Interpretable Neural Network Forecasting of Ocean State Transitions Using Saliency

Qi-fan Wu[1], **Dion Häfner**[1,2], **Roman Nuterman**[1], **Guido Vettoretti**[1,3], **Markus Jochum**[1]
[1]Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
[2]Pasteur Labs, United States
[3] Department of Physical and Environmental Sciences, University of Toronto, Canada
qifan.wu@nbi.ku.dk, dion.haefner@simulation.science, nuterman@nbi.ku.dk,
g.vettoretti@utoronto.ca, mjochum@nbi.ku.dk

Abstract

We study sudden transitions in a key component of the climate system, the Atlantic Meridional Overturning Circulation (AMOC). Exploiting simulation results from a fully coupled climate model, we train a convolutional neural network to predict the AMOC as a result of ocean subsurface density and freshwater forcing. We find that the model can forecast transition dynamics it has never seen. Furthermore, we show how saliency maps can be used to interpret black-box neural network models in climate dynamics and enhance their performance, and we demonstrate that high saliency on excitable regions enables out-of-sample prediction of large-scale transitions. This approach opens new perspectives for interpretable, long-term AMOC forecasting.

I Introduction

The Atlantic Meridional Overturning Circulation (AMOC) is the zonally integrated mean flow in the Atlantic Ocean that plays a key role in Earth's climate in the past and present, due to its control of heat transport, freshwater distribution, deepwater formation, and ocean stratification (Kuhlbrodt et al. 2007). Paleoclimate records provide evidence that during the last 100,000 years, parts of the North Atlantic ocean circulation have frequently collapsed and recovered, and these sudden phase transitions are recorded as Dansgaard–Oeschger (D–O) events characterized by abrupt warming into interstadial periods followed by gradual cooling into stadial periods (Dansgaard et al. 1993). Recently there has been progress in reproducing D-O events with numerical simulations of Earth system models, and these simulations suggest that such dramatic climate change events were linked to sudden transitions in AMOC (Jochum et al. 2022).

AMOC variability is a nonlinear problem, and although there are various hypotheses about the mechanisms that drive its changes, a unifying analytical theory has so far been lacking (Wunsch and Heimbach 2013). The collapse of the AMOC is a rapid and large-scale weakening of the overturning circulation amplified by internal ocean–climate feedbacks, and some recent studies suggest that this may currently be happening (Dijkstra and van Westen 2025). Reconstructing past variations of the AMOC is critical for assessing its variability beyond the few decades covered by modern observations, yet data availability for Holocene AMOC changes remains poorly constrained (Gerber et al. 2025), making it difficult to forecast potential AMOC collapse in the 21st century.

The recent blossoming of data-driven machine learning methods that distill dynamics from data and make nonlinear systems amenable to linear analysis (Brunton and Kutz 2019), can be considered one of the promising approaches for predicting the possible dynamics of AMOC. Very recently, neural networks have been used to identify key drivers of AMOC variability from Community Earth System Model (CESM) simulations with glacial boundary conditions that reproduce the observed structure of D–O events, showing that surface freshwater flux (SFWF) and potential density at 200 meters depth (PD_200m) are the main controls (Wu et al. 2025). In the present work, we use only CESM data without transitions to train a convolutional neural network (CNN) model to predict future AMOC collapses in the test set. Moreover, we apply a series saliency approach to jointly improve forecasting accuracy and interpretability by identifying the most influential temporal and feature components contributing to the CNN predictions.

This paper is organized as follows: In Sec. II we explain the data preprocessing method, construct the model for forecasting AMOC strength, and use a saliency map to interpret the prediction results. In Sec. III we evaluate the model's performance and combine it with the underlying dynamics to explain the saliency map. In Sec. IV we discuss potential applications and future research directions, and we summarize the main results.

II Method

7,000 years of CESM simulations (Vettoretti 2022) are used to construct the forecasting model for AMOC collapse. The time series of SFWF and PD_200m are the input features, and the AMOC strength is the target (Appendix A.1). The first 3000 years without sudden transition are used for training, the last 20 percent of data in the training set are used as validation set. The remaining time series that include one D–O event are used as the testing set to evaluate whether the model can predict the AMOC collapse without it being included in the training set.

We employ a sliding window approach that uses an input

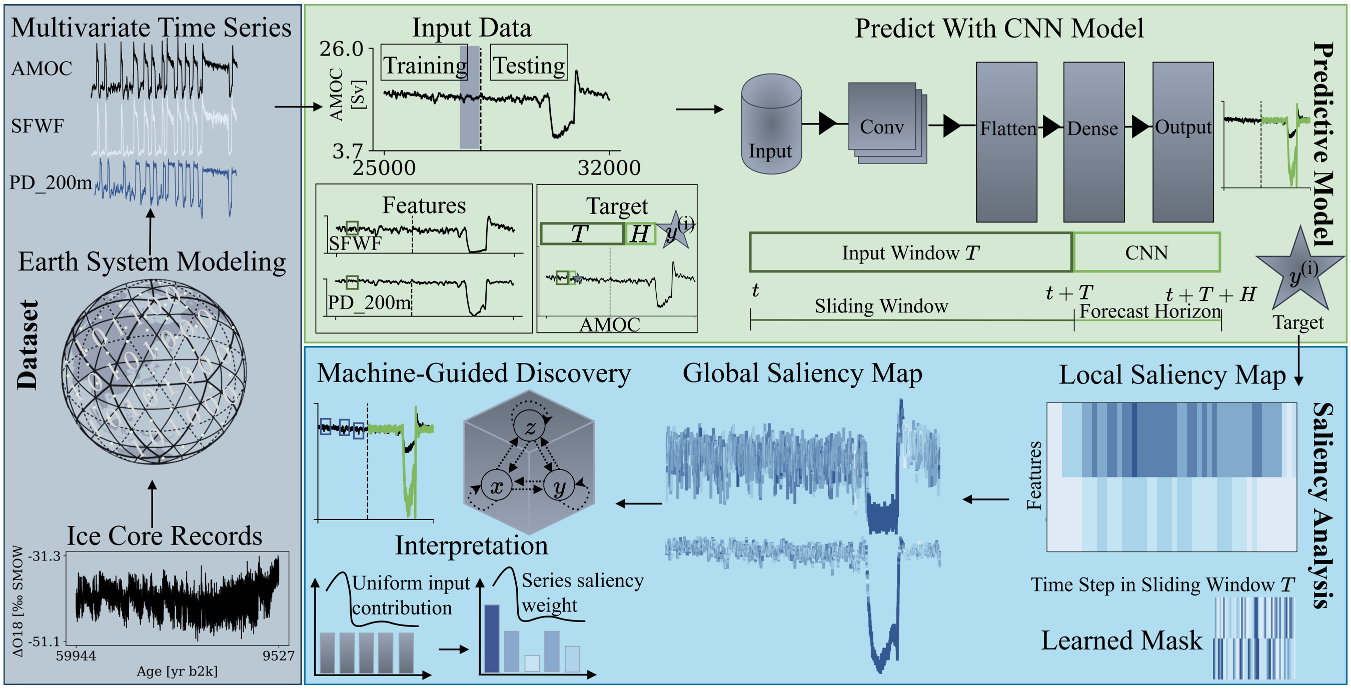


Figure 1: Schematic of the basic workflow. The CESM model reproduces the observed D–O event structure from ice core records and provides multivariate time series that capture the co-evolution of sudden transitions. We use only the interstadial data of two key variables before the last D–O event to predict the sudden AMOC collapse in the test set using a CNN with a sliding window approach. The gradient-based saliency map and the learned mask saliency map are used to enhance model performance and aggregate saliency globally to identify informative temporal–feature regions and investigate potential AMOC dynamics.

window of length T, starting at time t, with a forecast horizon H, to predict the target value $y^{(i)}$ at time $t + T + H$. In this study, we use 50 years of past SFWF and PD_200m data to predict the AMOC strength 20 years later, the same as the choices in Wu et al. (2025). The 1D CNN is used with three convolutional blocks and residual connections, followed by batch normalization, a flatten layer, and two dense layers for AMOC collapse forecasting. Limited by computational resources, we do not fine-tune hyperparameters or model architecture, as the goal is to successfully forecast the collapse rather than optimize the model. We design a custom loss function that combines mean squared error with sign-flip, integral consistency, and smoothness penalties to enhance the model's ability to capture both short-term fluctuations and long-term trends. The details are presented in Appendix A.2.

Due to their black-box nature, CNNs are still regarded as difficult to interpret in terms of their underlying mechanisms, making it hard to trust that they provide the optimal solution (Azam et al. 2023; Wang et al. 2020). Series saliency is a fascinating model-agnostic method that mixes original and perturbed time series with a learnable mask, achieving two birds with one stone by improving forecasting accuracy and providing temporal–feature interpretations (Pan, Hu, and Chen 2021). Gradient-based saliency and series saliency with a learnable mask are both methods that identify informative temporal-feature regions for forecasting (Pan, Hu, and Chen 2021; Pantiskas, Verstoep, and Bal 2020). Gradient-based saliency relies on gradient information after training, whereas series saliency adopts a learnable mixup strategy between original and perturbed inputs, integrating interpretability and adaptive augmentation (Mitrea, Lee, and Wu 2009; Guidotti et al. 2018; Rudin 2019; Pan, Hu, and Chen 2021; Serrano and Smith 2019; Dabkowski and Gal 2017). We investigate both methods for interpreting our CNN model. For each sliding window

$$X^{(i)} \in \mathbb{R}^{T \times F}, \quad i = 1, \ldots, k, \tag{1}$$

where F is the number of features and k is the total number of windows. The model output for each window is $y^{(i)}$ at time $t + T + H$. The gradient-based saliency map is computed as

$$S^{(i)} = \left| \frac{\partial y^{(i)}}{\partial X^{(i)}} \right|, \qquad S^{(i)} \in \mathbb{R}^{T \times F}. \tag{2}$$

Averaging over all k sliding windows gives

$$\bar{S} = \frac{1}{k} \sum_{i=1}^{k} S^{(i)}. \tag{3}$$

Each element $\bar{S}_{t,f}$ measures the sensitivity of the prediction to the input at time step t and feature f. We introduce a learnable mask

$$M \in \mathbb{R}^{T \times F}, \qquad 0 \leq M_{t,f} \leq 1, \tag{4}$$

which is optimized to minimize forecast loss when applied to perturbed inputs. The masked input is constructed as

$$X^{(i)}_{\text{masked}} = M \odot X^{(i)}_{\text{perturbed}}, \tag{5}$$

where $\odot$ denotes element-wise multiplication (i.e., feature-time specific weighting) and $X^{(i)}_{\text{perturbed}}$ is the noise-injected input sequence. $\bar{S}$ captures gradient-based sensitivity, while M encodes robustness-based importance under perturbation. Both methods capture key temporal–feature contributions to the forecast $y^{(i)}$. These global saliency patterns $\bar{S}*t, f$ and $M*t, f$ can be used to identify critical time steps $t \in [0, T]$ and dominant features $f \in \mathrm{SFWF}, \mathrm{PD}_2 00\mathrm{m}$ that most strongly influence the CNN's forecast $y^{(i)}$ at $t+T+H$, providing insight into the timing and mechanisms relevant to possible AMOC dynamics.

To make forecasts of AMOC strength more accurate and to evaluate the learned saliency map for interpretation, we separately apply the learned mask M and the gradient-based saliency map S to weight the input time-feature sequence before prediction.

$$\hat{y}^{(i)}_{\text{learned}} = f_\theta\big(M \odot X^{(i)}_{\text{perturbed}} \odot X^{(i)}\big), \tag{6}$$

$$\hat{y}^{(i)}_{\text{grad}} = f_\theta\big(S \odot X^{(i)}_{\text{perturbed}} \odot X^{(i)}\big). \tag{7}$$

f_θ is the trained CNN forecasting model, where θ represents the set of learned parameters of the model after training. The mask acts as an adaptive data augmentation and interpretation module, guiding the CNN to focus on informative temporal-feature regions when generating forecasts. We will discuss its role in machine-guided discovery of possible AMOC dynamics and the evaluation of the forecasting performance later.

III Results

Interestingly, the CNN prediction successfully captures the sudden transition in the testing set using only interstadial time series, with no D–O events present in the training set (Figure. 2(a)). This result is unexpected, given that CNNs are data-driven and not physically informed, usually interpolate within the learned data distribution, and our CNN relies on a purely supervised learning setup and is better at capturing local temporal correlations than rare or singular events (Raissi, Perdikaris, and Karniadakis 2019; Yang et al. 2024; Brunton and Kutz 2019; Hendrycks et al. 2020; Wang et al. 2020; Zhou et al. 2016; Cong, Yuan, and Liu 2013). We then compute the learned saliency mask M and the gradient-based saliency map S for the trained CNN model f_θ, and apply them to weight the input sequences $X^{(i)}$ and $X^{(i)}_{\text{perturbed}}$ to generate $\hat{y}^{(i)}_{\text{learned}}$ and $\hat{y}^{(i)}_{\text{grad}}$ for interpretable and more accurate AMOC forecasts (Figure. 2(b), Figure. 2(c), Equation (6), Equation (7)).

What is more interesting is that all three models successfully forecast the AMOC collapse, but with notable amplitude differences in their predicted transitions (Figure.2(c)). The results show that both masking strategies substantially reduce the prediction error compared to the no-mask baseline, with the gradient-based saliency map achieving the lowest Mean Squared Error (MSE) and custom loss overall (Table 1). However, because the onset time of D–O events is inherently difficult to define, these error metrics alone are insufficient for a complete assessment (Slattery et al. 2024). Therefore, evaluating the accuracy of the predicted onset timing is also crucial for a more comprehensive evaluation of the model.

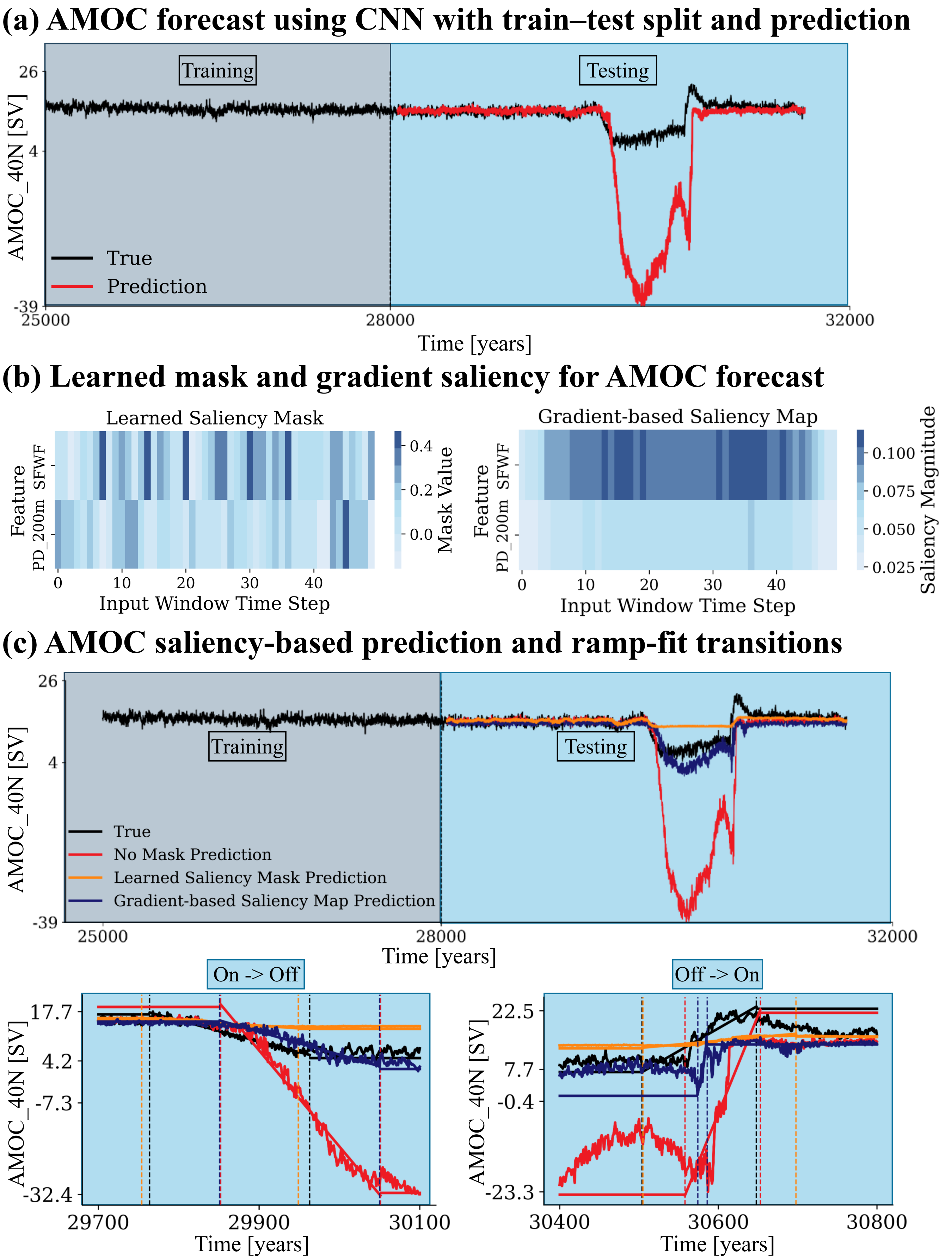
Method	**MSE**	**Custom Loss**
No Mask Prediction	5.88×10^{-2}	2.33×10^{5}
Learned Saliency Mask Prediction	2.19×10^{-3}	8.92×10^{3}
Gradient-based Saliency Map Prediction	1.74×10^{-3}	7.22×10^{3}

Table 1: MSE and Custom Loss error for different model predictions.

The ramp fitting method is a widely used tool for quantifying climate transitions in time series (Mudelsee 2000; Erhardt et al. 2019; Slattery et al. 2024; Capron et al. 2021). For a (climate) system at equilibrium disturbed by external forcing transitions to a new equilibrium state, the ramp fitting determines when a transition starts and ends, as well as the mean levels before and after the change (Mudelsee 2000). Consider a system with candidate ramp start and end points (t_0, t_1) and corresponding values (a_0, a_1), and we compute a piecewise-linear ramp $r(t; t_0, t_1, a_0, a_1)$ via nonlinear least squares and then select the pair that maximizes the absolute height change $|\Delta h| = |a_1 - a_0|$. The results show that the learned saliency mask prediction most closely matches the true transition timing for both downward and upward transitions, while the other models exhibit lagged transition timing regardless of the magnitude error in the height change (Table 2, Figure. 2(c)).

Method	t_0	t_1	**Δheight**
On $\rightarrow$ Off Transition (Down)			
True	29764	29963	-11.60
No Mask Prediction	29852	30051	-52.45
Learned Saliency Mask Prediction	29754	29949	-2.91
Gradient-based Saliency Map Prediction	29851	30050	-12.54
Off $\rightarrow$ On Transition (Up)			
True	30504	30648	$+16.54$
No Mask Prediction	30558	30653	$+46.93$
Learned Saliency Mask Prediction	30505	30698	$+3.51$
Gradient-based Saliency Map Prediction	30574	30586	$+12.73$

Table 2: The table compares the estimated transition times (t_0 and t_1) and height changes (Δheight) obtained from different predictive models for both downward (On $\rightarrow$ Off) and upward (Off $\rightarrow$ On) transitions.



(a) AMOC forecast using CNN with train–test split and prediction

(b) Learned mask and gradient saliency for AMOC forecast

(c) AMOC saliency-based prediction and ramp-fit transitions

Figure 2: (a) AMOC forecast using CNN with train–test split and prediction performance. (b) Learned saliency mask and gradient-based saliency map indicating informative temporal–feature regions. (c) AMOC forecast comparison with and without saliency masks. The two lower subplots show ramp-fit detection of On → Off and Off → On transition timing for D–O events. Dashed lines indicate detected transition start t_0 and end t_1.

As the CNN model with a learned saliency map performs well in predicting the collapse, we further aim to interpret the model by analyzing saliency values in the dynamical phase space to link saliency structure with system dynamics and model performance.

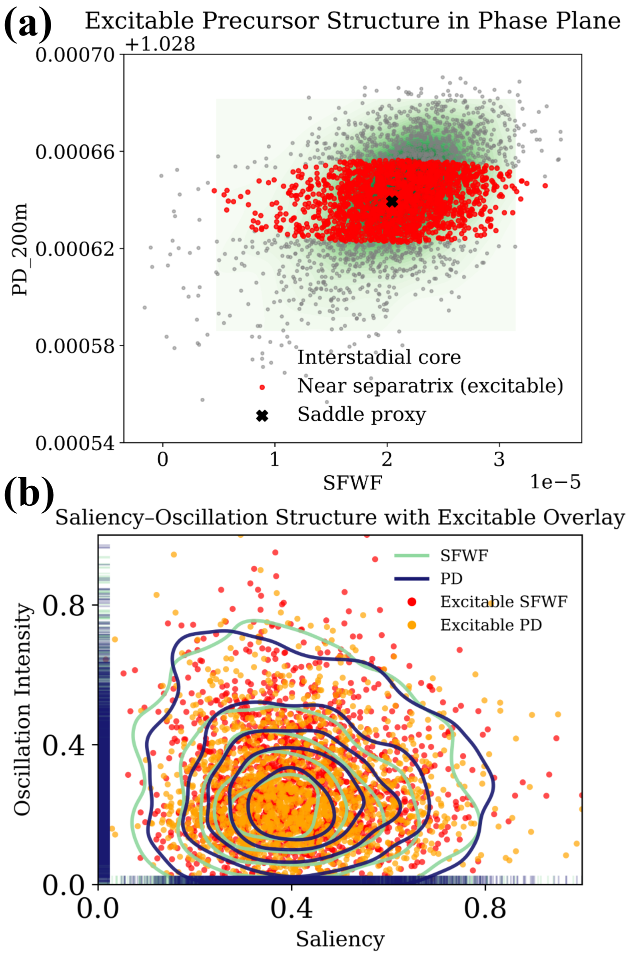


Figure 3: (a) Phase-plane structure showing excitable precursors near the separatrix and the saddle proxy. (b) Saliency–oscillation structure for SFWF and PD, showing the distribution of excitable states. The overlap indicates that excitable states are concentrated within regions of high saliency–oscillation density, suggesting that they align with the dominant dynamical regime of the system as captured by the model, where critical transitions are likely to emerge.

We compute time-resolved saliency using input gradients from the trained CNN model, and oscillation intensity using the sliding standard deviation and Hilbert envelope methods (Ouergli 2002; Feldman 2011; Oppenheim and Schafer 1989). The excitable region is computed by estimating the distance to the saddle proxy in the SFWF–PD phase plane, identified via KMeans clustering (Ikotun et al. 2023; Jin and Han 2010). The distance to the separatrix is computed as the Euclidean distance between each system state in phase space and the saddle proxy, estimated from the mean of the cluster

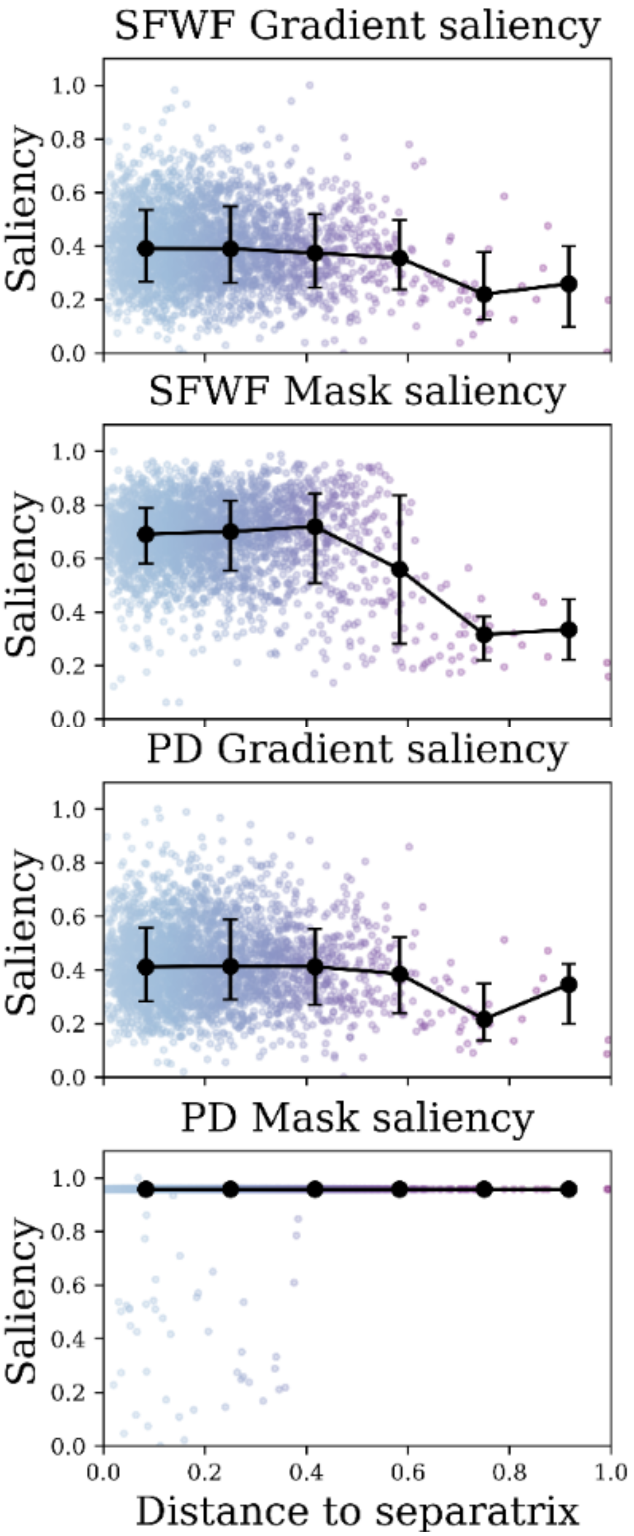


Figure 4: Saliency versus distance to the separatrix, showing increasing sensitivity near the saddle region (median ± 16–84% quantiles).

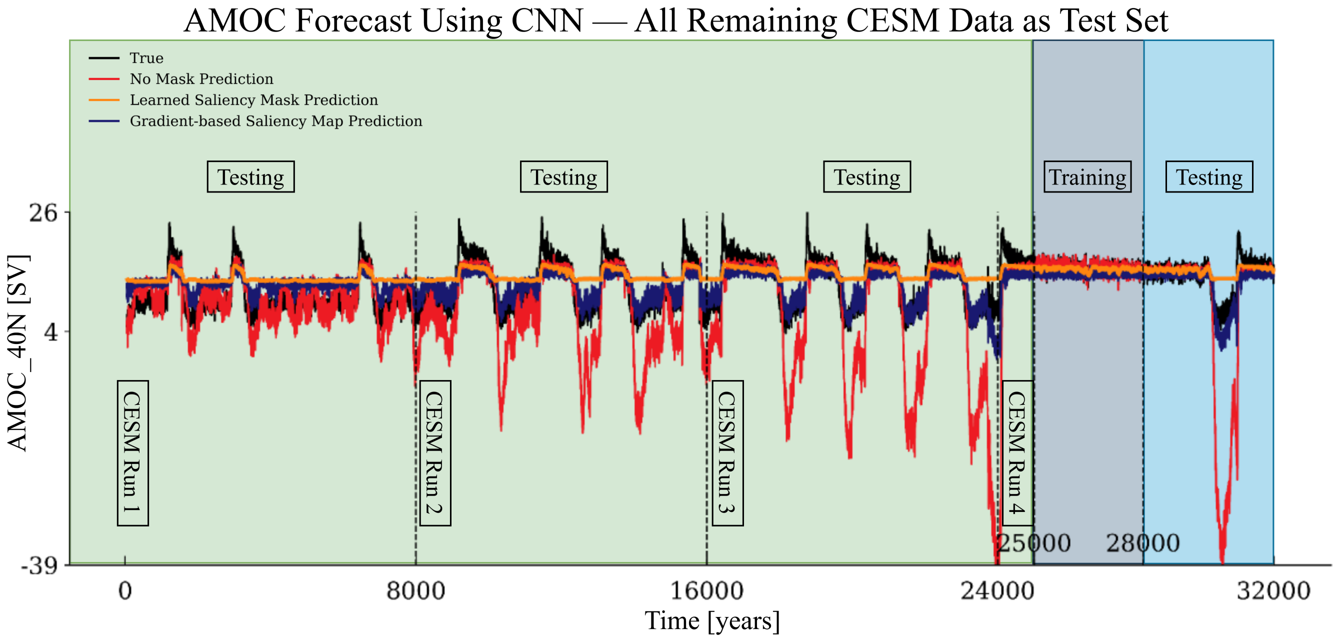


Figure 5: AMOC forecast using a CNN model applied to all remaining CESM runs as test data.

centers. This provides a quantitative measure of how close each state is to the separatrix, indicating excitable potential and guiding where the model assigns high saliency.

Both gradient and learned mask saliency for SFWF decrease as the distance to the separatrix increases, indicating that the model assigns higher saliency to states closer to the separatrix (i.e., more excitable states) (Figure. 4). It should be mentioned that for PD_200m, the overall saliency magnitude in the learned mask is lower, suggesting weaker model sensitivity to PD compared to SFWF (Figure. 2(b), Figure. 4). The process of analyzing AMOC dynamics through the model saliency described above in this study is given by Equation (A.9) to Equation (A.18) in Appendix A.4.

To conclude, the overlap between excitable states and the joint saliency–oscillation density indicates that excitable states predominantly cluster in regions of mid saliency and low oscillation intensity, suggesting that the system approaches critical transitions through subtle, early-stage precursors rather than strong oscillatory signals (Figure. 3). In addition, the distance-to-separatrix analysis reveals that the model assigns higher saliency to dynamically sensitive states near the separatrix (Figure. 4). These results suggest that the model not only forecasts the collapse in a dynamically consistent manner but also provides interpretable insights into how excitable states emerge as early-warning indicators of critical transitions.

IV Summary and Outlook

To summarize, our CNN model successfully predicts the AMOC collapse even though it is trained solely on interstadial time series without any D–O events. By incorporating both gradient-based and learned mask saliency maps, we not only enhance the forecasting performance but also provide a physically interpretable view of the black-box model. In particular, the learned saliency mask enables the model to predict the collapse onset with high temporal accuracy. Furthermore, the dynamical analysis reveals that states with high saliency are concentrated near the separatrix, corresponding to excitable regions in the phase space where the system is most sensitive to perturbations. This suggests that the model leverages these excitable precursors as early-warning indicators of critical transitions.

Overall, this work demonstrates a data-driven and interpretable framework for predicting AMOC collapses, bridging machine learning saliency methods with dynamical systems analysis. So far, the trained CNN models can predict all D-O events in our rest dataset (Figure. 5, Appendix A.3). A systematic investigation of the statistical properties of the forecasts, such as the maximum reliable forecast horizon, the amount of training data required, and the uncertainty in onset-time prediction, will be left for future work.

Moreover, as identified in Sec. III, the system's sensitivity near the separatrix, along with the clustering of excitable states at mid saliency and low oscillation intensity indicates inherent excitability and susceptibility to perturbations. High saliency on excitable regions enables out-of-sample extrapolation of large events. Incorporating noise analysis into this framework could further enhance forecast robustness, and thus improving the feasibility of long-term prediction and practical applications.

References

Azam, S.; Montaha, S.; Fahim, K. U.; Rafid, A. R. H.; Mukta, M. S. H.; and Jonkman, M. 2023. Using feature maps to unpack the CNN 'Black box' theory with two medical datasets of different modality. *Intelligent Systems with Applications*, 18: 200233.

Brunton, S. L.; and Kutz, J. N. 2019. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge: Cambridge University Press. ISBN 1108422098.

Capron, E.; Rasmussen, S. O.; Popp, T. J.; Erhardt, T.; Fischer, H.; Landais, A.; Pedro, J. B.; Vettoretti, G.; Grinsted, A.; Gkinis, V.; Vaughn, B.; Svensson, A.; Vinther, B. M.; and White, J. W. C. 2021. The anatomy of past abrupt warmings recorded in Greenland ice. *Nature communications*, 12(1): 2106–12.

Cong, Y.; Yuan, J.; and Liu, J. 2013. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 46(7): 1851–1864.

Dabkowski, P.; and Gal, Y. 2017. Real Time Image Saliency for Black Box Classifiers. In *Neural Information Processing Systems*.

Dansgaard, W.; Johnsen, S. J.; Clausen, H. B.; Dahl-Jensen, D.; Gundestrup, N. S.; Hammer, C. U.; Hvidberg, C. S.; Steffensen, J. P.; Sveinbjörnsdottir, A. E.; Jouzel, J.; and Bond, G. 1993. Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature (London)*, 364(6434): 218–220.

Dijkstra, H. A.; and van Westen, R. M. 2025. The Probability of an AMOC Collapse Onset in the Twenty-First Century. *Annual review of marine science*.

Erhardt, T.; Capron, E.; Rasmussen, S. O.; Schüpbach, S.; Bigler, M.; Adolphi, F.; and Fischer, H. 2019. Decadal-scale progression of the onset of Dansgaard–Oeschger warming events. *Climate of the Past*, 15(2): 811–825.

Feldman, M. 2011. *Hilbert Transform Applications in Mechanical Vibration*. Newark: Wiley, 1. aufl. edition. ISBN 9780470978276.

Gerber, L.; Lippold, J.; Süfke, F.; Valk, O.; Testorf, P.; Ehnis, M.; Tautenhahn, S.; Max, L.; Chiessi, C. M.; Regelous, M.; Szidat, S.; Friedrich, O.; and Pöppelmeier, F. 2025. Low variability of the Atlantic Meridional Overturning Circulation throughout the Holocene. *Nature communications*, 16(1): 6748–12.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5).

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T. L.; Parajuli, S.; Guo, M.; Song, D. X.; Steinhardt, J.; and Gilmer, J. 2020. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8320–8329.

Ikotun, A. M.; Ezugwu, A. E.; Abualigah, L.; Abuhaija, B.; and Heming, J. 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622: 178–210.

Jin, X.; and Han, J. 2010. *K-Means Clustering*, 563–564. Boston, MA: Springer US. ISBN 978-0-387-30164-8.

Jochum, M.; Chase, Z.; Nuterman, R.; Pedro, J.; Rasmussen, S.; Vettoretti, G.; and Zheng, P. 2022. Carbon Fluxes during Dansgaard–Oeschger Events as Simulated by an Earth System Model. *Journal of Climate*, 35(17): 5745 – 5758.

Kuhlbrodt, T.; Griesel, A.; Montoya, M.; Levermann, A.; Hofmann, M.; and Rahmstorf, S. 2007. On the driving processes of the Atlantic meridional overturning circulation. *Reviews of Geophysics*, 45(2).

Mitrea, C. A.; Lee, C. K. M.; and Wu, Z. 2009. A Comparison between Neural Networks and Traditional Forecasting Methods: A Case Study. *International Journal of Engineering Business Management*, 1: 11.

Mudelsee, M. 2000. Ramp function regression: a tool for quantifying climate transitions. *Computers Geosciences*, 26(3): 293–307.

Oppenheim, A. V.; and Schafer, R. W. 1989. *Discrete-time signal processing / Alan V. Oppenheim, Ronald W. Schafer.* Prentice-Hall signal processing series. Englewood Cliffs (NJ): Prentice-Hall. ISBN 0-13-216292-X.

Ouergli, A. 2002. Hilbert Transform from Wavelet Analysis to Extract the Envelope of an Atmospheric Mode: Examples. *Journal of atmospheric and oceanic technology*, 19(7): 1082–1086.

Pan, Q.; Hu, W.; and Chen, N. 2021. Two Birds with One Stone: Series Saliency for Accurate and Interpretable Multivariate Time Series Forecasting. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2884–2891. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Pantiskas, L.; Verstoep, K.; and Bal, H. 2020. Interpretable Multivariate Time Series Forecasting with Temporal Attention Convolutional Neural Networks. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1687–1694.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378: 686–707.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951. Florence, Italy: Association for Computational Linguistics.

Slattery, J.; Sime, L. C.; Muschitiello, F.; and Riechers, K. 2024. Estimating biases during detection of leads and lags between climate elements across Dansgaard–Oeschger events. *Climate of the Past*, 20(11): 2431–2454.

Vettoretti, G. 2022. guidov/Vettoretti_et_al_2022-NG: Vettoretti_et_al_2022.

Vettoretti, G.; Ditlevsen, P.; Jochum, M.; and Rasmussen, S. O. 2022. Atmospheric CO2 control of spontaneous millennial-scale ice age climate oscillations. *Nature geoscience*, 15(4): 300–306.

Wang, B.; Ma, R.; Kuang, J.; and Zhang, Y. 2020. How Decisions Are Made in Brains: Unpack “Black Box” of CNN With Ms. Pac-Man Video Game. *IEEE Access*, 8: 142446–142458.

Wu, Q.-F.; Jochum, M.; Avery, J. E.; Vettoretti, G.; and Nuterman, R. 2025. Machine Guided Derivation of the Atlantic Meridional Overturning Circulation (AMOC) Strength. *Geophysical Research Letters*, 52(3): e2024GL113454. E2024GL113454 2024GL113454.

Wunsch, C.; and Heimbach, P. 2013. Two Decades of the Atlantic Meridional Overturning Circulation: Anatomy, Variations, Extremes, Prediction, and Overcoming Its Limitations. *Journal of Climate*, 26(18): 7167 – 7186.

Yang, S.; Kim, H.; Hong, Y.; Yee, K.; Maulik, R.; and Kang, N. 2024. Data-driven physics-informed neural networks: A digital twin perspective. *Computer Methods in Applied Mechanics and Engineering*, 428: 117075.

Zhou, S.; Shen, W.; Zeng, D.; Fang, M.; Wei, Y.; and Zhang, Z. 2016. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*, 47: 358–368.

A Appendix

In this appendix, we present additional details and supplementary discussions of the dataset and model.

A.1 Dataset

This work uses the annual time series of three key variables (SFWF, PD_200m, and AMOC) from four 8000-year long Community Earth System Model (CESM) simulations with glacial boundary conditions that reproduce the observed structure of D–O events (Vettoretti et al. 2022; Vettoretti 2022; Wu et al. 2025).

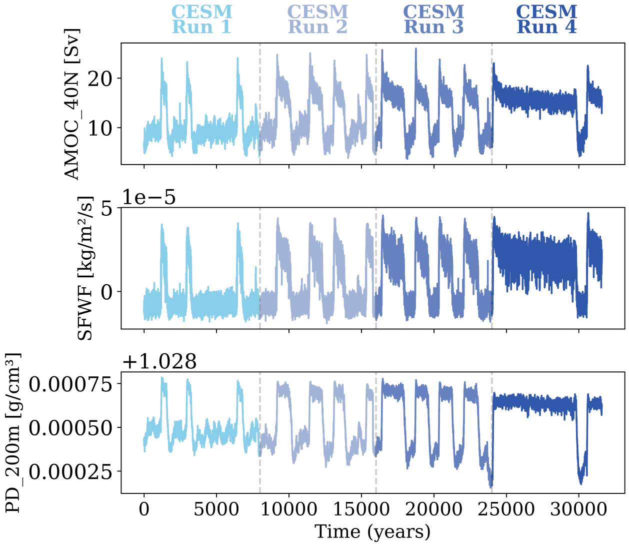


Figure A.1: The annual time series of three key variables (SFWF, PD_200m, and AMOC) from four CESM simulations reproduce the observed structure of D–O events, characterized by abrupt upward transitions and relatively gradual downward transitions.

In this work, only the time series from CESM Run 4 are used. The time index from the start of CESM Run 4 to 27,999 is used as the training set, and from 28,000 to the end as the test set.

A.2 Loss Function

In this section, we present our custom loss function that is designed to enforce both local and global coherence in predictions by integrating multiple constraints on the output dynamics. It begins by computing the first derivative of the predicted and true sequences, identifying abrupt transitions through a 2-sigma threshold. A penalty is then introduced for incorrect sign flips in these transitions, ensuring that predicted changes align with observed variations. To maintain integral consistency, the cumulative sum of predictions is compared to the true values, penalizing discrepancies that accumulate over time. Additionally, a smoothness constraint is imposed by minimizing the difference in second derivatives, discouraging abrupt oscillations and promoting gradual transitions. The final loss function Equation (A.8) balances these terms, combining squared error with structured penalties to refine the model's ability to capture both short-term fluctuations and long-term trends.

$$
\begin{aligned}
\mathcal{L} = & \underbrace{\mathbb{E}\left[\|y_{\text{true}} - y_{\text{pred}}\|^2\right]}_{\text{MSE}} \\
& + \underbrace{10\mathbb{E}\left[|\text{sign}(\Delta y_{\text{true}}) - \text{sign}(\Delta y_{\text{pred}})| \cdot 1_{\text{abrupt}}\right]}_{\text{Sign Flip Penalty (Abrupt Transitions)}} \\
& + \underbrace{5\mathbb{E}\left[\left\|\int_0^t y_{\text{true}}(s)ds - \int_0^t y_{\text{pred}}(s)ds\right\|^2\right]}_{\text{Integral Consistency Penalty}} \\
& + \underbrace{2\mathbb{E}\left[\left|\Delta^2 y_{\text{true}} - \Delta^2 y_{\text{pred}}\right|\right]}_{\text{Smoothness Penalty (Second Derivative)}}
\end{aligned}
\tag{A.8}
$$

The primary objective is captured by the MSE, defined as $\mathbb{E}\left[\|y_{\text{true}} - y_{\text{pred}}\|^2\right]$, which ensures that the predicted values y_{pred} closely follow the true values y_{true}. To regulate abrupt transitions, we introduce a sign-flip penalty, $\mathbb{E}\left[|\text{sign}(\Delta y_{\text{true}}) - \text{sign}(\Delta y_{\text{pred}})| \cdot 1_{\text{abrupt}}\right]$, where $\Delta y_t = y_t - y_{t-1}$ represents the first derivative and 1_{abrupt} is an indicator function that detects rapid changes. To preserve integral consistency, the term $\mathbb{E}\left[\left\|\int_0^t y_{\text{true}}(s)ds - \int_0^t y_{\text{pred}}(s)ds\right\|^2\right]$ penalizes discrepancies in the cumulative sum of predictions over time. Smoothness is enforced via the second derivative penalty, $\mathbb{E}\left[\left|\Delta^2 y_{\text{true}} - \Delta^2 y_{\text{pred}}\right|\right]$, where $\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$, mitigating sudden oscillations and promoting gradual transitions.

By using the custom loss function Equation (A.8) in the training process, our CNN model can successfully predict AMOC collapse with the correct sign when using only interstadial SFWF or PD_200m as input. In contrast, using the Mean Squared Error (MSE) loss leads to predictions that also capture sudden transitions but with inverted signs (Figure A.2).

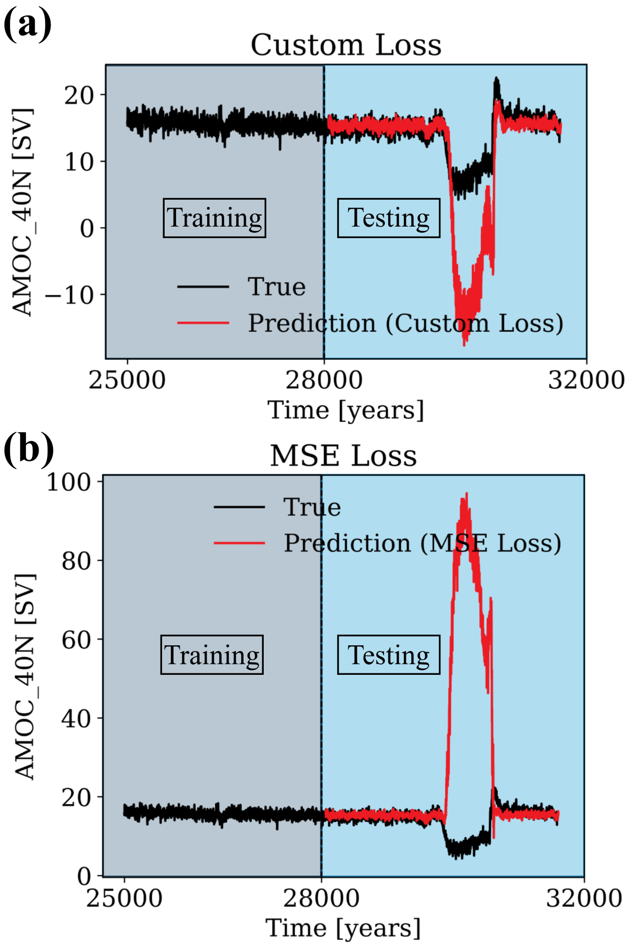


Figure A.2: Illustration of the proposed custom loss function structure and its effect on prediction sign correctness. (a) Prediction using the custom loss function accurately captures the collapse sign. (b) Prediction using the MSE loss captures transitions but with inverted signs.

A.3 Ramp Fitting Algorithms

In this study, we use the ramp fitting algorithms in the following Algorithm 1. We also applied this ramp-fitting algorithm to three transitions from different CESM runs, shown in Figure 5, as examples. Future research may extend this approach to quantify model sensitivity and uncertainty, as well as to determine the minimal data requirements.

Algorithm 1: Ramp Fitting Procedure for Transition Detection

Input: Time series t, signal y, ramp length range $[L_{\min}, L_{\max}]$, direction (up, down, or both)
Output: Best-fit ramp parameters (t_0, t_1, a_0, a_1)

```
Initialize best score $S^* \leftarrow -\infty$
for $i = 1$ to $|t| - L_{\min}$ do
    for $j = i + L_{\min}$ to $\min(i + L_{\max}, |t|)$ do
        Set candidate interval $(t_0, t_1) \leftarrow (t_i, t_j)$
        Fit piecewise linear ramp $R(t; t_0, t_1, a_0, a_1)$ to $y[i : j]$
        Compute ramp height $H = |a_1 - a_0|$ and slope $s = (a_1 - a_0)/(t_1 - t_0)$
        if direction constraints satisfied and $H > S^*$ then
            $S^* \leftarrow H$
            $(t_0^*, t_1^*, a_0^*, a_1^*) \leftarrow (t_0, t_1, a_0, a_1)$
        end if
    end for
end for
return $(t_0^*, t_1^*, a_0^*, a_1^*)$
```

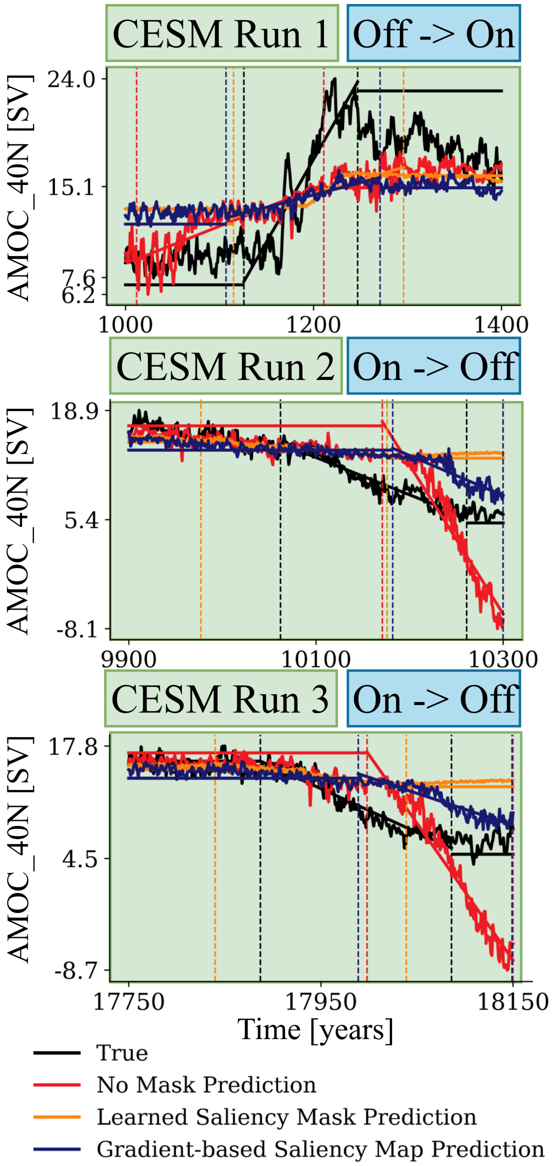


Figure A.3: Examples of ramp-fitting applied to three transitions from different CESM runs.

A.4 Hilbert Signal Envelope Method

For each input feature $f \in \{\mathrm{SFWF}, \mathrm{PD\_200m}\}$ and sliding window $X^{(i)} \in \mathbb{R}^{T\times F}$, the oscillation intensity is quantified using two complementary metrics. First, the local mean is computed as

$$\mu^{(i)}_{t,f} = \frac{1}{T} \sum_{\tau=t-\frac{T}{2}}^{t+\frac{T}{2}} X^{(i)}_{\tau,f}, \tag{A.9}$$

and the sliding standard deviation is then given by

$$\sigma^{(i)}_{t,f} = \sqrt{\frac{1}{T} \sum_{\tau=t-\frac{T}{2}}^{t+\frac{T}{2}} \left(X^{(i)}_{\tau,f} - \mu^{(i)}_{t,f}\right)^2}. \tag{A.10}$$

Next, the Hilbert envelope amplitude is computed from the analytic signal

$$z^{(i)}_{t,f} = X^{(i)}_{t,f} + i\,\hat{X}^{(i)}_{t,f}, \tag{A.11}$$

$$A^{(i)}_{t,f} = |z^{(i)}_{t,f}| = \sqrt{\left(X^{(i)}_{t,f}\right)^2 + \left(\hat{X}^{(i)}_{t,f}\right)^2}, \tag{A.12}$$

where $\hat{X}^{(i)}_{t,f}$ is the Hilbert transform of $X^{(i)}_{t,f}$

Finally, the oscillation intensity is expressed as

$$O^{(i)}_{t,f} \in \{\sigma^{(i)}_{t,f},\, A^{(i)}_{t,f}\}, \tag{A.13}$$

providing a time-resolved measure of oscillatory strength aligned with the saliency maps $\bar{S}_{t,f}$ and $M_{t,f}$.

In this work, we further quantify the system's dynamical sensitivity by estimating the distance of each state to a saddle proxy in the SFWF–PD phase plane using KMeans clustering. To identify the saddle proxy in the SFWF–PD phase plane, the state vectors $\mathbf{x}_n = \begin{bmatrix} \mathrm{SFWF}_n \\ \mathrm{PD}_{200m,n} \end{bmatrix} \in \mathbb{R}^2$ are clustered into two groups using KMeans. The cluster centers are

$$\mathbf{c}_j = \frac{1}{N_j} \sum_{\mathbf{x}_n \in \mathcal{C}_j} \mathbf{x}_n, \qquad j = 1, 2, \tag{A.14}$$

where $\mathcal{C}_j$ is the j-th cluster and N_j is the number of points in it. The saddle proxy is then estimated as the mean of the two cluster centers,

$$\mathbf{s} = \frac{1}{2}\big(\mathbf{c}_1 + \mathbf{c}_2\big). \tag{A.15}$$

The Euclidean distance of each state to the saddle proxy is given by

$$d_n = \big\|\mathbf{x}_n - \mathbf{s}\big\|_2, \tag{A.16}$$

which is then normalized to

$$\tilde{d}_n = \frac{d_n - d_{\min}}{d_{\max} - d_{\min}}. \tag{A.17}$$

In this study, we define excitable regions as the states that are geometrically close to a saddle proxy, within the 20% closest shell around it,

$$\text{Excitable region} = \left\{ \mathbf{x}_n \;\middle|\; \frac{\big\|\mathbf{x}_n - \mathbf{s}\big\|_2 - d_{\min}}{d_{\max} - d_{\min}} < 0.2 \right\}. \tag{A.18}$$

This provides a quantitative measure of how close each system state is to the separatrix, thereby allowing us to examine how model saliency varies with dynamical sensitivity.