JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2022MS003162

Special Section:
Machine learning application to Earth system modeling

Explainable Artificial Intelligence for Bayesian Neural Networks: Toward Trustworthy Predictions of Ocean Dynamics

Mariana C. A. Clare[1,2], Maike Sonnewald[3,4,5], Redouane Lguensat[6], Julie Deshayes[7], and V. Balaji[3,4,8]

[1]Imperial College London, London, UK, [2]ECMWF, Bonn, Germany, [3]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA, [4]Ocean and Cryosphere Division, NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA, [5]University of Washington, Seattle, WA, USA, [6]Institut Pierre-Simon Laplace, IRD, Sorbonne Université, Paris, France, [7]LOCEAN-IPSL, CNRS, Sorbonne Université, Paris, France, [8]Laboratoire des Sciences du Climat et de l'Environnement, CEA Saclay, Gif Sur Yvette, France

Key Points:

- Novel use of a Bayesian Neural Network (BNN) to quantify uncertainty in ocean dynamical regime classifications, giving a holistic prediction
- Explaining the skill of a BNN using two techniques originating from two different classes of explainable AI: SHapley Additive exPlanation (SHAP) and Layer-wise Relevance Propagation (LRP)
- Trustworthiness is evaluated by comparing similarities and differences between SHAP and LRP explanations with intuition from physical theory

Correspondence to:
M. C. A. Clare,
mariana.clare@ecmwf.int

Citation:
Clare, M. C. A., Sonnewald, M., Lguensat, R., Deshayes, J., & Balaji, V. (2022). Explainable artificial intelligence for Bayesian Neural Networks: Toward trustworthy predictions of ocean dynamics. *Journal of Advances in Modeling Earth Systems*, *14*, e2022MS003162. https://doi.org/10.1029/2022MS003162

Received 28 APR 2022
Accepted 5 OCT 2022

Author Contributions:
Conceptualization: Mariana C. A. Clare, Redouane Lguensat
Formal analysis: Mariana C. A. Clare, Maike Sonnewald, Redouane Lguensat
Investigation: Mariana C. A. Clare
Methodology: Mariana C. A. Clare
Project Administration: Julie Deshayes

Abstract The trustworthiness of neural networks is often challenged because they lack the ability to express uncertainty and explain their skill. This can be problematic given the increasing use of neural networks in high stakes decision-making such as in climate change applications. We address both issues by successfully implementing a Bayesian Neural Network (BNN), where parameters are distributions rather than deterministic, and applying novel implementations of explainable AI (XAI) techniques. The uncertainty analysis from the BNN provides a comprehensive overview of the prediction more suited to practitioners' needs than predictions from a classical neural network. Using a BNN means we can calculate the entropy (i.e., uncertainty) of the predictions and determine if the probability of an outcome is statistically significant. To enhance trustworthiness, we also spatially apply the two XAI techniques of Layer-wise Relevance Propagation (LRP) and SHapley Additive exPlanation (SHAP) values. These XAI methods reveal the extent to which the BNN is suitable and/or trustworthy. Using two techniques gives a more holistic view of BNN skill and its uncertainty, as LRP considers neural network parameters, whereas SHAP considers changes to outputs. We verify these techniques using comparison with intuition from physical theory. The differences in explanation identify potential areas where new physical theory guided studies are needed.

Plain Language Summary Understanding ocean dynamics and how they are affected by global heating is crucial for understanding climate change impacts. Neural networks are ideally suited to this problem, but do not explain how they make predictions nor express how certain they are of the predictions' accuracy, which considerably limits their trustworthiness for ocean science problems. Here, we address both issues by using a "Bayesian Neural Network" (BNN), which directly expresses prediction uncertainty, and applying explainable AI (XAI) techniques to explain how the BNN arrives at its prediction. The BNN provides a comprehensive overview more suited to addressing the core problem than that provided by classical neural networks. We also apply two XAI techniques (SHAP and LRP) to the BNN and evaluate their trustworthiness by comparing the similarities and differences between their explanations with intuition from physical theory. Any differences offer an opportunity to develop physical theory guided by what the BNN considers important.

1. Introduction

There is already scientific certainty that global heating is changing the climate, but understanding exactly how the climate will change and the potential impacts is an open problem. Increasingly, artificial intelligence techniques, such as neural networks, are being used to better understand climate change (e.g., Ham et al., 2019; Huntingford et al., 2019; Rolnick et al., 2019; Cowls et al., 2021), but as neural network techniques become evermore ubiquitous, there is a growing need for methods to quantify their trustworthiness and uncertainty (Li et al., 2021; Mamalakis et al., 2021). Following Sonnewald and Lguensat (2021), we define a method to be trustworthy if its results are explainable and interpretable, and therefore these two concepts are somewhat linked as improving uncertainty quantification also improves result interpretability. Quantifying uncertainty using classical neural networks is particularly difficult because they lack the ability to express it and are often overconfident in their results (Joo et al., 2020; Mitros & Mac Namee, 2019). A range of techniques have been used to address this

Software: Mariana C. A. Clare, Maike Sonnewald, Redouane Lguensat
Supervision: Maike Sonnewald, Redouane Lguensat, Julie Deshayes, V. Balaji
Validation: Mariana C. A. Clare
Visualization: Mariana C. A. Clare
Writing – original draft: Mariana C. A. Clare
Writing – review & editing: Maike Sonnewald, Redouane Lguensat, Julie Deshayes, V. Balaji

uncertainty quantification issue (Guo et al., 2017) and a particularly common one is to use an ensemble of deep learning models (e.g., Beluch et al., 2018). However, choosing a good ensemble of models is non-trivial (see Scher & Messori, 2021) and may be computationally expensive because it requires the network to be trained multiple times. This lack of uncertainty analysis limits the extent to which classical neural networks can be useful for ocean and climate science problems. For example, lack of knowledge of uncertainties in future projections of sea level rise limits how effective coastal protection measures can be for coastal communities (Sánchez-Arcilla et al., 2021). Measures of uncertainty are also important for out-of-sample predictions, which are common in climate change science because neural networks must be trained on historical data and applied to a changed climate scenario where the dynamics governing a region may have fundamentally changed. Thus, quantifying uncertainty within a climate application is of paramount importance as decisions based on neural network predictions could have wide ranging impacts. Moreover, there can be distrust of neural network predictions in the climate science community because of the potential for spurious correlations giving rise to predictions that are nonphysical. Predictions are more trustworthy if they are explainable (i.e., if the reason why the network predicted the result can be understood by members of the climate science community). However, adding explainability techniques to uncertainty analysis is an understudied area.

In this work, we address both issues of uncertainty and trustworthiness by implementing a Bayesian Neural Network (BNN) (Jospin et al., 2020) with novel implementations of explainable AI techniques (known as XAI) (Samek et al., 2021). We focus on applying this technique to assess uncertainty in dynamical ocean regime predictions due to a changing climate following the THOR (Tracking global Heating with Ocean Regimes) framework (Sonnewald & Lguensat, 2021). This is the first time Bayesian Neural Networks (BNNs) have been used to predict large-scale ocean circulations, although they have been used for localized streamflows in Rasouli et al. (2012, 2020). Our work is particularly pertinent with a recent IPCC Special Report (Hoegh-Guldberg et al., 2018) highlighting uncertainty in ocean circulation as a key knowledge gap area that must be addressed. Both (Sonnewald & Lguensat, 2021) and our work are designed to predict future changes to ocean circulation using data from the sixth phase of the Coupled Model Intercomparison Project (CMIP) (used in IPCC reports) (Eyring et al., 2015). We note however that, as CMIP6 is a large international collaboration, data dissemination and quality control can be difficult, which in turn limits the capability for good analysis. Sonnewald and Lguensat (2021) is an example of using sparse data in this context, and resolving this issue generally is an area of ongoing research (Eyring et al., 2019).

Unlike classical neural networks, BNNs make well-calibrated uncertainty predictions (Jospin et al., 2020; Mitros & Mac Namee, 2019) and clearly inform the user of how unsure the outcome is. This provides a more comprehensive description of the neural network prediction compared to a classical neural network and one which better meets the needs of climate and ocean science researchers. Furthermore, the uncertainty measures provided by the BNN approach can help reveal whether a new unseen datapoint is out-of-distribution relative to the training data (see e.g., Jospin et al., 2020). For example, it is known that the wind stress over the Southern Ocean will change in the future, with implications for the dynamics key to maintaining global scale heat transport. However, the region already has extreme conditions, so a change here could result in entirely new dynamical connections. The BNN outputs would allow us to understand if the new conditions are out-of-distribution compared to the original training data and thus whether the BNN's categorization of the dynamical regime for the new conditions can still be trusted. This uncertainty analysis is possible in BNNs because the weights, biases and/or outputs are distributions rather than deterministic point values. Moreover, these distributions mean BNNs can easily be used as part of an ensemble approach (a very common approach in climate science), by simply sampling point estimates from the trained distributions to generate an ensemble (Bykov et al., 2020).

Using BNNs is a large step toward trustworthy predictions, but results also gain considerable trustworthiness to climate researchers and practitioners if their skill is physically explainable. Note that throughout we define explaining skill to mean explaining the correlations between the input features that give rise to the predictions. Governments and regulatory bodies are also increasingly imposing regulations that require trustworthiness in AI processes used in certain decision-making (see Cath et al., 2018) and imposing large fines if the standards are not met (see e.g., recent directives from the European Commission 2021 and the USA government [E.O. 13960 of 3 December 2020]). XAI techniques can be used to explain the skill of neural networks (Arrieta et al., 2020; Samek et al., 2019, 2021), but there has been little work combining explainability with uncertainty analysis in part because the distributions in BNNs add extra complexity. In this work, we adapt two common XAI techniques so that they can be used to explain the skill in BNN results for one of the first times: Layer-wise Relevance Propagation (LRP) (Binder et al., 2016) (previously applied using the same approach in Bykov et al., 2020) and SHapley Additive exPlanation (SHAP) values (Lundberg & Lee, 2017) (previously applied in Cui et al. (2019); Yao et al. (2021) but using a different approach). These XAI methods reveal the extent to which the BNN is fit for purpose for our problem. Moreover, our approach means we can gain a reliable notion of the confidence of the explanation, which has been highlighted as a key area where XAI techniques must improve (Lakkaraju et al., 2022). Applying our XAI techniques to BNNs trained on real-world ocean circulation data in an application designed to understand future climate has the added benefit that we are able to validate and confirm these novel applications of XAI using physical understanding of ocean circulation processes, improving confidence in our BNN predictions. Thus, our novel framework is able to quantify uncertainty and improve trustworthiness (i.e., explainability and interpretability) in predictions, marking a significant step forward for using neural networks in climate and ocean science.

In this work, we choose to apply two different XAI techniques specifically to gain a holistic view of the skill of the BNN as LRP considers the neural network parameters whereas SHAP considers the impact of changing input features on the BNN outputs. This is important to ensure that what the BNN has learned is genuinely rooted in physical theory. The two different approaches also give a more overall impression of uncertainty as they capture different aspects with LRP capturing model uncertainty and SHAP capturing prediction sensitivity to this model uncertainty. Furthermore, by considering two different techniques, we can explore whether they agree as to which features are important in each area of the domain. This allows us test if the "disagreement problem" exists in this work, where two techniques explain network skill in different ways (Krishna et al., 2022), which is a growing area of interest in XAI research.

To summarize the main contributions of our work are that we present the first application of BNNs to quantify uncertainty in large-scale ocean circulation predictions, and explain the skill of these predictions through novel implementations of the XAI techniques, SHAP and LRP, thereby improving trustworthiness. The remainder of this paper is structured as follows: Section 2 explores the theory behind BNNs and applying XAI techniques to BNNs, Section 3 explores the data set used to train the BNN, Section 4 shows the results of applying the BNN and novel XAI techniques to the data set and finally Section 5 concludes this work.

2. Methods

The main methods used in this work are a BNN which we combine with two XAI techniques to explain the BNN's predictions. The general workflow is summarized in pictorial form in Figure 1. In this section, we detail each of the components in this figure, discussing BNNs in Section 2.1 and the two XAI techniques in Section 2.2.

2.1. Bayesian Neural Networks

Unlike classical deterministic neural networks, BNNs are capable of making well-calibrated uncertainty predictions, which provide a measure of the uncertainty of the outcome (Jospin et al., 2020). This is possible due to the fact that the weights and biases on at least some of the layers in the network are distributions rather than single point estimates (see Figure 2). More specifically, as BNNs use a Bayesian framework, once trained, the distributions of the weights and biases represent the posterior distributions based on the training data (Bykov et al., 2020). Note that for brevity in this section hereafter, we refer to the weights and biases as network parameters. The distributions in the output layer facilitate the assessment of aleatoric uncertainty (uncertainty in the data) and the distributions in the hidden layers facilitate the assessment of epistemic uncertainty (uncertainty in the model) (Salama, 2021). In this work, we choose to assess both types of uncertainty and use distributions for the output layer, as well as for the network parameters of the hidden layers. Our BNN approach therefore provides a more holistic view than previous work to assess uncertainty in large-scale ocean neural network predictions in Gordon and Barnes (2022) where a deterministic neural network is used to predict the mean and variance of the output distribution.

Following Jospin et al. (2020), the posterior distributions in the BNN (i.e., the distributions of the network parameters given the training data) are calculated using Bayes rule

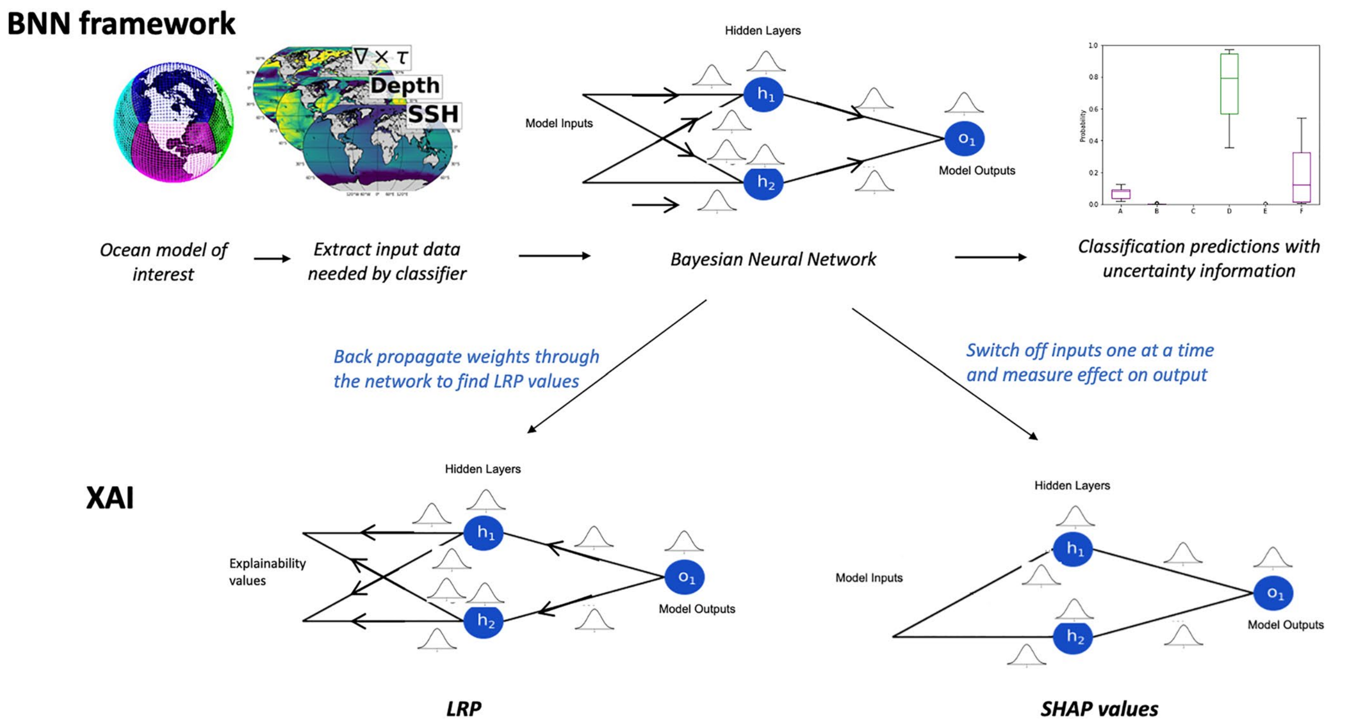

BNN framework
$\nabla \times \tau$
Depth
SSH
Hidden Layers
Model Inputs
h1
h2
o1
Model Outputs
Probability
A
B
C
D
E
F
Ocean model of interest
Extract input data needed by classifier
Bayesian Neural Network
Classification predictions with uncertainty information
Back propagate weights through the network to find LRP values
Switch off inputs one at a time and measure effect on output
XAI
Explainability values
LRP
SHAP values

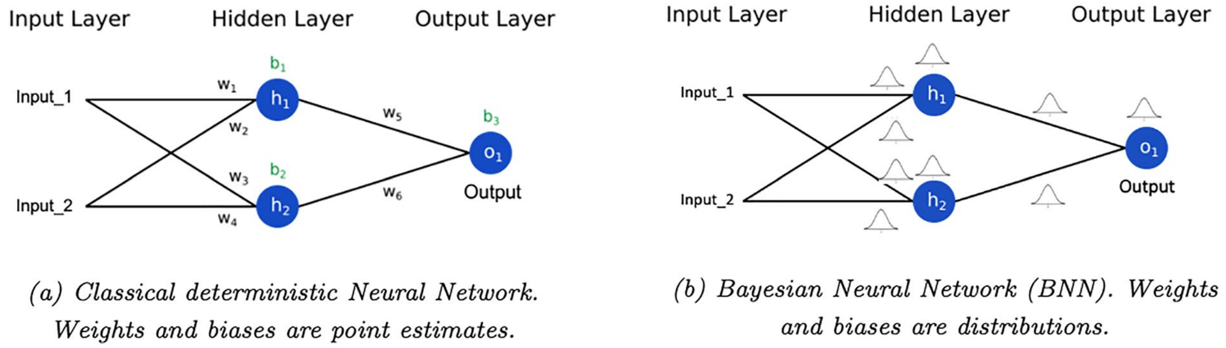
Figure 1. Detailed sketch of workflow used in this work to obtain and explain neural network predictions from ocean input data. This sketch includes both the Bayesian Neural Network and the two explainable AI techniques used. Note this figure adapts images used in Figure 5 of Sonnewald and Lguensat (2021).

$$p(W|D_{tr}) = \frac{p(D_{tr}|W)p(W)}{p(D_{tr})} = \frac{p(D_{tr}|W)p(W)}{\int_W p(D_{tr}|W)p(W)\,dW}, \tag{1}$$

where W are the network parameters, $D_{tr} = (x_n, y_n)$ the training data and $p(W)$ the prior distribution of the parameters. The probability of output y given input x is then given by the marginal probability distribution

$$p(y|x, D_{tr}) = \int_W p(y|f(x;W))p(W|D_{tr})\,dW, \tag{2}$$

where $f(\cdot;\, W)$ is the neural network. However, computing $p(W|D_{tr})$ directly is very difficult, especially due to the denominator in Equation 1 which is intractable (Bykov et al., 2020; Jospin et al., 2020). A number of methods have been proposed to approximate the denominator term including Markov Chain Monte Carlo sampling (Titterington, 2004) and variational inference (Osawa et al., 2019). We use the latter which approximates the posterior using a variational distribution, $q_\Phi(W)$, with a known formula dependent on the parameters, Φ, that



(a) *Classical deterministic Neural Network. Weights and biases are point estimates.*

(b) *Bayesian Neural Network (BNN). Weights and biases are distributions.*

Figure 2. Comparing a standard neural network to a Bayesian Neural Network.

define the distribution (e.g., for a normal distribution, Φ are its mean and variance). The BNN then learns the parameters Φ which lead to the closest match between the variational distribution and the posterior distribution that is, the parameters Φ which minimize the following Kullback–Leibler divergence (KL-divergence)

$$D_{KL}(q_{\Phi}||p) = \int_W q_{\Phi}\left(W'\right) \log\left(\frac{q_{\Phi}(W')}{p(W'|D_{tr})}\right) dW'. \tag{3}$$

This formula still requires the posterior to be computed and so following standard practice, we use the ELBO formula instead

$$\int_W q_{\Phi}\left(W'\right) \log\left(\frac{p(W', D_{tr})}{q_{\Phi}(W')}\right) dW', \tag{4}$$

which is equal to $\log(p(D_{tr})) - D_{KL}(q_{\Phi}\|p)$. Thus maximizing (Equation 4) is equivalent to minimizing (Equation 3) since $\log(p(D_{tr}))$ only depends on the prior (Jospin et al., 2020). In our work, we follow standard practice and assume that all variational forms of the posterior are normal distributions and thus the Φ parameters the neural network learns are the mean and variance of these distributions. Furthermore, for all priors in the BNN, we use the normal distribution $\mathcal{N}(0, 1)$, which is again standard practice because of the normal distribution's mathematical properties and simple log-form (Silvestro & Andermann, 2020). We note briefly here that whilst probabilistic predictions could be achieved by instead using an ensemble of deterministic neural networks, this would not only take much longer to train but also lead to overconfident results (Joo et al., 2020). By contrast, the Bayesian approach in the BNN results in a more accurate representation of confidence.

In our work, we also calculate the entropy of the final distribution as a measure of uncertainty. In information theory, entropy is considered as the expected information of a random variable and for each sample i is given by

$$H_i = -\sum_{j=1}^{N_l} p_{ij} \log\left(p_{ij}\right), \tag{5}$$

where N_l is the number of possible variable outcomes and p_{ij} is the probability of each outcome j for sample i (Goodfellow et al., 2016). Hence, the larger the entropy value, the less skewed the distribution and the more uncertain the model is of the result.

Finally, for the layer architecture of the BNN, we use the same architecture as in Sonnewald and Lguensat (2021), who use a deterministic neural network to predict ocean regimes from the same data set as ours (see Section 3). Thus, our BNN has four layers with [24, 24, 16, 16] nodes and “tanh” activation, where the layers are “DenseVariational” layers from the TensorFlow probability library (Dillon et al., 2017), rather than the “Dense” layers used in Sonnewald and Lguensat (2021). For the output layer of the network, we use the “OneHotCategorical” layer from the TensorFlow probability library instead of a “SoftMax” layer and thus use the negative log-likelihood function as the loss function. The network is compiled with an Adam Optimizer (Kingma & Ba, 2014) with an initial learning rate of 0.01, which is reduced by a factor of four if the loss metric on the validation data set does not decrease after 15 epochs (i.e., after the entire training data set has passed through the neural network 15 times). The network is trained for 100 epochs and the best model network parameters over all epochs are recorded and saved as the trained parameters.

2.2. Explainable AI

Whilst using a BNN enables scientists to determine how certain the network is of its results, being able to explain the source of the predictive skill is also of key importance particularly because of the potential for spurious correlations in neural networks giving rise to nonphysical predictions. As discussed in Section 1, XAI techniques have recently been developed to explain the skill of neural networks (i.e., explain the correlations between the input features that give rise to the predictions). These techniques can then be used to reveal the extent to which neural networks are fit for purpose for a given problem (Arrieta et al., 2020; Samek et al., 2019). However, there has been little research into combining XAI techniques with uncertainty analysis. In this section, we outline how to adapt the two common XAI techniques, LRP and SHAP, so that they can be applied to BNNs. We remind the reader that we selected two XAI techniques originating from two different classes to gain a holistic view of the skill of the BNN. This is important to ensure that what the BNN has learned is genuinely rooted in physical theory, and we compare the outcomes of these methods with intuition from that theory.

2.2.1. Layer-Wise Relevance Propagation

LRP explains network skill by calculating the contribution (or *relevance*) of each input datapoint to the output score (Binder et al., 2016). This leads to the construction of a "heatmap" where a positive/negative "relevance" means a feature contributes positively/negatively to the output (Bach et al., 2015). For a neural network, this relevance is calculated by back-propagating the relevance layer-by-layer from the output layer to the input layer.

LRP has been successfully used to explain neural network skill in fields as diverse as medicine (Böhle et al., 2019), information security (Seibold et al., 2020) and text analysis (Arras et al., 2017), and has also already been applied to deterministic neural networks in climate science (Mamalakis et al., 2022; Sonnewald & Lguensat, 2021; Toms et al., 2020). However, there has been little research into applying LRP to BNNs, because the formulae used to calculate the relevance are difficult to apply when the network parameters are distributions.

BNNs do however have the advantage that it is easy to generate a deterministic ensemble of networks from them, simply by sampling network parameters from the distributions. We therefore follow the novel methodology in Bykov et al. (2020) and use LRP on this ensemble of networks, efficiently generating an ensemble of LRP values which serve as a proxy for explaining the skill of the BNN. Each datapoint has its own distribution of LRP values and own level of uncertainty. If a datapoint has positive or negative relevance for every ensemble member, we can be increasingly confident about this point's relevance for explaining the skill of the BNN. For the remaining points, still following (Bykov et al., 2020), quantile heatmaps of the ensemble of LRP values can be used to visualize how many ensemble members have positive relevance and how many have negative.

There are many different formulae for calculating the relevance score with LRP (see Montavon et al., 2019), but in this work, we follow Sonnewald and Lguensat (2021) and use the LRP-ϵ rule which is good for handling noise. The relevance at layer l of a neuron i is then the sum of $R^{(l,l+1)}_{i\leftarrow j}$ for all neurons j in layer $l+1$ where

$$R^{(l,l+1)}_{i\leftarrow j} = \frac{z_{ij}}{z_j + \epsilon \ \mathrm{sign}\,(z_j)} R^{(l+1)}_j . \tag{6}$$

Here z_{ij} is the activation at neuron i multiplied by the weight from neuron i to j, $z_j = \sum_i z_{ij}$ and ϵ is an arbitrary small positive number which is here chosen to be 10^{-7} (see Montavon et al. (2019) for more details).

2.2.2. SHapley Additive exPlanation Values

For our second XAI technique, we consider SHAP values, known more commonly as SHAP values. These were first proposed in the context of game theory in Shapley (1953), but have since been extended to explaining skill in neural networks (Lundberg & Lee, 2017) and have been applied in climate science to deterministic neural networks in Dikshit and Pradhan (2021); Mamalakis et al. (2022). There has been work adding uncertainty to the SHAP values of deterministic neural networks by adding noise (Slack et al., 2021), but this work represents one of the first times SHAP values are used to explain the skill of a BNN.

SHAP values are designed to compute the contribution of each input datapoint to the neural network output using a type of occlusion analysis. They test the effect of removing/adding a feature to the final output that is, calculating $f_F(x) - f_{F\backslash i}(x)$, where f is the model, F is the set of all features and i the feature being considered (Lundberg & Lee, 2017). To calculate the SHAP value, we must combine this for all features in the model with a weighted average meaning the SHAP value of feature i for output $y = f_F(x)$ is

$$\phi_i(x) = \sum_{S\subset F\backslash i} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S\cup\{i\}}(x) - f_S(x) \right], \tag{7}$$

where S are all the sub-sets of F excluding feature i. Note that summing the SHAP value for every feature i gives the difference between the model prediction and the null model that is,

$$f_F(x) = \mathbb{E}[y] + \sum_i \phi_i(x), \tag{8}$$

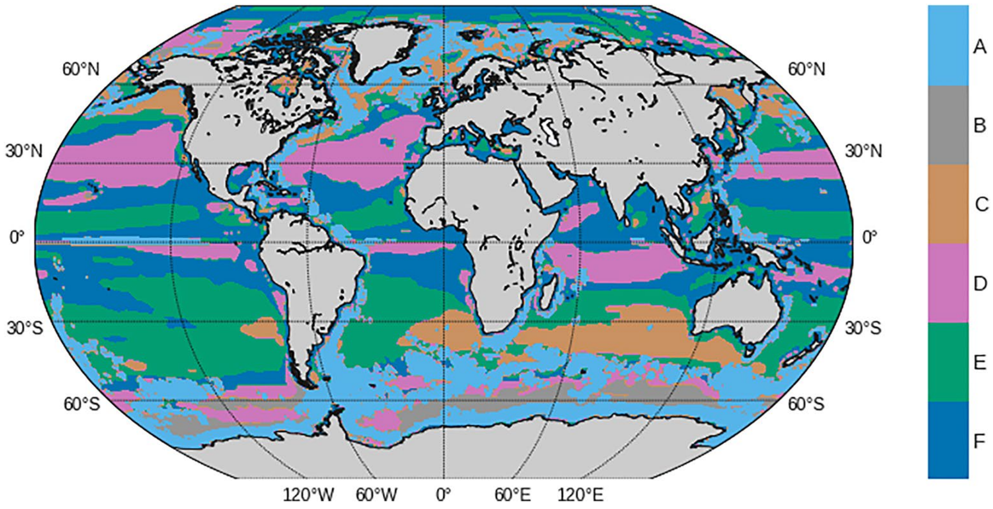


Figure 3. Global representation of dynamical ocean regimes in Estimating the Circulation and Climate of the Ocean data. For a full description of the ocean regimes see Sonnewald and Lguensat (2021).

where $\mathbb{E}[y]$ is the average of all outputs y in the training data set (Mazzanti, 2020). We remark here that evaluating (7) for every feature can be computationally expensive; the complexity of the problem scales by $2^{|F|}$. Therefore various techniques have been proposed to speed up the evaluation of SHAP values, the most popular of which is KernelSHAP (Lundberg & Lee, 2017). In this work, however, we choose to calculate the exact SHAP values because we only have eight features (see Section 3) and these more efficient techniques assume feature independence (which our data set does not have), and can lead to compromises on accuracy if not handled appropriately (Aas et al., 2021).

Like with LRP, we apply SHAP to an ensemble of deterministic neural networks generated from the BNN. We note here that SHAP is model agnostic so in the future, with changes to implementation, it may be possible to apply SHAP directly to the BNN itself. We expect the SHAP results to differ from the LRP results because the LRP ensemble captures the model uncertainty as LRP values are a weighted sum of the network weights, whereas SHAP captures the sensitivities of the outputs as a result of these uncertainties.

3. Data

A recent IPCC Special report highlights the need for a better understanding of uncertainty in ocean circulation patterns (Hoegh-Guldberg et al., 2018). An understanding of emergent circulation patterns can be gained using a dynamical regime framework (Sonnewald et al., 2019). These regimes simplify dynamics and each regime is then defined to be the solution space where the simplification is justifiable (Kaiser et al., 2021). Sonnewald et al. (2019) show that unsupervised clustering techniques such as *k*-means clustering can be used to identify and partition dynamical regimes if the equations governing the dynamics are known. Specifically they use *k*-means clustering of model data from the numerical ocean model ECCOv4 (Estimating the Circulation and Climate of the Ocean) to identify dynamical regimes and develop geoscientific utility criteria. In our work, we follow Sonnewald and Lguensat (2021) and use this regime deconstruction framework as the labeled target data that the BNN seeks to predict at each point on the grid. Because the dynamical regimes were found in the model equation space, we have an automatic way to verify the XAI results. Figure 3 shows a global representation of these six dynamical ocean regimes, which we have labeled A, B, C, D, E, and F corresponding to the regimes “NL,” “SO,” “TR,” “N-SV,” “S-SV,” and “MD” in Sonnewald and Lguensat (2021). We have made this label simplification because the aim of this work is to develop a neural network technique to improve the trustworthiness of neural network analyses of ocean model outputs. Thus anything other than a high-level understanding of the physics is beyond the scope of this work and we refer the reader to Sonnewald et al. (2019) and Sonnewald and Lguensat (2021) for a more in-depth discussion.

Table 1
Approximate Importance of Features for Predicting Each Regime According to the Equation Space, Using Analysis From Figure 1 in Sonnewald et al. (2019)

	Features					
	Wind stress curl	Bathymetry	Dynamic sea level	Coriolis	Gradient bathymetry	Gradient dynamic sea level
A	High	High	High	High	High	High
B	High	High	High	High	High	High
C	High	Med	Med	High	Med	Med
D	Low	Low	Low	Med	Low	Low
E	Med	Med	Med	High	Med	Med
F	Med	Med	Med	Med	Med	Med

For our input features, we follow Sonnewald and Lguensat (2021) and use data from the numerical ocean model ECCOv4, but the framework is set up so that it can be readily trained on CMIP6 data in the future (Forget et al., 2015). The following features are then used for prediction: wind stress curl, Coriolis (deflection effect caused by the Earth's rotation), bathymetry (measurement of ocean depth), dynamic sea level, and the latitudinal and longitudinal gradients of the bathymetry and the dynamic sea level. These features are chosen following the dynamical regime decomposition in Sonnewald et al. (2019) and Table 1 shows which features are important for each regime according to the clustering of the equation space based on theoretical intuition. The specific composition of these features into terms in the equation space then manifests as different key ocean circulation patterns. Note that, following standard practice, all input features are normalized before being used in the BNN.

For the training and test data set split, we split by ocean basin. The Atlantic Ocean basin (80°W to 20°E) is the test data set. The rest of the data set is split into a training data set (80%) and a validation data set (20%) with the validation data set being used to compute loss metrics during BNN training thus helping to prevent overfitting (see Section 2.1). In summary, the training data set has approximately 85,000 gridpoints, the validation data set approximately 20,000 gridpoints and the test data set approximately 40,000 gridpoints. During BNN training, we shuffle the data set after each epoch to ensure that the neural network does not memorize the order of the data input, and thus that we are truly using a gridpoint-by-gridpoint approach where the BNN does simply learn spatial correlations. In this way, we also decrease the correlation between the training and validation data sets.

3.1. Remark on *k*-Means Clustering and Data Selection

We conclude this section with a brief remark on the *k*-means clustering approach used to identify the regimes that are the target data for the BNN. The clustering was conducted on equation terms within the barotropic vorticity budget (Sonnewald et al., 2019). However, determining these balances from CMIP6 model data is very difficult, and understanding these difficulties is the subject of ongoing work (see e.g., Waldman & Giordani, 2022). Therefore it is to-date impractical to train our BNN using the closed baratropic vorticity equation terms as input features. Therefore, as discussed, we instead use surface fields and depth. Thus, here, as in Sonnewald and Lguensat (2021), we are conducting sub-surface dynamics inference; a highly underdetermined problem whose non-linearity means that progress with standard approaches such as linearization has been limited. Our work aims not only to show that a BNN can help solve this underdetermined problem, but also to understand why it makes these predictions and how uncertain these predictions are. This will help take the first steps toward understanding the uncertainties and the correlation between the surface input features and the in-depth dynamical ocean regimes. The general framework could also have implications in other fields for the solving of highly underdetermined problems.

We are confident that the *k*-means clustering in Sonnewald et al. (2019) identifies the correct ocean regimes as they have been independently verified in Appendix C in Sonnewald and Lguensat (2021) using unseen data from a different model. They also agree with physical intuition and knowledge of ocean dynamics. However, even if there are biases in the regimes, this will have little to no impact on the classification accuracy of our BNN relative to the given outputs because a neural network simply emulates what it is given. In further work, we could consider using a Gaussian Mixture Model (GMM) (Valletta et al., 2017) instead of *k*-means clustering.

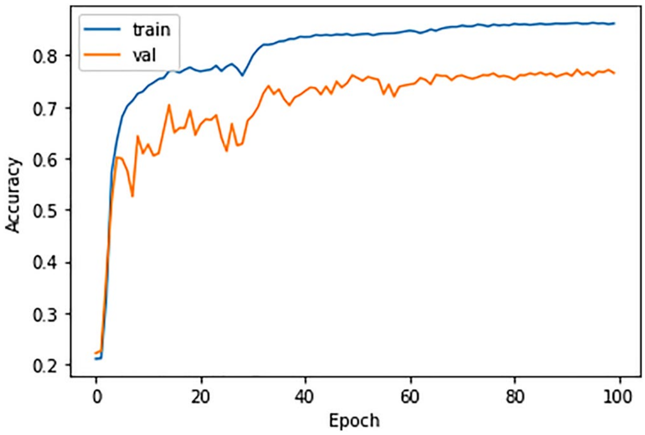


Figure 4. Training accuracy and loss metrics for the BNN showing that the training has converged. See Section 3 for exact descriptions of the training and validation data sets.

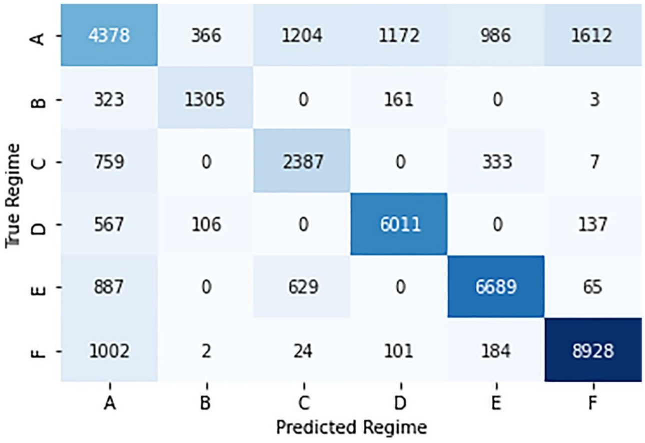
For a given datapoint, GMMs predict the probability distribution across all the identified clusters. Thus their main advantage over *k*-means clustering is that they reveal information on the uncertainty of the clusters themselves. However to the best of our knowledge using a BNN with GMMs would require the construction of new functionality and/or layers in the Tensorflow probability library, because BNNs are designed to be trained on deterministic outputs. Furthermore, given how robust our *k*-means clusters are (see Sonnewald et al., 2019), GMMs will likely give similar results and add potentially needless complexity.

4. Results

In this section, we first use a BNN to make a probabilistic forecast of ocean circulation regimes and show the value added by the uncertainty analysis that can be conducted through using a BNN instead of a deterministic neural network. We then use two modified XAI techniques to explain the skill of this network, comparing the two techniques with each other and with physical understanding.

4.1. Bayesian Neural Networks

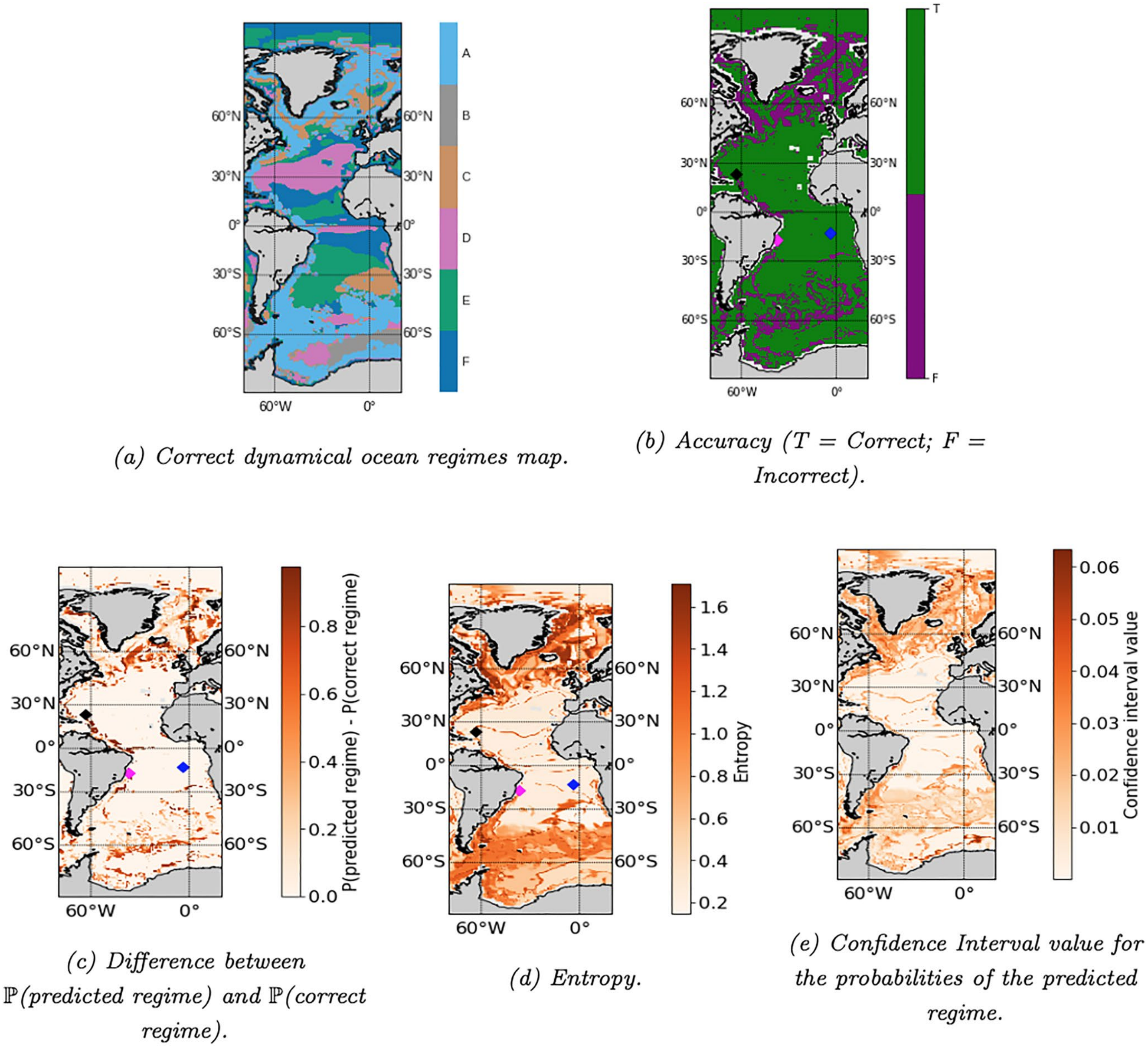
The advantage of BNNs over deterministic neural networks is that they provide an good uncertainty estimate efficiently. However, for BNNs to be of value they must also make accurate predictions. Figure 4 compares the accuracy metrics of the BNN applied to the training data set (the global ocean, excluding the Atlantic Ocean basin) and the validation data set during training. The accuracy metric clearly converges and the level of accuracy is high, indicating that the architecture and learning rates chosen are appropriate for this data set. When the trained BNN is applied to the test data set (the Atlantic Ocean basin), the accuracy is 74%, which is approximately the same as the accuracy achieved by the deterministic neural network in Sonnewald and Lguensat (2021) on the same data. Thus, by using a BNN we have not lost accuracy. For multi-classification tasks, accuracy can be insufficient for fully reflecting the model performance. Therefore in Figure 5 we show the confusion matrix for our BNN. This shows that most incorrect predictions occur for regime A for which errors are not unexpected—it is a composite regime with a less Gaussian structure meaning it is less clearly defined and less easily determined by *k*-means (Sonnewald et al., 2019). The confusion matrix also shows that the BNN sometimes struggles to differentiate between Regimes C and E. This can be seen again in Figure 6b (the spatial distribution of the correct and incorrect regime predictions), which shows inaccuracies around 30°S and 0°W where Regime C transitions to Regime E. These inaccuracies will be discussed later in this section when we analyze the added uncertainty information provided by the BNN.



	A	B	C	D	E	F
A	4378	366	1204	1172	986	1612
B	323	1305	0	161	0	3
C	759	0	2387	0	333	7
D	567	106	0	6011	0	137
E	887	0	629	0	6689	65
F	1002	2	24	101	184	8928

Figure 5. Confusion matrix for Bayesian Neural Network ocean regime predictions.

As we are considering aleatoric uncertainty (uncertainty in the input data), the BNN output is not deterministic but is instead a distribution. Moreover, as we are also considering epistemic uncertainty (uncertainty in the model parameters), the network parameters are distributions, the full output is an ensemble of distributions. In Figure 7, we show both types of uncertainty using a box-and-whisker plot for the predictions for three example datapoints. The narrower the box and whisker, the lower the epistemic uncertainty in the prediction for this regime. For example, in Figure 7a there is almost no width to the box and whisker indicating low epistemic uncertainty, whereas for Figure 7b there are a range of possible probabilities of the most likely regime occurring, indicating epistemic uncertainty. In both Figures 7a and 7b the highest probability is high (almost 1 for Figure 7a and just under 0.8 on average for Figure 7b), which indicates that the aleatoric uncertainty is low. Therefore, practitioners can be confident in the results for both these datapoints, with Figure 7a being a more trustworthy neural network regime prediction than Figure 7b. By contrast, Figure 7c has high levels of epistemic uncertainty and fairly high levels of aleatoric uncertainty meaning that



(a) Correct dynamical ocean regimes map.

(b) Accuracy (T = Correct; F = Incorrect).

(c) Difference between $\mathbb{P}$(predicted regime) and $\mathbb{P}$(correct regime).

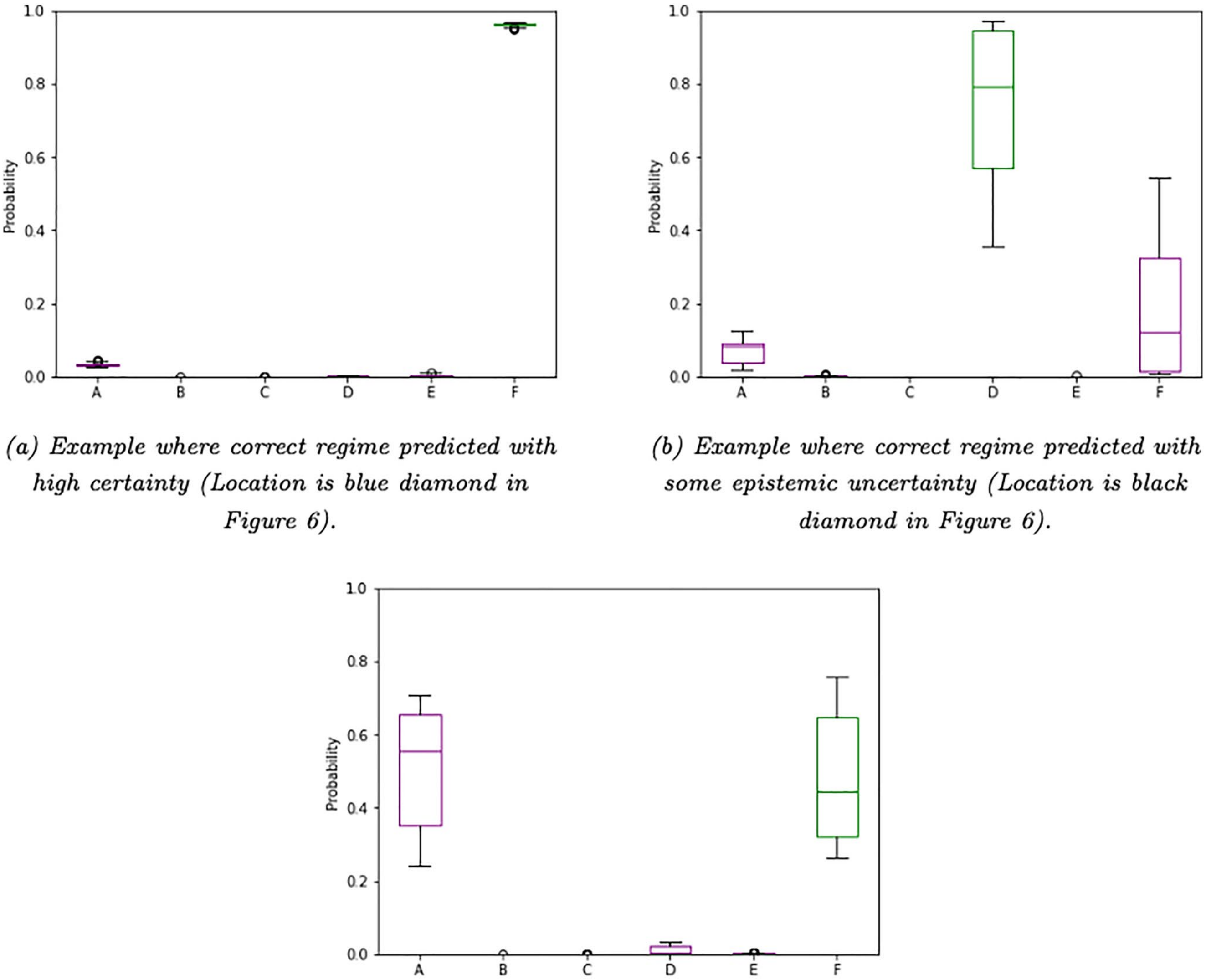
(d) Entropy.

(e) Confidence Interval value for the probabilities of the predicted regime.

Figure 6. Spatial distribution of key metrics calculated from the Bayesian Neural Network predictions for the test data set (Atlantic Ocean basin), as well as the correct regimes for this data set. The diamonds are the three locations of the example datapoints in Figure 7.

although the practitioner can trust that the regime is either A or F, the overall neural network regime prediction for this datapoint is not very trustworthy.

Using these distributions, we can calculate the difference between the probability the BNN assigns to the predicted regime and the probability it assigns to the correct regime. If the BNN has predicted the correct regime then this difference is zero, and, if the BNN is very certain in its prediction of the incorrect regime, the maximum possible probability difference is one. The spatial distribution of this value is shown in Figure 6c and unsurprisingly corresponds closely with the spatial distribution of the correct and incorrect BNN predictions in Figure 6b. The probability difference map adds value compared to the accuracy map because we can see where errors are more substantial. For example, although the BNN appears to perform poorly in the accuracy statistics around Greenland (especially around 50°W and 50°N and 20°W and 70°N), the difference between the probability of the correct regime and the highest probability is low. Therefore the BNN is still assigning a high probability to the correct regime here which is useful for practitioners. In contrast, off the north coast of South America, the



(a) Example where correct regime predicted with high certainty (Location is blue diamond in Figure 6).

(b) Example where correct regime predicted with some epistemic uncertainty (Location is black diamond in Figure 6).

(c) Example where incorrect regime predicted with both epistemic and aleatoric uncertainty (Location is magenta diamond in Figure 6).

Figure 7. Box-and-whisker plot of Bayesian Neural Network predictions of ocean regimes, generated using an ensemble of outputs. Here, as standard, the boxes indicate the interquartile range. The correct regime is colored green and the incorrect regimes are colored purple.

probability difference is almost 1 meaning the BNN is doing a poor job here and should not be used in its current state for predictions here. Comparing Figure 6c with Figure 6a reveals that almost all the high probability differences occur at the boundaries between regime A and other regimes (e.g., in the Southern Ocean at the boundary between regimes B and D with regime A). This reveals that a key weakness of our BNN is predicting whether the input features indicate Regime A or a different regime in borderline cases. Thus by analyzing this probability difference, we have gained valuable information for future predictions and learned that to improve the BNN accuracy, we should provide more training data on the boundaries between regime A and other regimes. This could be achieved by, for example, running more model simulations with different perturbations to the initial conditions.

The distributions outputted by the BNN can also be used to numerically quantify the uncertainty in the network predictions. We can calculate the entropy value using Equation 5, where we recall that the higher the value the more uncertain the result. Figure 6d shows the spatial distribution of this entropy and comparing with Figure 6b

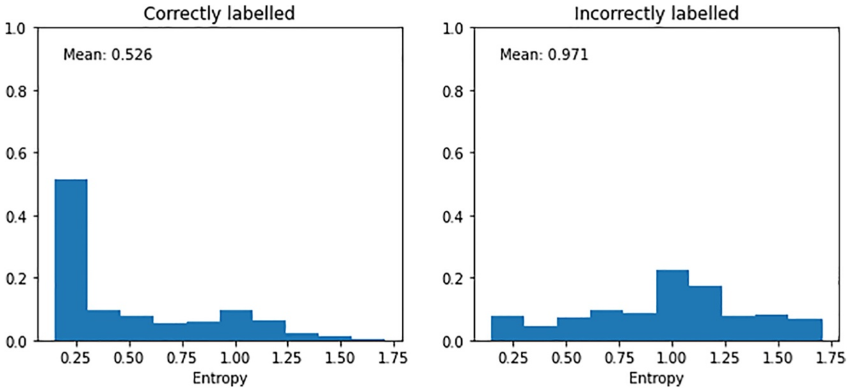


Figure 8. Distribution of entropy values for the correct and incorrect regime predictions. Recall that the lower the entropy, the more certain the result.

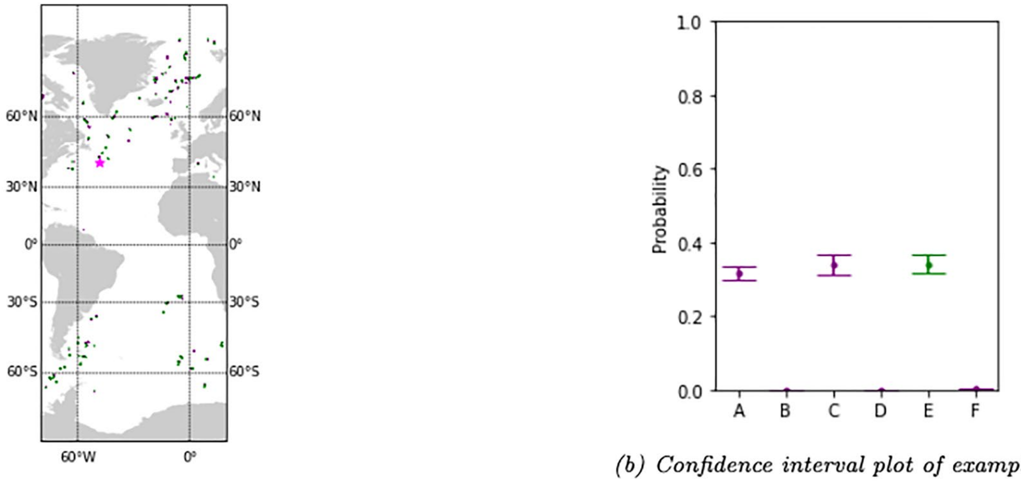
shows that the higher entropy values tend to be where the BNN prediction is incorrect. For example, there is high levels of entropy around 30°S and 0°W where Regime C transitions to Regime E, which, as previously discussed, is an area of high inaccuracy (see Figures 5 and 6b). More precisely, Figure 8 compares the distribution of the entropy when the BNN predictions are correct and when they are incorrect, and clearly shows that the entropy for the correct predictions is skewed toward lower values, whereas the entropy for the incorrect predictions is skewed higher. This is a good result because it means that the predictions are notably more uncertain when they are incorrect than when they are correct, that is, the correct regime classifications are also those that the BNN informs the practitioner are the most trustworthy.

Finally, Figure 7 show that there can be substantial overlap between the box-and-whisker for each regime. However this can be misleading as box-and-whisker plots consider upper and lower quartiles which are not useful for assessing statistical significance. Therefore, we also consider the confidence intervals and in Figure 6e show the spatial distribution of the confidence level size for the probabilities of the predicted regime. Note that, the spatial distribution for the confidence intervals is very similar to that for the entropy. Broadly speaking, the higher the value of the confidence intervals the higher the epistemic uncertainty, and the higher the entropy values the higher the aleatoric uncertainty. Therefore this similarity in spatial distributions suggests that for our framework, aleatoric and epistemic uncertainty are highly linked. Using confidence intervals, we find that for the majority of cases, the probabilities for the most likely regime are statistically significantly different from the probabilities for the other regimes. Figure 9a highlights the datapoints for which this is not the case, and that these datapoints correspond to points for which there is high entropy (see Figure 6d). For the vast majority of the datapoints in Figure 9a, the top two most likely regimes are statistically significantly different from the other regimes and the correct regime is one of the two regimes. Therefore although the neural network is uncertain for these datapoints, it is still predicting a high probability for the correct regime. Finally, there are approximately 20 datapoints where only the top three most likely regimes are significantly different from the others. An example of one such datapoint is shown in Figure 9b, where half the regimes have the same probability. Although this is not ideal, this is an example of where a BNN is better than a deterministic neural network, because it clearly informs the user that it is very uncertain of its prediction and that using this BNN on this datapoint is inappropriate.

Therefore, in this section we have shown that by looking at the probabilities and confidence intervals produced by the BNN, practitioners can make an informed decision as to whether to trust the BNN prediction for the dynamical regime or whether further analysis is required for these datapoints.

4.2. Explainable AI

To explain the BNN's skill, we sapply two common XAI techniques, LRP and SHAP, to an ensemble of deterministic neural networks generated from the BNN. We consider LRP in Section 4.2.1 and SHAP in Section 4.2.2, and then compare results from the two techniques in Section 4.2.3 to test the “disagreement problem” discussed in Section 2.2. If LRP and SHAP largely agree with each other as to which features are relevant in each area (i.e.,



(a) Spatial distribution of points where the most likely regime is not statistically significantly different from other regimes.

(b) Confidence interval plot of example datapoint, where the probabilities for the top three regimes are not statistically significantly different.

Figure 9. Considering whether the differences between the probabilities for each regime are statistically significantly different. The star on (a) is the location of the example datapoint in (b). In both figures, incorrect predictions are colored purple and correct predictions green.

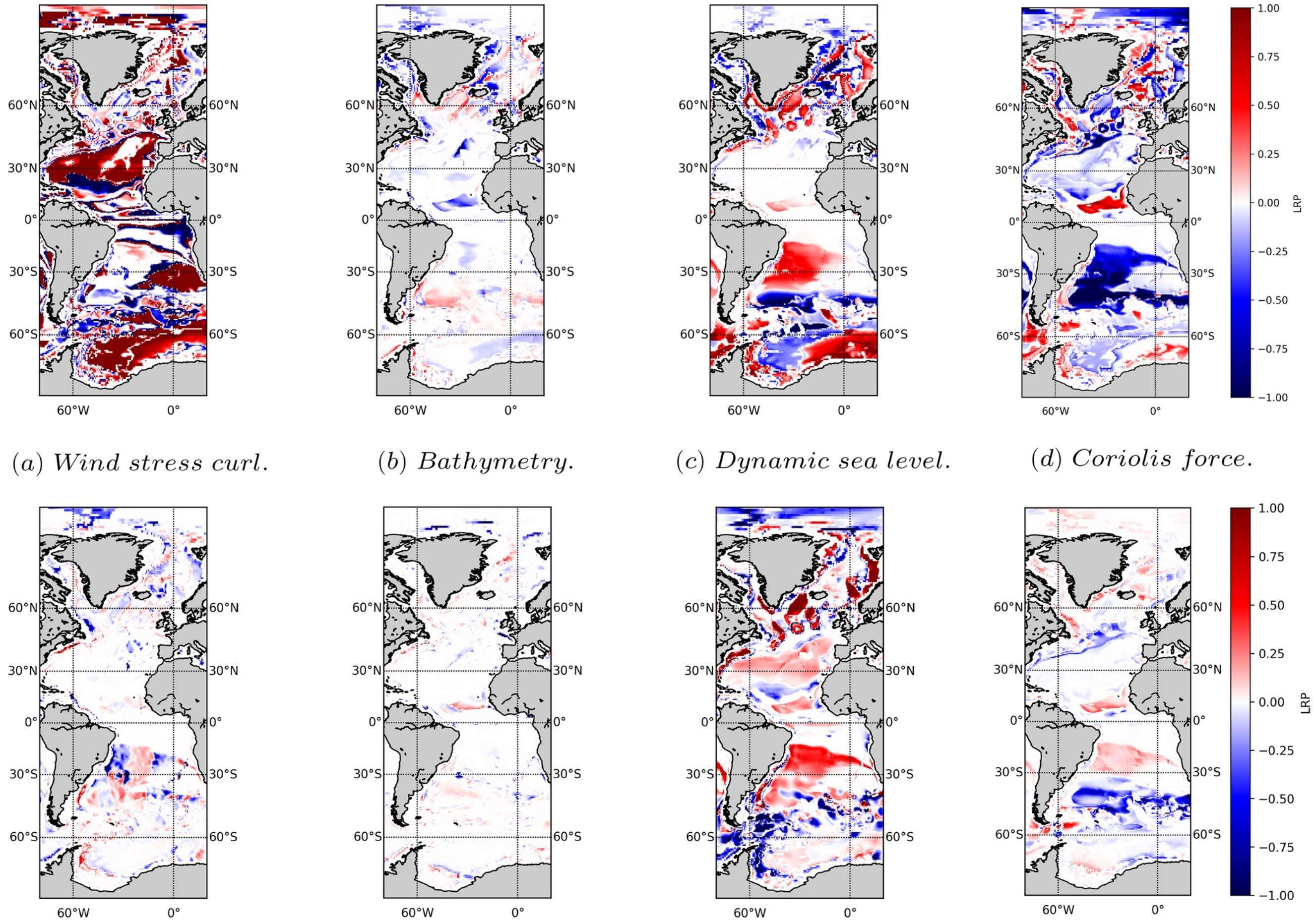
there is no disagreement problem) and also agree with our intuition from physical theory then this increases the trust in our XAI results. This is important to ensure that what the BNN has learned is genuinely rooted in physics. Moreover, the use of a BNN allows us to explore whether disagreement between SHAP and LRP is more likely to occur when predictions have higher entropy (i.e., higher uncertainty).

4.2.1. Layer-Wise Relevance Propagation

Applying LRP using our ensemble approach means that each input variable has its own distribution of LRP values and own level of uncertainty. Figure 10 shows the values for which the sign of the LRP value (i.e., the relevance) remains the same between the 25%–75% quantiles of the ensemble. Note that throughout the LRP values are scaled by the maximum absolute LRP value for any variable across the ensemble. If the LRP value consistently has the same sign across the quantiles, then we can be confident of the effect this feature has on the output; the piece of information of most interest to practitioners in a recent survey in Lakkaraju et al. (2022). We recall from Section 2.2.1 that a positive/negative LRP value means a feature contributes positively/negatively to the output.

In Figure 10, red indicates that the variable in this area is helpful for the BNN in making its predictions, blue that it is unhelpful, and white that it is too uncertain to have consistent relevance. Note that certain areas of white may also be because the variable does not contribute (see Figure A1 in Appendix A which shows the actual LRP values for the 25%, 50%, and 75% quantiles of the ensemble). An important point to note when interpreting these trends is that our network predicts using a gridpoint-by-gridpoint approach and does not see the overall global map, thus making the spatial coherence striking in its consistency. To aid with the interpretation of the LRP values for each regime, we include Figure 11 (which shows the most probable ocean regime predicted by the BNN) to help qualitatively see the trends, and Table 2 which highlights the general trends in the relevance and variance of the LRP values for each regime with respect to each feature. By comparing Table 1 with Table 2, we can compare the general trends of the LRP values with what is expected from the clustering of the equation space. A strong difference is that according to LRP the gradients of the bathymetry are irrelevant to the BNN predictions with high certainty (apart from for key processes discussed in Table 3), whereas the equation space suggests the bathymetry gradients are relevant for some regimes.

Of particular interest when comparing Tables 1 with Table 2 are the differences for Regimes A and B. From the equation space (see Table 1), we would expect all features to be helpful for these regimes. However, in the case of Regime A, the LRP values conclude that both the wind stress curl and the longitudinal gradient of the dynamic



(*a*) *Wind stress curl.*

(*b*) *Bathymetry.*

(*c*) *Dynamic sea level.*

(*d*) *Coriolis force.*

(*e*) *Gradient bathymetry (lon).*

(*f*) *Gradient bathymetry (lat).*

(*g*) *Gradient dynamic sea level (lon).*

(*h*) *Gradient dynamic sea level (lat).*

Figure 10. Layer-wise Relevance Propagation values which are consistent across the whole ensemble. Red indicates that the variable in this area is helpful for regime prediction, blue that it is unhelpful, and white that it is too uncertain to have consistent relevance.

sea level are unhelpful. Figure 6 shows that both the highest areas of inaccuracy and the highest areas of entropy (i.e., uncertainty) in the BNN occur for datapoints which should be Regime A. These LRP values suggest that the reason for the errors and uncertainty in the predictions for these datapoints is that the BNN is incorrectly weighting the wind stress curl and the longitudinal gradient of the dynamic sea level there. By contrast, for Regime B, there are no features which are unhelpful. Instead, there are some features for which the BNN has no relevance (gradients of both the bathymetry and the dynamic sea level). The BNN predictions for Regime B are generally accurate and certain, and therefore this implies that, despite the conclusions from the equation space, the BNN can rely on certain key features it has identified to make accurate certain predictions. There is therefore scope for learning about the physical ocean processes guided by understanding of what the BNN determines as important and unimportant.

For reasons of brevity, we do not detail all the physical interpretations in Figure 10 and Table 2 but instead focus on the key dynamical processes of the North Atlantic Drift, the Gulf Stream leaving the continental shelf, and the North Atlantic wind gyre; and the key physical characteristic of the mid-Atlantic ridge specifically as it crosses the wind gyre (hereafter simply referred to as the mid-Atlantic ridge). The location of these processes is shown in Figure 12 and the variance and relevance of the LRP values for them are summarized in Table 3. The table highlights that for the North Atlantic Drift, there are no features which have strong positive relevance; in fact, the Coriolis force and latitudinal gradient of the sea level have strong negative relevance. Instead, it is regimes other than the regime of the North Atlantic Drift (Regime F), which have highly relevant

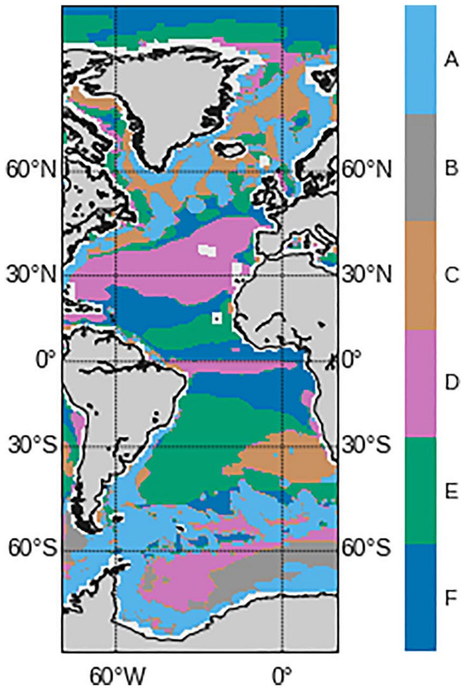


Figure 11. Most probable ocean regime predicted by Bayesian Neural Network.

features in the area; for example, both the dynamic sea level and its longitudinal gradient are strongly positively relevant for Regime A in this area. This is also noted in Sonnewald and Lguensat (2021), who suggest this could be because of multiple inputs contributing medium importance to predictions for Regime F (see Table 1). In contrast, where the Gulf Stream leaves the continental shelf, the Coriolis effect and wind stress curl are both strongly helpful. This conclusion greatly agrees with physical intuition, which states that these features are the key drivers for the Gulf Stream's movement across the North Atlantic (Webb, 2021). Table 3 also shows that the bathymetry gradient is unhelpful for this process. Before leaving the coast, physical intuition suggests that the gradient of the bathymetry is the key driver of the Gulf Stream and this can be seen in the LRP values, (particularly for the latitudinal gradient in Figure A1h). It is therefore likely that the BNN is using the same weightings for the bathymetry gradient as the Gulf Stream leaves the continental shelf, but the key drivers have changed meaning the bathymetry gradient is no longer helpful. Also of interest is the longitudinal gradient of the sea level, which is unhelpful for the North Atlantic Drift, very uncertain for the Gulf Stream leaving the continental shelf (an area which has high entropy in Figure 6d) and then helpful for the wind gyre. This suggests the this feature is acting as an indicator between the three regimes discussed here. For the wind gyre, the wind stress curl is also strongly helpful, which agrees with the intuition from physical theory of gyres, which states that they are largely driven by the wind stress curl (see Munk, 1950). Note however that the theory also indicates that Coriolis should be somewhat helpful but it is unhelpful. This variation may be because the BNN does not seem to be able to accurately weight low values of Coriolis (near the equator). Nevertheless the general agreement with physical intuition for the dynamical processes discussed here highlights our BNN's ability to learn key physical processes.

Unlike the other processes highlighted, the mid-Atlantic ridge is a physical characteristic of the bathymetry that will remain unchanged by a future climate. The ridge is clearly identifiable in the features in Figure 10 and it is therefore interesting to highlight the differences between the relevance of this ridge and the relevance of the other gridpoints in the wind gyre around it. The most noticeable difference is that the ridge adds uncertainty to the BNN predictions—for almost all features, the relevance of the mid-Atlantic ridge is more uncertain than that of the wind gyre. The exception is the bathymetry, which becomes strongly unhelpful with high certainty at the mid-Atlantic ridge. Added to the fact that the bathymetry gradients are also more unhelpful at the ridge than at

Table 2
General Trends in the Variance and Relevance of Layer-Wise Relevance Propagation Values for Each Regime and Each Feature

	Features													
	Wind stress curl		Bathymetry		Dynamic sea level		Coriolis		Gradient bathymetry		Gradient dynamic sea level (lon)		Gradient dynamic sea level (lat)	
	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.
A	Med	Med –	Med	Med +	Med	High +	Med	Med +	Low	Low	Med	High –	Med	Med –
B	Low	High +	Low	Med –	Med	High +	Low	Low	Low	Low	Low	Low	Low	Low
C	Med	High +	Low	Med – (NH) Med + (SH)	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low
D	Med	High +	Med	Med –	Low	Med + (NH) Med – (SH)	Med	Med –	Low	Low	Med	Med +	Low	Low
E	High	High +	Low	Low	Low	High +	Med	High –	Low	Low	Low	High +	Low	Med +
F	Med	Med –	Med	Med –	Low	Med –	High	Med –	Low	Low	High	Med +	High	Med (– >+)

Note. Here + indicates that the feature is helpful for regime prediction and—that it is unhelpful (so high + indicates high positive relevance). (– >+) indicates that between the 25th and 75th quantiles, the variable changes from unhelpful to helpful.

Table 3

Variance and Relevance of Layer-Wise Relevance Propagation Values for the Key Dynamical Processes of the North Atlantic Drift (NAD); the Gulf Stream Leaving the Continental Shelf (GS), the Wind Gyre and the Key Physical Feature of the Mid-Atlantic Ridge as It Crosses the Wind Gyre (MAR) (See Figure 12)

	Features													
	Wind stress curl		Bath.		Dynamic Sea Level		Coriolis		Gradient bath.		Gradient sea level (lon)		Gradient sea level (lat)	
	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.
NAD	Low	Med +	Low	Low	Med	Med +	Low	High –	Low	Low	Med	Med –	Low	High –
GS	Med	High +	Med	Med –	Low	Low	Med	High +	Med	Med –	High	High (– >+)	Med	Med +
Gyre	Low	High +	Med	Med –	Low	Low	Med	High –	Low	Low	Med	Med +	Low	Low
MAR	High	Med (– >+)	Low	High –	Med	Med –	Med	High –	Med	Med –	Med	High +	High	Med +

Note. Here + indicates that the feature is helpful for regime prediction and—that it is unhelpful (so high + indicates high positive relevance). (– >+) indicates that between the 25th and 75th quantiles, the variable changes from unhelpful to helpful.

the surrounding gridpoints, this suggests that the BNN is able to identify the ridge in the bathymetry but unable to weight it correctly, which leads to uncertainty in the relevance of the other features. We observe that, in contrast to bathymetry, both gradients of the dynamic sea level increase in helpfulness at the ridge, in particular the longitudinal gradient. Moreover, Figure 6 shows the BNN predicts the correct regime for the mid-Atlantic ridge with high certainty. Therefore, this suggests that reliable and accurate predictions for regimes at the mid-Atlantic ridge should be based more on the gradient of the dynamic sea level than the bathymetry itself.

To summarize, our discussion of LRP values in this section has highlighted both our BNN's ability to identify known physical characteristics and the potential scope to advance physical theory through analyzing its skill. We have also shown that LRP values are highly uncertain. In Appendix B, we show that this uncertainty in LRP values is also present in prediction of day-ahead 2m temperature biases. This reinforces the point that considering uncertainty greatly improves our ability to correctly interpret LRP values for geoscience problems.

4.2.2. SHapley Additive exPlanation Values

Whereas LRP considers the relevance of a feature for all regimes simultaneously, the SHAP approach sees the problem as binary for each regime: including a feature at a gridpoint either increases the probability of the specific regime being considered there or decreases it. There is therefore a SHAP value for each gridpoint for each regime, meaning we have six times the number of SHAP values as we do LRP. Moreover our ensemble approach means each input variable and regime pairing has its own distribution of SHAP values and own level of uncertainty. Table 4 summarizes the general trends in the SHAP values and in particular highlights that for all regimes and features the variance in the ensemble is low, and most features considered are helpful for predictions. The main exceptions to the latter are the latitudinal gradient of the dynamic sea level and both bathymetry gradients, which are not important for regime predictions (apart from in certain key areas discussed later).

Figure 13 shows the gridpoints for which the sign of the SHAP value remains the same between the 25% and 75% quantiles of the ensemble. Note that even though our BNN uses a gridpoint-by-gridpoint approach, for ease of interpretation, we display the SHAP results using a spatial representation, as if SHAP had been applied to a full image. For simplicity, we focus here on Figure 13a which shows the SHAP values for Regime A, although note that the following statements hold true for the regimes for the other figures too. In Figure 13a, red indicates that the probability of Regime A is increased here by including this feature, blue that the probability is decreased and white mainly that this feature has no effect on the probability of predicting

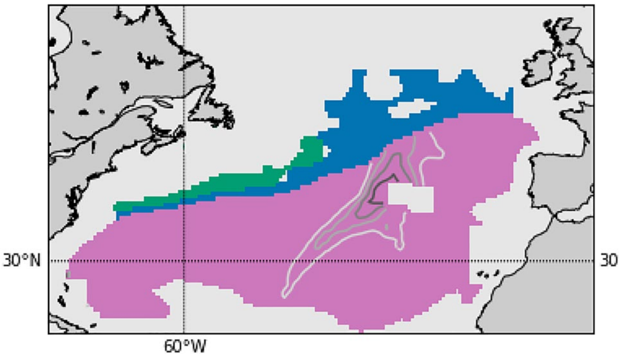
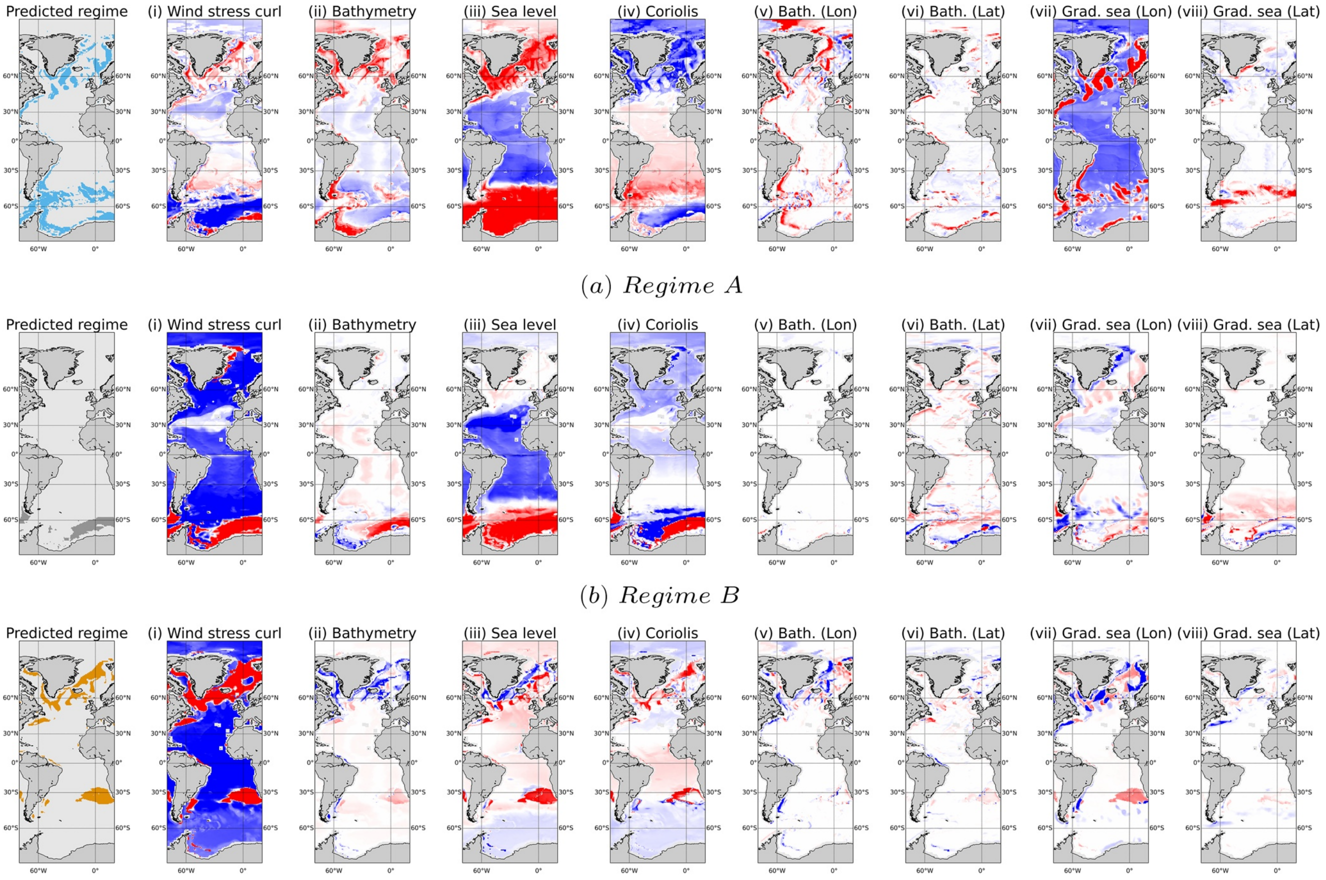


Figure 12. Locations of key dynamical processes and physical features of interest in Table 3: the North Atlantic Drift is the blue area at ~40°N; the Gulf Stream leaving the continental shelf is the green area near coastline at ~70°W and 40°N; the wind gyre is the pink area at ~0° and 30°S; and the part of the Mid-Atlantic Ridge we are focusing on is the gray-scale contours crossing the wind gyre at ~30°W.

Table 4
General Trends in the Variance and Relevance of SHapley Additive exPlanation (SHAP) Values for Each Regime and Each Feature, Where NH Refers to the Values in the Northern Hemisphere and SH to Those in the Southern Hemisphere

	Features													
	Wind stress curl		Bathymetry		Dynamic sea level		Coriolis		Gradient bathymetry		Gradient sea level (lon)		Gradient sea level (lat)	
	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.
A	Low	Med +	Low	High +	Low	High +	Low	Med +	Low	Low	Low	High +	Low	Low
B	Low	High +	Low	Med +	Low	High +	Low	High +	Low	Low	Low	Low	Low	Low
C	Low	High +	Low	Med − (NH) Med + (SH)	Low	High +	Low	Med +	Low	Low	Low	High +	Low	Low
D	Low	High +	Low	Low	Low	Med + (NH) Med − (SH)	Low	Med +	Low	Low	Low	Med +	Low	Low
E	Low	High +	Low	Low	Low	High +	Low	Med −	Low	Low	Low	High +	Low	Low
F	Low	High +	Low	Low	Low	Med −	Low	Med +	Low	Low	Low	Med +	Low	Low

Note. To allow direct comparison with Layer-wise Relevance Propagation, for each regime, we only consider the SHAP values in the regime location rather than in the whole domain. Therefore + means the feature is helpful for the prediction here and—that it is unhelpful.

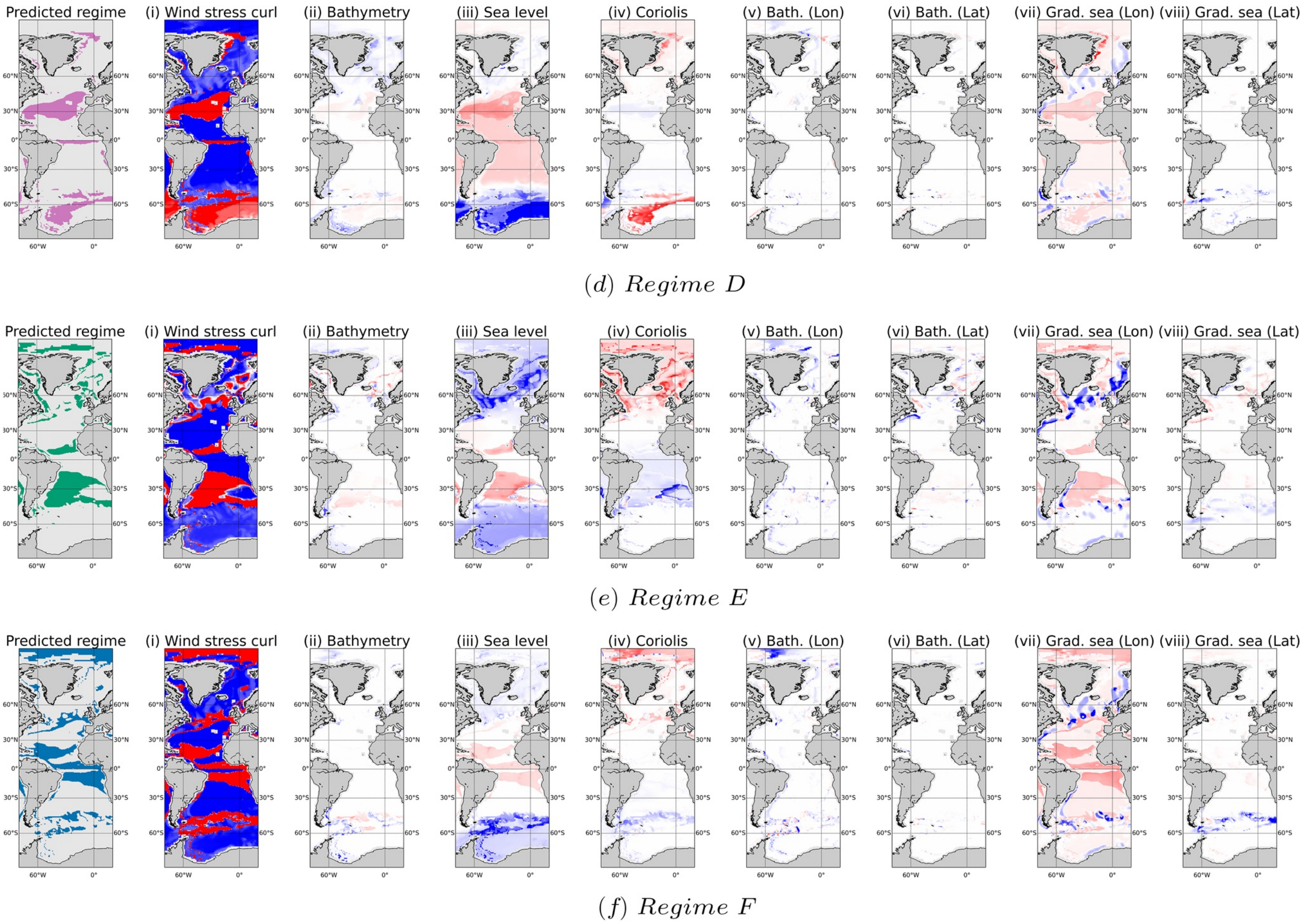


(*a*) *Regime A*

(*b*) *Regime B*

(*c*) *Regime C*

Figure 13. SHapley Additive exPlanation (SHAP) values which are consistent across the whole ensemble for Regimes A (a), B (b), C (c), D (d), E (e) and F (f). Red indicates that the probability of the Regime here is increased by including this feature and blue that the probability is decreased. White means that the SHAP value is either too uncertain or that the variable has no effect.



(d) Regime D

(e) Regime E

(f) Regime F

Figure 13. (Continued)

Regime A here (although it can also mean there is uncertainty in the SHAP value). If the red matches with the area where the BNN predicts Regime A or the blue matches with the area where the BNN does not predict Regime A, this means that including this feature is helpful for predicting this regime in this location. An example of this in Figure 13a is the SHAP values for the longitudinal gradient of the sea level. If, however, the red matches with a area where the BNN does not predict Regime A or the blue matches with the area where the BNN does predict Regime A, then including this feature is unhelpful for predicting this regime. An example of this in Figure 13a is the dynamic sea level where including it increases the probability of Regime A everywhere below 40°S and above the North Atlantic Drift, but Regime A is only predicted in certain parts of this area. Notably, Figure 6d shows that at the latitudes where the dynamic sea level is unhelpful, the BNN predictions have high entropy (i.e., high uncertainty) suggesting that the dynamic sea level may be a key contributing factor to the uncertainty here.

As in the LRP section, we also consider the key dynamical processes of the North Atlantic Drift, the Gulf Stream leaving the continental shelf and the North Atlantic wind gyre, as well as the physical characteristic of the mid-Atlantic ridge where it crosses the wind gyre (see Figure 12). The variance and relevance of the SHAP values for these processes are summarized in Table 5. For the North Atlantic Drift, the SHAP values show that the wind stress curl is strongly helpful, and that the Coriolis, dynamic sea level and the longitudinal gradient of the sea level are also helpful. The North Atlantic Drift is a geostrophic current and therefore this feature relevance agrees strongly with the physical theory which governs these types of currents (Webb, 2021). It is also in contrast to the conclusions from the LRP values where no feature is strongly helpful, only the dynamic sea level and the wind stress are at all helpful and the Coriolis is strongly unhelpful. This difference in the relevance of the Coriolis is also seen for the gyre, which SHAP values say is irrelevant and the LRP values say

Table 5
Variance and Relevance of SHapley Additive exPlanation Values for the Key Dynamical Processes of the North Atlantic Drift (NAD); the Gulf Stream Leaving the Continental Shelf (GS), the Wind Gyre and the Key Physical Feature of the Mid-Atlantic Ridge as It Crosses the Wind Gyre (MAR) (See Figure 12)

	Features													
	Wind stress curl		Bathymetry		Dynamic sea level		Coriolis		Gradient bathymetry		Gradient sea level (lon)		Gradient sea level (lat)	
	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.	Var	Rel.
NAD	Low	High +	Low	Low	Low	Med +	Low	Med +	Low	Low	Low	Med +	Low	Low
GS	Low	High +	Low	Med –	Low	Med –	Low	Med +	Low	Low	Low	High –	Low	Med +
Gyre	Low	High +	Low	Low	Low	Low	Low	Low	Low	Low	Low	Med +	Low	Low
MAR	Low	High +	Med	Med–	Low	Low	Low	Low	Med	Med –	Low	Med +	Med	Med +

is strongly unhelpful. Neither agree with intuition from physical theory, which suggests that Coriolis should have some relevance for the gyre. The SHAP values and LRP values do however both identify that for the gyre, the wind stress curl is strongly helpful and the longitudinal gradient of the sea level is helpful, which we recall from Section 4.2.1 agrees with physical intuition. The SHAP and LRP relevance patterns for where the Gulf Stream leaves the continental shelf are also similar to each other. Furthermore, the increased certainty in the SHAP values makes it clear that the longitudinal gradient of the sea level is strongly unhelpful for predictions of this process, whereas for LRP the relevance is very uncertain. Like with LRP, there is also a clear distinction in the SHAP values between the North Atlantic Drift, the Gulf Stream leaving the continental shelf and the wind gyre, strengthening the hypothesis that this feature is an indicator between the three regimes. Finally, the mid-Atlantic ridge is not as prominent in the SHAP values as it is in the LRP values, but the SHAP values still have increased uncertainty there, which is particular significant when the general uncertainty in the ensemble of SHAP values is so low. Furthermore, like the LRP values, the SHAP values also show that both bathymetry and its gradients are more unhelpful at the mid-Atlantic ridge than for the surrounding gridpoints. This supports the conclusions made in Section 4.2.1 that the BNN is able to identify the ridge but not weight it properly.

To summarize, we have shown that SHAP values provide further evidence of the BNN's ability to identify known physical processes. We have also begun to demonstrate the benefit of using two different XAI techniques, and in the next section compare the findings from the two different techniques more systematically.

4.2.3. LRP Versus SHAP

As discussed in Section 2.2.2, LRP and SHAP use two very different approaches to explain skill and hence different types of uncertainty are reflected in their values: LRP considers the neural network parameters and therefore captures the model uncertainty, whereas SHAP captures the sensitivities of the outputs as a result of the uncertainties. Comparing Tables 2 and 4 clearly shows that this different approach results in SHAP values being more certain in their assessment of feature relevance than LRP values. This difference suggest that our BNN is fairly robust because the uncertainty in the network is greater than the uncertainty in the predictions. This is equivalent to the findings in Section 2.1 where our BNN predictions have low entropy (i.e., low uncertainty) despite the weights in the BNN being distributions (see Figure 6d).

Table 6 directly compares the trends in the relevances of LRP and SHAP. Some differences between SHAP and LRP are due to the fact that SHAP values separate out the relevance of each feature for each regime, whereas LRP values consider the relevance of a feature for all regimes simultaneously. For example, in the upper part of the Atlantic (~60°N), the SHAP values for Regime A (Figure 13a) show that the wind stress curl is helpful for predicting that regime. However, the SHAP values for regimes C and E (Figures 13c and 13e respectively) show that the wind stress curl also increases the probability of regimes C and E at that location. Therefore when the SHAP values for all regimes are considered, the wind stress curl may actually be more unhelpful than helpful, agreeing with LRP.

Table 6
Comparing the General Trend in the Relevances of Layer-Wise Relevance Propagation > SHapley Additive exPlanation.

	Features						
	Wind stress curl	Bathymetry	Dynamic sea level	Coriolis	Gradient bathymetry	Gradient sea level (lon)	Gradient sea level (lat)
A	Med − >Med +	Med + >High +	=	=	=	High − >High +	Med − >Low
B	=	Med − >Med +	=	Low >High +	=	=	=
C	=	=	Low >High +	Low >Med +	=	Low >Med +	=
D	=	Med − >Low	=	Med − >Med +	=	=	=
E	=	=	=	High − >Med −	=	=	Med + >Low
F	Med − >High +	Med − >Low	=	Med − >Med +	=	=	Med >Low

Note. If the relevance changes sign, the change is colored red.

As in Sections 4.2.1 and 4.2.2, for brevity we do not discuss all differences between SHAP and LRP. Instead, we summarize the key comparisons for each regime in the following list:

Regime A

- Wind stress curl is helpful in SHAP but unhelpful in LRP (see discussion in text previously).
- The locations where the dynamic sea level has strong relevance in the LRP values coincides directly with the areas where regime A is predicted. The dynamic sea level is also helpful in SHAP, but SHAP shows that this feature also increases the probability of Regime A in areas where Regime A is not predicted. Note that the latter are areas of high entropy (see Figure 6d).
- The longitudinal gradient of the dynamic sea level is strongly unhelpful in LRP and strongly helpful in SHAP. Again the areas where SHAP and LRP differ correspond to areas of high entropy in the BNN predictions.

Regime B

- Wind stress curl is strongly helpful in both LRP and SHAP, but along the east coast of Greenland, in the SHAP values, the wind stress curl increases the probability of regime B, but the BNN does not predict this regime nor would regime B be accurate there. This area has high entropy and in the LRP values the relevance of the wind stress curl switches here from unhelpful in the 25th quantile to helpful in the 75th quantile. This suggests that the BNN has high uncertainty in the relevance of this input feature here.
- In the SHAP values, the bathymetry is helpful but in LRP it is unhelpful. This is despite the fact that areas where this regime is predicted by the BNN, generally have low entropy
- Coriolis is strongly helpful in SHAP (as would be expected from physical intuition) but has low relevance in the LRP values, apart from around the tip of South America where it is strongly helpful.

Regime C

- In regime C, particularly in the southern hemisphere, most features have no relevance in the LRP values but a medium or high relevance in the SHAP values. In particular, the dynamic sea level and its longitudinal gradient have no relevance with high certainty in the LRP values but strong positive relevance with high certainty in the SHAP values. Note that entropy is low for this regime, particularly in the southern hemisphere
- Wind stress curl is strongly helpful in both LRP and SHAP. This likely explains the irrelevance in other features in the LRP values: LRP values consider the weightings in the BNN, and the wind stress curl has such a strong weighting that all other features are comparatively close to zero. In contrast, SHAP values consider the sensitivity of the output to other features, which does change

Regime D

- In both SHAP and LRP, the dynamic sea level is helpful in the northern hemisphere but unhelpful in the southern hemisphere.
- Coriolis is strongly helpful at high latitudes in the SHAP values and irrelevant at mid-latitudes. In contrast, Coriolis is unhelpful in the LRP values especially at the mid-latitudes. This variation suggests the BNN does

not accurately weight low values of Coriolis (near the equator), resulting in unhelpful LRP values. Nearer the poles, the weighting improves enough for SHAP to become helpful but not enough for LRP to become helpful.
- The wind stress curl is strongly helpful in both the SHAP and LRP values but the SHAP values for wind stress curl do not have increased uncertainty at the mid-atlantic ridge. This reflects the general trend of greater certainty in SHAP values than LRP values.

Regime E

- Wind stress curl is strongly helpful for SHAP and LRP, but the LRP values in the southern hemisphere have high variance especially around 35°S where the BNN entropy is highest.
- Coriolis is strongly unhelpful in LRP especially at mid-latitudes but only slightly unhelpful in SHAP (see discussion for Regime D).
- The latitudinal gradient of the dynamic sea level is irrelevant in the SHAP values but has relevance in the LRP values. There is however a split in the LRP relevance at 35°S—above this latitude the relevance is positive and below the relevance is negative. This split corresponds with an increase in entropy, where entropy is higher below this latitude.

Regime F

- Wind stress curl is strongly helpful in SHAP but unhelpful in LRP. We would expect wind stress curl to be helpful from Table 1 so this is an example where SHAP agrees more closely with physical intuition than LRP.
- Bathymetry is unhelpful for this regime in LRP but in SHAP only has relevance at the coastlines.
- Coriolis is unhelpful in LRP at mid-latitudes but has no relevance in SHAP except at high latitudes (see discussion for Regime D).
- The latitudinal gradient of the dynamic sea level is very uncertain in LRP changing from unhelpful to helpful, despite the fact that the entropy is low for predictions of this regime. This gradient has no relevance according to SHAP, and thus the mean of the SHAP and LRP values agree for this feature. This reflects the general trend of greater certainty in SHAP values than LRP values.

In general, SHAP and LRP agree on how to explain the skill of the BNN, thus meaning that in our work we do not have a "disagreement problem." There are however some small differences, which can either be explained by the different ways in which these two techniques interpret skill or by the fact that they occur where there is high entropy in the BNN predictions reflecting the BNN's uncertainty in feature relevance. We have thus demonstrated that both techniques are helpful for understanding the BNN's interpretations of physical processes. Moreover, where the two techniques agree with each other and in particular also agree with physical intuition, this greatly improves the trustworthiness of the feature relevance explanations in the BNN and where the techniques differ between themselves and/or with physical intuition there is scope for further analysis and learning of both BNN and physical ocean processes.

5. Discussion and Conclusion

In this work, we have successfully applied a BNN and two different XAI techniques to explore the trustworthiness of neural network analyses of ocean dynamical regime classifications. We have shown that using a BNN rather than a classical deterministic neural network adds considerable value to predictions, by making uncertainty analysis possible and allowing practitioners to make informed decisions as to whether to trust a prediction or conduct further investigation. Furthermore, our analysis of the entropy (i.e., uncertainty) of the BNN predictions shows the promising result that the predictions are notably more certain when they are correct than when they are incorrect.

Through our novel applications of the XAI techniques, LRP and SHAP, we have also shown that it is possible to explain the skill of a BNN, conduct uncertainty analysis of explainability values, and hence use XAI techniques to understand the extent to which the BNN is fit for purpose, where we here demonstrate this using comparison with theory. Our spatial representation of both the SHAP and LRP values means that the relevance of specific important dynamical processes such as the North Atlantic Drift can be identified, thereby improving the interpretability and hence trustworthiness of the neural network predictions. This comparison with physical theory is important to ensure that what the BNN has learned is genuinely rooted in physical theory and in turn the skill of our BNN for sub-surface inference shows there is fundamental insight from surface variables to the in-depth ocean. The latter is a very hard theoretical problem because it is highly underdetermined but our work shows that BNNs can make progress toward solving it. Specifically our work also takes the first steps toward understanding the uncertainties and the correlation between the input features that lead to this skill in sub-surface inference. Moreover, the spatial coherency of both the uncertainty and XAI assessments suggest that our framework could be leveraged to identify potential new physical hypotheses in areas of interest, guided by the BNN's ability to highlight hitherto unrecognized correlations in the input space. However, we stress that these correlations do not necessarily imply causation (Samek et al., 2021). Therefore for deployment of developed neural network applications for high-stakes decision making within geoscience, these correlations should only be used to postulate new hypotheses, which must then be explored using a well-conducted study driven by physical theory.

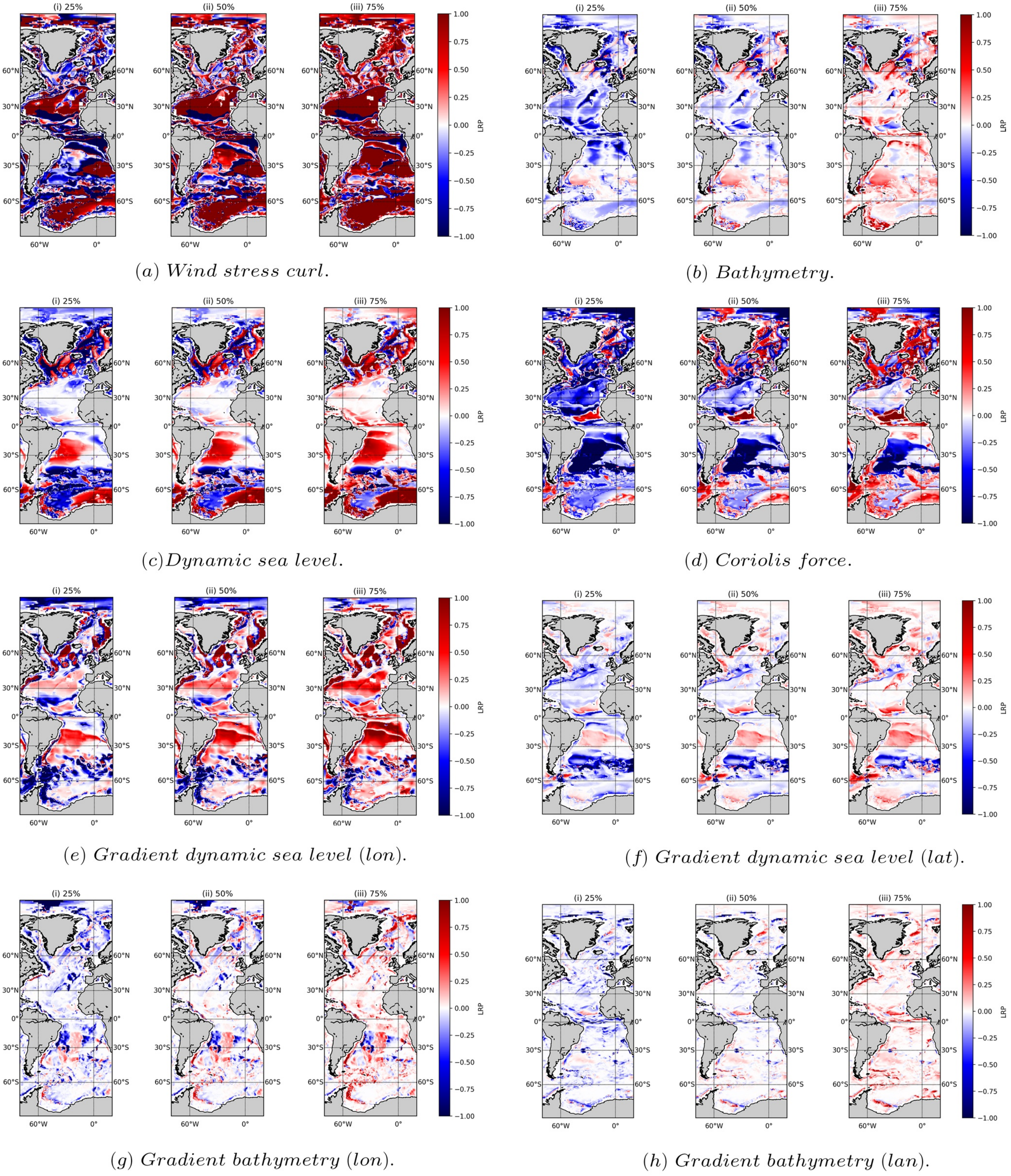
Our comparison of LRP and SHAP values has shown that in general they agree with each other as to which features are relevant in each area of the domain, building trust in the BNN predictions and their explanations. This is particularly striking given that SHAP is model-agnostic and does not consider any internal architecture of the network, exploring only how sensitive the predictions are to the removal of input features, whereas LRP uses a model-intrinsic approach based on the internal architecture of the network. These two different XAI techniques do result in different levels of uncertainty in the feature relevances because LRP better captures the neural network model uncertainty and SHAP better captures BNN prediction sensitivity. Any disagreements in feature relevance also tend to occur due to these different approaches and/or in areas of high entropy. Knowledge of these disagreements is useful to practitioners as it highlights areas where the explanation of the BNN's skill is less trustworthy and may require further analysis. Furthermore the use of an ocean dynamical framework allows the accuracy of the XAI results in this work to be verified with physical intuition. It also enables a better understanding of how SHAP and LRP explain skill which is beneficial to the machine learning community. Where there are differences between the XAI techniques and physical intuition, this provides another potential opportunity to learn more about physical theory, although with the same caveats discussed above.

We hypothesize that the good agreement with physical intuition demonstrated in this work is in part due to the overall normally distributed covariance structure of the problem, which is helpful for the *k*-means clustering and thus directly beneficial for the BNN training (Sonnewald et al., 2019). The methodology outlined in this work has many potential applications in geoscience and beyond, for more complex and nonlinear covariance structures. We note that whilst *k*-means clustering is an inherently linear algorithm, kernel tricks can be used to solve non-linear problems (Tzortzis & Likas, 2009). Besides classification problems, where the re-application of our methodology is straightforward, a promising research avenue is the use of XAI, augmented with uncertainty quantification, for regression problems. An example of high interest to the climate modeling community is subgrid scale parametrization efforts for numerical models. So far, subgrid scale parametrizations based on neural networks have limited generalization capacities, especially in areas of the numerical model space that they are not explicitly trained on (Bolton & Zanna, 2019). A regression based XAI framework could thus accelerate the development of such techniques, because the reasons why the networks fail to generalize might be better understood for both specific local scale features such as where the Gulf Stream leaves the continental shelf and larger scale processes. In further work, we will benefit from the ongoing recent research developments in XAI for regression, for example, in Letzgus et al. (2021), and aim to apply our methodology to this more challenging problem.

Finally, we recommend that for trustworthy explainability results for more complex covariance structures, a BNN should be used along with one model-intrinsic XAI technique, like LRP and one model-agnostic XAI technique like SHAP, so as to consider both neural network model properties and output sensitivity. For an accurate and robust network, we would expect the similarities between the two XAI techniques to dominate and the differences to highlight areas that require further analysis, thus being of valuable use to practitioners and might hint at new scientific hypotheses.

Appendix A: LRP Figures

Figure 10 in Section 4.2.1 reveals the Layer-wise Relevance Propagation (LRP) values which have a consistent sign across the 25%, 50%, and 75% quantiles. However, there is also considerable variability across the ensemble of LRP values and thus to give a better idea of this uncertainty, we also include Figure A1 which shows the 25%, 50%, and 75% quantiles of the LRP ensemble. Using this figure, we see, for example, that for many areas, the bathymetry gradients go from being strongly unhelpful at the 25% quantile to strongly helpful at the 75% quantile,



(a) *Wind stress curl.*

(b) *Bathymetry.*

(c)*Dynamic sea level.*

(d) *Coriolis force.*

(e) *Gradient dynamic sea level (lon).*

(f) *Gradient dynamic sea level (lat).*

(g) *Gradient bathymetry (lon).*

(h) *Gradient bathymetry (lan).*

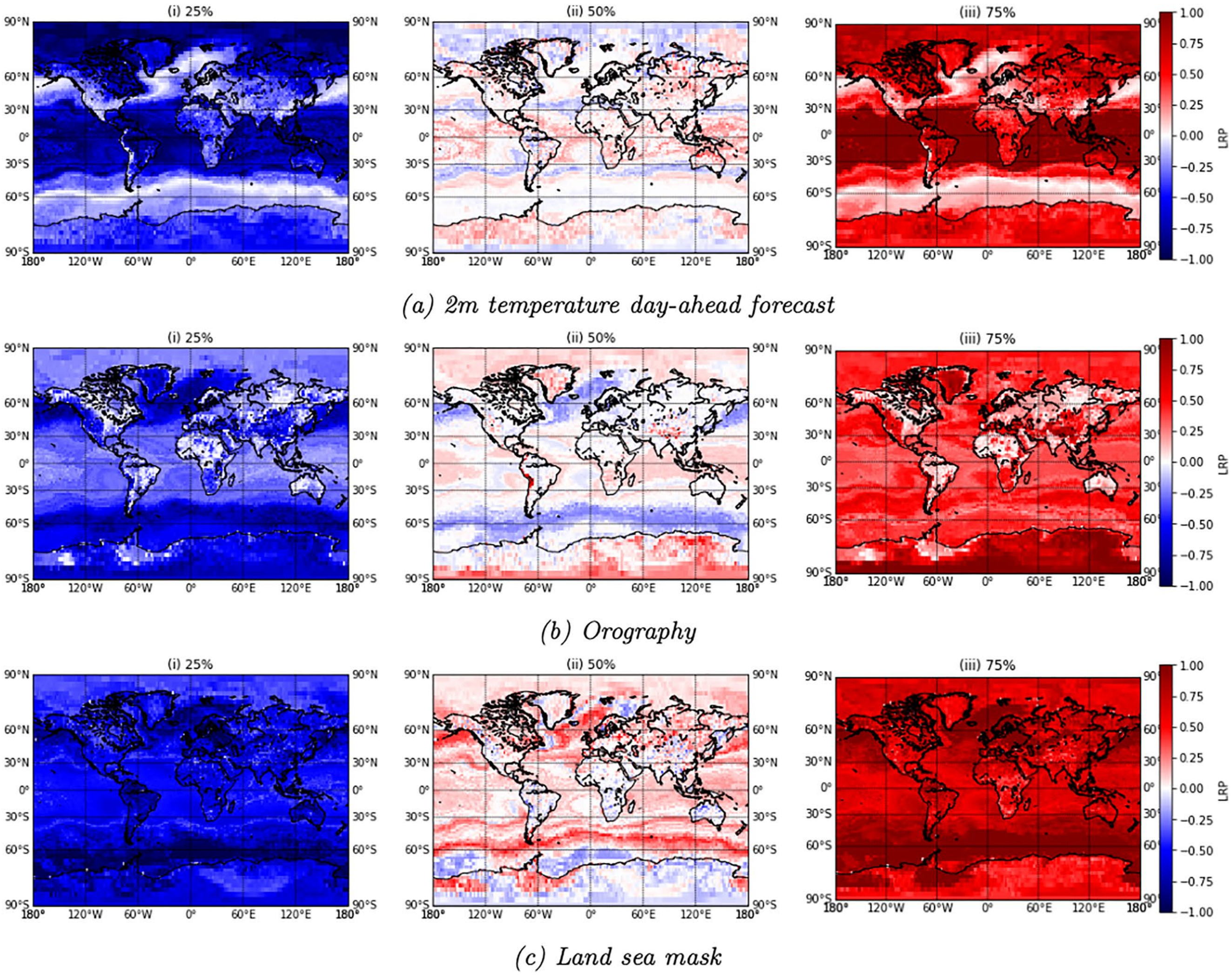
Figure A1. Layer-wise Relevance Propagation (LRP) values at the 25th, 50th (median), and 75th quantile of the ordered ensemble. Note that the ensemble is ordered separately for each feature, so as to more clearly show the uncertainty and range of the LRP values for each individual feature.

showing a high degree of uncertainty. The figure also illustrates better the areas which are irrelevant to Bayesian Neural Network (BNN) predictions (i.e., where the LRP value is zero).

Appendix B: Applying Bayesian Neural Network-explainable AI Methodology to a 2 m Temperature Day-Ahead Biases

For completeness, we also apply our Bayesian Neural Network-explainable AI (BNN-XAI) methodology to the problem of predicting 2 m temperature day-ahead bias. This allows us to check whether the uncertainty shown in the LRP values in Figure 10 is as a result of the specific problem considered in the main body of this work, or if it is present in other geoscience problems.

For our data set, we consider the 2 m temperature day-ahead bias between ECMWF's IFS day-ahead operational forecast and operational analysis. Bias-corrections are part of the standard statistical post-processing that is conducted to optimize the predictions of numerical weather prediction models. There is ongoing research into using statistical and machine learning methods for post-processing, summarized in Vannitsem et al. (2020). Moreover, neural networks have been successfully used to predict 2 m temperature biases in Ben Bouallègue et al. (2022), for example, To the best of our knowledge no thorough explainability analysis of these neural



(a) 2m temperature day-ahead forecast

(b) Orography

(c) Land sea mask

Figure B1. Layer-wise Relevance Propagation (LRP) values at the 25th, 50th (median), and 75th quantile of the ordered ensemble, for the prediction of day-ahead 2 m temperature bias. Note that the ensemble is ordered separately for each feature, so as to more clearly show the uncertainty and range of the LRP values for each individual feature.

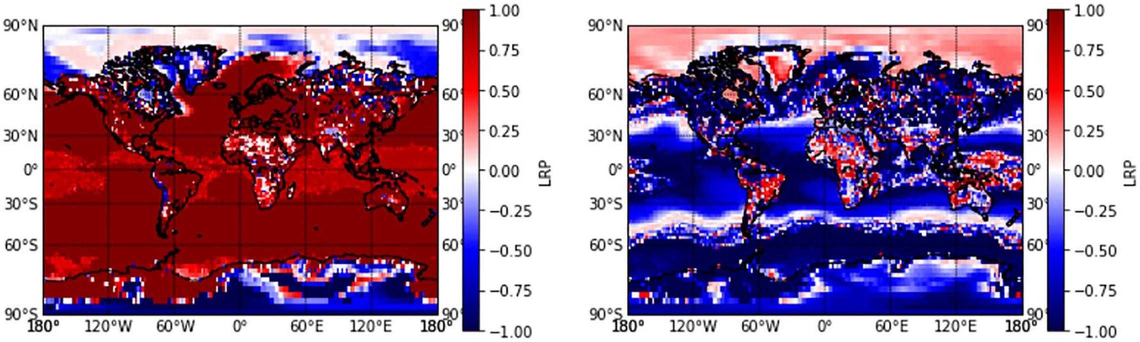


Figure B2. Two examples of Layer-wise Relevance Propagation (LRP) values for the 2 m temperature day-ahead forecast input feature from the Bayesian Neural Network ensemble. Note that in the example on the left, the day-ahead forecast is mostly helpful for predicting the day-ahead 2 m temperature bias and in the one on the right it is mostly unhelpful, highlighting the importance of considering uncertainty when interpreting LRP values.

network predictions has been conducted. Using LRP with regression problems is an under-researched area and thus we change the problem into a classification problem by binning the output data as in Clare et al. (2021). The input features for the BNN are then the 2 m temperature day-ahead forecast, orography and the land-sea mask. The training data set is 2011–2015, the validation data set is 2016 and the test data set is 2017. Figure B1 shows the LRP values at the 25th, 50th (median), and 75th quantile of the ordered ensemble for all three features. They show that the ensemble of LRP values for this data set are even more uncertain than for the ocean data set used in the main body of this work. To highlight this uncertainty, in Figure B2, we show two examples of LRP values from the BNN ensemble, one where the 2m temperature day-ahead forecast is mostly helpful and the other where it is mostly unhelpful. Thus, without knowledge of the uncertainty in the LRP values, we could easily infer the wrong conclusions from the values. Hence we have once again shown that considering uncertainty greatly improves our ability to correctly interpret LRP values.

Data Availability Statement

The relevant code for the explainable Bayesian THOR framework presented in this work is preserved at Clare et al. (2022), available via CC-BY license. The ECCOv4r3 data is available to download at NASA (2022).

Acknowledgments

MCAC acknowledges funding from the Higher Education, Research and Innovation Department at the French Embassy in the United Kingdom. MS acknowledges funding from the 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 acknowledge the Make Our Planet Great Again (MOPGA) fund from the Agence Nationale de Recherche under the "Investissements d'avenir" programme with reference ANR-17-MPGA-0010. Additionally VB is supported by the Cooperative Institute for Modeling the Earth System, Princeton University (Award NA18OAR4320123 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce).

References

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, *298*, 103502. https://doi.org/10.1016/j.artint.2021.103502

Arras, L., Horn, F., Montavon, G., Müller, K.-R., & Samek, W. (2017). "What is relevant in a text document?": An interpretable machine learning approach. *PLoS One*, *12*(8), e0181142. https://doi.org/10.1371/journal.pone.0181142

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

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

Beluch, W. H., Genewein, T., Nürnberger, A., & Köhler, J. M. (2018). The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9368–9377).

Ben Bouallègue, Z., Cooper, F., Chantry, M., Düben, P., Bechtold, P., & Sandu, I. (2022). *Statistical modelling of 2 m temperature and 10 m wind speed forecast errors (Technical Report)*. ECMWF.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., & Samek, W. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. In *International conference on artificial neural networks* (pp. 63–71).

Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in Aging Neuroscience*, *11*, 194. https://doi.org/10.3389/fnagi.2019.00194

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376–399. https://doi.org/10.1029/2018ms001472

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

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': The US, EU, and UK approach. *Science and Engineering Ethics*, *24*(2), 505–528. https://doi.org/10.1007/s11948-017-9901-7

Clare, M., Lguensat, R., & Sonnewald, M. (2022). THOR Bayesian approach. Retrieved from https://github.com/maikejulie/DNN4Cli/tree/main/THOR/bayesianapproach%20%2010.5281/zenodo.6479249

Clare, M. C., Jamil, O., & Morcrette, C. J. (2021). Combining distribution-based neural networks to predict weather forecast probabilities. *Quarterly Journal of the Royal Meteorological Society*, *147*(741), 4337–4357. https://doi.org/10.1002/qj.4180

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). *The AI gambit—Leveraging artificial intelligence to combat climate change: Opportunities, challenges, and recommendations*. AI & Society.

Cui, T., Marttinen, P., & Kaski, S. (2019). Learning global pairwise interactions with Bayesian neural networks. *arXiv preprint*. arXiv:1901.08361.

Dikshit, A., & Pradhan, B. (2021). Interpretable and explainable AI (XAI) model for spatial drought prediction. *Science of the Total Environment*, *801*, 149797. https://doi.org/10.1016/j.scitotenv.2021.149797

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., et al. (2017). Tensorflow distributions. *arXiv preprint* arXiv:1711.10604.

European Commission. (2021). Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Retrieved from https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206

Eyring, V., Bony, S., Meehl, G., Senior, C., Stevens, B., Stouffer, R., & Taylor, K. (2015). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organisation. *Geoscientific Model Development Discussions*, *8*(12), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., et al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110. https://doi.org/10.1038/s41558-018-0355-y

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

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Retrieved from http://www.deeplearningbook.org

Gordon, E. M., & Barnes, E. A. (2022). Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321–1330).

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568–572. https://doi.org/10.1038/s41586-019-1559-7

Hoegh-Guldberg, O., Jacob, D., Bindi, M., Brown, S., Camilloni, I., Diedhiou, A., et al. (2018). Impacts of 1.5 C global warming on natural and human systems. Global warming of 1.5 C. An IPCC Special Report.

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, *14*(12), 124007. https://doi.org/10.1088/1748-9326/ab4e55

Joo, T., Chung, U., & Seo, M.-G. (2020). Being Bayesian about categorical probability. In *International conference on machine learning* (pp. 4950–4961).

Jospin, L. V., Buntine, W., Boussaid, F., Laga, H., & Bennamoun, M. (2020). Hands-on Bayesian neural networks—A tutorial for deep learning users. *arXiv preprint*. arXiv:2007.06823.

Kaiser, B. E., Saenz, J. A., Sonnewald, M., & Livescu, D. (2021). Objective discovery of dominant dynamical processes with intelligible machine learning. *arXiv preprint*. arXiv:2106.12963.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint*. arXiv:1412.6980.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint*. arXiv:2202.01602.

Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint*. arXiv:2202.01875.

Letzgus, S., Wagner, P., Lederer, J., Samek, W., Müller, K.-R., & Montavon, G. (2021). Toward explainable AI for regression models. *arXiv preprint*. arXiv:2112.11407.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., et al. (2021). Trustworthy ai: From principles to practices. *arXiv preprint*. arXiv:2110.01167.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777).

Mamalakis, A., Barnes, E. A., & Ebert-Uphoff, I. (2022). Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *arXiv preprint*. arxiv:2202.03407.

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2021). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *arXiv preprint*. arXiv:2103.10005.

Mazzanti, S. (2020). Shap values explained exactly how you wished someone explained to you. *Towards Data Science*, *3*, 2020.

Mitros, J., & Mac Namee, B. (2019). On the validity of bayesian neural networks for uncertainty estimation. *arXiv preprint*. arXiv:1912.01530.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise relevance propagation: An overview. In *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 193–209).

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

NASA. (2022). ECCOv4r3 dataset. Retrieved from https://ecco-group.org/products.htm

Osawa, K., Swaroop, S., Jain, A., Eschenhagen, R., Turner, R. E., Yokota, R., & Khan, M. E. (2019). Practical deep learning with Bayesian principles. *arXiv preprint*. arXiv:1906.02506.

Rasouli, K., Hsieh, W. W., & Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, *414*, 284–293. https://doi.org/10.1016/j.jhydrol.2011.10.039

Rasouli, K., Nasri, B. R., Soleymani, A., Mahmood, T. H., Hori, M., & Haghighi, A. T. (2020). Forecast of streamflows to the arctic ocean by a Bayesian neural network model with snowcover and climate inputs. *Hydrology Research*, *51*(3), 541–561. https://doi.org/10.2166/nh.2020.164

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., et al. (2019). Tackling climate change with machine learning. *arXiv preprint*. arXiv:1906.05433.

Salama, K. (2021). Keras documentation: Probabilistic Bayesian neural networks. Retrieved from https://keras.io/examples/keras_recipes/bayesian_neural_networks/

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247–278. https://doi.org/10.1109/jproc.2021.3060483

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700). Springer Nature.

Sánchez-Arcilla, A., Gracia, V., Mösso, C., Cáceres, I., González-Marco, D., & Gómez, J. (2021). Coastal adaptation and uncertainties: The need of ethics for a shared coastal future. *Frontiers in Marine Science*, *8*, 1222. https://doi.org/10.3389/fmars.2021.717781

Scher, S., & Messori, G. (2021). Ensemble methods for neural network-based weather forecasts. *Journal of Advances in Modeling Earth Systems*, *13*(2), 1–17. https://doi.org/10.1029/2020ms002331

Seibold, C., Samek, W., Hilsmann, A., & Eisert, P. (2020). Accurate and robust neural networks for face morphing attack detection. *Journal of Information Security and Applications*, *53*, 102526. https://doi.org/10.1016/j.jisa.2020.102526

Shapley, L. S. (1953). A value for N-person games. In *Contributions to theory games (AM-28)* (Vol. 2). Princeton University Press.

Silvestro, D., & Andermann, T. (2020). Prior choice affects ability of Bayesian Neural Networks to identify unknowns. *arXiv preprint* arXiv:2005.04987.

Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. In *Advances in Neural Information Processing Systems* (Vol. *34*).

Sonnewald, M., & Lguensat, R. (2021). Revealing the impact of global heating on North Atlantic circulation using transparent machine learning. *Journal of Advances in Modeling Earth Systems*, *13*(8), e2021MS002496. https://doi.org/10.1029/2021MS002496

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

Titterington, D. (2004). Bayesian methods for neural networks and related models. *Statistical Science*, *19*(1), 128–139. https://doi.org/10.1214/088342304000000099

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. https://doi.org/10.1029/2019ms002002

Tzortzis, G. F., & Likas, A. C. (2009). The global kernel k-means algorithm for clustering in feature space. *IEEE Transactions on Neural Networks*, *20*(7), 1181–1194. https://doi.org/10.1109/tnn.2009.2019722

Valletta, J. J., Torney, C., Kings, M., Thornton, A., & Madden, J. (2017). Applications of machine learning in animal behaviour studies. *Animal Behaviour*, *124*, 203–220. https://doi.org/10.1016/j.anbehav.2016.12.005

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G., Flowerdew, J., Hemri, S., et al. (2020). Statistical postprocessing for weather forecasts—Review, challenges and avenues in a big data world. *Bulletin of the American Meteorological Society*, *102*(3), 1–44. https://doi.org/10.1175/BAMS-D-19-0308.1

Waldman, R., & Giordani, H. (2022). Ocean barotropic vorticity balances: Theory and application to numerical models.

Webb, P. (2021). *Introduction to oceanography*. Roger Williams University.

Yao, L., Leng, Z., Jiang, J., & Ni, F. (2021). Modelling of pavement performance evolution considering uncertainty and interpretability: A machine learning based framework. *International Journal of Pavement Engineering*, 1–16. https://doi.org/10.1080/10298436.2021.2001814