18}O$ isotope record) and the Oligocene/Miocene timeseries ($\delta^{13}C$ and $\delta^{18}O$ isotope record) were all correctly identified by our binary classifier as genuine fold bifurcations. We evaluated longer empirical timeseries of up to 1500 data points, i.e. the PETM timeseries ($\delta^{13}C$ and $\delta^{18}O$ isotope record), against the length 1500 based energy density histogram. The calculated energy scores for both the PETM $\delta^{18}O$ and $\delta^{13}C$ timeseries fall below the energy threshold, flagging both PETM timeseries as misclassified transcritical bifurcations instead of a genuine fold bifurcation.

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

4 Discussion

4.1 Performance of our binary classifier with OOD in comparison to other EWS methods

We introduce a new binary classifier focused on the detection of fold bifurcations in Earth systems. Fold bifurcations are the most dominant transitions in these systems and are widely associated with Earth system tipping points (e.g. in the AMOC (Rahmstorf, 1995; Van Westen et al., 2025; Stommel, 1961) or the polar ice sheets (Robinson et al., 2012; Garbe et al., 2020)). Our new classifier is able to confidently identify fold bifurcations in Earth system applications, as shown by our empirical test cases. This constitutes a significant improvement over generic EWS (autocorrelation and variance based), as it allows to quantitatively detect approaching critical thresholds and identify their bifurcation type instead of providing only qualitative indication of an otherwise unspecified approaching transition.

Compared to other already exiting multiclass classifiers (Bury et al., 2021), our binary classifier also shows improved performance (see Table 4). It requires significantly less computational power and time to train, as it only contains two classes instead of several. For the length 500 classifier its performance metrics are better than the multiclass version (see Table 4), owing to its more focused nature. The specific training only on fold bifurcation and null timeseries prevents a 'dilution' of the classifier's performance from the overlap of features of several different bifurcation types. For the length 1500 classifier version the multiclass approach shows better performance indicating that longer timeseries might benefit from classifiers that were trained on more than two classes. The disadvantage of the multiclass models at shorter lengths potentially balances out at length 1500 due to learning from more representations, that allow for better separation. However, as the geological record tends to be rather incomplete and a majority of geological high-resolution timeseries are relatively short (less than 500 datapoints), our binary classifier constitutes a valuable new tool for tipping point detection tailored to this kind of geological records.

Another advantage of our binary classifier over the existing multiclass classifiers is its ability to flag misclassified transcritical bifurcations using an energy-based OOD detection approach. Transcritical and fold bifurcations are mathematically very similar (Kuznetsov, 1998) and thus, current DL frameworks have issues picking up on their subtle differences, often misclassifying transcritical bifurcations as fold (see our softmax confidence score results). This has consequences for the interpretation of how abrupt and reversible an investigated transition is, an important characteristic especially when studying and trying to mitigate Earth system tipping points. Transcritical bifurcations are mostly gradual and easily reversible (Kuznetsov, 1998); misclassifying them as fold bifurcations leads to a high rate of 'false alarms', in which timeseries are erroneously flagged as abrupt and irreversible tipping points. The energy-based OOD detection approach implemented with our binary classifier is able to reduce the misclassification rate to about 32% (length 500) and 34% (length 1500). This is a significant improvement over current multiclass classifiers, that have a significantly higher rate of misclassification of about 47% (Bury et al., 2021). Thus, considering its performance gain and lower rates of misclassification, our binary classifier together with the OOD approach constitute an improvement on already existing DL multiclass classifiers.

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

4.2 Comparison of generic EWS and our binary classifier for the Cenozoic proxy records

We compare the results of our binary classifier with generic EWS from the literature for the different Cenozoic proxy records, including the Paleocene/Eocene isotope record, the CENOGRID isotope record and the Eastern Mediterranean sapropel record. Generic EWS from the literature pick up on several critical transitions in these records related to major events associated with excursions in carbon and oxygen isotopes (Boettner et al., 2021; Setty et al., 2023) or the anoxia proxy Mo (Hennekam et al., 2020a). We use our binary classifier to identify which of these transitions are fold bifurcations and thus constitute an abrupt tipping point.

4.2.1 The Paleocene/Eocene isotope record

The PETM, ETM2 and ETM3 events are characterised by negative $\delta^{13}C$ as well as $\delta^{18}O$ excursions, that indicate strong carbon cycle perturbations and warming (see Fig. 1), with the PETM being considered the strongest, followed by the ETM2 and ETM3. For the PETM and ETM2 events generic EWS (autocorrelation and variance based) show a rising trend (both autocorrelation and variance) in the $\delta^{13}C$ record, for the $\delta^{18}O$ isotope record only autocorrelation shows a rising trend (Setty et al., 2023). A rise in variance and autocorrelation strongly implies an approaching transition (associated with critical slowing down), a rise in autocorrelation only does not constitute a reliable bifurcation indicator (Scheffer et al., 2009; Ditlevsen and Johnsen, 2010). The generic EWS results indicate approaching transitions for both the PETM and ETM2 in the carbon isotope record, with an ambiguous signal for the oxygen isotope record. For the PETM, our binary classifier length 1500 identifies fold bifurcations (tipping points) in both isotope records, with a strong signal in $\delta^{13}C$ and to a lesser degree also in $\delta^{18}O$. However, our length 500 classifier does not recover any fold bifurcations for the PETM. Depending on the choice of classifier, there is evidence for the PETM event being driven by a tipping point in the carbon cycle (with the rise in temperature being a consequence rather than a driver of the perturbation). Using our binary classifier, the ETM2 shows approaching fold bifurcations in both the $\delta^{13}C$ and $\delta^{18}O$ isotope records, both with low to medium DL probabilities. This supports the ETM2 as being driven by perturbations in the carbon cycle and temperature like the PETM, but with lower overall intensity. For the ETM3 generic EWS results for both the $\delta^{13}C$ and $\delta^{18}O$ isotope records show no significant increase in autocorrelation and variance (except for a rise in autocorrelation in the $\delta^{18}O$ record), implying no critical transitions associated with the ETM3 event (and an ambiguous signal for the oxygen isotope record). Using our binary classifier we also do not recover any fold bifurcation predictions; the weak ETM3 event is not associated with any tipping point. Our binary classifier generally corroborates previous generic EWS results (for the PETM depending on the classifier length), that indicate approaching transitions in the Paleocene/Eocene, and is able to identify these critical transitions as fold bifurcations (tipping points).

4.2.2 The CENOGRID isotope record

The Eocene/Oligocene Transition, the Oligocene/Miocene Transition and the Middle Miocene Transition are associated with positive $\delta^{13}C$ as well as $\delta^{18}O$ excursions indicating a general cooling of the Earth system and ice volume increase (see Fig. 2). For the Eocene/Oligocene Transition generic EWS do not indicate any combined rising trends in autocorrelation and variance of the $\delta^{13}C$ and $\delta^{18}O$ record (only rising autocorrelation alone indicating an ambiguous signal in these records) (Boettner et al., 2021). Our DL classifier returns null predictions for both isotope records,

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

corroborating the results of the generic EWS. For the Oligocene/Miocene Transition, generic EWS indicate critical transitions in the $\delta^{13}C$ and $\delta^{18}O$ records based on rising trends in autocorrelation and variance for both records. Our binary classifier length 500 also picks up on fold bifurcations in both isotope records with a higher DL probability for the $\delta^{18}O$ record. However, using our length 1500 we do not recover any fold bifurcations. Based on the choice of classifier, there is supportive evidence for the system going through a rapid and catastrophic tipping point at the Oligocene/Miocene Transition driven mainly by perturbations in temperature (likely related to the development of a stable Antarctic ice sheet). For the Middle Miocene Climate Transition, generic EWS do not return any combined rising trends in autocorrelation and variance of the $\delta^{13}C$ and $\delta^{18}O$ record (except for a rise in the $\delta^{13}C$ autocorrelation), indicating no approaching critical transitions (except for an ambiguous signal in the carbon isotope record). Our binary classifier agrees with this by returning null predictions for both isotope records. Generally, our binary classifier corroborates generic EWS results from the literature (for the Oligocene/Miocene transition depending on the choice of classifier), that predict approaching transitions in the CENOGRID record, and is able to additionally identify these transitions as fold bifurcations (tipping points).

4.2.3 Comparison of DL classifier results and generic EWS results from the Eastern Mediterranean sapropel record

To assess the oxic-to-anoxic transitions associated with Mediterranean sapropels, the anoxia proxy Mo is used (see Hennekam et al., 2020a). Generic EWS identify increases in variance of the Mo records prior to all anoxic events, all within the 95% confidence level. Increases in autocorrelation are also found in all Mo records (with only the autocorrelation results for S4 and S8 being within the 90% confidence level). Generic EWS indicate critical transitions in all analysed sapropel events based on combined rising trends in variance and autocorrelation. Our binary classifier supports this by returning fold predictions for all sapropels with consistently high DL predictions indicating that sapropel events were actual tipping point events.

4.3 Factors influencing classifier performance and results

The performance of our DL classifiers on our empirical study systems can be influenced by a variety of factors. Complex multivariate systems like the Paleocene/Eocene record and the CENOGRID record are more difficult to analyze and can deliver conflicting or ambiguous DL results. They are influenced by various underlying coupled processes, changing climatic boundary conditions and forcings (e.g. atmospheric CO_2 concentrations and existence of polar ice sheets; see Westerhold et al., 2020). In addition, the CENOGRID record represents a composite record, based on several cores from different latitudes and ocean basins. Every record from this composite could be influenced by different forcings leading to varying tipping dynamics; combining these records might lead to ambiguous results in our DL classifier. Generally, it seems that relatively simple systems with only few forcings like the Mediterranean sapropel record (that is mostly driven by sea surface productivity and water column ventilation in a restricted basin; see Rohling et al., 2015) tend to deliver more clear and unambiguous results supported by high DL probabilities. Our analyses show, that results for timeseries of complex systems can be sensitive to the chosen classifier and the length of the time interval studied prior to the bifurcation, which might lead to different results for the same timeseries analysed with different classifier lengths. For example, for the PETM our binary classifier length 1500 returns fold

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

bifurcation predictions for both isotope records, whereas our binary classifier length 500 does not predict any approaching transitions. In this case, the early warning and resilience loss of the system might have started long before the bifurcation. Shortly before the bifurcation, other additional mechanisms than gradual resilience loss (which our DL classifiers is trained to detect) could have been at play, ultimately pushing the system beyond its tipping point without detectable early warning signals in the last 500 datapoints. Alternatively, the drivers of resilience loss related to the PETM simply acted on longer timescales with lower frequencies and therefore need a longer classifier to be picked up. We observe the opposite situation with the Oligocene/Miocene Transition, for which our length 1500 classifier returns null predictions, whereas our length 500 classifier predicts fold bifurcations for both isotope records. Other processes than gradual resilience loss (that our DL classifier is trained to detect) might act in the earlier parts of the timeseries and obscure the signal leading to null predictions from our length 1500 DL classifier. Generally, our analyses indicate, that there does not exist a 'one size fits all' classifier for all use cases, especially when considering complex multivariate systems. A process-based understanding of the study system aids with the choice of classifier length. This includes being aware of which forcings act on the investigated system and if they change over time. Through the right choice of classifier length it is possible to include the forcings relevant for the tipping of the system. Additionally, it is also important to have knowledge about the timeframes on which the relevant forcings act, in order to choose a suitable classifier length capturing the relevant processes.

4.4 Evaluation of OOD detection methods

4.4.1 OOD analyses on synthetically generated datasets

Our softmax confidence score analyses reveals that 76% of the synthetically generated transcritical timeseries are classified as fold bifurcations by our binary classifier. This indicates a high degree of overconfidence of our DL classifier and a high rate of transcritical timeseries being misclassified as fold. Compared to a misclassification rate of 47% of already existing multiclass classifiers (Bury et al., 2021), this is comparably high and warrants alternative mitigation strategies. Thus, we implemented our energy-based out-of-distribution (OOD) detection approach. This approach is successful in separating in-distribution (fold) and out-of-distribution (transcritical) timeseries, especially in the length 500 energy score diagram. Compared to using the softmax confidence scores, energy-based OOD detection reduces the percentage of transcritical timeseries wrongly identified as fold bifurcations significantly from 76% to about 32% (length 500) and 34% (length 1500) respectively. This is also a significant improvement on the 47% misclassification rate of already existing multiclass classifiers (Bury et al., 2021). We therefore propose the energy-based OOD detection approach in combination with our binary classifier (especially the length 500 energy score diagram and classifier) as an improved tool for the detection of fold bifurcations and flagging of potentially misclassified transcritical transitions. The length 1500 energy score diagram does not show such a clear distinction between ID and OOD as the length 500 energy score diagram (Fig. 5B). This might be caused by the smaller number of synthetically generated timeseries it is based on; only 500 fold and transcritical timeseries each for length 1500 compared to 1250 timeseries each for length 500. This number was limited by the amount of the synthetically generated timeseries in the original test dataset. This could potentially be improved upon in the future by generating more length 1500 test timeseries, in order to better separate ID and OOD in the length 1500 energy score diagram. However, as the

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

geological record is often incomplete and available timeseries rather short (below 500 datapoints), we focus on the length 500 classifier and energy score diagram here.

We also identified noise as a potential cause for the bimodality of the ID and OOD energy score curves. There is a significant difference in the noise levels, with timeseries with higher energy scores (above the energy threshold) showing higher noise than the ones with lower energy scores (below the energy threshold). This means, that a higher confidence in identification (indicated by higher energy scores) coincides with higher noise in the timeseries. Generally, the detection of early warning signals is linked to the 'noise' portion of the timeseries. More stochastic perturbation (noise) means that a system explores a larger portion of its local potential landscape (Scheffer et al., 2009; Dakos et al., 2008) making it easier for a classifier to determine the type of bifurcation, thus being more confident and returning higher energy scores. Low confidence energy scores are generally associated with less noisy data. In our case, the binary classifier is less confident in identifying fold timeseries with little noise (below the energy threshold) resulting in them being misclassified as transcritical. On the other hand, exceptionally noisy OOD transcritical timeseries (above the energy threshold) are more similar to the ID fold bifurcations and have a higher chance of getting misclassified as such. These findings indicate that our binary classifier is generally better at identifying fold bifurcations, if they contain more noise. This can be an advantage over generic EWS approaches, that are typically not well suited to handle noisy timeseries (Perretti and Munch, 2012; Dakos et al., 2015). Many empirical Earth systems are open 'noisy' systems, governed by many (unknown) forcings and feedbacks (Dakos et al., 2015). Thus, for detecting critical transitions in Earth systems with inherently higher amounts of noise, our new binary classifier provides a powerful alternative to existing generic EWS.

4.4.2 OOD analyses on empirical datasets

Our energy-based OOD detection method to flag misclassified transcritical bifurcations identified all length 500 empirical timeseries previously classified as fold as genuine fold bifurcations. However, the length 1500 PETM timeseries are flagged as misclassified transcritical bifurcations. As mentioned above, the length 1500 energy score diagram is based on significantly less synthetically generated timeseries than the length 500 version and thus could yield misleading results due to lacking data basis. Another reason could be, that the empirical PETM timeseries contain only little amounts of noise and are therefore erroneously identified as OOD (transcritical) by our energy-based OOD detection method. The empirical case of the PETM needs further investigation and testing to be able to unambiguously identify whether this event was an actual catastrophic tipping point (fold bifurcation).

5 Conclusions

We developed and successfully tested a new deep learning classifier, focused on the detection of fold bifurcations in the Earth system. Our binary classifier shows clear performance gains in comparison to generic EWS as well as existing DL classifiers. We additionally implemented an energy-based OOD detection method that is able to significantly reduce the amount of misclassified transcritical bifurcations. Comparing our DL approach with the results of previous studies (using generic EWS) we are generally able to corroborate the existence of previously identified critical transitions. We

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

add to this by identifying most of the critical transitions from the empirical datasets as genuine fold bifurcation tipping points using our binary classifier paired with energy-based OOD detection. There's scope for future investigation into how exactly factors like classifier length or noise levels influence the results. Nevertheless, our binary classifier together with the energy-based OOD detection method provides a powerful new tool for the reliable detection of tipping points geared specifically towards Earth system applications.

Code and data availability

All data and code, as well as instructions for reproducing the results and figures in this manuscript are available on Zenodo (Grohganz et al., 2026). To facilitate practical adoption, we will release an open-source implementation of our classifier and out-of-distribution (OOD) detection method, along with detailed usage instructions and tutorials, in a public GitHub repository upon publication.

Author contribution

M.G.: investigation, formal analysis, methodology, software, data curation, validation, visualization, writing – original draft, writing – review and editing; T.M.B: methodology, software, data curation, writing – review and editing; B.v.d.B.: formal analysis, writing – review and editing; G.J.R.: conceptualization, formal analysis, writing – review and editing; R.H.: conceptualization, formal analysis, project administration, supervision, funding acquisition, writing – review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Competing interests

The authors declare that they have no conflict of interests.

Acknowledgements

This work was supported by the Netherlands Organisation for Scientific Research (NWO) through the Vidi grant (grant no. VI.Vidi.223.138) awarded to R.H. Additional support was provided by EMBRACER (Summit grant SUMMIT.1.034), also financed by NWO.

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

References

Armstrong McKay, D. I., Staal, A., Abrams, J. F., Winkelmann, R., Sakschewski, B., Loriani, S., Fetzer, I., Cornell, S. E., Rockström, J., and Lenton, T. M.: Exceeding 1.5°C global warming could trigger multiple climate tipping points, Science, 377, eabn7950, https://doi.org/10.1126/science.abn7950, 2022.

Arnold, V. I.: Geometrical methods in the theory of ordinary differential equations, Springer, New York2012.

Barnet, J. S. K., Littler, K., Westerhold, T., Kroon, D., Leng, M. J., Bailey, I., Röhl, U., and Zachos, J. C.: Late Cretaceous-Early Paleogene stable isotope and coarse fraction record of ODP Site 208-1262, PANGAEA [dataset], https://doi.org/10.1594/PANGAEA.884588, 2018.

Barnet, J. S. K., Littler, K., Westerhold, T., Kroon, D., Leng, M. J., Bailey, I., Röhl, U., and Zachos, J. C.: A High-Fidelity Benthic Stable Isotope Record of Late Cretaceous–Early Eocene Climate Change and Carbon-Cycling, Paleoceanography and Paleoclimatology, 34, 672-691, https://doi.org/10.1029/2019PA003556, 2019.

Boers, N.: Early-warning signals for Dansgaard-Oeschger events in a high-resolution ice core record, Nature Communications, 9, 2556, https://doi.org/10.1038/s41467-018-04881-7, 2018.

Boers, N.: Observation-based early-warning signals for a collapse of the Atlantic Meridional Overturning Circulation, Nature Climate Change, 11, 680-688, https://doi.org/10.1038/s41558-021-01097-4, 2021.

Boers, N. and Rypdal, M.: Critical slowing down suggests that the western Greenland Ice Sheet is close to a tipping point, Proceedings of the National Academy of Sciences, 118, e2024192118, https://doi.org/10.1073/pnas.2024192118, 2021.

Boettner, C., Klinghammer, G., Boers, N., Westerhold, T., and Marwan, N.: Early-warning signals for Cenozoic climate transitions, Quaternary Science Reviews, 270, 107177, https://doi.org/10.1016/j.quascirev.2021.107177, 2021.

Bury, T. M.: ewstools: a Python package for early warning signals of bifurcations in time series data, Journal of Open Source Software, 8, 5038, 2023.

Bury, T. M., Dylewsky, D., Bauch, C. T., Anand, M., Glass, L., Shrier, A., and Bub, G.: Predicting discrete-time bifurcations with deep learning, Nature Communications, 14, 6331, https://doi.org/10.1038/s41467-023-42020-z, 2023.

Bury, T. M., Sujith, R. I., Pavithran, I., Scheffer, M., Lenton, T. M., Anand, M., and Bauch, C. T.: Deep learning for early warning signals of tipping points, Proceedings of the National Academy of Sciences, 118, e2106140118, https://doi.org/10.1073/pnas.2106140118, 2021.

Crawford, J. D.: Introduction to bifurcation theory, Reviews of Modern Physics, 63, 991-1037, https://doi.org/10.1103/RevModPhys.63.991, 1991.

Dablander, F. and Bury, T. M.: Deep learning for tipping points: Preprocessing matters, Proceedings of the National Academy of Sciences, 119, e2207720119, https://doi.org/10.1073/pnas.2207720119, 2022.

Dakos, V., Carpenter, S. R., van Nes, E. H., and Scheffer, M.: Resilience indicators: prospects and limitations for early warnings of regime shifts, Philosophical Transactions of the Royal Society B: Biological Sciences, 370, 10.1098/rstb.2013.0263, 2015.

Dakos, V., Scheffer, M., van Nes, E. H., Brovkin, V., Petoukhov, V., and Held, H.: Slowing down as an early warning signal for abrupt climate change, Proceedings of the National Academy of Sciences, 105, 14308-14312, doi:10.1073/pnas.0802430105, 2008.

Deb, S., Sidheekh, S., Clements, C. F., Krishnan, N. C., and Dutta, P. S.: Machine learning methods trained on simple models can predict critical transitions in complex natural systems, Royal Society Open Science, 9, 211475, https://doi.org/10.1098/rsos.211475, 2022.

Ditlevsen, P. D. and Johnsen, S. J.: Tipping points: Early warning and wishful thinking, Geophysical Research Letters, 37, https://doi.org/10.1029/2010GL044486, 2010.

Drijfhout, S.: Competition between global warming and an abrupt collapse of the AMOC in Earth's energy imbalance, Scientific Reports, 5, 14877, https://doi.org/10.1038/srep14877, 2015.

Galeotti, S., DeConto, R., Naish, T., Stocchi, P., Florindo, F., Pagani, M., Barrett, P., Bohaty, S. M., Lanci, L., Pollard, D., Sandroni, S., Talarico, F. M., and Zachos, J. C.: Antarctic Ice Sheet variability across the Eocene-Oligocene boundary climate transition, Science, 352, 76-80, https://doi.org/10.1126/science.aab0669, 2016.

Garbe, J., Albrecht, T., Levermann, A., Donges, J. F., and Winkelmann, R.: The hysteresis of the Antarctic Ice Sheet, Nature, 585, 538-544, https://doi.org/10.1038/s41586-020-2727-5, 2020.

Grohganz, M., Bury, T. M., van der Bolt, B., Reichart, G.-J., and Hennekam, R.: An Earth system deep learning classifier for tipping point detection: supplementary data and code, Zenodo [dataset], https://doi.org/10.5281/zenodo.19629897, 2026.

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

Hawkins, E., Smith, R. S., Allison, L. C., Gregory, J. M., Woollings, T. J., Pohlmann, H., and de Cuevas, B.: Bistability of the Atlantic overturning circulation in a global climate model and links to ocean freshwater transport, Geophysical Research Letters, 38, https://doi.org/10.1029/2011GL047208, 2011.
Hendrycks, D. and Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks, arXiv preprint arXiv:1610.02136, 2016.
Hennekam, R., van der Bolt, B., van Nes, E. H., de Lange, G. J., Scheffer, M., and Reichart, G.-J.: Early-Warning Signals for Marine Anoxic Events, Geophysical Research Letters, 47, e2020GL089183, https://doi.org/10.1029/2020GL089183, 2020a.
Hennekam, R., Van der Bolt, B., Van Nes, E. H., de Lange, G. J., Scheffer, M., and Reichart, G.-J.: Calibrated XRF-scanning data (mm resolution) and calibration data (ICP-OES and ICP-MS) for elements Al, Ba, Mo, Ti, and U in Mediterranean cores MS21, MS66, and 64PE406E1, PANGAEA [dataset], https://doi.org/10.1594/PANGAEA.923197, 2020b.
Jackson, L. C., Kahana, R., Graham, T., Ringer, M. A., Woollings, T., Mecking, J. V., and Wood, R. A.: Global and European climate impacts of a slowdown of the AMOC in a high resolution GCM, Climate Dynamics, 45, 3299-3316, https://doi.org/10.1007/s00382-015-2540-2, 2015.
Kéfi, S., Guttal, V., Brock, W. A., Carpenter, S. R., Ellison, A. M., Livina, V. N., Seekell, D. A., Scheffer, M., van Nes, E. H., and Dakos, V.: Early Warning Signals of Ecological Transitions: Methods for Spatial Patterns, PLOS ONE, 9, e92097, https://doi.org/10.1371/journal.pone.0092097, 2014.
Kermack, W. O. and McKendrick, A. G.: A contribution to the mathematical theory of epidemics, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 115, 700-721, https://doi.org/10.1098/rspa.1927.0118, 1927.
Kriegler, E., Hall, J. W., Held, H., Dawson, R., and Schellnhuber, H. J.: Imprecise probability assessment of tipping points in the climate system, Proceedings of the National Academy of Sciences, 106, 5041-5046, https://doi.org/10.1073/pnas.0809117106, 2009.
Kuznetsov, Y. A.: Elements of applied bifurcation theory, Springer1998.
Lauretano, V., Zachos, J. C., and Lourens, L. J.: Benthic stable isotope record of ODP Site 208-1263, PANGAEA [dataset], https://doi.org/10.1594/PANGAEA.893894, 2018a.
Lauretano, V., Zachos, J. C., and Lourens, L. J.: Orbitally Paced Carbon and Deep-Sea Temperature Changes at the Peak of the Early Eocene Climatic Optimum, Paleoceanography and Paleoclimatology, 33, 1050-1065, https://doi.org/10.1029/2018PA003422, 2018b.
Lear, C. H., Bailey, T. R., Pearson, P. N., Coxall, H. K., and Rosenthal, Y.: Cooling and ice growth across the Eocene-Oligocene transition, Geology, 36, 251-254, https://doi.org/10.1130/g24584a.1, 2008.
Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J.: Tipping elements in the Earth's climate system, Proceedings of the National Academy of Sciences, 105, 1786-1793, https://doi.org/10.1073/pnas.0705414105, 2008.
Lenton, T. M., Milkoreit, M., Willcock, S., Abrams, J. F., Mc Kay, D. L. A., Buxton, J. E., Donges, J., Loriani, S., Wunderling, N., and Alkemade, F.: Global tipping points report 2025, 2025.
Liu, W., Wang, X., Owens, J., and Li, Y.: Energy-based out-of-distribution detection, Advances in neural information processing systems, 33, 21464-21475, 2020.
Milkoreit, M., Hodbod, J., Baggio, J., Benessaiah, K., Calderón-Contreras, R., Donges, J. F., Mathias, J.-D., Rocha, J. C., Schoon, M., and Werners, S. E.: Defining tipping points for social-ecological systems scholarship—an interdisciplinary literature review, Environmental Research Letters, 13, 033005, https://doi.org/10.1088/1748-9326/aaaa75, 2018.
Perretti, C. T. and Munch, S. B.: Regime shift indicators fail under noise levels commonly observed in ecological systems, Ecological Applications, 22, 1772-1779, https://doi.org/10.1890/11-0161.1, 2012.
Rahmstorf, S.: Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycle, Nature, 378, 145-149, https://doi.org/10.1038/378145a0, 1995.
Robinson, A., Calov, R., and Ganopolski, A.: Multistability and critical thresholds of the Greenland ice sheet, Nature Climate Change, 2, 429-432, https://doi.org/10.1038/nclimate1449, 2012.
Rohling, E. J., Marino, G., and Grant, K. M.: Mediterranean climate and oceanography, and the periodic development of anoxic events (sapropels), Earth-Science Reviews, 143, 62-97, https://doi.org/10.1016/j.earscirev.2015.01.008, 2015.
Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., and Sugihara, G.: Early-warning signals for critical transitions, Nature, 461, 53-59, https://doi.org/10.1038/nature08227, 2009.

https://doi.org/10.5194/egusphere-2026-2223
Preprint. Discussion started: 23 April 2026
© Author(s) 2026. CC BY 4.0 License.

Setty, S., Cramwinckel, M. J., van Nes, E. H., van de Leemput, I. A., Dijkstra, H. A., Lourens, L. J., Scheffer, M., and Sluijs, A.: Loss of Earth system resilience during early Eocene transient global warming events, Science Advances, 9, eade5466, https://doi.org/10.1126/sciadv.ade5466, 2023.

Sexton, P. F., Norris, R. D., Wilson, P. A., Pälike, H., Westerhold, T., Röhl, U., Bolton, C. T., and Gibbs, S.: Eocene global warming events driven by ventilation of oceanic dissolved organic carbon, Nature, 471, 349-352, https://doi.org/10.1038/nature09826, 2011.

Seydel, R. U.: Practical bifurcation and stability analysis, Springer2009.

Shevenell, A. E., Kennett, J. P., and Lea, D. W.: Middle Miocene Southern Ocean Cooling and Antarctic Cryosphere Expansion, Science, 305, 1766-1770, https://doi.org/10.1126/science.1100061, 2004.

Stommel, H.: Thermohaline Convection with Two Stable Regimes of Flow, Tellus, 13, 224-230, https://doi.org/10.1111/j.2153-3490.1961.tb00079.x, 1961.

Van Breedam, J., Huybrechts, P., and Crucifix, M.: Hysteresis and orbital pacing of the early Cenozoic Antarctic ice sheet, Climate of the Past, 19, 2551-2568, https://doi.org/10.5194/cp-19-2551-2023, 2023.

van Nes, E. H., Arani, B. M. S., Staal, A., van der Bolt, B., Flores, B. M., Bathiany, S., and Scheffer, M.: What Do You Mean, 'Tipping Point'?, Trends in Ecology & Evolution, 31, 902-904, https://doi.org/10.1016/j.tree.2016.09.011, 2016.

van Westen, R. M. and Dijkstra, H. A.: Asymmetry of AMOC Hysteresis in a State-Of-The-Art Global Climate Model, Geophysical Research Letters, 50, e2023GL106088, https://doi.org/10.1029/2023GL106088, 2023.

van Westen, R. M., Kliphuis, M., and Dijkstra, H. A.: Physics-based early warning signal shows that AMOC is on tipping course, Science Advances, 10, eadk1189, https://doi.org/10.1126/sciadv.adk1189, 2024.

van Westen, R. M., Vanderborght, E., and Dijkstra, H. A.: A saddle-node bifurcation may be causing the AMOC collapse in the Community Earth System Model, Earth Syst. Dynam., 16, 2063-2085, https://doi.org/10.5194/esd-16-2063-2025, 2025.

Westerhold, T.: Cenozoic global reference benthic carbon and oxygen isotope dataset (CENOGRID) PANGAEA [dataset], https://doi.org/10.1594/PANGAEA.917503, 2020.

Westerhold, T., Marwan, N., Drury, A. J., Liebrand, D., Agnini, C., Anagnostou, E., Barnet, J. S. K., Bohaty, S. M., De Vleeschouwer, D., Florindo, F., Frederichs, T., Hodell, D. A., Holbourn, A. E., Kroon, D., Lauretano, V., Littler, K., Lourens, L. J., Lyle, M., Pälike, H., Röhl, U., Tian, J., Wilkens, R. H., Wilson, P. A., and Zachos, J. C.: An astronomically dated record of Earth's climate and its predictability over the last 66 million years, Science, 369, 1383-1387, https://doi.org/10.1126/science.aba6853, 2020.

Wilson, G. S., Pekar, S. F., Naish, T. R., Passchier, S., and DeConto, R.: Chapter 9 The Oligocene–Miocene Boundary – Antarctic Climate Response to Orbital Forcing, in: Developments in Earth and Environmental Sciences, edited by: Florindo, F., and Siegert, M., Elsevier, 369-400, https://doi.org/10.1016/S1571-9197(08)00009-8, 2008.

Wissel, C.: A universal law of the characteristic return time near thresholds, Oecologia, 65, 101-107, https://doi.org/10.1007/BF00384470, 1984.

Wunderling, N., Donges, J. F., Kurths, J., and Winkelmann, R.: Interacting tipping elements increase risk of climate domino effects under global warming, Earth Syst. Dynam., 12, 601-619, https://doi.org/10.5194/esd-12-601-2021, 2021.

Zachos, J., Pagani, M., Sloan, L., Thomas, E., and Billups, K.: Trends, Rhythms, and Aberrations in Global Climate 65 Ma to Present, Science, 292, 686-693, https://doi.org/10.1126/science.1059412, 2001.

Zachos, J. C., Röhl, U., Schellenberg, S. A., Sluijs, A., Hodell, D. A., Kelly, D. C., Thomas, E., Nicolo, M., Raffi, I., Lourens, L. J., McCarren, H., and Kroon, D.: Rapid Acidification of the Ocean During the Paleocene-Eocene Thermal Maximum, Science, 308, 1611-1615, https://doi.org/10.1126/science.1109004, 2005.