Revealing the impact of global warming on climate modes using transparent machine learning and a suite of climate models

Maike Sonnewald [1 2 3] **Redouane Lguensat** [4 5] **Aparna Radhakrishnan** [2] **Zouberou Sayibou** [6]
Andrew T. Wittenberg [2] **V. Balaji** [1]

Abstract

The ocean is key to climate through its ability to store and transport heat and carbon. From studies of past climates, it is clear that the ocean can exhibit a range of dramatic variability that could have catastrophic impacts on society, such as changes in rainfall, severe weather, sea level rise and large scale climate patterns. The mechanisms of change remain obscure, but are explored using a transparent machine learning method, Tracking global Heating with Ocean Regimes (THOR) presented here. We investigate two future scenarios, one where CO_2 is increased by 1% per year, and one where CO_2 is abruptly quadrupled. THOR is engineered combining interpretable and explainable methods to reveal its source of predictive skill. At the core of THOR, is the identification of dynamically coherent regimes governing the circulation, a fundamental question within oceanography. Three key regions are investigated here. First, the North Atlantic circulation that delivers heat to the higher latitudes is seen to weaken and we identify associated dynamical changes. Second, the Southern Ocean circulation, the strongest circulation on earth, is seen to intensify where we reveal the implications for interactions with the ice on Antarctica. Third, shifts in ocean circulation regimes are identified in the tropical Pacific region, with potential impacts on the El Niño Southern Oscillation, Earth's dominant source of year-to-year climate variations affecting weather extremes, ecosystems, agriculture, and fisheries. Together with revealing these climatically relevant ocean dynamics, THOR also constitutes a step towards trustworthy machine learning called for within oceanography and beyond because its predictions are physically tractable. We conclude with by highlighting open questions and potentially fruitful avenues of further machine learning applications to climate research.

1. Introduction

The ocean within 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). Ocean circulation changes could have catastrophic impacts on society, such as changes in rainfall, severe weather, sea level rise and large scale climate patterns, because of the vast amounts of heat and carbon it governs. Tools from machine learning (ML) are well placed to help address challenges towards understanding ocean and climate variability. Another key tool to understand future changes is the Coupled Model Intercomparison Project (CMIP6) (Meehl et al., 2000; 2007; Taylor et al., 2012; Eyring et al., 2015). The CMIP6 ensemble is a comprehensive ensemble of climate models, and the simulations are vital tools for understanding variability. However, the model variability often is reduced to bulk metrics e.g. summarizing complex dynamics with single numbers leaving specific mechanisms opaque (Weijer et al., 2020). This is because the complexity and size of the CMIP6 model ensemble can hinder data dissemination and analysis. Such hurdles are examples 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 Tracking global Heating with Ocean Regimes (THOR) method 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.

[1] Program in Atmospheric and Oceanic Sciences, Princeton University, USA [2] NOAA/OAR Geophysical Fluid Dynamics Laboratory, Ocean and Cryosphere Division, USA [3] University of Washington, School of Oceanography, USA [4] Laboratoire des Sciences du Climat et de l'Environnement, CEA Saclay, Institut Pierre Simon Laplace, France [5] LOCEAN-IPSL, Sorbonne Université, Institut Pierre Simon Laplace, France. [6] Bronx Community College, USA. Correspondence to: Maike Sonnewald <maikes@princeton.edu>.

Tackling Climate Change with Machine Learning Workshop at ICML 2021.

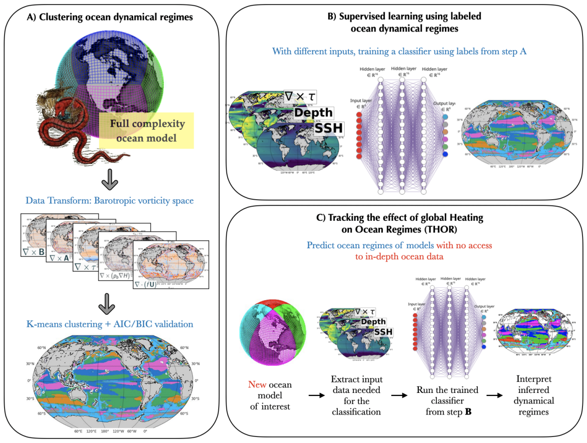
A) Clustering ocean dynamical regimes
Full complexity ocean model
Data Transform: Barotropic vorticity space
$\nabla \times \mathbf{B}$
$\nabla \times \mathbf{A}$
$\nabla \times \tau$
$\nabla \times (p_b \nabla H)$
$\nabla \cdot (f\mathbf{U})$
K-means clustering + AIC/BIC validation
B) Supervised learning using labeled ocean dynamical regimes
With different inputs, training a classifier using labels from step A
$\nabla \times \tau$
Depth
SSH
Input layer $\in \mathbb{R}^8$
Hidden layer $\in \mathbb{R}^{16}$
Hidden layer $\in \mathbb{R}^{16}$
Hidden layer $\in \mathbb{R}^{16}$
Output layer $\in \mathbb{R}^6$
C) Tracking the effect of global Heating on Ocean Regimes (THOR)
Predict ocean regimes of models with no access to in-depth ocean data
New ocean model of interest
Extract input data needed for the classification
Run the trained classifier from step B
Interpret inferred dynamical regimes

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 towards trustworthy ML. Figure taken from (Sonnewald & Lguensat, 2021)

Here, the THOR method is in focus, in particular the way in which a transparent ML application is crafted combining ML and oceanographic domain knowledge. We also demonstrate the power of THOR on three ocean features key to climate application. First, the North Atlantic circulation that delivers heat to the higher latitudes is seen to weaken and we identify associated dynamical changes. Second, the Southern Ocean circulation, the strongest circulation on earth, is seen to intensify where we reveal the implications for interactions with the ice on Antarctica. Third, the El Nino Southern Oscillation, the dominant climate mode on Earth, is seen to shift its mean state with implications for the monsoon and fisheries. The paper concludes with open questions for the ML community towards making applications more directly applicable to climate related problems, with a focus on transparent ML.

2. The Tracking global Heating with Ocean Regimes (THOR) method

2.1. A brief overview

THOR overcomes two common problems with ML applications to climate: a lack of labelled data, and the difficulty of understanding of the applications' source of predictive power. Fig. 1 shows the different components of THOR. Step A creates a labelled dataset. A label effectively constitutes defining consistent phenomena of interest. THOR uses an unsupervised clustering ML algorithm, namely a k-means algorithm, to identifies coherent structures within data from a state estimate (ocean model fit to observational data, (Wunsch & Heimbach, 2013; Forget et al., 2015)). Six regimes are identified, and their physical interpretation is detailed in (Sonnewald & Lguensat, 2021). For brevity in discussion, the dynamical impacts are described rather than the regimes in general. The knowledge of which dynamical drivers are key allows the regimes to be matched with input features taken only from the surface, that theory suggests could be good proxies for the dominant terms identified by the unsupervised ML.

Step B in THOR uses this labelled dataset to train an Ensemble MLP to predict in depth physics using only surface fields. Step B also addresses the lack of understanding of the source of predictive skill, a core hurdle to adoption within climate science (Rudin, 2019; Irrgang et al., 2021; Sonnewald et al., 2021). This is critical for climate applications and there is no observational 'out of sample' data to train on, making generalization (Balaji, 2020), and avoiding underspecification (D'Amour et al., 2020) central hurdles. THOR is deemed transparent using ML that is both interpretable and explainable (known as IAI and XAI, respectively where AI stands for artificial intelligence), specifically using the interpretable first step to feature engineer the second supervised step. For NN and other 'black-box' models, Additive Feature Attribution (AFA) methods to explain skill retrospectively are increasing in popularity, and here the method Layerwise Relevance Propagation is used for reasons discussed below (Olden et al., 2004; Toms et al., 2020; Lapuschkin et al., 2015; Ribeiro et al., 2016; Lundberg & Lee, 2017; Montavon et al., 2017; Zeiler & Fergus, 2013; Rumelhart et al., 1986; Simonyan et al., 2014). The relevances are then combined with domain expertise to ensure the predictions are rooted in physics. Step C applies THOR to data from CMIP6.

2.2. Predictions with a neural network

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 methods to infer the source of predictive skill. The data used is comprised of labeled input variables referred to as features, with the dynamical regimes (obtained in Step A) 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 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; 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. 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 {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 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 capacity and to weaken the dependence on the initial training parameters. The THOR Ensemble MLP is composed of 50 MLP renditions using the same architecture, as described above. When predicting the classes, an average over the 50 softmax probabilities for each pixel was done. 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 (Pedregosa et al., 2011; Hoyer & Hamman, 2017; Harris et al., 2020; Hamman et al., 2018).

2.3. Explaining the prediction skill: LRP + oceanographic theory

Using supervised ML, being able to explain the source of predictive skill and move beyond a 'black box' approach, to achieve 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., 2021; Balaji, 2020; Irrgang et al., 2021; McGovern et al., 2019; Toms et al., 2020). 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). Other methods were also tested, but overall the LRP method was most robust to local perturbations, and deemed most reliable. 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). 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 in Figure 5 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 2 illustrates the complete pipeline when using THOR, the Ensemble MLP is already trained and the user can use it to obtain the dynamical ocean regime classification on their ocean model of interest. Similarly, The user can also get the LRP predictions that explain the positive/negative relevances of the input features.

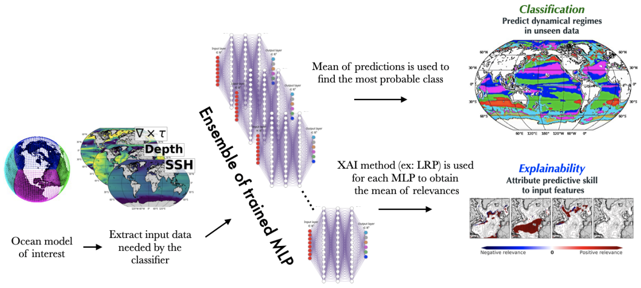
Classification
Predict dynamical regimes in unseen data
Mean of predictions is used to find the most probable class
$\nabla \times \tau$
Depth
SSH
Ensemble of trained MLP
XAI method (ex: LRP) is used for each MLP to obtain the mean of relevances
Explainability
Attribute predictive skill to input features
Ocean model of interest
Extract input data needed by the classifier
Negative relevance
0
Positive relevance

Figure 2. THOR as used in the "test phase". For a new ocean model, the predictions and relevances are averaged for each point (lat, lon) over the Ensemble MLP. Figure taken from (Sonnewald & Lguensat, 2021)

For each dynamical regime, the relevance contributions are assessed as the mean and standard deviation across region spatially. Positive and negative relevance contributions are treated separately. 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 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 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.

3. Application to understand the ocean under climate change

We now apply THOR to the CMIP6 models Geophysical Fluid Dynamics Laboratory (GFDL) Earth System Model 4.1 (ESM4.1 (Dunne et al., 2020; Krasting et al., 2018)) and the IPSL-CM6A-LR model (Boucher et al., 2018; 2020). THOR is applied to a 'historical' (1992-2011) scenario, and the 60 last years of scenarios where the CO_2 is 1% per year, and one where CO_2 is abruptly quadrupled (Fig. 3, white indicates inconsistent recognition).

The North Atlantic Overturning Circulation (AMOC) Overall, it is known that the AMOC weakens in a warmer climate. Here in Fig. 3 rows 1-3, the dark blue region (MD) shifts south and east, signifying less heat being transported north. This is stronger in IPSL-CM6A-LR than ESM4.1. There is a distinct shift with the orange region (TR) towards the west, signifying less dense waters forming, also stronger in IPSL-CM6A-LR. For the 4xAbrupt CO_2 the changes are bigger.

The El Nino Southern Oscillation (ENSO) The ENSO (Fig. 3 rows 4-6) change in mean state shows a widening of the pink (N-SV) regime stretching westwards from South America, signifying strengthened delivery of cold waters. This change is bigger for IPSL-CM6A-LR. The green regimes (S-SV) also widen, which are a signature of waters being pushed towards the equator by winds. These regime shifts and inter-model differences are linked to the time-mean patterns of surface winds and mixed layer depths in this region, which strongly modulate ENSO's behavior, impacts, and predictability (Guilyardi et al., 2020; Fedorov et al., 2020; Ding et al., 2020; Stevenson et al., 2021). Changes seen between the forcing scenarios are less marked than for the AMOC.

The Southern Ocean circulation The Southern Ocean is shown for a section called the Weddell Sea. Here a large wind gyre interacts with the current circling Antarctica (Fig. 3 rows 7-9), which does not change its position and strength markedly. The gray region (SO) widens, and moves into the pink region (N-SV). The gray regime increasing in size signifies a strengthening of the current system bringing up waters from great depths to be cooled and returned northwards. This is known to compensate for a weakened AMOC. This shift appears greater in ESM4.1 than in the IPSL-CM6A-LR model, but for both increases with increased forcing, as expected.

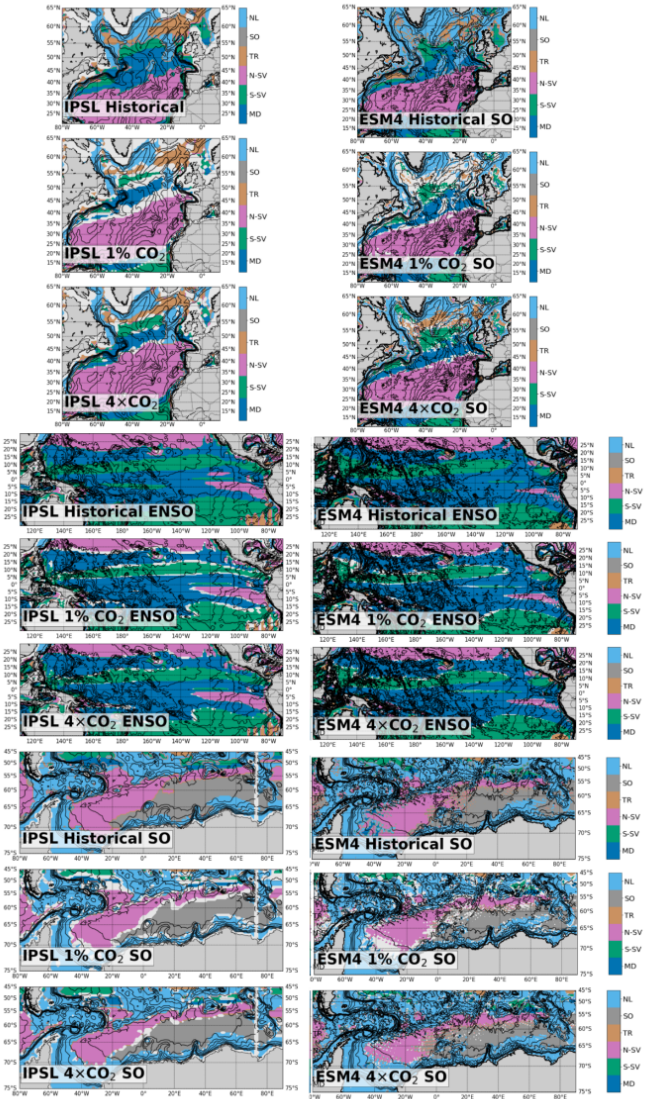

Figure 3. THOR applied to IPSL-CM6A-LR and ESM4.1 for the North Atlantic (top 3 rows), the tropical Pacific/ENSO region (row 4-6) and the southern Ocean (rows 6-9). Left column in the IPSL-CM6A-LR model and right column is the ESM4.1 model.

4. Conclusion

4.1. Discussion

Understanding how the climate has changed in the past and may change in the future is one the main reasons why several modeling groups around the world develop and analyze climate models. THOR could accelerate analysis and dissemination of climate model data needing only depth, sea level and wind stress which are in largely part of standard outputs of model efforts such as CMIP6. To ensure generalization, THOR has been designed and developed with interpretability and explainability in mind, towards a transparent method tailored for climate applications. THOR is a product of cross-fertilization between researchers in ML and in climate modeling, specifically oceanography. Domain expertise was key to validate the results of the k-means clustering and of the Ensemble MLP training + LRP relevance. However, assuring generalization means also that THOR should only be applied to similar horizontal resolutions than of the ECCO model on which it was trained. We note that the transparency of the predictive skill associated with THOR underscores its applicability to climate related research. This is because the predictions are rooted in known physics, and as such are less vulnerable to misclassification in out-of-sample applications.

4.2. Future directions and open questions for the machine learning community

Future work will include the application of THOR to several other CMIP6 models under climate change scenarios. Note that our method is fast and scalable (since we use it in a test mode), and could help accelerate large scale analysis of climate models.

From a ML perspective, we state here some of the important research questions where the ML community can have valuable contribution:

- The use of neural networks was motivated by their flexibility and because they can be trained efficiently with a large amount of data. Other ML techniques such as Random Forests (RF) might be interesting to consider especially when also using XAI methods specifically tailored for them (see Appendix A for an example)

- Investigating other XAI techniques such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) can be useful to measure how robust they are, and if they lead to the same conclusions as the ones found with LRP.

- Most importantly, several works in the literature have pointed out limitations of XAI techniques (Slack et al., 2020; Alvarez-Melis & Jaakkola, 2018). In this work, we used an ensemble to ensure more robust LRP predictions. Reseach efforts in making XAI techniques more robust and reliable, combined with incorporation of physical knowledge needs to be prioritized.

- The idea of using ensembles was also done to estimate the uncertainties of the predictions, and of the LRP explanations. Uncertainty quantification is without a doubt important for climate model applications. Developments in ML techniques that quantify the uncertainty associated with predictions are of pivotal importance. A potentially fruitful avenue of research might be the use of Bayesian Neural Networks (BNNs) instead of the Ensemble MLP, but then XAI methods adapted for BNNs need to be explored (Bykov et al., 2020).

.

Software and Data

The Earth System Grid Federation portal (ESGF) was used to get the CMIP6 data. 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`.

Acknowledgements

Authors would like to thank Kristopher Rand, Serguei Nikonov and Youngrak Cho for their valuable help. Authors would like also to thank the two anonymous reviewers. MS, AR and ZS and VB 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 and VB 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. We acknowledge the CMIP6 climate modeling groups for producing and making their model output available.

A. Random Forest based THOR

Step 2 of the THOR method can be conducted using other ML techniques, we developed also a Random Forest based version using LightGBMs. Confusion matrices of the prediction results can be found in Figure 4. Using split-based feature importance we notice that the wind stress curl feature is the most important feature for the classification which confirms results found the MLP used originally in THOR.

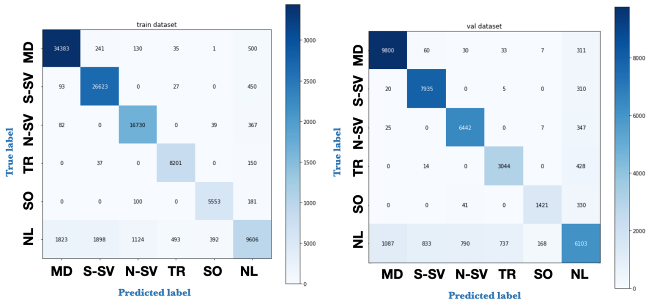

Figure 4. Confusion matrices for train and validation data when using the LightGBM based THOR.

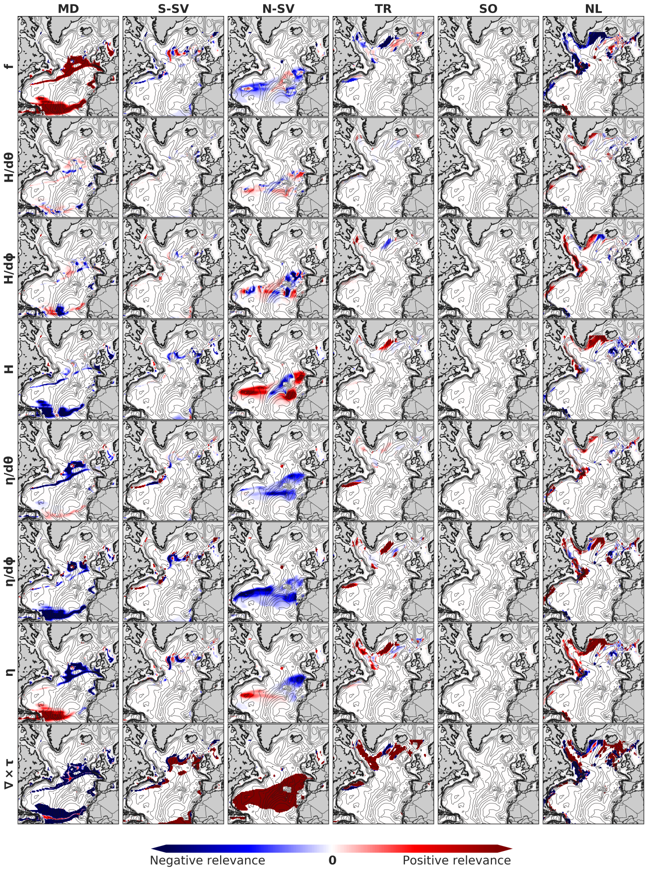

Figure 5. The spatial relevance maps. The dynamical regimes (columns) and the input features (rows) illustrating the contributions to the skill of the Ensemble MLP. The relevances are averaged for each point (lat, lon) over the Ensemble MLP. Figure from (Sonnewald & Lguensat, 2021)

B. LRP: Spatial representation of interpretability/explainability

The spatial maps of the interpretability (Figure 5) show intricate detail of what contributes to the Ensemble MLP learning. The mapping between feature relevance and the oceanographic 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 ocean 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.

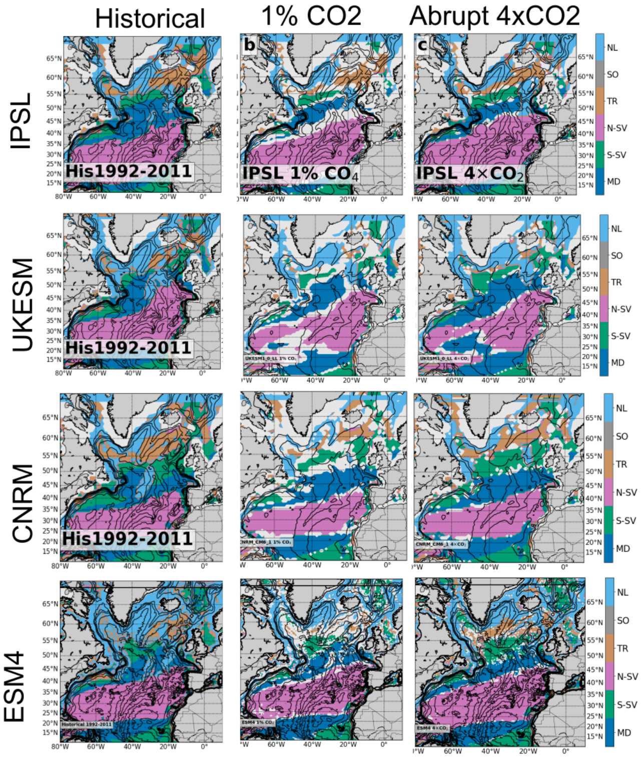

Figure 6. THOR applied to the North Atlantic.

C. Further THOR CMIP6 model applications

As an example of further applications of THOR, models from the CMIP6 ensemble were analysed and results presented for the North Atlantic (Figure 6), the Southern Ocean (Figure 7) and the Tropical Pacific (Figure 8). The data from these models is available through in the Amazon cloud as discussed and demonstrated in github.com/maikejulie/DNN4Cli.

The models used in addition to the ESM4 and IPSL model are the UK Earth System Model UKESM1-0-LL (Tang et al., 2019) and the French National Centre for Meteorological Research model CNRM-ESM2-1 (Seferian, 2018). Deeper analysis of present physics are the subject of present work.

References

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., and Kindermans, P.-J. innvestigate neural networks! *Journal of Machine Learning Research*, 20(93):1–8, 2019. URL http://jmlr.org/papers/v20/18-540.html.

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Balaji, V. Climbing down charney's ladder: Machine learning and the post-dennard era of computational climate science, 2020.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pp. 63–71. Springer, 2016.

Bingham, R. J. and Hughes, C. W. Signature of the atlantic meridional overturning circulation in sea level along the east coast of north america. *Geophysical Research Letters*, 36 (2), 2009. doi: https://doi.org/10.1029/2008GL036215. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008GL036215.

Boucher, O., Denvil, S., Levavasseur, G., Cozic, A., Caubel, A., Foujols, M.-A., Meurdesoif, Y., Cadule, P., Devilliers, M., Ghattas, J., Lebas, N., Lurton, T., Mellul, L., Musat, I., Mignot, J., and Cheruy, F. Ipsl ipsl-cm6a-lr model output prepared for cmip6 cmip, 2018. URL https://doi.org/10.22033/ESGF/CMIP6.1534.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., et al. Presentation and evaluation of the ipsl-cm6a-lr climate model. *Journal of Advances in Modeling Earth Systems*, 12(7): e2019MS002010, 2020.

Buckley, M. W. and Marshall, J. Observations, inferences, and mechanisms of the atlantic meridional overturning circulation: A review. *Reviews of Geophysics*, 54(1):5–63, 2016. doi: https://doi.org/10.1002/2015RG000493. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015RG000493.

Bykov, K., Höhne, M. M.-C., Müller, K.-R., Nakajima, S., and Kloft, M. How much can i trust you?–quantifying uncertainties in explaining neural networks. *arXiv preprint arXiv:2006.09000*, 2020.

Cheng, W., Chiang, J. C. H., and Zhang, D. Atlantic Meridional Overturning Circulation (AMOC) in CMIP5 Models: RCP and Historical Simulations. *Journal of Climate*, 26 (18):7187–7197, 09 2013. ISSN 0894-8755. doi: 10.1175/JCLI-D-12-00496.1. URL https://doi.org/10.1175/JCLI-D-12-00496.1.

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

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4): 303–314, 1989.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., and Sculley, D. Underspecification presents challenges for credibility in modern machine learning, 2020.

Ding, H., Newman, M., Alexander, M. A., and Wittenberg, A. T. Relating cmip5 model biases to seasonal forecast skill in the tropical pacific. *Geophysical Research Letters*, 47(5): e2019GL086765, 2020.

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., Krasting, J. P., Malyshev, S., Naik, V., Paulot, F., Shevliakova, E., Stock, C. A., Zadeh, N., Balaji, V., Blanton, C., Dunne, K. A., Dupuis, C., Durachta, J., Dussin, R., Gauthier, P. P. G., Griffies, S. M., Guo, H., Hallberg, R. W., Harrison, M., He, J., Hurlin, W., McHugh, C., Menzel, R., Milly, P. C. D., Nikonov, S., Paynter, D. J., Ploshay, J., Radhakrishnan, A., Rand, K., Reichl, B. G., Robinson, T., Schwarzkopf, D. M., Sentman, L. T., Underwood, S., Vahlenkamp, H., Winton, M., Wittenberg, A. T., Wyman, B., Zeng, Y., and Zhao, M. 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, 2020. doi: https://doi.org/10.1029/2019MS002015. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002015. e2019MS002015 2019MS002015.

Eyring, V., Bony, S., Meehl, G., Senior, C., Stevens, B., Ronald, S., and Taylor, K. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organisation. *Geoscientific Model Development Discussions*, 8:10539–10583, 12 2015. doi: 10.5194/gmdd-8-10539-2015.

Eyring, V., Cox, P., Flato, G., Gleckler, P., Abramowitz, G., Caldwell, P., Collins, W., Gier, B., Hall, A., Hoffman, F., Hurtt, G., Jahn, A., Jones, C., Klein, S., Krasting, J., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G., and Williamson, M. Taking climate model evaluation to the next level. *Nature Climate Change*, 9:102–110, 02 2019.

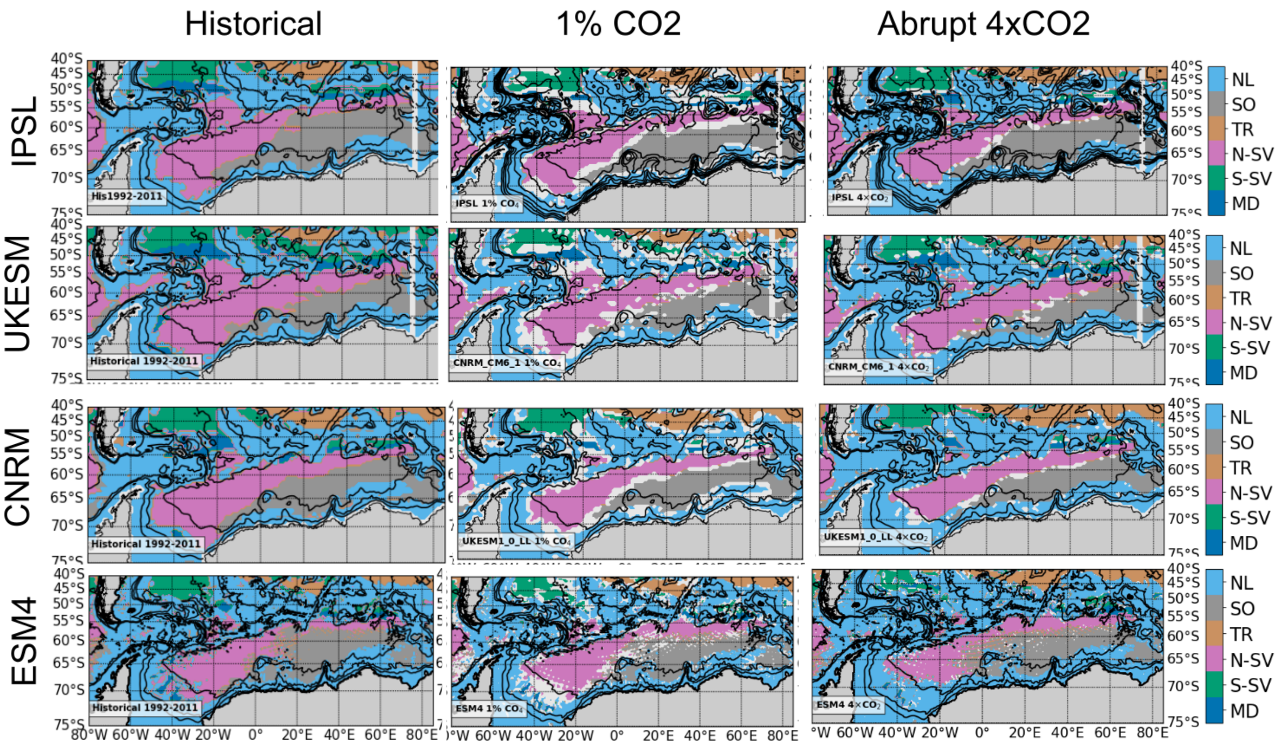

Figure 7. THOR applied to the Southern Ocean.

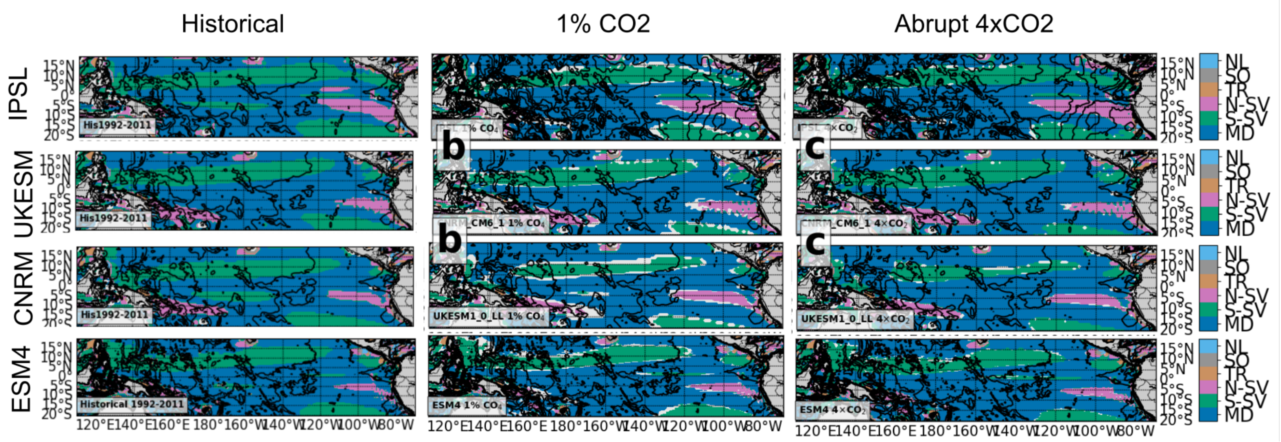

Figure 8. THOR applied to the Tropical Pacific.

Fedorov, A. V., Hu, S., Wittenberg, A. T., Levine, A. F., and Deser, C. Enso low-frequency modulation and mean state interactions. *El Niño Southern Oscillation in a Changing Climate*, pp. 173–198, 2020.

Forget, G., Campin, J.-M., Heimbach, P., Hill, C., Ponte, R., and Wunsch, C. Ecco version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation. *Geoscientific Model Development Discussions*, 8:3653–3743, 05 2015. doi: 10.5194/gmdd-8-3653-2015.

Guilyardi, E., Capotondi, A., Lengaigne, M., Thual, S., and Wittenberg, A. T. Enso modeling: History, progress, and challenges. *El Niño Southern Oscillation in a Changing Climate*, pp. 199–226, 2020.

Hamman, J., Rocklin, M., and Abernathy, R. Pangeo: a big-data ecosystem for scalable earth system science, 2018.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

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

Hughes, C. W. and de Cuevas, B. A. 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, 10 2001. ISSN 0022-3670. doi: 10.1175/1520-0485(2001)031⟨2871:WWBCIR⟩2.0.CO;2. URL 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., and Saynisch-Wagner, J. Will artificial intelligence supersede earth system and climate models?, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014.

Krasting, J. P., John, J. G., Blanton, C., McHugh, C., Nikonov, S., Radhakrishnan, A., Rand, K., Zadeh, N. T., Balaji, V., Durachta, J., Dupuis, C., Menzel, R., Robinson, T., Underwood, S., Vahlenkamp, H., Dunne, K. A., Gauthier, P. P., Ginoux, P., Griffies, S. M., Hallberg, R., Harrison, M., Hurlin, W., Malyshev, S., Naik, V., Paulot, F., Paynter, D. J., Ploshay, J., Schwarzkopf, D. M., Seman, C. J., Silvers, L., Wyman, B., Zeng, Y., Adcroft, A., Dunne, J. P., Dussin, R., Guo, H., He, J., Held, I. M., Horowitz, L. W., Lin, P., Milly, P., Shevliakova, E., Stock, C., Winton, M., Xie, Y., and Zhao, M. Noaa-gfdl gfdl-esm4 model output prepared for cmip6 cmip, 2018. URL https://doi.org/10.22033/ESGF/CMIP6.1407.

Lapuschkin, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10:e0130140, 07 2015. doi: 10.1371/journal.pone.0130140.

Larson, S. M., Buckley, M. W., and Clement, A. C. Extracting the Buoyancy-Driven Atlantic Meridional Overturning Circulation. *Journal of Climate*, 33(11):4697–4714, 05 2020. ISSN 0894-8755. doi: 10.1175/JCLI-D-19-0590.1. URL https://doi.org/10.1175/JCLI-D-19-0590.1.

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

Losch, M., Adcroft, A., and Campin, J.-M. How sensitive are coarse general circulation models to fundamental approximations in the equations of motion? *Journal of Physical Oceanography*, 34(1):306 – 319, 2004. doi: 10.1175/1520-0485(2004)034⟨0306:HSACGC⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/phoc/34/1/1520-0485_2004_034_0306_hsacgc_2.0.co_2.xml.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predicti pdf.

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

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., and Stouffer, R. J. The coupled model intercomparison project (cmip). *Bulletin of the American Meteorological Society*, 81(2):313–318, 2000.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F., Stouffer, R. J., and Taylor, K. E. The wcrp cmip3 multimodel dataset: A new era in climate change research. *Bulletin of the American meteorological society*, 88(9):1383–1394, 2007.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, May 2017. ISSN 0031-3203. doi: 10.1016/j.patcog.2016.11.008. URL http://dx.doi.org/10.1016/j.patcog.2016.11.008.

Olden, J. D., Joy, M. K., and Death, R. G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3):389 – 397, 2004. ISSN 0304-3800. doi: https://doi.org/10.1016/j.ecolmodel.2004.03.013. URL http://www.sciencedirect.com/science/article/pii/S0304380004001565.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

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

Ribeiro, M. T., Singh, S., and Guestrin, C. "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, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144, 2016.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 05 2019. doi: 10.1038/s42256-019-0048-x.

Rumelhart, D., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

Schlund, M., Eyring, V., Camps-Valls, G., Friedlingstein, P., Gentine, P., and Reichstein, M. Constraining uncertainty in projected gross primary production with machine learning. *Journal of Geophysical Research: Biogeosciences*, 125(11):e2019JG005619, 2020. doi: https://doi.org/10.1029/2019JG005619. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JG005619. e2019JG005619 2019JG005619.

Seferian, R. Cnrm-cerfacs cnrm-esm2-1 model output prepared for cmip6 cmip, 2018. URL https://doi.org/10.22033/ESGF/CMIP6.1391.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Sonnewald, M. and Lguensat, R. Revealing the impact of global heating on north atlantic circulation using transparent machine learning. *Journal of Advances in Modeling Earth Systems*, n/a(n/a):e2021MS002496, 2021. doi: https://doi.org/10.1029/2021MS002496. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002496. e2021MS002496 2021MS002496.

Sonnewald, M., Lguensat, R., Jones, D. C., Dueben, P. D., Brajard, J., and Balaji, V. Bridging observation, theory and numerical simulation of the ocean using machine learning. *arXiv preprint arXiv:2104.12506*, 2021.

Stevenson, S., Wittenberg, A. T., Fasullo, J., Coats, S., and Otto-Bliesner, B. Understanding diverse model projections of future extreme el niño. *Journal of Climate*, 34(2):449–464, 2021.

Tang, Y., Rumbold, S., Ellis, R., Kelley, D., Mulcahy, J., Sellar, A., Walton, J., and Jones, C. Mohc ukesm1.0-ll model output prepared for cmip6 cmip historical, 2019. URL https://doi.org/10.22033/ESGF/CMIP6.6113.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498, 2012.

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

Wang, Z., Lu, Y., Dupont, F., W. Loder, J., Hannah, C., and G. Wright, D. Variability of sea surface height and circulation in the north atlantic: Forcing mechanisms and linkages. *Progress in Oceanography*, 132:273 – 286, 2015. ISSN 0079-6611. doi: https://doi.org/10.1016/j.pocean.2013.11.004. URL http://www.sciencedirect.com/science/article/pii/S0079661113002231. Oceanography of the Arctic and North Atlantic Basins.

Weaver, A. J., Sedláček, J., Eby, M., Alexander, K., Crespin, E., Fichefet, T., Philippon-Berthier, G., Joos, F., Kawamiya, M., Matsumoto, K., Steinacher, M., Tachiiri, K., Tokos, K., Yoshimori, M., and Zickfeld, K. Stability of the atlantic meridional overturning circulation: A model intercomparison. *Geophysical Research Letters*, 39(20), 2012. doi: https://doi.org/10.1029/2012GL053763. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012GL053763.

Weijer, W., Cheng, W., Garuba, O. A., Hu, A., and Nadiga, B. T. Cmip6 models predict significant 21st century decline of the atlantic meridional overturning circulation. *Geophysical Research Letters*, 47(12):e2019GL086075, 2020. doi: https://doi.org/10.1029/2019GL086075. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL086075. e2019GL086075 10.1029/2019GL086075.

Wunsch, C. and Heimbach, P. Chapter 21 - dynamically and kinematically consistent global ocean circulation and ice state estimates. In Siedler, G., Griffies, S. M., Gould, J., and Church, J. A. (eds.), *Ocean Circulation and Climate*, volume 103 of *International Geophysics*, pp. 553 – 579. Academic Press, 2013. doi: https://doi.org/10.1016/B978-0-12-391851-2.00021-0. URL http://www.sciencedirect.com/science/article/pii/B9780123918512000210.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks, 2013.

Zhang, R., Sutton, R., Danabasoglu, G., Kwon, Y.-O., Marsh, R., Yeager, S. G., Amrhein, D. E., and Little, C. M. 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, 2019. doi: 10.1029/2019RG000644. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019RG000644.