Prototype-Based Methods in Explainable AI and Emerging Opportunities in the Geosciences

Anushka Narayanan [1 2] **Karianne J. Bergen** [1 2 3]

Abstract

Prototype-based methods are intrinsically interpretable XAI methods that produce predictions and explanations by comparing input data with a set of learned prototypical examples that are representative of the training data. In this work, we discuss a series of developments in the field of prototype-based XAI that show potential for scientific learning tasks, with a focus on the geosciences. We organize the prototype-based XAI literature into three themes: the development and visualization of prototypes, types of prototypes, and the use of prototypes in various learning tasks. We discuss how the authors use prototype-based methods, their novel contributions, and any limitations or challenges that may arise when adapting these methods for geoscientific learning tasks. We highlight differences between geoscientific data sets and the standard benchmarks used to develop XAI methods, and discuss how specific geoscientific applications may benefit from using or modifying existing prototype-based XAI techniques.

1. Introduction

Machine learning (ML) and deep learning (DL) are powerful tools for modeling, classification, and prediction. In recent years, ML/DL has been widely adopted by scientists because these tools outperform the best domain-specific, non-ML methods with low computational costs and improved scalability for many tasks in scientific data analysis (Wang et al., 2023b). Weather and climate science is an example of a scientific domain that has recently seen a rapid adoption of ML and DL methods (Rolnick et al., 2022; Reichstein et al., 2019). ML/DL has driven impressive advances in weather forecasting models (Pathak et al., 2022; Lam et al., 2023; Bi et al., 2023), climate emulators (Beucler et al., 2021), extreme event prediction (Griffin et al., 2022), spatio-temporal forecasting (Nguyen et al., 2023), and remote sensing image analysis (Jakubik et al., 2023). However, a common argument against the unrestrained use of ML in scientific applications is that most ML/DL methods operate as "black-box" models that lack "interpretability" (Rudin, 2019; de Burgh-Day & Leeuwenburg, 2023). In scientific research, specifically in geoscientific disciplines, interpretable ML/DL models and explainable AI (XAI) techniques are critical to verifying the model has learned the correct underlying physical principles and patterns governing the data (Yang et al., 2024). This is especially important to build trust in models deployed in operational settings, such as early warning systems for extreme weather (Kuglitsch et al., 2022). Interpretability is also key when ML is used with the aim of discovering novel scientific insights or patterns in data to advance scientific research (McGovern et al., 2019). Accounting for interpretability allows researchers to go beyond predictive accuracy and showcase the equally important insight of how or why the model makes specific predictions.

In scientific research, including in the fields of climate and weather, the most commonly used interpretability methods are *post hoc* techniques (Toms et al., 2020; Ham et al., 2019). Post hoc XAI techniques such as SHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), or LRP (Bach et al., 2015) allow scientists to investigate feature patterns, quantify predictor importance, or generate saliency maps. However post hoc techniques have certain known limitations and drawbacks (Zhou et al., 2021; Neely et al., 2021; Adebayo et al., 2018). Lipton (2018) highlights the difficulty of interpreting predicitions when different XAI methods yield conflicting explanations. In an example from the geosciences, Mamalakis et al. (2022) compare a series of post hoc XAI methods to explain the decisions of a convolutional neural network (CNN) for a climate-related prediction task. The authors show that each post hoc method included in the analysis resulted in inconsistent and complex model explanations. McGovern et al. (2019) caution against using

[1]Data Science Institute, Brown University, Providence, RI, USA. [2]Department of Earth, Environmental and Planetary Sciences, Brown University, Providence, RI, USA. [3]Department of Computer Science, Brown University, Providence, RI, USA.. Correspondence to: Anushka Narayanan <anushka_narayanan@brown.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

post hoc XAI methods to derive scientific conclusions as rigorous hypothesis testing is required to ensure the viability of the insights.

In contrast, *inherently interpretable* models can alleviate some of the concerns associated with post hoc methods. An inherently interpretable model aims to explain its reasoning as it makes its prediction and does not require additional techniques to interpret the prediction after-the-fact (Rudin, 2019). Linear regression (interpretable feature weights) and decision trees (binary decision rules) (Gilpin et al., 2018) are two examples of ML methods that can be classified as inherently interpretable. However, more complex models involving DL architectures, such as CNNs, ANNs and LSTMs, generally do not share a similar inherent interpretability.

Case-based reasoning can be used as a way to integrate interpretability into complex model architectures. Case-based reasoning uses the comparison between an input and a particular instance to explain a prediction (Keane & Kenny, 2019), such as instances from the training data or counterfactual examples. A popular example of case-based reasoning is the use of learned *prototypes* embedded in the model architecture to reason out the model's predictions. Generally, the model identifies *prototypical representations* of the data during training and uses features representing the similarity between an input and the learned prototypes to make a prediction (Li et al., 2018). In this approach, the user can interpret the model's prediction by inspecting the most similar prototype(s), which represent typical features or patterns in the data. This approach to generating explanations aims to mimic the human reasoning process; for example, one might identify a bird's species by comparing key traits: its beak, wing, and feet, to an example of a 'prototypical' beak, wing or feet for a sparrow (and other candidate species) (Chen et al., 2019). Prototype-based explanation techniques offer an alternative to common post hoc XAI techniques, with the added benefit that the reasoning provided by prototype-based explanations is inherently built into the model decision making process.

Prototype-based XAI techniques are an underutilized approach that can provide inherently interpretable ML alternatives for the scientific research community already using ML. This review paper is organized as follows. In Section 2 we present a brief overview of prototypes. In Section 3, we review methods for prototype-based XAI methods and categorize the literature into (1) studies that focus on the development and visualization of prototypes, (2) studies that derive different types of prototypes, and (3) studies that use prototypes for different types of learning tasks. In Section 4, we present a perspective on how scientists can leverage and extend these methods in their research. Specifically, we highlight research tasks in the domain of climate and weather that may benefit from leveraging case-based XAI techniques where explanations are derived from computing a similarity metric to a case or previous instance. Many climate and weather research studies currently use post hoc XAI techniques to interrogate ML models (Yang et al., 2024). We argue that the application and development for prototype-based inherently interpretable XAI approaches for geoscientific data is a promising avenue for future research toward advancing scientific discovery in weather and climate science.

2. Prototype-based XAI Models

Before examining specific prototype XAI methods, we describe the general architecture of prototype-based neural networks. A prototype-based model architecture consists of standard neural network layers (e.g., convolution, recurrent, dense) that learn representative features of the dataset, a *prototype layer* that generates latent prototypes associated with each class in training and computes a similarity metric between the input and the prototypes, and a fully connected component that converts the similarity scores to the prototypes to a final output specific to the learning task (e.g., classification, regression, prediction) (Li et al., 2018). The model also needs a component that allows visualization of the latent prototypes; this latent space visualization component is necessary because prototypes are representations in the latent space, which can be difficult to interpret directly, so we project these representations to visualize them in the same space as the data.

In addition to the standard trainable parameters a neural network learns, the prototypes themselves are represented by learnable parameters that are iteratively updated according to a specified optimization procedure (Chen et al., 2019), which typically includes additional terms in the loss function. Generally, the loss function is designed to ensure explanations are meaningful: the prototypes should be human-interpretable, represent diverse representations of a target class (Chen et al., 2019), and be similar to instances in the training dataset (Li et al., 2018). This is essential to the case-based reasoning process, where a sample's prediction derives its explanations from its similarity to previously seen instances (i.e., in the case of ML, from the training dataset). When a prototype network is deployed on test samples, the model's final prediction is derived from the prototype with the highest similarity (i.e., predicting the target label of that specific prototype with the highest similarity), thus providing an explanation for its prediction.

3. Case Studies

The studies discussed below are organized into three categories. Section 3.1 traces the development of prototype-based XAI networks and how we visualize the prototypes to make them human-readable. Section 3.2 focuses on case studies that develop novel additions to the prototype network to generate different prototype-based explanations. Section 3.3 presents studies focused on developing novel ways to use the similarity metrics and discusses how they can inform the learning task and the relevant explanation. In each of the case studies, we will discuss the study's novel contribution to the field, their specific methodology and application, and any potential limitations or challenges that can arise when applying these tools for geoscience-related research.

3.1. Development and Visualization of Prototypes

3.1.1. Image-Sized Prototypes

Li et al. (2018) develop a prototype-based image classifier, as defined in Section 2, for XAI with case-based reasoning as a primary objective. The authors train their model to classify handwritten digits in the benchmark MNIST image dataset using an encoder-decoder architecture with an additional classifier network (see Figure 1). The encoder learns useful and relevant features in the image for both the classification task and image reconstruction. The prototype layer learns prototypical vectors in the latent space from the encoder. These prototypes are updated via the loss constraints described in Section 2. The prototype layer outputs a similarity metric between the input image and the set of learned prototypes. The similarity scores are used in final softmax classifier. The prototypes are visualized with the decoder network, which maps the latent prototype vectors back to the original input dimension such that the decoded prototypes are of the original input image size. The authors successfully demonstrate their prototype-classifer network achieve competitive accuracy on benchmark datasets such as MNIST in comparison with a standard CNN without the use of prototypes. The authors demonstrate a limitation of this work: the decoded prototypes are not guaranteed to correspond to realistic or plausible instances. As a result, this approach may produce indecipherable or incomprehensible prototypes images, which may not provide a human-understandable explanation.

3.1.2. ProtoPNet: Patch-based Prototypes

Chen et al. (2019) follow up on the previous study and introduce ProtoPNet, a prototypical-part-based network that is presently widely cited as the standard prototype-based network in case-based reasoning XAI. The key innovation of Chen et al. (2019) is an architecture that allows the prototypes to represent "parts" (or patches) of the input images,

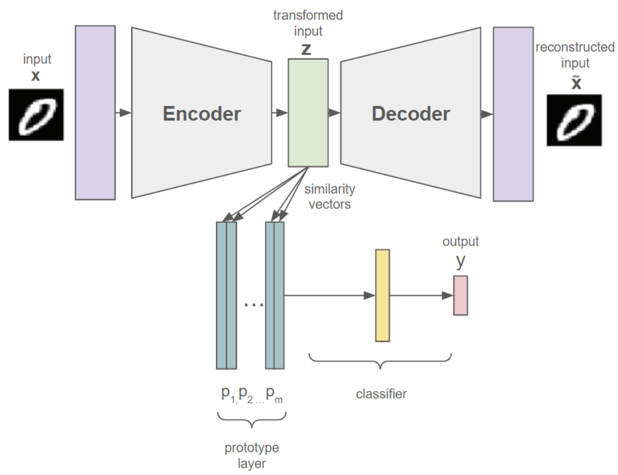

Figure 1. Diagram of the prototype-based XAI model (Li et al., 2018) with encoder-decoder architecture and classifier for an MNIST example.

rather than the whole image as in (Li et al., 2018). The latent space consists of the output of the convolution layers broken up into a grid of patches. The loss function ensures the training images contain latent patches similar to prototypes of the correct class and ensures the prototype patches are dissimilar to prototypes of other classes. In ProtoPNet, the authors constrain the prototype patches to be $1 \times 1 \times D$ where D is the number of channels from the convolution output. This allows for prototypes to represent typical attributes of local features rather than the entire image itself. Another contribution of the work is the use of a projection technique that requires the latent prototype patch to be identical to a training sample rather than just maximizing the similarity to the training sample. In doing so, the authors remove the necessity of the decoder network and rather project the latent prototypical patch onto a training sample. Since each prototype is constrained to correspond to part of the feature map from a training sample, we can trace the latent prototype back to an actual patch in the training images, which is easily visualizable and human-interpretable. The authors demonstrate their prototype network model on a bird image classification task, as the example in the introduction, where prototypes of each class (i.e., sparrow) may represent a red beak, a black eye, or other distinct, localized visual characteristics. However, the authors state in the cases where test images contain similarities to prototypical parts of multiple classes, those prototypical parts tend to correspond to patterns in the background of the image. This mandates users to employ pruning approaches to ensure prototypes consist of patterns from the object of interest and not necessarily the background.

3.1.3. Spatially Deformable Prototypes

Donnelly et al. (2022) extend ProtoPNet to produce a deformable prototype-based architecture that consists of similarly derived patch-based prototypes from the latent space as mentioned above in Chen et al. (2019) that can be organized in a spatially flexible manner. The authors demonstrate the method on the benchmark bird classification dataset used in Chen et al. (2019). In this work, the authors allow a prototype to consist of smaller rectangular prototypical patches that can change their relative spatial positioning for an input image (see Figure 2). The smaller rectangular prototypical patches are additionally constrained to be orthogonal to each other along with the orthogonal constraints among the larger prototypes themselves. The spatially flexible prototypes allow for similarities between features in the input image and the prototypical patches to vary in their orientation, allowing for flexibility and learning of distorted or obscured images. However, the additional constraints may increase training complexity, and the resulting explanations have added complexity making them potentially less human-interpretable.

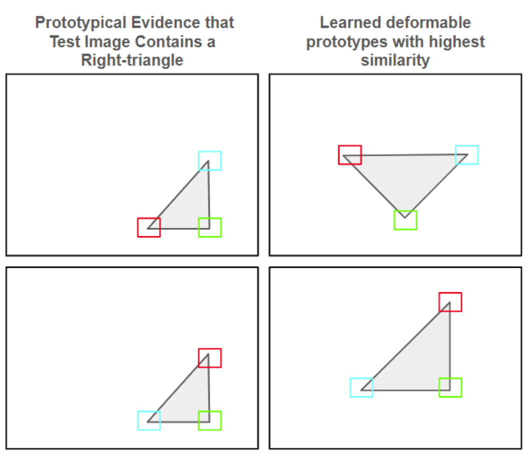

Figure 2. Notional diagram of deformable prototypes (Donnelly et al., 2022) where the prototypical patches (bounding boxes) (Chen et al., 2019) vary in their spatial organization for one input image (see Fig. 5 in original paper (Donnelly et al., 2022)).

3.2. Types of Prototypes

3.2.1. ST-ProtoPNet: Trivial and Support Prototypes

Wang et al. (2023a) present a technique where in addition to finding prototypes that are most representative of the classes in a dataset, termed *trivial prototypes*, they allow prototypes to represent instances close to the classification boundary between the target classes. These prototypes are analogous to the support vectors in SVMs and thus termed *support prototypes*. The trivial and support prototypes are notionally represented in Figure 3 for binary classification, but this approach can be extended for multi-classification. The support prototypes of different classes are constrained to be close to the decision boundary between classes, in contrast with maximizing the distance between trivial prototypes of different classes. This removes the assumption that expects a test image to contain standard trivial prototypical parts of a class. Rather, samples that contain parts resembling support prototypes may contain subtle salient, distinguishing characteristics of the sample's target class. In an object classification example shown by the authors, trivial prototypes tend to focus on obvious distinct characteristics (i.e if the background is different between classes) compared to support prototypes that focused on salient characteristics of the foreground (i.e specific distinguishing features of the object). The support and trivial prototypes are alternatively optimized within each training epoch, and a weighted combination of similarity to the trivial prototypes and the support prototypes is used in the final classification. In the author's approach, the model is limited to learn an equal number of support and trivial prototypes and the same number of total prototypes for each class. This method may need to be modified for class-imbalanced data where a flexible number of support and trivial prototypes may be more useful.

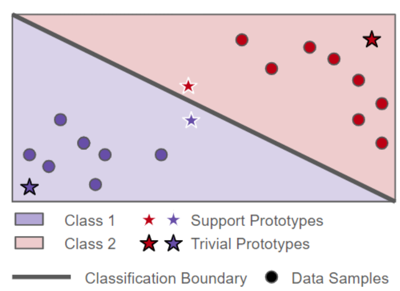

Figure 3. Notional diagram of support prototypes found closer to the classification boundary and trivial prototypes found far from each other (see Fig. 1 in original paper (Wang et al., 2023a)).

3.2.2. ProtoSeNet: Sequential Prototypes

Ming et al. (2019) develop a network designed for analyzing sequential data in which prototypes represent a prototypical series of sequential events or ordered events in the training data. The similarities between sub-sequences in the input sample and the prototypical sub-sequences from previous cases are the explanations for the classification task. To develop prototypical sequences, the authors use a recurrent DL architecture that can capture important temporal and sequential information in the data. They ensure only critical sequences are generated as the final prototypes, avoiding duplicate prototypical sequences. The authors include a projection step that assigns the prototypical sub-sequence to the closest observed sub-sequence in the training data. In order to search through all possible sub-sequences of all lengths in each training sample, the authors employ a greedy search algorithm. However, the training complexity grows quadratically with increases in the observed sample's sequence length, which indicates this method may only be feasible for classifying sequences of shorter lengths. The authors apply their technique to text classification and protein sequence classification. The authors also include a human-in-the-loop approach in which prototypes can be manually updated using user knowledge. In an experiment, they recruit non-ML experts to select prototypes that are most similar to a given target sequence (e.g., a sequence of text). The prototype options include both the highest similarity prototype from the model and other randomly chosen prototypes. If enough participants choose a prototype that the model did not choose, the prototype for that target class is manually updated.

3.2.3. Multi-variable Prototypes

Ghosal & Abbasi-Asl (2021) present a network that aims to learn representative prototypes from multi-variable time series data. They develop single-variable prototypes by extracting prototypical features from each variable separately, similar to ProtoPNet in Chen et al. (2019). They also investigate the interactions between the single-variable prototypes to develop a multi-variable prototypical representation of the data. They first generate individual single-variable prototypical patterns, then they concatenate and extract relationships between the individual single-variable patterns to generate a comprehensive multi-variable prototype-based explanation. The authors test their approach on synthetically generated multi-variable time series data and on a benchmark Epilepsy time series dataset. They reveal that the prototypes capture patterns within each feature and relationships between multiple features. However a major limitation of this method is that the number of prototypes generated increases exponentially with the number of features in a dataset. For example, the authors use a synthetic dataset of 4 single-variable features (3 relevant to the target class, 1 random noise feature) in which the number of multi-variable prototypes is 4^3, or 64, to account for all the combinations of the 3 relevant single-variables in the multi-variable prototype. This may be computationally infeasible to use for higher-dimensional data and can result in increasingly numerous explanations to navigate through for the user.

3.2.4. ProtoAD: Anomaly Detection in Time series Prototypes

Li et al. (2023) use a prototype-based approach to explain anomalies in time series data. The authors initially compute prototypical sequential data from training data consisting of "normal" (non-anomalous) time series data using an LSTM-autoencoder in an unsupervised manner, using a reconstruction-based loss function. The authors detect anomalies in the time-series via anomaly scores. The anomaly score is computed by the model by comparing sequences in the time series to the prototypical sequences learned from the normal data. The authors do not prescribe a threshold for when an anomaly is detected; rather they evaluate performance metrics on the quality of the model-generated anomaly scores with real-valued anomaly scores. The authors apply their method to find anomalous behaviour in a synthetic test case of a sine wave function with random noise injections as well as various real-world benchmark datasets with competitive accuracies to non-prototype based anomaly detection methods. However, the authors' method requires the training data to be free of anomalies, which may prove difficult to guarantee for non-benchmark observed datasets without extensive domain-specific knowledge.

3.3. Prototype Reasoning Across Learning Tasks

3.3.1. ProtoTree: Decision Trees for Learning Tasks

Nauta et al. (2021) present a combination of a prototype-based network and a decision-tree-based classification method, termed ProtoTree. The similarity scores to prototypes are passed as features into a decision tree that makes the final class prediction. For example, a simplified tree-based reasoning may look like a series of binary questions asking whether a bird image has high similarity to particular prototypes, which may be associated with different classes. In this example, the model first checks whether the input image contains the red beak prototype. If true, then it looks for a blue wing prototype. Through this reasoning process, the decision tree will output the classification. This allows for target classes to share similar prototypes (i.e both sparrow and robin have black eye color prototypes) and does not require the test image to contain all prototypes associated with a particular class. The tree-structure also allows for pruning of prototypes by removing leaves that have little discriminative power in the learning task. Pruning reduces the chance that multiple similar prototypes are included in the model. Compared to ProtoPNet, ProtoTree achieves similar accuracies with a significantly reduced number of prototypes, around 90% of the prototypes needed for ProtoPNet. However, training must be performed in two stages: first to learn the optimal prototypes and then to learn the optimal splits in the decision tree. The authors point out that learning these parameters simultaneously resulted in an overly complex and inaccurate model. ProtoTree is able to generate local explanations by tracing a specific test instance's path through the decision tree. Global explanations of the model's reasoning for each target class can also be generated from the full decision tree.

3.3.2. NP-ProtoPNet: Learning Tasks with Negative Reasoning

Singh & Yow (2021) present NP-ProtoPNet, a method that closely resembles ProtoPNet (Chen et al., 2019) with the additional of a novel reasoning technique. ProtoPNet relies heavily on positive reasoning, where a network makes a classification based on how similar the patches in the input image are with a prototypical patch. In addition to positive reasoning, NP-ProtoPNet network uses negative reasoning: it can reject classes based on the absence of matching prototypical patterns between other classes and the input image. Similarity scores will be positive if the input contains to prototypical patterns of certain classes or negative if the input does not contain prototypical patterns of other classes. An equal combination of the positive and negative similarity scores is used to make final prediction. However, in this equal combination approach, models can tend to make predictions with using more negative reasoning, if the absolute value of the negative similarity scores are higher than the absolute value of the positive similarity scores (i.e the model is very confident the input does not look like Class A,B, or C but only mildly confident the input sample looks like Class D.) The authors demonstrate this method on the classification on X-ray images for the presence of Covid-19.

3.3.3. ProtoLNet: Learning Tasks with Location Scaling

Barnes et al. (2022) extend ProtoPNet (Chen et al., 2019) to generate prototypes along with learned *location scaling* components for the classification of the phases of the Madden Julian Oscillation (MJO), an important atmospheric phenomenon impacting tropical weather at monthly scales, using environmental features, such as wind speed and longwave radiation. In many geoscience applications, the location of features or phenomena is important. This may not hold true for standard image classification applications using a standard ProtoPNet model. For example, in bird classification, the input image can share similarities to a learned prototypical part such as a red beak. However the location of the red beak in the input image does not matter, it is only necessary the input image contain parts similar to a red beak. By contrast, in climate applications in which the inputs are gridded climate data rather than natural images, the absolute location (e.g., near the poles vs. near the equator) may be critical for classification. The authors develop ProtoLNet, a method that learns prototype patches along with the location in the image where the prototype is important. For each phase of the MJO, the authors generate prototypical patterns of the environmental features and the associated location where these patterns are important. Figure 4 represents a notional diagram of ProtoLNet's reasoning for a simple location-relevant learning task. The location scaling grid shows where these prototypes are important, however, an important consideration is that, one location scaling grid is identically applied for all features in the data; in the MJO use-case, prototypical patterns of both wind speed and radiation are required to be present in the same location rather than in unique locations for each feature.

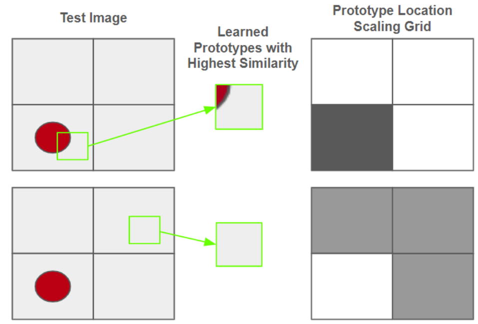

Figure 4. Notional representation of ProtoLNet's (Barnes et al., 2022) prototype-based reasoning for an example location-relevant learning task classifying which quadrant the object is located. We find those prototypes with the highest similarity to the test image. The location scaling grid shows where each prototype is important for this classification.)

4. Discussion

In this section, we first discuss the unique properties of geoscientific data and explore how potential geoscience learning tasks may benefit from the use of the prototype-based XAI methods highlighted above. We also discuss literature that addresses the general limitations of prototype-based explanations and XAI in general, considerations for evaluating their level of interpretability, and different cautions to be aware of when using these methods in the scientific domain.

Geoscientific data has unique characteristics that differ from the standard natural image or language data typically used in prototype-based XAI research and thus requires particular attention when using prototype-based techniques. For example, geoscientific data often include both a temporal component, gridded spatial data contain spatial autocorrelations, datasets are often high-dimensional or contain data from multiple spectral channels. Data sets may be multimodal, with information combined from multiple sources or sensing modalities, and both observational and simulation-generated data sets can be massive in volume (Karpatne et al., 2019). Unlike natural images that usually contain objects of sharp edges, colors or features over a distinct background, geoscientific phenomena tend to be multi-spatial scale and have ambiguous boundaries such as weather fronts at the meso-scale level ($\approx 10^{-1}$ to 10^2 km) and tropical cyclones at the synoptic-scale level ($\approx 10^3$ to

10^4 km). Typically, when we treat gridded climate data (for example, a latitude-longitude grid of temperature, wind speed, and other meteorological variables) like natural images in computer vision, each channel (temperature, wind speed, etc.) is often a separate, independent feature in a particular sample (Barnes et al., 2022). In contrast, natural RGB images use three channels to represent one color. Thus it may not be appropriate to process or interpret gridded climate data in the same manner as RGB natural images. In addition, many geoscientific events of scientific interest lack adequate training data (Ham et al., 2005); for example in the case natural disasters such as tropical cyclones and flash floods, observational datasets will typically contain very few instances of extreme events (Kuglitsch et al., 2022; Karpatne et al., 2019).

The applications of image-sized prototypes (Li et al., 2018) can be appropriate for climate and weather learning tasks that require understanding global modes of a target class. Modes derived using Principal Component Analysis (PCA; also referred to as Empirical Orthogonal Function, or EOF, analysis), are commonly used in climate science to analyze climate patterns (Hannachi et al., 2007; Kao & Yu, 2009) such as El Niño (Wang et al., 2017). Image-sized prototypes offer an alternative method for deriving climate modes, without the constraint of orthogonality imposed in PCA. In contrast, patch prototypes (Li et al., 2018) can be used to investigate localized features or patterns within a larger input grid. For example in climate prediction tasks, an image-sized prototype can represent a global climate mode (e.g., El Niño phase), while a patch prototype may point to localized features in a specific region (e.g., sea surface temperature patterns in the eastern Pacific) (Rivera Tello et al., 2023). Deformable prototypes (Donnelly et al., 2022) enable a flexible configuration of prototype patch clusters within the images. Deformable prototypes would be appropriate for modeling multi-scale interactions such as precipitation events (Tan et al., 2024; Prein et al., 2023) where differently sized and organized prototype patches can inform prototypical patterns for a range of spatial scales.

Geoscientific tasks often involve using DL for spatio-temporal forecasting such as short-term weather variables (Suleman & Shridevi, 2022), turbulent flow (Wang et al., 2020) and climate oscillations (Geng & Wang, 2021; Wang et al., 2023c) where generating sequential or temporal explanations can help identify critical events for different forecast lead-times. Originally tested on sequence classification, the sequential prototype approach of Ming et al. (2019) could be adapted for time-series forecasting. In this approach, a DL model would be able to generate explanations for a forecast depending on the presence of key events, identified by their similarity to salient prototypical sub-sequences. Extending the work in Barnes et al. (2022) to explore prototypes with individual feature (channel)-specific location scaling grids may help gain more insight into the role specific individual meteorological variables play in climate phenomena. While prototype-based methods have been developed for spatial and temporal data, there has been limited work on combining these approaches for interpretable spatio-temporal data analysis.

Prototype-anomaly scores (Li et al., 2023) can be useful in detecting anomalous changes in the testing data or whether it is out-of-distribution in an unsupervised manner. Anomaly scores may be useful in the case of climate research, where data is usually non-stationary given natural and anthropogenic climate change (i.e true distribution of the data is always evolving). Learning support prototypes (Wang et al., 2023a) can help distinguish harder to discern patterns, which may be useful in detecting climate and weather phenomena that contain more subtle distinguishing characteristics and are harder to predict, such as atmospheric rivers (Chapman et al., 2019) and tropical cyclones (Galea et al., 2023). Data points with characteristics that are more similar to support prototypes compared to trivial prototypes can point to data points with salient characteristics. For extreme event prediction such as cyclone intensification (Xu et al., 2021) or extreme precipitation events (Franch et al., 2020), one can also consider using the techniques developed in Singh & Yow (2021) and Nauta et al. (2021) to further understand the decision making process when learning tasks may rely on negative feature contributions or benefit from binary tree-like decision making. In standard prototype approaches, models make decisions that are solely based on positive similarity of the input image to a prototype. However in applications where an input image does not contain a crucial prototypical pattern, using negative reasoning or binary tree-like reasoning may benefit the model's prediction accuracy. This can be especially important in extreme event detection and related high-impact decision-making (McGovern et al., 2017).

In the case of hyperspectral satellite imagery, a common data-type in geoscientific research (Rolf et al., 2024), techniques are developed to reduce their large dimensionality (Santara et al., 2017). The motivation behind this is to gain insight into distinguishing features that may not be visible in all spectral bands of a satellite image and find specific salient bands that contain relevant information for a task, in contrast with standard RGB image channels. Generating multi-variable prototype-based explanations as described in Ghosal & Abbasi-Asl (2021) may be a promising alternative method to investigate prototypical patterns within individual spectral channels and the relevant multi-spectral prototypical patterns that are salient for a model's decision process. In general, standard feature importance techniques point to how single-variable features contribute to a model prediction, a common research motivation in geosciences (O'Gorman & Dwyer, 2018; Qiu et al., 2018; Wei et al., 2023). Generating multi-variable explanations can provide additional information on how features both individually and in combination contribute to a model prediction, which can be especially important for geoscientific phenomena that tend to be driven by complex multiple environmental features or forcings.

While we discuss the potential opportunities for adapting prototype-based XAI techniques in the geosciences, we must also consider evaluating the utility of prototype-based explanations when using them in any scientific domain. For prototype-based explanations, in particular, it is important to avoid generating a large number of prototypes to prevent over-complicating explanations. Though we need the prototypes to be generally representative of the data, prototypes should also be sparse in nature and provide a simple 'explanation'. One can consider including pruning approaches noted in Nauta et al. (2021) and Chen et al. (2019) that can generate simpler and non-duplicative prototype based explanations. Domain scientists should play integral roles in evaluating the robustness and reliability of the generated prototype-based explanations for a learning task. For example, including a domain specialist in the human-in-the-loop approach described in Ming et al. (2019) may help reduce the redundancy of the model-generated prototypes to making the explanation easier to comprehend. The case studies discussed above vary in the types of prototypes generated, model and training complexity and the generated final explanation. Depending on the specific learning task and the level of complexity of the explanation that is needed, domain scientists can aid in choosing the appropriate method such that prototypes provide an informative yet concise explanation during the model decision-making process.

Hoffmann et al. (2021) demonstrate the potential pitfalls of using prototype-based networks for generating explanations. The authors find that introducing artifacts into the input data that are imperceptible to humans, can drastically alter the explanations despite the image and the altered image looking virtually identical to each other. This is an essential critique of prototype-based methods because they rely on comparing the similarity of the latent representation of a training sample with the test sample to establish an 'explanation'. Artifacts can alter the latent representation significantly which can alter the 'explanation' even if a human would mark the altered training sample as similar to the testing sample. To combat these issues, the authors recommend implementing data augmentation techniques and adversarial training to prevent the introduction of artifacts and train the model to recognize the samples augmented with artifacts. Similar pitfalls arise in using traditional post hoc XAI techniques. Ghorbani et al. (2019) and Alvarez-Melis & Jaakkola (2018) show that perturbations to input data produce varied 'explanations' when using post hoc XAI techniques indicating a similar risk in their robustness and stability. Kim et al. (2022) establish a framework, HIVE: Human Interpretability of Visual Explanations that assesses how explanations derived from AI models can help or hinder human decision making. Huang et al. (2023) establish a quantitative interpretability benchmark metric to follow when interrogating the consistency and stability of the prototype-based explanations. Rigorous testing and robustness evaluation of the prototype-based XAI methods should be included in best practices when employing these methods in geosciences, especially when these methods are involved with generating scientific insight and/or involved with high-stakes decision-making.

5. Conclusion

In this review, we present case studies from the emerging field of prototype-based XAI, a set of methods that aim to develop built-in interpretability for complex deep learning architectures. These case studies show promising avenues for further development for use-cases in geoscientific research, where there has been significant recent progress through the use of AI. We include case studies that focus on the the development of the prototypes and their visualizations, studies that explore the different form of prototypes, and studies that focus on using prototypes in different types of learning tasks. By adapting these methods and tailoring them for geoscience specific domain tasks, we can generate accurate, reliable, and insightful models. When considering the methods showcased for geosciences applications, one should always consider what the end-goal task is and what form of explanation would be most useful for the specific task. We include potential avenues to use prototype-based methods for geoscientists to explore in their respective research domains. We also include a discussion on studies that have developed rigorous testing and evaluation frameworks to evaluate the robustness and test the integrity of prototype-based explanations when employing these techniques.

Though this review focuses on learning tasks with geoscientific data, the points raised in this review also apply more generally to other scientific disciplines that share similar research tasks. Most prototype-based XAI approaches have been developed for standard natural image, text and natural language benchmark datasets. With the increasing adoption of ML in the sciences, there is greater need for XAI tools designed with the unique properties of scientific data and needs of scientific researchers in mind. Such tools, especially when developed in collaboration between scientists and XAI researchers, have the potential to advance both fields.

Acknowledgements

This work was supported by the SciAI Center, and funded by the Office of Naval Research (ONR), under Grant Number N00014-23-1-2729. The authors thank the students in EEPS-DATA 1720 (Brown University, Spring 2024) for their peer feedback on early drafts of this review.

Impact Statement

The prototype-based XAI techniques discussed in this work may be implemented with the goal of increasing trustworthiness in scientific applications of ML, especially those that involve high stakes decision-making in societal contexts. Broader Impacts are discussed in the Discussion section of the main text, including our recommendation that generated prototype-based explanations be thoroughly evaluated, with inputs from domain experts, for robustness and reliability before using them as justifications in real-world scenarios.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31, 2018.

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

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

Barnes, E. A., Barnes, R. J., Martin, Z. K., and Rader, J. K. This looks like that there: Interpretable neural networks for image tasks when location matters. *Artificial Intelligence for the Earth Systems*, 1(3):e220001, 2022.

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., and Gentine, P. Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, 126(9):098302, 2021.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023.

Chapman, W., Subramanian, A., Delle Monache, L., Xie, S., and Ralph, F. Improving atmospheric river forecasts with machine learning. *Geophysical Research Letters*, 46 (17-18):10627–10635, 2019.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This looks like that: deep learning for interpretable image recognition. *Advances in Neural Information Processing systems*, 32, 2019.

de Burgh-Day, C. O. and Leeuwenburg, T. Machine learning for numerical weather and climate modelling: a review. *Geoscientific Model Development*, 16(22):6433–6477, 2023.

Donnelly, J., Barnett, A. J., and Chen, C. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10265–10275, 2022.

Franch, G., Nerini, D., Pendesini, M., Coviello, L., Jurman, G., and Furlanello, C. Precipitation nowcasting with orographic enhanced stacked generalization: Improving deep learning predictions on extreme events. *Atmosphere*, 11(3):267, 2020.

Galea, D., Kunkel, J., and Lawrence, B. N. Tcdetect: a new method of detecting the presence of tropical cyclones using deep learning. *Artificial Intelligence for the Earth Systems*, 2(3):e220045, 2023.

Geng, H. and Wang, T. Spatiotemporal model based on deep learning for enso forecasts. *Atmosphere*, 12(7):810, 2021.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019.

Ghosal, G. R. and Abbasi-Asl, R. Multi-modal prototype learning for interpretable multivariable time series classification. *arXiv preprint arXiv:2106.09636*, 2021.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89. IEEE, 2018.

Griffin, S. M., Wimmers, A., and Velden, C. S. Predicting rapid intensification in North Atlantic and eastern North Pacific tropical cyclones using a convolutional neural network. *Weather and Forecasting*, 37(8):1333–1355, 2022.

Ham, J., Chen, Y., Crawford, M., and Ghosh, J. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):492–501, 2005.

Ham, Y.-G., Kim, J.-H., and Luo, J.-J. Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775):568–572, 2019.

Hannachi, A., Jolliffe, I. T., and Stephenson, D. B. Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 27(9):1119–1152, 2007.

Hilburn, K. A., Ebert-Uphoff, I., and Miller, S. D. Development and interpretation of a neural-network-based synthetic radar reflectivity estimator using goes-r satellite observations. *Journal of Applied Meteorology and Climatology*, 60(1):3–21, 2020.

Hoffmann, A., Fanconi, C., Rade, R., and Kohler, J. This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. *arXiv preprint arXiv:2105.02968*, 2021.

Huang, Q., Xue, M., Huang, W., Zhang, H., Song, J., Jing, Y., and Song, M. Evaluation and improvement of interpretability for self-explainable part-prototype networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2011–2020, 2023.

Iglesias-Suarez, F., Gentine, P., Solino-Fernandez, B., Beucler, T., Pritchard, M., Runge, J., and Eyring, V. Causally-informed deep learning to improve climate models and projections. *Journal of Geophysical Research: Atmospheres*, 129(4):e2023JD039202, 2024.

Jakubik, J., Roy, S., Phillips, C., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

Kao, H.-Y. and Yu, J.-Y. Contrasting eastern-pacific and central-pacific types of enso. *Journal of Climate*, 22(3): 615–632, 2009.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1544–1554, 2019.

Keane, M. T. and Kenny, E. M. How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems. In *Case-Based Reasoning Research and Development: 27th International Conference, ICCBR 2019, Otzenhausen, Germany, September 8–12, 2019, Proceedings 27*, pp. 155–171. Springer, 2019.

Kim, J.-H., Ham, Y.-G., Kim, D., Li, T., and Ma, C. Improvement in forecasting short-term tropical cyclone intensity change and their rapid intensification using deep learning. *Artificial Intelligence for the Earth Systems*, 3 (2):e230052, 2024.

Kim, S. S., Meister, N., Ramaswamy, V. V., Fong, R., and Russakovsky, O. HIVE: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision*, pp. 280–298. Springer, 2022.

Kuglitsch, M., Albayrak, A., Aquino, R., Craddock, A., Gill, J. E., Kanwar, R., Koul, A., Ma, J., Marti, A., Menon, M., et al. Artificial intelligence for disaster risk reduction: Opportunities, challenges, and prospects. *World Meteorological Organization (WMO) Bulletin*, 71(1):30–37, 2022.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Li, B., Jentsch, C., and Müller, E. Prototypes as explanation for time series anomaly detection. *arXiv preprint arXiv:2307.01601*, 2023.

Li, O., Liu, H., Chen, C., and Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Lipton, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Mamalakis, A., Ebert-Uphoff, I., and Barnes, E. A. Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pp. 315–339. Springer, 2020.

Mamalakis, A., Barnes, E. A., and Ebert-Uphoff, I. Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, 1(4):e220012, 2022.

Mayer, K. J. and Barnes, E. A. Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophysical Research Letters*, 48(10):e2020GL092092, 2021.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T., and Williams, J. K. Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, 98(10):2073–2090, 2017.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T. Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, 100(11):2175–2199, 2019.

Ming, Y., Xu, P., Qu, H., and Ren, L. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 903–913, 2019.

Molina, M. J., O'Brien, T. A., Anderson, G., Ashfaq, M., Bennett, K. E., Collins, W. D., Dagon, K., Restrepo, J. M., and Ullrich, P. A. A review of recent and emerging machine learning applications for climate variability and weather phenomena. *Artificial Intelligence for the Earth Systems*, 2(4):220086, 2023.

Nauta, M., Van Bree, R., and Seifert, C. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14933–14943, 2021.

Neely, M., Schouten, S. F., Bleeker, M. J., and Lucic, A. Order in the court: Explainable AI methods prone to disagreement. *arXiv preprint arXiv:2105.03287*, 2021.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 25904–25938, 23–29 Jul 2023.

O'Gorman, P. A. and Dwyer, J. G. Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, 10(10):2548–2563, 2018.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Prabhat, Kashinath, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., Karaismailoglu, E., von Kleist, L., Kurth, T., Greiner, A., et al. Climatenet: An expert-labelled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development Discussions*, 2020:1–28, 2020.

Prein, A. F., Mooney, P. A., and Done, J. M. The multi-scale interactions of atmospheric phenomenon in mean and extreme precipitation. *Earth's Future*, 11(11): e2023EF003534, 2023.

Qiu, C., Schmitt, M., Mou, L., Ghamisi, P., and Zhu, X. X. Feature importance analysis for local climate zone classification using a residual convolutional neural network with multi-source datasets. *Remote Sensing*, 10(10):1572, 2018.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, f. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

Rivera Tello, G. A., Takahashi, K., and Karamperidou, C. Explained predictions of strong eastern pacific el niño events using deep learning. *Scientific Reports*, 13(1): 21150, 2023.

Rolf, E., Klemmer, K., Robinson, C., and Kerner, H. Mission critical–satellite data is a distinct modality in machine learning. *arXiv preprint arXiv:2402.01444*, 2024.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Santara, A., Mani, K., Hatwar, P., Singh, A., Garg, A., Padia, K., and Mitra, P. BASS net: Band-adaptive spectral-spatial feature learning neural network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5293–5301, 2017.

Singh, G. and Yow, K.-C. These do not look like those: An interpretable deep learning model for image recognition. *IEEE Access*, 9:41482–41493, 2021.

Suleman, M. A. R. and Shridevi, S. Short-term weather forecasting using spatial feature attention based lstm model. *IEEE Access*, 10:82456–82468, 2022.

Tan, J., Huang, Q., and Chen, S. Deep learning model based on multi-scale feature fusion for precipitation nowcasting. *Geoscientific Model Development*, 17(1):53–69, 2024.

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

Wang, C., Deser, C., Yu, J.-Y., DiNezio, P., and Clement, A. *El Niño and Southern Oscillation (ENSO): A Review*, pp. 85–106. Springer Netherlands, Dordrecht, 2017.

Wang, C., Liu, Y., Chen, Y., Liu, F., Tian, Y., McCarthy, D., Frazer, H., and Carneiro, G. Learning support and trivial prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2062–2072, 2023a.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023b.

Wang, H., Hu, S., and Li, X. An interpretable deep learning enso forecasting model. *Ocean-Land-Atmosphere Research*, 2:0012, 2023c.

Wang, R., Kashinath, K., Mustafa, M., Albert, A., and Yu, R. Towards physics-informed deep learning for turbulent flow prediction. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1457–1466, 2020.

Wei, M., Fang, G., and Ge, Y. Tropical cyclone genesis prediction based on support vector machine considering effects of multiple meteorological parameters. *Journal of Wind Engineering and Industrial Aerodynamics*, 243: 105591, 2023.

Xu, W., Balaguru, K., August, A., Lalo, N., Hodas, N., DeMaria, M., and Judi, D. Deep learning experiments for tropical cyclone intensity forecasts. *Weather and Forecasting*, 36(4):1453–1470, 2021.

Yang, R., Hu, J., Li, Z., Mu, J., Yu, T., Xia, J., Li, X., Dasgupta, A., and Xiong, H. Interpretable machine learning for weather and climate prediction: A survey. *arXiv preprint arXiv:2403.18864*, 2024.

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

A. Appendix

Table 1. Column 1 depicts characteristics and properties of geoscientific data. Column 2 states the specific ML tasks in which they're commonly used. Column 3 states the challenges in using them in standard ML and XAI methods. Column 4 includes an example from the literature for the geoscientific task. For a more in-depth review on the development and usage of machine learning and post-hoc XAI analysis in the geosciences, one can refer to Molina et al. (2023), Yang et al. (2024), de Burgh-Day & Leeuwenburg (2023) and Mamalakis et al. (2020).

Geoscientific Data Characteristics	**Geoscientific ML Task**	**Challenges with Standard ML and XAI techniques**	**Literature Example**
Phenomena of interest have ambiguous boundaries and can blend with image background	geoscientific feature detection	Image recognition models rely on features like defined boundaries or sharp edges over a distinct background	tropical cyclone and atmospheric river segmentation (Prabhat et al., 2020)
Gridded data contains separate, independent environmental features	climate prediction	Often treated as RGB channels in an image representing a singular color	prediction of strong El Niño events (Rivera Tello et al., 2023)
Hyper- or multi-spectral satellite data	geoscientific feature/channel selection, dimensionality-reduction	High dimensionality with relevant information spread across multiple channels in contrast with RGB natural images	estimating radar reflectivity using remote-sensing (Hilburn et al., 2020)
Spatio-temporal, location-specific patterns, temporal patterns	climate mode variability, spatio-temporal forecasting	Sequential methods are developed to work with text or natural language data (tokens vs. continuous signals)	sub-seasonal tropical-extratropical circulation forecasting (Mayer & Barnes, 2021)
Domain-shift, anomalous data	climate prediction	Poor performance on out of distribution data	climate projections under different climate regimes or distributions (Iglesias-Suarez et al., 2024)
Limited observed samples of a rare or extreme event of interest	extreme weather, climate prediction	Class imbalanced data, poor prediction performance	prediction of extreme rapid-intensification cyclone events (Kim et al., 2024)

Table 2. For each case study in Column 1, Column 2 states the explanation each study provides using prototypes and Column 3 states potential geoscientific learning tasks for this method.

Case Study	Summary Explanation	Geoscientific Learning Tasks
Li et al. (2018)	This input image is overall resembles a learned prototypical image of the target class.	climate phase prediction, dimensionality reduction
Chen et al. (2019)	This specific patch in input image resembles a learned prototypical local patch or pattern directly from a patch in a training image.	generating local feature patterns
Donnelly et al. (2022)	This organized cluster of prototypical patches resembles a set of prototypical patches within an image of the target class.	organization of feature patterns, detecting multi-scale feature patterns
Wang et al. (2023a)	This input may resemble prototypes representative of a target class or the target class boundary.	hard-to-learn feature detection, extreme event detection
Ming et al. (2019)	A window of this input sequence resembles a prototypical window sequence in the target class.	temporal forecasting, climate oscillations prediction
Ghosal & Abbasi-Asl (2021)	This multi-variable input shares similarities with single variables prototypes and their relationships with each other.	multi-forcing classification, multi-variable feature attribution, multi-spectral imagery classification
Li et al. (2023)	This input sequence of data significantly differs from prototypical window sequences in the training data.	anomaly detection, extreme event detection
Nauta et al. (2021)	The prediction is derived from a tree-like reasoning process on whether the input contains similarities to certain prototypes.	multi-variable prediction tasks
Singh & Yow (2021)	This input image contains similar prototypical parts of the target class and does not contain prototypical parts from another class.	climate prediction and classification, extreme event prediction
Barnes et al. (2022)	This input image contains similar prototypical parts of the target class only in specific regions of the image.	generating spatially relevant feature patterns