https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

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

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

[1]Institute for Marine and Atmospheric Research Utrecht, Department of Physics, Utrecht University, Utrecht, the Netherlands
[2]Center for Complex Systems Studies, Utrecht University, Utrecht, the Netherlands

Correspondence: Francesco Guardamagna (f.guardamagna@uu.nl)

Abstract. The Atlantic Meridional Overturning Circulation (AMOC) is an important tipping element of the climate system, with the potential to undergo an abrupt transition from its present strong state to a weak state. Such a collapse would have severe global consequences, including regional cooling, sea-level rise, altered precipitation patterns, and cascading impacts on other climate tipping elements. Both statistical and physics-based early warning signals (EWS) of an approaching AMOC tipping event have been proposed. Here, we introduce a convolutional neural network (CNN)–based framework designed to predict the distance of an AMOC state to its tipping point under imposed freshwater flux forcing. We first evaluate the CNN model using simulations from the Earth System Model of Intermediate Complexity CLIMBER-X. We then test its generalization capabilities by applying the CNN model, trained on CLIMBER-X data, to the AMOC tipping trajectory obtained recently in the Community Earth System Model (CESM). Explainable AI methods are used to identify the spatiotemporal features most relevant to the predictions. Our results demonstrate the potential of deep learning to provide reliable estimates of the distance to the AMOC tipping point and generalize across models of varying complexity.

1 Introduction

The Atlantic Meridional Overturning Circulation (AMOC) is a key component of the global climate system, transporting heat and salt from the subtropical to the northern Atlantic (Johns et al. (2011)). The present-day AMOC may be in a multi-stable regime where at least two stable equilibrium states exist: a strong circulation state and a weak or collapsed state. The mechanism leading to abrupt transitions between these two states is the salt-advection feedback (Weijer et al. (2019)). When the subpolar North Atlantic becomes fresher, such as from increased meltwater discharge from the Greenland Ice Sheet, the surface water density decreases. This reduces deep-water formation, weakening the AMOC and diminishing the northward transport of salinity, which further freshens the surface waters.

The multi-stability of the AMOC has been demonstrated in models of varying complexity. This includes conceptual models (Stommel (1961); Sinet et al. (2024)), Earth System Models of Intermediate Complexity (EMICs) (Hofmann and Rahmstorf (2009)), and fully coupled General Circulation Models (GCMs) (Hawkins et al. (2011); van Westen et al. (2024b)). The AMOC is hence classified as a potential tipping element of the climate system (Lenton et al. (2008)). The consequences of such a transition would be dramatic and global in scale. These include a cooling of several degrees in Western Europe (Liu et al. (2017); van Westen et al. (2024b); Rahmstorf (2024)), a significant sea level rise along the Atlantic coast of North America (Rahmstorf (2024)), and a southward shift of the Intertropical Convergence Zone (ITCZ), potentially increasing the probability

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

of severe droughts and extensive flooding (Cerato et al. (2025)). A collapse of the AMOC could also trigger a cascading effect, potentially destabilizing other tipping elements of the Earth system, such as the Antarctic ice sheet, tropical monsoon systems, and the Amazon rainforest (Wunderling et al. (2024)).

Given these potentially strong and dramatic consequences, efforts have been made to identify early warning signals (EWS) that could anticipate an approaching AMOC tipping event. Previous studies have used both physics-based approaches (van Westen et al. (2024b)) and statistical indicators (Ditlevsen and Ditlevsen (2023); Boers (2021)), the latter based on the tendency for the variance and lag-1 autocorrelation of a state variable to increase when the AMOC approaches a tipping point. However, considerable uncertainty remains in estimating the timing of the tipping onset (Ben-Yami et al. (2024)). A recent machine learning (ML)–based model employing a Reservoir Computing architecture (Panahi et al. (2024)) has demonstrated promising results when applied to an AMOC tipping simulated in the Community Earth System Model (CESM) shown in van Westen et al. (2024b). Despite its success, the approach has notable limitations. First, it requires prior knowledge about the freshwater forcing rate, which is a significant constraint for real-world applications. Second, to achieve high predictive performance, the model must be trained on data from the simulation itself, extending up to 100 years before the onset of collapse. Third, as the model relies solely on one-dimensional indices, such as the AMOC strength at 26°N as input, it precludes the application of explainable AI techniques, thereby offering little insight into whether the ML model captures dominant physical processes causing the tipping event.

Here we propose a new AMOC tipping prediction scheme based on a Convolutional Neural Network (CNN) deep learning architecture. The goal is to predict the distance to the AMOC tipping point under surface freshwater flux forcing, without prior knowledge of the forcing rate. Section 2 provides an overview of the CLIMBER-X and CESM datasets used in this study, as well as a description of the CNN architecture.

In Section 3, we first evaluate the performance of the CNN using different AMOC tipping simulations from CLIMBER-X. The network is provided with sea surface temperature (SST) and sea surface salinity (SSS) fields over the Atlantic Ocean as input, as well as the full-depth salinity section at 35°S (S_z^{35S}). We also compare the performance of the CNN with that of a linear regressor (LR) model trained on the same input data.

Second, we assess the generalization capability of the deep-learning approach by applying the CNN framework, trained on CLIMBER-X data, to the CESM AMOC collapse simulation from van Westen et al. (2024b). We again compare the performance of the CNN with that of the LR model, showing that while the CNN successfully generalizes to the CESM simulation, the LR model fails to do so.

In Section 4, we apply an explainable AI technique to identify the regions and variables that contribute most to the CNN and LR predictions, allowing us to assess the physical consistency of the proposed methodology. Finally, Section 5 provides a summary and discussion.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

2 Models and Methods

In this section, we describe the models used to generate the simulation data analyzed in this study, as well as the CNN architecture, its hyperparameters, and the validation and testing procedures.

2.1 CLIMBER-X

CLIMBER-X (Willeit et al. (2022a)) is an EMIC, with a climate core consisting of multiple coupled components: the Semi-Empirical Dynamical-Statistical Atmosphere Model (SESAM), the frictional–geostrophic 3D ocean model GOLDSTEIN, the Simple Sea Ice Model (SISIM), and the land surface–vegetation model PALADYN. Each of these interacting modules are discretized on a $5^\circ \times 5^\circ$ horizontal resolution. The GOLDSTEIN ocean model includes 23 unequally spaced vertical levels and simulates the dynamics of various physical variables, including salinity and temperature. A comprehensive description of all model components can be found in Willeit et al. (2022a). CLIMBER-X provides an efficient framework to study AMOC stability. In particular, it reproduces the typical AMOC bi-stability seen across the climate model hierarchy and matches many aspects of state-of-the-art CMIP6 models across diverse forcings and boundary conditions (Willeit et al. (2022a)).

In particular, CLIMBER-X exhibits hysteresis behavior in the AMOC when the freshwater balance in the North Atlantic is perturbed. In our experiments, we identify the tipping point for this kind of forcing scenario by applying a slowly increasing surface freshwater forcing F_H in the Atlantic between 20°N and 50°N (compensated over the rest of the surface ocean). For the slow constant forcing rate $r_F = 10^{-5}$ Sv yr^{-1} and starting from preindustrial conditions with greenhouse gas concentrations fixed at preindustrial levels, the AMOC collapses once the freshwater forcing reaches $F_H^C = 0.22$ Sv. This threshold will be used as an estimate for the AMOC tipping point in CLIMBER-X for this type of quasi-equilibrium freshwater forcing experiment.

2.2 CESM

The Community Earth System Model (CESM) is a fully coupled GCM. Its ocean component is the Parallel Ocean Program version 2 (POP2) (Smith et al. (2010)), the atmosphere component is the Community Atmosphere Model version 4 (CAM4) (Neale et al. (2013)), and the sea-ice component is the Community Ice Code version 4 (CICE4) (Hunke et al. (2015)). For a complete description of the CESM model, refer to Hurrell et al. (2013).

The CESM simulation used in this study is the hosing experiment conducted by van Westen et al. (2024b) using CESM version 1.0.5, with a horizontal resolution of 1° for the ocean and sea-ice components and 2° for the atmosphere and land components. The simulation is initialized from a preindustrial control simulation, on which a slowly increasing surface freshwater flux is applied in the North Atlantic between 20°N and 50°N (compensated over the rest of the surface ocean) at a rate $r_F = 3 \times 10^{-4}$ Sv yr^{-1}. 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.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

2.3 CNN

Convolutional Neural Networks (CNN) are a class of deep learning models originally designed for image analysis tasks (Lecun et al. (1998)). CNNs are particularly well-suited for analyzing two-dimensional data that exhibit spatial patterns and correlations across input fields, such as climate variables over geographic grids. CNN's capture the spatial hierarchy of features by progressively learning patterns from low-level details to high-level representations (Goodfellow (2016)). A CNN typically consists of three key building blocks: the Convolutional Layer, which extracts local features using learnable filters; the Pooling Layer, which reduces spatial dimensions while retaining relevant information; and the Fully Connected Layer, which integrates the learned features for final predictions.

In this study, we employ a CNN architecture (Fig. 1) comprising three convolutional layers with a progressively increasing number of 3×3 filters: 32 filters in the first layer, 64 in the second, and 128 in the third. Each convolutional layer is immediately followed by batch normalization to enhance training stability and accelerate convergence, and by a 2×2 max-pooling layer. Throughout the network, LeakyReLU activation functions are employed to model nonlinear relationships, allowing the network to approximate complex and rich nonlinear functions. These initial convolutional blocks act as a feature extractor, transforming the input data into a set of high-level representations. The resulting feature maps are then flattened and passed through a fully connected layer with 256 nodes. Finally, a single output neuron produces the CNN's prediction.

During the training phase of the CNN, weights are optimized via back propagation. At each iteration, a mini-batch of data is processed, and the weights are updated to minimize the loss between the CNN's predictions and the target values. The learning rate controls the magnitude of these weight updates. In our case, the loss function is the Mean Squared Error (MSE), and the optimizer is the Adam method. To improve training efficiency and prevent overfitting, we employ both an early stopping policy and a reduce-on-plateau policy. According to the early stopping policy, training ends if the loss fails to decrease for a specified number of epochs (n_{es}), where one epoch is completed when the entire training dataset is processed once. Meanwhile, the reduce-on-plateau policy lowers the learning rate by a factor of 10 if the loss does not improve for a specified number of epochs (n_{rop}), always set to $n_{es}/2$. Consequently, the main CNN hyperparameters are the mini-batch size (bs), the patience of the early stopping policy (n_{es}), and the initial learning rate (lr).

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°S) from freshwater-forced AMOC tipping trajectories obtained in CLIMBER-X. We also use the combination of these two variables as input. Additionally, the full-depth salinity profile at 35°S (S_z^{35S}) is considered as an input configuration. For all input variables, we extracted annual-mean fields at 10-year intervals and retained only ocean grid points, setting all non-ocean cells to zero.

The CNN output is a normalized index d_F specifically designed to eliminate the need to explicitly provide information about the freshwater forcing rate during training and testing. This index represents the distance to the AMOC tipping point and is defined as:

$$d_F(t) = \frac{F_H(t_p) - F_H(t)}{F_H(t_p)} \tag{1}$$

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

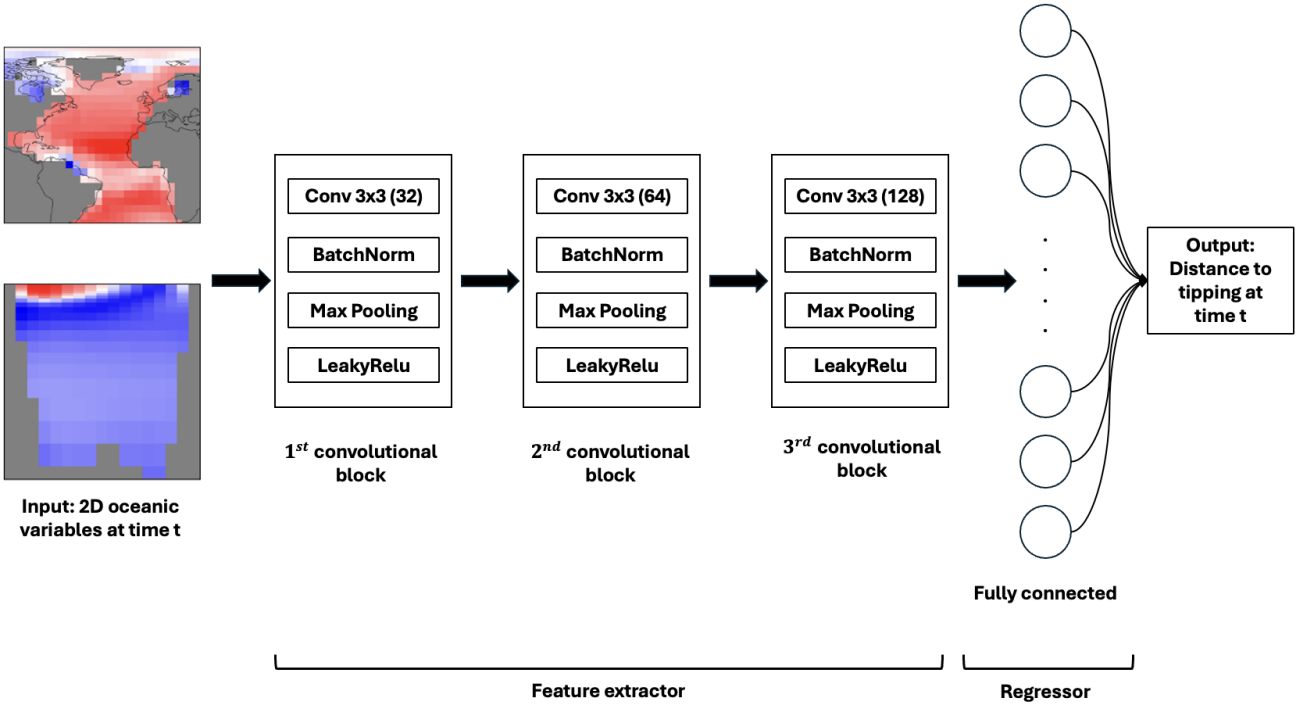


Figure 1. *Schematic representation of the CNN architecture. For each convolutional layer, the filter dimension is reported, with the number of filters indicated in parentheses. The fully connected layer consists of 256 nodes.*

where $F_H(t)$ denotes the freshwater flux value F_H at time step t, and $F_H(t_p)$ represents its value at tipping. Specifically, in CLIMBER-X, $F_H(t_p) = F_H^C = 0.22$ Sv (see section 2.1) under pre-industrial conditions. The index d_F ranges between 0 and 1, corresponding to the AMOC in pre-industrial conditions and at its tipping point, respectively. This definition follows from restricting the dataset to $F_H < F_H^C$, such that only pre-collapse states are used for training and testing. As the aim is to estimate the distance to the tipping point under increasing freshwater forcing (F_H), post-collapse data are excluded.

2.4 Validation and testing

The performance of our CNN approach is initially evaluated using seven distinct AMOC tipping simulations conducted with the CLIMBER-X model, each generated by applying a uniform freshwater forcing perturbation between 20°N and 50°N in the North Atlantic, with a different forcing rate r_F. The different forcing rates considered are r_F =[10^{-5}, 10^{-4}, 2×10^{-4}, 3×10^{-4}, 4×10^{-4}, 5×10^{-4}, 6×10^{-4}] Sv yr^{-1}. For all forcing rates, $d_F(t)$ is defined with respect to a freshwater flux at tipping $F_H(t_p) = F_H^C = 0.22$ Sv, which corresponds to the tipping point identified for the slowest forcing experiment ($r_F = 10^{-5}$).

Following a leave-one-out strategy, one simulation is set aside for testing while the remaining six are used to determine the optimal set of hyperparameters through a K-fold cross-validation approach. The complete list of hyperparameter values evaluated during validation is reported in table A1. Once the optimal hyperparameters are identified, the CNN is retrained on

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

the six simulations used for validation and then evaluated on the trajectory excluded from both training and validation. This validation and evaluation procedure, carried out separately for each input variable configuration explored in our study, ensures that each of the seven simulations serves as a test set once, providing a robust assessment of model performance.

By testing on data unseen during training and validation, and with distinct characteristics due to different forcing rates, we rigorously evaluate the CNN's ability to generalize across different forcing conditions. A complete summary of the optimal hyperparameter configurations, determined independently for each CLIMBER-X test simulation and for each input variable configuration, is provided in A2.

3 Results: Performance and Generalization

In this section, we present the outcomes of two complementary analyses. Section 3.1 reports the results of training and testing our CNN model on CLIMBER-X data. Section 3.2 examines the generalization capability of our framework by applying a CNN trained exclusively on CLIMBER-X simulations to data from the CESM. It also outlines the modifications to the validation and training procedure required for this generalization. For both experiments, we also report the performance of a linear regression (LR) model trained and evaluated under the same conditions, providing a baseline for comparison.

3.1 CLIMBER-X data

The AMOC strength at 26°N from two CLIMBER-X simulations with the slowest and fastest forcing rates, $r_F = [10^{-5}, 6 \times 10^{-4}]$ Sv yr^{-1}, is shown in Fig. 2a and b, respectively.

As described in Section 2.4, we evaluate the performance of the CNN using seven independent AMOC collapse simulations produced with CLIMBER-X. Four input configurations are considered: Atlantic sea surface temperature (SST) fields, Atlantic sea surface salinity (SSS) fields, their combination (SST+SSS), and the full-depth salinity profile at 35°S (S_z^{35S}). For each configuration, the CNN is evaluated on all seven CLIMBER-X simulations following the procedure outlined in Section 2.4. To account for stochasticity arising from random weight initialization, the entire evaluation procedure is repeated 20 times for each simulation.

The LR model is trained using the same input variables and target (d_F), again following the procedure outlined in Section 2.4. Unlike the CNN, the LR model does not require random initialization or hyperparameter tuning via cross-validation. In the following, we first present and discuss the CNN results, and subsequently analyze the performance of the LR model.

Figure 2c–f presents the CNN predictions of the distance to tipping for the slowest and fastest CLIMBER-X simulations, with $r_F = [10^{-5}, 6 \times 10^{-4}]$ Sv yr^{-1}, across all four input configurations. The reported predictions are the median across 20 independent training trials; variability across trials is negligible and therefore not shown. To assess the overall performance of the CNN, Fig. 3 shows the distribution of mean squared error values obtained by testing the CNN on each of the seven CLIMBER-X AMOC collapse simulations, each characterized by a different forcing rate r_F.

The results indicate that the CNN performs well across all CLIMBER-X simulations, regardless of the forcing rate r_F. However, a consistent improvement in performance is found as r_F increases and when SSS fields are included during training,

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

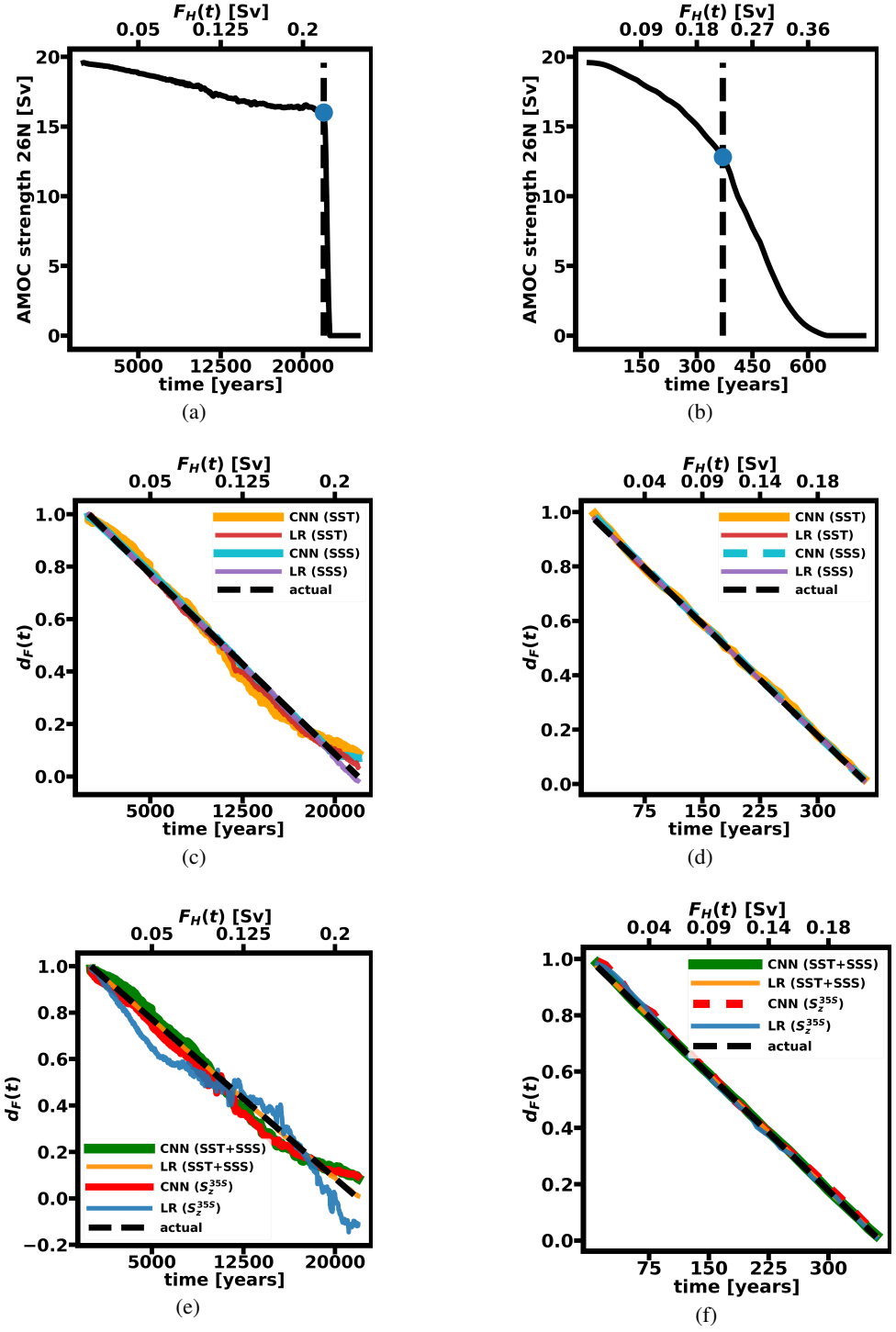


Figure 2. *(a)-(b) Evolution of AMOC strength at 26°N in the CLIMBER-X model, under forcing rates* r_F=[10^{-5}, 6×10^{-4}] *Sv* yr^{-1}, *respectively. (c)-(e) CNN predictions vs actual distances to tipping for* r_F=10^{-5} *Sv* yr^{-1}. *The CNN is trained using SST or SSS fields over the Atlantic individually in (c), while in (e), the combination of SST and SSS fields over the Atlantic and* S_z^{35S} *is adopted. (d)-(f) Same as (c)-(e) but for* r_F=6×10^{-4} *Sv* yr^{-1}. *In all cases, the median predictions out of 20 trained CNN realizations are shown. For each input-variable configuration and for both test simulations, the corresponding predictions from the LR model are also included for comparison.*

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

with the best results obtained adopting the SSS-only configuration and the worst with the S_z^{35S} configuration. This behavior is consistent with the fact that SSS fields are directly influenced by the applied freshwater forcing in the Atlantic Ocean (see section 2.1).

The worst results occur for the slowest forcing rate ($r_F = 10^{-5}$ Sv yr^{-1}), when the CNN is trained on S_z^{35S} fields, with the SST-only and SST+SSS configurations performing slightly better but still comparable (Fig. 2c,e). Specifically, the S_z^{35S} input configuration provides a median MSE of 1.93×10^{-3}, corresponding to a prediction uncertainty of 961 years (9.61×10^{-3} Sv if we express the error in terms of freshwater forcing). In contrast, training solely on SSS fields produces the best results, with a median MSE of 1.79×10^{-4}, corresponding to a prediction uncertainty of 293 years (2.93×10^{-3} Sv). These errors remain acceptable given that, under a freshwater forcing of 10^{-5} Sv yr^{-1}, the CLIMBER-X model requires 21,890 years to reach the tipping point. Thus, an error of 961 years corresponds to 4.39%, while 293 years corresponds to 1.34% of the total temporal span of the simulation.

The CNN's accuracy improves consistently as r_F increases. From $r_F = 2 \times 10^{-4}$ Sv yr^{-1} onward, the CNN achieves an almost perfect fit between the predicted and actual distances to tipping, with very low percentage errors relative to the total span of the test simulations. This improvement holds across all input configurations but is particularly clear when the CNN is trained solely on SSS fields. For example, for $r_F = 6 \times 10^{-4}$ Sv yr^{-1} (Fig. 2d,f), training on SSS fields results in a median MSE of 1.7×10^{-5}, corresponding to an uncertainty of 1.5 years (9.03×10^{-4} Sv). In contrast, training on S_z^{35S} fields, which represents the worst-performing configuration also at $r_F = 6 \times 10^{-4}$ Sv yr^{-1}, produces a median MSE of 2.04×10^{-4}, corresponding to an uncertainty of 5.2 years (3.13×10^{-3} Sv). For $r_F = 6 \times 10^{-4}$ Sv yr^{-1}, the SST-only configuration performs slightly better but remains comparable to S_z^{35S}. Given that under a freshwater forcing rate of $r_F = 6 \times 10^{-4}$ Sv yr^{-1}, the CLIMBER-X model requires 365 years to reach the tipping point, errors of 5.2 and 1.5 years correspond to low percentage errors of 1.42% and 0.44%, respectively, relative to the total simulation length.

Having analyzed the performance of the CNN, we now turn to that of the LR model. Figures 2 and 3 show that the LR provides skillful predictions that are comparable to those of the CNN when using the S_z^{35S} input configuration, and that outperform the CNN for the SST, SSS, and SST+SSS configurations. The LR exhibits its weakest performance for the slowest forcing rate $r_F = 10^{-5}$ Sv yr^{-1}, when trained on S_z^{35S}. Specifically, for the test simulation obtained for $r_F = 10^{-5}$ Sv yr^{-1}, training on S_z^{35S} fields results in an MSE of 5.96×10^{-3}, corresponding to an uncertainty of 1690 years (1.69×10^{-2} Sv).

Unlike the CNN, LR performance improves substantially when trained on SST fields for the same slow-forcing simulation. In this case, the MSE decreases to 4.5×10^{-4}, corresponding to an uncertainty of 464 years (4.64×10^{-3} Sv). The best LR performance for $r_F = 10^{-5}$ Sv yr^{-1} is obtained using the SST+SSS configuration, which yields an MSE of 1.06×10^{-5}, corresponding to an uncertainty of 71.4 years (7.14×10^{-4} Sv), with the SSS-only configuration performing slightly worse.

As found for the CNN, LR accuracy improves consistently with increasing forcing rate, and, in general, the LR performs best when trained using the SSS-only input configuration. However, unlike the CNN, the SST+SSS configuration for the LR yields performance comparable to that of the SSS-only case and, for two forcing rates ($r_F = 10^{-4}$ and $r_F = 10^{-6}$ Sv yr^{-1}), slightly outperforms it.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

As previously discussed, the LR provides skillful predictions that are comparable to those of the CNN when using the S_z^{35S} input configuration, and that outperform the CNN for the SST, SSS, and SST+SSS configurations. The largest relative improvement of the LR with respect to the CNN is obtained for $r_F = 5 \times 10^{-4}$ Sv yr^{-1}. In this case, the LR model trained on SST, SSS, and SST+SSS fields yields MSE values of 1.13×10^{-5}, 4.34×10^{-7}, and 7.65×10^{-7}, corresponding to uncertainties of 1.45 (7.37×10^{-4} Sv), 0.283 (1.44×10^{-4} Sv), and 0.376 (1.91×10^{-4} Sv) years, respectively. These results correspond to improvements of 83.43%, 94.64%, and 95.65% relative to the CNN MSE. For all other forcing rates, LR performance improvements for the SST, SSS, and SST+SSS configurations remain above 65%, with the only exception being the SST+SSS configuration at $r_F = 10^{-4}$ Sv yr^{-1}, where the improvement is approximately 50%.

Taken together, the LR results presented in this section suggest that, within the CLIMBER-X model, the relationship between the input variables (SST, SSS, $SST + SSS$ and S_z^{35S}) and the normalized distance-to-tipping index d_f can be well approximated by a linear mapping. Consequently, when both models are trained and evaluated solely on CLIMBER-X data, there is no clear advantage in using the more complex CNN architecture.

The primary advantage of the CNN lies in its generalization capability. When trained on CLIMBER-X data and evaluated on the more complex CESM model, the LR model fails to provide reliable predictions, whereas the CNN demonstrates robust generalization performance (see Section 3.2). The limited generalization ability of the LR is further illustrated by its behavior when tested on the CLIMBER-X simulation with $r_F = 10^{-4}$ Sv yr^{-1}. In this case, the LR maintains strong predictive performance for most of the simulation but exhibits a clear degradation in the vicinity of the actual tipping point. This performance decrease is particularly found when the model is trained with the S_z^{35S} input configuration.

For $r_F = 10^{-4}$ Sv yr^{-1}, the collapse of the AMOC is initiated approximately 200 years before the system reaches its actual tipping point, marking the onset of a regime shift. While the CNN can adapt and generalize across this transition, the LR fails to do so, resulting in a pronounced performance degradation after collapse initiation. A detailed analysis of this limitation of the LR, which does not occur for the CNN, is provided in the Appendix (B).

A complete summary of the results obtained from training, validation, and testing on CLIMBER-X data for both the CNN and the LR models is reported in Table C1.

3.2 Generalization on CESM data

After assessing the CNN's performance when trained, validated, and tested exclusively on CLIMBER-X data, we evaluate its generalization capabilities. To this end, the CNN is trained using data from all seven CLIMBER-X simulations described in Section 3.1, and its performance is subsequently evaluated on the CESM AMOC collapse simulation presented in van Westen et al. (2024b). In this section, we first describe the preprocessing steps adopted to ensure compatibility between CLIMBER-X and CESM data. We then outline the modifications introduced to the CNN validation procedure for the generalization experiments. Finally, we present and discuss the CNN generalization results and compare them with those obtained using an LR model trained and evaluated under the same conditions.

For this analysis, we adopt the same input variable configurations as in the CLIMBER-X experiments. To ensure compatibility between CLIMBER-X and the CESM datasets, several preprocessing steps are applied. First, CESM SST and SSS fields

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

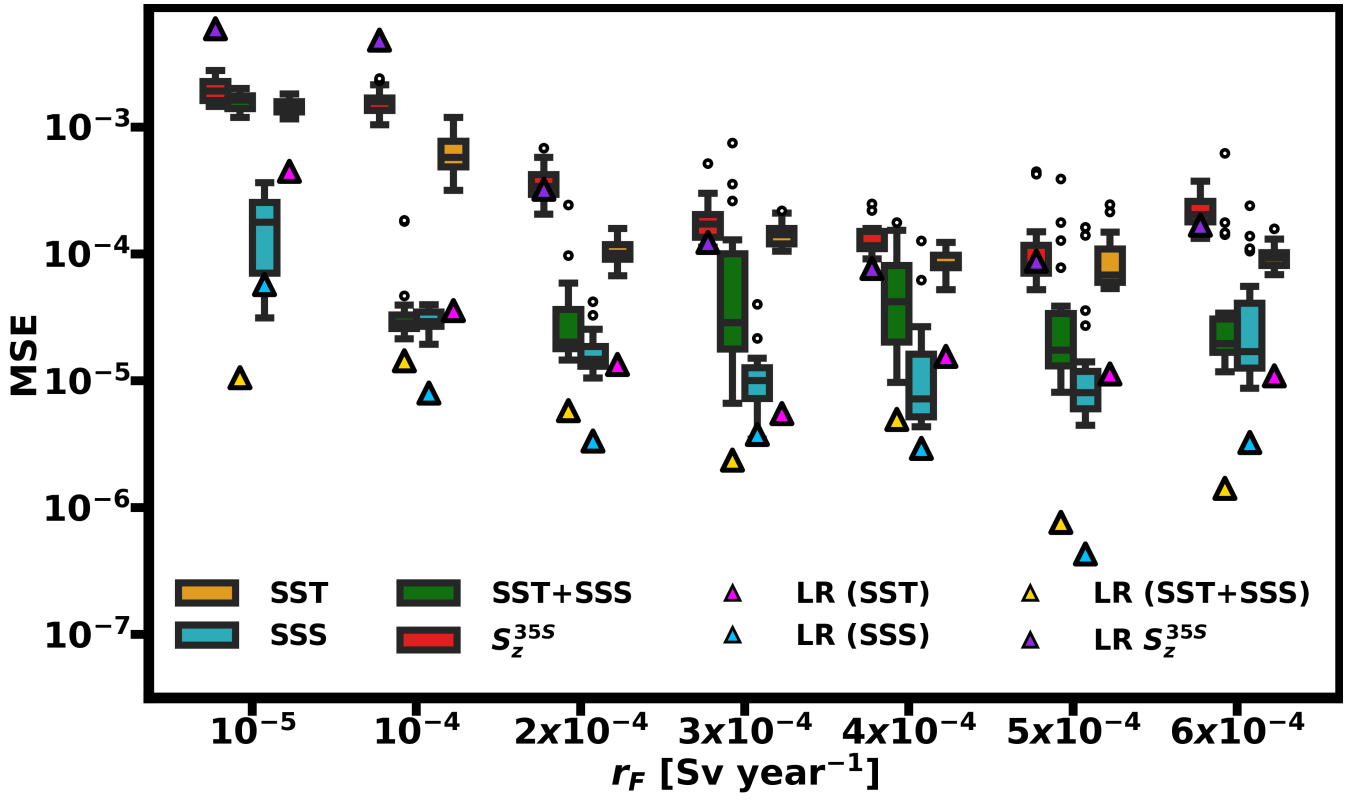


Figure 3. *Box plot showing the distribution of the results obtained using as test each of the seven distinct CLIMBER-X AMOC collapse simulations, each generated with a different forcing rate r_F. The boxes represent the distributions over 20 trials: for each trial, the model was reinitialized, retrained on all simulations except the one held out for test, and re-evaluated on that simulation. The black line inside each box marks the median; the lower and upper edges correspond to the 25th and 75th percentiles, spanning the interquartile range (IQR). Whiskers extend to the most extreme results within 1.5×IQR. Dots indicate outliers beyond this range. Triangles correspond to the performance of linear regressors trained using the same input-variable configurations as the convolutional neural networks and evaluated on the same test simulations.*

over the Atlantic are bi-linearly re-gridded to the 5° × 5° horizontal resolution of CLIMBER-X, and the 35°S salinity profile is interpolated onto CLIMBER-X's vertical levels using the same scheme. Second, for SST and SSS fields, the Baltic Sea and Hudson Bay are masked by setting the corresponding grid points to zero. These regions exhibit anomalously low surface salinities compared to the Atlantic mean (∼ 30 PSU), particularly the Baltic Sea, which has an average value of only ∼ 7 PSU. While CESM represents these low salinity features, CLIMBER-X does not, and their inclusion would produce strong outliers relative to CLIMBER-X's salinity distribution, thereby limiting cross-model prediction skill. For the same reason, the region along the Russian coast in front of the East Siberian Sea is also masked, since its salinity values are comparably low to those of the Hudson Bay and similarly deviate from the Atlantic mean. Third, in all generalization experiments, both CLIMBER-X and CESM inputs are normalized using the mean and standard deviation of the CLIMBER-X training data. Finally, the normalized

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

distance to tipping is computed relative to $F_H(t_p) = F_H^E = 0.53$, identified by van Westen et al. (2024b) as the AMOC tipping threshold in their CESM freshwater forcing experiment.

For CNN's hyperparameter selection, a different validation strategy than in the previous section is adopted. Instead of the leave-one-out scheme used previously (see section 2.4), the CESM data are partitioned into three sets: the first 430 years are discarded to remove transient effects, the following 440 years are used for validation to tune the hyperparameters, and the remaining 880 years serve as the test set (see Figure 4a). As in the CLIMBER-X experiments, validation and performance evaluation are performed separately for each input variable configuration. The complete list of hyperparameter values evaluated during validation is reported in table A1.

During these experiments, we found that combining early stopping with a reduce-on-plateau policy led to stronger overfitting on CLIMBER-X data and reduced the CNN's ability to generalize. To avoid this, the CNN is trained for a fixed number of epochs (n_{ep}) with a constant learning rate. The hyperparameters are therefore identical to those used in the CLIMBER-X experiments, except that n_{es} is replaced by n_{ep}. To ensure robustness, we perform 50 independent trials for each hyperparameter configuration. In each trial, the mean squared error (MSE) and the Pearson correlation between predicted and true distances to tipping on the validation set are computed and then averaged across trials. The optimal hyperparameter set is defined as the one that minimizes the median MSE, while also maintaining the mean correlation above a minimum threshold, $\rho_{\min}$ (we use the median rather than the mean for MSE to reduce the influence of outliers). This criterion ensures that the selected hyperparameters not only minimize error but also capture the expected decreasing trend in the distance to tipping. Thresholds are set to $\rho_{\min} = 0.8$ for SSS-only, SST+SSS, and S_z^{35S}, and to $\rho_{\min} = 0.5$ for SST-only, since no higher correlations are consistently achieved with this input setup.

Despite these measures, some variability remains in the validation results. For the SST+SSS and SSS-only configurations, the hyperparameter search does not always converge to the same optimal set. In contrast, for SST-only and S_z^{35S}, the search consistently yields a unique configuration. To account for this instability in the SST+SSS and SSS-only cases, we identify the top five hyperparameter configurations during validation, rather than relying on a single best configuration. In what follows, we present results obtained with the best-performing configuration on the CESM test set (last 880 years).

The optimal hyperparameters for the CESM generalization experiments, along with additional sets for the SSS-only and SST+SSS configurations, are listed in Tables A3 and A4, respectively. The results for the additional hyperparameter sets are shown in Figures A1 and A2.

Compared to the CLIMBER-X experiments, the optimal configurations for CESM generalization consistently favor a much smaller number of epochs. This suggests that successful generalization from the simpler CLIMBER-X model to the more complex CESM requires limiting the number of training epochs to prevent overfitting the CLIMBER-X data and to encourage the learning of more general patterns shared across models.

After hyperparameter selection, the CNN's generalization performance is tested separately for each input variable configuration on the last 880 years of CESM data. Since CESM performance shows greater variability than CLIMBER-X, we assess it more robustly by running 500 independent trials per input variable configuration. In each trial, the CNN is randomly reinitialized, trained on all seven CLIMBER-X simulations with the previously identified optimal hyperparameters, and then evaluated

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

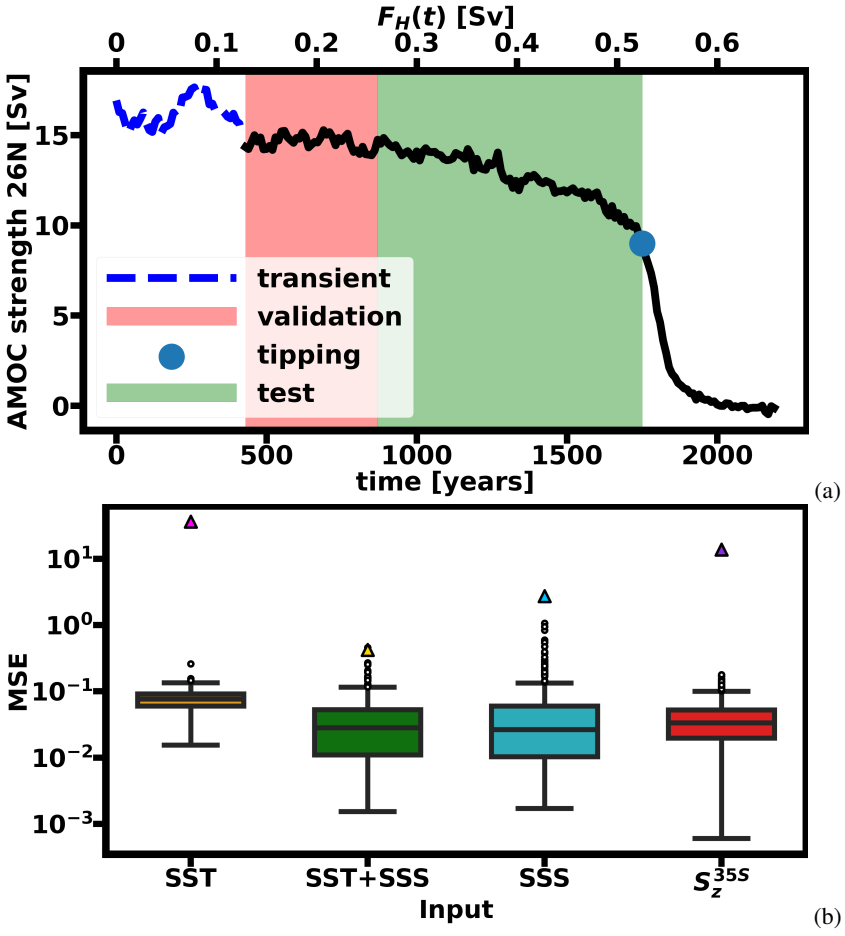


Figure 4. *(a) Plot illustrating how the CESM data is partitioned for the generalization experiments. The first 430 years (transient behavior) are discarded, the subsequent 440 years of data are used for validation, and the remaining 880 years are used for testing. (b) Box plot showing, for each input variable configuration, the distribution of results obtained by testing the CNN generalization performance on the CESM test dataset. Each box summarizes 500 trials: in each trial, the model was reinitialized, retrained on all seven CLIMBER-X simulations (see section 3.1), and then evaluated on the CESM test dataset. The black line inside each box marks the median; the lower and upper edges represent the 25th and 75th percentiles, defining the interquartile range (IQR). Whiskers extend to the most extreme values within 1.5×IQR, and dots indicate outliers beyond this range. Triangles indicate the corresponding LR performances included for comparison.*

on the CESM test set. Figure 4b presents the distribution of the MSE scores across the 500 trials for each input configuration. Figure 5 shows the median and the 90% confidence interval (5th to 95th percentile of the distribution) of the predicted distances to tipping across all 500 trials, along with the predictions obtained by the best-performing CNN realizations, compared to the actual distances to tipping. For comparison, the performances obtained from training and evaluating a linear regression (LR) model under the same conditions are also reported.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

From both Fig. 4(b) and Fig. 5, it is clear that the LR model fails to generalize to the CESM data and does not provide reliable predictions. This behavior is consistent across all input-variable configurations considered. Although the LR performance improves when trained on SST+SSS fields, its generalization ability remains poor. Even for this configuration, the LR results are consistently worse than those obtained by all 500 CNN realizations trained on SST+SSS fields (see Fig. 5).

In contrast, the CNN demonstrates the ability to generalize to a more complex climate model, maintaining reliable performance on the CESM data despite being trained exclusively on simulations from CLIMBER-X. This task cannot be achieved using the simple LR model considered here. Given the poor generalization performance of the LR model when trained on CLIMBER-X and evaluated on CESM data, the remainder of this section focuses exclusively on a more detailed analysis of the CNN performances obtained with the different input-variable configurations considered in this study.

The CNN shows the worst generalization perfomances among all input variables configuration when is trained solely on SST fields. In this case, the CNN is, on average, unable to capture the decreasing distance to tipping, as shown when examining the median of the 500 CNN realizations in Fig. 5b. The median prediction decreases only from 0.5 (model year 870) at the first time step to 0.48 (model year 918) at the final time step, whereas the actual value corresponds to model year 1750, the last time step before tipping at a 10-year temporal resolution, with tipping occurring at model year 1758. The median MSE between predictions and true distances is also relatively high, equal to 0.032, which corresponds to an uncertainty of 354 years (or 0.1 Sv in terms of freshwater forcing), as shown in Fig. 4b. Additionally, a significant variance in predictive performance is found among the different CNN realizations. This variability can be quantified as the interquantile range (IQR), defined as the distance between the 25th and 75th percentiles of the distribution of the 500 MSE scores. For the SST configuration, the IQR is 0.03, corresponding to an uncertainty of 100 years (0.03 Sv).

Also, when considering the best-performing realization among the 500 trials, the SST-only input configuration shows the worst results. The corresponding MSE is 1.5×10^{-2} for SST, equivalent to an uncertainty of 218 years (0.065 Sv). In contrast, for the [SST+SSS, SSS, S_z^{35S}] input configurations, the best-performing CNN realizations achieved MSE scores of $[1.4 \times 10^{-3}, 1.7 \times 10^{-3}, 6 \times 10^{-4}]$, corresponding to uncertainties of [69, 73, 43] years ([0.02,0.022,0.013] Sv), respectively. These results demonstrate that for these input configurations, there exist CNN realizations capable of strong generalization performance, closely matching the actual distances to tipping, even for the CESM data.

Considering average performance, the SSS-only configuration provides the best results among all the input configurations. As shown in Figure 4b, the median MSE score among the 500 CNN realizations is 0.026, corresponding to an uncertainty of 286 years (0.086 Sv). This aligns with the results obtained when the CNN was trained, validated, and tested only on CLIMBER-X data (see section 3.1). This consistency is expected, as both the CLIMBER-X and CESM datasets are generated through the same type of freshwater forcing experiments (see section 2.1 and 2.2). Therefore, in both datasets, the SSS fields are directly influenced by the applied freshwater perturbations, making them highly informative input variables for the CNN.

The SST+SSS and S_z^{35S} configurations yield a median MSE of 0.028 and 0.03, respectively, corresponding to 295 and 320 years (0.088, 0.096 Sv).

Overall, all three configurations outperform the SST-only setup in terms of median MSE, indicating substantially improved predictive skill. As discussed earlier, the CNN trained on SST only fails to capture any meaningful variability in the distance

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

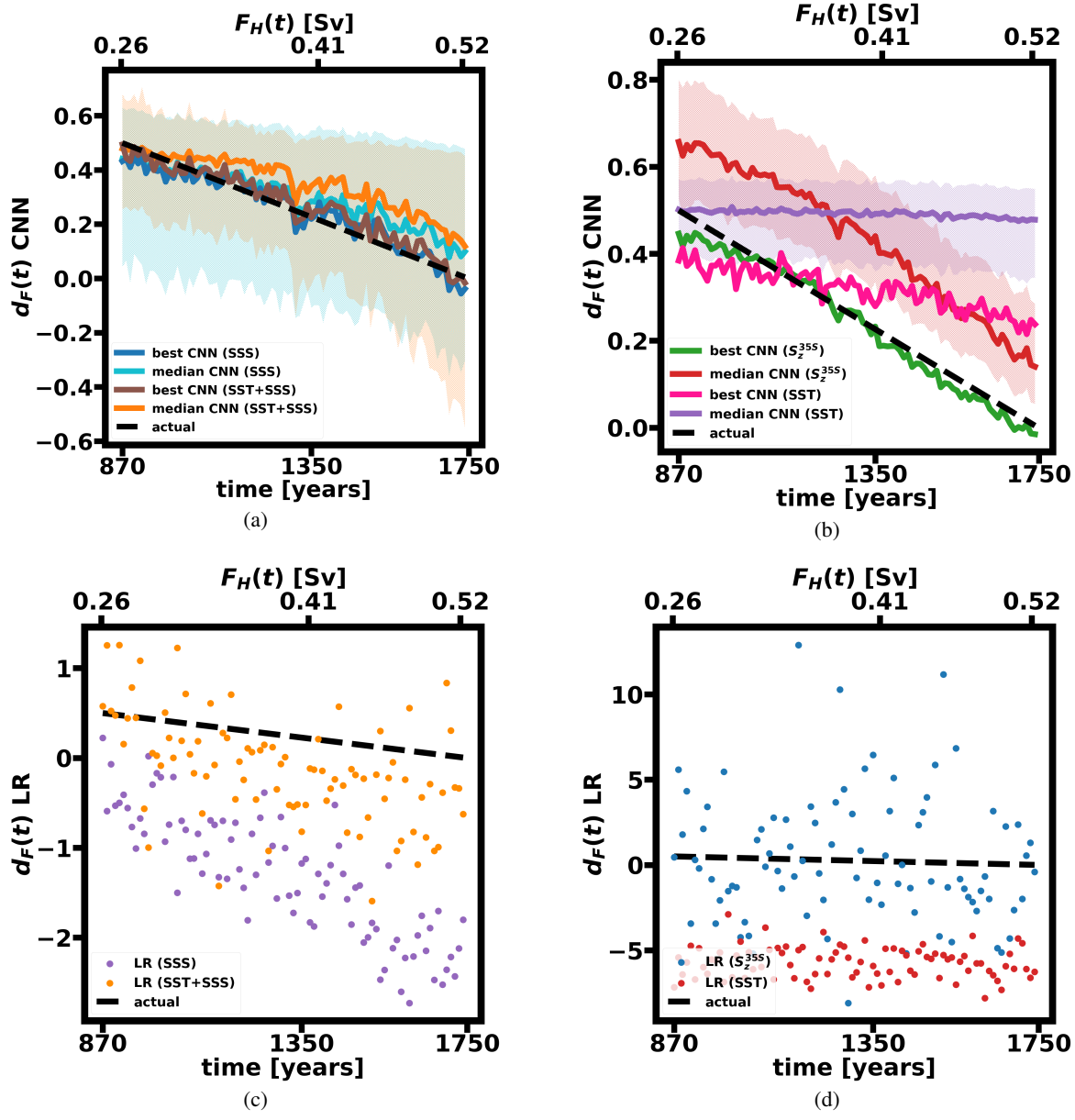


Figure 5. *CNN predictions of the distance to tipping versus the true values for the CESM test dataset. Panel (a) shows results for the CNN trained on the combined SST and SSS fields over the Atlantic and the CNN trained on SSS fields only. Panel (b) shows results for the CNN trained on SST fields only and the CNN trained on* S_z^{35S}. *Shaded areas indicate the* 90% *confidence interval computed over 500 independent trials. In each trial, the network was randomly reinitialized, retrained using all seven CLIMBER-X simulations (see Section 3.1), and subsequently evaluated on the CESM test dataset. For comparison, panels (c) and (d) present the results obtained for an LR model trained and evaluated on the same datasets. Panel (c) shows results for the LR trained on SSS fields only and on the combined SST+SSS fields over the Atlantic, while panel (d) shows results for the LR trained on SST fields only and on* S_z^{35S}.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

to the tipping point. In contrast, when trained on SST+SSS, SSS-only, or S_z^{35S} fields, the CNN is able, on average, to predict a clear and consistent decreasing trend as the system approaches the tipping point (Fig. 5). At the first time step of the test period, the median predicted distances are 0.48 (corresponding to model year 918) for the SST+SSS configuration, 0.657 (model year 603) for the S_z^{35S}, and 0.44 (model year 982) for the SSS-only configuration. At the last time step, the median predictions are approximately 0.12 (model year 1550) for SST+SSS, 0.14 (model year 1511) for the S_z^{35S} configurations, and 0.1 (model year 1584) for SSS-only. These results demonstrate that, when trained with any of these three input configurations, the CNN is capable of producing, on average, a coherent and consistent decreasing d_F over time. The variability in predictive performance across the 500 realizations differs among the configurations. The SST+SSS and SSS-only setups show relatively high interquartile ranges (IQR) of 0.042 and 0.05, corresponding to uncertainties of 221 years (0.066 Sv) and 252 years (0.076 Sv), respectively. By contrast, the S_z^{35S} configuration exhibits much lower variance, with an IQR of 0.033, corresponding to 155 years (0.047 Sv) of uncertainty.

Finally, we discuss the biases and variability in the results associated with the three input configurations, also considering the predictions obtained using the other optimal hyperparameter configurations identified during validation for the SSS and $SST + SSS$ input configurations (see Figs. A1 and A2 in the Appendix).

The $SST + SSS$ configuration exhibits greater variability in performance compared to both the SSS-only and S_z^{35S} configurations, as shown in Fig. A2. Among the optimal hyperparameter configurations identified during validation, two produce results comparable to those obtained with the best configuration reported in this section. In contrast, one configuration performs poorly and fails to capture the decreasing distance to the tipping point, while another systematically overestimates the distance to tipping, although it still reproduces a coherent decreasing signal as the system approaches the tipping point.

Although not immediately clear from the predictions obtained using the best-performing configuration shown in this section, the SSS-only configuration tends to systematically underestimate the distance to tipping. This behavior is found when examining predictions obtained with the other optimal hyperparameter configurations identified during validation (see Fig. A1). Across these configurations, the CNN trained solely on SSS consistently underestimates the distance to tipping, while still capturing a reliable decreasing signal over time. As a result, the SSS configuration appears more robust than $SST + SSS$, which exhibits greater variability in performance across different hyperparameter configurations.

For the S_z^{35S} input configuration, the CNN exhibits a systematic overestimation bias that gradually decreases as the tipping point is approached. Despite this tendency and its slightly lower performance compared to the SSS and $SST + SSS$ configurations, the S_z^{35S} configuration can be considered more stable, as only one hyperparameter configuration consistently performs best during validation.

Overall, these results highlight the potential of all three input configurations to enable reliable generalization to more complex climate-model data. In particular, the good performance of the S_z^{35S} configuration is consistent with the findings of van Westen et al. (2024b), which identified the freshwater transport of the AMOC at 35°S as a physics-based early warning indicator of AMOC tipping in the CESM model.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

4 Results: Explainability

4.1 Temporal trends and relevance maps computation

To identify the spatial patterns in the input data that are most relevant to the CNN and LR predictions for the CLIMBER-X data, and to the CNN predictions for the CESM data, we apply the SHAP methodology (see Appendix D). Shapley values assign a relevance score to each input feature relative to a reference prediction. This reference prediction corresponds to the expected value of the model output evaluated over a subset of the input data, referred to as the background state.

For the CLIMBER-X explainability analysis, the background state is constructed from the first 1000 years (model years 0–1000) of the simulation with the slowest forcing rate, $r_F = 10^{-5}$ Sv yr^{-1}. For the CESM analysis, the background state is defined using the first 100 years of the CESM test dataset, corresponding to model years 870–970 of the AMOC collapse simulation presented in van Westen et al. (2024b).

This choice reflects the experimental setup: the first 430 years of the CESM simulation are discarded as transient, the following 440 years are used for validation, and model performance is evaluated only on the remaining 880 years (corresponding to model years 870-1750). Accordingly, the explainability analysis is restricted to this test period, and the background state is selected from the same subset to ensure consistency.

With this choice of background state, the reference prediction corresponds to different regimes in the two experiments. In the CLIMBER-X case, the background samples correspond to conditions close to a stable AMOC state. In the CESM case, the background samples correspond to intermediate states of the simulation, with predicted distances to tipping of approximately $d_F \approx 0.5$, reflecting conditions roughly halfway to collapse.

Since the samples analyzed in both the CLIMBER-X and CESM explainability analyses correspond to system states that progressively approach the tipping point relative to the chosen reference state, the sign of the Shapley values can be interpreted in terms of AMOC stability. Negative Shapley values highlight input regions associated with contributions pointing toward AMOC destabilization, whereas positive Shapley values correspond to contributions pointing toward AMOC stabilization.

To assess the robustness of the SV relevance maps, we examine whether regions identified as highly relevant exhibit consistent and physically meaningful temporal evolution as the AMOC approaches the tipping point. For this purpose, we analyze the temporal trends of the input variables at the grid points highlighted by the explainability analysis. The underlying rationale is that, if a location consistently receives high relevance scores, variations in the corresponding variable strongly influence the predicted distance to tipping. Consequently, the variable at that location is expected to display a consistent temporal evolution as the system approaches tipping, in the form of a systematic increasing or decreasing trend that reflects the progressive destabilization of the AMOC.

To identify the grid points most relevant for the CNN predictions, meaning those consistently associated with strong relevance scores, we follow a three-step procedure. First, for each time step prior to tipping, and for each input variable configuration and CLIMBER-X or CESM collapsing trajectory used for testing, we compute the corresponding relevance map. Second, the relevance values at each grid point are normalized by the maximum absolute relevance across the entire map at that time

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

step. Third, we compute the temporal average of the normalized relevance for each grid point and retain only those whose mean relevance in absolute value exceeds a specified threshold, denoted by ρ_{rel}.

For the grid points identified as most relevant, we then examine whether the associated input variables evolve consistently as the system approaches tipping. To do so, for each grid point we extract the associated time series, we isolate its trend component, and apply the Mann–Kendall test to determine whether the trend is significantly increasing or decreasing at the 95% confidence level. For grid points with a significant trend, the magnitude is quantified using Sen's slope estimator.

4.2 Explainability results: CLIMBER-X

To ensure that the analysis of input-field relevance is not affected by poorly performing CNN realizations, we select, for each input-variable configuration and for each CLIMBER-X simulation used for testing, the best-performing model among the 20 independent training trials (see Section 3.1). Since the LR model does not involve random weight initialization, its training procedure is deterministic, and therefore, this selection step is not required for the LR model. We first present the explainability results obtained for the CNN model and subsequently those obtained for the LR model. Figure 6 shows the results of the trend analysis for the SST and SSS fields derived from CLIMBER-X simulations performed under two freshwater forcing rates, $r_F = [10^{-5}, 6 \times 10^{-4}]$ Sv yr^{-1}. Superimposed contours highlight regions where the CNN Shapley value relevance scores remain consistently high over time, defined as grid points with a mean absolute relative relevance score $\rho_{rel} \geq 0.2$. These relevance scores are obtained from CNN models trained separately on the SST and SSS fields. For visualization purposes, Arctic Ocean regions located north of Siberia and Canada are not displayed, although they were included during training, as they do not exhibit significant contributions to the model predictions.

The relevance maps for intermediate forcing rates (r_F=[10^{-4}, 2×10^{-4}, 3×10^{-4}, 4×10^{-4}, 5×10^{-4}] Sv yr^{-1}) are not shown for SST and SSS inputs, as they closely resemble those obtained for the slowest ($r_F = 10^{-5}$ Sv yr^{-1}) and fastest ($r_F = 6 \times 10^{-4}$ Sv yr^{-1}) forcing cases. Similarly, relevance maps obtained when the CNN is trained on combined SST and SSS inputs are also omitted, as they display patterns very similar to SSS alone. This outcome is expected, given that the network achieves significantly better performance when trained solely on SSS fields (see section 3.1). As a result, during training, the network tends to assign greater relevance to the SSS fields.

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. Their perturbation experiment applied a freshwater flux of 0.1 Sv over 100 years in the North Atlantic between 50°N and 70°N. Consistent with their findings, our results indicate that a weakening of the AMOC is accompanied by a decrease in SST in the North Atlantic, with particularly strong cooling south of Greenland. In the South Atlantic, a modest SST increase is occurs, peaking along the coasts of West Africa, South America, and near the equator. Our analysis also reveals SSS patterns similar to those reported by Stouffer et al. (2006). In the North Atlantic, surface salinity decreases significantly in response to the freshwater perturbation, which in our case is applied between 20°N and 50°N. This freshening is particularly pronounced around 25°N near the coasts of Africa, where CLIMBER-X simulations show a strong and persistent decline in salinity. The most substantial reduction occurs in the Greenland-Iceland-Norwegian Seas (GINS), a key region for deep water formation. Although less intense than in the

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

GINS, the 25°N freshening remains sustained over time. We also find an increase in SSS in the South Atlantic, linked to reduced northward salt transport, with the largest anomalies around 30°S extending from South America to the central basin. An increase in salinity is also found in the Gulf of Mexico.

Turning to the explainability results obtained when the CNN is trained solely on SSS fields, we find that under both the slowest ($r_F = 10^{-5}$ Sv yr^{-1}) and fastest ($r_F = 6 \times 10^{-4}$ Sv yr^{-1}) forcing scenarios the network assigns high relevance to regions located within and near the hosing area. Strong negative contributions originate from areas around 25°N, close to the African coast, where a consistent decrease in salinity occurs. This signal clearly reflects the progressive destabilization of the system as it approaches the tipping point.

For the slower forcing scenario ($r_F = 10^{-5}$ Sv yr^{-1}), strong negative relevance scores are also assigned to regions around 15°N, just outside the hosing area, where salinity exhibits a persistent increase. In the faster forcing case, a single grid point associated with a relevant positive contribution appears close to the region around 25°N that exhibits strong negative relevance scores. Although this point is identified as positively contributing, its relevance score is much smaller than those associated with the surrounding negative contributions. Its overall influence on the model output can therefore be considered negligible and effectively compensated by the stronger negative contributions in the surrounding area, allowing the network to produce a coherent decreasing prediction of the distance to the tipping point.

Overall, these results indicate that the CNN correctly focuses on regions where the most consistent salinity changes occur within and close to the hosing area.

According to the SHAP results for SST, grid points over the Greenland, Norwegian, and Barents Seas consistently provide the strongest negative contributions to the predicted distance to tipping under both forcing scenarios. This result is consistent with our trend analysis and previous studies (e.g., Willeit and Ganopolski (2024)), which show that in CLIMBER-X a slowdown of the AMOC is accompanied by a pronounced and sustained SST decline across the Nordic Seas. The CNN captures this relationship by assigning high negative relevance to regions where cooling signals a weakening of the AMOC.

Under both forcing scenarios, strong negative scores are also assigned to the Labrador Sea. SST anomalies in the subpolar North Atlantic, including the Labrador Sea, have been identified as a characteristic fingerprint of AMOC weakening (Caesar et al. (2018)).

For the faster-forcing case, additional negative contributions appear along the European coast around 45°N and in a small area off the African coast at 15°N, two regions where the simulations exhibit a persistent decrease in SST over time. In contrast, for the slower forcing case, additional positive contributions emerge in parts of the Nordic Seas, i.e., in regions that are otherwise strongly associated with AMOC destabilization. These positive contributions are comparable in magnitude to the negative contributions in the same region but are spatially less extensive. As a result, they are compensated by the more widespread negative contributions, producing an overall coherent decreasing signal in the model output.

Moreover, when trained on SST and tested on the $r_F = 10^{-5}$ Sv yr^{-1} scenario, the CNN exhibits the worst predictive performance among all experiments performed on CLIMBER-X data. The presence of strong positive contributions in regions physically linked to AMOC weakening may therefore reflect the reduced predictive skill of the model in this configuration, leading to less consistent attribution patterns.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

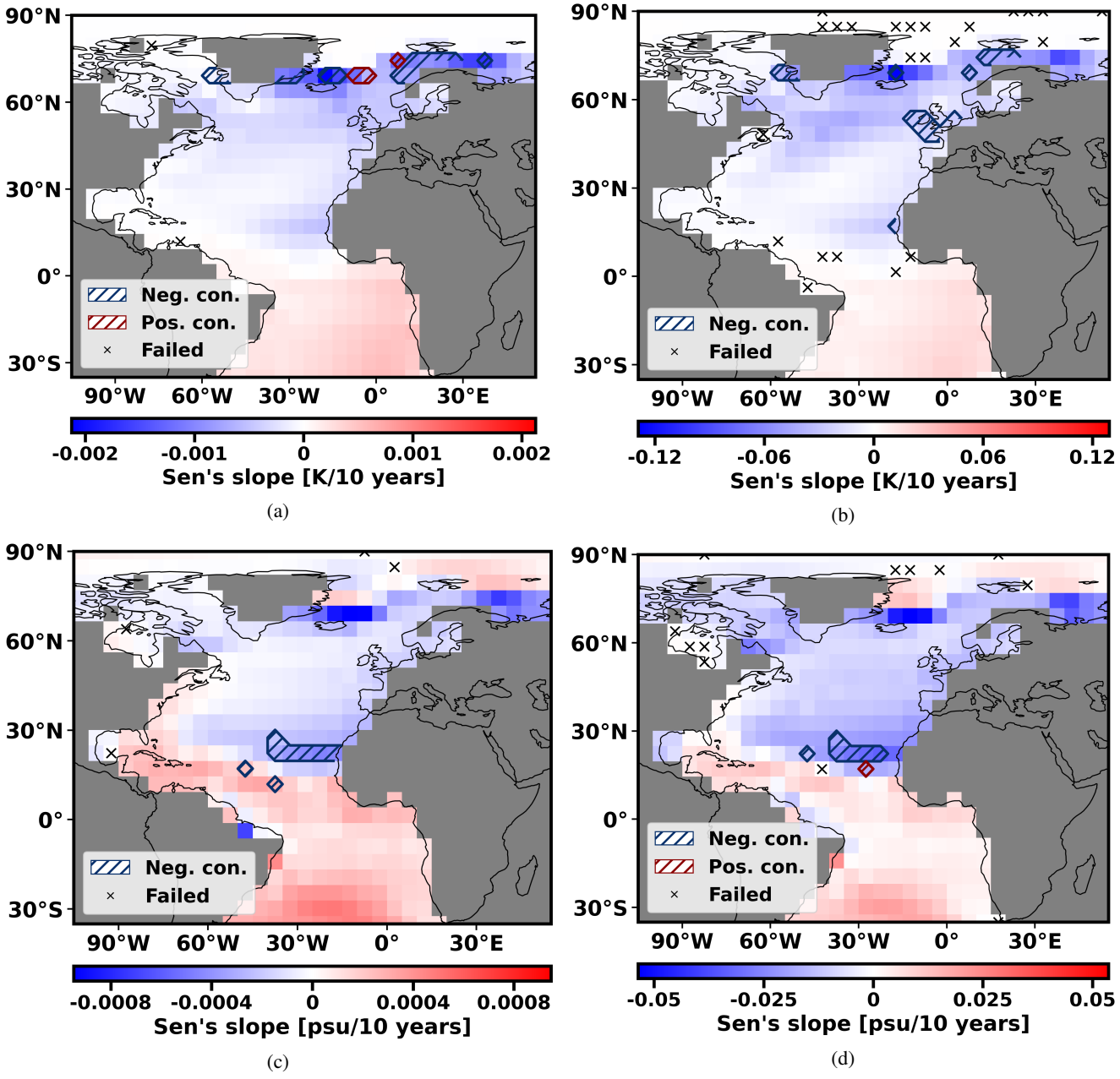


Figure 6. *Maps showing temporal trends in SST and SSS fields from CLIMBER-X simulations as the system approaches the AMOC tipping point, based on Sen's slope estimator. Two freshwater forcing rates are considered:* $r_F = 10^{-5}$ *Sv yr*$^{-1}$ *(panels a, c) and* $r_F = 6 \times 10^{-4}$ *Sv yr*$^{-1}$ *(panels b, d). Panels (a–b) show SST trends; panels (c–d) show SSS trends. Crosses mark locations where trends are not statistically significant (p-value from the Mann-Kendall test* ≥ 0.05*). Grey areas denote grid points excluded from the CNN input (set to zero). Red and blue contours highlight grid points with consistent positive or negative contributions to the CNN output over time. A point is considered relevant if its mean absolute relevance over time (*ρ_{rel}*) exceeds* 0.2*, with relevance at each time step normalized by the maximum across all grid points. Relevance scores are computed using the SHAP methodology, applied to the best-performing CNN realization (out of 20), trained solely on either SST or SSS fields.*

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Having examined the SST and SSS relevance patterns, we now turn to the vertical salinity structure at 35°S. As shown in Fig. 7, the temporal trends in S_z^{35S} are broadly consistent with those reported by Stouffer et al. (2006). On average, positive salinity anomalies extend down to approximately 1000 m. This increase results from the reduced northward transport of high-salinity surface waters as the AMOC weakens. Below this depth, salinity decreases. This deep freshening reflects a reduction in North Atlantic Deep Water formation, which weakens the southward transport of saline deep waters into the South Atlantic and allows Antarctic Bottom Water to penetrate farther northward.

We next examine the S_z^{35S} SHAP scores, also shown in Fig. 7. As in the previous analysis, we present results only for the forcing rates $r_F = 10^{-5}$ and 6×10^{-4} Sv yr^{-1}, as the intermediate forcing rates exhibit similar relevance patterns.

For both forcing scenarios, strong negative relevance scores are associated with grid points in the upper ocean layers, down to roughly 500 m depth. These regions extend from approximately 15°E to 30°W for the slower forcing case ($r_F = 10^{-5}$ Sv yr^{-1}) and from about 10°E to 15°W for the faster forcing case ($r_F = 6 \times 10^{-4}$ Sv yr^{-1}). In these areas, salinity increases substantially over time as a consequence of the reduced northward export of high-salinity surface waters during the AMOC weakening. The CNN assigns strong negative relevance to these regions, correctly highlighting an area where variability in upper-ocean salinity provides a strong signal of AMOC destabilization.

Positive relevance contributions are also present for both forcing scenarios. In the case of $r_F = 10^{-5}$ Sv yr^{-1}, these positive contributions occur in regions that exhibit relatively weak temporal variability as the tipping point is approached. It is therefore plausible that the network associates these areas with conditions that remain closer to the stable regime, since salinity values in these regions vary much less over time than those highlighted by negative contributions.

For both forcing rates, these positive contributions are smaller in magnitude and spatial extent than the negative contributions discussed above. Consequently, their overall influence on the CNN prediction is limited, and the combined relevance pattern remains consistent with the network producing a coherent decrease in the predicted tipping distance as the system approaches the critical threshold.

Having analyzed which spatial patterns in the input data lead to the most important CNN predictions, we now perform the same explainability analysis for the LR model. Figure 8 presents the results of this analysis for LR models trained individually on SSS and SST fields. As before, the two forcing rates considered are $r_F = [10^{-5}$ Sv yr^{-1},6×10^{-4} Sv yr$^{-1}]$. To identify the most relevant features in the input data, the SHAP methodology is applied again, and a feature is considered relevant if its mean normalized relevance score over time (ρ_{rel}) exceeds 0.2.

For the LR model, regions of the Arctic Ocean north of Siberia and Canada are also displayed, since the LR model tested on $r_F = 10^{-5}$ Sv yr^{-1} and trained solely on SSS fields tends to assign large relevance scores to some grid points in this area. Overall, the LR explainability maps for models trained on SSS and SST fields appear more scattered and difficult to interpret than those obtained for the CNN, and they also vary substantially across different r_F values (not shown).

The only stable patterns found for the SST fields across the different forcing rates correspond to regions with a higher density of relevant contributions. In the South Atlantic, these contributions are concentrated along the African coast and extend toward the central basin, while in the North Atlantic they appear close to the African coast around 15°N. These regions indeed

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

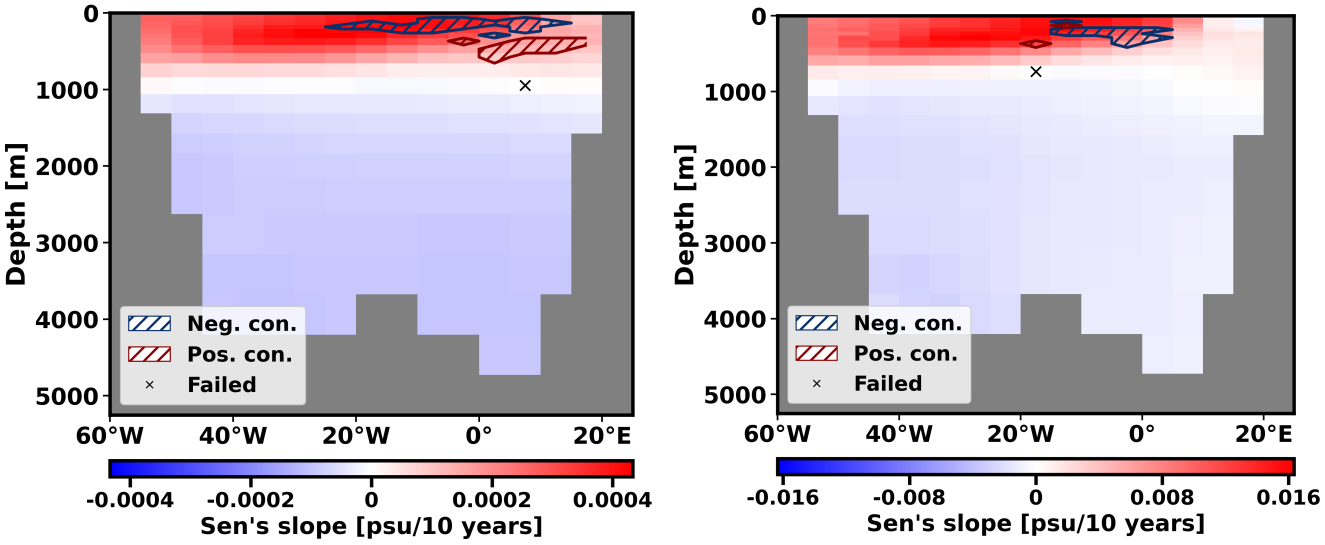


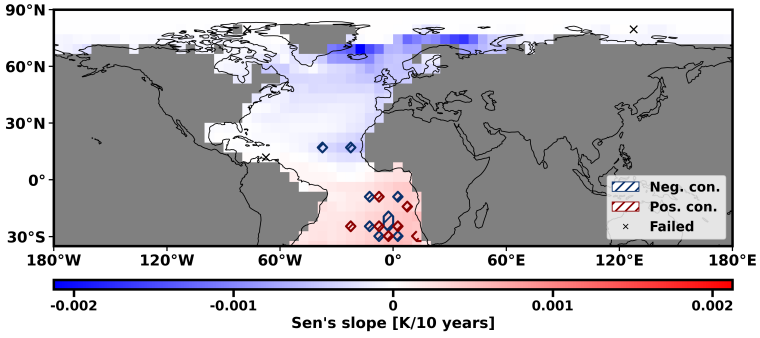
Figure 7. *Maps showing temporal trends in S_z^{35S} fields from CLIMBER-X simulations as the system approaches the AMOC tipping point, based on Sen's slope estimator. The freshwater forcing rates considered are: $r_F = 10^{-5}$ Sv yr^{-1} (panel a) and $r_F = 10^{-6}$ Sv yr^{-1} (panel b). Crosses mark locations where trends are not statistically significant (p-value from the Mann-Kendall test ≥ 0.05). Grey areas denote grid points excluded from the network input (set to zero). Red and blue contours highlight grid points showing consistent positive or negative contributions to the CNN output over time. A point is considered relevant if its mean absolute relevance over time (ρ_{rel}) exceeds 0.2, with relevance at each time step normalized by the maximum across all grid points. Relevance scores are computed using the SHAP methodology, applied to the best-performing CNN realization (out of 20).*

exhibit consistent temporal trends as the system approaches the tipping point, with an increase in SST in the South Atlantic region and a decrease in SST in the North Atlantic region.

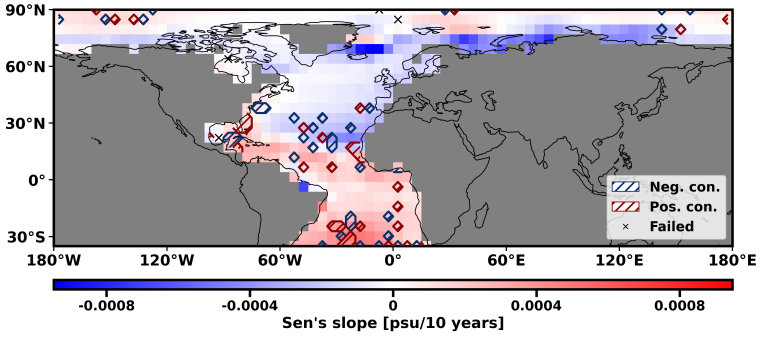
For the SSS fields, the only robust patterns across the different forcing rates are regions with a high density of relevant contributions around 30°S in the South Atlantic, where salinity increases over time, and within the hosing region near 25°N, where salinity consistently decreases. When the LR model is trained on the combined $SST + SSS$ fields, these same regions also emerge as areas of high relevance density. However, within these regions, the model assigns positive and negative contributions to a comparable number of grid points. As the tipping point is approached and the predicted distance decreases, no clear dominance of negative contributions is found. This indicates that the LR model does not clearly distinguish which spatial patterns contribute to the destabilization of the AMOC.

The relevance patterns identified for the S_z^{35S} fields appear more stable across the different forcing rates r_F. The only exception is the case $r_F = 10^{-5}$ Sv yr^{-1}, for which the LR model also assigns relevance to grid points in the deeper layers. Starting from $r_F = 2 \times 10^{-4}$ Sv yr^{-1}, the LR explainability maps begin to resemble the relevance map obtained for $r_F = 6 \times 10^{-4}$ Sv yr^{-1}, which is shown in Fig. 9.

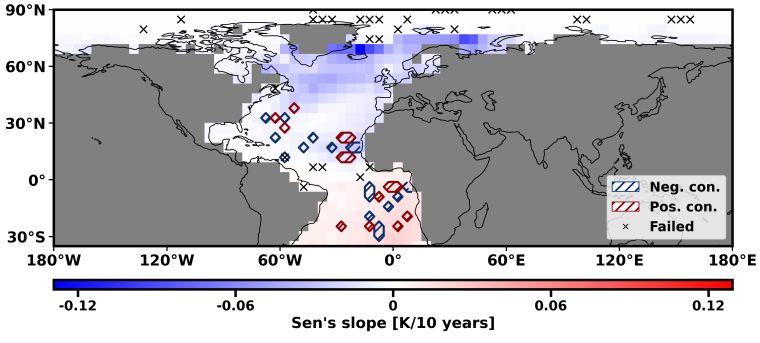
https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.



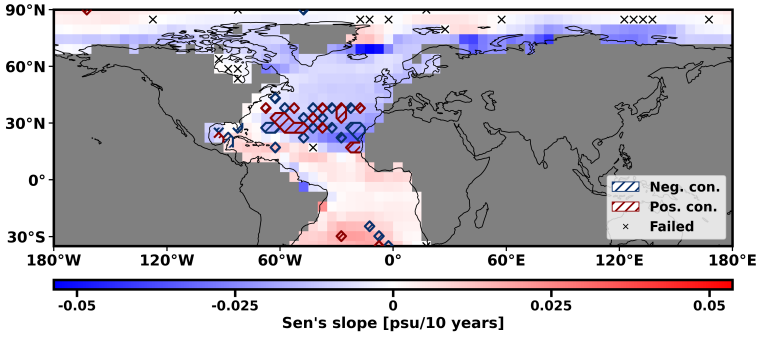
(a)



(b)



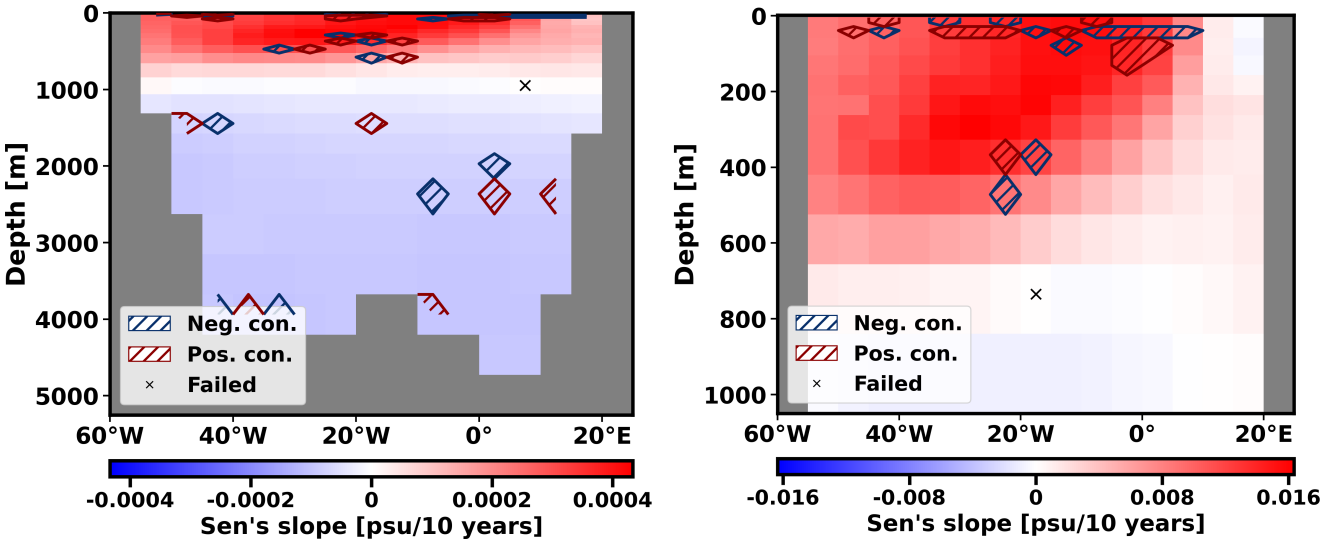
(c)



(d)

Figure 8. As in Fig. 6, but here red and blue contours highlight grid points with consistent positive and negative contributions to the LR output over time. As in Fig. 6, the panels refer to simulations obtained for $r_F = 10^{-5}$ Sv yr^{-1} in (a)–(b) and $r_F = 6 \times 10^{-4}$ Sv yr^{-1} in (c)–(d). The variables analyzed are SST fields in (a) and (c) and SSS fields in (b) and (d) over the Atlantic.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.



(a) (b)

Figure 9. As in Fig. 7, but here red and blue contours highlight grid points with consistent positive and negative contributions, respectively, to the LR output over time. As in Fig. 7, the panels refer to simulations obtained for $r_F = 10^{-5}$ Sv yr^{-1} in (a) and $r_F = 6 \times 10^{-4}$ Sv yr^{-1} in (b). In both panels, the variable analyzed is S_z^{35S}.

In these cases, the LR model tends to focus primarily on the upper layers down to about 500 m depth, with most of the relevant contributions concentrated in the top layers above 200 m and extending across the entire basin. As found for the SSS and SST fields, we do not find a clear predominance of negative contributions. This again suggests that the LR model is unable to clearly identify which spatial patterns contribute to destabilizing the AMOC.

The LR relevance maps reveal that, in contrast to the CNN model, the LR is not able to clearly identify coherent and stable large-scale spatial patterns associated with the destabilization of the AMOC. Instead, the LR tends to assign importance to individual grid points rather than capturing extended spatial structures in the input fields. As a consequence, the relevance patterns appear more scattered and less physically interpretable than those obtained with the CNN. This behavior likely contributes to the poorer generalization capabilities of the LR model, both when applied to the more complex CESM simulations (see Section 3.2) and when evaluated under different dynamical regimes within the CLIMBER-X model itself (see Appendix B).

4.3 Explainability results CESM data

The same explainability analysis applied to the Climber data and described in section 4.1 has also been conducted to interpret the predictions made on the CESM data. For the CESM relevance analysis, we focus on the best-performing CNN realization among 500 (see section 3.2). Figure 10 presents the SHAP relevance maps obtained from the best-performing CNN realizations, when the model was trained either on SSS fields alone or on S_z^{35S} fields from all seven CLIMBER-X simulations used in the preliminary experiments (see section 3.1). Relevance maps for the CNN trained solely on SST fields are not shown, as this

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

configuration exhibited poor generalization performance (section 3.2), making the resulting patterns unreliable. For this reason, the relevance scores for the LR model are also not shown, as it is unable to generalize on the more complex CESM data. In the SST+SSS configuration, most relevance is attributed to the SSS component. Consequently, the relevance maps for the SSS-only and SST+SSS configurations (not shown) display very similar patterns, since the network largely focuses on SSS features during training.

As in the CLIMBER-X case, relevance scores are overlaid on temporal trend maps, which illustrate the evolution of variable values across different locations. The temporal trends of the CESM SSS fields resemble those obtained from the CLIMBER-X data. We find a consistent decrease in SSS in the North Atlantic as a result of the freshwater perturbation, which, consistent with the CLIMBER-X simulations, was applied between 20°N and 50°N. As in CLIMBER-X, a pronounced salinity decrease occurs in the Greenland-Iceland-Norwegian Seas (GINS), corresponding to key regions of deep water formation. However, unlike the CLIMBER-X data, the most pronounced SSS decrease in the CESM data occurs within the hosing area itself. Similar to CLIMBER-X, the CESM SSS fields also display a consistent salinity increase in the South Atlantic, due to a reduction in northward salt transport, as well as a consistent increase in the Gulf of Mexico.

The SSS relevance maps for the CESM data reveal strong negative contributions in the hosing region, extending from 30°N to 60°N and spanning from 0 to 30°W. This area exhibits the largest decrease in salinity over time within the hosing region in the CESM simulation, and the network correctly identifies it as strongly associated with AMOC destabilization. Additional negative contributions are also present farther north, near the coast of Scandinavia.

The SSS relevance maps for CESM also show consistent negative contributions in the South Atlantic near the coast of South America and in the Gulf of Mexico. Both regions display a persistent increase in salinity over time in both CLIMBER-X and CESM as the system approaches the tipping point.

Overall, the relevance maps computed for the CESM SSS fields indicate that the CNN trained on CLIMBER-X is able to successfully generalize to the CESM simulations. In particular, the network identifies in CESM the regions exhibiting the most coherent and physically meaningful salinity trends associated with the approach to the tipping point. This behavior likely reflects the convolutional architecture of the network, which detects spatial structures rather than relying on exact geographic coordinates. As a result, the CNN can recognize similar large-scale patterns even when their precise spatial location differs slightly between models due to differences in model dynamics, provided that these patterns occur within broadly comparable regions of the basin.

The temporal patterns of the S_z^{35S} fields closely resemble those found in the CLIMBER-X simulations, with a positive salinity trend in the upper layers down to approximately 1000 m depth and a negative trend below this level. When applied to the more complex CESM model, the CNN primarily focuses on the upper ocean layers, extending to about 1000 m depth across the Atlantic basin. In this region, the relevance maps highlight strong negative relevance scores.

Within the same region, additional positive contributions are present but are more scattered and characterized by smaller relevance scores and a more limited spatial extent. Consequently, their influence on the model output is much weaker compared to the dominant negative relevance patterns described above. The predominance of these highly relevant negative regions allows the network to produce a consistently decreasing d_F as the AMOC approaches its tipping point.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

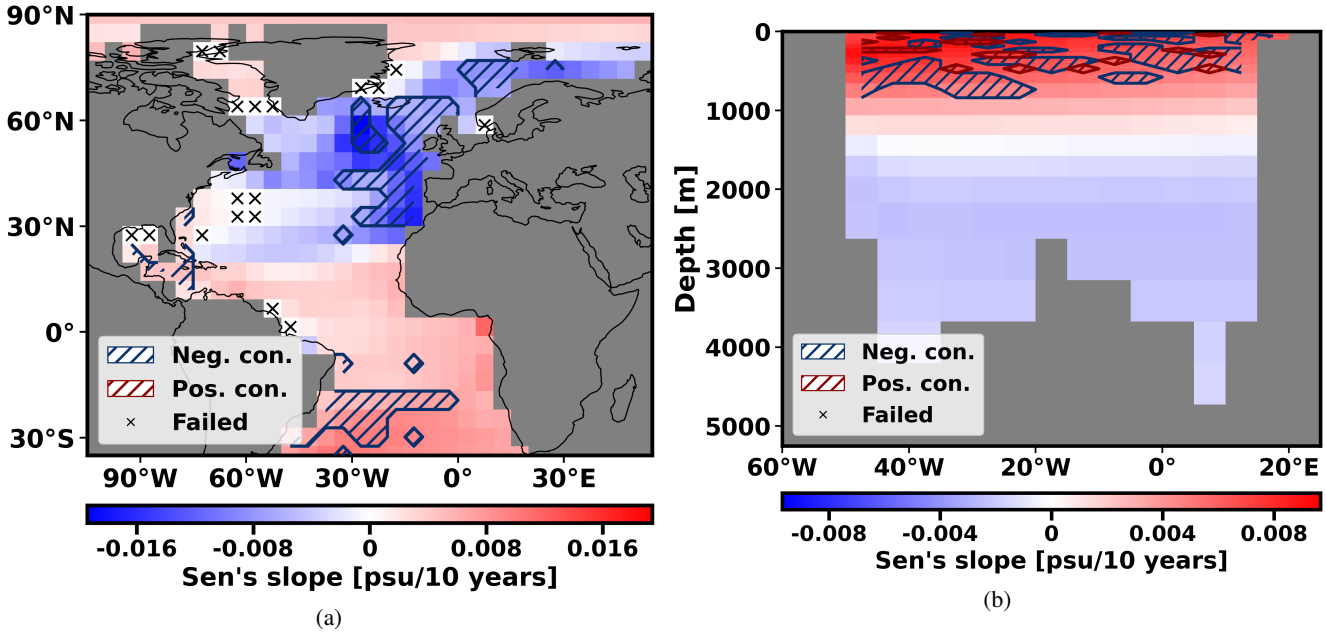


Figure 10. *Temporal trends in sea surface salinity (SSS) and full-depth salinity along 35°S (S_z^{35S}) as the Atlantic Meridional Overturning Circulation (AMOC) approaches the tipping point, based on the CESM AMOC collapse simulation (van Westen et al. (2024b)). Trends are estimated using Sen's slope estimator. Panel (a) presents SSS trends, panel (b) shows S_z^{35S} trends. Crosses mark areas where trends are not statistically significant, based on the Mann-Kendall test at the 95% confidence level. Red and blue contours indicate regions of consistent positive and negative relevance, respectively, derived from SHAP scores produced by the best-performing convolutional neural network (CNN) out of an ensemble of 500 models. This CNN was trained on SSS fields for panel (a) and on S_z^{35S} fields for panel (b), using data from all seven AMOC-collapse simulations analyzed in the CLIMBER-X experiments (see section 3.1). A grid point is considered relevant if its mean absolute relevance score over time (ρ_{rel}) exceeds 0.2. All relevance scores are normalized by the maximum absolute value at each time step.*

As discussed earlier, the upper layers down to approximately 1000 m depth exhibit consistent increases in salinity in both the CLIMBER-X and CESM simulations. Moreover, the increase in salinity in these upper layers is stronger in amplitude and more persistent over time than the decrease observed in the deeper layers. The CNN therefore correctly identifies the upper ocean layers as providing a more robust indicator of the AMOC's destabilization in both CLIMBER-X and CESM, compared to the weaker and less persistent signals present at greater depths.

5 Conclusions

Considerable efforts have been made in recent years to develop early warning signals (EWS) for detecting whether the AMOC is approaching a tipping point that could lead to its collapse. These include physics-based EWS (van Westen et al. (2024b)), statistical methods (Ditlevsen and Ditlevsen (2023); Boers (2021)), and, more recently, machine learning (ML)–based method-

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

ologies. A Reservoir Computing framework was proposed in Panahi et al. (2024) to assess the approach to an AMOC tipping point in models of varying complexity, including CESM, and from observations. ML methods have also been recently applied to related tasks, such as predicting short-term AMOC variability from observations (Zhai et al. (2024)), and estimating long-term AMOC variability in CESM using symbolic regression combined with different ML architectures (Wu et al. (2025)).

Here, we propose a Convolutional Neural Network (CNN) based method to estimate the distance to the AMOC tipping point, here indicated by the normalized quantity d_F. Such CNN approaches have previously been applied successfully to climate-related problems such as El Niño prediction (Ham et al. (2019)) and AMOC reconstruction from historical SST and SSS data (Michel et al. (2025)). Our CNN was first evaluated using freshwater-forced AMOC collapse simulations from the EMIC CLIMBER-X, generated under different rates of freshwater forcing. We tested several input configurations: SST fields over the Atlantic, SSS fields, their combination, and the full-depth salinity section at 35°S (S_z^{35S}). While the CNN demonstrates strong predictive performance across all input configurations and forcing rates, it is nevertheless outperformed by a simple ordinary least squares linear regressor (LR). This indicates that, when working solely with CLIMBER-X data, using a more complex CNN model does not provide a clear advantage over a simpler linear approach.

To evaluate the generalization capability of the CNN, we train it exclusively on CLIMBER-X simulations, validate it on one-third of the CESM AMOC collapse simulation from van Westen et al. (2024b), and test it on the remaining portion. The CNN achieves strong generalization performance on the more complex CESM data when using the SSS, SST+SSS, and S_z^{35S} input configurations, whereas the LR model fails completely when tested on the same generalization task. This demonstrates that a more complex CNN architecture is required for this type of cross-model generalization and is capable of predicting the distance to the AMOC tipping point in a fully coupled Earth System Model, even when trained solely on data from a less computationally demanding intermediate-complexity model.

Finally, we evaluated the physical consistency of both the CNN and the LR models using the Shapley Values (SHAP) methodology, an explainable AI technique. When applied to both CLIMBER-X and CESM data, the CNN consistently focuses on spatial patterns that exhibit significant temporal variability as the system approaches the tipping point and clearly identifies large-scale spatial structures associated with AMOC destabilization. This indicates that the CNN is capturing the effects of physically relevant processes.

In contrast, when applied to the CLIMBER-X data, the LR model produces relevance maps that appear more scattered and difficult to interpret, without clearly identifying coherent and stable spatial patterns associated with AMOC weakening. Instead, the LR tends to assign importance to individual grid points rather than capturing larger-scale spatial structures. Because these contributions are tied to specific geographical locations, the LR model fails to generalize to the more complex CESM simulation.

Our results demonstrate that the proposed CNN framework is a valuable tool for guiding the design of freshwater forcing experiments aimed at studying AMOC stability in complex Earth System Models. These simulations are computationally expensive and time-demanding, limiting the amount of available data. In contrast, EMICs like CLIMBER-X are less computationally demanding, enabling the generation of larger synthetic datasets. Our CNN approach leveraged this advantage by being trained on CLIMBER-X data and then evaluated directly on CESM data.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Although our current setup cannot yet be applied directly to real observations, as we focused on the case of linearly increasing freshwater forcing in the North Atlantic without incorporating the effects of anthropogenic forcing, it demonstrates strong potential. Future developments may extend our method by designing more sophisticated distance-to-tipping indices that account for anthropogenic drivers and adopting more advanced ML architectures. Another limitation for real-world applications is that generalizing directly from EMICs to real observations to determine d_F is challenging. However, previous studies (**?**, Michel et al. (2025)) have shown that generalization from Earth System Models based synthetic data to real-world observations is achievable. This suggests that, with the future availability of more data from collapse experiments in Earth System Models, the framework could be extended and adapted to observations.

Code and data availability. The code and data used in this study are available at Guardamagna (2026)

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Appendix A: Optimal Hyperparameter Configurations

Experiment	Hyperparameters explored during validation
CLIMBER-X data only	lr = [0.01, 0.001, 0.0001]
	bs = [64, 128]
	n_{es} = [100, 200, 500, 1000]
CESM generalization	lr = [0.01, 0.005, 0.001, 0.0005, 0.0001]
	bs = [64, 128, 256, 512]
	n_{ep} = [5, 10, 50, 100, 200]

Table A1. Hyperparameter values explored during validation for the two experimental setups considered in this study. Here, lr denotes the learning rate, bs the batch size, n_{es} the early stopping patience (i.e., the number of epochs without improvement before training is stopped), and n_{ep} the maximum number of training epochs.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

r_F test	Input Variables	Optimal Hyperparameters
10^{-5}	SST	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
10^{-5}	SSS	$lr = 10^{-3}$, $bs = 64$, $n_{es} = 10^3$
10^{-5}	SST + SSS	$lr = 10^{-3}$, $bs = 64$, $n_{es} = 5 \times 10^2$
10^{-5}	S_z^{35S}	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
10^{-4}	SST	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 5 \times 10^2$
10^{-4}	SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
10^{-4}	SST + SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
10^{-4}	S_z^{35S}	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
2×10^{-4}	SST	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
2×10^{-4}	SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
2×10^{-4}	SST + SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
2×10^{-4}	S_z^{35S}	$lr = 10^{-4}$, $bs = 128$, $n_{es} = 5 \times 10^2$
3×10^{-4}	SST	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 5 \times 10^2$
3×10^{-4}	SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
3×10^{-4}	SST + SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
3×10^{-4}	S_z^{35S}	$lr = 10^{-3}$, $bs = 64$, $n_{es} = 5 \times 10^2$
4×10^{-4}	SST	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 5 \times 10^2$
4×10^{-4}	SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
4×10^{-4}	SST + SSS	$lr = 10^{-4}$, $bs = 128$, $n_{es} = 10^3$
4×10^{-4}	S_z^{35S}	$lr = 10^{-3}$, $bs = 64$, $n_{es} = 10^3$
5×10^{-4}	SST	$lr = 10^{-4}$, $bs = 128$, $n_{es} = 10^3$
5×10^{-4}	SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
5×10^{-4}	SST + SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
5×10^{-4}	S_z^{35S}	$lr = 10^{-3}$, $bs = 64$, $n_{es} = 10^3$
6×10^{-4}	SST	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 5 \times 10^2$
6×10^{-4}	SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
6×10^{-4}	SST + SSS	$lr = 10^{-4}$, $bs = 64$, $n_{es} = 10^3$
6×10^{-4}	S_z^{35S}	$lr = 10^{-3}$, $bs = 64$, $n_{es} = 5 \times 10^2$

Table A2. Summary of optimal hyperparameter configurations identified during the CLIMBER-X experiments (see Section 2.4). Each of the seven CLIMBER-X AMOC collapsing runs, generated using different forcing rates ($r_F = [10^{-5}, 10^{-4}, 2\times10^{-4}, 3\times10^{-4}, 4\times10^{-4}, 5\times10^{-4}, 6\times10^{-4}]$ Sv yr^{-1}), was used once as the test set in a leave-one-out strategy. The remaining six runs were used to identify optimal hyperparameters via k-fold cross-validation. This procedure was repeated for each input configuration (see Section 3.1).

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Input Variables	Optimal Hyperparameters
SST	$lr = 5 \times 10^{-3}$, $bs = 128$, $n_{\text{ep}} = 5$
SSS	$lr = 5 \times 10^{-3}$, $bs = 512$, $n_{\text{ep}} = 10$
SST + SSS	$lr = 1 \times 10^{-2}$, $bs = 256$, $n_{\text{ep}} = 10$
S_z^{35S}	$lr = 5 \times 10^{-4}$, $bs = 512$, $n_{\text{ep}} = 5$

Table A3. Summary of the optimal hyperparameter configurations identified during the CESM data generalization experiments for each input variable configuration evaluated in the study (see Section 3.2).

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Input Variables	Hyperparameters
SSS	$lr = 5 \times 10^{-3}$, $bs = 128$, $n_{\text{ep}} = 5$
	$lr = 5 \times 10^{-3}$, $bs = 512$, $n_{\text{ep}} = 50$
	$lr = 5 \times 10^{-3}$, $bs = 64$, $n_{\text{ep}} = 10$
	$lr = 1 \times 10^{-2}$, $bs = 256$, $n_{\text{ep}} = 50$
SST + SSS	$lr = 5 \times 10^{-3}$, $bs = 256$, $n_{\text{ep}} = 10$
	$lr = 5 \times 10^{-3}$, $bs = 512$, $n_{\text{ep}} = 10$
	$lr = 5 \times 10^{-3}$, $bs = 128$, $n_{\text{ep}} = 5$
	$lr = 5 \times 10^{-4}$, $bs = 128$, $n_{\text{ep}} = 200$

Table A4. Additional hyperparameter configurations identified during the CESM data generalization experiments for the SSS and SST+SSS input setups (see Section 3.2).

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

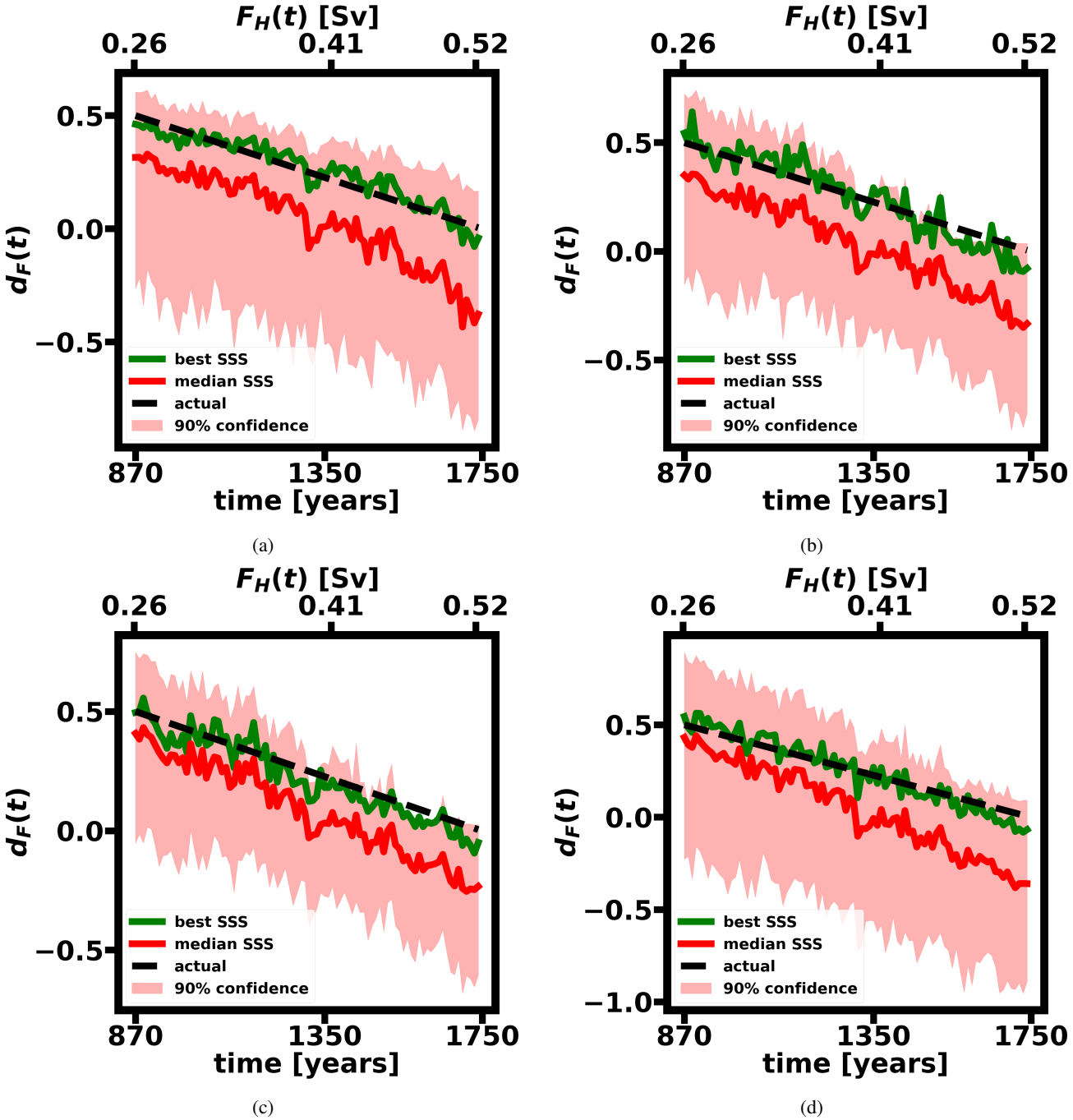


Figure A1. CNN predictions of the distance to tipping versus the true values for the CESM test dataset. Panel from (a) to (d) report the distribution of the predictions obtained training the CNN with the 4 hyperparameter configurations reported in table **??** for the SSS input variable configuration. Shaded areas indicate the 90% confidence interval computed over 500 independent trials. In each trial, the network was randomly reinitialized, retrained using all seven CLIMBER-X simulations (see Section 3.1), and subsequently evaluated on the CESM test dataset. In every case the CNN is trained using solely the SSS fields.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

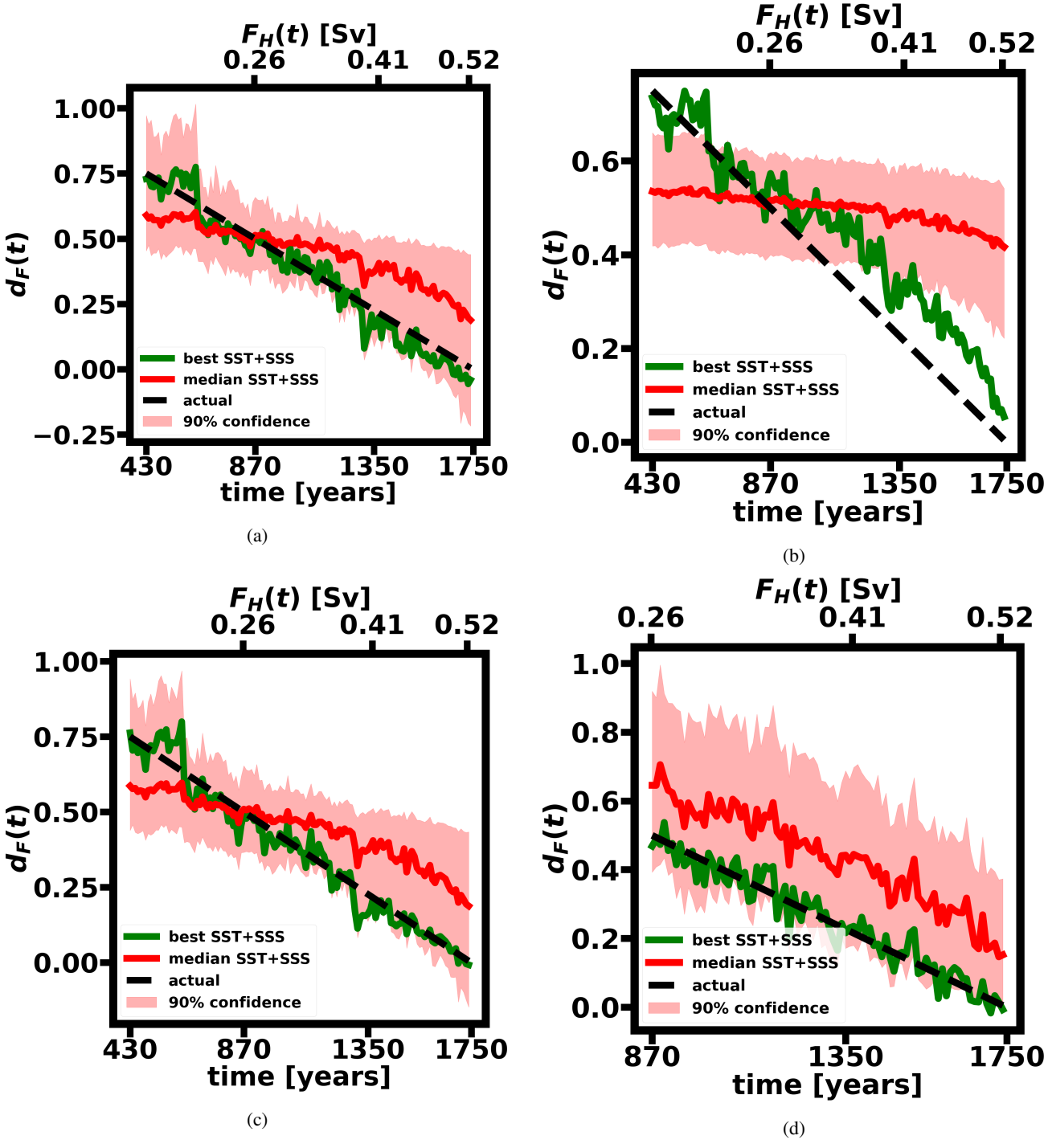


Figure A2. Same as figure A1, but for the CNN trained on SST+SSS fields.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Appendix B: LR model failure on regime shift

As introduced in Section 3.1, when a simple linear regression (LR) model is trained and evaluated exclusively on CLIMBER-X data, it achieves nearly perfect performance. This indicates an almost perfectly linear relationship between the input variables, Atlantic SST and SSS fields together with the full-depth salinity profile at 35°S (S_z^{35S}), and the distance to tipping, quantified by the index $d_F(t)$ (see Section 2.3).

An exception arises when the model is evaluated on the CLIMBER-X simulation forced at a rate of 10^{-4} Sv yr^{-1}. In this case, predictive performance deteriorates sharply in the vicinity of the tipping point (see Figure B1). The degradation is particularly severe when the LR is trained on S_z^{35S} fields. In this configuration, the model becomes completely unreliable close to tipping, exhibiting a dramatic loss of predictive skill as the system approaches collapse. In contrast, when surface variables (SST and SSS) are used as predictors, the LR performance remains overall strong, without a comparable breakdown near the tipping point.

For the slowest freshwater forcing rate considered in this study, 10^{-5} Sv yr^{-1}, the AMOC collapse is abrupt once the tipping point at $0.2189\,Sv$ is crossed. In contrast, for faster forcing rates, starting from 2×10^{-4} Sv yr^{-1} in the simulations analyzed here, the transition toward the OFF state becomes smoother and more gradual, and no abrupt collapse is found. At the intermediate forcing rate of 10^{-4} Sv yr^{-1}, the collapse remains abrupt, but it is initiated before the CLIMBER-X tipping threshold of 0.2189 Sv is reached. At this tipping value, the AMOC strength has already decreased to 7 Sv (see Figure B1). Similar rate-dependent behavior has been reported in different models of varying complexity (e.g., Chapman et al. (2024); Lohmann and Ditlevsen (2021)).

These results indicate that, before reaching the tipping threshold, the system enters a different dynamical regime. In this regime, the LR model fails to generalize appropriately, leading to a marked reduction in predictive skill shortly after the initiation of the AMOC collapse. In contrast, the CNN maintains strong predictive skill in the configuration where the LR fails, when trained on S_z^{35S}, demonstrating substantially greater generalization capabilities.

To further test this hypothesis, we analyze an additional AMOC collapse simulation forced at 5×10^{-5} Sv yr^{-1}. For this forcing rate, the AMOC collapse is initiated even earlier relative to the tipping threshold compared to the 10^{-4} Sv yr^{-1} simulation (see Figure B1). In fact, at $0.2189\,Sv$, the AMOC has already fully collapsed.

To compare model performance, both the CNN and LR were trained on the other seven CLIMBER-X simulations and subsequently tested on the 5×10^{-5} Sv yr^{-1} run. For the CNN, we retained the previously identified optimal hyperparameters for the 10^{-5} Sv yr^{-1} simulation. This choice was motivated by the proximity of the two forcing rates and by the fact that the CNN showed good predictive performance on the 5×10^{-5} Sv yr^{-1} test simulation even without re-validating the hyperparameters.

Under the 5×10^{-5} Sv yr^{-1} forcing, the LR trained on S_z^{35S} exhibits an even more pronounced failure after the initiation of the AMOC collapse than in the 10^{-4} Sv yr^{-1} case, rendering it effectively unreliable. When trained on surface variables (SST, SSS, or $SST + SSS$), the degradation is less severe than for S_z^{35S} but remains evident and more pronounced than in the 10^{-4} Sv yr^{-1} test case.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

By contrast, the CNN maintains substantially more reliable performance when evaluated on the 5×10^{-5} Sv yr^{-1} simulation. When trained on SSS, predictions remain nearly perfect. For the other input configurations, the CNN exhibits a mild underestimation of the distance to tipping d_F in the vicinity of collapse, but the degradation remains significantly smaller than that observed for the LR. Overall, these results confirm the superior generalization capability of the CNN, which successfully captures and extrapolates across the two distinct dynamical regimes of the AMOC system.

To further investigate the reasons for the LR model's failure, we computed the Pearson correlation coefficient between the input variables and the normalized distance to tipping, d_F (see figure B2 and B3). From this analysis, it appears clear that the failure of the LR model after the collapse is not due to the emergence of nonlinear features with respect to the distance index d_F. The input fields are strongly linearly correlated with the index d_F; also, after the onset of the collapse, the type of linear correlation changes.

From Fig. B2 and B3, we observe that the temporal patterns of the SST and SSS fields prior to the onset of the collapse resemble those discussed in Section 4.2. In the North Atlantic, a decrease in SST is observed, corresponding to a positive linear correlation with the decreasing index d_F over time. In contrast, the South Atlantic exhibits an increase in SST, reflected by a negative linear correlation with the decreasing d_F. The SSS fields display a similar spatial pattern.

After the onset of the collapse, the regions of decreasing SST and SSS shift southward, progressively involving the South Atlantic as well. For the SSS fields, this decreasing pattern extends considerably farther south compared to the SST signal. As shown in Fig. B3, once the collapse is reached, the signal affects almost the entire South Atlantic, except for a small region around 30°S extending from the African coast to approximately 37°W, and a limited area near the coast of South America around 20°N.

The southward expansion of the decreasing SST and SSS signals is consistent with the weakening of the AMOC. A reduction in AMOC strength leads to a decreased northward transport of heat and salt; once the collapse is reached, this transport effectively ceases. This alters the basin-scale distribution of heat and salinity, allowing the anomalies associated with the weakening circulation to extend progressively southward.

Furthermore, the collapse of the AMOC is associated with a southward shift of the Intertropical Convergence Zone (ITCZ) (Cerato et al. (2025)). The ITCZ marks the region of maximum tropical convection and precipitation; therefore, its southward displacement leads to increased rainfall over the tropical South Atlantic. The resulting freshwater input reduces surface salinity in this region, contributing to the decrease in SSS observed in the correlation patterns. The southward shift of the ITCZ in response to AMOC weakening has also been reported in previous freshwater-forcing AMOC tipping simulations with the Climber-X model (Willeit et al. (2022b)). Another notable difference between the pre- and post-collapse onset trends is that, following the onset of the collapse, the Arctic exhibits decreasing surface salinity alongside increasing surface temperature.

Looking at S_z^{35S}, we can notice that the pre-collapse patterns resemble again the patterns already reported in section 4.2, with positive salinity anomalies extending down to 1000 m and negative salinity anomalies under 1000 m, after the collapse initiation the negative salinity anomalies extend deeper, reaching 1500 m. The LR model, which captures linear relationships between the input fields and the target variable and is trained primarily on data from the pre-collapse regime, cannot generalize

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

to the post-collapse regime. The CNN, on the other hand, maintains strong predictive performance even after the collapse initiation, demonstrating superior generalization capabilities.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

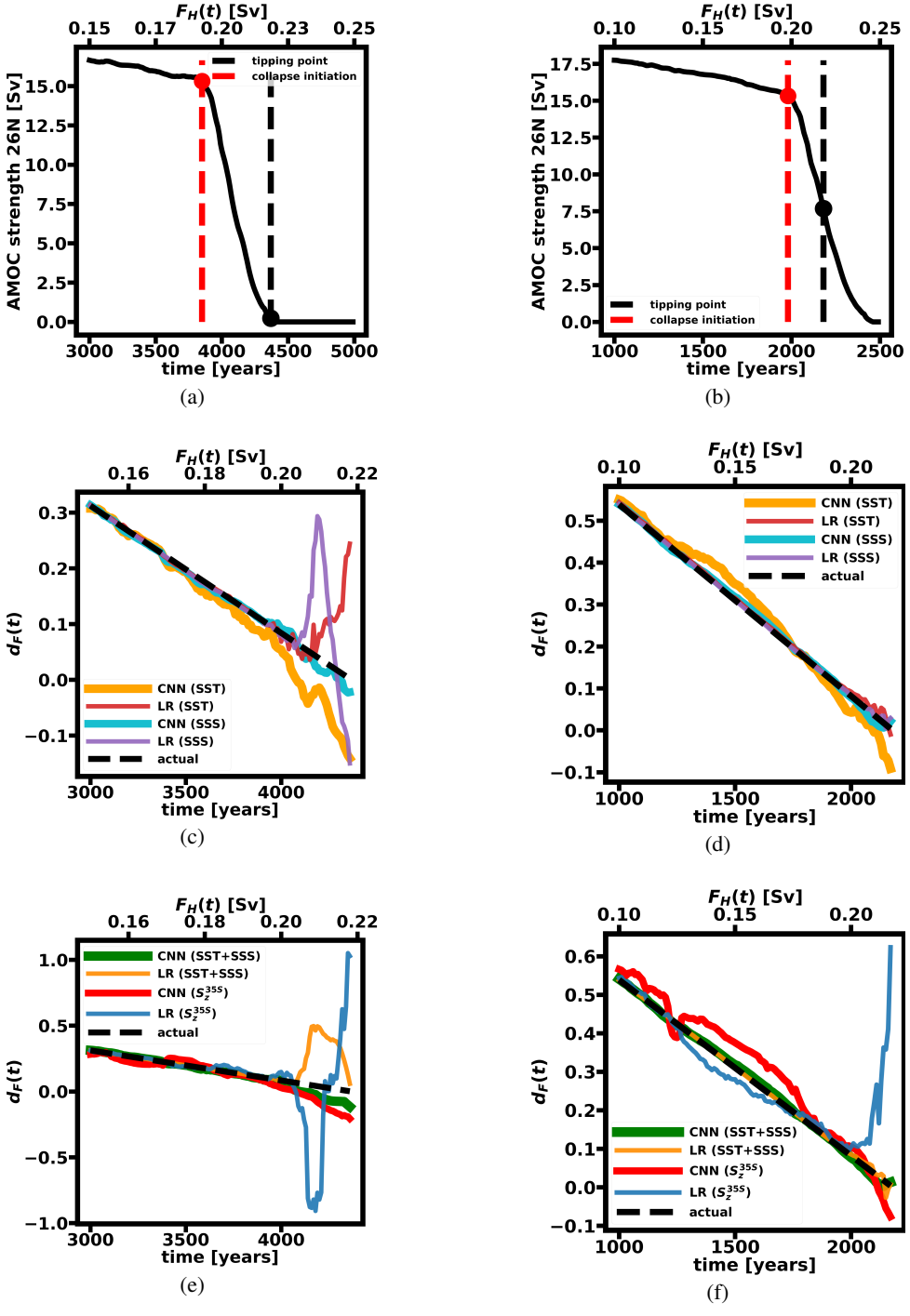
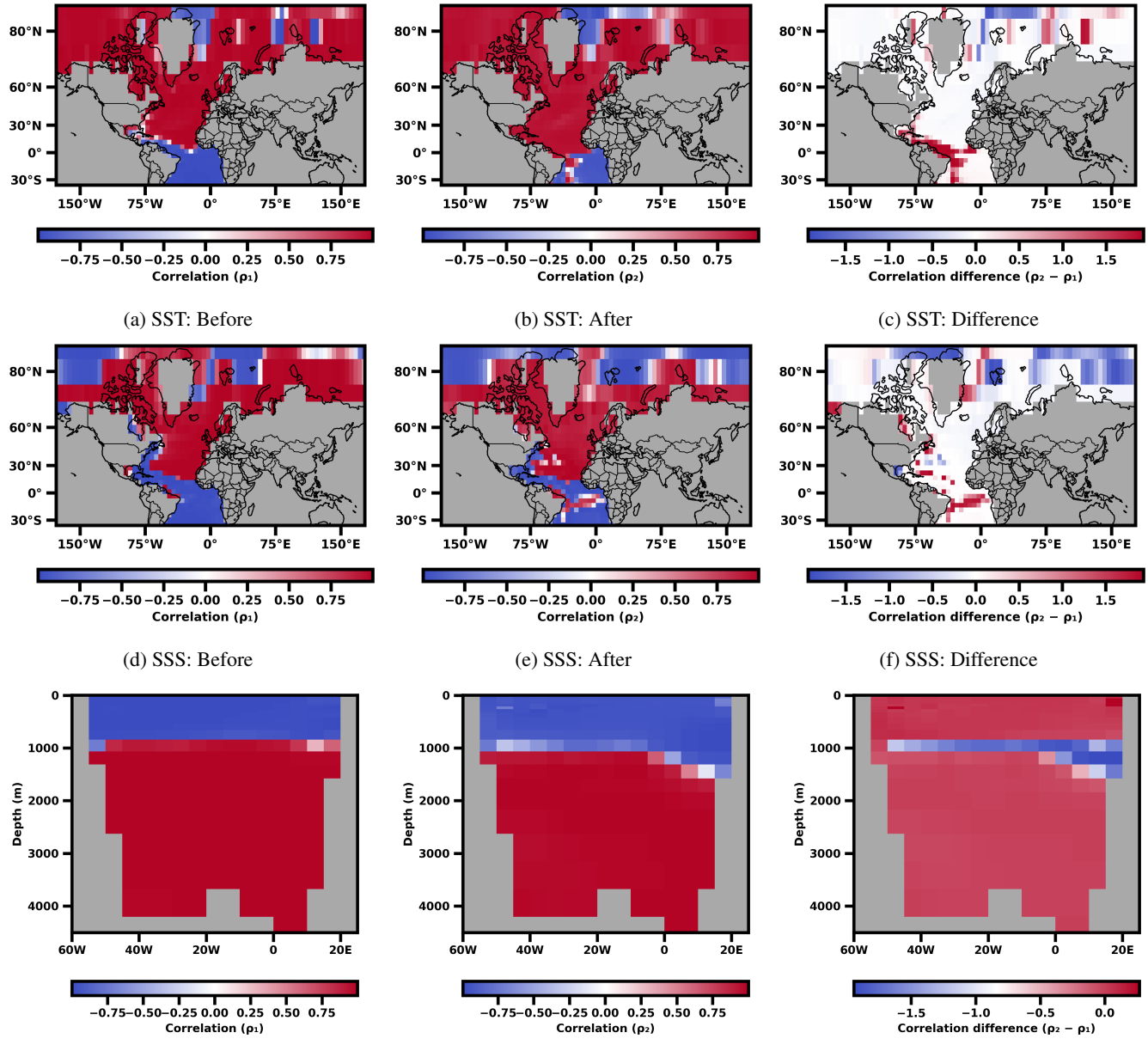


Figure B1. *(a)-(b) Evolution of AMOC strength at 26°N in the CLIMBER-X model, under forcing rates* r_F=[5×10^{-5}, 10^{-4}] *Sv* yr^{-1}, *respectively. (c)-(e) CNN predictions vs actual distances to tipping for* r_F=5×10^{-5} *Sv* yr^{-1}. *The CNN is trained using SST or SSS fields over the Atlantic individually in (c), while in (e), the combination of SST and SSS fields over the Atlantic and* S_z^{35S} *is adopted. (d)-(f) Same as (c)-(e) but for* r_F=10^{-4} *Sv* yr^{-1}. *In all cases, the median predictions out of 20 trained CNN realizations are shown. For each input-variable configuration and for both test simulations, the corresponding predictions from a linear regressor are also included for comparison.*

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.



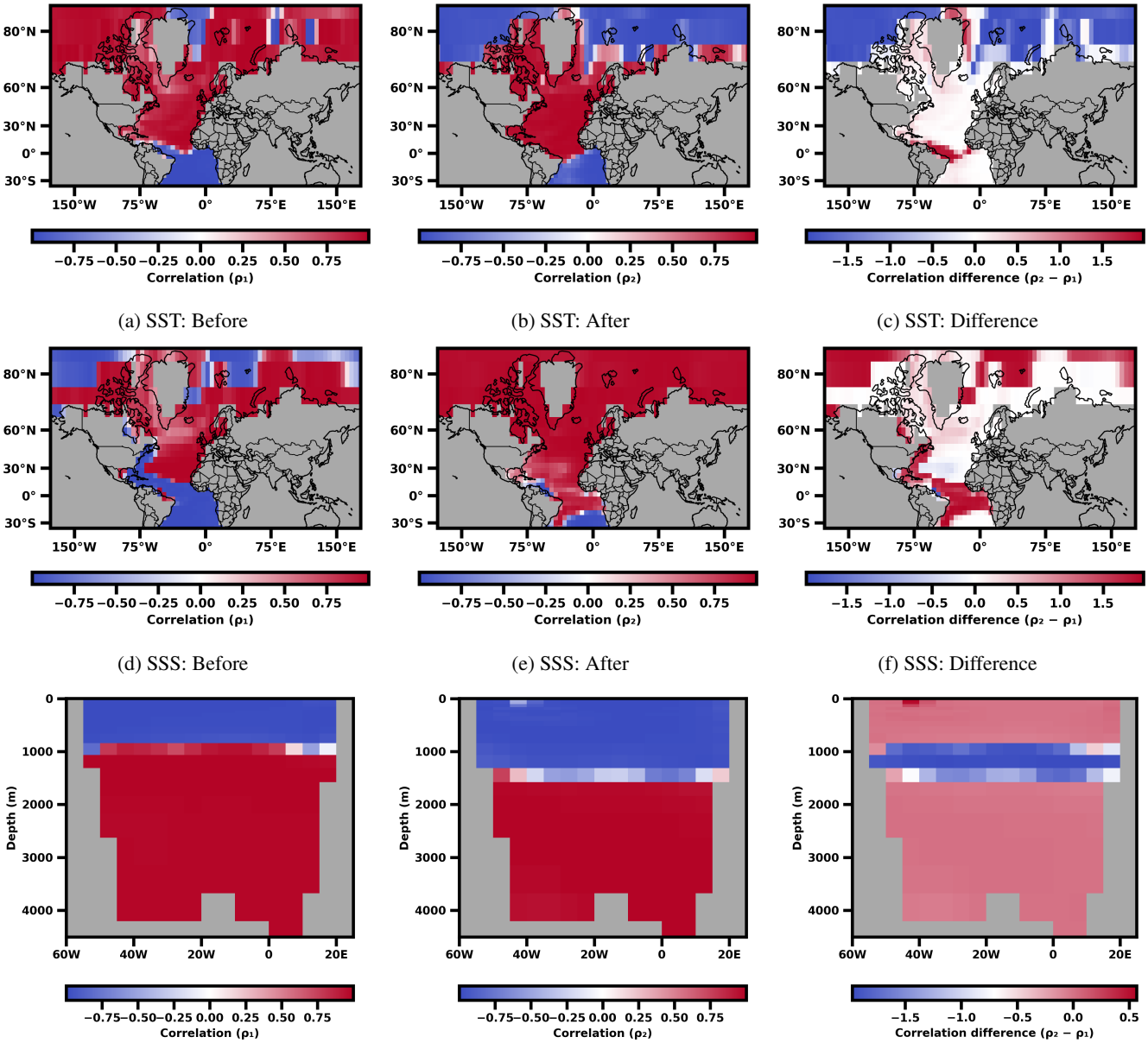
(a) SST: Before (b) SST: After (c) SST: Difference

(d) SSS: Before (e) SSS: After (f) SSS: Difference

(g) S_z^{35S}: Before (h) S_z^{35S}: After (i) S_z^{35S}: Difference

Figure B2. Spatial distribution of the Pearson correlation coefficient between the temporal evolution of SST (a–c), SSS (d–f), and the S_z^{35S} field (g–i) and the normalized distance-to-tipping index d_F. For each variable, the correlation computed before the onset of the AMOC collapse (ρ_1; a, d, g) and after the onset of the collapse (ρ_2; b, e, h) are shown. Panels (c, f, i) report the difference in correlation, $\rho_2 - \rho_1$. All panels correspond to the AMOC-collapsing CLIMBER-X simulation forced at a rate of 10^{-4} Sv yr^{-1}.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.



(a) SST: Before (b) SST: After (c) SST: Difference

(d) SSS: Before (e) SSS: After (f) SSS: Difference

(g) S_z^{35S}: Before (h) S_z^{35S}: After (i) S_z^{35S}: Difference

Figure B3. Same as Fig. B2, but for the AMOC-collapsing Climber-X simulation forced at a rate of 5×10^{-5} Sv yr^{-1}.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Appendix C: CNN performances

C1 Performances on Climber data

r_F (Sv yr^{-1})	Input	MSE CNN (IQR) LR	Error (yr) CNN (IQR) LR	Error (Sv) CNN (IQR) LR
10^{-5}	SST	1.42×10^{-3} (2.77×10^{-4}) 4.50×10^{-4}	824 (79.6) 464	8.24×10^{-3} (7.96×10^{-4}) 4.64×10^{-3}
10^{-5}	S_z^{35S}	1.93×10^{-3} (6.65×10^{-4}) 5.96×10^{-3}	961 (165) 1690	9.61×10^{-3} (1.65×10^{-3}) 1.69×10^{-2}
10^{-5}	SSS	1.79×10^{-4} (1.85×10^{-4}) 5.77×10^{-5}	293 (166) 166	2.93×10^{-3} (1.66×10^{-3}) 1.66×10^{-3}
10^{-5}	SST + SSS	1.63×10^{-3} (3.64×10^{-4}) 1.06×10^{-5}	883 (100) 71.4	8.83×10^{-3} (1.00×10^{-3}) 7.14×10^{-4}
10^{-4}	SST	5.79×10^{-4} (2.87×10^{-4}) 3.58×10^{-5}	52.7 (12.6) 13.1	5.27×10^{-3} (1.26×10^{-3}) 1.31×10^{-3}
10^{-4}	S_z^{35S}	1.59×10^{-3} (3.48×10^{-4}) 4.86×10^{-3}	87.2 (9.77) 152	8.72×10^{-3} (9.77×10^{-4}) 1.53×10^{-2}
10^{-4}	SSS	2.94×10^{-5} (8.04×10^{-6}) 8.02×10^{-6}	11.9 (1.59) 6.17	1.19×10^{-3} (1.59×10^{-4}) 6.20×10^{-4}
10^{-4}	SST + SSS	2.87×10^{-5} (7.15×10^{-6}) 1.44×10^{-5}	11.7 (1.44) 8.26	1.17×10^{-3} (1.44×10^{-4}) 8.30×10^{-4}
2×10^{-4}	SST	1.01×10^{-4} (2.79×10^{-5}) 1.35×10^{-5}	11.0 (1.50) 4.00	2.20×10^{-3} (2.99×10^{-4}) 8.03×10^{-4}
2×10^{-4}	S_z^{35S}	3.42×10^{-4} (1.27×10^{-4}) 3.24×10^{-4}	20.2 (3.68) 19.6	4.05×10^{-3} (7.35×10^{-4}) 3.94×10^{-3}
2×10^{-4}	SSS	1.48×10^{-5} (5.68×10^{-6}) 3.36×10^{-6}	4.2 (0.783) 2.00	8.41×10^{-4} (1.57×10^{-4}) 4.01×10^{-4}
2×10^{-4}	SST + SSS	2.02×10^{-5} (1.88×10^{-5}) 5.82×10^{-6}	4.9 (2.00) 2.63	9.84×10^{-4} (3.99×10^{-4}) 5.28×10^{-4}

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

r_F (Sv yr^{-1})	Input	MSE CNN (IQR) LR	Error (yr) CNN (IQR) LR	Error (Sv) CNN (IQR) LR
3×10^{-4}	SST	1.42×10^{-4} (3.87×10^{-5}) 5.52×10^{-6}	8.69 (1.20) 1.69	2.61×10^{-3} (3.59×10^{-4}) 5.14×10^{-4}
3×10^{-4}	S_z^{35S}	1.72×10^{-4} (6.97×10^{-5}) 1.23×10^{-4}	9.57 (1.96) 7.99	2.87×10^{-3} (5.88×10^{-4}) 2.43×10^{-3}
3×10^{-4}	SSS	1.01×10^{-5} (5.47×10^{-6}) 3.79×10^{-6}	2.32 (0.637) 1.40	6.95×10^{-4} (1.91×10^{-4}) 4.26×10^{-4}
3×10^{-4}	SST + SSS	2.88×10^{-5} (8.31×10^{-5}) 2.37×10^{-6}	3.9 (4.25) 1.11	1.17×10^{-3} (1.28×10^{-3}) 3.37×10^{-4}
4×10^{-4}	SST	8.24×10^{-5} (1.99×10^{-5}) 1.56×10^{-5}	5.0 (0.582) 2.13	1.99×10^{-3} (2.33×10^{-4}) 8.65×10^{-4}
4×10^{-4}	S_z^{35S}	1.16×10^{-4} (4.02×10^{-5}) 7.72×10^{-5}	5.9 (0.966) 4.74	2.36×10^{-3} (3.86×10^{-4}) 1.92×10^{-3}
4×10^{-4}	SSS	7.21×10^{-6} (1.11×10^{-5}) 2.95×10^{-6}	1.47 (0.960) 0.927	5.88×10^{-4} (3.84×10^{-4}) 3.76×10^{-4}
4×10^{-4}	SST + SSS	4.21×10^{-5} (6.12×10^{-5}) 4.96×10^{-6}	3.6 (2.48) 1.20	1.42×10^{-3} (9.91×10^{-4}) 4.87×10^{-4}

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

r_F (Sv yr^{-1})	Input	MSE CNN (IQR) LR	Error (yr) CNN (IQR) LR	Error (Sv) CNN (IQR) LR
5×10^{-4}	SST	6.82×10^{-5} (4.99×10^{-5}) 1.13×10^{-5}	3.6 (1.20) 1.45	1.81×10^{-3} (6.01×10^{-4}) 7.37×10^{-4}
5×10^{-4}	S_z^{35S}	8.01×10^{-5} (4.65×10^{-5}) 8.80×10^{-5}	3.9 (1.06) 4.03	1.96×10^{-3} (5.30×10^{-4}) 2.05×10^{-3}
5×10^{-4}	SSS	8.09×10^{-6} (5.83×10^{-6}) 4.34×10^{-7}	1.24 (0.433) 0.283	6.23×10^{-4} (2.16×10^{-4}) 1.44×10^{-4}
5×10^{-4}	SST + SSS	1.75×10^{-5} (2.09×10^{-5}) 7.65×10^{-7}	1.8 (0.969) 0.376	9.16×10^{-4} (4.85×10^{-4}) 1.91×10^{-4}
6×10^{-4}	SST	9.31×10^{-5} (2.14×10^{-5}) 1.10×10^{-5}	3.5 (0.407) 1.19	2.11×10^{-3} (2.44×10^{-4}) 7.25×10^{-4}
6×10^{-4}	S_z^{35S}	2.05×10^{-4} (8.16×10^{-5}) 1.69×10^{-4}	5.2 (1.01) 4.68	3.13×10^{-3} (6.05×10^{-4}) 2.85×10^{-3}
6×10^{-4}	SSS	1.70×10^{-5} (2.82×10^{-5}) 3.28×10^{-6}	1.5 (1.04) 0.652	9.04×10^{-4} (6.22×10^{-4}) 3.96×10^{-4}
6×10^{-4}	SST + SSS	1.97×10^{-5} (1.40×10^{-5}) 1.43×10^{-6}	1.6 (0.531) 0.430	9.71×10^{-4} (3.19×10^{-4}) 2.61×10^{-4}

Table C1. Summary table reporting the median and interquartile range (IQR) of the MSE distribution, computed between the predictions of 20 different CNN realizations and the actual distances to tipping, for each test run and input variable configuration used in the CLIMBER-X experiments (see Section 3.1). The table also includes the corresponding errors expressed in years and freshwater forcing. Results obtained using a Linear Regressor trained on the same input variables and evaluated on the same test simulations are also reported for comparison.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

C2 Performances on CESM data

Input	Median MSE (IQR)	Median Error (years) (IQR)	Median Error (Sv) (IQR)	Predicted Distance at Final Time Step (Year 1750)
SST only	$0.077\,(0.03)$	487 (100)	0.147 (0.03)	0.48 (year 917)
SSS only	$0.0265\,(0.06)$	286 (252)	0.086 (0.076)	0.1 (year 1584)
SST + SSS	$0.028\,(0.042)$	295 (221)	0.088 (0.067)	0.12 (year 1550)
S_z^{35S}	$0.0334\,(0.03)$	321 (155)	0.1 (0.047)	0.14 (year 1511)

Table C2. Summary table showing the median mean squared error (MSE) between the predicted and actual distances to tipping, computed over 500 independent CNN realizations evaluated on the CESM test dataset (see Section 3.2). The variability in model performance is indicated in parentheses and corresponds to the interquartile range (IQR). For both the median and the IQR, the corresponding uncertainties are also expressed in years and in freshwater forcing.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Appendix D: Shapley Additive exPlanations

Shapley Additive exPlanations (SHAP) (Lundberg and Lee (2017)) is a model-agnostic explanability method that can be applied to any machine learning model, from simple linear models to complex deep neural networks. SHAP can be used to estimate which input features a model relies on to produce a specific output.

The SHAP method is based on Shapley values, a concept from cooperative game theory. In this framework, a group of players jointly produces a single outcome, and the Shapley value of each player quantifies its average contribution to that outcome. In the context of machine learning, the players correspond to the input features of a given sample, while the outcome corresponds to the model's prediction.

Given an input sample x with m features, the Shapley value ϕ_i associated with the i^{th} feature x_i is defined as the average marginal contribution of that feature to the model output:

$$\phi_i = \sum_{S \subseteq F \setminus i} \frac{|S|!, (m - |S| - 1)!}{m!} \left[f(S \cup i) - f(S) \right], \tag{D1}$$

where S is a subset of features that does not include feature x_i, F is the set of all input features, and $f(\cdot)$ denotes the model being analyzed.

In practice, computing the model output when a feature is missing requires an approximation. This is done using a background dataset $X_{\text{background}}$, which is a subset of the dataset X. When a feature i is not included, its value in the input sample is replaced with a value randomly drawn from $X_{\text{background}}$, while the other feature values are kept unchanged. This procedure is repeated multiple times, and the average model output is used as an estimate. In the SHAP framework, the average prediction over the background dataset is used as a reference value. Feature contributions are computed as deviations from this average prediction.

Due to its interpretability and robustness, SHAP has been used in previous ocean science studies involving convolutional neural networks (CNNs) to better understand model behavior. For example, Smith et al. (2023) used SHAP to identify which surface variables are most important for a CNN reconstructing subsurface ocean temperature and salinity profiles. Similarly, Park et al. (2025) and Yao et al. (2024) applied SHAP to identify the main drivers of CNN-based predictions of surface chlorophyll concentration and the Pacific Decadal Oscillation, respectively.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Author contributions. All authors contributed to the conceptual design of the study. Sacha Sinet provided support for generating the CLIMBER-X data. Francesco Guardamagna developed the code for the machine learning architecture and carried out the data analysis and interpretation of the results. Henk A. Dijkstra supervised the project. The manuscript was written by Francesco Guardamagna, with all co-authors providing feedback and contributing to revisions throughout the writing process.

Competing interests. The authors declare that they have no conflicts of interest.

Acknowledgements. The work of Francesco Guardamagna was supported by the Netherlands Organization for Scientific Research (NWO) under Grant OCENW.M20.277. The work of Henk Dijkstra was supported by the European Research Council through the ERC-AdG project TAOC (PI: Dijkstra, project 101055096).

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

References

Ben-Yami, M., Morr, A., Bathiany, S., and Boers, N.: Uncertainties too large to predict tipping times of major Earth system components from historical data, Science Advances, 10, eadl4841, https://doi.org/10.1126/sciadv.adl4841, 2024.

Boers, N.: Observation-based early-warning signals for a collapse of the Atlantic Meridional Overturning Circulation, Nature Climate Change, 11, 680–688, 2021.

Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G., and Saba, V.: Observed fingerprint of a weakening Atlantic Ocean overturning circulation, Nature, 556, 191–196, 2018.

Cerato, G., Bellomo, K., D'Agostino, R., and von Hardenberg, J.: Multi-model evidence of future tropical Atlantic precipitation change modulated by AMOC decline, Journal of Climate, https://doi.org/10.1175/JCLI-D-24-0333.1, 2025.

Chapman, R., Sinet, S., and Ritchie, P. D. L.: Tipping mechanisms in a conceptual model of the Atlantic Meridional Overturning Circulation, Weather, 79, 316–323, https://doi.org/https://doi.org/10.1002/wea.7609, 2024.

Ditlevsen, P. and Ditlevsen, S.: Warning of a forthcoming collapse of the Atlantic meridional overturning circulation, Nature Communications, 14, 4254, 2023.

Goodfellow, I.: Deep learning, vol. 196, MIT press, 2016.

Guardamagna, F.: Code for reproducing the results of the paper "Predicting the Distance of the AMOC to Its Tipping Point Using CNNs", https://doi.org/10.5281/zenodo.19369578, 2026.

Ham, Y.-G., Kim, J.-H., and Luo, J.-J.: Deep learning for multi-year ENSO forecasts, Nature, 573, 568–572, https://doi.org/10.1038/s41586-019-1559-7, 2019.

Hawkins, E., Smith, R. S., Allison, L. C., Gregory, J. M., Woollings, T. J., Pohlmann, H., and de Cuevas, B.: Bistability of the Atlantic overturning circulation in a global climate model and links to ocean freshwater transport, Geophysical Research Letters, 38, https://doi.org/https://doi.org/10.1029/2011GL047208, 2011.

Hofmann, M. and Rahmstorf, S.: On the stability of the Atlantic meridional overturning circulation, Proceedings of the National Academy of Sciences, 106, 20 584–20 589, https://doi.org/10.1073/pnas.0909146106, 2009.

Hunke, E. C., Lipscomb, W. H., Turner, A. K., Jeffery, N., and Elliott, S.: CICE: The Los Alamos Sea ice model documentation and software user's manual version 5.1 LA-CC-06-012, T-3 Fluid Dynamics Group, Los Alamos National Laboratory, 675, 15, 2015.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model: A Framework for Collaborative Research, Bulletin of the American Meteorological Society, 94, 1339 – 1360, https://doi.org/https://doi.org/10.1175/BAMS-D-12-00121.1, 2013.

Johns, W. E., Baringer, M. O., Beal, L. M., Cunningham, S. A., Kanzow, T., Bryden, H. L., Hirschi, J. J. M., Marotzke, J., Meinen, C. S., Shaw, B., and Curry, R.: Continuous, Array-Based Estimates of Atlantic Ocean Heat Transport at 26.5°N, Journal of Climate, 24, 2429 – 2449, https://doi.org/10.1175/2010JCLI3997.1, 2011.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86, 2278–2324, https://doi.org/10.1109/5.726791, 1998.

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J.: Tipping elements in the Earth's climate system, Proceedings of the National Academy of Sciences, 105, 1786–1793, https://doi.org/10.1073/pnas.0705414105, 2008.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Liu, W., Xie, S.-P., Liu, Z., and Zhu, J.: Overlooked possibility of a collapsed Atlantic Meridional Overturning Circulation in warming climate, Science Advances, 3, e1601 666, https://doi.org/10.1126/sciadv.1601666, 2017.

Lohmann, J. and Ditlevsen, P. D.: Risk of tipping the overturning circulation due to increasing rates of ice melt, Proceedings of the National Academy of Sciences, 118, e2017989 118, https://doi.org/10.1073/pnas.2017989118, 2021.

Lundberg, S. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, https://arxiv.org/abs/1705.07874, 2017.

Michel, S. L. L., Dijkstra, H. A., Guardamagna, F., Jacques-Dumas, V., van Westen, R. M., and von der Heydt, A. S.: Deep learning based reconstructions of the Atlantic meridional overturning circulation confirm twenty-first century decline, Environmental Research Letters, 20, 064 036, https://doi.org/10.1088/1748-9326/add7f0, publisher: IOP Publishing, 2025.

Neale, R. B., Richter, J., Park, S., Lauritzen, P. H., Vavrus, S. J., Rasch, P. J., and Zhang, M.: The mean climate of the Community Atmosphere Model (CAM4) in forced SST and fully coupled experiments, Journal of Climate, 26, 5150–5168, 2013.

Panahi, S., Kong, L.-W., Moradi, M., Zhai, Z.-M., Glaz, B., Haile, M., and Lai, Y.-C.: Machine learning prediction of tipping in complex dynamical systems, Phys. Rev. Res., 6, 043 194, https://doi.org/10.1103/PhysRevResearch.6.043194, 2024.

Park, J.-S., Park, J.-Y., Ham, Y.-G., Kim, J.-H., and Jeon, W. J.: A Deep Learning Framework for Chlorophyll Prediction in Large Marine Ecosystems: Benchmarking with a Dynamic Model and Implications for Fish Catch Forecasts, EGUsphere, 2025, 1–19, https://doi.org/10.5194/egusphere-2025-5673, 2025.

Rahmstorf, S.: Is the Atlantic Overturning Circulation Approaching a Tipping Point?, Oceanography, https://doi.org/10.5670/oceanog.2024.501, 2024.

Sinet, S., Ashwin, P., von der Heydt, A. S., and Dijkstra, H. A.: AMOC stability amid tipping ice sheets: the crucial role of rate and noise, Earth System Dynamics, 15, 859–873, https://doi.org/10.5194/esd-15-859-2024, 2024.

Smith, P. A. H., Sørensen, K. A., Buongiorno Nardelli, B., Chauhan, A., Christensen, A., St. John, M., Rodrigues, F., and Mariani, P.: Reconstruction of subsurface ocean state variables using Convolutional Neural Networks with combined satellite and in situ data, Frontiers in Marine Science, Volume 10 - 2023, https://doi.org/10.3389/fmars.2023.1218514, 2023.

Smith, R., Jones, P., Briegleb, B., Bryan, F., Danabasoglu, G., Dennis, J., Dukowicz, J., Eden, C., Fox-Kemper, B., Gent, P., et al.: The parallel ocean program (POP) reference manual ocean component of the community climate system model (CCSM) and community earth system model (CESM), LAUR-01853, 141, 1–140, 2010.

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

Stouffer, R. J., Yin, J., Gregory, J. M., Dixon, K. W., Spelman, M. J., Hurlin, W., Weaver, A. J., Eby, M., Flato, G. M., Hasumi, H., Hu, A., Jungclaus, J. H., Kamenkovich, I. V., Levermann, A., Montoya, M., Murakami, S., Nawrath, S., Oka, A., Peltier, W. R., Robitaille, D. Y., Sokolov, A., Vettoretti, G., and Weber, S. L.: Investigating the Causes of the Response of the Thermohaline Circulation to Past and Future Climate Changes, Journal of Climate, 19, 1365 – 1387, https://doi.org/10.1175/JCLI3689.1, 2006.

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

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

Weijer, W., Cheng, W., Drijfhout, S. S., Fedorov, A. V., Hu, A., Jackson, L. C., Liu, W., McDonagh, E. L., Mecking, J. V., and Zhang, J.: Stability of the Atlantic Meridional Overturning Circulation: A Review and Synthesis, Journal of Geophysical Research: Oceans, 124, 5336–5375, https://doi.org/https://doi.org/10.1029/2019JC015083, 2019.

https://doi.org/10.5194/egusphere-2026-1872
Preprint. Discussion started: 10 April 2026
© Author(s) 2026. CC BY 4.0 License.

Willeit, M. and Ganopolski, A.: Generalized Stability Landscape Of The Atlantic Meridional Overturning Circulation, Earth System Dynamics, 15, 1417–1434, https://doi.org/10.5194/esd-15-1417-2024, 2024.

Willeit, M., Ganopolski, A., Robinson, A., and Edwards, N. R.: The Earth system model CLIMBER-X v1.0 – Part 1: Climate model description and validation, Geoscientific Model Development, 15, 5905–5948, https://doi.org/10.5194/gmd-15-5905-2022, 2022a.

Willeit, M., Ganopolski, A., Robinson, A., and Edwards, N. R.: The Earth system model CLIMBER-X v1.0 – Part 1: Climate model description and validation, Geoscientific Model Development, 15, 5905–5948, https://doi.org/10.5194/gmd-15-5905-2022, 2022b.

Wu, Q.-F., Jochum, M., Avery, J. E., Vettoretti, G., and Nuterman, R.: Machine Guided Derivation of the Atlantic Meridional Overturning Circulation (AMOC) Strength, Geophysical Research Letters, 52, e2024GL113 454, https://doi.org/https://doi.org/10.1029/2024GL113454, e2024GL113454 2024GL113454, 2025.

Wunderling, N., von der Heydt, A. S., Aksenov, Y., Barker, S., Bastiaansen, R., Brovkin, V., Brunetti, M., Couplet, V., Kleinen, T., Lear, C. H., Lohmann, J., Roman-Cuesta, R. M., Sinet, S., Swingedouw, D., Winkelmann, R., Anand, P., Barichivich, J., Bathiany, S., Baudena, M., Bruun, J. T., Chiessi, C. M., Coxall, H. K., Docquier, D., Donges, J. F., Falkena, S. K. J., Klose, A. K., Obura, D., Rocha, J., Rynders, S., Steinert, N. J., and Willeit, M.: Climate tipping point interactions and cascades: a review, Earth System Dynamics, 15, 41–74, https://doi.org/10.5194/esd-15-41-2024, 2024.

Yao, Z., Xu, D., Wang, J., Ren, J., Yu, Z., Yang, C., Xu, M., Wang, H., and Tan, X.: Predicting and Understanding the Pacific Decadal Oscillation Using Machine Learning, Remote Sensing, 16, https://doi.org/10.3390/rs16132261, 2024.

Zhai, Z.-M., Moradi, M., Panahi, S., Wang, Z.-H., and Lai, Y.-C.: Machine-learning nowcasting of the Atlantic Meridional Overturning Circulation, APL Machine Learning, 2, 036 103, https://doi.org/10.1063/5.0207539, _eprint: https://pubs.aip.org/aip/aml/article-pdf/doi/10.1063/5.0207539/20082574/036103_1_5.0207539.pdf, 2024.