Revealing the impact of global heating on North Atlantic circulation using transparent machine learning

Maike Sonnewald\textsuperscript{1,1} and Redouane Lguensat\textsuperscript{2,2}

\textsuperscript{1}Princeton University
\textsuperscript{2}LOCEAN-IPSL

November 30, 2022

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 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 dataset 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\textsubscript{2} quadrupling, the circulation weakens due to a shift in deep water formation regions, a northward shift of the Gulf stream and an eastwards shift in the North Atlantic Current. If CO\textsubscript{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 towards trustworthy machine learning called for within oceanography and beyond.
Revealing the impact of global heating on North Atlantic circulation using transparent machine learning

Maike Sonnewald\textsuperscript{1,2,3} and Redouane Lguensat\textsuperscript{4,5}

\textsuperscript{1}Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ 08540, USA
\textsuperscript{2}NOAA/OAR Geophysical Fluid Dynamics Laboratory, Ocean and Cryosphere Division, Princeton, NJ 08540, USA
\textsuperscript{3}University of Washington, School of Oceanography, Seattle, WA 98195, USA
\textsuperscript{4}Laboratoire des Sciences du Climat et de l’Environnement, CEA Saclay, Institut Pierre Simon Laplace, Gif-Sur-Yvette, France
\textsuperscript{5}LOCEAN-IPSL, Sorbonne Université, Institut Pierre Simon Laplace, Paris, France.

Key Points:

\begin{itemize}
  \item The transparent machine learning method Tracking global Heating with Ocean Regimes (THOR) is presented and applied to the North Atlantic circulation.
  \item Global heating shifts regimes of the Gulf Stream north, the North Atlantic Current east and deep watermass formation towards lighter waters.
  \item Widely applicable, THOR could accelerate analysis and dissemination of climate model data needing only depth, sea level and wind stress.
\end{itemize}

Corresponding author: Maike Sonnewald, maikes@princeton.edu
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 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 dataset 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 towards trustworthy machine learning 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 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 towards trustworthy machine learning 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 (Zhang et al., 2019; Lohmann et al., 2014; Lozier 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 (Tsujino et al., 2020; Roberts et al., 2004; Marshall & Schott, 1999; Lohmann et al., 2014). Due to the complicated and non-linearly 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 towards 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; Schlund et al., 2020; Reichstein et al., 2019), 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 (Zhang et al., 2019; Larson et al., 2020; Cheng et al., 2013; Weaver et al., 2012; Weijer et al., 2020; Meehl et al., 2000). 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 (Meehl et al., 2000, 2007; Taylor et al., 2012; Eyring et al., 2015). 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 hinder 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 e.g. 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 (Marshall & Schott, 1999; Lohmann et al., 2014; Böning et al., 2006). 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 (Zhang et al., 2019; Joyce & Zhang, 2010; Nye et al., 2011; Sanchez-Franks & Zhang, 2015; Yeager, 2015; Zhang, 2008). These mechanisms can be seen as governing the circulation or being a direct product of its strength (Wunsch & Ferrari, 2004; Kuhlbrodt et al., 2007). Overall, the field of oceanography is increasingly starting to use advanced ML methods, as reviewed in (Sonnewald et al., 2021a). To infer subsurface dynamics, ML has been employed to predict currents at 1000m from satellites (Chapman & Charantonis, 2017), and subsurface structure from idealised 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 labelled 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 labelled data from which to learn. Such data is scarce, and labelling is often complicated by the data being some combination of highly nonlinear, chaotic, high-dimensional, non-stationary 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 la-
bels 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 (Rudin, 2019; Irrgang et al., 2021; Sonnewald et al., 2021a). 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 [ALTAI]) and in a mandate in the United States of America (E.O. 13960 of Dec 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) (Olden et al., 2004; Toms et al., 2020; Lapuschkin et al., 2015; Ribeiro et al., 2016; Lundberg & Lee, 2017). 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 (Sonnewald et al., 2019, 2020; Callaham et al., 2021; Kaiser et al., In prep).

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

Identifying dynamical regimes

The first step of THOR identifies 2D dynamical regimes (Fig. 1a) in the realistic 1° numerical ocean model Estimating the Circulation and Climate of the Ocean (EC-COv4r3 (Forget et al., 2015; Adcroft et al., 2004; 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 non-unique 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: 1) the wind and bottom stress curl, 2) the advection of planetary vorticity, 3) bathymetric interactions through bottom pressure torque, 4) curl of non-linear interactions between terms and 5) lateral viscous dissipation from within the ocean interior (Sonnewald et al., 2019; Munk, 1950; Hughes & de Cuevas, 2001) (Appendix B). The five terms form a closed budget, and a 5-dimensional vector field, \( \mathbf{x} \), on the model grid. Each element \( x_i \) 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 \( (\text{lon}, \text{lat}) = (\theta_i, \phi_i) \). Within \( x \), 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 (Fig. 2a). The global area averaged term balances (Fig. 2b) demonstrate which dynamical drivers are important and which are negligible. Here, the North Atlantic is discussed (Fig. 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) 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
Figure 2: ECCO 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 colours of the dynamical regimes are: Non-linear (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).
the higher latitudes. The NL regime is notable as it is made up of a collection of smaller
regimes that all have a large non-linear torque component, but make up a very small com-
ponent 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 dynam-
ical 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 con-
sidering 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 large scale circulation, the AMOC
is an overall clockwise feature, with surface waters travelling 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 resul-
tant streamfunction. The sum of the streamfunctions associated with each regime com-
prises the AMOC. Decomposing the AMOC into dynamical regimes shows the local con-
tribution of each regime individually to the AMOC, and reveals a complex interplay of
dynamical features. The density structure can be decomposed by dynamical regime sim-
ilarly. Together, the density structure and the meridional overturning are thus decom-
posed by dynamical regime. Overarching coherent and in-depth physical regimes emerge
(Fig. 3). The overall transport in the N-SV regime is clockwise (red, Fig. 3a). It trans-
ports relatively light watermasses northwards in the surface (<1000m) as seen by the
light coloured isopycnals overlying the transport. It coincides with the large subtropi-
cal gyre thought to be in Sverdrup balance (Thomas et al., 2014; Wunsch, 2011). In the
S-SV regime the transport is largely anticlockwise (blue, Fig. 3b), taking place also in
the predominantly lighter watermasses with northward transport confined to the sur-
face (<500m). The S-SV regime is largely seen in the subpolar gyre. The TR regime also
transports waters anticlockwise (Fig. 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 2500m. The SO
regime is largely confined to the Southern Ocean (Fig.3d), and absent in the North At-
tlantic. The NL regime (Fig.3e) contributes clockwise between 50-80°N, reaching depths
of approximately 2500m. This regime also has dense waters close to the surface but they
are lighter than in the TR regime. Together, the TR and NL regime are thought to gov-
ern 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 (Fig.3f), the transports are both clockwise and anticlockwise, with stronger
transports largely confined to lower latitudes (< 30°N and S). The MD regime acts pre-
dominantly 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 approximately
1000m at 47-53°N. The latitudes where the clockwise/anticlockwise circulation is stacked,
coincides with the region occupied by the North Atlantic Current.

Fig. 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 dens-
Figure 3: Physical interpretation of regimes from 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 Fig. 2. Note the TR regimes contribution of dense watermasses that move southwards at depth, the stacked (red over blue in depth) contributions of the MD 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 grey areas the regime was not present.
Est waters come from the DSOW and ISOW, and creation of denser waters would over-
all act to invigorate the AMOC. If there were a shift in the location of deep water for-
mation towards 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 sim-
plified representation of the highly complicated full structure, which highlights the un-
derlying processes that constitute the dynamical regimes. The motivation behind using
ML for this strategy is that it can identify such areas within the equation space that con-
stitute the dominant term balances in an unbiased manner. THOR independently iden-
tifies the expected dynamics, but importantly adds geographic precision. The simplifi-
cation 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.

**Prediction with a NN**

The second step of THOR trains a NN (Fig. 1b) to infer in-depth dynamics from
data that is largely readily available from for example CMIP6 models, using NN meth-
ods to infer the source of predictive skill (Fig. 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 bottom pressure
torque 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 ($\eta$) are used, with
$\eta$ 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 ($\nabla \times \tau$), Cori-
olis ($f$) and $\eta$ (Buckley & Marshall, 2016; Bingham & Hughes, 2009; Z. Wang et al.,
2015). The $f$ and gradients of the $\eta$ term reflect the surface geostrophic velocity. In sum the
features are: wind stress torque, $H$, $f$ and $\eta$, and the latitudinal and longitudinal gra-
dients of $H$ and $\eta$.

A fully connected multilayer perceptron (MLP) NN is used. The motivation to em-
ploy a NN is to determine relationships between the input features and the labels within
a training dataset, so these relationships can be leveraged to make similar predictions
for unseen data. MLPs are powerful universal function approximators, and particularly
suited for 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 $\{2 \text{ to } 5\}$, we
manually tested different activation function from $\{\text{ReLU, SeLU, Tanh}\}$ 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 stochas-
tic 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 5 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 ca-
pacity 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 archi-
tecture as mentioned above. When predicting the classes, an average over the 50 soft-
max 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 posi-
tion is then the one with the maximum probability.
Figure 4: Sketch of dynamics. Upper layer illustrates the dynamical regimes (colours as in Fig. 2). Cube below shows density contours (colours as in Fig. 3), with the overall currents that are coloured from warm (red) to cold (blue). The DSOW and ISOW are strictly overflows, but depicted with arrows for simplicity.
Figure 5: Detailed sketch of THOR workflow step C illustrating the Ensemble 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 Fig. 1) is run to get the probabilities of belonging to one of the 6 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.

Code was written using the Python-based Keras library (Chollet et al., 2015) and makes use of several other open source libraries (Pedregosa et al., 2011; Hoyer & Hamman, 2017; Harris et al., 2020; Hamman et al., 2018). 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, Fig. 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.

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 non-trivial. 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 (Sonnewald et al., 2021b; Balaji, 2020; Irrgang et al., 2021; McGovern et al., 2019; 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 towards the resulting prediction. These often produce a ‘heatmap’ rendition of NN classification decisions (Montavon et al., 2017; Zeiler & Fergus, 2013; Rumelhart et al., 1986; Simonyan et al., 2014). The LIME method was also used to assess the source of predict
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
 calculates 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” \( R_{i \leftarrow j}^{(l+1)} \) 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 \( \epsilon \)-rule was used that helps avoid numerical issues when dividing by small
numbers:

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

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 En-
semble MLP. For geoscientific applications, the positive component of LRP have previ-
ously 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-
\( \epsilon \) implementation provided by the iNNvestigate (Alber et al., 2019) library was used, that
supports Keras-written models. Fig. A3 illustrates the spatial distributions of the rel-
levances.

For each dynamical regime, the relevance contributions are assessed as the mean
and standard deviation across the North Atlantic region spatially. Note the initial la-
beles and not the predicted clusters were used. Positive and negative relevance contribu-
tions are treated separately (Fig. 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 pos-
tive relevance. There is considerable spatial variability, as reflected by the standard de-
viation, 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
\( \eta \) contribute positively in the S-SV, TR and NL regimes, which could be due to a merid-
ional 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 Fig. A3, 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 con-
tribute 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 sec-
tor considered. The ability to explain the Ensemble MLPs skill lends confidence to its
subsequent predictions. Assessing the relevance metric highlights the physical underpin-
ing of the Ensemble MLP skill, and means that THOR can be applied with more con-
fidence in previously unseen models or under different climate forcing.
Interpreting physical regimes in a climate model

The final step of THOR (Fig. 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 CO₂ concentration in the atmosphere is increased by 1% over 140 years (1pctCO₂), representing a still transient climate state, and an abrupt quadrupling of CO₂ (abrupt4xCO₂) that has had more time to stabilise. The AMOC weakens as expected (Weijer et al., 2020) in the 1pctCO₂, and decreases further in the abrupt4xCO₂ (Fig. 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 year 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 (Fig. 6a), dynamical regime distributions similar to ECCO are seen. The MD regime occupies a large area stretching east and northwards from the 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 35-45°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 centre 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 1pctCO₂ scenario (Fig. 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 abrupt4xCO₂ 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 abrupt4xCO₂ scenario (Fig. 6c), the climate has had more time to stabilise, 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 eastward shift which could indicate a change in the North Atlantic Current. Concurrently, the S-
Figure 6: The dynamical regimes predicted under global heating using THOR. 

a) The historical scenario shows dynamical regimes that are very similar to ECCO (Fig. 2). 

b) The 1pctCO2 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 abrupt4xCO2 scenarios shows a distinct shift of the downwelling areas (TR) towards 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 AMOC’s progressive weakening from historical, 1pctCO2 and abrupt4xCO2 scenarios, as observed (Fig. A1).
SV regime extends further South. These observed changes point towards a weakened AMOC.

Interpreting the changes between the historical and future forcing scenarios, the declining AMOC can be put into context. The 1pctCO2 scenario is still stabilising, 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 abrupt4xCO2 scenario, the dynamical regime configuration is more stable, in concurrence with the climate having had more time to stabilise. Under abrupt4xCO2, the Gulf Stream has shifted further north, and the North Atlantic Current has shifted east. The deep water formation regions move towards 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 favoured. 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 recognise 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 dataset using interpretable unsupervised ML, clustering on transformed realistic 3D ocean model momentum fields (Sonnewald et al., 2019) (Fig. 1). The labels constitute six dynamical regimes, representing northward and southward surface transport, northern hemisphere deep water formation and southern hemisphere upwelling, a momentum driven regime and a composite dynamical regime where non-linear processes dominate (Fig. 2, 3).

Using THOR, the evaluated forcing scenarios are the historical and the future projections 1pctCO2 and abrupt4xCO2 (Fig. 6). In the North Atlantic (Fig. 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 abrupt4xCO2 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 1pctCO2 and further in the abrupt4xCO2, 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 (Fig. 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 non-local 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 were 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 (Wunsch & Ferrari, 2004; Griffies et al., 2015; Kuhlbrodt et al., 2007). 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 (Sonnewald et al., 2021a; Balaji, 2020; Irrgang 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 (Fig. 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 the-
ory suggests could be helpful, such as 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. Towards 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.

References


Eyring, V., Bony, S., Meehl, G., Senior, C., Stevens, B., Ronald, S., & Taylor, K.


doi: 10.5194/gmd-8-3071-2015

doi: 10.1175/2011JCLI4083.1

doi: 10.1175/JCLI-D-14-00353.1


doi: 10.5334/jors.148


doi: 10.22033/ESGF/CMIP6.1407

doi: 10.1029/2004RG000166


doi: 10.1016/j.patcog.2016.11.008

doi: 10.1175/1520-0469(1950)007⟨0080:OTWDOC⟩2.0.CO;2


doi: https://doi.org/10.1016/j.ecolmodel.2004.03.013


doi: https://doi.org/10.1002/2015JC011346


Smith, R. D., & Collaborators. (2010). *The parallel ocean program (pop) reference manual: Ocean component of the community climate system model (ccsm) and community earth system model (cesm).*


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...
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 Fig. 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 National de Recherche under the “Investissements d’avenir” program with the reference ANR-17-MPGA-0010. 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.
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 Fig. 3 is defined as:

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

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 approximately 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. Fig. A1 illustrates the weakening of the AMOC, as expected, where the historical is strongest and the 1pctCO2 shows a weakening that is more pronounced in the abrupt4xCO2 scenario.

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:

$$\begin{align*}
\partial_t u + f k \times u &= -\frac{1}{\rho_0} \nabla_h p + \frac{1}{\rho_0} \partial_z \tau + a + b, \\
\partial_z p &= -g \rho,
\end{align*}$$

$$\nabla_h \cdot u + \partial_z w = 0,$$

Pressure, gravity, density, and vertical shear stress are $p$, $g$, $\rho$, and $\tau$, respectively, with $\rho_0$ the reference density; the three-dimensional velocity field $\mathbf{v} = (u, v, w) = (u, w)$; the gradient $\nabla = (\nabla_h, \partial_z)$; the unit vector is denoted $\mathbf{k}$; planetary vorticity is a function of latitude $\theta$ in $f k = (0, 0, 2\Omega \sin(\theta))$; the viscous forcing from vertical shear is $\partial_z \tau$; the nonlinear torque is $a$, and the horizontal viscous forcing $b$ 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
Figure A1. The AMOC in ESM4.1. a) The abrupt4xCO2 scenario where the AMOC is weakest, b) the 1pctCO2 scenario has a slightly stronger AMOC and c) shows the modern AMOC. The grey area shows the bathymetry.
$$\beta V = \frac{1}{\rho_0} \nabla p_b \times \nabla H + \frac{1}{\rho_0} \nabla \times \tau + \nabla \times A + \nabla \times B$$  \(\text{(B3)}\)

where $$\nabla U = 0$$, $$U \cdot \nabla f = \beta V$$, the bottom pressure is denoted $$p_b$$, $$A = \int_H^\eta \eta d\zeta$$, and

$$B = \int_H^\eta \eta d\zeta$$. The curl operator $$\nabla \times$$ produces a scalar, that represents the vertical component of the operator. The left-hand side of equation 3 is the planetary vorticity advection term, while the right-hand side of equation 3 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 3 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 bottom pressure torque, the wind and bottom stress curl, curl of non-linear 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:

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

where $$uu$$ is a second-order tensor. The right-hand side of equation 4 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) (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 3 are at hand is a stringent test of underspecification. Fig. 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 & Collaborators, 2010)), in the coupled ocean-sea ice configuration (Gent et al., 01 Oct. 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 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 parameterizations 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 D2. Together they show that
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 Fig. 2b.
Figure D1. Confusion matrices for the Ensemble MLP classification on train (left) and validation (right) data

Table D1. Classification report of the performance of the Ensemble MLP on the validation set

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>0.89</td>
<td>0.93</td>
<td>0.91</td>
<td>10241</td>
</tr>
<tr>
<td>N-SV</td>
<td>0.90</td>
<td>0.93</td>
<td>0.92</td>
<td>8270</td>
</tr>
<tr>
<td>S-SV</td>
<td>0.87</td>
<td>0.94</td>
<td>0.90</td>
<td>6821</td>
</tr>
<tr>
<td>TR</td>
<td>0.79</td>
<td>0.85</td>
<td>0.82</td>
<td>3486</td>
</tr>
<tr>
<td>SO</td>
<td>0.82</td>
<td>0.88</td>
<td>0.85</td>
<td>1792</td>
</tr>
<tr>
<td>NL</td>
<td>0.73</td>
<td>0.60</td>
<td>0.66</td>
<td>9718</td>
</tr>
<tr>
<td>Average</td>
<td>0.83</td>
<td>0.85</td>
<td>0.84</td>
<td>40328</td>
</tr>
</tbody>
</table>

our ML classifier reaches an average F-score of 0.84 (an ideal F-score is 1). Individually, the Ensemble MLP reaches a good F-score ($\geq 0.8$) for all the classes except the NL class, which is unsurprisingly the hardest to classify.

Appendix E LRP: Spatial representation of interpretability/explainability

The spatial maps of the interpretability (Fig. D1) 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 $\eta$ 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 $\eta$ and $H$ term, and their latitudinal and longitudinal gradients, that adds valuable information to the 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 $\eta$ 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., 2021b).

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 dataset created by LIME is obtained by perturbing the 8 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 [GitHub](https://github.com/marcotcr/lime)

The results of applying LIME to the Ensemble MLP are displayed in Fig. ?? Over-all, 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
Figure E1. The spatial relevance maps. The dynamical regimes (columns) and the input features (rows) illustrating the contributions to the skill of the Ensemble MLP. The relevances are averaged for each point (lat, lon) over the Ensemble MLP.
Figure F1. The spatial relevance maps computed using LIME. The dynamical regimes (columns) and the input features (rows) illustrating the contributions to the skill of the Ensemble MLP. The relevances are averaged for each point (lat, lon) over the Ensemble MLP, see Fig. 5.
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 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 LRPCs 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 e.g. 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.

Appendix H Detailed maps THOR ESM predictions for 1pctCO2 and abrupt4xCO2 scenarios

The change in area occupied by the dynamical regimes from the historical to the 1pctCO2 (Fig. F1) and abrupt4xCO2 (Fig. F2) scenarios. The figures serve to demonstrate the changes taking place in more detail than Fig. 4. The figures demonstrate where a regime has displaced another (left columns in F1 and 2), or where the regime has been replaced by another (right columns in F1 and F2). For the North Atlantic Current, the eastwards shift is seen by the area occupied by the MD regime in the 1pctCO2 scenario having expanded into the NL and S-SV areas in F1a. Historical expanses to the north are being occupied by the S-SV regime (Fig. F1b). Similarly, the North Atlantic Current shift in the abrupt4xCO2 scenario is seen particularly also in F2b where large expanses of the S-SV regime have moved into the area occupied by the S-SV regime in the historical scenario. For the 1pctCO2 scenario in F1, the shift in the Gulf Stream location can be seen in F1c, where the TR regime has been displaced, and F1d where the MD regime has replaced the S-SV regime. This is similarly visible in abrupt4xCO2 scenario for Fig. F2c and d, 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 1pctCO2 scenario (Fig. F1g, h, k and l), likely because the areas that were shifting were not deemed significant, and still could be influenced by natural variability. In the abrupt4xCO2 scenario, a marked shift is observed (Fig. F2g, h, k and l), where the TR regime partially displaces the NL regime in the sub-polar region.
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 1pctCO2 scenario, where the colour 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 colours different from the regime in question signify that the area was displaced by a different regime of the respective colour.
Figure H2. Maps of expanse addition and loss by regime for scenario abrupt4xCO2. As for Fig. F1.