JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2021MS002496

Special Section:
Machine learning application to Earth system modeling

Revealing the Impact of Global Heating on North Atlantic Circulation Using Transparent Machine Learning

Maike Sonnewald[1,2,3] **and Redouane Lguensat**[4,5]

[1]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA, [2]NOAA/OAR Geophysical Fluid Dynamics Laboratory, Ocean and Cryosphere Division, Princeton, NJ, USA, [3]University of Washington, School of Oceanography, Seattle, WA, USA, [4]Laboratoire des Sciences du Climat et de l'Environnement, CEA Saclay, Institut Pierre Simon Laplace, Gif-Sur-Yvette, France, [5]LOCEAN-IPSL, Sorbonne Université, Institut Pierre Simon Laplace, Paris, France

Key Points:

- The transparent machine learning method Tracking global Heating with Ocean Regimes (THOR) is presented, applied to the North Atlantic ocean
- Global heating shifts regimes of the Gulf Stream north, the North Atlantic Current east and deep water mass formation toward lighter waters
- Widely applicable, THOR could accelerate analysis and dissemination of climate model data needing only depth, sea level and wind stress

Correspondence to:
M. Sonnewald,
maikes@princeton.edu

Citation:
Sonnewald, M., & Lguensat, R. (2021). Revealing the impact of global heating on North Atlantic circulation using transparent machine learning. *Journal of Advances in Modeling Earth Systems*, *13*, e2021MS002496. https://doi.org/10.1029/2021MS002496

Received 5 FEB 2021
Accepted 29 JUN 2021

Abstract The North Atlantic ocean is key to climate through its role in heat transport and storage. Climate models suggest that the circulation is weakening but the physical drivers of this change are poorly constrained. Here, the root mechanisms are revealed with the explicitly transparent machine learning (ML) method Tracking global Heating with Ocean Regimes (THOR). Addressing the fundamental question of the existence of dynamical coherent regions, THOR identifies these and their link to distinct currents and mechanisms such as the formation regions of deep water masses, and the location of the Gulf Stream and North Atlantic Current. Beyond a black box approach, THOR is engineered to elucidate its source of predictive skill rooted in physical understanding. A labeled data set is engineered using an explicitly interpretable equation transform and k-means application to model data, allowing theoretical inference. A multilayer perceptron is then trained, explaining its skill using a combination of layerwise relevance propagation and theory. With abrupt CO_2 quadrupling, the circulation weakens due to a shift in deep water formation regions, a northward shift of the Gulf Stream and an eastward shift in the North Atlantic Current. If CO_2 is increased 1% yearly, similar but weaker patterns emerge influenced by natural variability. THOR is scalable and applicable to a range of models using only the ocean depth, dynamic sea level and wind stress, and could accelerate the analysis and dissemination of climate model data. THOR constitutes a step toward trustworthy ML called for within oceanography and beyond, as its predictions are physically tractable.

Plain Language Summary The North Atlantic circulation is key to climate through heat transport and storage, and is projected to weaken under global heating. The mechanisms of change remain obscure, but are addressed here using a transparent machine learning (ML) method, engineered combining interpretable and explainable methods to reveal its source of predictive skill. Tackling the fundamental question of identifying dynamically coherent regimes governing the circulation, the Tracking global Heating with Ocean Regimes (THOR) method reveals a weakened circulation under abrupt CO_2 quadrupling, seeing a shift in deep water formation, the Gulf Stream and North Atlantic Current. If CO_2 is increased 1% yearly, similar but weaker patterns emerge. THOR is readily applicable to other models needing only depth, wind stress and sea surface height fields as input, and could accelerate discovery and analysis. THOR is a step toward trustworthy ML called for within oceanography and beyond because its predictions are physically tractable.

1. Introduction

The North Atlantic Meridional Overturning Circulation (AMOC) is defined as a zonally integrated stream function of meridional volume transport in the Atlantic Basin. AMOC is central to the global climate and particularly that of northwestern Europe, bringing warm waters north where they become dense and sink (Lohmann et al., 2014; Lozier et al., 2019; Zhang et al., 2019). Emerging from a myriad of interacting dynamics, the AMOC acts as a primary mechanism for North Atlantic storage of heat and carbon (Lohmann et al., 2014; Marshall & Schott, 1999; Roberts et al., 2004; Tsujino et al., 2020). Due to the complicated and nonlinearly interacting governing features of the AMOC, in-depth and often unavailable data is necessary

to understand potential sources of variability. Here, a transparent Machine learning (ML) method that elucidates the governing mechanisms of AMOC is presented called Tracking global Heating with Ocean Regimes (THOR). THOR is engineered to use only limited and readily available data to predict the governing mechanisms. Here, "transparent" is defined as having the source of predictive skill not only being retrospectively explainable, but also to be interpretable using established theory. Developing transparent ML is seen as key toward building trustworthy ML applications for oceanography and beyond. While globally applicable, here the variability of key underpinning dynamics contributing to the AMOC variability are assessed in a climate model under global heating. THOR addresses the known capability gap of analysis tools for climate models (Eyring et al., 2019; Reichstein et al., 2019; Schlund et al., 2020), while opening the "black box" often associated with ML applications.

The AMOC, and indeed the global climate, exhibits an array of changes in response to anthropogenic forcing, with variability poorly constrained by models (Cheng et al., 2013; Larson et al., 2020; Meehl et al., 2000; Weaver et al., 2012; Weijer et al., 2020; Zhang et al., 2019). To understand likely future changes in the AMOC and indeed the Earth system, the Coupled Model Intercomparison Project (CMIP) now in its' sixth phase is often used (Eyring et al., 2015; Meehl et al., 2000, 2007; Taylor et al., 2012). The CMIP6 ensemble members overall show a decline in the AMOC with global heating, but presents the circulation as a bulk metric leaving specific mechanisms opaque (Weijer et al., 2020). The complexity and size of the CMIP6 model ensemble can hinders data dissemination and analysis, limiting the ability to discern specific mechanisms underpinning variability such as the AMOC decline because necessary data is unavailable. This is an example of an emerging class of problems in CMIP6 and beyond, where researchers must handle data that is increasingly large, potentially sparse, and due to logistics of for example dissemination, often unavailable (Eyring et al., 2019).

The rate and direction of northward transport of warm waters and the density and depth of the southward return flow comprise the AMOC. The formation of North Atlantic Deep Water (NADW) from intense surface cooling returns dense watermasses south (Böning et al., 2006; Lohmann et al., 2014; Marshall & Schott, 1999). The Gulf Stream and the North Atlantic Current (also referred to as the North Atlantic Drift or Trans Atlantic Current) are major sources of warm surface waters through the horizontal gyre circulation. The gyre circulation is coupled to AMOC, modulated by the NADW through bathymetric interactions (Yeager, 2015; Zhang, 2008; Zhang & Vallis, 2007), and dense deep water can be associated with a vigorous AMOC. Three locations are mainly seen as NADW source regions; the Labrador Sea deep water (LSDW) from the basin between Canada and Greenland, the Denmark Strait Overflow Water (DSOW) entering the Atlantic from the area between Greenland and Iceland and the Iceland-Scotland Overflow Water (ISOW) coming from the east of the Reykjanes ridge. The Reykjanes Ridge, stretching south and into the mid Atlantic Ridge from Iceland, forms an obstacle for the deep waters that largely flow counterclockwise to head south at depth. Due to its higher characteristic temperature deep water from the LSDW is lighter. On decadal timescales, a northward shift in the Gulf Stream signals a weaker AMOC. After leaving the western boundary of the continental US around the Grand Banks, the flow is referred to as the North Atlantic Current, which shifts eastwards under a weaker AMOC (Joyce & Zhang, 2010; Nye et al., 2011; Sanchez-Franks & Zhang, 2015; Yeager, 2015; Zhang, 2008; Zhang et al., 2019). These mechanisms can be seen as governing the circulation or being a direct product of its strength (Kuhlbrodt et al., 2007; Wunsch & Ferrari, 2004). Overall, the field of oceanography is increasingly starting to use advanced ML methods, as reviewed in Sonnewald et al. (2021). To infer subsurface dynamics, ML has been employed to predict currents at 1000m from satellites (Chapman & Charantonis, 2017), and subsurface structure from idealized simulations (Manucharyan et al., 2021).

THOR overcomes two common problems with ML applications, and a demonstration of how these can be overcome is also a core motivation of the work. These problems are centered around a lack of labeled data, and the difficulty of understanding of the applications' source of predictive power. First, supervised ML algorithms such as neural networks (NN), are particularly useful for regression/classification problems, but need labeled data from which to learn. Such data is scarce, and labeling is often complicated by the data being some combination of highly nonlinear, chaotic, high-dimensional, nonstationary or multi-scale. A label effectively constitutes defining consistent phenomena of interest. THOR uses unsupervised ML and identifies coherent structures within data to use these as labels. Unsupervised ML is particularly useful in

this context, as the labels can be assigned without bias. Second, adoption of ML within the physical sciences suffers from a lack of trust that stems from a lack of a transparent understanding of the source of predictive skill (Irrgang et al., 2021; Rudin, 2019; Sonnewald et al., 2021). Ensuring that what is learned by the machine is physically meaningful, and not due to trivial coincidences, is important for example for reliability and generalization (Balaji, 2020), and to avoid underspecification (D'Amour et al., 2020). Trustworthy ML has also been called for in government guidelines from the European Union (Assessment List for Trustworthy Artificial Intelligence) and in a mandate in the United States of America (E.O. 13960 of December 3, 2020). This transparency can be achieved by either building specifically interpretable ML models (interpretable artificial intelligence or "IAI"), or retrospectively explaining predictive skill (explainable artificial intelligence or "XAI"). THOR is deemed transparent being both interpretable and explainable, specifically using the interpretable first step to feature engineer the second supervised step. For NN and other "black-box" models, methods to explain skill retrospectively include connection weight approaches, Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive Explanation (SHAP) and Layer-wise Relevance Propagation (LRP) (Lapuschkin et al., 2015; Lundberg & Lee, 2017; Olden et al., 2004; Ribeiro et al., 2016; Toms et al., 2020). Together, this class of method is referred to as Additive Feature Attribution (AFA). They aim to attribute the predictive skill to specific input features given for example to the NN, which can then for example be used by a domain expert to ensure the predictions are not due to chance. Other methods rooted in 'saliency' mapping also exist (McGovern et al., 2019). For unsupervised ML, leveraging theoretical knowledge in both the design and interpretation of results can be fruitful, which also motivated its use here (Callaham et al., 2021; Sonnewald et al., 2019, 2020).

THOR provides rapid and comprehensive evaluation of climate model simulations, using ML to objectively identify shifts in physics that modulate the AMOC variability. Here, key shifts in different future forcing scenarios reveal that a shift in the Gulf Stream and North Atlantic Current, together with a change in the deep water formation regions, are suggestive of a weakening AMOC. A focus on transparent ML underpins the study, both through the experiment design and a subsequent analysis of the source of predictive skill. This predictive skill is importantly rooted in physical understanding.

2. Methods and Results

2.1. Identifying Dynamical Regimes

The first step of THOR identifies 2D dynamical regimes (Figure 1a) in the realistic 1° numerical ocean model Estimating the Circulation and Climate of the Ocean (ECCOv4r3 (Adcroft et al., 2004; Forget et al., 2015; Wunsch & Heimbach, 2013), 1992–2013). Approached naively, finding robust regimes is intractable due to the high dimensionality of the complex numerical model, with a high likelihood of nonunique solutions conflating interpretation. THOR uses a model data transformation into equation space, reducing the dimensionality to five and enhancing interpretability (Sonnewald et al., 2019). The five dynamical drivers/terms are the fundamental sources of depth integrated (barotropic) vorticity: (a) the wind and bottom stress curl, (b) the advection of planetary vorticity, (c) bathymetric interactions through bottom pressure torque (BPT), (d) curl of nonlinear interactions between terms and (e) lateral viscous dissipation from within the ocean interior (Hughes & de Cuevas, 2001; Munk, 1950; Sonnewald et al., 2019) (Appendix B). The five terms form a closed budget, and a 5-dimensional vector field, $\mathbf{x}$, on the model grid. Each element $\mathbf{x}_i$ represents the 5-dimensional vector defined on the model's global horizontal grid. Each index i uniquely identifies a grid point on the sphere, with (lon, lat) = (θ_i, ϕ_i). Within $\mathbf{x}$, six distinct and unique dynamical regimes are identified as clusters using the unsupervised ML k-means algorithm and information criteria model selection. The dynamical regimes used in THOR were original presented in Sonnewald et al. (2019), where more details on the method can be found.

The six dynamical regimes are back projected onto the globe, with the geographical area covered signifying the unique balance of dynamical drivers present there (Figure 2a). The global area averaged term balances (Figure 2b) demonstrate which dynamical drivers are important and which are negligible. Here, the North Atlantic is discussed (Figure 2c). The "Northern Hemisphere Sverdrupian" dynamical regime (N-SV, pink) represents a region where the vorticity input by the wind is largely negative, and the input by advection is positive. The term "Sverdrupian" refers to a canonical dynamical balance between the wind stress curl and the advection (Sverdrup, 1947). The "Southern Hemisphere Sverdrupian" dynamical regime (S-SV, green)

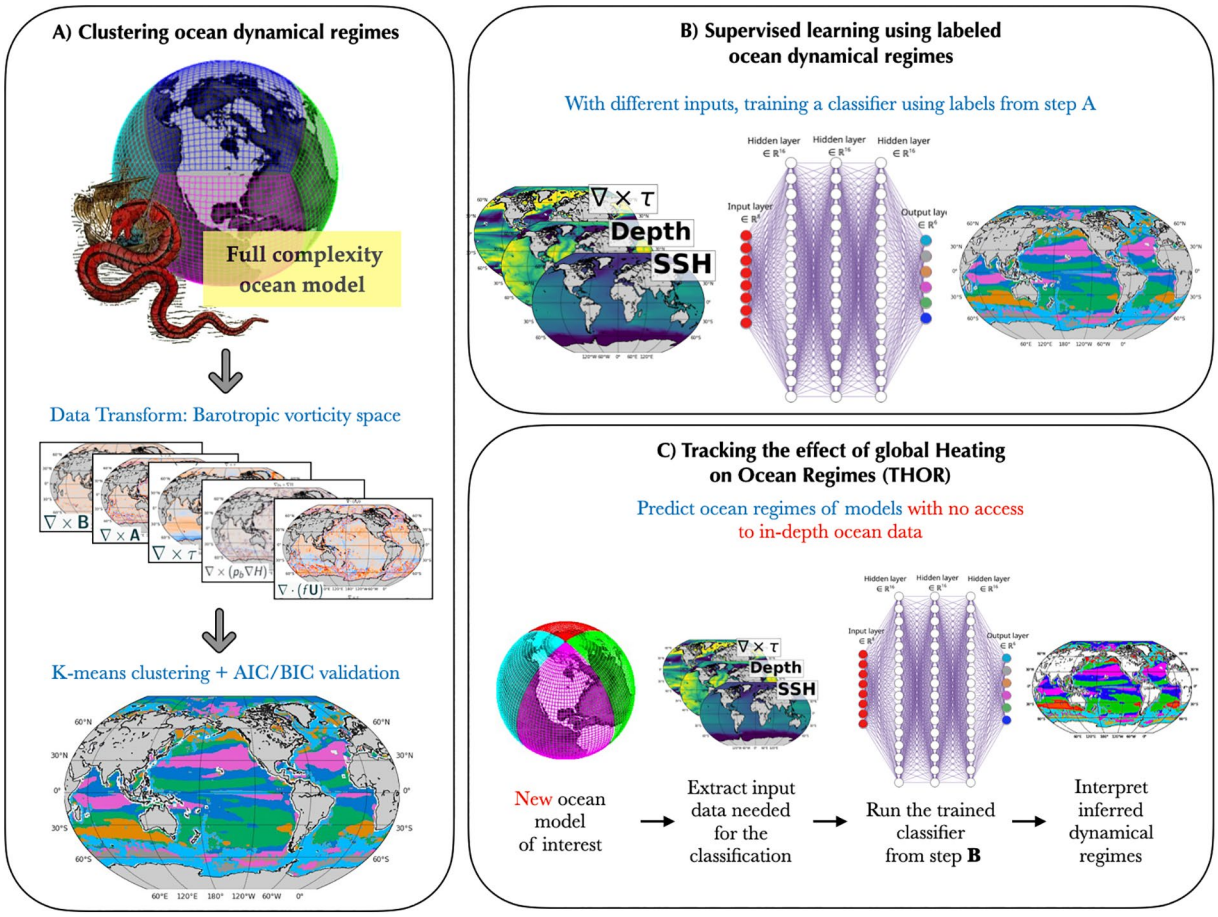

Figure 1. Sketch of THOR workflow. Method to identify dynamical regimes that are indicative of dynamics contributing to the AMOC variability. THOR is engineered for interpretability and explainability of ML predictive skill for transparent, and as such to move toward trustworthy ML. A more detailed sketch of step B can be seen in Figure 5. Globe from Forget et al. (2015).

has a largely positive vorticity input by the wind, while the advection is a source of negative vorticity. In the North Atlantic, the S-SV is found north of the "Momentum Driven" regime (MD, dark blue). The MD regime has area averaged vorticity inputs that are of much smaller magnitude than the other regimes. The wind stress curl adds negative vorticity, while interactions with the bathymetry contribute positive vorticity. The MD dynamical regime occupies a region associated with the North Atlantic Current. The "Transitional" dynamical regime (TR, burnt orange) is found north of the S-SV regime. The TR regime has positive vorticity input by the wind stress curl, and negative vorticity input by the advection and interactions with the bathymetry. This balance is expected from a region associated with deep watermass formation (Zhang et al., 2011). The "Southern Ocean" dynamical regime (SO, gray) is negligible in the North Atlantic. The "Non-linear" regime (NL, light blue), is associated with western boundaries and areas of rough bathymetry, and it is particularly prevalent in the higher latitudes. The NL regime is notable as it is made up of a collection of smaller regimes that all have a large nonlinear torque component, but make up a very small component of the ocean area (Sonnewald et al., 2019).

To interpret a regime's role in the North Atlantic circulation, the co-local density structure and the contribution to the meridional circulation are used. The 2D dynamical regimes allow a partitioning of the in-depth ocean physics by regime. This is achieved by using the dynamical regimes' latitude and longitude spatial extent as a mask, and considering only the depth information covered by this mask. In this manner, it is possible to consider only the properties in the ocean volume (surface to seafloor) delineated by the geographical area covered by a regime. The meridional overturning circulation (Appendix A) captures the bulk meridional movement of watermasses at a fixed latitude, and in the North Atlantic constitutes the AMOC. As a

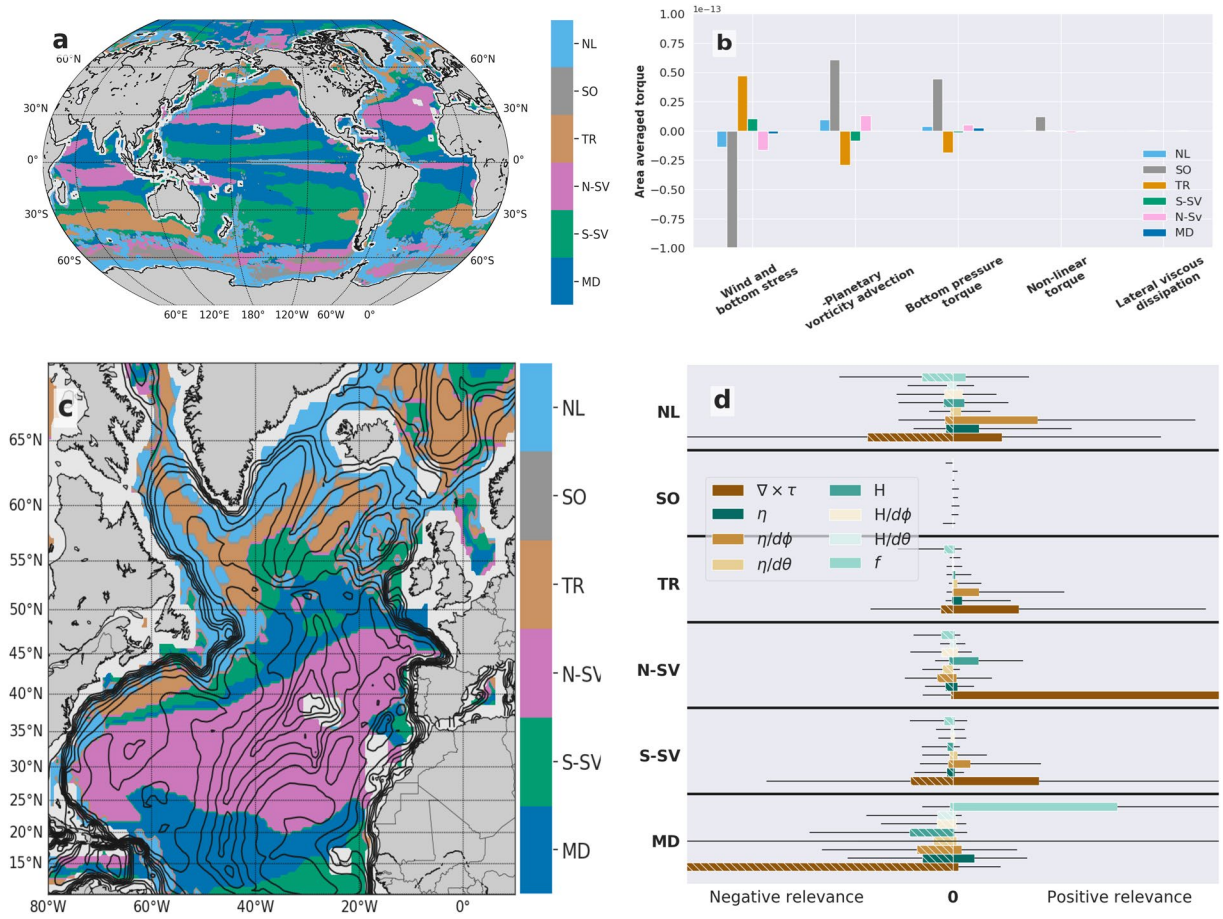

Figure 2. Estimating the Circulation and Climate of the Ocean dynamical regimes geographical expanse, area averaged term magnitudes and learning contributions. (a) is a spatial projection of regimes (adapted from Sonnewald et al., 2019), (b) is a closeup of the North Atlantic, (c) are the area averaged terms (m^2s^{-2}), (d) are the features' respective contributions to the learning in the regimes (black bar: standard deviation). The negative and positive relevances are computed separately for each regime, with the resulting mean and standard deviation presented on the left (negative) and right (positive), as indicated by the *x*-axis labels. The names and colors of the dynamical regimes are: Nonlinear (NL, light blue), Southern Ocean (SO, gray), Transitional (TR, burnt orange), Northward-Sverdrupian (N-SV, pink), Southward-Sverdrupian (S-SV, green) and Momentum Driven (MD, dark blue).

large scale circulation, the AMOC is an overall clockwise feature, with surface waters traveling northward to return south at depth. The individual dynamical regimes' contributions to the AMOC can be assessed by decomposing the overall transport by dynamical regime, and calculating the resultant streamfunction. The sum of the streamfunctions associated with each regime comprises the AMOC. Decomposing the AMOC into dynamical regimes shows the local contribution of each regime individually to the AMOC, and reveals a complex interplay of dynamical features. The density structure can be decomposed by dynamical regime similarly. Together, the density structure and the meridional overturning are thus decomposed by dynamical regime. Overarching coherent and in-depth physical regimes emerge (Figure 3). The overall transport in the N-SV regime is clockwise (red, Figure 3a). It transports relatively light watermasses northwards in the surface (<1000 m) as seen by the light colored isopycnals overlying the transport. It coincides with the large subtropical gyre thought to be in Sverdrup balance (Thomas et al., 2014; Wunsch, 2011). In the S-SV regime the transport is largely anticlockwise (blue, Figure 3b), taking place also in the predominantly lighter watermasses with northward transport confined to the surface (<500 m). The S-SV regime is largely seen in the subpolar gyre. The TR regime also transports waters anticlockwise (Figure 3c). The TR regime is associated with the creation of deep watermasses, with doming of isopycnals in the higher latitudes constituting dense waters close to the surface, and also transports reaching depths below 2500 m. The SO regime is largely confined to the Southern Ocean (Figure 3d), and absent in the North Atlantic. The NL regime (Figure 3e) contributes clockwise between 50 and 80°N, reaching depths of ~2500 m. This regime also has dense waters

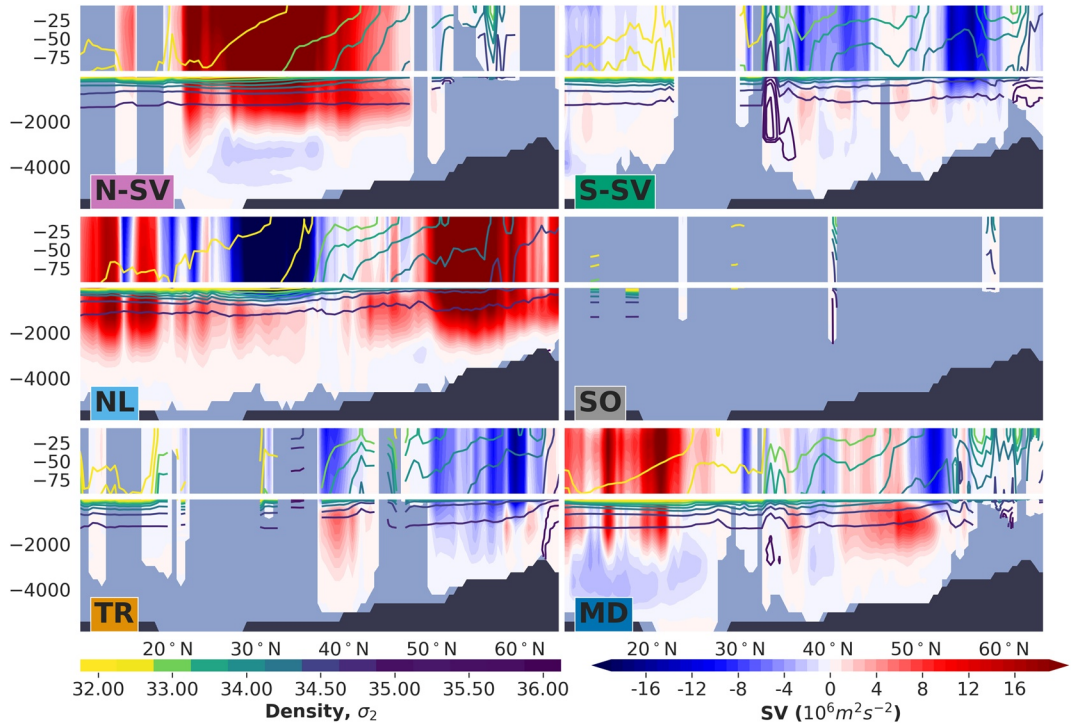

Figure 3. Physical interpretation of regimes from Atlantic Meridional Overturning Circulation (AMOC) contribution and density structure. The individual dynamical regime contributions to the AMOC (filled contours, red northwards, blue southwards) and isopycnal structure (density, contours). Name abbreviations as in Figure 2. Note the Transitional regimes contribution of dense watermasses that move southwards at depth, the stacked (red over blue in depth) contributions of the Momentum Driven regime and the distinct partitioning between northwards and southwards light surface waters contribution for the N-SV and S-SV regimes respectively. The sum of figures (a–f) comprises the AMOC. In gray areas the regime was not present.

close to the surface but they are lighter than in the TR regime. Together, the TR and NL regime are thought to govern the creation and advection of dense waters in the higher latitudes that return south at depth, constituting convection, and the resulting overall clockwise circulation. In the MD regime (Figure 3f), the transports are both clockwise and anticlockwise, with stronger transports largely confined to lower latitudes (<30°N and S). The MD regime acts predominantly in lighter waters. Notably, the MD regime has vertically stacked clockwise/anticlockwise transports, which is only also present in the NL regime. The MD regime is largely found in regions where there is a sign change in the forcing, such as the S-SV and N-SV, where continuity through the convergence between the two suggests a strong eastwards current in the surface waters could be found. This is seen in the MD regime, allowing stacked meridional transports, particularly with a core of clockwise transport at ~1000 m at 47–53°N. The latitudes where the clockwise/anticlockwise circulation is stacked, coincides with the region occupied by the North Atlantic Current.

Figure 4 shows a cartoon of how the dynamical regimes map onto the 3D isopycnal and current structures. The currents at the western boundary, through the Gulf Stream and North Atlantic Current, bring warm and light waters northward hugging the coast until they separate around the Grand Banks (where S-SV and MD regime coincide). As these waters are brought east and north they cool, in the North Atlantic Current (MD regime). Some are transformed to denser watermasses by intense cooling (TR and NL regime). There are several locations where the denser watermasses can be formed, but they are largely brought to depth as LSDW, DSOW or ISOW (marked arrows). The densest waters come from the DSOW and ISOW, and creation of denser waters would overall act to invigorate the AMOC. If there were a shift in the location of deep water formation toward the Labrador Sea, this could incur an AMOC weakening as less dense waters would result. The partitioning of the overall dynamics into the regimes is a simplified representation of the highly complicated full structure, which highlights the underlying processes that constitute the dynamical regimes. The motivation behind using ML for this strategy is that it can identify such areas within

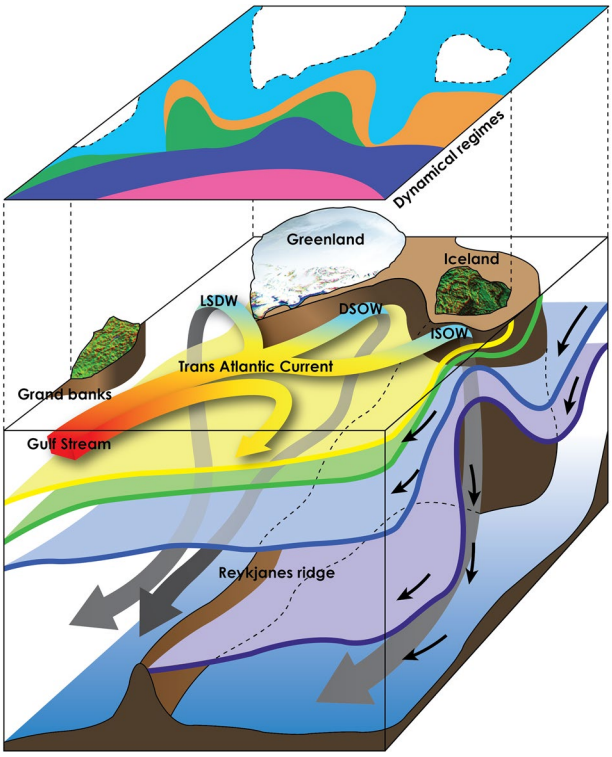

Figure 4. Sketch of dynamics. Upper layer illustrates the dynamical regimes (colors as in Figure 2). Cube below shows density contours (colors as in Figure 3), with the overall currents that are colored from warm (red) to cold (blue). The Denmark Strait Overflow Water and Iceland-Scotland Overflow Water are strictly overflows, but depicted with arrows for simplicity.

the equation space that constitute the dominant term balances in an unbiased manner. THOR independently identifies the expected dynamics, but importantly adds geographic precision. The simplification of the full ECCO data facilitated by the step A of THOR is not only helpful for process understanding, it also rephrases the development of a NN from a continuous to a class based framework.

2.2. Prediction With a NN

The second step of THOR trains a NN (Figure 1b) to infer in-depth dynamics from data that is largely readily available from for example CMIP6 models, using NN methods to infer the source of predictive skill (Figure 5). The data used is comprised of labeled input variables referred to as features, with the dynamical regimes as labels for each point on the model grid. The input features are engineered using the knowledge of the most important dynamical terms from step A: the advective component, the BPT and the wind stress torque. The wind stress torque is largely an available model output, and used as a feature. To approximate the torques from interactions of bottom pressure with the bathymetry, the depth (H) and dynamic sea level (η) are used, with η as a proxy for the pressure at the bottom (Hughes & de Cuevas, 2001; Losch et al., 2004). The advective component is influenced by the wind stress torque ($\nabla \times \tau$), Coriolis (f) and η (Buckley & Marshall, 2016; Bingham & Hughes, 2009; Z. Wang et al., 2015). The f and gradients of the η term reflect the surface geostrophic velocity. In sum the features are: wind stress torque, H, f and η, and the latitudinal and longitudinal gradients of H and η.

A fully connected multilayer perceptron (MLP) NN is used. The motivation to employ a NN is to determine relationships between the input features and the labels within a training data set, so these relationships can be leveraged to make similar predictions for unseen data. MLPs are powerful universal function approximators, and particularly suited for

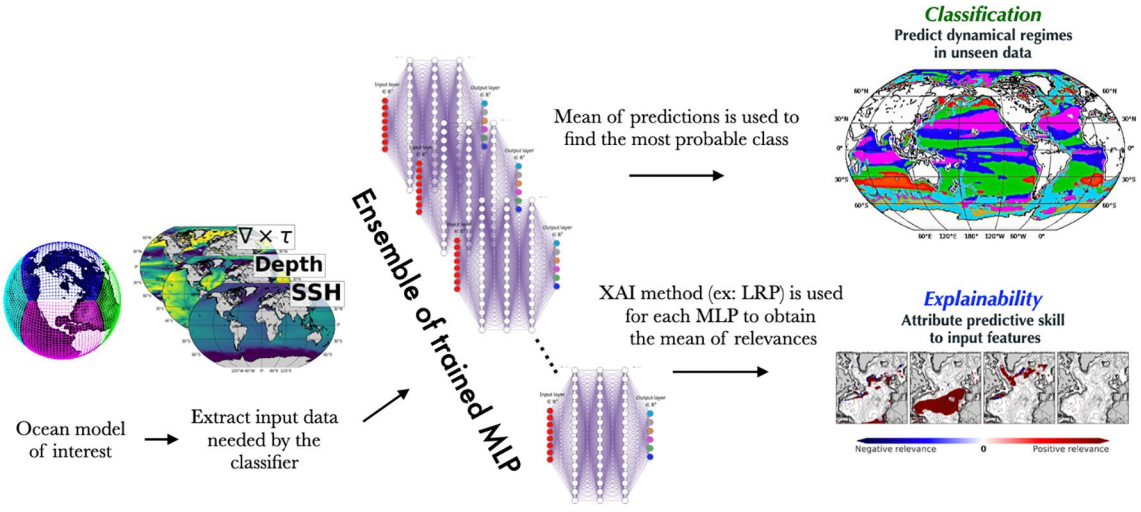

Figure 5. Detailed sketch of Tracking global Heating with Ocean Regimes (THOR) workflow step C illustrating the Ensemble multilayer perceptron (MLP). The last step of THOR applies the Ensemble MLP to unseen data. After extracting input data needed for the classification, the Ensemble of trained MLPs (Step B in Figure 1) is run to get the probabilities of belonging to one of the six classes signifying the dynamical regimes, the mean of the predictions is used to find the most probable class for each (lon,lat) sample. In principle, the same Ensemble MLP can be used to find the most relevant inputs that led to the prediction at a particular (lon, lat) sample in the unseen data using the trained Ensemble MLP for example with LRP.

multi-class classification applications (Cybenko, 1989). Testing, training and validation data were split by ocean basin, ensuring independence. Training input data were normalized to have a zero-mean and a unit-variance. The MLP retained in this work was the result of a hyperparameter search using Hyperband (Li et al., 2017), based on the implementation provided in Keras-Tuner (O'Malley et al., 2019). The search space was the number of neurons {8, 16, 32, 64, 128} and the number of layers {from 2 to 5}, we manually tested different activation function from {ReLU, SeLU, Tanh} and found Tanh to lead to slightly better performances. The hyperparameter search resulted in a 4-layer MLP with respectively 24-24-16-16 neurons and Tanh activations, a softmax layer is used for the final layer. Training was done using backpropagation combined with a stochastic gradient descent algorithm, here ADAM (Kingma & Ba, 2014), with a learning rate of 10^{-4} and early stopping if the validation loss stops improving after five iterations. In order to improve the robustness of the ML method an Ensemble MLP was used, where many instances of the MLP are trained. This is known to improve the generalization capacity and to weaken the dependence on the initial training parameters (Appendix G). The Ensemble MLP considered in this work is composed by 50 MLP with same architecture as mentioned above. When predicting the classes, an average over the 50 softmax probabilities for each pixel was done, and then a new softmax function is applied to constrain the sum of the outputs to be equal to one. The predicted class for a position is then the one with the maximum probability.

Code was written using the Python-based Keras library (Chollet et al., 2015) and makes use of several other open source libraries (Hamman et al., 2018; Harris et al., 2020; Hoyer & Hamman, 2017; Pedregosa et al., 2011). The good performance of the Ensemble MLP is illustrated in Appendix D, where the NL regime was most difficult to classify. An independent validation on an unseen model of similar resolution and access to the barotropic vorticity terms to assess performance was done. This serves as a stringent test to avoid underspecification (D'Amour et al., 2020), and confirms the skill of THOR (CESM1 at 1° horizontal resolution, Figure C1 and Appendix C). For application to further unseen data from CMIP6, the wind stress is taken from the ocean, and simplifying assumptions were made with respect to the curl operator due to a lack of grid-metadata.

2.3. The Source of Predictive Skill

Using supervised ML, being able to explain the source of predictive skill and move beyond a "black box" approach, to create transparency, is often nontrivial. This difficulty should not detract from the importance of transparent ML applications, as leveraging the combination of domain knowledge and emerging ML techniques such as AFA could be of pivotal importance for applications within the physical sciences (Balaji, 2020; Irrgang et al., 2021; McGovern et al., 2019; Sonnewald et al., 2021; Toms et al., 2020). When used as a "black box", a NN will be trained to make desired prediction, and while it can be skillful in making these predictions, it could have skill rooted in chance more than physics. Step B of THOR assesses which features in the input vector give rise to the predictive skill using LRP (Bach et al., 2015; Binder et al., 2016). The LRP method belongs to a growing family of techniques aiming to attribute relevance to the input features toward the resulting prediction. These often produce a "heatmap" rendition of NN classification decisions (Montavon et al., 2017; Rumelhart et al., 1986; Simonyan et al., 2014; Zeiler & Fergus, 2013). The LIME method was also used to assess the source of predictive skill, with similar results. Overall, the LRP method was most robust to local perturbations, and deemed most reliable (see Appendix F for details). Methods for AFA such as the LRP method are distinct from other 'saliency' methods reviewed in McGovern et al. (2019). To construct the "heatmap" individual contributions (called relevance) are calculated from input nodes to the output classification score. A positive/negative relevance suggests that a feature contributes positively/negatively to NN decision (Lapuschkin et al., 2015). In the case of an Ensemble MLP, the contributions are calculated layer by layer from the output layer to the input layer. To illustrate, at layer l, the relevance of a neuron i is the sum of "messages" $R_{i\leftarrow j}^{(l,l+1)}$ from all the neurons j belonging to layer $l+1$ (Binder et al., 2016). These messages are calculated using different variants of the LRP, here an ϵ-rule was used that helps avoid numerical issues when dividing by small numbers:

$$R_{i\leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j + \epsilon \cdot \mathrm{sign}(z_j)} R_j^{(l+1)},$$

where z_{ij} are weighted activations (multiplication of the activation at neuron i with the NN weight from neuron i to j), and z_j is the sum of weighted activation linked to neuron j. A scaling of the relevance maps to lie between -1 and 1 is standard. The relevance maps shown here are the average of the 50 LRP relevance maps calculated using the Ensemble MLP. For geoscientific applications, the positive component of LRP have previously been used to demonstrate different sources of relevance for El Niño event patterns from the eastern Pacific and the central Pacific (Toms et al., 2020). In this work, the LRP-ϵ implementation provided by the iNNvestigate (Alber et al., 2019) library was used, that supports Keras-written models. Figure A1 illustrates the spatial distributions of the relevances.

For each dynamical regime, the relevance contributions are assessed as the mean and standard deviation across the North Atlantic region spatially. Note the initial labels and not the predicted clusters were used. Positive and negative relevance contributions are treated separately (Figure 2d). The information the LRP provides should not be interpreted directly in terms of the theoretical rationale used to select the input features. Rather, the LRP provides an a posteriori assessment of the detailed adjustments of the Ensemble MLP at each location, where the absence of a term can also contribute positive relevance. There is considerable spatial variability, as reflected by the standard deviation, but it is notable that all terms contribute positively. The wind stress curl is the dominant positive feature across all but the NL and MD regimes, although the S-SV regime also features large negative contributions. The longitudinal and latitudinal gradients of η contribute positively in the S-SV, TR and NL regimes, which could be due to a meridional flow facilitated by such a gradient e.g. the Gulf Stream. The f parameter contributes positive relevance to the MD and NL regimes, but largely negative relevance in the S-SV, N-SV and TR regimes. The importance of f in the MD regime could be associated with the geostrophic currents. The H term contributes significant positive relevance in the N-SV regime, as the regions where there is little variability in H within the deep and flat ocean (abyss) are recognized (spatial maps in Figure A1, discussed further in Appendix E). The N-SV regime is notably sheltered from the bathymetry dynamically, and thus a range of H values constituting the abyss would facilitate recognition. While H can contribute to the relevances, the gradient of H in latitude and longitude was not seen to have large relevance, outside of the NL regime. This could be due to the smaller variability in the ranges of the gradient of H as compared to the H term in the North Atlantic sector considered. The ability to explain the Ensemble MLPs skill lends confidence to its subsequent predictions. Assessing the relevance metric highlights the physical underpinning of the Ensemble MLP skill, and means that THOR can be applied with more confidence in previously unseen models or under different climate forcing.

2.4. Interpreting Physical Regimes in a Climate Model

The final step of THOR (Figures 1c and 5) is to apply the trained Ensemble MLP to a climate model in order to assess circulation changes under global heating. This application provides direct knowledge of the dynamical source of the weakening in the AMOC. The model used is the Geophysical Fluid Dynamics Laboratory (GFDL) Earth System Model 4 (ESM4.1 (Dunne et al., 2020; Krasting et al., 2018)) featuring in CMIP6. ESM4.1 is chosen as it is recognized to perform well, having the highest weighting among other CMIP6 models when explaining the historical record (Brunner et al., 2020). ESM4.1 has a horizontal ocean resolution of 1/2° which is comparable to ECCO, containing similar physical processes. Data from the historical scenario was used (1990–2010, comparable to ECCO), which has been forced with observations. Two future forcing scenarios were used, that were run for 150 years. One where the CO_2 concentration in the atmosphere is increased by 1% over 140 years (1pctCO_2), representing a still transient climate state, and an abrupt quadrupling of CO_2 (abrupt4xCO_2) that has had more time to stabilize. The AMOC weakens as expected (Weijer et al., 2020) in the 1pctCO_2, and decreases further in the abrupt4xCO_2 (Figure A1). These are designed to reflect two distinct strategies for how society could move forwards without strong mitigation. To ensure results are not due to natural variability, consistent classifications on 20 years sections of the final 60 years are used and dynamical regime assignments are only given if >75% of predictions agree. If an assignment is given, the dynamical regime classification is described as "robust" to natural variability.

Applying THOR to the ESM4.1 model with historical forcing (Figure 6a), dynamical regime distributions similar to ECCO are seen. The MD regime occupies a large area stretching east and northwards from the

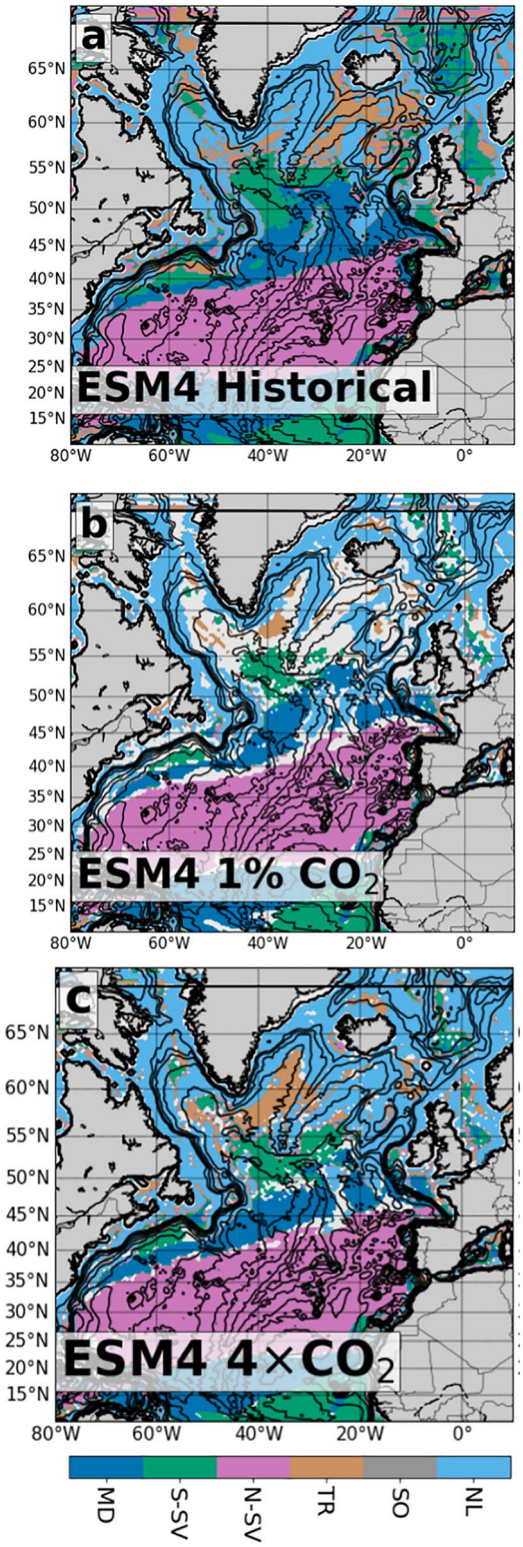

Figure 6. The dynamical regimes predicted under global heating using Tracking global Heating with Ocean Regimes. (a) The historical scenario shows dynamical regimes that are very similar to Estimating the Circulation and Climate of the Ocean (Figure 2). (b) The 1pctCO_2 scenario shows expanses of ocean areas that were not robustly/consistently classified (white), particularly in the areas associated with downwelling. This could suggest episodic downwelling. (c) The abrupt4xCO_2 scenarios shows a distinct shift of the downwelling areas (TR) toward areas where lighter waters are created, an eastwards shift in the North Atlantic Current (MD), and a northward shift in the Gulf Stream. This illustrates dynamical changes that are associated with Atlantic Meridional Overturning Circulation's progressive weakening from historical, 1pctCO_2 and abrupt4xCO_2 scenarios, as observed (Figure A1).

eastern coast of America. It reaches a latitude of ca 55°N, and stays eastwards of the Reykjanes ridge. The N-SV regime is seen spanning the Atlantic Ocean. On the western boundary from 35 to 45°N, a sliver of the MD regime separates the S-SV regime, with patches of the TR regime. The NL regime is prevalent along large parts of the subpolar gyre, somewhat confined to the West of the Reykjanes ridge. In the center of the subpolar gyre (50°N and 40°W) there is a large area of S-SV, with the TR regime extending northward into the Labrador and Irminger basin. The dominant TR area is in the Iceland basin to the east of the Reykjanes ridge.

In the 1pctCO_2 scenario (Figure 6b), the unclassified white areas highlight that the climate could still have a large component of natural variability. The locations associated with deep watermass creation in the historical and abrupt4xCO_2 are not well classified, which is ascribed to natural variability. Interpreting the unclassified regions to be due to episodic shifts in deep watermass creation, the TR regime now occupies the Irminger and Labrador basins periodically. A slower equilibration timescale (less robust classification) of the TR regime is expected, as this would be an advective process rather than a fast Kelvin wave process (Zhang et al., 2011). There is an expansion of the S-SV. The TR and S-SV regime patterns could be associated with different episodic deep watermass formation of upper and lower NADW, with formation in the Labrador Sea likely creating lighter watermasses. The area on the western boundary south of the Grand Banks sees a northward shift of the MD and S-SV regimes, interpreted as a northward shift in the Gulf Stream path. The MD regime has contracted somewhat moving northward, but surrounding areas are poorly classified. This indicates that the North Atlantic Current could be changing, but has not shifted drastically. More detailed discussion and figures can be found in Appendix H.

In the abrupt4xCO_2 scenario (Figure 6c), the climate has had more time to stabilize, and most of the ocean area is robustly assigned to a dynamical regime. The TR regime has shifted to the west of the Reykjanes ridge, and markedly widened its expanse in the Labrador Sea compared to the historical baseline. The northward shift of the MD and S-SV regimes south of the Grand Banks persists, shifting even further North. This suggests that the Gulf Stream also shifts further northward. The MD regime heading across the basin does not make it further north than 50°N, demonstrating a distinct eastwards shift which could indicate a change in the North Atlantic Current. Concurrently, the S-SV regime extends further South. These observed changes point toward a weakened AMOC. More detailed discussion and figures can be found in Appendix H.

Interpreting the changes between the historical and future forcing scenarios, the declining AMOC can be put into context. The 1pctCO_2 scenario is still stabilizing, and has an AMOC that is more highly variable and not quite as weak overall. The mechanisms identified are the meridional shift in the Gulf Stream and the change in location of the deep watermass formation areas. This shift suggests that UNADW is being created. For the abrupt4xCO_2 scenario, the dynamical regime configuration is more stable, in concurrence with the climate having had more time to stabilize. Under abrupt4xCO_2, the Gulf Stream has shifted further north, and the North Atlantic Current has shifted east. The deep water formation regions move toward locations where less dense waters could result. These three factors are associated with a weakening of AMOC (Zhang et al., 2019). It is of note that most of the CMIP6 models predict a weakening of the AMOC (Weijer et al., 2020). Using THOR, comparing the historical scenario to the future scenarios, illustrates that this weakening could be indicated by an eastward shift of the North Atlantic Current, a northward shift of the Gulf Stream, and a likely slower shift of the regions where dense waters are formed to areas where lighter watermasses could be favored. Consistent identification of regimes can help identify the potential dominant mechanisms causing AMOC variability.

3. Discussion and Conclusion

The THOR method is presented, engineered as a trustworthy ML application to recognize dynamical regimes that are tied to dynamical ocean features governing AMOC such as the Gulf Stream, North Atlantic Current and the formation regions of deep watermasses. THOR is grounded in basic understanding of ocean physics, which allows the ML components of THOR to be explicitly evaluated against physical

intuition. While applicable globally, the present focus is on climatically key AMOC. THOR is devised using the ECCO state estimate. Key features modulating the strength and variability of AMOC are localized and assessed in the CMIP6 model ESM4.1 to understand their response under global heating. Dominant drivers are the deep water formation areas and the Gulf Stream and North Atlantic Current transporting heat northward. Elucidating such underlying dynamics in, for example the CMIP6 ensemble, is often hindered by the difficulty of data dissemination for analysis. In response to this difficulty, THOR is developed, and engineered to use readily available climate model data: the mean η and H, their lateral gradients, the wind stress curl and f. The dynamical regimes are predicted using an explainable Ensemble MLP. The Ensemble MLP has been trained by constructing a labeled data set using interpretable unsupervised ML, clustering on transformed realistic 3D ocean model momentum fields (Sonnewald et al., 2019) (Figure 1). The labels constitute six dynamical regimes, representing northward and southward surface transport, northern hemisphere deep water formation and southern hemisphere upwelling, a MD regime and a composite dynamical regime where nonlinear processes dominate (Figures 2 and 3).

Using THOR, the evaluated forcing scenarios are the historical and the future projections 1pctCO_2 and abrupt4xCO_2 (Figure 6). In the North Atlantic (Figure 6), the location of deep water formation (TR) moves from the east of the Reykjanes ridge to the west, and into the Labrador Sea where less dense watermasses are formed. The regime associated with the North Atlantic Current (MD) reduces its reach northwards and is seen to shift eastwards, particularly in the abrupt4xCO_2 scenario. South of the Grand Banks, the latitudinally stacked S-SV and MD regimes, associated with the Gulf Stream path, shift north, particularly in the abrupt4xCO2 scenario. The AMOC decreases from the historical to the 1pctCO_2 and further in the abrupt4xCO_2, and THOR elucidates the dynamics that could underpin this change. Identifying such in-depth dynamics is difficult in CMIP, both due to the prohibitively large volumes of data, with their associated dissemination hurdles, as well as the lack of all necessary fields being saved to close the ocean momentum budget. The source of predictive skill for the Ensemble MLP (Figure 2d, Appendix E) illustrates the importance of the change in the wind stress in future climate, but also stresses the role of ocean dynamics in shaping the distribution of the dynamical regimes through the role of other input features. THOR scales readily, and can elucidate the dynamical features in ocean models of similar horizontal resolution.

Assessment reports such as the IPCC rely on intercomparisons of models such as the CMIP6 ensemble, that largely have ocean components of 1° resolution. The spread between projections of features such as the AMOC in CMIP6 (Weijer et al., 2020) highlight the need to understand its source. THOR could help understand both the dynamical source and also guide model development. Assessing the source of the spread in AMOC weakening in CMIP5 points to a number of dynamics, and the weakening may have been underestimated (Saba et al., 2016). One feature in CMIP5 models impacting AMOC was a differing cold biases in the entire Northern Hemisphere (C. Wang et al., 2014). The deep convection was also largely too far south and reaching too deep (Heuzé, 2017). Such process variability would be apparent using THOR. Structural model errors are a key source of the spread of projections of AMOC, and the dynamics can partially be seen as emerging from these. Identification of processes form a drive to guide climate model development using process oriented diagnostics (Maloney et al., 2019). THOR could be used as such a process identification method, diagnosing specific features leading to structural model errors. Because THOR is scalable and uses only few input fields, it could provide a rapid and comprehensive analysis of process representation and identification of gaps in phenomena.

The dynamical regimes identified using THOR demonstrate clear spatial changes under different climate forcing. THOR by construction, relies on the identification of dynamical regimes on the basis of those found in ECCO. This implicitly assumes that ECCO represents six dynamical regimes that will only be spatially different in location and expanse in a different model and under different climate forcing. The highly robust nature of the dynamical regimes identified in ECCO in the first part of THOR lends confidence to this underlying assumption, as very large changes in the basic configuration of ocean dynamics would be necessary to arrive at a novel dynamical regime (Sonnewald et al., 2019). However, THOR should only be applied to similar horizontal resolutions. If the horizontal resolution of the ocean model changes significantly, for example to eddy-resolving, more physical processes can be explicitly represented

and the clear distinction between regimes could erode. Another assumption made in THOR is the use of a depth integral in the dynamical regime identification. This implies, for example, that nonlocal changes in deep advection in bottom currents could be missed. A caveat related to what a shift in mechanisms would lead to in terms of driving the AMOC strength, is if a thermohaline framework used or a mixing/wind driven framework. If a strictly thermohaline framework were used (similar to a heat engine) they would be driving, rather than governing forces (Griffies et al., 2015; Kuhlbrodt et al., 2007; Wunsch & Ferrari, 2004). Note that a weaker/stronger AMOC would exhibit the same changes in the highlighted mechanisms.

To be truly appropriate for application to the physical sciences, the source of skill from ML should be transparent. At the root of this need is a necessity that the ML is based on something physical and not random chance (Balaji, 2020; Irrgang et al., 2021; Sonnewald et al., 2021). The interpretability and explainability of THOR comes from a combination of the equation transform at its core (Sonnewald et al., 2019), the engineering of its input features, and the LRP explanation of its predictive skill. First, the equation transform reduces complicated full model data to a form that enables identified regimes to be dynamically interpretable (Figures 2c and 3). Second, the knowledge of the dominant terms provides a rationale for the engineering of input features, as these form a proxy of the key dynamical drivers. Third, the LRP provides detailed information about the source of the predictive skill. The explanation of predictive skill was seen as crucial to THOR, but importantly restricted the NN architecture available. For example, the Ensemble MLP did not encode explicit mathematical formulations that theory suggests could be helpful, such as a the Jacobian operator. The structural changes needed would preclude the LRP application. This is because the original LRP was designed for regular feed forward MLPs and not bilinear MLPs (comprising two paths whose outputs are multiplied). This is an example of NN development that would be meaningful for ML applications to the physical sciences, that to the authors' knowledge are lacking as of date. Interpreting the relevances with this additional information could make the sources of the skill less abstract. Other methods for AFA such as LIME are also available, as well as SHAP based on game theory (Lundberg & Lee, 2017; Ribeiro et al., 2016). It should be noted, that many perturbation-based methods that exist to explain the predictive skill of "black-box" ML models are still not robust to local perturbations on inputs (Alvarez-Melis & Jaakkola, 2018). The ideal desired outcome of an AFA method is that the feature attribution will remain similar when input features surrounding the sample being explained are perturbed slightly, with no change in the NN prediction. Highlighting their brittle nature, techniques such as LIME or SHAP can also be tweaked to intentionally lead to misleading interpretations (Slack et al., 2020). The LRP method which is not perturbation-based was deemed most reliable for the present work, also because it does not treat the NN completely as a "black-box". This is because unlike LIME, it has access to weights and biases of the NN.

For the interpretability of THOR, the LRP method was deemed appropriate as it was found to give physically plausible results. This raises the question of whether interpretable techniques are meant to confirm a priori held notions or gain new insights. The main purpose here was to use LRP as a means to confirm that what the NN learned is not due to chance. Toward gaining new insight, making methods of AFA more robust to local perturbation and improving the performance would be highly beneficial, particularly for application within oceanography and the broader physical sciences. Gaining an appreciation of the specific features that gave rise to the predictive skill can importantly help to avoid underspecification (D'Amour et al., 2020) (Appendix G). Underspecification is a problem where several ML models perform well, but may not for example represent the pertinent physics and therefore fail if tested on data beyond the scope of the initial testing. Such underspecification is particularly important to avoid for example in a climate model parameterization setting where data with which to validate results are not available. The combination of ML techniques and feature engineering that forms the base of THOR is generally applicable, and could serve as a blueprint for other studies.

Future work will assess the variability in key AMOC drivers in further CMIP6 models, with identification of structural model errors in focus. A further goal is to assess other ocean areas key to climate, such as the Southern Ocean where deep waters are brought to the surface closing the loop of water mass transformation.

Appendix A: Numerical Models: ECCO and GFDL-ESM4

The ECCOv4 (Forget et al., 2015) global state estimate (Wunsch & Heimbach, 2013) has a nominal 1° resolution. A least squares with Lagrange multipliers approach is used to obtain observationally adjusted initial and boundary conditions as well as internal model parameters. This results in a free-running version of the MIT General Circulation Model (MITgcm, (Adcroft et al., 2004)) that has been optimized to track observations. Adjoint methods are used to create the state estimate, allowing both the optimization to data, but also the closure of the momentum budget. This is because "nudging" terms that are often applied to bring models closer to observations are not needed. This budget closure is seen as an important component of the success of step 1 in THOR. The overall meridional overturning ($\Psi_{z\theta}$) from Figure 3 is defined as:

$$\Psi_{z\theta}(\theta, z) = -\int_{-H}^{z}\int_{\phi_2}^{\phi_1} v(\phi, \theta, z')d\phi dz',$$

where z is the relative level depth and v is the meridional (north-south) component of velocity. For the regimes, the relevant velocity fields were then used. A positive $\Psi_{z\theta}$ signifies a clockwise circulation, while a negative $\Psi_{z\theta}$ signifies an anticlockwise circulation.

The GFDL-ESM4.1 model (Dunne et al., 2020) consists of the AM4.0 atmosphere model, at ~1° resolution, with 49 vertical levels of comprehensive, interactive chemistry and aerosols (including aerosol indirect effect) from precursor emissions. The OM4 MOM6-based ocean model is used, with a resolution of 1/2°, 75 vertical levels, and a hybrid pressure/isopycnal vertical coordinate system. The ESM4.1 uses the SIS2 sea ice model, with radiative transfer and C-grid dynamics for compatibility with MOM6. The land model is LM4.1, that has vegetation dynamics tiles that explicitly treat plant age and height structure and soil microbes, with daily fire, crops, pasture, and grazing. The COBALTv2 ocean biogeochemical component represents the ocean ecology and biogeochemistry. The dust and iron cycling between land-atmosphere and ocean is fully interactive. The AMOC in depth space (depth + latitude) is available for ESM4.1, another reason why it is a good test case for THOR. Figure A1 illustrates the weakening of the AMOC, as expected, where the historical is strongest and the 1pctCO_2 shows a weakening that is more pronounced in the abrupt4xCO_2 scenario.

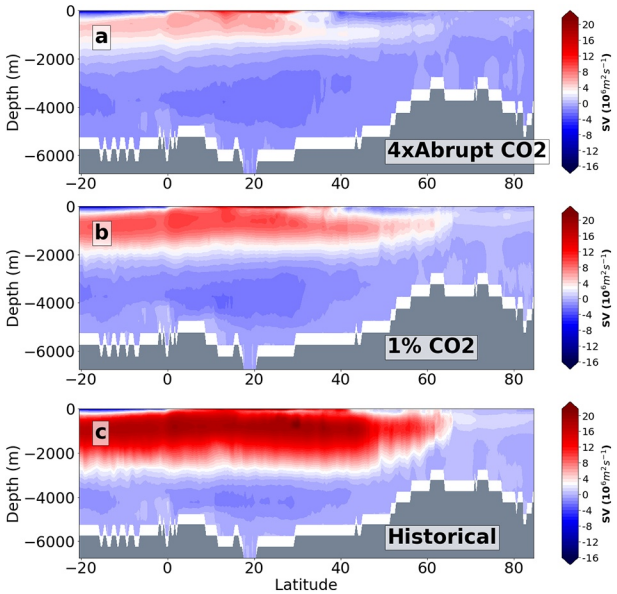

Figure A1. The Atlantic Meridional Overturning Circulation (AMOC) in ESM4.1. (a) The abrupt4xCO_2 scenario where the AMOC is weakest, (b) the 1pctCO_2 scenario has a slightly stronger AMOC and (c) shows the modern AMOC. The gray area shows the bathymetry.

Appendix B: Equation Transform

To arrive at the five dimensional field from the full 3D model fields, a closed momentum budget was used. The discussion below is adapted from Sonnewald et al. (2019). The momentum and continuity equations of the ocean are seen as a thin shell sitting on a rotating sphere:

$$\partial_t \mathbf{u} + f\mathbf{k} \times \mathbf{u} = -\frac{1}{\rho_0}\nabla_h \mathrm{p} + \frac{1}{\rho_0}\partial_z \tau + \mathbf{a} + \mathbf{b}, \partial_z \mathrm{p} = -\mathrm{g}\rho \tag{B1}$$

$$\nabla_h \cdot \mathbf{u} + \partial_z w = 0, \tag{B2}$$

Pressure, gravity, density, and vertical shear stress are p, g, ρ, and τ, respectively, with ρ_0 the reference density; the three-dimensional velocity field $\boldsymbol{v} = (u, v, w) = (\boldsymbol{u}, w)$; the gradient $\nabla = (\nabla_h, \partial z)$; the unit vector is denoted $\mathbf{k}$; planetary vorticity is a function of latitude θ in $f\boldsymbol{k} = (0, 0, 2\Omega\sin(\theta))$; the viscous forcing from vertical shear is $\partial_z \tau$; the nonlinear torque is $\boldsymbol{a}$, and the horizontal viscous forcing $\boldsymbol{b}$ includes subgrid-scale parameterizations. Under steady state, the vertical integral from the surface $z = \eta(x, y, t)$ to the water depth below the surface $z = H(x, y)$ is

$$\beta V = \frac{1}{\rho_0}\nabla \mathrm{p}_b \times \nabla H + \frac{1}{\rho_0}\nabla \times \boldsymbol{\tau} + \nabla \times \mathbf{A} + \nabla \times \mathbf{B} \tag{B3}$$

where $\nabla \boldsymbol{U} = 0$, $\boldsymbol{U} \cdot \nabla f = \beta V$, the bottom pressure is denoted p_b, $\mathbf{A} = \int_H^\eta \mathbf{a} dz$, and $\mathbf{B} = \int_H^\eta \mathbf{b} dz$. The curl operator $\nabla\times$ produces a scalar, that represents the vertical component of the operator. The left-hand side of Equation B3 is the planetary vorticity advection term, while the right-hand side of Equation B3 is the bottom pressure torque (BPT), the wind and bottom stress curl, the nonlinear torque, and the viscous torque, respectively. The five terms in Equation B3 constitute the dynamical drivers/terms are the fundamental sources of depth integrated (barotropic) vorticity: on the LHS, the advection of planetary vorticity, on the RHS from left to right, bathymetric interactions through BPT, the wind and bottom stress curl, curl of nonlinear interactions between terms and the lateral viscous dissipation from within the ocean interior.

The subgrid-scale parameterization introduces a torque, which is included in the viscous torque term. Nonlinear torque is composed of three terms:

$$\nabla \times \mathbf{A} = \nabla \times \left[\int_{-H}^{\eta} \nabla \cdot (\mathbf{uu}) dz\right] + [w\zeta]_{z=H}^{z=\eta} + [\nabla w \times \mathbf{u}]_{z=H}^{z=\eta} \tag{B4}$$

where $\boldsymbol{uu}$ is a second-order tensor. The right-hand side of Equation B4 represents the curl of the vertically integrated momentum flux divergence, the nonlinear contribution to vortex tube stretching, and the conversion of vertical shear to barotropic vorticity. Horizontal viscous forcing includes that induced by subgrid-scale parameterizations. In Sonnewald et al. (2019), twenty-year averaged fields (1992–2013) are used after a Laplacian smoother is applied, with an effective averaging range of three grid cells.

Appendix C: Independent Validation of the Ensemble MLP With Model CESM1 POP2

An independent test of THOR with an unseen model, but where the terms in Equation B3 are at hand is a stringent test of underspecification. Figure C1 illustrates the independent validation with CESM1 POP2 (1948–2011). The ocean component of the CESM1 is the Parallel Ocean Program version 2 (POP2 (Smith et al., 2010)), in the coupled ocean-sea ice configuration (Gent et al., 2011, October 01, 2011). CESM1 POP2 is a non-eddy-resolving version at nominal 1° horizontal resolution with 60 vertical levels. The horizontal resolution is comparable to ECCO, making CESM1 POP2 a good candidate for comparison.

In CESM1 POP2, the balances of the terms is seen to show similar balances to those in ECCO, with two Sverdrupian regimes and two topographic Sverdrupian regimes of opposing signs. The area recognized as the MD regime again has little area averaged torque. There is a notable NL regime presence 40–50°N

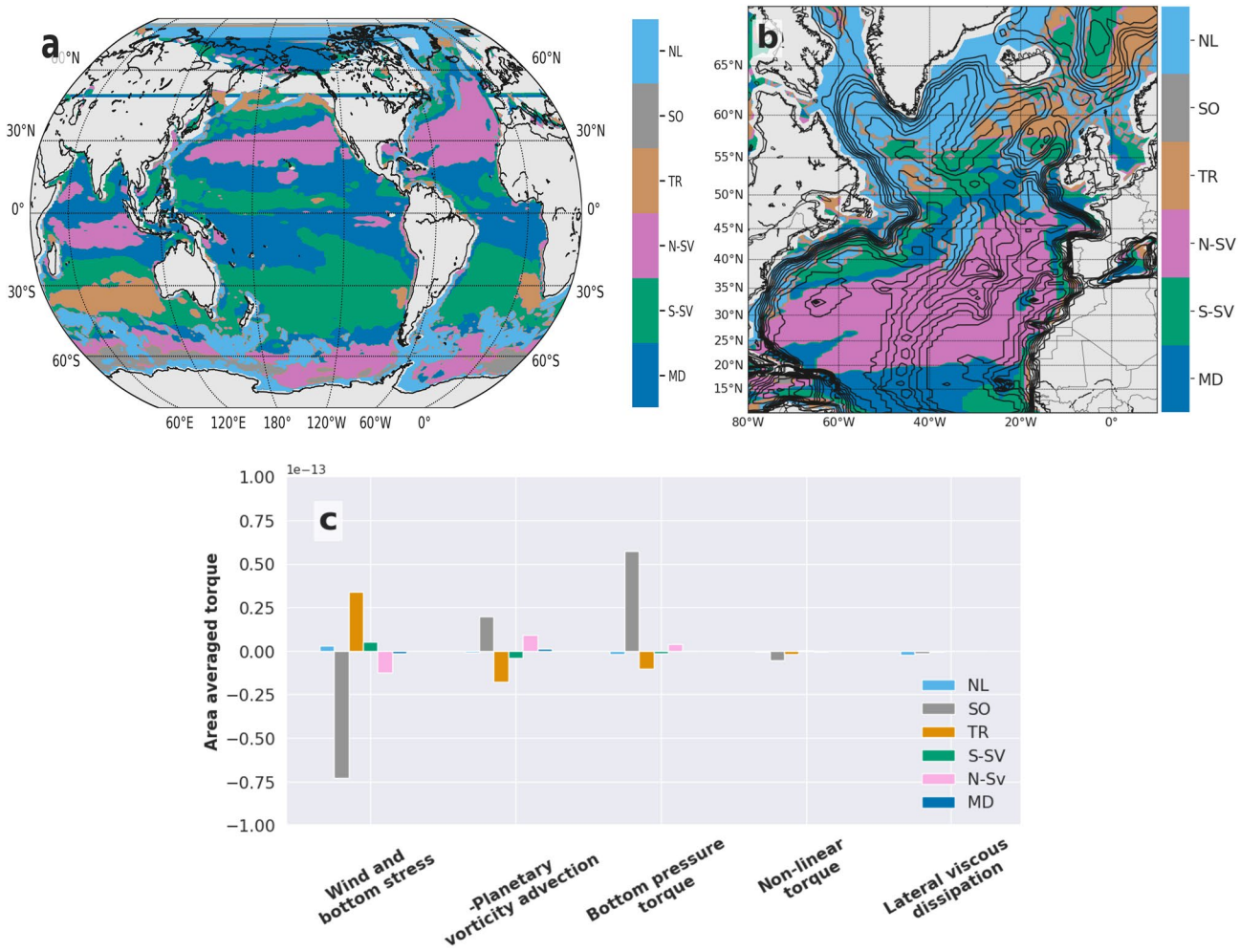

Figure C1. Validation with CESM1 POP2. THOR applied to CESM1 POP2, where the terms used for initial dynamical regime identification are available. The area averaged balances in the predicted dynamical regimes exhibit similar fundamental balances as Figure 2b.

along the Mid-Atlantic ridge, potentially due to bathymetric interactions. Relevance maps confirm the importance of such features. Of particular interest is the difference in parameterisations that impact deep water formation. In CESM1 POP2, dense waters are injected from the Nordic Seas into the abyss directly. The longer integration could mean that the model is not entirely equivalent to ECCO, but still represents a "modern" climate. There is a difference in the NL regime, but this is not surprising as the regime is influenced by cancellations.

Appendix D: Performance Metrics for Ensemble MLP

Figure D1 shows the confusion matrices of the true/predicted classes using the Ensemble MLP (step B in Figure 1) for training data and validation data, while detailed classification performance metrics are reported in Table D1. Together they show that our ML classifier reaches and average F-score of 0.84 (an ideal F-score is 1). Individually, the Ensemble MLP reaches a good F-score (≥0.8) for all the classes except the NL class, which is unsurprisingly the hardest to classify.

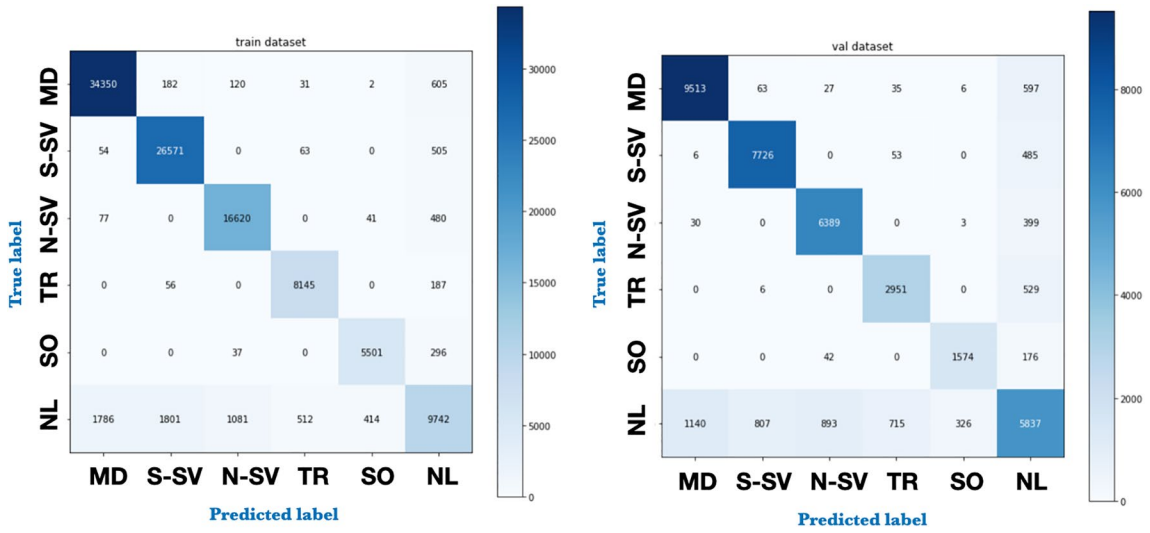

Figure D1. Confusion matrices for the Ensemble multilayer perceptron classification on train (left) and validation (right) data.

Appendix E: LRP: Spatial Representation of Interpretability/Explainability Using

The spatial maps of the interpretability (Figure E1) show intricate detail of what contributes to the Ensemble MLP learning. The mapping between feature relevance and the barotropic vorticity equation terms is not direct, but through the equation transform component of step A in THOR, the relevance maps can be evaluated in terms of an interpretation based on known physics. What is meant by the mapping not being direct is that there is important information also in what the Ensemble MLP found unhelpful. It is also interesting to note that the role of the bathymetry is evident in all but the wind stress curl feature. Bathymetry here is distinct from the feature H. For the feature H, both the latitudinal and longitudinal gradients show equivalent patterns in longitude and latitude. The bathymetry also seems largely absent from the η feature importance overall. It is interesting that the N-SV regime has positive relevance to the west of the Mid-Atlantic ridge, and negative to the east. Overall, the spatial relevances reveal that the standard deviations of the relevances can serve as a proxy for rich spatial structure.

The dominant positive relevance of the wind stress could suggest that this feature alone would give good predictions of five of the six dynamical regimes. The nature of the term balances in for example the N-SV and TR regimes demonstrate the added relevance that the other features add. Referring back to the terms in the dynamical regimes, it is possible to deconstruct what information is added by the η and H term, and their latitudinal and longitudinal gradients, that adds valuable information to the MLP.

Table D1
Classification Report of the Performance of the Ensemble MLP on the Validation Set

Class	Precision	Recall	F-score	Support
MD	0.89	0.93	0.91	10,241
N-SV	0.90	0.93	0.92	8,270
S-SV	0.87	0.94	0.90	6,821
TR	0.79	0.85	0.82	3,486
SO	0.82	0.88	0.85	1,792
NL	0.73	0.60	0.66	9,718
Average	0.83	0.85	0.84	40,328

Abbreviations: MD, momentum driven; MLP, multilayer perceptron; NL, nonlinear; N-SV, Northward-Sverdrupian; SO, Southern Ocean; S-SV, Southward-Sverdrupian; TR, transitional.

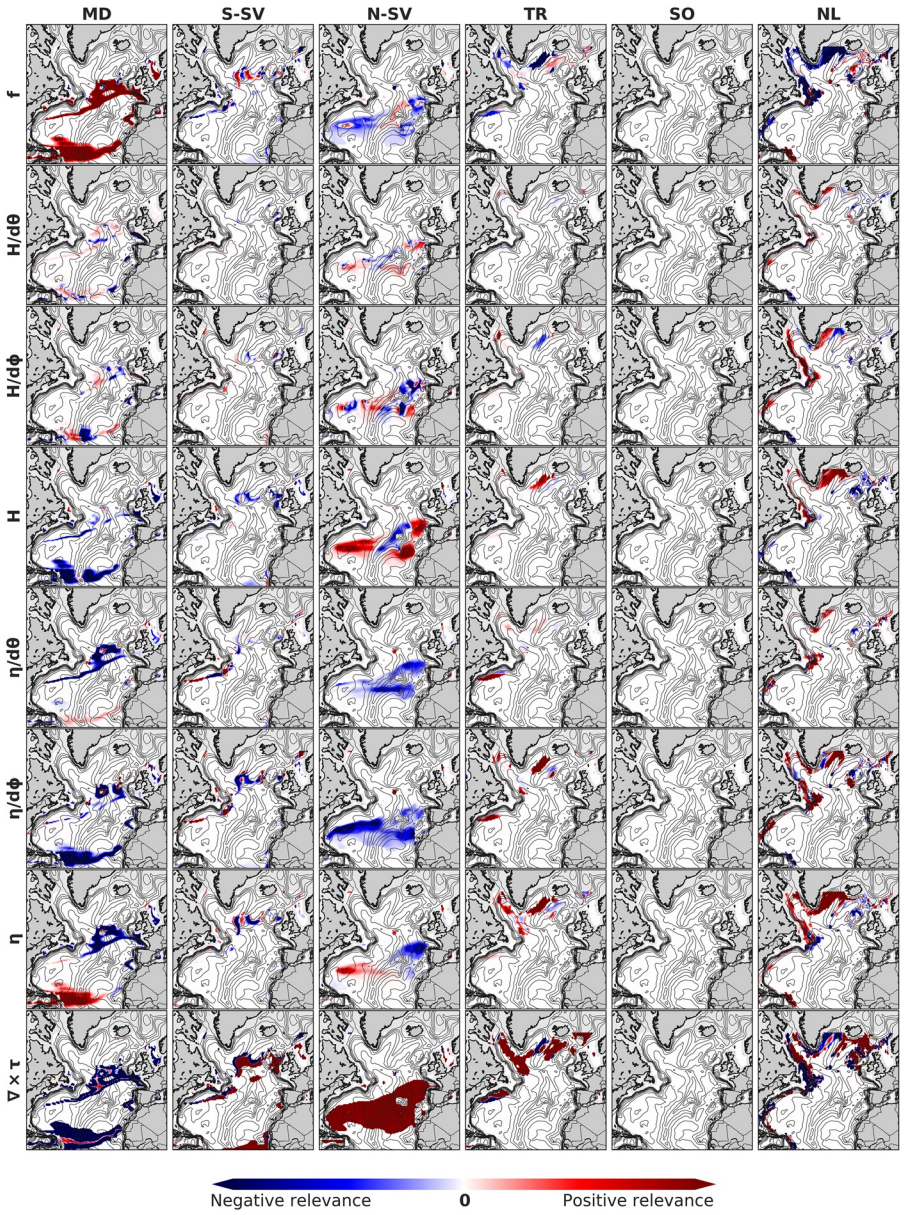

Figure E1. The spatial relevance maps. The dynamical regimes (columns) and the input features (rows) illustrating the contributions to the skill of the Ensemble multilayer perceptron (MLP). The relevances are averaged for each point (lat, lon) over the Ensemble MLP.

Highlighting these expected physics, the N-SV regime serves as a good example. One expects the wind stress to be a dominant feature, which is confirmed. The interactions with bathymetry through the bottom pressure torque term are recognised as not being a dominant component of the regime. The negative contribution of H where the mid-Atlantic ridge is present, but positive contribution beyond this therefore confirm what one expects. Thus when there is a large change in the magnitude of H this contributes negatively to learning. A similar expected feature is the importance of f in the MD regime. This regime can be seen as largely geostrophic, and thus f is expected to be influential. What is more surprising for the MD term is for example that the wind stress largely contributes negatively apart from in a narrow strip in the center of the regime sections (one in the equatorial area and one constituting the North Atlantic Current area). There is also a clearly positive relevance associated with η in the equatorial region of the MD regime, which turns largely negative further north. Note that the latitude and longitude information

was not included. The NL regime is an example where the relevance maps from the North Atlantic clearly show that the *f* feature is unhelpful. This is interesting because in all other input features, the relevance maps results are less clear.

While it is encouraging that expected feature importance emerged form the Ensemble MLP relevance assessment, there could be more to learn studying the LRP results. An exciting avenue for future work will be a deeper analysis into the exact mappings of the input variable ranges onto the regimes. While the impact of bathymetry is more straightforward to interpret, subtleties within other terms, such as the gradients of bathymetry could prove fruitful in terms of potentially even gaining a deeper understanding of the regimes themselves. This was similarly suggested by Toms et al. (2020). In this sense, it could be possible to develop a feedback to the theoretical components of THOR, and use the relevance maps to gain a better understanding of the unifying features. Such approaches are discussed further in Sonnewald et al. (2021).

Appendix F: LIME: Alternative Interpretability/Explainability Rendering

One other popular method to explain black box ML models is using Local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016). As an example of how LIME functions, suppose you have a sample of data *s* for which you are seeking an explanation of its prediction by the ML model. LIME is based on the idea of generating new samples in the "neighborhood" of the input features of *s* and passing them through the ML model to get their predictions. Then an interpretable model is fit to the results, such as a sparse linear model where the new samples are weighted by their distance to *s*.

Here, the new neighborhood data set created by LIME is obtained by perturbing the eight input features individually by drawing from a normal distribution whose mean and standard deviation are calculated from the feature. This application of LIME is standard, using the code written by the original authors of LIME https://github.com/marcotcr/lime

The results of applying LIME to the Ensemble MLP are displayed in Figure F1. Overall, LIME results are very similar to LRP results, with similar spatial patterns seen. Particularly the NL regime exhibited very similar relevances between the LRP and LIME assessments, which was unexpected as this dynamical regime posed the biggest classification challenge. A main difference is that the negative relevances that appeared using the LRP method appear very weak using LIME, or as having no relevance. An example of this is the mid-Atlantic ridge region in the N-SV regime for the input features associated with depth. These were largely a source of negative relevance in the LRP application, but appear neutral using LIME. Another difference is that there is largely only positive relevance coming from the $\nabla \times \tau$. The similarity between the LIME and LRP application is encouraging. Discrepancies could be due to the statistical assumptions surrounding the feature perturbation, and fitting of a linear model to obtain the relevances. As such, the difference between the LRP and LIME methods could suggest that assuming an underlying Gaussian distribution for feature space exploration as well as linearity are inappropriate. The LRP method, while having its own shortcomings, does not make such assumptions. The LRP method, which is not perturbation-based, was deemed most reliable for the present work, also because it does not treat our NN as a complete black-box, i.e. it has access to weights and biases of the NN. The LRP explanation was also more appealing because we found it to be plausible physically. This does also raise the question of whether interpretable techniques are meant to confirm our a priori beliefs or gain new insight. Here, we use LRP as a means to confirm that what the NN learned is not due to chance. Making methods of AFA more robust to local perturbation and improving the performance would be highly beneficial to their wider application, but particularly for application within oceanography and the larger physical sciences.

SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) is another example of an AFA that is gaining popularity. It is an attempt to unify the field on interpretable machine learning and answer the question of when one method is preferable over another. Authors of the seminal paper link established and theoretically accepted concepts from coalitional game theory such as the use of Shapely values, with additive feature attribution methods such as LRP and LIME and other techniques. A notable example is the

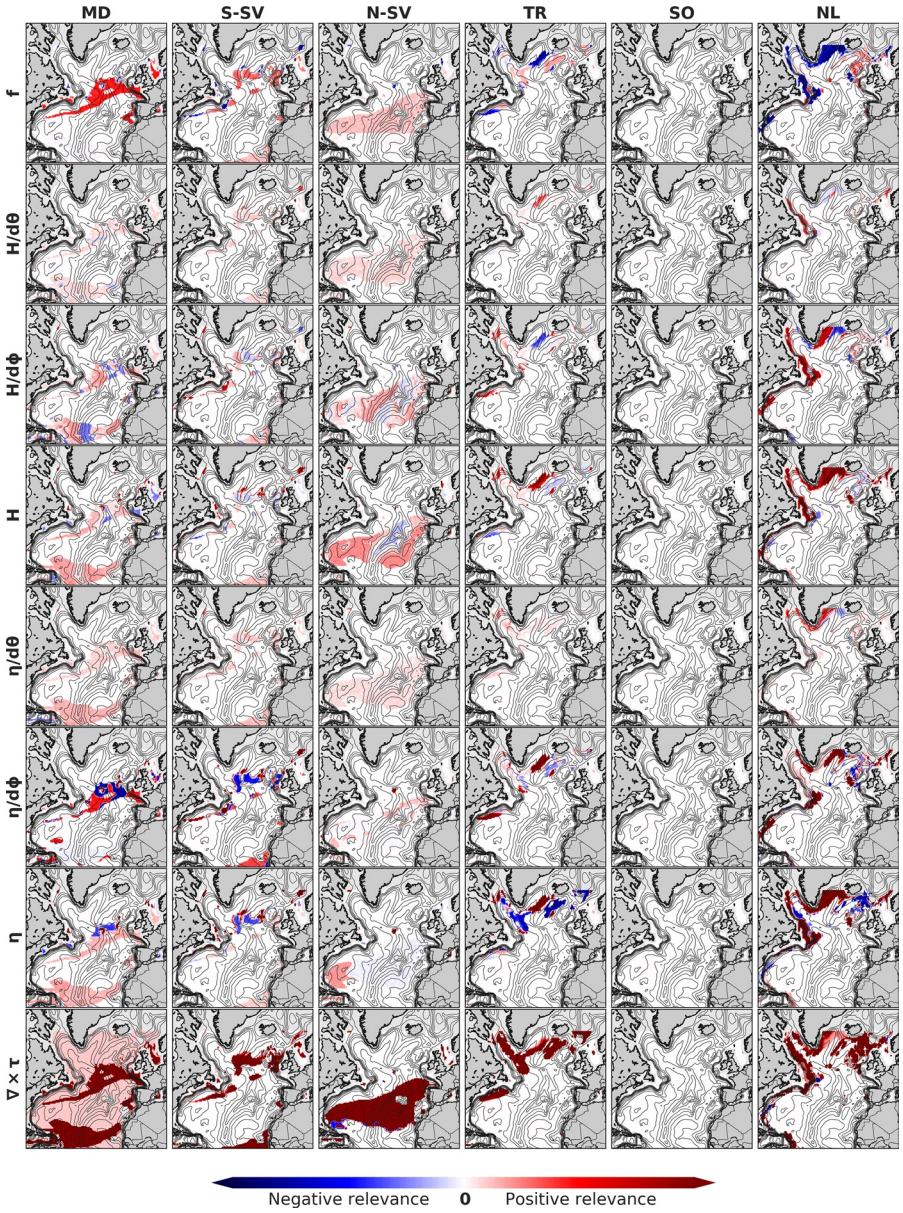

Figure F1. The spatial relevance maps computed using Local interpretable model-agnostic explanations. The dynamical regimes (columns) and the input features (rows) illustrating the contributions to the skill of the Ensemble multilayer perceptron (MLP). The relevances are averaged for each point (lat, lon) over the Ensemble MLP, see Figure 5.

introduction of the KernelSHAP method that uses Shapely values in a LIME context where the surrogate model is a linear regression model. While KernelSHAP is a model agnostic technique that can be used for neural networks, authors developed DeepSHAP where use shapely values and DeepLIFT (a technique similar to LRP) to leverage the compositional nature of deep neural networks and optimize the computational performance. While SHAP has potential, it is also an application very sensitive to its hyperparameters, and exploring SHAP is part of future work.

Appendix G: THOR in the Context of Explainability and Overcoming Underspecification

The application of the LRPs highlighted the importance of designing the MLP using ensemble training. This is because of the stochastic element in the MLP training, which need not always arrive at a global minima, implying both that the LRP interpretation can be skewed and that predictions from an MLP trained only once would likely be lacking. The relevances of the individual trainings revealed ambiguity beyond the very strong features for example the wind stress importance and the role of the bathymetry, seen in the spatial representations of the relevances. However, the nature of the LRP could also exaggerate the impact of the stochastic aspect of the MLP training. Note that the LRP used is not designed specifically for geoscientific applications. The heatmapping procedure that the LRP applies satisfies certain predefined properties, that are then stored as "relevances". What definitions would be optimal for oceanographic applications is a topic of future study.

In principle, the training of a NN, and largely applications of ML, are an optimization problem. For NNs, the optimization problem has connotations for the rate of learning, but also for the way the NN is able to explore, and "fit" itself, to the parameter space imposed by its architecture and the data. For geoscientific data, the nature of the optimization could have significant impact, as finding an appropriate global minima is more complicated than for example for an Ising model. A danger is that the NN model "underspecifies", meaning that the given data leave enough ambiguity as to the true global minimum, that the NN model would not be generally applicable to the problem at hand (D'Amour et al., 2020). If an NN model is underspecified, it will not have "learned" a true representation of the underlying system (e.g., an adequate representation of ocean physics) with which to make predictions. THOR's inherent interpretability could be a way to address this, as one first reduces complexity by creating a categorical problem for the NN, and then also is able to ensure generality by assessing the explanation of the prediction skill. See Figures H1 and H2 for spatial details.

Appendix H: Detailed Maps THOR ESM Predictions for 1pctCO_2 and Abrupt4xCO_2 Scenarios

The change in area occupied by the dynamical regimes from the historical to the 1pctCO_2 (Figure H1) and abrupt4xCO_2 (Figure H2) scenarios. The figures serve to demonstrate the changes taking place in more detail than Figure 4. The figures demonstrate where a regime has displaced another (left columns in H1 and H2), or where the regime has been replaced by another (right columns in H1 and H2). For the North Atlantic Current, the eastwards shift is seen by the area occupied by the MD regime in the 1pctCO_2 scenario having expanded into the NL and S-SV areas in Figure H1a. Historical expanses to the north are being occupied by the S-SV regime (Figure H1b). Similarly, the North Atlantic Current shift in the abrupt4xCO_2 scenario is seen particularly also in Figure H2b where large expanses of the S-SV regime have moved in to the area occupied by the S-SV regime in the historical scenario. For the 1pctCO_2 scenario in Figure H1, the shift in the Gulf Stream location can be seen in Figure H1c, where the TR regime has been displaced, and in Figure H1d where the MD regime has replaced the S-SV regime. This is similarly visible in abrupt4xCO_2 scenario for Figure H2c and H2d, but the areal expanse of the MD replacement is larger signifying a greater shift. The change in the TR and NL regime are not very large in the sub-polar region associated with deep water formation for the 1pctCO_2 scenario (Figure H1g, H1h, H1k and H1l), likely because the areas that were shifting were not deemed significant, and still could be influenced by natural variability. In the abrupt4xCO_2 scenario, a marked shift is observed (Figure H1g, H1h, H1k and H1l), where the TR regime partially displaces the NL regime in the sub-polar region. In H1e and H1f little change is observed.

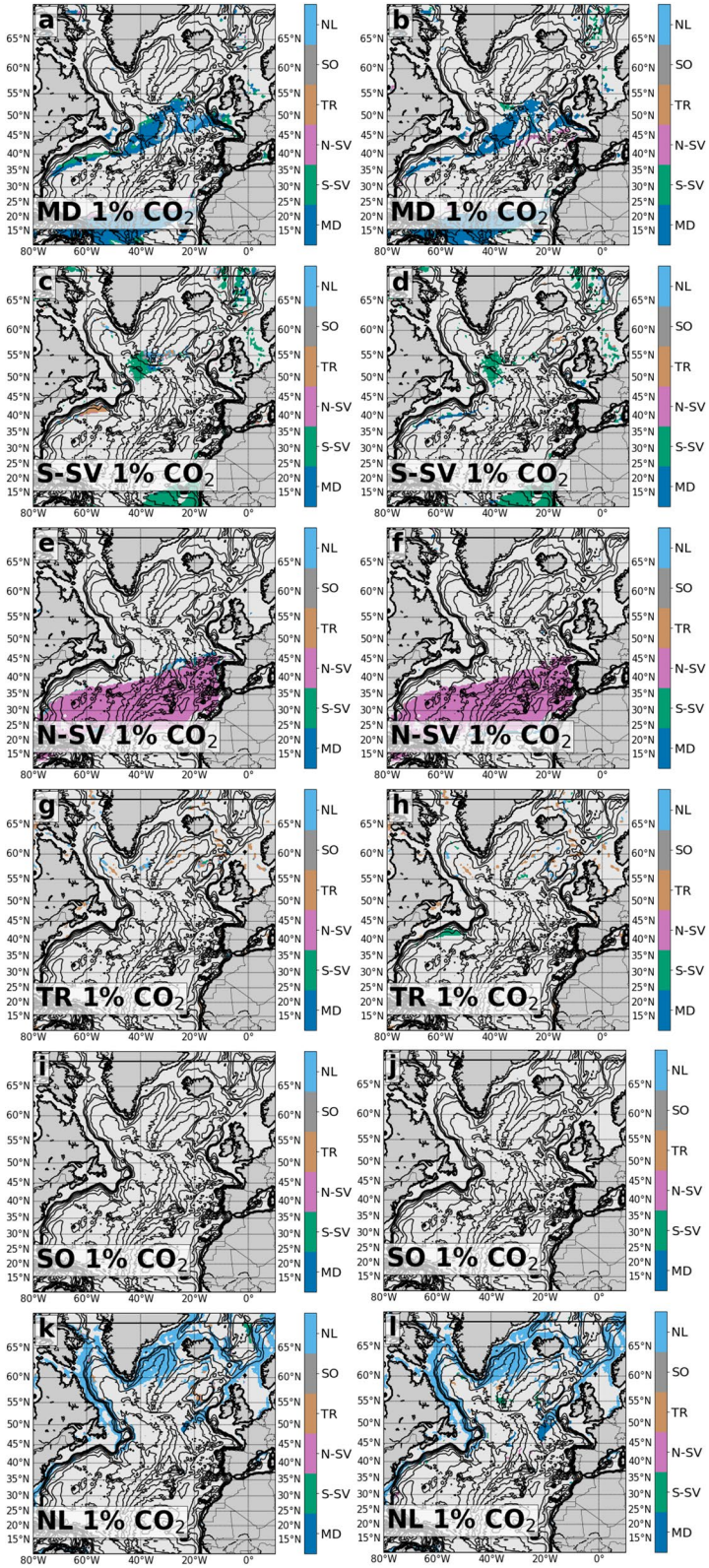

Figure H1. Maps of expanse addition and loss by regime for scenario. The left column shows the area where the regime expanded into for the 1pctCO_2 scenario, where the color that is different from the regime in question shows what regime was replaced. The right column shows the regime expanse for the historical scenario, where colors different from the regime in question signify that the area was displaced by a different regime of the respective color.

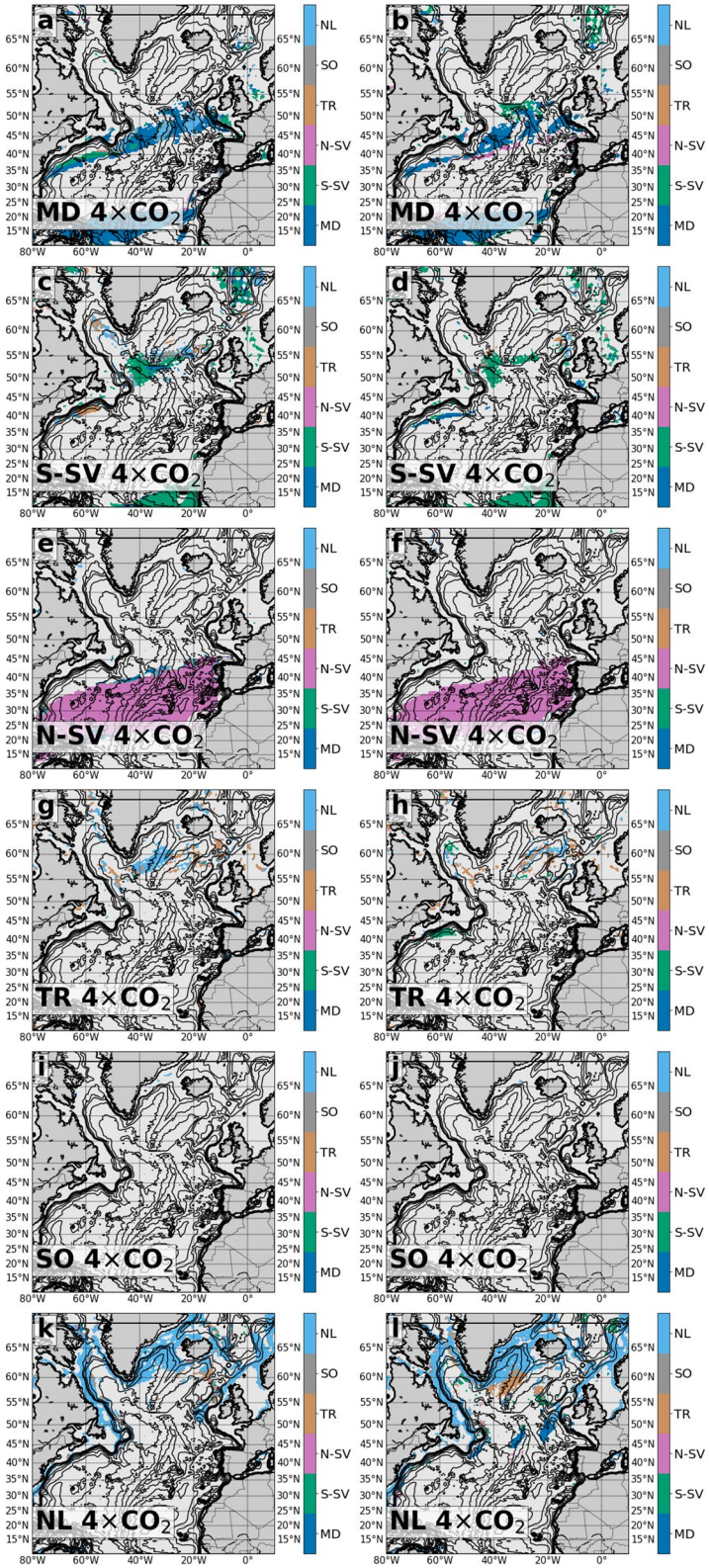

Figure H2. Maps of expanse addition and loss by regime for scenario abrupt4xCO_2. As for Figure H1.

Data Availability Statement

ESM4.1 data available at https://esgf-node.llnl.gov/search/cmip6/. Code is available through GitHub https://github.com/maikejulie/DNN4Clim and fully reproducible also within the Amazon Cloud. ECCOv4r3 data available at: https://ecco-group.org/products.htm.

Acknowledgments
The authors are grateful for conversations with a number of colleagues including A. Adcroft, V. Balaji, S. Griffies, J. Le Sommer, S. Yeager and R. Zhang, and assistance in acquiring data from S. Nikonov, A. Radhakrishnan and K. Rand. A. Radhakrishnan offered assistance with the Amazon Cloud infrastructure (grant from ASDI), and CMIP infrastructure (Balaji et al., 2018). The Earth System Grid Federation portal (ESGF) was used to get the CMIP 6 data. Y. Cho offered assistance with Figure 4. The authors also thank the two anonymous reviewers. MS funding: Cooperative Institute for Modeling the Earth System, Princeton University, under Award NA18OAR4320123 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of Princeton University, the National Oceanic and Atmospheric Administration, or the U.S. Department of Commerce. RL funding: The Make Our Planet Great Again (MOPGA) funding to Project Hermès from the Agence Nationale de Recherche under the "Investissements d'avenir" program with the reference ANR-17-MPGA-0010.

References

Adcroft, A., Hill, C., Campin, J.-M., Marshall, J., & Heimbach, P. (2004). Overview of the formulation and numerics of the MIT GCM. In *Proceedings of the ECMWF seminar series on numerical methods, recent developments in numerical methods for atmosphere and ocean modelling* (pp. 139–149).

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., et al. (2019). Innvestigate neural networks! *Journal of Machine Learning Research*, *20*(93), 1–8. Retrieved from http://jmlr.org/papers/v20/18-540.html

Alvarez-Melis, D., & Jaakkola, T. S. (2018). *On the robustness of interpretability methods*. https://arxiv.org/abs/1806.08049

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Balaji, V. (2020). Climbing down charney's ladder: Machine learning and the post-dennard era of computational climate science. *Philosophical Transactions A*, *379*, 20200085. http://doi.org/10.1098/rsta.2020.0085

Balaji, V., Taylor, K. E., Juckes, M., Lawrence, B. N., Durack, P. J., Lautenschlager, M., et al. (2018). Requirements for a global data infrastructure in support of CMIP6. *Geoscientific Model Development*, *11*(9), 3659–3680. https://doi.org/10.5194/gmd-11-3659-2018

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., & Samek, W. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. In *International conference on artificial neural networks* (pp. 63–71). https://doi.org/10.1007/978-3-319-44781-0_8

Bingham, R. J., & Hughes, C. W. (2009). Signature of the Atlantic meridional overturning circulation in sea level along the east coast of North America. *Geophysical Research Letters*, *36*, L02603. https://doi.org/10.1029/2008gl036215

Böning, C. W., Scheinert, M., Dengg, J., Biastoch, A., & Funk, A. (2006). Decadal variability of subpolar gyre transport and its reverberation in the north Atlantic overturning. *Geophysical Research Letters*, *33*, L21S01. https://doi.org/10.1029/2006gl026906

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., & Knutti, R. (2020). Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth System Dynamics*, *11*(4), 995–1012. https://doi.org/10.5194/esd-11-995-2020

Buckley, M. W., & Marshall, J. (2016). Observations, inferences, and mechanisms of the Atlantic meridional overturning circulation: A review. *Reviews of Geophysics*, *54*(1), 5–63. https://doi.org/10.1002/2015rg000493

Callaham, J. L., Koch, J. V., Brunton, B. W., Kutz, J. N., & Brunton, S. L. (2021). Learning dominant physical processes with data-driven balance models. *Nature Communications*, *12*, 1016.

Chapman, C., & Charantonis, A. A. (2017). Reconstruction of subsurface velocities from satellite observations using iterative self-organizing maps. *IEEE Geoscience and Remote Sensing Letters*, *14*(5), 617–620. https://doi.org/10.1109/LGRS.2017.2665603

Cheng, W., Chiang, J. C. H., & Zhang, D. (2013). Atlantic Meridional Overturning Circulation (AMOC) in CMIP5 Models: RCP and Historical Simulations. *Journal of Climate*, *26*(18), 7187–7197. https://doi.org/10.1175/JCLI-D-12-00496.1

Chollet, F., et al. (2015). *Keras*. https://keras.io

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314. https://doi.org/10.1007/bf02551274

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., et al. (2020). *Underspecification presents challenges for credibility in modern machine learning*. https://arxiv.org/abs/2011.03395

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., et al. (2020). The GFDL Earth System Model Version 4.1 (gfdl-esm 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2019MS002015. https://doi.org/10.1029/2019MS002015

Eyring, V., Bony, S., Meehl, G., Senior, C., Stevens, B., Ronald, S., & Taylor, K. (2015). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organisation. *Geoscientific Model Development Discussions*, *8*, 10539–10583. https://doi.org/10.5194/gmdd-8-10539-2015

Eyring, V., Cox, P., Flato, G., Gleckler, P., Abramowitz, G., Caldwell, P., et al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*, 102–110. https://doi.org/10.1038/s41558-018-0355-y

Forget, G., Campin, J.-M., Heimbach, P., Hill, C. N., Ponte, R. M., & Wunsch, C. (2015). Ecco version 4: Ann integrated framework for non-linear inverse modeling and global ocean state estimation. *Geoscientific Model Development*, *8*(10), 3071–3104. https://doi.org/10.5194/gmd-8-3071-2015

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., et al. (2011). The community climate system model version 4. *Journal of Climate*, *24*(19), 4973–4991. https://doi.org/10.1175/2011jcli4083.1

Griffies, S. M., Winton, M., Anderson, W. G., Benson, R., Delworth, T. L., Dufour, C. O., et al. (2015). Impacts on ocean heat from transient mesoscale eddies in a hierarchy of climate models. *Journal of Climate*, *28*(3), 952–977. https://doi.org/10.1175/jcli-d-14-00353.1

Hamman, J., Rocklin, M., & Abernathy, R. (2018). *Pangeo: A big-data ecosystem for scalable earth system science*. European Geosciences Union 20th EGU General Assembly, EGU2018, Proceedings from the conference held 4–13 April, 2018 in Vienna, Austria, p. 12146.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Heuzé, C. (2017). North Atlantic deep-water formation and amoc in CMIP5 models. *Ocean Science*, *13*(4), 609–622. https://doi.org/10.5194/os-13-609-2017

Hoyer, S., & Hamman, J. (2017). Xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, *5*(1). https://doi.org/10.5334/jors.148

Hughes, C. W., & de Cuevas, B. A. (2001). Why Western Boundary Currents in Realistic Oceans are Inviscid: A Link between Form Stress and Bottom Pressure Torques. *Journal of Physical Oceanography*, *31*(10), 2871–2885. https://doi.org/10.1175/1520-0485(2001)031<2871:WWBCIR>2.0.CO;2

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Will artificial intelligence supersede earth system and climate models? *Nature Machine Intelligence*.

Joyce, T., & Zhang, R. (2010). On the path of the Gulf Stream and the Atlantic meridional overturning circulation. *Journal of Climate*, *23*. https://doi.org/10.1175/2010JCLI3310.1

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. https://arxiv.org/abs/1412.6980

Krasting, J. P., John, J. G., Blanton, C., McHugh, C., Nikonov, S., Radhakrishnan, A., et al. (2018). Noaa-gfdl gfdl-esm4 model output prepared for CMIP6 CMIP. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1407

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

Lapuschkin, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, *10*, e0130140. https://doi.org/10.1371/journal.pone.0130140

Larson, S. M., Buckley, M. W., & Clement, A. C. (2020). Extracting the buoyancy-driven atlantic meridional overturning circulation. *Journal of Climate*, *33*(11), 4697–4714. https://doi.org/10.1175/JCLI-D-19-0590.1

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, *18*(1), 6765–6816.

Lohmann, K., Jungclaus, J. H., Matei, D., Mignot, J., Menary, M., Langehaug, H. R., et al. (2014). The role of subpolar deep water formation and nordic seas overflows in simulated multidecadal variability of the atlantic meridional overturning circulation. *Ocean Science*, *10*(2), 227–241. https://doi.org/10.5194/os-10-227-2014

Losch, M., Adcroft, A., & Campin, J.-M. (2004). How sensitive are coarse general circulation models to fundamental approximations in the equations of motion? *Journal of Physical Oceanography*, *34*(1), 306–319. https://doi.org/10.1175/1520-0485(2004)034⟨0306:HSACGC⟩2.0.CO;2

Lozier, M., Li, F., Bacon, S., Bahr, F., Bower, A., Cunningham, S., et al. (2019). A sea change in our view of overturning in the subpolar north atlantic. *Science*, *363*, 516–521. https://doi.org/10.1126/science.aau6592

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, et al. (Eds.) *Advances in neural information processing systems* (pp. 4765–4774). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Maloney, E. D., Gettelman, A., Ming, Y., Neelin, J. D., Barrie, D., Mariotti, A., et al. (2019). Process-oriented evaluation of climate and weather forecasting models. *Bulletin of the American Meteorological Society*, *100*(9), 1665–1686. https://doi.org/10.1175/bams-d-18-0042.1

Manucharyan, G. E., Siegelman, L., & Klein, P. (2021). A deep learning approach to spatiotemporal sea surface height interpolation and estimation of deep currents in geostrophic ocean turbulence. *Journal of Advances in Modeling Earth Systems*, *13*(1), e2019MS001965. https://doi.org/10.1029/2019MS001965

Marshall, J., & Schott, F. (1999). Open-ocean convection: Observations, theory, and models. *Reviews of Geophysics*, *37*(1), 1–64. https://doi.org/10.1029/98rg02739

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., & Stouffer, R. J. (2000). The coupled model intercomparison project (CMIP). *Bulletin of the American Meteorological Society*, *81*(2), 313–318. https://doi.org/10.1175/1520-0477(2000)081<0313:TCMIPC>2.3.CO;2

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F., et al. (2007). The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, *88*(9), 1383–1394. https://doi.org/10.1175/bams-88-9-1383

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, *65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008

Munk, W. H. (1950). On the wind-driven ocean circulation. *Journal of Meteorology*, *7*(2), 80–93. https://doi.org/10.1175/1520-0469(1950)007<0080:OTWDOC>2.0.CO;2

Nye, J., Joyce, T., Kwon, Y.-O., & Link, J. (2011). Silver hake tracks changes in northwest Atlantic circulation. *Nature Communications*, *2*, 412. https://doi.org/10.1038/ncomms1420

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, *178*(3), 389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al. (2019). *Keras Tuner*. Retrieved from https://github.com/keras-team/keras-tuner

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.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). San Francisco, CA, USA. august 13–17, 2016.

Roberts, M. J., Banks, H., Gedney, N., Gregory, J., Hill, R., Mullerworth, S., et al. (2004). Impact of an eddy-permitting ocean resolution on control and climate change simulations with a global coupled GCM. *Journal of Climate*, *17*(1), 3–20. https://doi.org/10.1175/1520-0442(2004)017<0003:IOAEOR>2.0.CO;2

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rumelhart, D., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536. https://doi.org/10.1038/323533a0

Saba, V. S., Griffies, S. M., Anderson, W. G., Winton, M., Alexander, M. A., Delworth, T. L., et al. (2016). Enhanced warming of the northwest Atlantic Ocean under climate change. *Journal of Geophysical Research: Oceans*, *121*(1), 118–132. https://doi.org/10.1002/2015jc011346

Sanchez-Franks, A., & Zhang, R. (2015). Impact of the Atlantic meridional overturning circulation on the decadal variability of the Gulf Stream path and regional chlorophyll and nutrient concentrations. *Geophysical Research Letters*, *42*(22), 9889–9887. https://doi.org/10.1002/2015gl066262

Schlund, M., Eyring, V., Camps-Valls, G., Friedlingstein, P., Gentine, P., & Reichstein, M. (2020). Constraining uncertainty in projected gross primary production with machine learning. *Journal of Geophysical Research: Biogeosciences*, *125*, e2019JG005619. https://doi.org/10.1029/2019JG005619

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). *Deep inside convolutional networks: Visualising image classification models and saliency maps*. https://arxiv.org/abs/1312.6034

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM conference on AI, ethics, and society.(Aies)*. https://arxiv.org/abs/1911.02508

Smith, R. D., Jones, P., Briegleb, B. P., Bryan, F. O., Danabasoglu, G., Dennis, J. M., et al. (2010). *The parallel ocean program (pop) reference manual: Ocean component of the community climate system model (CCSM) and community earth system model (CESM)*. http://n2t.net/ark:/85065/d70g3j4h

Sonnewald, M., Dutkiewicz, S., Hill, C., & Forget, G. (2020). Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Science Advances*, *6*(22). https://doi.org/10.1126/sciadv.aay4740

Sonnewald, M., Lguensat, R., Jones, D. C., Dueben, P. D., Brajard, J., & Balaji, V. (2021). Bridging observation, theory and numerical simulation of the ocean using machine learning. *Environmental Research Letters*, *16*, 073008. https://doi.org/10.1088/1748-9326/ac0eb0

Sonnewald, M., Wunsch, C., & Heimbach, P. (2019). Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science*, *6*(5), 784–794. https://doi.org/10.1029/2018ea000519

Sverdrup, H. U. (1947). Wind-driven currents in a baroclinic ocean; with application to the equatorial currents of the eastern pacific. *Proceedings of the National Academy of Sciences*, *33*(11), 318–326. https://doi.org/10.1073/pnas.33.11.318

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/bams-d-11-00094.1

Thomas, M. D., Boer, A. M. D., Johnson, H. L., & Stevens, D. P. (2014). Spatial and temporal scales of Sverdrup balance. *Journal of Physical Oceanography*, *44*(10), 2644–2660. https://doi.org/10.1175/jpo-d-13-0192.1

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. https://doi.org/10.1029/2019MS002002

Tsujino, H., Urakawa, L. S., Griffies, S. M., Danabasoglu, G., Adcroft, A. J., Amaral, A. E., et al. (2020). Evaluation of global ocean–sea-ice model simulations based on the experimental protocols of the ocean model intercomparison project phase 2 (OMIP-2). *Geoscientific Model Development Discussions, 2020*, 1–86. https://doi.org/10.5194/gmd-2019-363

Wang, C., Zhang, L., Lee, S.-K., Wu, L., & Mechoso, C. (2014). A global perspective on CMIP5 climate model biases. *Nature Climate Change*, advance online publication. https://doi.org/10.1038/nclimate2118

Wang, Z., Lu, Y., Dupont, F., Loder, J. W., Hannah, C., & Wright, D. G. (2015). Variability of sea surface height and circulation in the north Atlantic: Forcing mechanisms and linkages. *Progress in Oceanography*, *132*, 273–286. https://doi.org/10.1016/j.pocean.2013.11.004

Weaver, A. J., Sedláček, J., Eby, M., Alexander, K., Crespin, E., Fichefet, T., et al. (2012). Stability of the Atlantic meridional overturning circulation: A model intercomparison. *Geophysical Research Letters*, *39*, L20709. https://doi.org/10.1029/2012gl053763

Weijer, W., Cheng, W., Garuba, O. A., Hu, A., & Nadiga, B. T. (2020). CMIP6 models predict significant 21st century decline of the atlantic meridional overturning circulation. *Geophysical Research Letters*, *47*, e2019GL086075. https://doi.org/10.1029/2019GL086075

Wunsch, C. (2011). The decadal mean ocean circulation and Sverdrup balance. *Journal of Marine Research*, *69*, 417–434. https://doi.org/10.1357/002224011798765303

Wunsch, C., & Ferrari, R. (2004). Vertical mixing, energy, and the general circulation of the oceans. *Annual Review of Fluid Mechanics*, *36*(1), 281–314. https://doi.org/10.1146/annurev.fluid.36.050802.122121

Wunsch, C., & Heimbach, P. (2013). Chapter 21 - Dynamically and kinematically consistent global ocean circulation and ice state estimates. In G. Siedler, S. M. Griffies, J. Gould, & J. A. Church (Eds.), *Ocean circulation and climate* (pp. 553–579). Academic Press. https://doi.org/10.1016/b978-0-12-391851-2.00021-0

Yeager, S. (2015). Topographic coupling of the Atlantic overturning and gyre circulations. *Journal of Physical Oceanography*, *45*, 1258–1284. https://doi.org/10.1175/JPO-D-14-0100.1

Zeiler, M. D., & Fergus, R. (2013). *Visualizing and understanding convolutional networks*. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, R. (2008). Coherent surface-subsurface fingerprint of the Atlantic meridional overturning circulation. *Geophysical Research Letters*, *35*, L20705. https://doi.org/10.1029/2008gl035463

Zhang, R., Delworth, T. L., Rosati, A., Anderson, W. G., Dixon, K. W., Lee, H.-C., & Zeng, F. (2011). Sensitivity of the North Atlantic Ocean circulation to an abrupt change in the Nordic sea overflow in a high resolution global coupled climate model. *Journal of Geophysical Research*, *116*, C12024. https://doi.org/10.1029/2011jc007240

Zhang, R., Sutton, R., Danabasoglu, G., Kwon, Y.-O., Marsh, R., Yeager, S. G., et al. (2019). A review of the role of the Atlantic meridional overturning circulation in Atlantic multidecadal variability and associated climate impacts. *Reviews of Geophysics*, *57*(2), 316–375. https://doi.org/10.1029/2019rg000644

Zhang, R., & Vallis, G. (2007). The role of bottom vortex stretching on the path of the north Atlantic western boundary current and on the northern recirculation gyre. *Journal of Physical Oceanography*, *37*, 2053–2080. https://doi.org/10.1175/jpo3102.1