Predicting the distance of the AMOC to its tipping point using CNNs

Submitted to

Nonlinear Processes in Geophysics

Francesco Guardamagna[1], Sacha Sinet[1] and Henk A. Dijkstra[1,2]

[1]Institute for Marine and Atmospheric Research Utrecht (IMAU), Utrecht, The Netherlands
[2] Center for Complex Systems Studies (CCSS), Utrecht, The Netherlands

Overall response

We are pleased to submit our responses to the comments provided by the reviewers for our manuscript entitled *"Predicting the distance of the AMOC to its tipping point using CNNs"* (manuscript number EGUSPHERE-2026-1872).

We would like to sincerely thank the Editor, Balasubramanya Nadiga, and all three reviewers for their careful evaluation of our manuscript and for their constructive and insightful comments. We greatly appreciate the time and effort devoted to reviewing our work. Their feedback has helped us identify several aspects of the manuscript that can be improved. In this document, we provide a detailed point-by-point response to each comment and describe the corresponding revisions that will be incorporated into the revised manuscript.

One of the main points raised by Reviewer 1, and similarly by Reviewers 2 and 3, concerns the use of a simple linear regressor as the baseline model. We fully understand and appreciate this concern and agree that, in principle, comparisons against other baselines are desirable.

However, to the best of our knowledge, there are currently no established benchmark models or widely adopted baseline methods for the specific prediction task considered in this study. Consequently, selecting an alternative baseline would necessarily be somewhat arbitrary and would not constitute a comparison against a recognized reference approach. Moreover, adding a single arbitrarily selected baseline would provide limited additional scientific insight and is unlikely to affect the main conclusions of the manuscript.

For these reasons, we do not intend to include additional baseline models in the revised manuscript. Instead, we will strengthen the discussion of the limitations of the Linear Regression baseline and explicitly acknowledge its role as a simple reference model.

While preparing our response, we identified a minor imprecision in the manuscript. The freshwater hosing value corresponding to the CLIMBER-X tipping point is not exactly $F_H^C = 0.22$ Sv, as currently stated, but rather $F_H^C \approx 0.22$ Sv, with the precise value being $F_H^C = 0.2189$ Sv. This correction does not affect any of the analyses or conclusions of the manuscript and will be incorporated into the revised version.

Authors' Response to Reviewer 1

General Comments. The Atlantic meridional overturning circulation (AMOC) is known to exhibit multiple equilibria and tipping points, which can be observed in climate model simulations where a freshwater flux is applied in the North Atlantic. In this draft, the authors propose inferring the accumulated freshwater flux applied in the North Atlantic from the patterns of sea surface temperature, sea surface salinity and salinity cross-section at 35°S in the Atlantic Ocean. This inference uses data from CLIMBER-X (an earth system model of intermediate complexity) and CESM (a low-resolution climate model). The accumulated freshwater flux is normalized in each model by an estimate of the total accumulated freshwater flux required to trigger a tipping point. Here, the method employs convolutional neural networks (CNNs). The results obtained are discussed using a sensitivity analysis.

The draft presents original research; however significant work is needed to make it publishable. Specifically, the scientific question and the research hypotheses are not clearly presented in the introduction. The results presented are overly technical and lack synthesis. The sensitivity analysis fails to reveal consistent patterns. The conclusion and discussion are underdeveloped.

Response:

We thank the reviewer for carefully reading our manuscript and for the constructive and thoughtful comments. We appreciate the time and effort devoted to evaluating our work. Below, we provide a detailed point-by-point response to each of the reviewer's comments and describe the corresponding changes that will be incorporated into the revised manuscript.

Major comments

Comment 1.1

The data from only two models are used. How representative are these two models? Are their outputs consistent with those of other models concerning the AMOC tipping and its SST and salinity signatures?

Response:

According to [1], CLIMBER-X exhibits a response to different forcing scenarios and boundary conditions that is comparable to state-of-the-art CMIP6 models. Moreover, the model reproduces the present-day AMOC overturning profile at 26°N in good agreement with observations. AMOC stability in CLIMBER-X has been extensively investigated in previous work [2], demonstrating that the model reproduces the characteristic AMOC hysteresis behavior found across a hierarchy of climate models [3, 4]. These characteristics make CLIMBER-X well suited for investigating the long-term stability of the AMOC under freshwater forcing. The main limitation of the model is its inability to adequately reproduce interannual climate variability, owing to its coarse spatial resolution and simplified atmospheric dynamics [1]. However, these limitations do not affect its suitability for studying long-term AMOC stability.

Furthermore, the Atlantic SST and SSS anomalies associated with AMOC weakening in CLIMBER-X are consistent with those simulated in North Atlantic freshwater-hosing experiments using other coupled global climate models [5, 6].

CESM is a state-of-the-art coupled global climate model that has been extensively used to investigate AMOC stability under North Atlantic freshwater forcing [4, 7, 8]. In addition, the Atlantic SST and SSS response to AMOC weakening simulated by CESM is consistent with that obtained in CLIMBER-X and in freshwater-hosing experiments using other coupled global climate models [5, 6].

To clarify these points, we will revise the manuscript to better motivate the choice of CLIMBER-X and CESM for this study. We already compare the SST and SSS responses of the models with the SST and SSS trends reported by Stouffer et al. [5]. We will also compare these responses with the more recent results of Pontes and Menviel [6].

Comment 1.2

Is the accumulated freshwater flux leading to a tipping AMOC well estimated in the simulations? The manuscript lacks clear explanations of how the AMOC collapse or tipping is defined or estimated. Would it be valuable to incorporate some uncertainties?

Response:

In the present study, the reference AMOC tipping point is estimated from a freshwater hosing simulation performed with a very slow forcing rate of 10^{-5} Sv yr^{-1}. This slow forcing rate is used to approximate the quasi-equilibrium response of the system and thereby obtain a reliable estimate of the tipping point.

For this simulation, we identify the tipping point at $F_H^C = 0.2189$ Sv. At this forcing value, the AMOC undergoes a clear and abrupt transition toward the collapsed state (see Figure 2a), characterized by a pronounced change in the slope of the AMOC-strength curve.

In the revised manuscript, we will explicitly explain the criterion used to determine the tipping point and discuss the limitations of the adopted methodology, emphasizing that the identified tipping point is an approximation of the true bifurcation point.

Comment 1.3

Data from hosing experiments may differ significantly from real-word. For instance, hosing experiments often involve a shallowing of the mixed layer in the hosing region, whereas rea-world freshwater flux from melting ice sheets has a different pattern. This limitation needs to be discussed, as the machine learning method developed may fail when applied to observational data.

Response:

We are aware of this limitation and acknowledge that the stability of the real AMOC is governed by a range of processes that are far more complex than those represented in our framework, which currently estimates the distance to the AMOC tipping point only in terms of the accumulated freshwater forcing applied in the simulations. Moreover, the synthetic data used to train the machine-learning model cannot perfectly reproduce the processes occurring in the real climate system. Consequently, we did not apply the framework to observational data.

We already discussed this limitation in the Conclusions section, but we will expand this discussion in the revised manuscript to make this point clearer and to emphasize that the proposed framework should currently be regarded as a proof of concept based on idealized freshwater-hosing experiments.

Comment 1.4

A linear regression model is given as a baseline. However, given the high dimensionality of the input data, the linear regression model may overfill. I suggest using a ridge regression or a random forest model instead.

Another point concerns the normalization. L243 : ''CESM inputs are normalized using the mean and standard deviation of the CLIMBER-X training data" In climate science, the model data often show large differences, as they all have different biases. I understand that the mean state of CLIMBER-X was removed from CESM data to define anomalies. If this is correct, this approach would emphasize the difference between CESM and CLIMBER-X rather than the distance to tipping. A better approach might be to use anomalies defined with a reference period relative to each model. Additionally, dividing by the standard deviation of CLIMBER-X could lead to a high variability in CESM data if the standard deviation of CLIMBER-X is much smaller than the one of CESM. Can the authors compare the standard deviations of the two models?

Response:

Regarding the suggestion of using ridge regression or a random forest as an additional baseline, we refer to the explanation provided in the overall response, where we discuss the rationale for not including additional baseline models in the revised manuscript.

In our transfer-learning framework, CLIMBER-X represents the source domain and CESM the target domain. Consequently, the CESM data are normalized using the CLIMBER-X training statistics to ensure that the model is applied under the same feature scaling used during training. We acknowledge that this does not remove the intrinsic biases between the two climate models. On the contrary, these differences are intentionally preserved, making the transfer-learning task more challenging and providing a more stringent assessment of the model's ability to generalize across models.

To clarify this point, we will revise the manuscript to better explain the rationale behind the adopted normalization strategy and explicitly discuss its implications for cross-model transfer.

Comment 1.5

The authors should explicitly refer to each figure panel, line and symbol to help reader verify statements and hypotheses. Currently, it is difficult to follow what is being discussed. Are the claims based on figures, hypotheses or suggestions? The results are presented as an exhaustive list of figures, which is overwhelming. I suggest reducing the figures. For instance, Figs. 6, 7, 8, 9 and 10 can be reduced to show the sensitivity analysis for a signle freshwater forcing and CLIMBER-X, with the other results summarized in the text. The same remark applies to appendices B and C that could be removed or significantly condensed. For appendix C, graphs might be more effective than tables for visual clarity.

The unclear presentation and difficult readability made the review challenging, so I primarily focused on the first 3 sections for minor comments.

Response:

In the revised manuscript, we will explicitly refer to the relevant figure panels, lines, and symbols throughout the Results section, making it easier for the reader to identify exactly which elements of each figure support the corresponding statements. We will also revise the text to more clearly distinguish statements directly supported by the figures from our interpretation of those results.

Regarding the number of figures, we agree that the presentation can be better streamlined. Figure 10 already reports the explainability analysis for the CESM experiment, which consists of a single freshwater-forcing simulation, with the two panels corresponding to two different input variables. For the CLIMBER-X experiments, the explainability analysis is currently shown for the slowest and fastest freshwater-forcing rates considered. In the revised manuscript, we will retain only the results for the slowest forcing rate ($10^{-5}\,\mathrm{Sv\,yr^{-1}}$) and discuss the results obtained for the other forcing rates in the text.

Regarding Appendices B and C, we believe that the reported analyses provide useful complementary information. Appendix B presents an additional assessment of the CNN generalization capabilities, while Appendix C summarizes the predictive performance across all experimental configurations considered in our study. Given the amount of information reported, we believe that replacing the tables in Appendix C with graphical representations would not necessarily improve readability and could instead make the comparison more difficult. However, we will condense both Appendices B and C and focus on only essential results.

Comment 1.6

The training, validation and testing strategy concerning the CESM results is unclear and needs to be clarified.

Response:

We agree that the description of the training, validation, and testing strategy for the CESM experiments can be improved. In the revised manuscript, we will improve this section to provide a clearer explanation of the adopted methodology.

Comment 1.7

The introduction needs to better explain the experiments used in the paper, and their limitations. The conclusion needs to provide comparisons with related recent findings, suggest future perspectives and discuss limitations.

Response:

In the revised manuscript, we will improve the description of the experiments and their limitations in the Introduction, and expand the Conclusions to include a more comprehensive discussion of the limitations of our study, comparisons with recent related work, and possible directions for future research.

Minor comments

Comment 1.8

L37-38: "First, it requires prior knowledge about the freshwater forcing rate, which is a significant constraint for real-world applications." The term freshwater forcing is ambiguous. Clarify whether it refers to a freshwater imposed from in hosing experiments, or the climatological freshwater forcing from the surface water budget (e.g., precipitation minus evaporation and runoff)?

Response:

Here, we refer specifically to the imposed freshwater forcing used in the hosing experiments, rather than the climatological freshwater forcing. We will clarify this terminology in the revised manuscript to avoid any ambiguity.

Comment 1.9

L38-30 : " the model must be trained on data from the simulation itself, extending up to 100 years before the onset of collapse. " Please specify the total number of year used, the physical field or time series used in the training data.

Response:

In [9], the authors applied their AMOC tipping prediction framework to the same CESM tipping simulation analyzed in our study. Although the exact amount of training data is not explicitly reported in the paper, it can be inferred from their Fig. 6, which illustrates the training and validation split, that the machine-learning framework was trained using approximately 1650 years of data from the simulation itself. As we specify in line 39 of our manuscript, the input variable is the AMOC strength time series at 26°N. In addition, the model is provided with the imposed freshwater-forcing rate as an input feature. We will clarify these details in the revised manuscript.

Comment 1.10

L39-41: ‘’ as the model relies solely on one-dimensional indices, such as the AMOC strength at $26 \circ$N as input, it precludes the application of explainable AI techniques

‘’ Why do the authors argue IA technique cannot be applied to indices? Provide justification or revise the statement

Response:

We agree that the original statement was not sufficiently precise. Explainable AI techniques can indeed be applied to models trained on one-dimensional indices. However, when the input consists solely of one-dimensional time series, such as the AMOC strength at 26°N, the explainability analysis is necessarily limited to identifying the importance of temporal features. In contrast, our framework is trained on 2-dimensional SST and SSS fields, allowing explainable AI techniques to identify the spatial patterns that the model exploits to estimate the distance to the AMOC tipping point. This provides a richer and more physically interpretable analysis of the underlying processes associated with AMOC weakening than can be obtained from one-dimensional indices alone. We will revise the manuscript to clarify this distinction.

Comment 1.11

L68-69:‘’ matches many aspects of state-of-the-art CMIP6 models across diverse forcings and boundary conditions (Willeit et al. (2022a)). ‘’ Can the authors be more specific and describe which processes have been validated and for what types of experiments?

Response:

We agree that the original statement was too general. In the revised manuscript, we will specify that CLIMBER-X has been evaluated against state-of-the-art CMIP5 and CMIP6 models, as well as observations, across a range of experiments and boundary conditions. In particular, the model performs similarly to state-of-the-art CMIP5 and CMIP6 models for standard CO_2 forcing experiments (abrupt-4×CO_2 and 1% yr^{-1} CO_2 increase), equilibrium climate sensitivity, transient climate response, climate feedbacks (including water vapour, lapse rate, cloud, and albedo feedbacks), hydrological sensitivity, and sea surface salinity responses to global warming. The model has also been evaluated for Last Glacial Maximum simulations and freshwater-hosing experiments investigating AMOC stability.

Comment 1.12

L74 : ‘’ the AMOC collapses once the freshwater forcing reaches $F_H^C = 0.22$ Sv”. How was the threshold determined? The manuscript does not provide sufficient explanation. Also define AMOC collapse. This term is loosely used in the literature. Specify the criteria for collapse (threshold, or detection of bifurcation). I suggest that the authors show and illustrate the results supporting this value.

Response:

As explained in our response to Comment 1.2, the value $F_H^C = 0.2189$ Sv is estimated from the freshwater hosing simulation with a forcing rate of 10^{-5} Sv yr^{-1}. At this forcing value, the AMOC undergoes a clear and abrupt transition from a vigorous circulation state to a collapsed state with an AMOC strength close to 0 Sv.

In the revised manuscript, we will clarify this point, describe the criterion used to identify the collapse, and explicitly discuss the limitations of the adopted methodology.

Comment 1.13

L78: ‘’The Community Earth System Model (CESM) is a fully coupled GCM. ‘’ Clarify the differences between CESM and CLIMBER-X. What processes are included in each model?

Response:

We will revise the manuscript to better clarify the differences between CESM and CLIMBER-X. Both CESM and CLIMBER-X are fully coupled Earth system models in which the atmosphere, ocean, sea ice, and land surface interact through the exchange of momentum, heat, and freshwater. However, they differ substantially in the complexity with which these processes are represented.

Comment 1.14

L86-87: ‘’ Under this forcing, van Westen et al. (2024a) estimated that the AMOC reaches its tipping point at model year 1758, when the freshwater input into the

North Atlantic reaches $F_H^E = 0.53$ Sv." How was the tipping point estimated? Is it equivalent to say that the AMOC collapsed and that the AMOC reached a tipping point? What does the model year 1758 refer to? Does the simulation start at year 0? Provide contextfor the timeline.

Response:

The AMOC tipping simulation performed in [8] is initialized from a statistically equilibrated pre-industrial control simulation of CESM, which corresponds to model year 0. Throughout the simulation, greenhouse gas concentrations, solar forcing, and aerosol forcing are kept constant, while the freshwater forcing in the Atlantic between 20°N and 50°N is gradually increased and compensated over the rest of the basin.

The AMOC tipping point is estimated by applying a breakpoint regression analysis to the AMOC strength time series. Specifically, the method identifies the model year at which the change in the slope of the AMOC time series is largest, corresponding to the onset of the transition towards the weak AMOC state. Using this approach, [8] identified model year 1758, at which the imposed freshwater forcing reaches 0.53 Sv, as the AMOC tipping point.

Reaching the tipping point is not equivalent to the collapse of the AMOC. Rather, model year 1758 marks the onset of the transition towards the collapsed (OFF) state, while the complete collapse occurs later in the simulation. We will revise the manuscript to clarify the method used to identify the tipping point, based on the breakpoint regression analysis of the AMOC time series, and provide additional context regarding the simulation timeline.

Comment 1.15

L112-113 and Fig. 1: ‘’ The CNN is trained using Sea Surface Temperature (SST) and Sea Surface Salinity (SSS) fields across the Atlantic Ocean (spanning from 90°N to 35 $\circ$ S)" Define the Atlantic Ocean boundaries. Most definition does not extend beyond 80°N (e.g., Fram straight). The Arctic included (see Fig. 1) justify this choice. Why do the authors choose to use SST and SST as input? Why not include subsurface ocean data, which may provide additional predictive skill?

Response:

The domain considered in this study extends from 35°S to 90°N and therefore includes the Arctic Ocean. We deliberately included the Arctic because it is dynamically connected to the North Atlantic. Exchanges of freshwater and heat through the Fram Strait, Davis Strait, and the Barents Sea Opening, together with dense-water formation in the Nordic Seas, play a key role in regulating the AMOC and its variability [10, 11]. Consequently, Arctic SST and SSS fields may contain additional information relevant for estimating the distance to the AMOC tipping point.

Regarding the choice of input variables, we selected SST and SSS because they are the most readily observable large-scale ocean variables and can be obtained from satellite and in situ observational products. While subsurface ocean variables could indeed provide additional predictive skill, they are much less comprehensively observed and would substantially reduce the applicability of the framework to observational datasets. We will clarify these motivations in the revised manuscript.

Comment 1.16

L114: ''the full-depth salinity profile at 35 $\circ$ S '' Do the authors mean that they used the cross-section at 35°S in the Atlantic Ocean in the depth-longitude space? Justify the choice of this latitude and its relevance to AMOC dynamics and tipping point.

Response:

We indeed refer to the full-depth Atlantic salinity cross-section at 35°S in the longitude–depth plane, and we will clarify this description in the revised manuscript.

The freshwater transport by the Atlantic Meridional Overturning Circulation at 34°S (F_{ovS}) has been proposed as a possible indicator of AMOC stability because it provides information on the strength of the salt-advection feedback [8, 12]. Hence, the salinity structure at this latitude can potentially provide valuable information on the stability of the AMOC.

Owing to the coarse spatial resolution of CLIMBER-X, the latitude closest to the standard 34°S Atlantic section used for AMOC diagnostics is 35°S. We therefore selected the full-depth salinity cross-section at 35°S as input to our framework. We will better clarify the choice of this latitude and its link to AMOC stability in the revised manuscript.

Comment 1.17

L129-130 ‘’ For all forcing rates, $d_F(t)$ is defined with respect to a freshwater flux at tipping $F_H(t_p) = F_H^C = 0.22Sv$, which corresponds to the tipping point identified for the slowest forcing experiment ‘’ How was the tipping point identified? Why use the same $F_H(t_p)$ for all forcing rates? Would the results differ if $F_H(t_p)$ varied? Just a question: would the results be different if $d_F(t)$ were defined as $(t - t_0)/(t_p - t)$, where t_0 is the time of the initial conditions?

Response:

The tipping point for the slowest forcing experiment ($r_f = 10^{-5}, \mathrm{Sv\,yr^{-1}}$) was identified directly from the AMOC strength time series (see Fig. 2a), where a clear change in slope marks the onset of an abrupt collapse. At $F_H(t_p) = F_H^C = 0.2189\,\mathrm{Sv}$, the AMOC undergoes a rapid transition towards the collapsed (OFF) state, marking the onset of the collapse. Since the freshwater forcing is increased very slowly, rate-induced effects are minimized and the system evolves close to quasi-equilibrium. Therefore, this experiment provides the best approximation of the intrinsic tipping point of the system.

The same value of $F_H(t_p)$ is adopted for all forcing rates because our objective is to estimate the distance to this intrinsic tipping point rather than to the onset of the transition in each individual simulation. For faster forcing rates, rate-induced effects cause the transition towards the collapsed state to occur before the intrinsic tipping point is reached. Consequently, adopting a different value of $F_H(t_p)$ for each forcing rate would change the target variable by incorporating rate-induced effects rather than measuring the distance to the underlying tipping point of the system.

Defining $d_F(t)$ in terms of time instead of freshwater forcing would also alter the interpretation of the target variable. Since different forcing rates require different times to reach the same freshwater forcing, a time-based definition would depend on the prescribed forcing rate rather than preserving a consistent normalization across experiments. Defining $d_F(t)$ in terms of accumulated freshwater forcing therefore provides a dimensionless quantity that is directly comparable across experiments with different forcing rates. We will better clarify all these aspects in the revised manuscript.

Comment 1.18

L130: Why is the unit of a freshwater forcing given in $Sv\,yr^{-1}$? Clarify the units. A freshwater flux is typical expressed in Sv. Is the forcing a rate of change?

Response:

The freshwater forcing itself is expressed in Sverdrups (Sv), while $\mathrm{Sv\,yr^{-1}}$ refers to the rate at which the imposed freshwater forcing is increased over time. To generate the seven CLIMBER-X simulations, a spatially uniform freshwater forcing was applied over the North Atlantic between 20°N and 50°N, with its magnitude increasing linearly in time. Therefore, the quantity expressed in $\mathrm{Sv\,yr^{-1}}$ is the rate of change of the imposed freshwater forcing, rather than the freshwater forcing itself. We will better clarify this distinction in the revised manuscript.

Comment 1.19

L135: ‘' then evaluated on the trajectory excluded from both training and validation ‘' Define trajectory. Does it refer to a time series from a single simulation? Please specify the data used for evaluation.

Response:

By trajectory, we refer to the complete sequence of samples generated by a single CLIMBER-X tipping simulation at a specific forcing rate. Our CNN framework was evaluated using a leave-one-out strategy across seven CLIMBER-X simulations obtained with different freshwater-forcing rates r_f. At each iteration, one complete simulation trajectory was excluded from both training and validation and used exclusively for testing, while the remaining six simulations were used for training and validation. We will better clarify this terminology in the revised manuscript.

Comment 1.20

L161-162: ‘' The reported predictions are the median across 20 independent training trials; variability across trials is negligible and therefore not shown. ‘' Justify the need for multiple trials, and reduce the number of trial if results are deterministic.

Response:

Multiple training trials are required because the CNN weights are randomly initialized before training, making the optimization process non-deterministic. Repeating the training, therefore, allows us to verify that the reported performance is robust with respect to different random initializations. We report the median prediction across 20 independent trials to reduce the influence of any individual initialization. Since the variability across trials is negligible, we do not report it in the manuscript. We will clarify this point in the revised version.

Comment 1.21

L157-159 : ‘’ The LR model is trained using the same input variables and target (dF), again following the procedure outlined in Section 2.4. ‘’ The LR model may overfit when using such high dimensional input. Did the authors try to apply regularization or to reduce the dimensionality of the input? If not, address this limitation and switch to a more robust baseline.

Response:

As explained in the overall response, we propose not to include an additional baseline in the revised manuscript.

Comment 1.22

L172-173: "corresponding to a prediction uncertainty of 961 years (9.61×10^{-3} Sv if we express the error in terms of freshwater forcing)" Clarify the calculation. How was the 961 years derived? How does this translate to 9.61 10^{-3} Sv?

Response:

For the slowest forcing experiment ($r_F = 10^{-5}\,\mathrm{Sv\ yr^{-1}}$), an MSE of 1.93×10^{-3} corresponds to a root mean squared error (RMSE) of $\sqrt{1.93 \times 10^{-3}} \approx 4.39 \times 10^{-2}$ in the normalized target variable. Since the interval between the beginning of the simulation and the tipping point spans 21,890 years, this corresponds to an uncertainty of $4.39 \times 10^{-2} \times 21,890 \approx 961$ years. Multiplying this time uncertainty by the applied forcing rate ($10^{-5}\,\mathrm{Sv\ yr^{-1}}$) yields an equivalent uncertainty in accumulated freshwater forcing of $961 \times 10^{-5} = 9.61 \times 10^{-3}$, Sv. We will clarify this calculation in the revised manuscript.

We also agree that reporting the prediction uncertainty in both years and accumulated freshwater forcing may be unnecessarily confusing. Since the target variable is defined in terms of accumulated freshwater forcing, in the revised manuscript we will report the prediction uncertainty only in Sv, providing a more direct and consistent measure of the model performance.

Comment 1.23

L179-180: "with very low percentage errors relative to the total span of the test simulations." Can the authors explain better? What is the length of the test simulation then?

Response:

We agree that the wording was ambiguous. By "the total span of the test simulations", we mean the interval from the initial condition to the tipping point in each test simulation. Using a leave-one-out strategy, the CNN is evaluated on seven different CLIMBER-X tipping simulations generated with freshwater-forcing rates ranging from 10^{-5} to 6×10^{-4} Sv yr^{-1}. Since the forcing rate differs among these simulations, the time required to reach F_H^C also varies; therefore, the total span of each test simulation is not the same.

As already discussed in our response to Comment 1.22, to avoid this ambiguity, we will remove the prediction uncertainty expressed in years and report it only in terms of freshwater forcing (Sv), which is directly related to the normalized target variable predicted by our framework and provides a more consistent measure of the model performance.

Comment 1.24

L188-189: ‘’ the LR provides skillful predictions ‘’ Define skillful. Is this based on a statistical metric?

Response:

We agree that the statement anticipated results that were presented only later in the manuscript. In the revised version, we will remove this claim from Section 3.1 and introduce it only in Section 3.2, where the CESM results are presented and discussed.

Comment 1.25

L219-220: '' For $r_F = 10^{-4}\, Sv\, yr^{-1}$, the collapse of the AMOC is initiated approximately 200 years before the system reaches its actual 220 tipping point, marking the onset of a regime shift" Which figure illustrates this result? Define AMOC collapse. Explain ''initiated". Does this refer to the start of a decline?

Response:

This statement is supported by Fig. B1b in Appendix B, and we will explicitly refer to this figure in the revised manuscript.

By "the collapse is initiated", we mean that the AMOC strength exhibits a clear change in its rate of decline, marking the onset of the rapid transition towards the collapsed (OFF) state. For the forcing rate $r_F = 10^{-4}\,\mathrm{Sv\,yr^{-1}}$, this transition begins approximately 200 years before the system reaches the intrinsic tipping point F_H^C identified from the quasi-static simulation with $r_F = 10^{-5}\,\mathrm{Sv\,yr^{-1}}$. This earlier onset is a consequence of rate-induced effects. We will revise the manuscript to clarify what is meant by the "initiation of the collapse" and to explicitly distinguish it from the system's intrinsic tipping point.

Comment 1.26

Appendix B : reduce the text and figures to focus on key results. Improve explanations to highlight the most important findings.

Response:

In the revised manuscript, we will reduce both the text and the number of figures of Appendix B, focusing on the key results and emphasizing the main findings.

Comment 1.27

Fig. 3, legend : no need to explain what is a boxplot, this is common knowledge. However, I suggest to keep the definition of extremes and error bars.

Response:

In the revised manuscript, we will simplify the figure caption by removing this explanation while retaining the definitions of the extremes and the error bars to facilitate the interpretation of the results.

Comment 1.28

Fig. 3: I suggest to use consistent color for models using the same inputs. For instance, CNN using SST-only in yellow boxplot and LR using SST-only with a yellow triangle…

Response:

We intentionally adopted different colors for the linear regression markers to increase visual contrast when they overlap with the corresponding CNN boxplots, making them easier to distinguish. No changes in the manuscript.

Comment 1.29

Fig. 3: What is shown here? Evaluation of the model when using the test simulation with the six other simulations used for training and validation?

Response:

This is indeed what Fig. 3 shows. The figure reports the performance of both the CNN and LR models when each of the seven CLIMBER-X simulations is used as the test simulation, while the remaining six simulations are used for training and validation following the leave-one-out strategy described in Section 2.4 of the manuscript. We will clarify this better in the revised figure caption.

Comment 1.30

L246-251: The authors explained L226-232 that CESM data was only used for evaluation, why then the hyperparameter are modified using CESM data?

Response:

We agree that this point requires further clarification. The CNN is trained exclusively using CLIMBER-X data, and no CESM data are used during the training phase. A small portion of the CESM simulation is used only as a validation set to recalibrate the model hyperparameters for the domain adaptation step. The final model performance is then evaluated on the remaining portion of the CESM simulation, which is never used during either training or hyperparameter tuning. We will better clarify this training, validation, and testing strategy in the revised manuscript.

Comment 1.31

L248 : ''the first 430 years are discarded to remove transient effects" Define ''transient effects".

Response:

By "transient effects", we refer to the initial period of the simulation immediately after the freshwater forcing is applied, during which the climate system is still adjusting to the imposed forcing and has not yet approached its quasi-equilibrium state. During this phase, the model response is dominated by the initial adjustment to the forcing and is therefore not representative of the subsequent long-term evolution. We will clarify this in the revised manuscript.

Comment 1.32

L263-268 : '' Despite these measures, some variability remains in the validation result" and '' In what follows, we present results obtained with the best-performing configuration on the CESM test set (last 880 years).'' the large variability in the results obtained suggest that there is a large uncertainty in the inference. I suggest to quantify the related uncertainty.

Response:

The variability mentioned in lines 236-268 refers to the fact that, for the SST+SSS and SSS-only configurations, several hyperparameter combinations achieve very similar validation performance, and no single configuration consistently outperforms the others. To assess the uncertainty associated with this variability, we evaluate the five best-performing hyperparameter configurations identified during validation. The results obtained with the best-performing configuration on the CESM test set are presented in the main text, while the results for the remaining four configurations are reported graphically in Appendix A. To make this uncertainty more explicit, in the revised manuscript we will also quantify the variability in the CESM performance across these five configurations, providing a clearer assessment of the sensitivity of the results to the choice among the top-performing hyperparameter configurations.

Comment 1.33

Fig 4 : Where is the training data in panel (a)? How is the tipping detected in panel (a)?

Response:

In Fig. 4, the training data are not shown because the framework is trained exclusively on data from the CLIMBER-X model and not on CESM data. The tipping point shown in panel (a) was identified by [8] using the procedure already described in our response to Comment 1.14.

Comment 1.34

L277 : here 500 trials are mentioned while 50 trials are mentioned at L255? Why 500 ? Clarify the purpose of such a large number of trials.

Response:

The two numbers refer to different stages of the analysis. During the hyperparameter search, each hyperparameter configuration is evaluated over 50 independent training trials to obtain a robust estimate of its validation performance. Once the optimal hyperparameter configuration has been selected, the final evaluation on the CESM test set is performed over 500 independent training trials, each initialized with different random CNN weights, following common practice to characterize the variability associated with the stochastic training procedure. We identified 500 trials as the minimum number required for the distribution of the prediction results to converge, providing a reliable estimate of the model performance. We will clarify this in the revised manuscript.

Comment 1.35

L344-347 : here and more generally in this subsection, can the authors refer to the figure / panel for each statement.

Response:

We agree that explicitly referring to the relevant figures, panels, and symbols throughout this subsection will improve the readability of the manuscript and make it easier for the reader to follow the discussion and verify each statement. We will revise the text accordingly in the revised manuscript.

Comment 1.36

L374 : ''SV relevance maps" Explain the acronym SV.

Response:

The acronym "SV" is incorrect. We intended to refer to "SHAP" (SHapley Additive exPlanations) relevance maps and will correct this error in the revised manuscript.

Comment 1.37

L390 : '' the magnitude is quantified using Sen's slope estimator." Justify the choice of Sen's slope estimator over a linear trend.

Response:

We chose Sen's slope estimator because it provides a robust, non-parametric estimate of the magnitude of monotonic trends and is less sensitive to outliers and short-term variability, which are common in climate model output. Since the trend is computed independently at every grid point of the SST, SSS, and S_z^{35S} input fields, using a robust estimator ensures that the estimated trend magnitude reflects the underlying increasing or decreasing tendency while reducing the influence of local fluctuations or extreme values. We will clarify this motivation in the revised manuscript.

Comment 1.38

L409-410: ‘’ First, we note that patterns are consistent with those reported by Stouffer et al. (2006), who conducted an inter-comparison of several EMICs, including an earlier version of the CLIMBER-X model used here. ‘’ Update the discussion to include more recent papers for context.

Response:

To the best of our knowledge, there are relatively few studies providing a systematic intercomparison of the Atlantic SST, Atlantic SSS, and full-depth Atlantic salinity section at 35°S responses to AMOC weakening across an ensemble of climate models. This is why we chose to cite [5], which represents one of the few exceptions by providing such an intercomparison across multiple Earth System Models of Intermediate Complexity (EMICs), including an earlier version of CLIMBER-X. A more recent study [6] investigates the Atlantic SST and SSS responses to AMOC weakening using a fully coupled Earth System Model (ESM) and an eddy-permitting ocean–sea-ice model, distinct from CESM and CLIMBER-X, and reports similar response patterns. We will include this more recent reference in the revised manuscript to provide additional context.

Comment 1.39

Section 4.2 and 4.3 : refer ro figure panels when describing results to improve readability. Maybe focus on only one freshwater forcing, and show the sensitivity or the SST, SSS and S35S for CLIBBER-X only.

Response:

In the revised manuscript, we will explicitly refer to the relevant figure panels throughout the discussion to make it easier for the reader to follow the results.

Regarding the number of figures in Section 4.2, we agree that they can be reduced. We will show the explainability results only for one freshwater-forcing rate and discuss in the text that the remaining forcing rates exhibit qualitatively similar sensitivity patterns. However, we will retain the CESM explainability analysis presented in Section 4.3, as it is highly relevant to assessing the physical consistency of our framework when generalizing to a state-of-the-art climate model.

Comment 1.40

The relevance score in Figs. 6, 7, 8, 9 and 10 are quite noisy and suggest that the CNN may not be learning physically consistent features. Discuss the implication for the skill of the CNN. Can it be linked to the large variability obtained?

Response:

We would first like to clarify that Figs. 8 and 9 show the explainability maps obtained for the LR model, not the CNN. Unlike the CNN, the LR model produces noisy, spatially fragmented relevance maps that do not reveal coherent large-scale physical patterns. We interpret this behavior as evidence that the LR model is overfitting the input data.

In contrast, the CNN explainability maps shown in Figs. 6, 7, and 10 consistently identify coherent and physically meaningful large-scale patterns across all the input variables considered and for both the CLIMBER-X and CESM data. These results indicate that the CNN learns physically consistent features rather than isolated local patterns.

We are not showing variability in the relevance maps. Following common practice in machine learning, the explainability analysis is performed only on the best-performing CNN realization on the test data among all the independent training trials. Consequently, the relevance maps are not affected by poorly performing training realizations, and the reported patterns reflect the behavior of the best-trained model.

Comment 1.41

L600 : one reference is missing.

Response:

Thanks for spotting this. The missing reference will be added to the revised manuscript.

Authors' Response to Reviewer 2

General Comments. This manuscript presents a CNN-based approach for predicting the distance of the AMOC from its tipping point. The model uses spatial fields as input, including sea surface temperature (SST), sea surface salinity (SSS), their combination, and the full-depth salinity profile, to predict a normalized distance-to-tipping metric ranging from 0 to 1. The authors futher employ SHAP analysis to identify the features that contribute most strongly the model predictions. In addition, the effort to transfer information learned from CLIMBER-X simulations to CESM simulations is potentially valuable. The topic is important and interesting. However, several important issues need to be addressed before the manuscript can be considered for publication.

Response:

We thank the reviewer for carefully reading our manuscript and for the constructive and thoughtful comments. We appreciate the time and effort devoted to evaluating our work. Below we provide a detailed point-by-point response to each of the reviewer's comments and describe the corresponding revisions that will be incorporated into the revised manuscript.

Comment 2.1

The CNN output is defined as a normalized index, which is designed to avoid explicitly providing information about the freshwater forcing rate during training and testing. However, in CLIMBER-X, different types of tipping may occur, including bifurcation-induced, noise-induced, and rate-induced tipping. The current definition of the distance to the tipping point appears to be mainly applicable to bifurcation-induced tipping under deterministic conditions. Therefore, the authors should provide a clearer explanation of the applicability and limitations of this distance to tipping definition. In particular, it would be helpful to clarify whether this definition is intended to characterize only the distance to a bifurcation threshold, or whether it can also meaningfully describe proximity to noise-induced or rate-induced tipping events. This distinction is important because the timing of noise-induced and rate-induced tipping events can be highly stochastic and may not have a simple linear relationship with the freshwater forcing value.

Response:

We agree that our framework and the definition of the normalized distance index d_f are specifically designed to estimate the distance to a bifurcation-induced tipping point. In our approach, d_f is defined as the normalized distance, measured in terms of freshwater forcing, to the intrinsic bifurcation threshold of the system. This definition provides a rate-independent target variable for the CNN, allowing it to estimate the proximity to the bifurcation without explicitly relying on the freshwater forcing rate during training or testing. Although the forcing rate can influence the timing of the AMOC collapse, the target quantity remains the distance to the underlying bifurcation point.

We agree that this definition is not intended to characterize the proximity to noise-induced or rate-induced tipping events. In these cases, the transition may occur before or after the bifurcation threshold, and therefore cannot be uniquely described by the normalized distance d_f. Accordingly, our framework is not designed to predict the onset of noise-induced or rate-induced tipping. We will revise the manuscript to explicitly state that the proposed distance metric and the CNN framework are only applicable to estimating the distance to bifurcation-induced tipping points, and to clearly discuss this limitation.

Comment 2.2

In the CLIMBER-X experiments, the LR model performs comparably to, and in some cases even better than, the CNN. This suggests that the relationship between the input variable fields and the target index may be relatively simple, rather than requiring complex nonlinear spatial features learned by the deep neural networks. To better justify the use of CNNs, the authors should include additional baseline models, such as shallow CNNs or other lightweight machine-learning methods. In addition, the training dataset appears to be relatively small for a deep learning approach. The authors should provide a clearer description of the dataset, including the size and construction of training samples. This is important because temporally adjacent fields are often highly correlated, meaning that the nominal sample size may overestimate the amount of independent information available for training and evaluation. The manuscript also reports a relatively large number of hyperparameter settings across different experiments, which may partly reflect the limited sample size and raises questions about the consistency and robustness of the model configuration.

Response:

Regarding the first point concerning the inclusion of additional baseline models, we have already addressed this issue in the overall response, where we explain the rationale for not including additional baseline methods.

Regarding the size and construction of the training dataset, we agree that a clearer description is warranted and will revise the manuscript accordingly. The CLIMBER-X experiments are based on seven independent AMOC collapse simulations generated using freshwater forcing rates ranging from 10^{-5} to $6 \times 10^{-4}\,\mathrm{Sv\,yr^{-1}}$. Since each simulation reaches the critical freshwater forcing F_H^C at a different time, their temporal lengths differ. The model performance is evaluated using a leave-one-out strategy, in which one simulation is reserved for testing and the remaining six are used for training. Consequently, the number of training samples varies between 532 (when the slowest-forcing simulation, $10^{-5}\,\mathrm{Sv\,yr^{-1}}$, is excluded from training) and 2684 (when the fastest-forcing simulation, $6 \times 10^{-4}\,\mathrm{Sv\,yr^{-1}}$, is excluded). Each training sample consists of a two-dimensional field corresponding to one of the considered input configurations: SST and SSS fields over the Atlantic Ocean on a $5° \times 5°$ spatial grid, the combination of both fields (SST+SSS), or the full-depth salinity section at latitude 35°S, defined over the entire Atlantic basin in the longitudinal direction and over the full ocean depth using 23 vertical levels. The fields are sampled every 10 years. For the CESM generalization experiments, the model is trained on all seven CLIMBER-X simulations, corresponding to a total of 2720 training samples for the same input-variable configurations. We will revise the manuscript to include this information and provide a clearer description of the construction of the training dataset.

Finally, regarding the reported hyperparameter configurations, we would like to clarify that they result from independent optimization procedures performed for different experimental settings. In the CLIMBER-X experiments, the selected optimization hyperparameters are highly consistent across both the leave-one-out folds and the different input-variable configurations. In particular, the learning rate is almost always equal to 10^{-4}, the batch size is predominantly 64, while the early-stopping patience shows only moderate variation. Overall, these results indicate that the optimization procedure is robust when both training and evaluation are performed within the same climate model.

The greater variability observed in the CESM experiments should not be interpreted as evidence of an unstable optimization procedure, but rather reflects two expected sources of variability. First, each input-variable configuration (SST, SSS, SST+SSS, and S^z_{35S}) defines a different transfer-learning task. Since these variables represent different physical fields, their transferability from CLIMBER-X to the more complex CESM model is expected to differ, and therefore different optimization hyperparameters may be preferred across input configurations. Second, for a given input-variable configuration, the optimal hyperparameter combination is not necessarily unique. As illustrated by the additional SSS and SST+SSS experiments reported in Table A4 and discussed in Section 3.2, different hyperparameter configurations can lead to comparable predictive performance on the independent CESM test dataset. Therefore, the observed variability reflects the existence of multiple competitive hyperparameter configurations for the same transfer-learning task, rather than an unstable optimization procedure. Importantly, these differences concern only the optimization hyperparameters, whereas the CNN architecture and training methodology remain unchanged throughout all experiments. We will revise the manuscript to clarify these aspects and provide a more detailed discussion of the reported hyperparameter configurations.

Comment 2.3

The manuscript emphasizes that the CNN is trained on CLIMBER-X and then generalized to CESM. However, since part of the CESM simulation is used for validation and hyperparameter selection, the model selection procedure has already incorporated information from the target model. Therefore, this experiment should not be presented as a fully independent cross-model extrapolation. The authors should moderate the relevant claims in the abstract and conclusions. In particular, statements such as “reliable estimates” and “generalize across models” should be softened unless additional validation using fully independent target model simulations is provided.

Response:

We agree with the reviewer that these statements should be moderated. Accordingly, we will revise the manuscript by softening the relevant claims in the Abstract, Conclusions, and throughout the text to better reflect the scope of the CESM experiments and avoid implying fully independent cross-model extrapolation.

Comment 2.4

The explainability analysis is potentially valuable: however, the SHAP maps appear to be derived from the best performing CNN realizations selected using test set performance. This may introduce selection bias, particularly for CESM, where the model shows substantial variability across realizations.

Response:

The SHAP analysis is indeed performed on the best-performing realization. This is common practice in explainability studies, as the objective is to interpret the model that best captures the relationship between the input variables and the target. Performing the explainability analysis on suboptimal realizations could instead lead to attribution patterns that primarily reflect prediction errors or incomplete learning, making it difficult to distinguish physically meaningful patterns from artifacts associated with reduced predictive performance. We will clarify this rationale in the revised manuscript.

Comment 2.5

Some typos should be corrected throughout the manuscript. For example, in line 64, "Each of these interacting modules are discretized" should be revised to "Each of these interacting modules is discretized". In line 413, "a modest SST increase is occurs" should be revised to "a modest SST increase occurs". There also appears to be an incorrectly formatted reference citation around line 600.

Response:

We will correct the typographical errors, grammar, and formatting inconsistencies throughout the manuscript, including the examples pointed out by the reviewer and the incorrectly formatted reference citation.

Authors' Response to Reviewer 3

General Comments. The manuscript investigates the use of a convolutional neural network CNN to predict the distance of a AMOC state to a tipping point using training data from a Earth System Model of Intermediate Complexity, CLIMBER-X, forced by imposed freshwater flux. The results are tested for compatibility with a Community Earth System Model (CESM), and the most relevant spatio temporal features are investigated using the SHAP methodology.

The problem of identifying the AMOC tipping point and provide early warning is of paramount importance. The paper is well written and competently done, but its emphasis is mostly technical while the physical set up and processes, as well as the relevance of the results, are not very clearly stated. Revisions are necessary prior to publication, as detailed in the following.

Response:

We thank the reviewer for carefully reading our manuscript and for the constructive and thoughtful comments. We appreciate the time and effort devoted to evaluating our work. Below, we provide a detailed point-by-point response to each of the reviewer's comments and describe the corresponding revisions that will be incorporated into the revised manuscript.

Comment 3.1

The choice of the specific models (CLIMBER-X and CESM) and of their settings has to be motivated and explained in more details. Why those models? And why the specific forcing scenario consisting of a slowly increasing surface freshwater input? As discussed also by the authors the forcing is not realistic. Its choice also determines a non realistic index d_F of the distance to the tipping point, that is defined in terms of the values of freshwater forcing reached by the model at the tipping time.

These choices and limitations needs careful explanation and motivation, up-front in the Introduction and in the Models and Methods sections.

Response:

Regarding the choice of climate models, we have already addressed this point in our response to Reviewer 1 (Comment 1.1), where we explain why CLIMBER-X and CESM are particularly well-suited for investigating the long-term stability of the AMOC under freshwater forcing. Moreover, the CESM experiment by van Westen et al. (2024) is currently the only published quasi-equilibrium AMOC collapse simulation performed with a state-of-the-art fully coupled GCM. We will revise the manuscript to better motivate the choice of these models in the Introduction and the Models sections.

Regarding the choice of the slowly increasing freshwater forcing, this experimental protocol is widely adopted in studies investigating the stability and bifurcation structure of the AMOC, as it enables the system to evolve quasi-statically toward its critical transition and characterize the associated hysteresis behavior (e.g., [3], [8], [13]). Although this forcing scenario is not intended to represent a realistic future evolution of the climate system, it provides a controlled environment in which the bifurcation point can be approximated and the normalized distance-to-tipping index d_f can be consistently defined. The objective of our work is not to develop a framework that can be directly applied to estimate the distance to tipping in the real AMOC. Rather, our aim is to demonstrate the viability of the proposed framework in a controlled modeling environment. We recognize that extending this approach to real observations will require substantial methodological developments, including the definition of an appropriate normalized distance index analogous to d_f that accounts for the complexities of the real climate system. We will revise the manuscript to make the scope of the present study and these limitations more explicit.

Comment 3.2

I am not completely convinced by the author's claim that the results obtained with CLIMBER-X are effectively generalized to the CESM outputs. The reason is that the CNN results are indeed trained using CLMBER-X, but as the authors state, the results are validated on a third of the CESM simulation and then tested on the rest. This implies, if I understand correctly, a different hyperparameter selection, a limited number of training epochs, and a normalization of the index and the use of values from CESM.

These points and their consequences on the result interpretation needs further explanation and if needed further testing and generalizations.

Response:

We agree that the CESM experiments do not constitute a fully independent cross-model generalization, as a portion of the CESM simulation is used for hyperparameter selection and validation. We will revise the manuscript to more explicitly discuss the resulting limitations of this approach and soften the claims regarding the generalization capabilities of the proposed framework to better reflect the scope of the cross-model experiments.

We would, however, like to clarify one point regarding the normalization of the index. The normalized distance-to-tipping index d_f for the CESM simulation is defined based on the tipping point identified by van Westen et al. (2024). This CESM-specific information is used exclusively to define the reference normalized index for the validation and testing stages and is not used during training to optimize the CNN weights. We will revise the manuscript to clarify this aspect more explicitly.

Comment 3.3

The choice of the Linear Regression (LR) method as benchmark is also dubious to me, as it can be expected to overfit the CLIMBER-X data. The authors should look into alternative and more convincing statistical tools.

Response:

Regarding the choice of the Linear Regression benchmark and the inclusion of additional baseline models, we have already addressed this point in the overall overview section of our response, where we explain the rationale for not including further baseline methods.

Comment 3.4

The relevance of the results needs a more complete and convincing discussion in the Conclusion section. I understand that in this new and complex field, even small steps can be complex and relevant, but given all the points above, is not clear how the results can be considered new and useful.

Response:

We thank the reviewer for this comment. We agree that the significance of our results should be discussed more clearly, and we will revise the Conclusions accordingly. In particular, we will better emphasize that the main contribution of this work is the demonstration that a deep-learning framework can provide a continuous estimate of the distance to an AMOC tipping point, rather than simply indicating that a tipping event is approaching. We further show that this estimate can be obtained from routinely observed surface ocean variables, such as SST and SSS. Although the present framework is validated in a controlled modeling environment, it lays the foundation for future extensions to real observational datasets and to other climate tipping phenomena, with appropriate methodological adaptations.

We will also better highlight the significance of the cross-model experiments. Although their limitations will be discussed more explicitly, the results demonstrate that our framework, trained on simulations from an intermediate-complexity climate model, can be successfully transferred to a substantially more complex GCM. This suggests that such an approach could serve as a valuable supporting tool for future investigations of AMOC tipping using comprehensive climate models, where generating training data is computationally expensive and time-consuming.

References

[1] M. Willeit, A. Ganopolski, A. Robinson, and N. R. Edwards, "The earth system model climber-x v1. 0–part 1: Climate model description and validation," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5905–5948, 2022.

[2] M. Willeit and A. Ganopolski, "Generalized stability landscape of the atlantic meridional overturning circulation," *Earth System Dynamics*, vol. 15, no. 6, pp. 1417–1434, 2024.

[3] S. Rahmstorf et al., "Thermohaline circulation hysteresis: A model intercomparison," *Geophysical Research Letters*, vol. 32, no. 23, 2005.

[4] R. M. van Westen and H. A. Dijkstra, "Asymmetry of amoc hysteresis in a state-of-the-art global climate model," *Geophysical Research Letters*, vol. 50, no. 22, e2023GL106088, 2023.

[5] R. J. Stouffer et al., "Investigating the causes of the response of the thermohaline circulation to past and future climate changes," *Journal of climate*, vol. 19, no. 8, pp. 1365–1387, 2006.

[6] G. M. Pontes and L. Menviel, “Weakening of the atlantic meridional overturning circulation driven by subarctic freshening since the mid-twentieth century,” *Nature Geoscience*, vol. 17, no. 12, pp. 1291–1298, 2024.

[7] E. Vanderborght, R. M. van Westen, and H. A. Dijkstra, “Feedback processes causing an amoc collapse in the community earth system model,” *Journal of Climate*, vol. 38, no. 19, pp. 5083–5102, 2025.

[8] R. M. Van Westen, M. Kliphuis, and H. A. Dijkstra, “Physics-based early warning signal shows that amoc is on tipping course,” *Science advances*, vol. 10, no. 6, eadk1189, 2024.

[9] S. Panahi et al., “Machine learning prediction of tipping in complex dynamical systems,” *Physical Review Research*, vol. 6, no. 4, p. 043 194, 2024.

[10] W. Weijer, T. W. Haine, A. H. Siddiqui, W. Cheng, M. Veneziani, and P. Kurtakoti, “Interactions between the arctic mediterranean and the atlantic meridional overturning circulation,” *Oceanography*, vol. 35, no. 3/4, pp. 118–127, 2022.

[11] S. Bacon, A. C. N. Garabato, Y. Aksenov, N. J. Brown, and T. Tsubouchi, “Arctic ocean boundary exchanges,” *Oceanography*, vol. 35, no. 3/4, pp. 94–102, 2022.

[12] C. Arumí-Planas, S. Dong, R. Perez, M. J. Harrison, R. Farneti, and A. Hernández-Guerra, “A multi-data set analysis of the freshwater transport by the atlantic meridional overturning circulation at nominally 34.5 s,” *Journal of Geophysical Research: Oceans*, vol. 129, no. 6, e2023JC020558, 2024.

[13] S. Rahmstorf, “Bifurcations of the atlantic thermohaline circulation in response to changes in the hydrological cycle,” *Nature*, vol. 378, no. 6553, pp. 145–149, 1995.