Geophysical Research Letters

RESEARCH LETTER
10.1029/2024GL113454

Machine Guided Derivation of the Atlantic Meridional Overturning Circulation (AMOC) Strength

Qi-Fan Wu[1,2], Markus Jochum[1], James E. Avery[3,4], Guido Vettoretti[1], and Roman Nuterman[1]

[1]Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark, [2]Department of Atmospheric Sciences, University of Helsinki, Helsinki, Finland, [3]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark, [4]Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark

Key Points:

- AMOC strength in a climate model is predicted with neural networks
- A causal framework is introduced to determine the key parameters
- A functional form of AMOC strength is then determined with symbolic regression

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Q.-F. Wu,
gtj420@alumni.ku.dk

Citation:

Wu, Q.-F., Jochum, M., Avery, J. E., Vettoretti, G., & Nuterman, R. (2025). Machine guided derivation of the Atlantic Meridional Overturning Circulation (AMOC) strength. *Geophysical Research Letters*, *52*, e2024GL113454. https://doi.org/10.1029/2024GL113454

Received 5 NOV 2024
Accepted 15 JAN 2025

Abstract A machine learning based methodology is developed to determine the strength of the Atlantic Meridional Overturning Circulation (AMOC) in the Community Earth System Model (CESM). Neural networks capture relationships between various climate variables and AMOC. We then identify which of the various are the most important to control the AMOC, and then perform symbolic regression to transform complex interactions into a simple closed-form approximation. A sensitivity analysis for this equation reveals that surface freshwater flux and potential density at 200 m depth are the main controls of the AMOC.

Plain Language Summary We create a method to predict AMOC strength by learning from data sets generated by the CESM using data-driven approaches. By training neural networks, we identified important variables that affect AMOC, particularly surface freshwater flux and potential density at 200 m depth. Using symbolic regression, we simplified complex relationships into a readable formula that matches traditional theories of the AMOC. This approach allows for a more transparent understanding of AMOC's dynamics and opens up a lot of possibilities of using machine learning in climate studies.

1. Introduction

The Atlantic Meridional Overturning Circulation (AMOC, the zonally integrated mean flow in the Atlantic ocean) brings heat from the tropics to the high-latitudes and ensures a relatively mild European climate (Kuhlbrodt et al., 2007). Atlantic Meridional Overturning Circulation instability is associated with dramatic changes in climate, as evidenced by the Dansgaard-Oeschger (D-O) events during the last glacial period, which were marked by sudden shifts in Greenland temperatures between warm interstadial and cold stadial conditions (Dansgaard et al., 1993; Henry et al., 2016).

There are various hypotheses about the mechanisms that drive AMOC changes, which can be split into changes to stratification and changes to surface fluxes. In the latter camp fall Kleppin et al. (2015) and Vettoretti and Peltier (2018), who show that transitions can be initiated by North Atlantic Oscillation (NAO) extrema or Arctic sea-ice export and melt, respectively. The importance of stratification for AMOC strength is highlighted by Brix and Gerdes (2003) and van Westen et al. (2024), who show that the cross-equatorial transport of Antarctic Bottom Water (AABW) or freshwater, respectively, can weaken the AMOC. In particular, Vettoretti and Peltier (2018) show that during stadials heat builds up underneath the sea-ice of the North Atlantic and is discharged in sudden large-scale polynyas. All these studies are based on numerical simulations; the transitions happen in less than a decade and proxy data from ice or sediment cores has still too much uncertainty to allow unambiguous identification of cause and effect (Capron et al., 2021; Erhardt et al., 2019). Unfortunately, ambiguity is still present in the analysis of numerical simulations. This is because standard determination of cause and effect is based on leads and lags, and the transition point in a noisy system is difficult to determine with absolute confidence (A. Slattery et al., 2024). The studies of Kleppin et al. (2015) and Vettoretti and Peltier (2018) are an exception because they rely on the painstaking analysis of numerous 4-dimensional fields to arrive at the anatomy of a single transition in one particular model. This is something for which one is unlikely to ever have enough observations.

To make progress, we use the concept of Granger Causality and determine whether one time series is useful in forecasting another: if the prediction of Y(t) does not get worse by removing X(t) in the prediction process, then X (t) does not cause Y(t) (Granger, 1969). The challenge then in the present context is to develop a model to predict

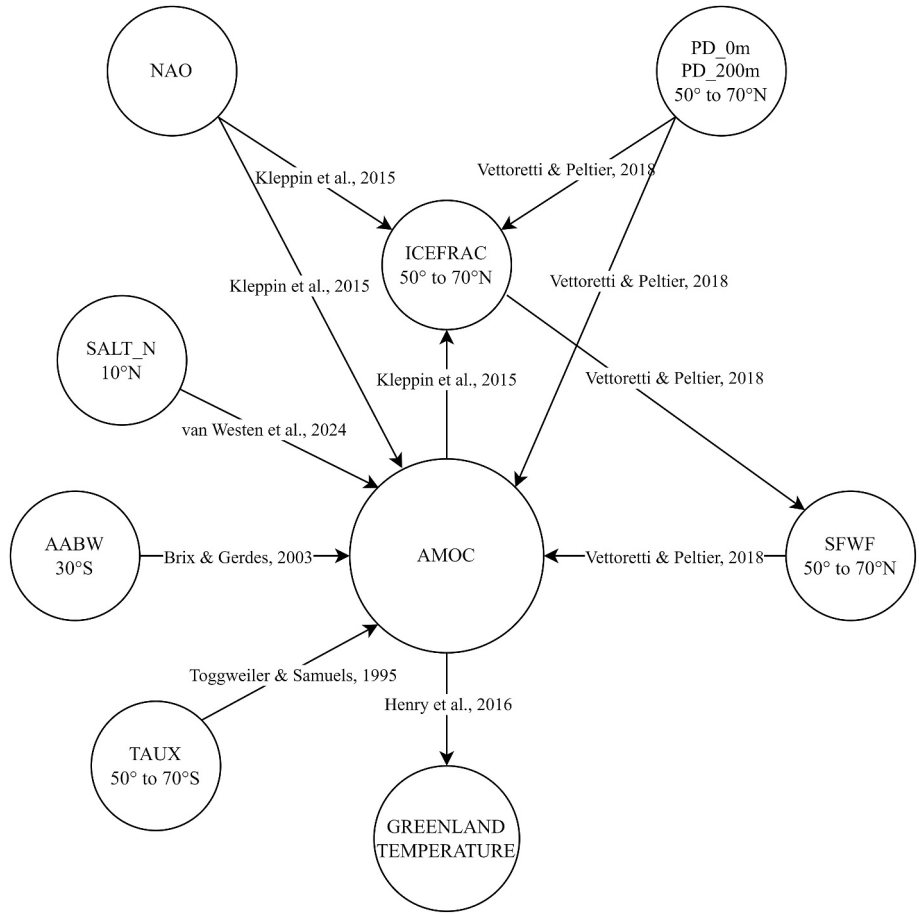

Figure 1. Causal graph illustrating the hypothesized mechanisms influencing the AMOC and its impact on Greenland temperatures during the last glaciation. Key variables include the North Atlantic Oscillation (NAO), salt transport at 10°N (SALT_N), Antarctic Bottom Water (AABW), sea ice fraction (ICEFRAC), windstress in grid-x direction (TAUX), surface freshwater flux (FW) inversely mirrored by the virtual salt flux in FW Flux formulation (SFWF), and potential density at 0 and 200 m depths (PD_0 m, PD_200 m) in the North Atlantic. Arrows indicate the dominant direction of influence, with citations referencing the studies supporting each relationship. Abbreviations for variable names are the original CESM variable names (Kay et al., 2015).

the AMOC. If one can build a successful prediction model, then causality can be simply established by its repeated application.

Recent years have witnessed the success of Neural Networks (NN) to predict Earth system dynamics by capturing non-linear relationships among numerous climate variables. Input variables (e.g., wind (x,y,t) or temperature (x,y, z,t) at a particular location) are connected to a future output variable (e.g., AMOC(y,t + τ)) through one or several hidden layers, and the way the nodes in the layers are connected are learned from available observations or model results. This can be surprisingly effective, although the "black-box" nature rarely enables scientific understanding (Irrgang et al., 2021). Thus, we employ yet another machine learning tool: Symbolic Regression (SR). After using NNs to determine possible causes of changes to AMOC, we will use SR to derive human-readable equations from the complex patterns learned by Neural Networks, making these predictions more interpretable (Murari et al., 2023). Thus, we will follow the path recently described by Häfner et al. (2023).

1. Develop a causal graph based on published hypotheses (Figure 1)
2. Reject hypotheses with the help of NNs and Granger Causality
3. Use parsimony to determine a minimum set of parameters to avoid overfitting
4. Use SR to determine AMOC strength as a function of climate states

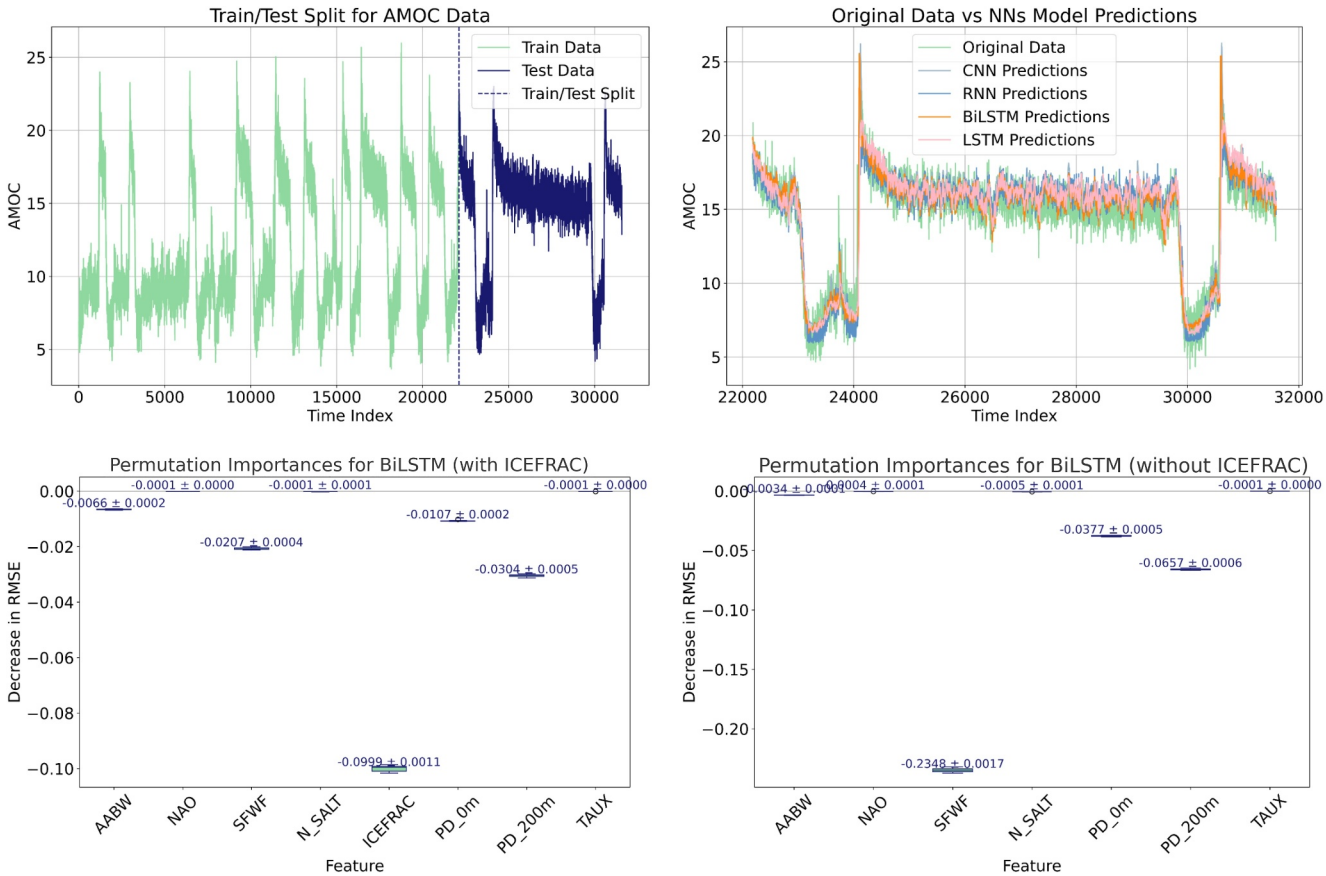

Figure 2. The top left panel shows the combined AMOC strength data set split into training and testing sets. The training set covers 70% of the data, and the testing set covers the remaining 30%. The top right panel compares the NN's predictive performance. The bottom panels display the permutation importance of the BiLSTM model's features, with the left panel including the feature ICEFRAC and the right panel excluding them.

The present study is based on long integrations of Community Earth System Model (CESM) with glacial boundary conditions that reproduce the observed structure of D-O events (Jochum et al., 2022; Vettoretti et al., 2022a). Future work will try to attempt to extend this methodology to the ice and sediment core record, but for the sake of brevity we will for the remainder of this study use "data" when we refer to the output of the numerical simulations.

This paper is organized as follows: In Section 2, we describe the data sets used in this study and outline the methodology, including the neural network architectures and symbolic regression techniques employed to analyze AMOC dynamics. In Section 3, we present the results, detailing the performance of the NN models and equations derived from the analysis. In Section 4, the main results are summarized, and the potential directions for future research are suggested.

2. Data and Methodology

2.1. Data

This study is based on annual means of four 8,000 years long integrations with the CESM. The data contain 13 D-O events and are described in detail in Vettoretti et al. (2022a). From the output we extract time series of 9 variables that are discussed in the literature as affecting AMOC strength or even cause AMOC collapse (Figure 1). They were combined into a unified data set, which was then split into training and testing sets for model development, with the training set covering 70% of the data and the testing set covering the remaining 30% (Figure 2).

Our prediction target is the maximum AMOC strength at 40°N. The minimum MOC strength at 30°S latitude is chosen to monitor AABW's influence on global ocean circulation (Brix & Gerdes, 2003). For salt transport (van Westen et al., 2024), the 10°N latitude is selected due to its position south of the subtropical gyre, without being influenced by the zonal currents and intertropical convergence zone dynamics that dominate at the equator (Treguier et al., 2014). The NAO index is calculated by determining the difference in sea level pressure between Icelandic Low and the Azores High (Kleppin et al., 2015). Previous study suggested that modifying wind stress in the Southern Ocean could impact the AMOC, potentially linking stronger Southern Hemisphere winds to increased deep water formation in the North Atlantic (Toggweiler & Samuels, 1995), and lastly, Vettoretti and Peltier (2018) find that AMOC transitions are triggered by surface freshwater anomalies or subsurface density anomalies.

2.2. Applying Neural Networks

For the NN prediction model, several choices need to be made (Rainio et al., 2024; Zeng et al., 2021; Zhang, 2012).

1. The type of NN
2. The number of hidden layers
3. The forecast horizon
4. The amount of data used for the forecast
5. Smoothing window of the data
6. Metric to judge the model performance

Precise predictions are not the main goal, instead one wants to know which variables change the prediction. Still, we find that our methods are robust to changes in details, and a quantitative analysis for changes to (1)–(4) is provided in Supporting Information S1.

We tested several neural network (NN) models: Convolutional Neural Networks (CNNs) efficiently capture local temporal patterns and can be processed in parallel, making them fast; Recurrent Neural Networks (RNNs) are designed for sequential data, capturing dependencies but may struggle with long sequences; Long Short-Term Memory networks (LSTMs) overcome RNN limitations by using gates to manage information over extended sequences; and Bidirectional Long Short-Term Memory networks (BiLSTMs) extend LSTMs by processing data in both forward and backward directions, thus providing a richer context (Gers et al., 2000; Ismail Fawaz et al., 2019; Lin et al., 2017; Schuster & Paliwal, 1997; Siami-Namini et al., 2019; Sutskever et al., 2014). It should be noted that the BiLSTM model only uses the past time steps as input to predict one future value and does not access values beyond the input window. The input features were scaled using the MinMaxScaler to enhance numerical stability during training (Pedregosa et al., 2011). The models were trained on the training data set using a combination of early stopping and learning rate reduction strategies to prevent overfitting and to optimize performance (Pedregosa et al., 2011).

A custom loss function to synchronize the abrupt transitions is applied, combining Root Mean Squared Error (RMSE) with penalties on mismatches in gradients, emphasizing abrupt changes using 5-sigma and 2-sigma thresholds and aligning large gradient variations in the predictions with the true values (Text S2 in Supporting Information S1). In addition to RMSE, it is natural to design a similar performance metric that incorporates these penalties in the loss function for abrupt changes. Moreover, as the custom loss function focuses heavily on gradient matching to emphasize abrupt changes, this may lead to overlooking small deviations in regions with higher variability by averaging them out when used as a performance metric, we also use Edit Distance on Real Sequence (EDR) to evaluate how well the model aligns with both overall prediction accuracy and abrupt transitions in the data (Chen et al., 2005). The following three performance metrics are used to assess the general performance of models.

- Root Mean Squared Error (RMSE): Measures the overall prediction accuracy, emphasizing large errors by calculating the square root of the mean squared differences between predicted and actual values.
- Custom Performance Metric with Gradient-Based Penalty: Applies penalties for mismatches in gradients between predicted and actual values, focusing on abrupt transitions. This includes both a 5-sigma penalty for large, abrupt changes and a 2-sigma penalty for moderate transitions, enhancing the model's sensitivity to abrupt changes.

- Edit Distance on Real Sequence (EDR): Focuses on aligning predicted and actual data patterns, handling small distortions and phase drift in time series data, providing a flexible error measure for comparing sequences and capturing alignment beyond simple magnitude differences (Chen et al., 2005; Sadiq et al., 2020).

Permutation importance is a model-agnostic global explanation method that provides insights into a machine learning model's behavior. It ranks feature importance based on the increase in prediction error when each feature's values are randomly shuffled, breaking its relationship between the feature and the true outcome (Breiman, 2001; Fisher et al., 2019). To assess the permutation importance of different features, the input time series were shuffled randomly, one at a time. The bottom of Figure 2 shows how the performance of the NN declines in turn. For example, reshuffling the ICEFRAC time series has a big impact on the performnace, but NAO has not. Also, the literature suggests that the sea-ice fraction is only a mediator (i.e., Figure 1), and one can test this with the NNs by removing it from the set of predictors, and indeed, we do not find reduced predictive skills, but instead SFWF, which causes freshening and freezing, is assigned even more importance (Figure 2, bottom). All four models performed similarily well, but we ended up choosing BiLSTM (RMSE: 1.17, Custom Performance Metric: 11.88, EDR: 7,946) for its slightly superior performance compared to CNN (RMSE: 1.38, Custom Performance Metric: 12.29, EDR: 8,032), RNN (RMSE: 1.31, Custom Performance Metric: 12.79, EDR: 8,400), and LSTM (RMSE: 1.32, Custom Performance Metric: 12.09, EDR: 8,234). In this case, the BiLSTM model better captures the general trends and transitions in the AMOC data, particularly where there are shifts or patterns (Figure 2). Since the models could all predict the onset of stadials and interstadials with just one hidden layer, no efforts were made to add additional layers or to experiment with data smoothing. Increasing the forecast horizon and reducing the data used for it both reduce the performance, without an obvious cut-off point (Figures S1 and S2 in Supporting Information S1). Thus, we chose a window size of 50 years and a forecast horizon of 20 years as a compromise between quality of forecast and length of computation. The model uses the past 50 time steps as input to predict a single value 20 steps into the future. A similar choice has also been made to address bias and uncertainty in abrupt climate event timing according to J. Slattery et al. (2023), as time lags of less than 20 years are challenging to distinguish reliably. While very similar in performance, close inspection reveals that the BiLSTM model better captures the general trends and transitions in the AMOC data, particularly where there are shifts or patterns, indicating its effectiveness at detecting time lags and preserving the overall structure of the time series, whereas the other models may minimize the average magnitude of errors but do not capture the dynamics of the rapid transitions as effectively as the BiLSTM model (Figure 2).

2.3. Symbolic Regression

The permutation analysis allows us to remove irrelevant variables from the initial list of possible AMOC controllers. They do not, however, allow us to understand the physics behind the AMOC variability. Thus, symbolic regression (SR) is applied to derive human-readable symbolic expressions from the best-performing model's output. Symbolic Regression is yet another machine learning tool, from a given set of operators and functions it determines which combination best captures the time series under investigation. We used a symbolic regression package PySR to achieve this (Cranmer, 2023). The equations derived from SR were evaluated for their ability to approximate the underlying AMOC dynamics. The non-linear least squares optimization is applied to search for a non-linear combination of the SR model with optimized coefficients to unify its output expressions (Dennis et al., 1996; Moré et al., 1980).

Finally, we employed yet another set of tools to help us interpret the results.

- Gradient-based sensitivity analysis is applied to quantify how infinitesimal changes in input features impact predictions of the combined model (Text S3 in Supporting Information S1). It evaluates the effect of each input individually, focusing on local sensitivity around specific points in the input space (Wang et al., 2024).
- Sobol Indices are applied for identifying causal relationships by quantifying the contribution of each input variable and their interactions to the output variance (Text S3 in Supporting Information S1). A high first-order Sobol index S_i indicates direct causality, and a high total index ST_i, with a lower first-order index, suggests complex or confounded interactions (Fel et al., 2021; Prieur & Tarantola, 2017; Rabitz, 2010; Saltelli et al., 2007; Sobol, 2001). Despite Sobol indices being a variance and correlation-based method, they are still effective in this context because sensitive features are typically considered as roots in the underlying causal graph, meaning they are assumed to have a direct or indirect influence on the outcome (Bénesse et al., 2021).

- Volterra expansions are used to compare the behavior of different equations (Text S3 in Supporting Information S1), helping to quantify to what extent an equation found through symbolic regression matches a known physical law (Barrett, 1976; Blom & Brunner, 1987; Boyd & Chua, 1985; Flake, 1963; Volterra, 1959). Another method with a similar goal is Polynomial Chaos Expansion (PCE), which transforms our results into orthogonal polynomial representations and calculates the cosine similarity between their PCE coefficients, and this approach is realized by the Chaospy package (Feinberg & Langtangen, 2015). It quantifies the similarity of their responses to input variations (Branicki & Majda, 2013; Mara & Becker, 2021).

This bewildering set of tools is needed to tie our final result, a formula that connects AMOC strength to environmental parameters, to existing theories. The NNs make clear that density and freshwater variations are key to understand AMOC variability in the present simulations. Unfortunately there is still little theoretical work done to connect density and AMOC strength, simply because of the nonlinear nature of the Navier-Stokes equations. There is, however, a simple scaling law that suggests that AMOC is proportional to $\Delta\rho^{1/3}$, with $\Delta\rho$ being the density difference between the tropics and the polar North Atlantic (Bryan, 1987). This connection appears intuitive enough, but its derivation ignores the effect of rotation (Straub, 1996). Thus, it has been one motivation behind the present study to quantify the relation between AMOC strength and density in a complex climate model, and check if not other important variables should be included in a model of the AMOC.

3. Results

3.1. Predictive Performance

The model employs a sliding window of 50 years of past data to predict the AMOC strength 20 years later. This configuration was identified as optimal through testing of various window sizes and forecasting horizons (Table S1 in Supporting Information S1). The best model is identified as the BiLSTM (Figure 2), and shown in Equation 1. In this equation, y_t represents the predicted AMOC strength at a future time, $\mathbf{x}_{t-50}$ to $\mathbf{x}_{t-1}$ are the past input data points used for the prediction, RNN is the model that processes these inputs along with previous internal states $\mathbf{h}_t$ (the hidden state) and $\mathbf{C}_t$ (the cell state), $\mathbf{W}$ is a weight matrix applied to the BiLSTM output, b is a bias added to this result, and ReLU is the activation function that ensures non-linearity in the model's predictions.

$$\mathrm{AMOC}_{t+20} = \mathrm{ReLU}\left(\mathbf{W}\cdot \mathrm{BiLSTM}\begin{pmatrix} \mathrm{AABW}_{t-50} & \dots & \mathrm{AABW}_{t-1} \\ \mathrm{NAO}_{t-50} & \dots & \mathrm{NAO}_{t-1} \\ \mathrm{SFWF}_{t-50} & \dots & \mathrm{SFWF}_{t-1} \\ \mathrm{N\_SALT}_{t-50} & \dots & \mathrm{N\_SALT}_{t-1} \\ \mathrm{TAUX}_{t-50} & \dots & \mathrm{TAUX}_{t-1} \\ \mathrm{PD\_0m}_{t-50} & \dots & \mathrm{PD\_0m}_{t-1} \\ \mathrm{PD\_200m}_{t-50} & \dots & \mathrm{PD\_200m}_{t-1} \end{pmatrix}, \mathbf{h}_t, \mathbf{C}_t + \mathbf{b}\right) \quad (1)$$

The performance of other NN permutation feature importances is shown in Figure S4 in Supporting Information S1. For all NN models shuffling the timeseries of AABW, TAUX, NAO and N_SALT lead to a minimal decrease in RMSE, indicating that the model does not need these features to make predictions. To identify the most effective subset of important features for predicting AMOC strength, different combinations of these features were evaluated using a BiLSTM model, and it is found that SFWF on its own already does perform quite well (Figure 3).

3.2. Symbolic Expressions

Symbolic Regression was applied to top-performing feature combinations to identify representative and parsimonious models for predicting the AMOC based on the selected features. Symbolic expressions of these feature combinations are obtained from the regression, the corresponding predictions are visualized (Figure 3 and Figure S4 in Supporting Information S1). While the SR model captures the overall trends well, it does not fully capture the extremes of the AMOC behavior, as shown in Figure 3.

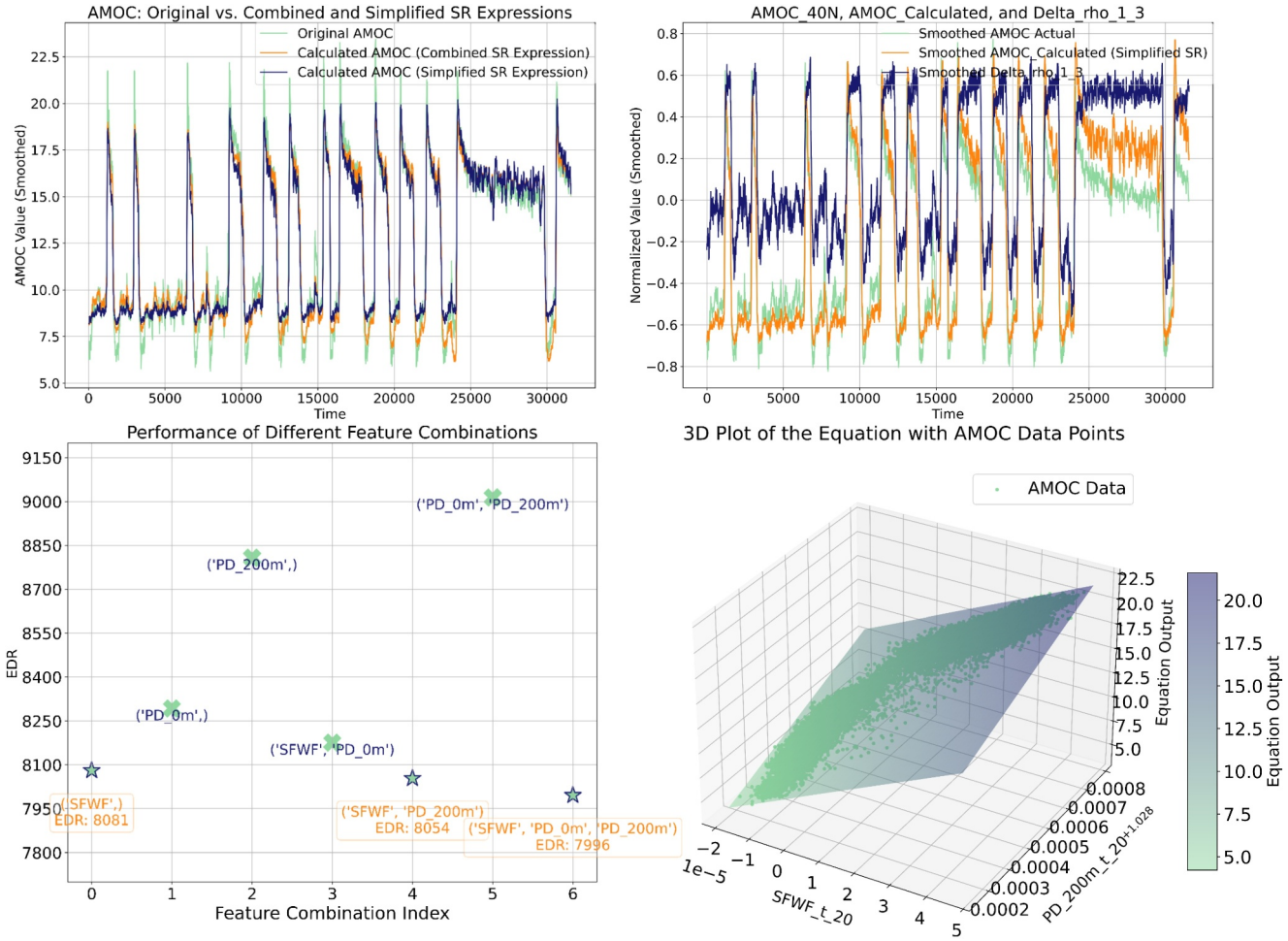

Figure 3. The top left panel compares original AMOC values with calculated AMOC values using the SR. The green line represents the original AMOC data, the orange line shows the calculated AMOC values before the simplification using Equation 2, and the blue line shows the calculated AMOC values after the simplification using Equation 3, a 30-year smoothing is applied for better visualization. The top right panel shows the normalized calculated AMOC values using the resulting equation of SR with the orange line, the normalized values calculated using the scaling in Bryan (1987) with the blue line, and the actual AMOC values with the green line, a 30-year box-car smoothing is applied for better visualization purpose. The bottom left panel shows the performance of different feature combinations in terms of EDR. The stars indicate the top-performing feature combinations at different numbers of features in each subset. The *x*-axis represents the feature combination index, and the *y*-axis represents the EDR value, with lower EDR indicating better performance. The figure suggests that freshwater forcing on its own is already quite important, and density anomalies add only little to performance. The bottom right panel shows Equation 3 with calculated AMOC data points.

We apply non-linear least squares optimization to search for a non-linear combination of the SR models with optimized coefficients, and the comparison of original with the calculated AMOC values is plotted in Figure 3. We compute the gradient of both actual and predicted AMOC time series, and identify the lag where the difference between their gradients is minimized, to ensure the prediction of SR aligned with the timing of key changes. This results in the following equation:

$$\begin{aligned}\mathrm{AMOC}_t \;\simeq\; & C_1 \cdot \mathrm{PD\_200m}_{t-20} \\ & + C_2 \cdot (\mathrm{SFWF}_{t-20})^{1/3} \\ & + C_3 \cdot \mathrm{SFWF}_{t-20} \\ & + C_4 \cdot \frac{\mathrm{SFWF}_{t-20}}{\mathrm{PD\_200m}_{t-20}} \\ & + C_5\end{aligned} \tag{2}$$

To make Equation 2 dimensionless and physically interpretable, coefficients determined by the least-square curve fitting method $C_1 = 8.407\ \mathrm{Sv}\cdot\mathrm{m}^3/\mathrm{kg}$, $C_2 = 45\ \mathrm{Sv}\cdot\left(\mathrm{s}\cdot\mathrm{m}^2/\mathrm{kg}\right)^{1/3}$, $C_3 = 38,383,842\ \mathrm{Sv}\cdot\mathrm{s}\cdot\mathrm{m}^2/\mathrm{kg}$, $C_4 = -39,368.043\ \mathrm{Sv}\cdot\mathrm{m/s}$, $C_5 = -8635.398\ \mathrm{Sv}$ are introduced to appropriately scale terms in the symbolic regression results, ensuring that all terms have consistent units. This adjustment aligns the CESM model's units in the data set, where the AMOC is measured in Sverdrups (Sv), SFWF is measured in kilograms per square meter per second $\left(\mathrm{kg}/\mathrm{m}^2/\mathrm{s}\right)$, and PD is measured in grams per cubic centimeter $\left(\mathrm{g}/\mathrm{cm}^3\right)$.

Sobol sensitivity analysis and gradient-based sensitivity analysis are used for assessing the importance of variables in SR expressions. Results show that $\mathrm{PD\_200m}_{t-20}$ and SFWF_{t-20} has an immediate, local impact on the model output according to gradient sensitivity. Sobol sensitivity analysis suggests that $\mathrm{PD\_0m}_{t-20}$ has higher-order effects and weak interaction with SFWF_{t-20}, while both $\mathrm{PD\_200m}_{t-20}$ and SFWF_{t-20} exhibit strong direct causality on AMOC, with SFWF_{t-20} having the highest first-order and total Sobol sensitivity as summarized below.

Variable	S_i	ST_i	Causality and confounding	Gradient sensitivity
$\mathrm{PD\_0m}_{t-20}$	0.0528	0.0538	Weak interactions with AMOC. Possible confounding with: $\mathrm{PD\_200m}_{t-20}$, SFWF_{t-20}	0.0638
$\mathrm{PD\_200m}_{t-20}$	0.1366	0.1370	Strong direct causality on AMOC	0.6341
SFWF_{t-20}	0.8095	0.8101	Strong direct causality on AMOC	0.4221

However, after further analysis, it was found that the terms involving C_1, C_2, and C_5 contribute relatively small to the overall equation. Therefore, we focus on the terms involving C_3, C_4 and C_5, which make the main contribution, and perform a least squares curve fit again. Equation 2 was simplified, the coefficients $C_6 = -24,277,902.33\ \mathrm{Sv}\cdot\mathrm{s}\cdot\mathrm{m}^2/\mathrm{kg}$, $C_7 = 25,223.123\ \mathrm{Sv}\cdot\mathrm{m/s}$, and $C_8 = 10.87\ \mathrm{Sv}$ were determined using a least-square curve fitting method. The curve-fitting process confirmed that C_6, C_7 and C_8 sufficiently capture the relationship between SFWF_{t-20}, $\mathrm{PD\_200m}_{t-20}$, and AMOC:

$$\mathrm{AMOC}_t \simeq C_6\cdot\mathrm{SFWF}_{t-20} + C_7\cdot\frac{\mathrm{SFWF}_{t-20}}{\mathrm{PD\_200m}_{t-20}} + C_8 \tag{3}$$

Note that CESM computes freshwater fluxes as virtual inverse salt fluxes, hence the positive signs.

The Overall PCE Similarity Score of 0.996, which quantifies similarity by representing functions as series of orthogonal polynomials based on input distributions, suggests that Equation 3 and the scaling in Bryan (1987) are strongly correlated, as shown in Figure 3. Also, similarity of 0.916 between the Volterra series of Equation 3 and Bryan (1987) indicates a good match, showing that these equations are structurally similar (Text S3 in Supporting Information S1).

4. Summary and Discussion

We used neural networks to predict AMOC strength based on a set of eight predictors. At least to the authors it came as a surprise how accurate these predictions are. Figure 4 of Vettoretti and Peltier (2018) show centennial-scale trends leading up to an AMOC transition, but being able to predict the exact timing is quite unexpected.

Shuffling and recombining the time series of these predictors showed that at least in the analyzed simulations the AMOC is mainly controlled by surface freshwater forcing and to a smaller extent through the advection of density anomalies. This not a new result, it corroborates the detailed analysis of two transitions in different CESM version by Vettoretti and Peltier (2018). The main contribution of the present study is that we arrived there using only standard ML tools, and that we analyzed 26 transitions rather than two. After identifying the two key predictors we used symbolic regression to establish a functional relationship between AMOC strength, freshwater forcing and subsurface density anomalies. The resulting function has structural similarities to the scaling law of Bryan (1987) and captures the AMOC strength quite well.

The methodology employed here requires, of course, that a system is at least to some extend predictable. The D-O events in the present models clearly are, but perturbation experiments in different model setup suggest they may not always be, stochastic forcing may be too strong (Kleppin et al., 2015). The discrepancies in precisely pinpointing abrupt transitions set a limitation on the model's ability to capture AMOC behavior accurately (Figure S6 in Supporting Information S1), as discrepancies may inevitably arise from stochastic factors, underscoring the potential influence of unpredictable atmospheric events. Furthermore, observational records and their associated data processing steps, including uneven temporal sampling and linear interpolation for regridding, may introduce artifacts that complicate accurate modeling and interpretation. However, a statistical analysis of ice core proxies suggests that the D-O events in the real world may indeed by predictable (Boers, 2018). Thus, a natural next step following the present study is the use of NNs on ice core data, where we use the CESM simulations as testbed to understand the effect of noise, dating uncertainties, and processing artifacts on the outcome.

Another promising avenue is the construction of reduced dimension emulators. Starting with Stommel (1961), they have a long history in climate research, because the reduced dimensionality allows the study of a wide parameter range and the relative ease of interpretation. So far, however, they have been built based on intuition as in Vettoretti et al. (2022a), whereas now we discover them via SR. This does not mean it is fully objective. As we have documented here, the use of SR and the path there is full of subjective choices, but their impact can be quantified and their robustness tested.

Data Availability Statement

Source data, extended data, and statements of data availability are available at Vettoretti et al. (2022b). The code and the converted data set can be found online at https://sid.erda.dk/sharelink/g0wQhlMT1b.

Acknowledgments
The CESM simulations and most of the analysis were performed at the Danish Center for Climate Computing (DC^3), but we are grateful for extra resources through LUMI project 465000815, awarded by the Danish e-Infrastructure Consortium (DeIC) and Aarhus University as project DeiC-AU-L5-0022. G. Vettoretti was funded through the EU project GreenFeedback, and R. Nuterman through the EU project NextGEMS. We appreciate the fruitful discussions about ML with Dion Häfner, Peisong Zheng and Niels Hansen and are indebted to the inspiring atmosphere of the Bornö research station. We are also grateful for the suggestions and guidance of two anonymous referees and the editor K. Karnauskas.

References

Barrett, J. (1976). Lectures on nonlinear systems. Technische Hogeschool Eindhoven. (Bevat ook: Bibliography on volterra series hermite functional expansions and related subjects)

Bénesse, C., Gamboa, F., Loubes, J.-M., & Boissin, T. (2021). Fairness seen as global sensitivity analysis. Retrieved from https://arxiv.org/abs/2103.04613

Blom, J. G., & Brunner, H. (1987). The numerical solution of nonlinear volterra integral equations of the second kind by collocation and iterated collocation methods. *SIAM Journal on Scientific and Statistical Computing*, *8*(5), 806–830. https://doi.org/10.1137/0908068

Boers, N. (2018). 07). Early-warning signals for dansgaard-oeschger events in a high-resolution ice core record. *Nature Communications*, *9*(1), 2556. https://doi.org/10.1038/s41467-018-04881-7

Boyd, S., & Chua, L. (1985). Fading memory and the problem of approximating nonlinear operators with volterra series. *IEEE Transactions on Circuits and Systems*, *32*(11), 1150–1161. https://doi.org/10.1109/tcs.1985.1085649

Branicki, M., & Majda, A. (2013). Fundamental limitations of polynomial chaos for uncertainty quantification in systems with intermittent instabilities. *Communications in Mathematical Sciences*, *11*(1), 55–103. https://doi.org/10.4310/CMS.2013.v11.n1.a3

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Brix, H., & Gerdes, R. (2003). North atlantic deep water and antarctic bottom water: Their interaction and influence on the variability of the global ocean circulation. *Journal of Geophysical Research*, *108*(C2), 3022. https://doi.org/10.1029/2002JC001335

Bryan, F. (1987). Parameter sensitivity of primitive equation ocean general circulation models. *Journal of Physical Oceanography*, *17*(7), 970–985. https://doi.org/10.1175/1520-0485(1987)017⟨0970:PSOPEO⟩2.0.CO;2

Capron, E., Rasmussen, S., Popp, T., Erhardt, T., Fischer, H., Landais, A., et al. (2021). The anatomy of past abrupt warmings recorded in Greenland ice. *Nature Communications*, *12*(1), 2106. https://doi.org/10.1038/s41467-021-22241-w

Chen, L., Özsu, M., & Oria, V. (2005). Robust and fast similarity search for moving object trajectories. In *Proceedings of the ACM SIGMOD international conference on management of data, 491–502. (SIGMOD 2005: ACM SIGMOD international conference on management of data; conference date: 14-06-2005 through 16-06-2005)*. https://doi.org/10.1145/1066157.1066213

Cranmer, M. (2023). Interpretable machine learning for science with pysr and symbolicregression.jl. Retrieved from https://arxiv.org/abs/2305.01582

Dansgaard, W., Johnsen, S. J., Clausen, H. B., Dahl-Jensen, D., Gundestrup, N. S., Hammer, C. U., et al. (1993). Evidence of general instability of climate from a 250-kyr ice-core record. *Nature*, *364*(6434), 218–220. https://doi.org/10.1038/364218a0

Dennis, J., Robert, J., & Schnabel, B. (1996). Numerical methods for unconstrained optimization and nonlinear equations. *SIAM*, *14*, 045–076. https://doi.org/10.1137/1.9781611971200

Erhardt, T., Capron, E., Rasmussen, S. O., Schüpbach, S., Bigler, M., Adolphi, F., & Fischer, H. (2019). Decadal-scale progression of the onset of dansgaard–oeschger warming events. *Climate of the Past*, *15*(2), 811–825. https://doi.org/10.5194/cp-15-811-2019

Feinberg, J., & Langtangen, H. P. (2015). Chaospy: An open source tool for designing methods of uncertainty quantification. *Journal of Computational Science*, *11*, 46–57. https://doi.org/10.1016/j.jocs.2015.08.008

Fel, T., Cadene, R., Chalvidal, M., Cord, M., Vigouroux, D., & Serre, T. (2021). Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. Retrieved from https://arxiv.org/abs/2111.04138

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Retrieved from https://arxiv.org/abs/1801.01489

Flake, R. H. (1963). Volterra series representation of nonlinear systems. *Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry*, *81*(6), 330–335. https://doi.org/10.1109/TAI.1963.6371765

Gers, F., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural Computation*, *12*(10), 2451–2471. https://doi.org/10.1162/089976600300015015

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438. https://doi.org/10.2307/1912791

Häfner, D., Gemmrich, J., & Jochum, M. (2023). Machine-guided discovery of a real-world rogue wave model. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(48). https://doi.org/10.1073/pnas.2306275120

Henry, L., McManus, J., Curry, W., Roberts, N., Piotrowski, A., & Keigwin, L. (2016). North atlantic ocean circulation and abrupt climate change during the last glaciation. *Science*, *353*(6298), 470–474. https://doi.org/10.1126/science.aaf5529

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Towards neural earth system modelling by integrating artificial intelligence in earth system science. *Nature Machine Intelligence*, *3*(8), 667–674. https://doi.org/10.1038/s42256-021-00374-3

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Jochum, M., Chase, Z., Nuterman, R., Pedro, J., Rasmussen, S., Vettoretti, G., & Zheng, P. (2022). Carbon fluxes during dansgaard–oeschger events as simulated by an earth system model. *Journal of Climate*, *35*(17), 5745–5758. https://doi.org/10.1175/JCLI-D-21-0713.1

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (2015). The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kleppin, H., Jochum, M., Otto-Bliesner, B., Shields, C. A., & Yeager, S. (2015). Stochastic atmospheric forcing as a cause of Greenland climate transitions. *Journal of Climate*, *28*(19), 7741–7763. https://doi.org/10.1175/JCLI-D-14-00728.1

Kuhlbrodt, T., Griesel, A., Montoya, M., Levermann, A., Hofmann, M., & Rahmstorf, S. (2007). On the driving processes of the atlantic meridional overturning circulation. *Reviews of Geophysics*, *45*(2). https://doi.org/10.1029/2004RG000166

Lin, T., Guo, T., & Aberer, K. (2017). Hybrid neural networks for learning the trend in time series proceedings of the twenty-sixth international joint conference on artificial intelligence (IJCAI-17). *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2273–2279. https://doi.org/10.24963/ijcai.2017/316

Mara, T. A., & Becker, W. E. (2021). Polynomial chaos expansion for sensitivity analysis of model output with dependent inputs. *Reliability Engineering and System Safety*, *214*, 107795. https://doi.org/10.1016/j.ress.2021.107795

Moré, J. J., Garbow, B. S., & Hillstrom, K. E. (1980). User guide for minpack-1. [in fortran]. *Interdisciplinary Journal of Information, Knowledge, and Management*, *14*, 045–076. https://doi.org/10.28945/4184

Murari, A., Rossi, R., & Gelfusa, M. (2023). Combining neural computation and genetic programming for observational causality detection and causal modelling. *Artificial Intelligence Review*, *56*(7), 6365–6401. https://doi.org/10.1007/s10462-022-10320-3

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Prieur, C., & Tarantola, S. (2017). Variance-based sensitivity analysis: Theory and estimation algorithms. In R. Ghanem, D. Higdon, & H. Owhadi (Eds.), *Handbook of uncertainty quantification* (pp. 1217–1239). Springer International Publishing. https://doi.org/10.1007/978-3-319-12385-1_35

Rabitz, H. (2010). Global sensitivity analysis for systems with independent and/or correlated inputs. *Procedia - Social and Behavioral Sciences*, *2*(6), 7587–7589. https://doi.org/10.1016/j.sbspro.2010.05.131

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Sadiq, M. U., Yousaf, M., Aslam, L., Aleem, M., Sarwar, D. S., & Jaffry, S. W. (2020). Nvpd: Novel parallel edit distance algorithm, correctness, and performance evaluation. *Cluster Computing*, *23*(2), 879–894. https://doi.org/10.1007/s10586-019-02962-w

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2007). *Global sensitivity analysis: The primer*. John Wiley and Sons, Ltd. https://doi.org/10.1002/9780470725184

Schuster, M., & Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The performance of lstm and bilstm in forecasting time series. In *IEEE international conference on big data (big data)* (pp. 3285–3292). https://doi.org/10.1109/BigData47090.2019.9005997

Slattery, A., Wen, Z., Tenblad, P., Sanjosé-Orduna, J., Pintossi, D., den Hartog, T., & Noël, T. (2024). Automated self-optimization, intensification, and scale-up of photocatalysis in flow. *Science*, *383*(6681), eadj1817. https://doi.org/10.1126/science.adj1817

Slattery, J., Sime, L. C., Muschitiello, F., & Riechers, K. (2023). The temporal phasing of rapid dansgaard–oeschger warming events cannot be reliably determined. *EGUsphere*, 1–27. https://doi.org/10.5194/egusphere-2023-2496

Sobol, I. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1–3), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

Stommel, H. (1961). Thermohaline convection with two stable regimes of flow. *Tellus*, *13*(2), 224–230. https://doi.org/10.1111/j.2153-3490.1961.tb00079.x

Straub, D. N. (1996). An inconsistency between two classical models of the ocean buoyancy driven circulation. *Tellus A: Dynamic Meteorology and Oceanography*, *48*(3), 477. https://doi.org/10.3402/tellusa.v48i3.12073

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th international conference on neural information processing systems - volume 2* (pp. 3104–3112). MIT Press.

Toggweiler, J., & Samuels, B. (1995). Effect of drake passage on the global thermohaline circulation. *Deep Sea Research Part I: Oceanographic Research Papers*, *42*(4), 477–500. https://doi.org/10.1016/0967-0637(95)00012-U

Treguier, A. M., Deshayes, J., Le Sommer, J., Lique, C., Madec, G., Penduff, T., et al. (2014). Meridional transport of salt in the global ocean from an eddy-resolving model. *Ocean Science*, *10*(2), 243–255. https://doi.org/10.5194/os-10-243-2014

van Westen, R. M., Kliphuis, M., & Dijkstra, H. A. (2024). Physics-based early warning signal shows that amoc is on tipping course. *Science Advances*, *10*(6), eadk1189. https://doi.org/10.1126/sciadv.adk1189

Vettoretti, G., Ditlevsen, P., Jochum, M., & Rasmussen, S. (2022a). Atmospheric co2 control of spontaneous millennial-scale ice age climate oscillations. *Nature Geoscience*, *15*(4), 300–306. https://doi.org/10.1038/s41561-022-00920-7

Vettoretti, G., Ditlevsen, P., Jochum, M., & Rasmussen, S. O. (2022b). Decadal average time series data from the ccsm4 model simulations. *University of Copenhagen Electronic Research Data Archive (ERDA)*. Retrieved from https://sid.erda.dk/cgi-sid/ls.py?share_id=Fo2F7YWBmv. [Dataset].

Vettoretti, G., & Peltier, W. R. (2018). Fast physics and slow physics in the nonlinear dansgaard–oeschger relaxation oscillation. *Journal of Climate*, *31*(9), 3423–3449. https://doi.org/10.1175/JCLI-D-17-0559.1

Volterra, V. (1959). *Theory of functionals and of integral and integro-differential equations/vito volterra; preface griffith c. evans*. Dower.

Wang, Y., Zhang, T., Guo, X., & Shen, Z. (2024). Gradient based feature attribution in explainable ai: A technical review. *arXiv*. preprint arXiv: 2403.10415. Retrieved from https://arxiv.org/abs/2403.10415

Zeng, S., Graf, F., Hofer, C., & Kwitt, R. (2021). Topological attention for time series forecasting. Retrieved from https://arxiv.org/abs/2107.09031

Zhang, G. P. (2012). Neural networks for time-series forecasting. In G. Rozenberg, T. Bäck, & J. N. Kok (Eds.), *Handbook of natural computing* (pp. 461–477). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-92910-9_14