

Probabilistic anticipation of AMOC transitions under critical forcing magnitudes and rates via deep learning

Wenjie Zhang^{a,b}, Yu Huang^{b,c},^{*} Sebastian Bathiany^{b,c}, Yechul Shin^d,
Maya Ben-Yami^{b,c}, Suiping Zhou^a, Niklas Boers^{b,c,e}

^a School of Aerospace Science and Technology, Xidian University, Xi'an, 710126, China

^b Earth System Modelling, School of Engineering and Design, Technical University of Munich, Munich, 85521, Germany

^c Complexity Science, Potsdam Institute for Climate Impact Research, Potsdam, 14473, Germany

^d School of Earth and Environmental Sciences, Seoul National University, Seoul, 08826, Republic of Korea

^e Department of Mathematics and Global Systems Institute, University of Exeter, Exeter, EX4 4QE, UK

ARTICLE INFO

Keywords:

AMOC instability
Deep learning
Bifurcation-induced transition
Rate-induced transition
Noise-induced transition
Complex systems

ABSTRACT

Key components of the Earth system can undergo abrupt transitions when the magnitude or rate of external forcing exceeds critical thresholds. In this study, we use the example of the Atlantic Meridional Overturning Circulation (AMOC) to demonstrate the challenges associated with anticipating such transitions when the complex system is susceptible to bifurcation-induced, rate-induced, and noise-induced tipping. Using a calibrated AMOC model, we conduct large ensemble simulations and show that transition behavior is inherently stochastic: under identical freshwater forcing scenarios, some ensemble members exhibit transitions while others do not. In this stochastic regime, traditional early warning indicators based on critical slowing down are unreliable in predicting impending transitions. To address this limitation, we develop a deep learning (DL)-based approach that identifies higher-order statistical differences between transitioning and non-transitioning trajectories within the ensemble realizations. This method enables the real-time prediction of transition probabilities for individual trajectories prior to the onset of tipping. Our results show that the DL-based indicator provides effective early warnings in a system where transitions can be induced by bifurcations, critical forcing rates, and noise. These findings underscore the probabilistic safe operating boundary and the potential in identifying early warning indicators for abrupt transitions of complex systems under uncertainty.

1. Introduction

Earth System Models (ESMs) and paleoclimate observations suggest that some major components of the Earth system are susceptible to abrupt transitions [1–4], sometimes labeled as tipping behavior. Major candidates of tipping elements in the Earth system are the Amazon rainforest [5], the continental ice sheets [6], and the Atlantic Meridional Overturning Circulation (AMOC) [7,8]. Previous studies have identified three primary mechanisms that can drive such transitions in nonlinear dynamical systems [9–11]. First, a gradual increase in external forcing may exceed a critical threshold, triggering a bifurcation-induced transition and thus causing a sudden shift to an alternative equilibrium [1,12–14]. Second, if the rate of external forcing is sufficiently rapid compared to the intrinsic time scale of the system, the system may not remain within the basin of attraction of its original

^{*} Corresponding author at: Earth System Modelling, School of Engineering and Design, Technical University of Munich, Munich, 85521, Germany.

E-mail address: y.huang@tum.de (Y. Huang).

<https://doi.org/10.1016/j.chaos.2026.118374>

Received 14 February 2026; Received in revised form 16 April 2026; Accepted 17 April 2026

Available online 30 April 2026

0960-0779/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

equilibrium, leading to a loss of stability and a subsequent transition [15–17]. Third, persistent stochastic perturbations can drive the system across basin boundaries even in the absence of changes in external forcing, leading to noise-induced transitions between alternative equilibria [9,18–22]. The rapid pace of anthropogenic forcing compared with the internal timescales of key Earth system components, such as the AMOC, poses significant risk and uncertainty associated with abrupt transitions from combined effects of these three mechanisms [23–29]. Safe operating spaces define the boundaries within which a system can function without undergoing undesired transitions. Therefore, identifying the boundaries of safe operating spaces and early warning signals of potential state transitions is an important and challenging task for the study of future climate change and climate mitigation policy [2,23,24,30,31].

The dominant approach to understanding the mechanisms and characteristics of critical transitions of the AMOC involves the use of process-based models. These models investigate AMOC instability and critical thresholds in the context of external forcing and internal climate variability. A common experimental setup applies a gradually changing external forcing to the AMOC until a critical threshold is reached, beyond which bifurcation-induced transition typically occurs [8,26,32–35]. However, recent ensemble simulation studies have demonstrated that under identical scenarios of increasing anthropogenic forcing, different internal climate variability between realizations can lead to divergent outcomes [22,26,36–38]. In some realizations, the AMOC undergoes sustained weakening followed by tipping, while in others, transitions do not occur. Because bifurcation-induced transitions are inherently deterministic, these results suggest that rate- and noise-induced effects can introduce stochasticity into the onset of AMOC tipping. A follow-up question is whether this mechanism also introduces uncertainty to the critical thresholds and safe operating spaces.

Despite its importance, conducting large ensemble simulations in comprehensive Earth system models to systematically investigate the stochastic nature of AMOC transitions is currently difficult to achieve due to the high computational cost. As an alternative, conceptual box models offer a more efficient framework for capturing the essential dynamics governing AMOC stability [35,39,40], and documenting the instability patterns evident in both comprehensive ESMs and paleoclimate archives [11,35,40–42]. Recent studies have utilized simulation data from comprehensive ESMs to calibrate a five-box model of the AMOC [11,40], with emphasis on the dynamical coupling and feedback mechanisms among the oceanic components of Northern Atlantic, Tropical Atlantic, North Atlantic Deep Water, Indo-Pacific, and Southern Ocean. The calibration enables the model to reproduce the double-equilibrium and critical transitions characteristic of the AMOC in comprehensive ESMs [11,28,35,40,43–46]. In this study, we aim to understand the characteristic of critical thresholds when the complex system is susceptible to bifurcation-induced, rate-induced, and noise-induced tipping. As a representative example, we employ this calibrated box model of Wood et al. [40] to generate a large ensemble of simulations under a range of forcing scenarios.

The collective behavior of a large ensemble of simulated realizations provides insights into the characteristics and uncertainties associated with tipping events. In practical applications, however, it is often necessary to identify early warning signals of pending transitions from a single noisy time series [12,30,31,47]. A key challenge, therefore, is to extract predictive insights from the empirical distribution of ensemble trajectories to inform the prediction of tipping in individual time series. Critical slowing down (CSD) theory suggests that as the local stability of a system weakens, even minor perturbations can have a prolonged effect on the system's dynamics [1,12,31,47]. This is typically reflected in increased variance and lag-1 autocorrelation of the noisy time series, serving as statistical early warning indicators preceding bifurcation-induced tipping. Nonetheless, CSD-based indicators are ineffective in detecting rate- and noise-induced tipping [17], whereas deep learning approaches [17,45,48,49] have shown enhanced predictive performance in predicting rate-induced and noise-induced tipping [17]. However, the deep learning model used in Huang et al. 2024 [17] is constrained to predicting transitions only with a fixed forecast lead time. In the stochastic regime, the time remaining before a transition is usually unknown and may vary substantially across realizations. A single predefined forecast lead time is therefore not necessarily very useful under practical circumstances. This caveat motivates a framework that updates transition risk continuously as the trajectory evolves. Moreover, the applicability of deep learning predictions has not yet been validated under scenarios involving combinations of the three types of tipping mechanisms. Here we aim to improve upon these deep learning approaches by removing the reliance on predefined lead times, thereby enabling real-time prediction of AMOC tipping probability. We demonstrate that convolutional neural networks (CNNs) can extract features related to higher-order statistical moments from the empirical distribution of ensemble realizations, allowing for the prediction of tipping events in a single time series under the combined effects of the three types of tipping mechanism.

2. Methods

2.1. AMOC box model

To efficiently explore the Atlantic Meridional Overturning Circulation (AMOC) tipping dynamics under stochastic forcing, we adopt the reduced three-box model of AMOC introduced by Chapman et al. [11], which is a modification of the five-box model proposed by Wood et al. [40]. The reduced three-box model retains the essential dynamics of the full five-box model, including bifurcation-, noise-, and rate-induced tipping behavior. It still consists of five conceptual oceanic boxes of the Northern Atlantic, Tropical Atlantic, Southern Ocean, North Atlantic Deep Water and Indo-Pacific. Following Chapman et al. [11], the five-box model is reduced by assuming constant salinity in the Southern Ocean and North Atlantic Deep Water boxes, with values inferred from calibration to HadGEM3-GC31-MM. A global salt conservation constraint is then applied to analytically determine the salinity in the Indo-Pacific box. This yields a two-dimensional system governing the evolution of salinity in the Northern Atlantic and Tropical Atlantic boxes. The AMOC strength q is modeled as:

$$q = \lambda(\alpha(T_S - T_{Nor.}) + \beta(S_{Nor.} - S_S)) \quad (1)$$

where α and β are linear coefficients from a simplified equation of state for seawater, and λ is a model parameter determined by the calibration. T and S represent box-averaged temperature and salinity, respectively. The subscripts indicate the box labels, where S denotes the Southern Ocean box and Nor denotes the Northern Atlantic box. Temperatures are held fixed in the simulations, so density dynamics are governed by salinity evolution. The salinity evolution equations differ depending on the sign of q .

For $q \geq 0$:

$$V_{Nor.} \frac{dS_{Nor.}}{dt} = q(S_{Trop.} - S_{Nor.}) + K_{Nor.}(S_{Trop.} - S_{Nor.}) - F_{Nor.}S_0 \quad (2)$$

$$V_{Trop.} \frac{dS_{Trop.}}{dt} = q(\gamma S_S + (1 - \gamma)S_{IP} - S_{Trop.}) + K_S(S_S - S_{Trop.}) + K_{Nor.}(S_{Nor.} - S_{Trop.}) - F_{Trop.}S_0 \quad (3)$$

and for $q < 0$:

$$V_{Nor.} \frac{dS_{Nor.}}{dt} = |q|(S_B - S_{Nor.}) + K_{Nor.}(S_{Trop.} - S_{Nor.}) - F_{Nor.}S_0 \quad (4)$$

$$V_{Trop.} \frac{dS_{Trop.}}{dt} = |q|(S_{Nor.} - S_{Trop.}) + K_S(S_S - S_{Trop.}) + K_{Nor.}(S_{Nor.} - S_{Trop.}) - F_{Trop.}S_0 \quad (5)$$

Salt conservation ensures the total salinity content C remains constant:

$$C = V_{Nor.}S_{Nor.} + V_{Trop.}S_{Trop.} + V_S S_S + V_{IP}S_{IP} + V_B S_B \quad (6)$$

allowing the salinity of Indo-Pacific box to be written as:

$$S_{IP} = \frac{C - V_{Nor.}S_{Nor.} - V_{Trop.}S_{Trop.} - V_S S_S - V_B S_B}{V_{IP}} \quad (7)$$

Here, V_i denotes the volume of each box, γ is the proportion of the cold water path; K_i represent wind driven transports between surface boxes, F_i are net surface freshwater fluxes. The S_0 is the reference salinity, set to 35 psu.

To simulate internal climate variability, we incorporate stochastic perturbations in the form of additive white noise:

$$dx(t) = f(x(t), t)dt + BdW_t \quad (8)$$

where $x(t) = (S_{Nor.}, S_{Trop.})$ is a vector that represents the salinities in Northern Atlantic and Tropical Atlantic boxes at a time step t and f is the model just described and calibrated to the piControl output from HadGEM3. B is a constant matrix of driving noise amplitudes and W_t is a vector of a standard independent Wiener process (each component of W_t is a Brownian motion with zero mean). Parameter values used in the simulations are summarized in Table S1, based on those in Alkhayun et al. [43], updated according to HadGEM3 [11,44].

2.2. Data and preprocessing

To comprehensively investigate the tipping behavior of the AMOC, we conducted extensive simulations across wide ranges of freshwater forcing magnitudes and rates, each subjected to stochastic perturbations in the form of additive white noise. As illustrated in Fig. 1(e), the freshwater forcing magnitudes ranged from 0.34 to 0.46 Sv, with higher resolution (0.005 Sv intervals) between 0.36 and 0.43 Sv to better capture critical transitions. In each scenario, the forcing increased linearly over a designated ramp-up period ranging from 50 to 1500 years in 50-year intervals, then remained fixed at its target value for an additional 2,000 years to allow the system to equilibrate. Each scenario was simulated 20,000 times with distinct noise realizations to ensure statistical robustness. In our algorithm to count the tipping proportion, the tipping point is defined as a persistent drop in AMOC strength below 0 Sv, which clearly separates irreversible collapse from transient weakening. This approach improves upon earlier tipping criteria (e.g., AMOC strength below 5 Sv [11]), which we found often included self-recovering trajectories that confound the subsequent analysis.

We first train a DL-based model using data from a representative scenario, in which forcing increased from 0 to 0.42 Sv over 800 years. This ensemble simulation yields a balanced dataset of 100,000 time series: 50,000 that exhibited AMOC collapse and 50,000 that did not. To identify dynamic and higher-order temporal precursors to the tipping of the AMOC, all time series are detrended and temporally aligned relative to the tipping point: For each time series exhibiting a tipping event, we truncate the series at the tipping point and retain the segment preceding it. This enables a focused analysis of the system's behavior leading up to the critical transition. For each time series without a tipping event, we randomly assign a pseudo-tipping point by sampling from the distribution of actual tipping times in the dataset. The time series is then truncated at that assigned point, and the preceding segment is retained. All truncated segments – whether from tipping or non-tipping cases – are subsequently aligned at time 0 (the tipping time) to explore common patterns preceding the occurrence of AMOC tipping. This dataset is subsequently divided into train, validation, and test using an 8:1:1 ratio to enable robust model training and evaluation. To highlight intrinsic fluctuations and suppress long-term trends, each time series was detrended by subtracting the ensemble mean of the non-tipping trajectories over the entire time span. This approach allows us to isolate early warning signals relative to the stable baseline that is unaffected by collapse dynamics.

2.3. Computation and evaluation of CSD indicators

Autocorrelation and variance are computed for all time series following the same procedure. Each time series undergoes nonlinear detrending using ordinary least squares regression to remove long-term trends. Subsequently, the CSD indicators lag-1 autocorrelation and variance are computed within a sliding window of 200 time steps, where both indicators are calculated using the standard way [3,12]. We experiment with different sizes for the sliding windows, ranging from 50 to 400 time steps. The resulting CSD indicators do not show significant divergence, and the insights for the inferences remain consistent.

To quantitatively compare the predictive performance of the CSD indicators with our DL-based classifier, we employ the area under the receiver operating characteristic (ROC) curve (AUC) as the evaluation metric [50]. The ROC curve captures the trade-off between true positive and false positive rates across varying discrimination thresholds. For the CSD indicators, the Kendall τ coefficient is used to summarize the temporal trend of the indicator and served as the discrimination threshold. For the detrended AMOC strength indicator, its raw value is used directly as the threshold variable. For the DL-based classifier, the predicted tipping probabilities are used directly as the threshold variable. This unified evaluation framework enables a fair and consistent comparison between classical early-warning indicators and the DL-based approach.

2.4. Configurations for deep learning

We design and train a CNN model to predict the probability of AMOC tipping in real time. As illustrated in Fig. S6, the architecture comprises two convolutional blocks. The first block contains two convolutional layers with a kernel size (3,1) and 32 output channels, each followed by a ReLU activation and a (3,1) max-pooling layer. The second block includes two convolutional layers with a kernel size (3,3) and 64 channels, also followed by ReLU activations and a (3,1) max-pooling layer. To reduce overfitting, a dropout layer with a rate of 0.4 is applied. The output feature maps are flattened and passed through two fully connected layers, with the final layer producing tipping probabilities via a softmax activation.

The model is implemented in PyTorch 2.5.0 and trained for up to 300 epochs with a batch size of 64, using the cross-entropy loss function and stochastic gradient descent (SGD) optimizer. The initial learning rate was set to 0.01 and decayed to 0.001 via a cosine annealing schedule. A weight decay of $1e-4$ was applied for regularization. Learning rate adjustments were performed with a patience of 5 epochs, and early stopping is triggered if validation performance did not improve for 50 consecutive epochs. These training configurations are applied consistently across all models in this study.

Model performance is evaluated using classification accuracy, defined as the proportion of correctly classified tipping and non-tipping cases. A prediction is considered correct if the model assigned a tipping probability greater than 0.5 to a tipping case, or less than 0.5 to a non-tipping case.

3. Results

3.1. Stochastic abrupt transition under critical magnitude and rate of forcing

We begin by analyzing the deterministic dynamics of AMOC instability, using a box model calibrated to the comprehensive ESM HadGEM3-GC31-MM [11,44]. By gradually increasing and then decreasing the freshwater forcing, we identify the bifurcation diagram at which the AMOC undergoes abrupt transitions between distinct equilibrium states (Fig. S1). Fig. S2(a) illustrates bifurcation- and rate-induced tipping, respectively, under noise-free conditions. Specifically, under a slowly increasing forcing from 0 to 0.6 Sv over 10,000 years, the AMOC abruptly tips once the forcing exceeds 0.53 Sv, indicating the critical threshold of bifurcation-induced tipping in case of the deterministic dynamics (Fig. S2(b)). In contrast, a rapidly changing forcing can also trigger the transition at a lower forcing magnitude [11,15]. For instance, when the freshwater forcing rapidly increases from 0 to 0.41 Sv over just 200 years (Fig. S2(c)), the system also collapses, which is a case of rate-induced tipping. This emphasizes that both the magnitude and rate of forcing determine when AMOC instability occurs [15]. To systematically identify the critical thresholds for both bifurcation- and rate-induced effects, we perform simulations across a broad range of forcing scenarios, with different combinations of the forcing magnitude and its changing rate. As a result, the deterministic boundary distinguishing tipping from non-tipping conditions is defined by joint thresholds from the bifurcation- and rate-induced effects (Fig. S2(d)).

To incorporate the stochastic nature, the box model is further extended by subjecting it to noise perturbations derived from the HadGEM3-GC31-MM [11,44]. To account for uncertainty arising from stochastic dynamics, we conduct 20,000 ensemble simulations, each subjected to identical freshwater forcing but with independent realizations of white noise. In the slow forcing scenario, where freshwater forcing increases from 0 to 0.6 Sv over 10,000 years, the two realizations q1 and q2 undergo tipping when the forcing exceeds an average threshold of 0.43 Sv (Figs. 1(a) and 1(c)). However, under a fast forcing scenario (increasing from 0 to 0.39 Sv over 200 years), realization q3 remains stable while q4 tips (Figs. 1(b) and 1(d)) - thus the tipping occurrence is stochastic. Fig. 1(e) presents the proportion of the tipping realizations within ensemble simulations at each combination of the forcing magnitude and its rate of change. This reveals a nonlinear and probabilistic transition boundary that depends not only on the magnitude of the forcing but also on its rate, suggesting that the boundary of the AMOC's safe operating space is not a fixed threshold, but rather a probabilistic zone. Furthermore, this result demonstrates that even forcing below deterministic tipping thresholds can lead to tipping with substantial probability under the combined effects of bifurcation-, rate- and noise-induced mechanisms. The contrasting critical threshold patterns between deterministic and stochastic dynamics are shown in Fig. S2(d) and Fig. 1(e). Consequently, the boundaries of the system's "safe operating space" become smeared out.

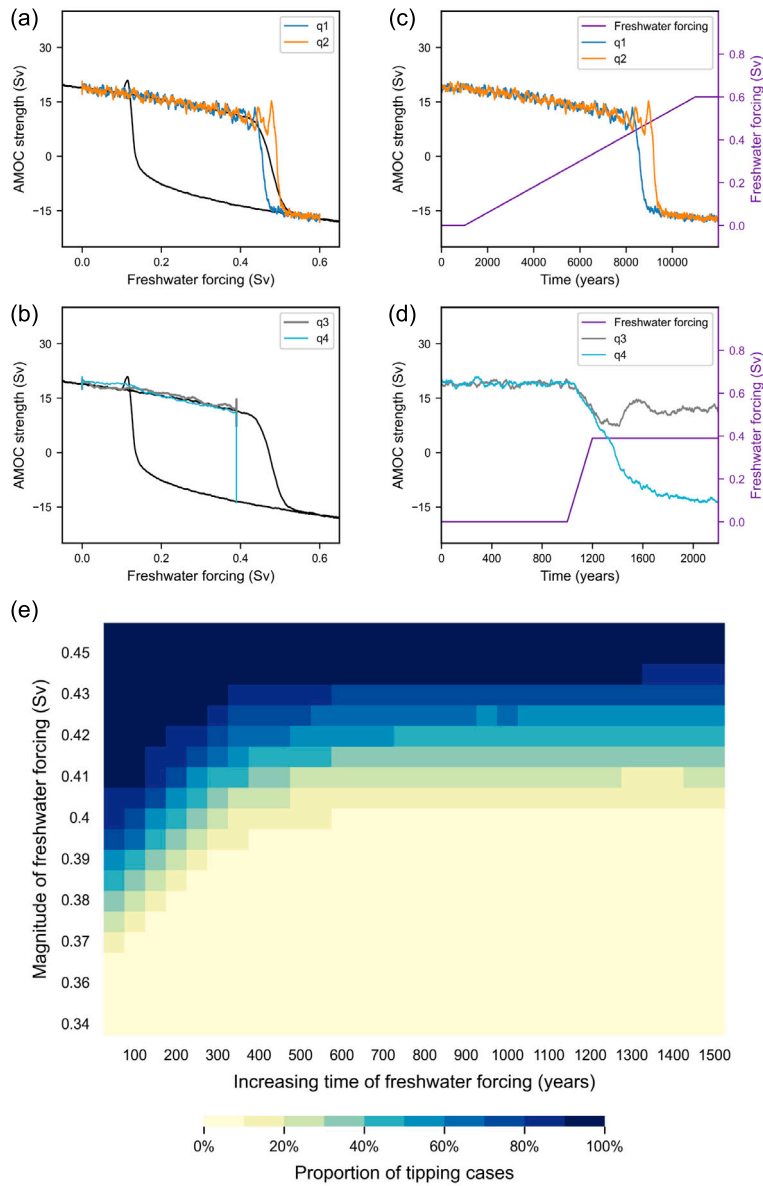


Fig. 1. Bistability, path-dependence, and collapse likelihood of the AMOC under freshwater forcing, based on simulations of the box model considered here [40,44]. (a, b) Bifurcation structure of the AMOC in response to freshwater forcing. The black line indicates the bistable regime of the AMOC strength as freshwater forcing linearly increases and then decreases, which is derived from an ensemble average over 50 realizations. Colored trajectories correspond to the transient responses shown in panels (c) and (d). (c) Two example trajectories (blue and orange) under a linear increase in freshwater forcing to 0.6 Sv over 10,000 years (purple line), both exhibiting collapse, indicative of bifurcation-induced tipping. (d) Two example trajectories under a faster linear increase in freshwater forcing to 0.39 Sv over 200 years (purple line). The gray trajectory remains in the strong AMOC state, while the cyan trajectory collapses, demonstrating rate-induced tipping. (e) Heat map showing the proportion of AMOC tipping cases across a range of freshwater forcing amplitudes and times over which the freshwater forcing increases. At each parameter setting, 20,000 stochastic simulations were performed under white noise perturbations. The AMOC tipping is defined as a persistent decline of the AMOC strength below 0 Sv, distinguishing it from temporary or reversible weakening. Color intensity indicates the tipping proportion, highlighting the combined influence of forcing value and rate on AMOC stability. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

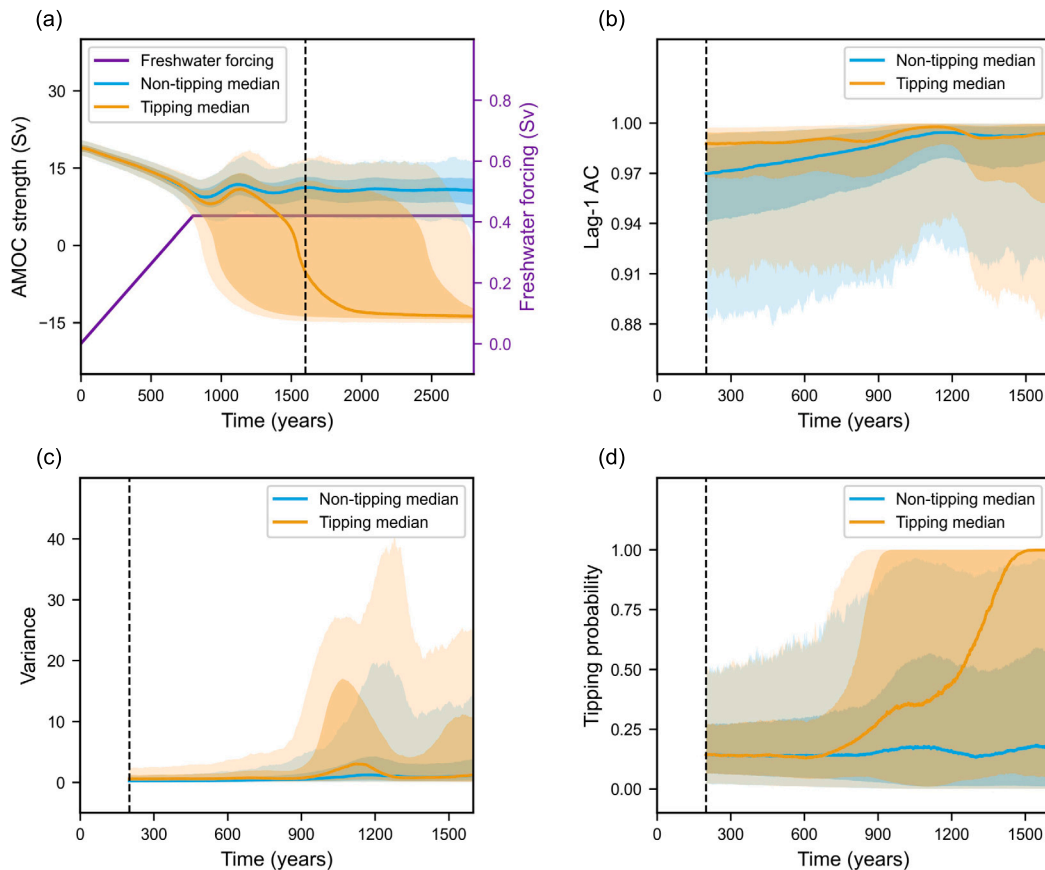


Fig. 2. Real-time evolution of early warning indicators in the AMOC three-box model. (a) Time series trajectories generated under a linearly increasing freshwater forcing up to 0.42 Sv over 800 years (purple line). The orange and blue lines denote the ensemble medians of trajectories that do and do not undergo tipping, respectively. Shaded regions indicate the 75% (dark) and 99% (light) confidence intervals. (b) estimated Lag-1 autocorrelation. (c) estimated variance. (d) tipping probabilities generated by the deep learning model. All indicators are computed using a rolling window of 200 time steps. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Large uncertainty for canonical early warnings

We respectively examine the CSD indicators and AMOC strength by testing them on the ensemble realizations under a representative freshwater forcing scenario. This scenario is characterized by a linear increase in forcing with an upper magnitude of 0.42 Sv over 800 years (Fig. 2(a)). Under this forcing scenario, tipping can be triggered through a combination of bifurcation-, rate-, and noise-induced tipping mechanisms (Fig. 1(e)). From the ensemble simulation, we extract 50,000 time series that exhibit tipping, where tipping is defined as the AMOC strength reaching 0 Sv, and the tipping time as the moment this threshold is first reached. All realizations that pass the 0 Sv eventually reach the AMOC-off state. In addition, we also pick 50,000 time series that do not tip, creating a balanced dataset of 100,000 samples. The ensemble time series in the tipping category exhibit a wide range of tipping times (Fig. 2(a)). This is consistent with recent findings that considerable uncertainty persists in estimating tipping times [30,38]. For subsequent analysis, each time series is detrended by subtracting the ensemble mean of the non-tipping time series, then we align the ends of all time series to their tipping times. As expected from theory, the computed CSD indicators and the detrended AMOC strength as an indicator, show substantial overlap between the tipping and non-tipping categories (Figs. S3(a)–S3(c)), indicating that they cannot be used to anticipate transitions. Indeed, the comparison of receiver operating characteristic (ROC) score [50] suggests that all of these canonical early warning indicators have no prediction skill at forecast lead times of 200 years before the tipping point (Fig. S4(a)). Although the AMOC strength typically decreases in the tipping realizations, the differences between the tipping and non-tipping ensembles remain statistically indistinguishable until shortly before the tipping point.

3.3. Disentangling the predictive signal with ensemble distributions

We further examine the ensemble probability distributions using a lead time of 200 years as an illustrative example, where a 200-year lead time means that data within 200 years prior to the tipping event are unavailable. Although the probability density functions

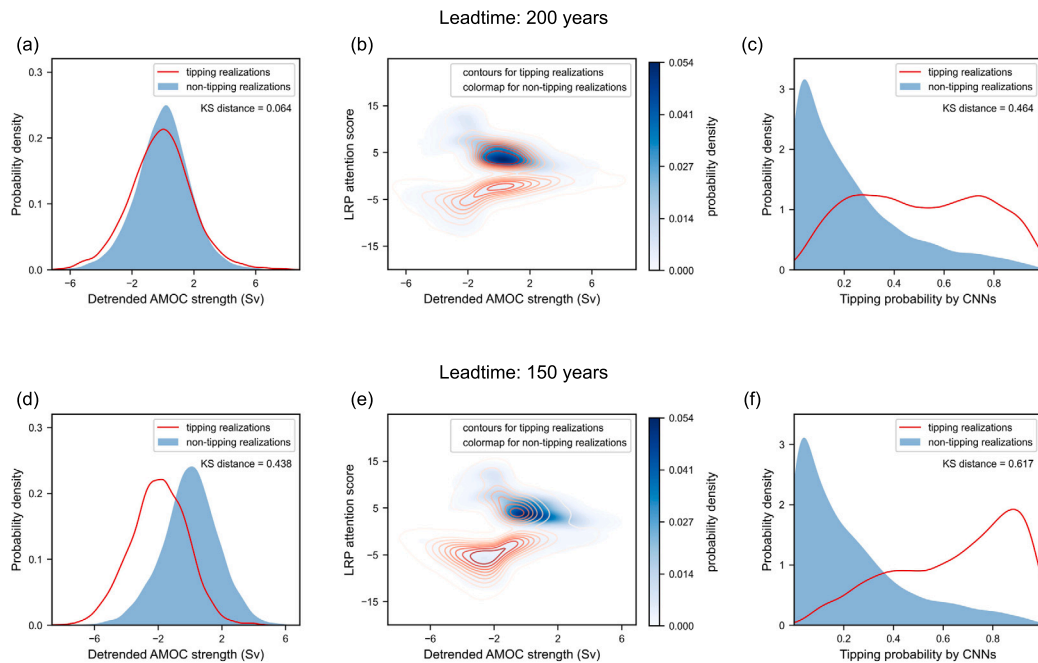


Fig. 3. Probability distributions of detrended AMOC strength, LRP attention score of CNNs, and tipping probability by CNNs for tipping and non-tipping realizations. Probability density functions are presented for three indicators: Detrended AMOC strength (a, d), LRP attention score of CNNs (b, e), and the tipping probability by CNNs (c, f). The distributions are shown for both tipping and non-tipping realizations at two forecast leadtimes: 200 years (a–c) and 150 years (d–f). The Kolmogorov–Smirnov (KS) distance, measuring the maximum distance between the cumulative distributions of the two ensembles, is reported for each panel. Higher KS distance indicate better distinguishing ability between tipping and non-tipping scenarios.

(PDFs) of the tipping and non-tipping categories overlap in their spreads, their distributional profiles reveal subtle differences in their kurtosis (Fig. 3(a)), suggesting that distinctions between the two groups of data may persist in higher-order statistical moments or within higher-dimensional feature spaces [17,51,52].

We develop a DL-based classifier to discriminate potential differences between time series with and without transitions. In previous studies, deep learning models were typically trained at a specific lead time prior to the onset of tipping events (Figs. S7(b) and S7(d)). Such approaches are inherently limited by their reliance on fixed forecast lead times, which constrains their applicability across varying temporal scenarios [17]. As illustrated in Fig. 4(a), to address this limitation, we adopt a sliding-window strategy (See Supplementary Note 1 for details) to extract multiple fixed-length time series segments preceding the tipping point for CNN training. Each segment has a fixed window length of 200 years, and the sampling windows slide backward along the time axis from the tipping point to capture features preceding a potential tipping event. In this study, we fix the sampling range to three partially overlapping windows, each 200 years long, with adjacent windows overlapping by 100 years, as shown in Fig. 4(a) (windows-3-stride-100). During sampling, all segments extracted from the same realization are assigned the same label: 1 if the realization undergoes a tipping event, and 0 otherwise. These time series segments are treated as independent training samples, allowing the neural networks to learn early-warning patterns across different lead times. Once trained, the neural networks can take any 200-year time series segment as input, and output the estimated probability of an impending transition. Fig. 4(b) shows that the neural networks can make skillful predictions across different forecast lead times. This approach allows the neural networks to learn from a wider range of precursory dynamics and removes the constraint of requiring inputs to align with a specific forecast lead time during inference. We test a range of windowing strategies that extract time series segments preceding the tipping time with varying lengths and strides (Fig. 4 and Fig. S5). Among these, the windows-3-stride-100 (window length=200 years) configuration yields the best predictive performance, particularly for forecast lead times between 100 and 200 years (Fig. 4(b)). Based on this result, we adopt the windows-3-stride-100 setting for all subsequent experiments unless otherwise specified.

To visualize the DL-based classifier's decision making process, we employ layer-wise relevance propagation (LRP), which identifies the attention weights most relevant to the predictions of neural networks in the corresponding target samples [17,53,54]. Fig. 3(b) presents the joint probability density of detrended AMOC strengths and the LRP attention scores from the DL-based classifier, where the tipping and non-tipping categories become separable. This demonstrates that the difference between tipping and non-tipping categories can be discriminated with a considerable lead time (see also Fig. 3(e)).

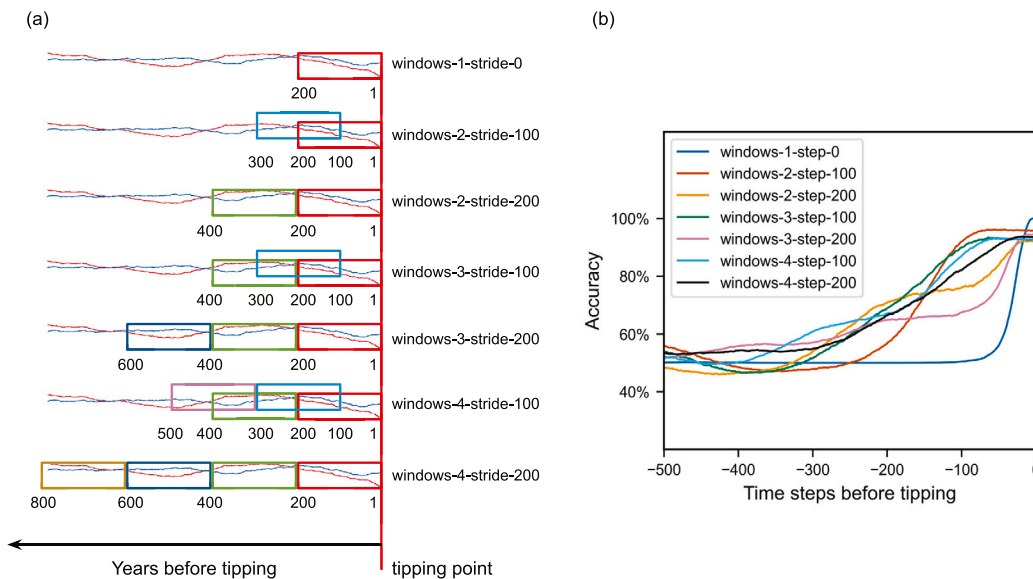


Fig. 4. (a) Schematic illustration of the sliding-window configurations used to train the CNN classifier. Seven different strategies are presented, each varying in the number of windows and stride lengths. The vertical red line marks the tipping point, and the colored boxes represent 200-year time series segments located at different temporal distances before the tipping point. Configurations include a single window (stride 0) as well as two, three, or four windows with strides of 100 or 200 years. During the training phase, a separate DL model is trained for each configuration using the corresponding time series segments. During inference, the trained model is applied to arbitrary 200-year segments from unseen time series to estimate tipping probabilities at different lead times. (b) Prediction accuracy of the trained models under the seven different training configurations. For each point on the x -axis (years before tipping), accuracy is computed as the proportion of correct predictions, calculated by pooling all predictions from the 200-year test segments that terminate at the corresponding lead time. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. A deep learning-based real-time indicator

From the above results we infer that the trained neural networks can estimate the probability of an impending AMOC tipping event in real time, without being constrained to a predefined forecast lead time. We apply the trained neural networks to continuously monitor the tipping probability based on the evolving AMOC strength time series. As shown in Fig. 2(d), the median of predicted tipping probability for tipping and non-tipping ensembles begins to diverge as early as year 700 in a 2800-year time series. For tipping realizations, the median of predicted probability steadily increases and approaches 1 around year 1500, while the non-tipping ensemble maintains a low and nearly constant probability below 0.2. This temporal separation demonstrates that the DL-based indicator is able to track evolving system dynamics and provide real-time early warnings. As illustrated in Fig. S3(d), approximately 100 years before the actual transition, the predicted tipping probability for the tipping realizations exceeds the 99% confidence interval of the non-tipping ensemble, providing a very early and statistically robust prediction signal. Accordingly, the ROC analysis (Fig. S4) shows that the DL-based indicator outperforms the traditional detrended AMOC strength-based indicator at forecast lead times of 200 and 150 years, where the CSD indicators exhibit no predictive skill.

3.5. Warning of tipping under combined bifurcation, rate and noise effects

To assess the feasibility of predicting stochastic tipping of AMOC in the general setting of combined bifurcation-, rate- and noise-induced effects, we generate a large ensemble of realizations for each freshwater forcing scenario shown in Fig. 1(e). Each scenario is parameterized by the magnitude and timescale of freshwater input, spanning a regime where tipping outcomes are intrinsically stochastic. In the following, we conduct two sets of experiments focusing on scenarios with stochastic tipping proportions between 5% and 95% at a 150-year forecast lead time.

We first examine whether the foresaid neural networks, trained solely on data from a single freshwater forcing scenario (linear increase to 0.42 Sv over 800 years), could generalize its prediction capability to other forcing scenarios. Despite this limited training domain, our neural network can achieve high accuracy across a wide region of forcing scenarios (Fig. 5(a)), particularly in scenarios characterized by strong forcing magnitudes and relatively slow forcing rates. The good transferability performance of the deep learning model suggests that predictive signatures preserved in the time-series samples are approximately consistent across different freshwater forcing scenarios. However, performance degrades under some fast forcing scenarios, likely due to the shortened pre-tipping trajectories that limit the available predictive information and the increased influence of rate-induced dynamics, which are harder to detect from early-time signals.

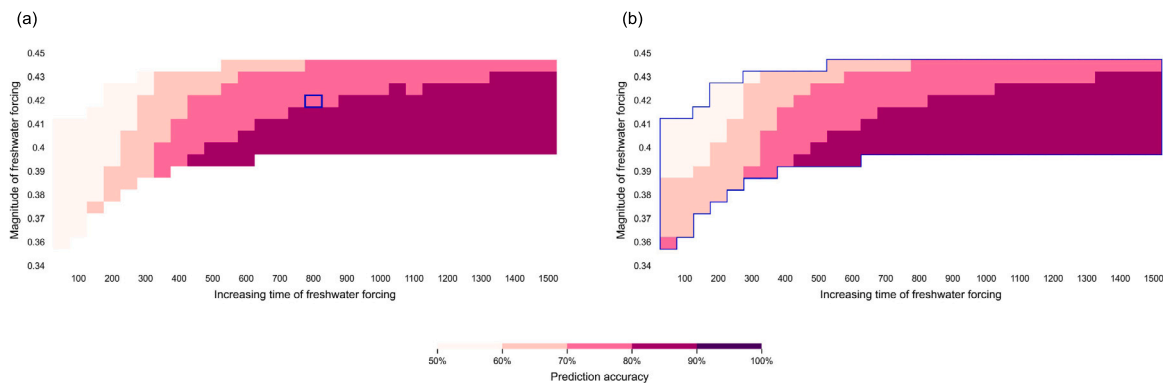


Fig. 5. Prediction accuracy of the DL model for AMOC tipping across freshwater forcing scenarios. (a, b) Heat maps showing the classification accuracy of the DL model in predicting AMOC tipping using the 150-year trajectory preceding each outcome. Each grid point corresponds to a freshwater forcing scenario, defined by its magnitude and increasing time, consistent with the parameter space in Fig. 1(e). Only scenarios with a tipping proportion between 5% and 95% were evaluated. (a) DL model was trained on 100,000 trajectories (50,000 tipping and 50,000 non-tipping) generated under a single representative freshwater forcing scenario (linear increase to 0.42 Sv over 800 years). (b) DL model was trained on a scenario-specific, balanced dataset, in which 1,000 tipping and 1,000 non-tipping trajectories were sampled for each evaluated forcing scenario. Color denotes the prediction accuracy for each scenario, revealing the generalization performance of DL models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

As a benchmark, we additionally train a second DL-based classifier using a balanced dataset sampled across all relevant forcing scenarios. For each scenario with tipping proportions between 5% and 95%, we draw 1,000 tipping and 1,000 non-tipping trajectories, aggregating them to form a comprehensive training dataset. The neural networks trained under this broader distribution demonstrated enhanced generalization, as shown in Fig. 5(b), achieving higher accuracy even in the rapid-forcing regimes where traditional statistical early warning indicators tend to fail. Notably, prediction accuracy remains robust even near the boundaries of the stochastic tipping regime, where the nonlinear intersection among bifurcation-, rate-, and noise-induced effects introduces substantial complexity and uncertainty.

4. Summary and discussion

Our study provides insights into the predictability of AMOC transitions in the general setting of combined bifurcation-, rate-, and noise-induced effects. Leveraging a recent AMOC box model calibrated to emulate a comprehensive ESM [40], we conduct large ensembles of stochastic simulations under diverse freshwater forcing scenarios. These simulations suggest that the safe boundary for AMOC collapse is not defined by deterministic thresholds, but instead emerges as a probabilistic zone due to internal variability, with the risk of transition further shaped by the magnitude and rate of external forcing, as well as stochastic perturbations. Critically, focusing only on the equilibrium case and associated CSD-based warning signals may lead to a severe underestimation of the actual tipping risk and associated false estimation of a safe operating space.

While bifurcation-induced transitions display early warning signals consistent with classical CSD, such indicators prove unreliable under rate- or noise-induced tipping. In contrast, our deep learning framework, which is trained on empirical distributions of ensemble realizations, overcomes this limitation by enabling real-time prediction of the tipping probability without requiring a predefined forecast lead time and without being confined to just one of the possible underlying tipping mechanisms (bifurcation, rate, or noise). Comparative analysis of detrended AMOC strength, LRP attention maps, and CNN-predicted tipping probabilities confirms that the neural networks capture emergent features beyond the low-order statistical moments used for detecting CSD and other classical early warning signals. Furthermore, the sliding-window approach used in this study relies on the implicit assumption that the signatures of early warning signals persist over a sufficiently long period prior to the tipping transition, thereby enabling the sliding windows to provide continuous early warnings before the transition occurs. The predictive skill of our model suggests that this assumption holds.

Our findings have two key implications. First, we show that the concept of a deterministic critical threshold may be inadequate for informing tipping risk assessments targeting the AMOC. Instead, probabilistic boundaries of safe operating spaces, derived from ensemble-based statistics, offer a more comprehensive framework for evaluating system stability. Second, our approach provides a scalable method for developing early warning indicators that account for the combined effects of bifurcation-, rate-, and noise-induced tipping. This methodology has broader applicability beyond the AMOC, extending to other Earth system components characterized by multistability and noise sensitivity, such as the Greenland Ice Sheet and the Amazon rainforest [4,36,55].

There remains scope for additional future work. The five-box model [40] employed in this study captures essential mechanisms of AMOC instability, is calibrated with respect to a comprehensive ESM, and facilitates large ensemble simulations, which are capable of reproducing the instability patterns observed in both comprehensive ESMs and paleoclimate archives [11,35,40,42]. While this five-box model necessarily abstracts spatial processes and feedbacks represented in full-complexity models [56], further integration with

intermediate-complexity or fully coupled climate models may help validate the robustness of our findings and assess their sensitivity to model architecture. However, the substantial computational cost associated with performing large-ensemble simulations with such high-complexity models would remain a challenge. Moreover, while LRP offers important interpretability, further integration with causal discovery tools [57] or physics-informed neural networks [58] might improve mechanistic attribution.

In summary, our results highlight the need to refine conceptual and computational tools for anticipating tipping events in the climate system. As anthropogenic forcing continues to accelerate, integrated approaches that blend dynamical theory, reduced-complexity modeling, and deep learning will be essential for managing deep uncertainty at the edge of climatic instability.

CRedit authorship contribution statement

Wenjie Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Yu Huang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Sebastian Bathiany:** Writing – review & editing, Validation, Methodology, Investigation, Funding acquisition, Conceptualization. **Yechul Shin:** Writing – review & editing, Validation, Investigation, Conceptualization. **Maya Ben-Yami:** Writing – review & editing, Validation, Investigation. **Suiping Zhou:** Writing – review & editing, Validation. **Niklas Boers:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Code availability

The code involved in the study is deposited in the Github public repository at <https://github.com/wenjio/Probabilistic-anticipation-of-AMOC-transitions-via-deep-learning>.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This is ClimTip contribution #157; the ClimTip project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101137601: Funded by the European Union. Y.H. also acknowledges Alexander von Humboldt Foundation for Humboldt Research Fellowship. N.B. acknowledges further support by the Volkswagen foundation, the Deutsche Forschungsgemeinschaft (DFG) and National Research Foundation of Korea (NRF): Climate Resilience under Zero-Emission Commitment Scenarios, BO 4455/2-1 (eBer-24-62950; RS-2024-00438471), and the advanced Research and Invention Agency (ARIA). Y.S. is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00334637).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.chaos.2026.118374>.

Data availability

Data will be made available on request.

References

- [1] Dakos V, Scheffer M, Van Nes EH, Brovkin V, Petoukhov V, Held H. Slowing down as an early warning signal for abrupt climate change. *Proc Natl Acad Sci* 2008;105:14308–12.
- [2] Armstrong McKay DJ, Staal A, Abrams JF, Winkelmann R, Sakschewski B, Loriani S, Fetzer I, Cornell SE, Rockström J, Lenton TM. Exceeding 1.5 c global warming could trigger multiple climate tipping points. *Science* 2022;377:eabn7950.
- [3] Boers N, Ghil M, Stocker TF. Theoretical and paleoclimatic evidence for abrupt transitions in the Earth system. *Environ Res Lett* 2022;17:093006.
- [4] Flores BM, Montoya E, Sakschewski B, Nascimento N, Staal A, Betts RA, Levis C, Lapola DM, Esquivel-Muelbert A, Jakovac C, et al. Critical transitions in the Amazon forest system. *Nature* 2024;626:555–64.
- [5] Boulton CA, Lenton TM, Boers N. Pronounced loss of Amazon rainforest resilience since the early 2000s. *Nat Clim Chang* 2022;12:271–8.
- [6] Garbe J, Albrecht T, Levermann A, Donges JF, Winkelmann R. The hysteresis of the Antarctic Ice Sheet. *Nature* 2020;585:538–44.
- [7] Boers N. Observation-based early-warning signals for a collapse of the Atlantic meridional overturning circulation. *Nat Clim Chang* 2021;11:680–8.
- [8] Van Westen RM, Kliphuis M, Dijkstra HA. Physics-based early warning signal shows that AMOC is on tipping course. *Sci Adv* 2024;10:eadk1189.
- [9] Ashwin P, Wieczorek S, Vitolo R, Cox P. Tipping points in open systems: bifurcation, noise-induced and rate-dependent examples in the climate system. *Philos Trans R Soc A: Math Phys Eng Sci* 2012;370:1166–84.
- [10] Brunetti M, Ragon C. Attractors and bifurcation diagrams in complex climate models. *Phys Rev E* 2023;107:054214.
- [11] Chapman R, Sinet S, Ritchie PD. Tipping mechanisms in a conceptual model of the Atlantic meridional overturning circulation. *Weather* 2024;79:316–23.

- [12] Scheffer M, Bascompte J, Brock WA, Brovkin V, Carpenter SR, Dakos V, Held H, Van Nes EH, Rietkerk M, Sugihara G. Early-warning signals for critical transitions. *Nature* 2009;461:53–9.
- [13] Kleinen T, Held H, Petschel-Held G. The potential role of spectral properties in detecting thresholds in the Earth system: application to the thermohaline circulation. *Ocean Dyn* 2003;53:53–63.
- [14] Held H, Kleinen T. Detection of climate system bifurcations by degenerate fingerprinting. *Geophys Res Lett* 2004;31.
- [15] Lohmann J, Ditlevsen PD. Risk of tipping the overturning circulation due to increasing rates of ice melt. *Proc Natl Acad Sci* 2021;118:e2017989118.
- [16] Alharthi T. Rate-dependent tipping points for forced dynamical systems. [Ph.D. thesis], University of Exeter; 2023.
- [17] Huang Y, Bathiany S, Ashwin P, Boers N. Deep learning for predicting rate-induced tipping. *Nat Mach Intell* 2024;1–10.
- [18] Hasselmann K. Stochastic climate models part I. theory. *Tellus* 1976;28:473–85.
- [19] Timmermann A, Lohmann G. Noise-induced transitions in a simplified model of the thermohaline circulation. *J Phys Oceanogr* 2000;30:1891–900.
- [20] Stocker TF, Schmittner A. Influence of CO₂ emission rates on the stability of the thermohaline circulation. *Nature* 1997;388:862–5.
- [21] Boers N. Early-warning signals for dansgaard-oeschger events in a high-resolution ice core record. *Nat Commun* 2018;9:2556.
- [22] Cini M, Zappa G, Ragone F, Corti S. Simulating AMOC tipping driven by internal climate variability with a rare event algorithm. *Npj Clim Atmos. Sci* 2024;7:31.
- [23] Lenton TM, Rockström J, Gaffney O, Rahmstorf S, Richardson K, Steffen W, Schellnhuber HJ. Climate tipping points—too risky to bet against. *Nature* 2019;575:592–5.
- [24] Ritchie PD, Clarke JJ, Cox PM, Huntingford C. Overshooting tipping point thresholds in a changing climate. *Nature* 2021;592:517–23.
- [25] Wunderling N, Winkelmann R, Rockström J, Loriani S, Armstrong McKay DJ, Ritchie PD, Sakschewski B, Donges JF. Global warming overshoots increase risks of climate tipping cascades in a network model. *Nat Clim Chang* 2023;13:75–82.
- [26] Romanou A, Rind D, Jonas J, Miller R, Kelley M, Russell G, Orbe C, Nazarenko L, Latto R, Schmidt GA. Stochastic bifurcation of the North Atlantic circulation under a mid-range future climate scenario with the NASA-GISS modele. *J Clim* 2023;1–49.
- [27] Klose AK, Donges JF, Feudel U, Winkelmann R. Rate-induced tipping cascades arising from interactions between the Greenland Ice Sheet and the Atlantic meridional overturning circulation. *Earth Syst Dyn* 2024;15:635–52.
- [28] Sinet S, Ashwin P, von der Heydt AS, Dijkstra HA. AMOC stability amid tipping ice sheets: the crucial role of rate and noise. *Earth Syst Dyn* 2024;15:859–73.
- [29] Baker J, Bell M, Jackson L, Vallis G, Watson A, Wood R. Continued Atlantic overturning circulation even under climate extremes. *Nature* 2025;638:987–94.
- [30] Ben-Yami M, Morr A, Bathiany S, Boers N. Uncertainties too large to predict tipping times of major Earth system components from historical data. *Sci Adv* 2024;10:eadi4841.
- [31] Rietkerk M, Skiba V, Weinans E, Hébert R, Laepple T. Ambiguity of early warning signals for climate tipping points. *Nat Clim Chang* 2025;1–10.
- [32] Zhang X, Knorr G, Lohmann G, Barker S. Abrupt north Atlantic circulation changes in response to gradual CO₂ forcing in a glacial climate state. *Nat Geosci* 2017;10:518–23.
- [33] Ben-Yami M, Skiba V, Bathiany S, Boers N. Uncertainties in critical slowing down indicators of observation-based fingerprints of the Atlantic overturning circulation. *Nat Commun* 2023;14:8344.
- [34] Ditlevsen P, Ditlevsen S. Warning of a forthcoming collapse of the Atlantic meridional overturning circulation. *Nat Commun* 2023;14:1–12.
- [35] Dijkstra HA. The role of conceptual models in climate research. *Phys D: Nonlinear Phenom* 2024;457:133984.
- [36] Lohmann J, Dijkstra HA, Jochum M, Lucarini V, Ditlevsen PD. Multistability and intermediate tipping of the Atlantic Ocean circulation. *Sci Adv* 2024;10:eadi4253.
- [37] Gu Q, Gervais M, Danabasoglu G, Kim WM, Castruccio F, Maroon E, Xie SP. Wide range of possible trajectories of North Atlantic climate in a warming world. *Nat Commun* 2024;15:4221.
- [38] Curtis PE, Fedorov AV. Collapse and slow recovery of the Atlantic meridional overturning circulation (AMOC) under abrupt greenhouse gas forcing. *Clim Dyn* 2024;62:5949–70.
- [39] Frierson DM, Hwang YT, Fučkar NS, Seager R, Kang SM, Donohoe A, Maroon EA, Liu X, Battisti DS. Contribution of ocean overturning circulation to tropical rainfall peak in the Northern Hemisphere. *Nat Geosci* 2013;6:940–4.
- [40] Wood RA, Rodríguez JM, Smith RS, Jackson LC, Hawkins E. Observable, low-order dynamical controls on thresholds of the Atlantic meridional overturning circulation. *Clim Dyn* 2019;53:6815–34.
- [41] Ganopolski A, Rahmstorf S. Abrupt glacial climate changes due to stochastic resonance. *Phys Rev Lett* 2002;88:038501.
- [42] Hobden B, Ashwin P, Ritchie PD. Regularization of a conceptual model for dansgaard-oeschger events. *Chaos: Interdiscip J Nonlinear Sci* 2025;35.
- [43] Alkhayyon H, Ashwin P, Jackson LC, Quinn C, Wood RA. Basin bifurcations, oscillatory instability and rate-induced thresholds for Atlantic meridional overturning circulation in a global oceanic box model. *Proc R Soc A* 2019;475:20190051.
- [44] Chapman R, Ashwin P, Baker J, Wood R. Quantifying risk of a noise-induced AMOC collapse from northern and tropical Atlantic Ocean variability. *Environ Res Commun* 2024;6:111003.
- [45] Panahi S, Kong LW, Moradi M, Zhai ZM, Glaz B, Haile M, Lai YC. Machine learning prediction of tipping in complex dynamical systems. *Phys Rev Res* 2024;6:043194.
- [46] Pals D, Bathiany S, Kuettel J, Wood R, Boers N. Targeted calibration to adjust stability biases in complex dynamical system models. *Phys Rev X* 2026;16(2):021007, APS.
- [47] Lenton TM. Early warning of climate tipping points. *Nat Clim Chang* 2011;1:201–9.
- [48] Bury TM, Sujith R, Pavithran I, Scheffer M, Lenton TM, Anand M, Bauch CT. Deep learning for early warning signals of tipping points. *Proc Natl Acad Sci* 2021;118:e2106140118.
- [49] Bury TM, Dylewsky D, Bauch CT, Anand M, Glass L, Shrier A, Bub G. Predicting discrete-time bifurcations with deep learning. *Nat Commun* 2023;14:6331.
- [50] Fawcett T. An introduction to roc analysis. *Pattern Recognit Lett* 2006;27:861–74.
- [51] Schölkopf B, Smola A, Müller KR. Kernel principal component analysis. In: International conference on artificial neural networks. 1997, p. 583–8.
- [52] He W, Wan S, Jiang Y, Jin H, Zhang W, Wu Q, He T. Detecting abrupt change on the basis of skewness: numerical tests and applications. *Int J Climatol* 2013;33.
- [53] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One* 2015;10:e0130140.
- [54] Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: an overview. In: Explainable AI: interpreting, explaining and visualizing deep learning. 2019, p. 193–209.
- [55] Bochov N, Poltronieri A, Robinson A, Montoya M, Rypdal M, Boers N. Overshooting the critical threshold for the greenland ice sheet. *Nature* 2023;622:528–36.
- [56] Shin Y, Geng X, Oh JH, Noh KM, Jin EK, Kug JS. Fast and slow responses of Atlantic meridional overturning circulation to antarctic meltwater forcing. *Geophys Res Lett* 2024;51. e2024GL108272.
- [57] Spirtes P, Zhang K. Causal discovery and inference: concepts and recent methodological advances. *Appl Inform* 2016;3:1–28.
- [58] Raissi M, Perdikaris P, Karniadakis G. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J Comput Phys* 2019;378:686–707.