Interpretable Machine Learning for Weather and Climate Prediction: A Survey

Ruyi Yang [1] *yangruyi853@gmail.com*
Jingyu Hu [2] *ym21669@bristol.ac.uk*
Zihao Li [3] *lizihao9885@gmail.com*
Jianli Mu [1] *mujl668@sina.com*
Tingzhao Yu [1] *tsingzao@hotmail.com*
Jiangjiang Xia [4] *xiajj@tea.ac.cn*
Xuhong Li [5] *jacqueslixuhong@gmail.com*
Aritra Dasgupta [6] *aritra.dasgupta@njit.edu*
Haoyi Xiong [5] *haoyi.xiong.fr@ieee.org*
[1] *Public Meteorological Service Center, China Meteorological Administration* [2] *University of Bristol*
[3] *Zhejiang University* [4] *Institute of Atmospheric Physics, Chinese Academy of Science* [5] *Baidu Inc.*
[6] *New Jersey Institute of Technology*

Abstract

Advanced machine learning models have recently achieved high predictive accuracy for weather and climate prediction. However, these complex models often lack inherent transparency and interpretability[1], acting as "black boxes" that impede user trust and hinder further model improvements. As such, interpretable machine learning techniques have become crucial in enhancing the credibility and utility of weather and climate modeling. In this survey, we review current interpretable machine learning approaches applied to meteorological predictions. We categorize methods into two major paradigms: 1) Post-hoc interpretability techniques that explain pre-trained models, such as perturbation-based, game theory based, and gradient-based attribution methods. 2) Designing inherently interpretable models from scratch using architectures like tree ensembles and explainable neural networks. We summarize how each technique provides insights into the predictions, uncovering novel meteorological relationships captured by machine learning. Lastly, we discuss research challenges around achieving deeper mechanistic interpretations aligned with physical principles, developing standardized evaluation benchmarks, integrating interpretability into iterative model development workflows, and providing explainability for large foundation models.

1 Introduction

Weather and climate change have a significant impact on social, economic, and environmental systems around the world. Therefore, accurate weather forecasting and climate prediction are crucial to hazard preparation, resource management, and understanding long-term climate change. Traditionally, these predictions have relied heavily on complex numerical models that solve fundamental physics equations influencing atmospheric dynamics (Richardson, 1922), such as Numerical Weather Prediction (NWP) models and General Circulation Models (GCMs). However, these physics-based numerical predictions have some major limitations, including uncertainties in initial conditions, incomplete representations of sub-grid processes, and constraints on spatial resolution and computing power. In recent years, machine learning (ML) techniques, particularly deep learning models, have achieved dramatic progress in processing massive datasets, characterizing spatial

[1]In this paper, the terms "explanation" and "interpretation," as well as "explainability" and "interpretability," and "explainable" and "interpretable" are used interchangeably.

features (Du et al., 2020), mining time correlations (Lee et al., 2020), super-resolution downscaling (Leinonen et al., 2020), and extracting spatial-temporal series in model predictions (Guo et al., 2021). As a result, an increasing number of scientific and business entities have incorporated machine learning into weather forecast and climate prediction (Yu & Yang, 2023; Qian & Jia, 2023; Chkeir et al., 2023; Yang et al., 2022; Yu et al., 2022; Arcomano et al., 2020; Weyn et al., 2019; Scher & Messori, 2018). The outstanding performance of large foundation models of weather forecast such as ClimaX (Nguyen et al., 2023) and GraphCast (Lam et al., 2023) shows the potential of machine learning-based prediction models in meteorological prediction.

Despite their predictive capabilities, most advanced ML models used for meteorology are usually regarded as "black boxes", lacking inherent transparency in their underlying logic and feature attributions (Du et al., 2019; Deng et al., 2021; Xiong et al., 2024). This lack of interpretability poses major challenges. First, it reduces trust from domain experts, such as meteorologists, who may be reluctant to rely on unexplained model outputs for high-stakes decision making. Second, it hinders further model refinement, as developers cannot easily diagnose errors or identify which relationships the models have captured. Third, opaque ML models provide limited insight into the fundamental atmospheric processes that lead to their predictions.

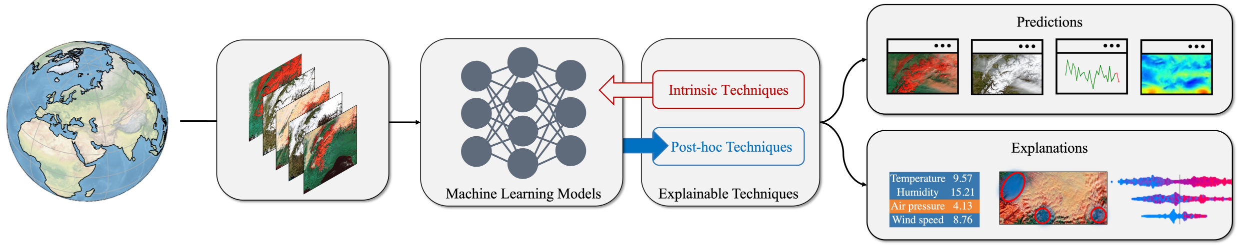
To address these limitations, explainable machine learning (Fig. 1a) techniques have become essential to enhance trust in predictions, facilitate further model improvements, and uncover new meteorological insights (Labe et al., 2023; Arrieta et al., 2020; McGovern et al., 2019). Despite initial progress in applying explainability techniques in weather and climate prediction applications, a systematic framework is still needed to summarize current research and challenges. However, existing surveys either focus mainly on reviewing explainability methods for general ML fields, as exemplified by recent surveys (Du et al., 2019; Murdoch et al., 2019), or review how to apply ML approaches to weather and climate predictions (Bochenek & Ustrnul, 2022; Ren et al., 2021).

Addressing this research gap, our work presents a comprehensive survey of the current advancements in applying explainability techniques across various methodological predictions. We categorize explainability into post-hoc explanations, which provide explainations for pre-trained models, and inherently interpretable models designed from scratch. For example, we examine representative post-hoc explanation methods such as SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017), Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), and Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) for weather and climate predictions. In addition, we also analyze the strengths and weaknesses of each explainability technique and highlight promising directions for future research.

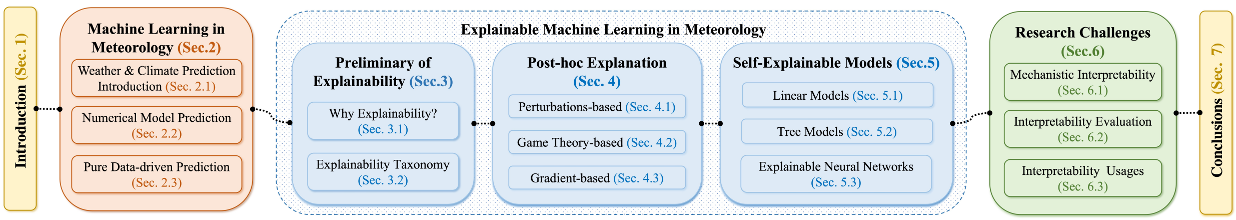
In conclusion, our survey provides a comprehensive review of interpretable machine learning applications in weather and climate prediction (Fig. 1b). Section 2 introduces machine learning in meteorology, breaking it down into two sub-categories: numerical model prediction improvement, and pure data-driven prediction. Section 3 addresses the preliminary aspects of explainability, including the reasons why we need explainability and a taxonomy of explainability techniques. Section 4 offers an overview of post-hoc explanation methods, such as perturbation-based, gradient-based, and game theory-based approaches. Section 5 introduces self-explainable models, including linear models, tree models, and explainable neural networks. Finally, Section 6 highlights the research challenges in the field, discussing mechanistic interpretability, evaluation of interpretability, the usages of interpretability, and interpretability for large foundation models in more detail.

2 Machine Learning in Weather and Climate Prediction

Machine learning has made significant progress in weather and climate prediction in recent years (Bochenek & Ustrnul, 2022; Kashinath et al., 2021; Yu et al., 2021). Weather prediction is generally defined as the forecast of various meteorological elements at a certain time for up to two weeks, whereas climate prediction operates on longer timescales, ranging from a month to decades. The latter mainly relies on simulations that incorporate a wide range of variables, including atmospheric chemistry, ocean currents, land surface processes, and ice dynamics (Giorgi & Mearns, 1991). In the following, we first introduce datasets for weather and climate prediction and then present the two categories of machine learning in weather and climate prediction: improvements of numerical model prediction and pure data-driven prediction.



(a) Explainable Machine Learning in weather and climate prediction



(b) The Structure of the Survey

Figure 1: Explainable machine learning in meteorological prediction and the structure of the survey

2.1 Weather and Climate Datasets

As an important part of geoscience data, meteorological data generally contain five dimensions: meteorological variables (e.g., temperature, wind speed, humidity and air pressure), time, longitude, latitude, and altitude (Wang, 2014; Daly, 2006). These data can be regarded as either multi-variable four-dimensional spatio-temporal data or multi-variable and multi-time three-dimensional spatial data (Wang, 2019). The meteorological data used for weather and climate prediction are generally divided into two categories, observation data and numerical model data. Observation data are mainly obtained by measuring and determining atmospheric conditions and variation through special sensors and detection equipment, including ground meteorological data, upper-level meteorological data, radar meteorological data, satellite meteorological data, etc. The first two are usually site data, showing non-uniform distribution in space. The latter two are quasi-grid data, which are usually processed into grid data through interpolation and other methods. Numerical model data, on the other hand, are primarily obtained by solving mathematical physics equations that can describe atmospheric motion through numerical prediction models. These include forecast data (Molteni et al., 1996) and reanalysis data from multi-source historical meteorological data fusion (Kalnay et al., 2018; Hersbach et al., 2020). The model data are usually grid data, with latitude and longitude evenly distributed in space. Meteorological data, characterized by their vast variety and long time series, can be considered as big data. Due to the strong spatial correlation and time continuity of meteorological data, machine learning technology can be used to analyze the underlying characteristics of the data, in which observation data and reanalysis data are often used as ground-truth labels.

2.2 Machine Learning for Numerical Model Prediction Improvements

The improvement of numerical model prediction using machine learning can be further divided into three subcategories: data assimilation of numerical model, physical process of numerical model and post-processing of numerical model prediction, as shown in Table 1.

Data Assimilation of Numerical Model: Data assimilation is a critical process in numerical prediction models that involves integrating observational data into a numerical model to provide a more accurate initial state for forecasts (Wang et al., 2022; Gustafsson et al., 2018; Anderson et al., 2009). This process relies on sophisticated algorithms that balance the latest observations with prior forecasts to correct inaccuracies in the model's initial conditions. Observational data can come from a myriad of sources such as satellites, sounding balloons, radar, and ground stations, and are crucial for capturing the state of the atmosphere at a given time. By continually incorporating real-time data, numerical prediction models can significantly improve their predictive accuracy, leading to more reliable weather forecasts. Since machine learning and data assimilation are both designed to extract the relationship between prediction and influential factors for minimizing the deviation of the predicted result from the ground-truth values, machine learning models can be used instead of traditional data assimilation schemes (He et al., 2022; Wu et al., 2021; Härter & de Campos Velho, 2012). Using the data assimilation method based on machine learning model can not only reduce the initial field error of numerical weather prediction, but also effectively improve the operation efficiency (Arcucci et al., 2021; Cintra & de Campos Velho, 2018).

Physical Process of Numerical Model: The physical processes within numerical models are governed by the fundamental laws of physics, specifically the dynamics, thermodynamics, and conservation of mass and momentum. These models simulate the behavior of the atmosphere by discretizing it into a grid and solving the equations for each grid cell. Some complex evolution processes of small- and medium-scale systems cannot be described by normal grid scale, but they are essential for achieving the prediction performance. It is required to approximate these physical processes such as convection, cloud formation and radiation transfer, which is called parameterization (Bauer et al., 2015). There always has a deviation between the parameterization scheme and the real physical process, which leads to the bias between the prediction and the observed values. Due to powerful feature extraction and nonlinear fitting capabilities for big data, machine learning algorithms, especially deep learning models, can be used to improve or replace the physical parameterization process (Bodini et al., 2020; Rasp et al., 2018). Relevant research shows that deep learning models can dramatically improve the computational efficiency of the physical process of numerical prediction models, but there are certain limitations (Seifert & Rasp, 2020). Stability and prediction performance can be improved by adding physical constraints to deep learning models.

Post-processing of Numerical model Prediction: Numerical prediction models obtain the atmospheric state in the future period mainly by solving the equations of fluid mechanics and thermodynamics which depicts the atmospheric evolution over time. Due to the uncertainty of the initial field, inaccurate representation of physical or dynamic processes and the chaos of atmospheric motion, the forecast data of the numerical model often have systematic deviation from the observation. Additionally, owing to limitations of computing resources, the spatiotemporal resolution of model data cannot meet the needs of practical applications. Traditional statistical methods have limited ability to improve the accuracy and spatiotemporal resolution of numerical predictions. The use of machine learning based post-processing technology for numerical prediction can effectively improve the forecast quality. Moreover, the online learning technology of machine learning can update the model as the observation data is updated without retraining the model. Post-processing of numerical prediction models involves several steps to refine the raw model output into usable forecasts (Ma et al., 2024). The first step is the numerical prediction model itself, where the raw data from the model are interpreted to predict the future states of the weather. Following this, post-processing technology is applied, which adjusts the model output to meet practical application needs. This process improves the performance of the model by aligning it more closely with the observed reality. The post-processing of the numerical model mainly includes bias correction and statistical downscaling. The bias correction is mainly used to correct the systematic deviation of the model forecast based on observation data (Han et al., 2021; Zhang et al., 2020; Watson, 2019). Statistical downscaling is used to translate large-scale information from the numerical model to a finer resolution that is more relevant for local forecasts (Pan et al., 2019; Sachindra et al., 2018). This can involve using historical observation data to adjust the model output to reflect local weather patterns more accurately.

2.3 Pure Data-driven Machine Learning Prediction

The exponential growth of high-resolution radar observation, satellite data, numerical model output, and other meteorological data provides the data foundation for machine learning, especially deep learning, and greatly promotes the development of pure data-driven weather and climate prediction. Data-driven weather forecasts represent a paradigm shift from traditional numerical model-based predictions, leveraging the power of machine learning and big data analytics. This paradigm uses historical meteorological data to train machine learning algorithms that can identify patterns and predict future atmospheric conditions. By analyzing large datasets that include past weather events, data-driven models can uncover complex relationships and dependencies that traditional methods might overlook. These models can provide valuable insights, espe-

Table 1: Different stages of machine learning in weather and climate prediction.

Stages	**Applications**
Pre-processing	Data assimilation in weather predictions (Wu et al., 2021; Arcucci et al., 2021; Cintra & de Campos Velho, 2018; Härter & de Campos Velho, 2012); Climate predictions: (He et al., 2022)
In-processing	Improve physical parameterization in weather predictions (Seifert & Rasp, 2020); Climate predictions (Rasp et al., 2018)
Post-processing	Bias correction: (Han et al., 2021; Zhang et al., 2020; Watson, 2019) ; Statistical downscaling: (Pan et al., 2019; Sachindra et al., 2018)

cially in situations where physical models struggle due to chaotic atmospheric behavior. Furthermore, the integration of machine learning techniques can enhance the speed and efficiency of forecast generation, making it possible to provide real-time updates with increased accuracy. According to the difference in driving data, it can be divided into observational data-driven prediction and numerical output-driven prediction.

Observational Data-driven Prediction: Observational data-driven prediction means that all input data come from observational data, and this type of prediction model does not rely on numerical models at all. One of the most common machine learning models driven by observational data is nowcasting (0–2h) (Chkeir et al., 2023; Ayzel et al., 2020; Foresti et al., 2019; Agrawal et al., 2019; Shi et al., 2015). Due to the fact that numerical model takes several hours to reach equilibrium state and has limited forecasting ability for mesoscale systems, there exist some errors between the model's nowcasting and observations. Moreover, traditional extrapolation methods based on historical observation data (e.g., radar echoes, satellite images) such as the optical flow and cross-correlation algorithm cannot effectively predict the generation and extinction of convective system. Therefore, machine learning methods are widely used in nowcasting because they show great potential in quickly fusing a large number of observational data and effectively extracting nonlinear features (Zhang et al., 2023). By transforming the nowcasting into a spatiotemporal series prediction problem, machine learning model can effectively predict the generation, evolution, and extinction of convective systems in central-eastern and southern China based on multi-source observation data including satellite infrared images, radar reflectivity, and lightning density (Zhou et al., 2020).

Numerical Output-driven Prediction: Numerical output-driven prediction is defined as a forecast in which the input data are partially or completely derived from the numerical model output. For prediction with longer forecast time than nowcasting, in addition to using observation data such as radar and satellite data, it is necessary to provide atmospheric circulation fields from the ground to upper-air. Specially, from short-to-medium-term weather forecast to decadal climate prediction, input factors are usually derived from numerical model reanalysis datasets (Price et al., 2023; Chen et al., 2023a; Bi et al., 2023a; Weyn et al., 2020). These predictors are analogous to the initial fields of numerical weather prediction from which machine learning models can extract dynamical features. Currently, large foundation models of meteorological forecast based on artificial intelligence released globally, such as Fuxi from Fudan University (Chen et al., 2023b), GraphCast developed by Google DeepMind (Lam et al., 2023), and FourCastNet developed by Nvidia (Pathak et al., 2022), are driven by atmospheric reanalysis data in numerical models. This type of model does not achieve true end-to-end forecasting, and the quality of its results depends to some extent on the results of numerical models. In addition, although the performance of this type of prediction model for general weather is comparable to that of state-of-the-art numerical models, the forecast ability for extreme weather can be further improved. To solve the problem, a feasible method is to incorporate physical mechanisms when building machine learning based prediction models (Kochkov et al., 2023).

3 Preliminary of Explainability

In this section, we discuss the preliminary of explainability, including why do we need explainability and the explainability definition and taxonomy.

3.1 Why Explainability?

Despite the promising progress of machine learning in various meteorology prediction tasks, most of these complex models act as "black boxes". Their inherent lack of transparency and interpretability reduces trust from end users, such as meteorologists and forecasters, who remain unsure how forecasts are made. It also hampers model debugging and diagnosis by developers seeking to further improve predictive performance. Explainability is crucial in meteorological modeling for several key reasons:

- **Increase trust from end users:** Complex machine learning models such as deep neural networks can achieve high accuracy but act as "black boxes", lacking transparency in their predictions. This makes meteorologists reluctant to fully trust and use the outputs. Interpretable machine learning can well visualize the decision-making principles of the model, which help end users understand the model logic. If the model logic is consistent with the domain knowledge of atmospheric science, it can increase the trust of users.
- **Help developers diagnose and improve models:** Interpretability methods highlight which meteorological factors and relationships drive predictions. This allows model developers to debug errors, check if meaningful patterns are learned, and improve model architecture and data pre-processing. The ultimate goal is to increase performance metrics such as accuracy and reliability.
- **Gain meteorological insights:** Explaining model behavior can reveal new discoveries about weather and climate prediction mechanism. For example, interpretable machine learning can be used to uncover the impact of global heating on North Atlantic circulation (Sonnewald & Lguensat, 2021). These insights further advance scientific knowledge in meteorology.
- **Integration of data-driven ML models with physical mechanism:** Understanding how machine learning models make predictions can help visualize whether the logic of a weather prediction model is consistent with the physical mechanisms of meteorology. It helps identify complementary strengths between statistical machine learning methods and methods based on physical mechanisms. This ultimately enables more efficient model fusion or ensembling of these two paradigms.

In summary, interpretable machine learning is key for the adoption of ML in meteorology by increasing user trust, helping model improvement, extracting new findings, and enabling the integration of machine learning models with physical mechanism. Tailored interpretability methodologies are needed to provide meaningful explanations that meet the needs of domain experts.

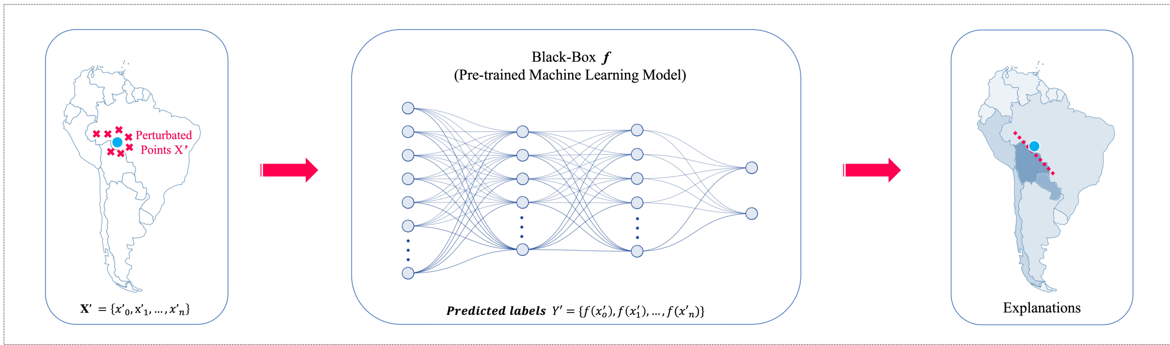
3.2 Explainability Definition and Taxonomy

Definition: Given the dataset D comprising pairs of input values X (e.g., relative humidity, wind speed, cloud amount, etc.) and the corresponding target outcomes Y (e.g., horizontal visibility), the model $f : \mathcal{X} \rightarrow \mathcal{Y}$ is used to predict outcomes based on input data $\mathcal{X} \in X$. For a specific test sample $x_i \in D$, the explanation to the model prediction $f(x_i)$ is represented as $g(f(x_i))$. It aims to clarify the reasoning behind the predictions made by the model.

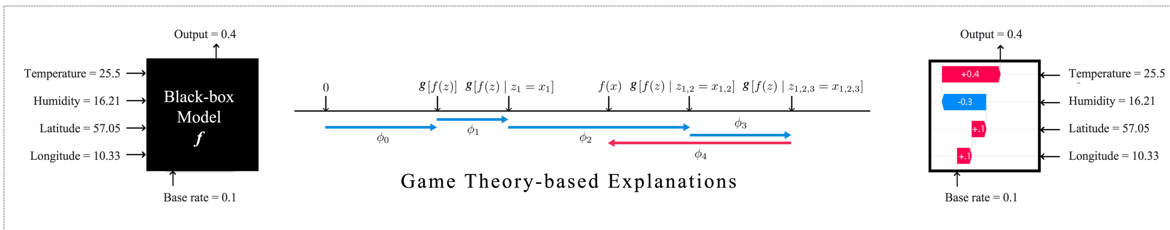
Taxonomy: Based on the above definition, we have the following taxonomy of explainability.

- Regarding explanation contents, they are called local explanations when explaining a specific test data point (x_i, y_i), and global explanations when referring to the entire dataset D.
- Regarding explanation method designs, explanations are categorized as model-specific and model-agnostic. Model-specific methods are especially designed for specific model(s), while model-agnostic methods can be applied to any machine learning model (Ribeiro et al., 2016a).
- Regarding explanation generation stages, explanations can be either post-hoc or intrinsic. Post-hoc explanations occur after f is trained, while intrinsic explanations imply that f is self-interpretable, explaining its predictions (Kakkad et al., 2023).

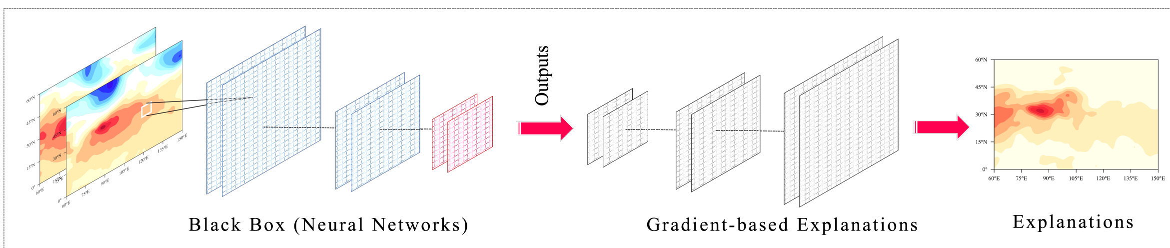
In the following sections, we categorize the explainability analysis techniques into two major types: post-hoc explanation methods presented in Section 4 and the inherently interpretable model design covered in



(a) Perturbation-based methods.



(b) Game Theory-based methods.



(c) Gradient-based methods.

Figure 2: Three major families of post-hoc explanation methods in weather and climate prediction: (a) Perturbation-based explanation, (b) Game Theory-based explanation (Referred & modified from SHAP (Lundberg & Lee, 2017)), (c) Gradient-based explanation.

Section 5. The former explains black-box models after they have already been trained, while the latter involves building interpretable prediction models.

4 Post-hoc Explanation

Post-hoc explanation techniques aim to interpret the predictions of machine learning models after they have already been trained, without changing the underlying models themselves (Deng et al., 2024; Fu et al., 2021; Wang et al., 2020). In the following, we categorize the local explanation methods into three main categories: perturbation-based methods, game theory based methods, gradient-based methods (see Fig. 2). The prediction tasks of these three categories of methods can be found in Table 2. Besides, we also briefly introduce explainability methods that cannot be categorized into these three categories. We summarize how these post-hoc attribution methods can be used to explain weather and climate ML predictions.

4.1 Perturbation-based Forward Propagation Methods

Perturbation techniques systematically alter input features to quantify the effect on predictions (Fig. 2a). For example, methods like Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016b) randomly generate perturbed inputs. The prediction difference when removing or perturbing a feature

Table 2: An overview of three categories of post-hoc explanation methods: Perturbation-based, Game Theory-based and Gradient-based methods.

Methods	Applcations	Prediction Tasks
	Perturbation Based Post-hoc Explanation	
Local Interpretable Model-agnostic Explanations (LIME)	(Valdés & Pou, 2021)	Water vapor patterns prediction
	(Gibson et al., 2021)	Seasonal precipitation prediction
	(Rajasekaran et al., 2023)	Irradiance, temperature and wind speed prediction
Permutation Feature Importance (PFI)	(Rasp & Lerch, 2018)	Ensemble weather forecasts
	(Molina et al., 2021)	Severe convective storms classification
	(Ghada et al., 2022)	Stratiform and convective rain classification
	(Shield & Houston, 2022)	Supercell thunderstorms prediction
	Game Theory Based Post-hoc Explanation	
Shapley Value	(Thanh Trieu et al., 2021)	Horizontal visibility predictions
	(Leinonen et al., 2023)	Thunderstorm nowcasting
	(Wang & Li, 2023)	Tropical cyclone wind radii prediction
SHapley Additive exPlanations (SHAP)	(Lu et al., 2021)	Heavy precipitation prediction
	(Gensini et al., 2021)	Severe weather events prediction
	(Dutta & Pal, 2022)	Short-term premonsoon thunderstorms prediction
	(Griffin et al., 2022)	Rapid intensification of tropical cyclones prediction
	(Silva et al., 2022)	Earth system model errors prediction
	Gradient Based Post-hoc Explanation	
Gradient-weighted Class-Activation Mapping (Grad-CAM)	(Higa et al., 2021)	Typhoon intensity classification
	(Rampal et al., 2022)	Daily rainfall prediction
	(Renault & Mehrkanoon, 2023)	Precipitation nowcasting
	(Reulen & Mehrkanoon, 2024)	Extreme precipitation nowcasting
Integrated Gradients (IG)	(Espeholt et al., 2022)	Precipitation prediction
	(González-Abad et al., 2023)	Daily temperature downscaling
	(Hu et al., 2023)	Probabilistic precipitation forecast
Layer-wise Relevance Propagation (LRP)	(Hilburn et al., 2020)	Synthetic radar reflectivity estimation
	(Barnes et al., 2020)	Annual-mean temperature and precipitation prediction
	(Toms et al., 2020)	Surface temperature anomalies prediction and ENSO phase identification
	(Toms et al., 2021a)	Surface temperature anomalies prediction
	(Toms et al., 2021b)	Madden-Julian Oscillation phase identification
	(Sonnewald & Lguensat, 2021)	Tracking global heating with ocean regimes
	(Zhuo & Tan, 2021)	Tropical cyclone intensity and size estimation
	(Retsch et al., 2022)	Convective area and organization prediction
	(Legler & Janjić, 2022)	Convective-scale model parameters estimation
	(Li et al., 2023c)	Probabilistic convective initiation nowcasting
	(Liu et al., 2023)	Short-term station precipitation prediction

highlights its importance. This type of model can be used to explain the predictions of any machine learning model. In the following, we focus on two representative perturbation-based methods, including LIME and permutation feature importance.

4.1.1 LIME Explanation

The LIME method mainly performs small perturbations on target input features within the local linear neighborhood and then uses a simple linear model to visualize important features to achieve the purpose of explanation (Ribeiro et al., 2016b). The method is easy to understand and can be applied to various machine learning models. For example, one study employs the LIME method to reflect the behavior of the random forest classifier in the water vapor patterns detected by meteorological satellites (Valdés & Pou, 2021). In particular, in 2013 and 2015, the importance of certain masks used in the model changes significantly, with masks 11 and 27 being the most important in both years, but with their ranks reversed in 2015. Additionally, some masks disappear from the top 15 subset when moving from 2013 to 2015, and others not previously in the top 15 appear in 2015, indicating significant temporal variations in the contributing factors to the model's predictions. Another study applies LIME to interpret the irradiance, temperature and wind speed predictions made by SRNN-LSTM hybrid models (Rajasekaran et al., 2023). The research highlights the significance of meteorological feature correlations and time steps towards model outputs, which are analyzed using LIME frameworks. It is concluded that the explanations generated by LIME comply with the theoretical constructs of the features involved in the model. Since this method works on a single prediction and can only be used for local explanation, it is often used in conjunction with other explainability methods to improve a more comprehensive explanation. For example, the LIME method, together with other explanation methods like partial dependence plots, is used to interpret seasonal precipitation prediction made by a random forest model for the western United States (Gibson et al., 2021). Specifically, the LIME method focuses on explaining the model's decision-making process for individual seasonal forecasts, and the latter are used to make global interpretation. Using LIME, the study analyzes two case studies: one where the model correctly predicts the seasonal cluster in 2005, and another where it incorrectly predicts the seasonal cluster in 2016. In the 2016 case, weak El Niño conditions largely contribute to this incorrect prediction. This analysis demonstrates the capacity of LIME to provide insights into the specific factors that influence individual weather forecasts, thus enhancing the interpretability of complex weather prediction models.

In this survey, we showcase the performance of LIME by using it to interpret and visualize an MLP-based temperature prediction task. The importance of input features in two cases is shown in Fig. 3a and 3b.

4.1.2 Permutation Feature Importance

The Permutation Feature Importance (PFI) method is first proposed to measure the importance of input variables in random forests (Breiman, 2001). The input variables are ranked based on the influence of random perturbations on prediction errors, with larger difference in skill indicating greater importance (Gagne II et al., 2019). In one study, the PFI is applied to determine the relative importance of different features in ensemble weather forecasts (Rasp & Lerch, 2018). This method involves randomly shuffling each predictor or feature in the validation set one at a time and observing the increase in mean Continuous Ranked Probability Score (CRPS) compared to the unpermuted features. The results of this interpretation method show that it is necessary to extract local information to improve the accuracy of probabilistic temperature forecast. Another study employs the PFI method to analyze the impact of specific meteorological variables on the performance of a Convolutional Neural Network (CNN) in classifying thunderstorms (Molina et al., 2021). This process involves 500 permutations for each of the 20 variables, allowing for a comprehensive assessment of variable importance in the context of thunderstorm classification. In addition, PFI is utilized to rank input features based on their importance for the performance of machine learning models in classifying rain types (Ghada et al., 2022). This approach permutes a specific feature to disrupt its association with the target variable, followed by predictions using the modified dataset. The resulting change in model performance, particularly in the Area Under the Curve (AUC), indicates the significance of the permuted feature for the model's predictive performance. The reflectivity at the lowest layer and the average pectral width in the layers below separation level are most influential in the model's predictions for classifying rain types.

This PFI approach mentioned above helps to identify which features significantly impact the model's performance without the need to re-estimate the model for each feature omitted. However, it is important to note that this method does not capture colinearities between features. To fully account for the correlation between input variables, the multipass permutation variable importance method (Lakshmanan et al., 2015) is developed to interpret machine learning models for forecasting supercell thunderstorms (Shield & Houston, 2022). This method involves initially assessing the model's performance on a test dataset, then permuting a single input variable and re-evaluating the model's performance with the altered data. The variable whose permutation results in the most significant decrease in performance is considered the most important. This procedure is run repeatedly, keeping previously ranked variables permuted, until all variables are ranked. This approach, while more computationally intensive, provides a more precise understanding of the variables' impact on the model, especially in the presence of correlations between variables.

4.2 Game Theory based Methods

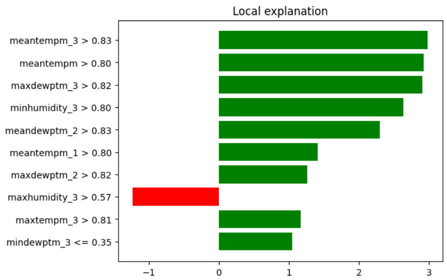
Game theory-based explanation methods like Shapley values and SHapley Additive exPlanations (SHAP) explain a model's predictions by attributing impact to input features (Fig. 2b). They do this by comparing a prediction to what it would be without each feature. These techniques can generate consistent, model-agnostic explanations based on game theory. However, they scale poorly computationally for complex models with many features. In the following, we introduce how both methods can be used to explain the meteorological predictions made by ML models.

4.2.1 Shapley Value Explanation

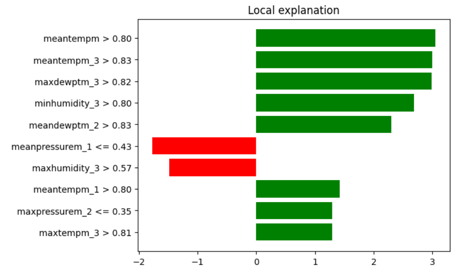
The Shapley values method from cooperative game theory quantifies the contribution of each feature to a model's output by comparing what the prediction would be with and without that feature. This method calculates the marginal contribution of a feature by iterating through all possible combinations, or coalitions, of features. For instance, the Shapley value method is utilized for local interpretation of specific weather predictions (Thanh Trieu et al., 2021). This method assesses the contribution of each meteorological feature to the difference between the actual prediction and the average prediction. In a case study of horizontal visibility in Kemi, Finland, air temperature is found to have the most positive contribution, while cloud amount had the most negative impact. Besides, the Shapley value method is used to assess the importance of each data source in predicting thunderstorm hazards, enhancing prediction explainability (Leinonen et al., 2023). This approach calculates the contribution of each predictor to improving a performance metric, employing game theory for fair interpretation. The Shapley values represent the weighted average of improvements achieved by adding a predictor to a set that previously did not contain it. This method requires computing the metric for each subset of predictors, including the empty set, which is feasible in this study due to the limited number of data sources. The sum of all Shapley values is equal to the total improvement provided by all predictors. The results demonstrate that the weather radar data play the most important role in prediction of the three types of disasters: lightning, hail and heavy precipitation. Furthermore, the Shapley value method is employed to interpret deep learning predictions of tropical cyclone wind radii (Wang & Li, 2023). This approach is used to generate heat maps, highlighting features with higher relevance to the tropical cyclone wind radius predictions. Larger values in these maps indicate features of greater significance to the prediction outcomes, providing insight into the individual contributions of the variables in the model.

4.2.2 SHAP Explanation

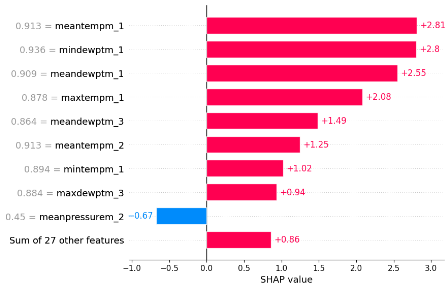
Although Shapley value interpretation can consider all feature coalitions comprehensively, directly computing Shapley values for all 2^n feature subsets is computationally expensive for high dimensional models. SHAP is built on this framework to efficiently estimate Shapley values through conditional expectations (Lundberg & Lee, 2017). This enables model explanations to scale to higher dimensions by approximating Shapley values, simplifying assumptions to enable faster computation. Due to the good trade-off between accuracy and feasibility, the method is widely used in weather and climate predictions. For example, the SHAP method is used to interpret the contributions of meteorological features in a deep learning model for forecasting heavy precipitation (Lu et al., 2021). For visual analysis, the SHAP values of the 20 most important features such as thermodynamic and dynamic parameters are analyzed using 500 minority samples. It offers insights into both the relative importance of convection parameters and their positive or negative contributions to the heavy precipitation predictions. Another study utilizes SHAP to interpret the output of machine learning models predicting severe weather events (Gensini et al., 2021). This method allows for the identification of thresholds where each predictor variable begins to notably impact the probabilistic contribution to the forecast outcome in severe or significant-severe weather classification, providing a deeper



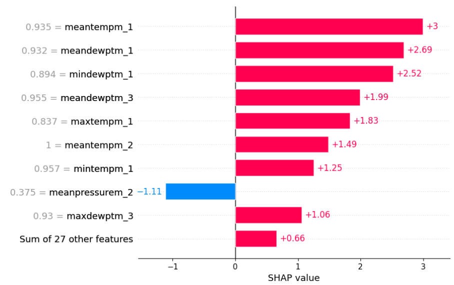
(a) LIME output for 2017/7/5



(b) LIME output for 2017/7/26



(c) SHAP output for 2017/7/5



(d) SHAP output for 2017/7/26

Figure 3: The feature attribution results of different input features in MLP prediction models using LIME and SHAP interpretation methods. The MLP model is used to forecast the weather temperature task. The data comes from the weather data of Lincoln, Nebraska, which is provided by Weather Undergroun and covers 997 days since January 4, 2015 (Here, MLP model and data are both from `https://github.com/Rite188/Simple-DNN-on-weather-forcast`).

understanding of the model's predictions for tornadoes and hail. Additionally, SHAP is applied to quantify the contributions of different predictor variables in machine learning models for short-term predictions of pre-monsoon thunderstorms (Dutta & Pal, 2022). The study uses SHAP to determine the relevance of different predictor variables, thereby interpreting the models' decisions toward thunderstorm prediction and validating the models with domain knowledge. Besides, the SHAP method is utilized to assess the impact of individual features on the probability of rapid intensification in tropical cyclones (Griffin et al., 2022). The sum of SHAP values for all input features is found to be slightly lower than the rapid intensification probability, a difference that grows with increasing rapid intensification probability. This methodology offers a clear and interpretable understanding of how individual features influence the model's performance in predicting rapid intensification in tropical cyclones. Lastly, SHAP values are used to analyze the contribution of input variables to earth system model errors predictions made by XGBoost model (Silva et al., 2022). Specifically, the SHAP framework evaluates the sum of contributions from each input feature and the average predicted value to understand their impact on lightning predictions errors. The result demonstrates the errors in lightning prediction are highly correlated with the effects of convective processes and surface heterogeneity, highlighting the potential of SHAP in characterizing and exploring errors in Earth system models.

In this survey, we also showcase the explanation performance of SHAP by providing the results of two cases using it to explain the importance of input variables, as shown in Fig. 3c and 3d.

4.3 Gradient-based Backpropagation Methods

Gradient-based feature attribution explains a prediction by analyzing input feature derivatives with model output (Fig. 2c). Saliency methods like Gradient-weighted class-activation mapping (Grad-CAM) (Selvaraju et al., 2017), Integrated Gradients (IG) (Sundararajan et al., 2017) and Layer-wise Relevance Propagation (LRP) determine influential input variables by computing gradients. Some representative gradient-based interpretation results are shown in Fig. 4. In the following, we introduce how these gradient-based attribution methods can be used to explain weather and climate predictions made by machine learning models.

4.3.1 Grad-CAM Explanation

Grad-CAM obtains the weight of each channel by calculating the gradient to identify the areas that influence the decision-making basis. We can use this method to visualize any layer without changing the model structure. Current research mainly applies this method to various CNN-based deep learning models. For example, one recent study implements Grad-CAM to identify spatial locations in input fields that significantly support daily rainfall predictions made by a CNN for a specific output location (Rampal et al., 2022). Grad-CAM does not specify which individual predictor variable had the strongest influence, but instead shows the importance of spatial locations in the aggregated predictor space. This approach aids in making the model's predictions more transparent and trustworthy, thereby potentially leading to new insights in complex systems. Another study discusses the integration of meteorological domain knowledge into deep learning models for improving typhoon intensity classification from satellite images (Higa et al., 2021). Specifically, it highlights the use of Grad-CAM to visualize which areas of the satellite images are most important for predicting typhoon intensity. By preprocessing images with fisheye distortion to emphasize the typhoon's eye and surrounding cloud distributions, the model achieved higher classification accuracy. Grad-CAM visualizations confirmed that the model focuses on meteorologically significant regions for intensity classification, aligning with expert domain knowledge.

Grad-CAM can also be utilized for global analysis via analyzing the importance and interaction of different layers in the machine learning model. For example, one study proposes a novel UNet-based architecture for precipitation and cloud cover nowcasting tasks (Renault & Mehrkanoon, 2023). To provide comprehensive explanations, they apply Grad-CAM to visualize activation heatmaps at different layers in the encoder and decoder. Analyzing these heatmaps reveals how the residual connections and depthwise separable convolutions interact, with the two paths switching importance between encoder and decoder. The heatmaps also show how the model progresses from detecting borders to focusing on cloud centers. For example, deeper network levels show activations in more abstract areas, with significant activations in select spots. Overall, using Grad-CAM on multiple layers provides global and layered explanations that increase understanding of the model's precipitation nowcasts. Furthermore, the same multiple layer analysis explanation framework has been applied to explain predictions made by a novel generative adversarial network designed for extreme precipitation nowcasting task (Reulen & Mehrkanoon, 2024). There are some interesting findings. For example, the precipitation map encoder shows higher activation in areas of high precipitation at shallow depths, while the precipitation mask encoder activation maps correlate with input precipitation at first and predicted precipitation at the next depth. At deeper levels, the activation areas become more abstract for both encoders. Overall, the encoder activation heatmaps become less directly linked to precipitation levels as depth increases.

4.3.2 Integrated Gradients Explanation

In addition to Grad-CAM, IG is another representative gradient-based explainability technique. It measures the impact of input changes on model predictions in deep learning, particularly in complex models such as U-Net. This technique is known for its ability to overcome issues like gradient saturation, common in standard gradient-based methods (Sundararajan et al., 2017). By integrating the product of input values and gradients, IG assesses the sensitivity of outputs to inputs, using a blurred input as a baseline for comparison. Since this method does not alter the model's architecture and is effective for visualizing how input perturbations affect forecasts, it is widely used in weather forecast. For example, IG has been used to understand MetNet-2 learning process, a physically independent probabilistic weather model based on deep neural networks, shedding light on the interaction between various meteorological variables in precipitation forecast (Espeholt et al., 2022). IG attributes the network's precipitation predictions to specific input variables, revealing that while the absolute vorticity's influence is minimal for near-term forecasts, it becomes more significant for predictions up to 12 hours ahead. This aligns with the geostrophic theory, where positive vorticity at higher altitudes correlates with upward motion lower in the atmosphere, setting the stage for potential convection, a precursor to precipitation. The IG-based analysis in another study

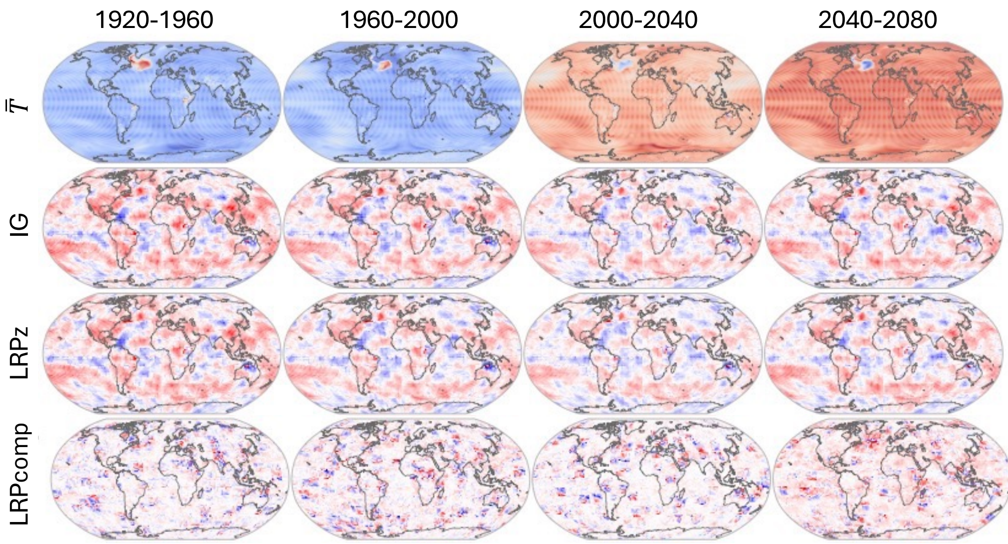


Figure 4: Visualization comparison between three explainability methods. Here LRP$_{comp}$ and LRP$_z$ use two different rules of backpropagation. The figure is adapted from Bommer et al. (Bommer et al., 2023).

indicates that precipitation intensity and uncertainty are highly responsive to input precipitation changes, with integrated vapor transport and integrated water vapor also identified as significant factors in U-net models for precipitation prediction (Hu et al., 2023). For climate prediction, one study uses IG to integrate gradients along a path from the input to a baseline, offering a more accurate representation of feature importance in climate-related CNN applications (González-Abad et al., 2023). They quantify the influence of input variables on a temperature downscaling, helping to identify critical predictor variables and their spatial regions of influence.

4.3.3 Layer-by-layer Backpropagation

Deep neural networks are often extremely complex, while each layer of the neural network is relatively simple (e.g., deep features are usually a linear summation of shallow features and nonlinear activation function). This facilitates the analysis of the importance of shallow features for deep features. Therefore, this type of algorithm estimates the importance of intermediate features and propagates these importance values layer by layer to the input layer, to determine the importance of the input units. This algorithm includes Layer-wise Relevance Propagation (LRP) (Bach et al., 2015; Ebert-Uphoff & Hilburn, 2020), deep Taylor decomposition (Montavon et al., 2017), etc. The fundamental difference between these backpropagation algorithms lies in the different rules they adopt for layer-by-layer propagation of importance. In the following, we will introduce the application of LRP in weather forecast and climate predictions.

For weather forecast, LRP is often used to account for predictors that have a significant impact on deep learning models. For example, one study employs LRP to investigate the interaction between deep convection in the tropics and the large-scale atmosphere when using Multi-Layer Perceptrons (MLP)[2] to predict convective area and organization (Retsch et al., 2022). It finds that large-scale vertical velocity significantly impacts both convective area and organization, with its influence more pronounced in predicting convective area. While thermodynamic factors like atmospheric moisture affect convective area predictions, they are less important for convective organization, where horizontal wind fields play a more significant role. Another article addresses the challenge of estimating parameters not well-defined physically in convection-permitting numerical weather prediction models, particularly for cloud representation (Legler & Janjić, 2022). It uses Bayesian Neural Networks (BNNs) and Bayesian approximations of point estimate MLP, to predict several

[2]Due to historical reasons, multilayer perceptrons (MLPs) are also referred to as artificial neural networks (ANNs) or feedforward neural networks (FNNs) in the literature. For consistency, this paper uses the term MLP.

parameters of a modified shallow-water model based on atmospheric state observations or analysis. The study utilizes LRP to understand how these neural networks learn, revealing that the networks selectively focus on a few grid points characterized by strong winds and rain for making parameter predictions. Additionally, one article applies the LRP method to a Station-based Precipitation Post-processing Model (SPPM) designed to enhance the accuracy of medium-range station precipitation forecast with deep-learning algorithms (Liu et al., 2023). Specifically, the LRP method is used to assess the sensitivity of the predictors, revealing that total precipitation from the NWP is the most crucial and sensitive factor, particularly for larger forecast grades. The importance of low-level (850 hPa) field, single-level field, and geographic variables is also examined, which is in agreement with the meteorological domain knowledge.

Not only can we use LRP to estimate the importance of predictors, we can also use it to show where the neural network focuses primarily when making predictive decisions. To visualize the underlying logic of a CNN-based deep learning model decisions when assimilating GOES-R series observations in precipitating scenes, a study employs LRP as an attribution method, demonstrating the CNN's synergistic use of radiance and lightning information (Hilburn et al., 2020). Lightning data particularly help in identifying important neighboring locations. The study investigates the sensitivity to radiance gradients, indicating that sharper gradients elicit stronger responses in predicted radar reflectivity, and highlights the unique value of lightning observations in pinpointing locations of strong radar echoes. This research also explored the application of LRP to regression tasks. Besides, DeepTCNet, a CNN-based model for estimating tropical cyclone intensity and wind radii, uses LRP to interpret its decision-making process (Zhuo & Tan, 2021). LRP attributes which input features most affect the model's output, overcoming the limitations of saliency maps by considering the redistribution of prediction values across input features, thereby providing a more coherent explanation for the model's predictions. This interpretative approach is crucial in enhancing trust in DeepTCNet's forecasts for tropical cyclone characteristics and has the potential to explore more complex tropical cyclone mechanisms. In addition, LRP can also be applied to more complex deep learning models. The CIUnet model, leveraging U-net architecture and Himawari-8 data, effectively forecasts convective initiation with a high detection rate and low false alarms (Li et al., 2023c). LRP analysis validates the model's accuracy in pinpointing crucial regions and features for precise convective initiation predictions. Key input factors include brightness temperature differences between spectral channels and terrain height.

In addition to weather forecast, there is also some research focus on the application of LRP in climate prediction. In particular, a study explores the use of LRP in MLP for seasonal surface temperature anomalies prediction and El Niño Southern Oscillation (ENSO) phase identification (Toms et al., 2020). LRP helps explain neural network decisions by identifying the relevance of each input feature for the network's output on a sample-by-sample basis, effectively creating a heatmap of relevance. Specifically, LRP is applied to study El Niño events, with the relevance values from LRP for each sample normalized to a range of 0 to 1. This normalization ensures equal weighting of relevances across samples when composing the overall relevance heatmap. The study demonstrates that LRP can trace the reasoning of a neural network's decisions, highlighting its potential to reveal meaningful geoscientific insights from neural network analysis. Furthermore, the LRP method is also applied to identify and interpret the spatial patterns enabling the MLP decision-making process in surface temperature anomalies predictions on decadal timescales (Toms et al., 2021a), Madden-Julian Oscillation phase identification (Toms et al., 2021b), and annual-mean temperature and precipitation predictions (Barnes et al., 2020). Additionally, LRP is employed to elucidate the predictive skills of ensemble MLP for tracking global heating with ocean regimes (Sonnewald & Lguensat, 2021). The application of LRP provides a post-hoc assessment of how ensemble MLP adjusts at each location in the North Atlantic region. This approach allows for a nuanced understanding of the contributions of various features like wind stress curl, latitudinal and longitudinal gradients, to the neural network's decision-making process, thereby enhancing confidence in the model's predictions and its application to unseen models or under different climate forcing scenarios.

4.4 Other Explanation Methods

Beyond the aforementioned three major categories of explanation techniques, there also exist some other post-hoc explanation methods such as activation maximization. Activation maximization generates archetypal model inputs that highly activate certain layers like hidden neurons. This probes what features models

Table 3: An overview of intrinsic self-explainable models in weather and climate prediction.

Self-explainable Models	Applcations	Prediction Tasks
Linear Models	(Herman & Schumacher, 2018)	Extreme weather forecast
Tree-based Models	(Mecikalski et al., 2015)	Convective initiation prediction
	(Loken et al., 2022)	Severe weather hazards prediction
	(Zhang et al., 2019)	Tropical cyclone genesis prediction
Attention-based Explainable Neural Networks	(Tekin et al., 2021)	High-resolution temperature forecasting
	(Suleman & Shridevi, 2022)	Short-term temperature forecasting

have learned to detect. For example, activation maximization is used as a visualization technique to identify patterns that maximize specific activation functions in deep learning model for wind speed forecast (Abdellaoui & Mehrkanoon, 2020). This post-hoc method focuses on finding new input data that maximizes the activation of a neuron, aiming to understand the model after training. The objective is to find input data that contribute the most to minimizing the error between the wind speed prediction and ground-truth data. For this purpose, the study defines a custom objective function as the inverse of the mean squared error, focusing on weather element forecasting as a regression problem. The interpreted results help further understand the most important features of weather forecasting in target cities.

Some other explainability tools have also been used to provide explanations for predictions made by machine learning models. For example, a study introduces a multi-variate wind speed forecasting model that leverages machine learning and a clustering-based multi-objective gravity search algorithm for improved accuracy and reliability (Li et al., 2023a). Explainability in wind speed forecasting is implemented by employing post-hoc attribution analysis and visualization tools like partial dependence plots and individual conditional expectation plots. These tools help in understanding the model's interpretability, assessing the robustness and reliability of forecasted results, and explaining the relationship between features and forecast outcomes.

5 Design Intrinsic Self-Explainable Model

In this section, we introduce techniques to discuss more inherently interpretable models, with the aim of making the logic behind the predictions more transparent while still maintaining high accuracy. This is achieved by designing model architectures and components that are intuitive for humans to understand. However, perfectly interpretable models typically sacrifice some prediction performance. Thus, there is a trade-off between accuracy and interpretability that depends on the specific application. In the following, we summarize some of the key techniques to enhance inherent model transparency, including linear models, tree-based models, and attention mechanisms. We have also summarized these techniques in Table 3.

5.1 Linear Models

Linear models such as linear regression and logistic regression (LR) have coefficients that directly indicate the strength of association between the input predictors and the prediction. This inherent linear additivity makes the prediction explanation clear. However, linearity is often a simplifying assumption that can significantly reduce prediction accuracy. In general, linear models trade some precision for interpretability. To solve this, linear models can be combined with other machine learning models to obtain a forecast result with ideal accuracy and interpretability. For example, one study investigates the application of LR to extreme weather forecast, to understand how it makes predictions (Herman & Schumacher, 2018). The work performs a principal component analysis (PCA) before performing the prediction task, and the extracted principal components are provided to the LR. The authors show how visualizing regression coefficients for LR provides insight into what dynamics and variables are the most predictive. For example, the models automatically learn that model precipitation forecasts are most predictive along the Pacific coast where large-scale dynamics dominate extreme events, while moisture and instability become more important in the central US where convection is key factor. Analyzing these models builds understanding of the captured relationships, reveals systematic model biases like displacing precipitation features, and helps guide meteorological forecasters in using and correcting the guidance.

5.2 Tree Based Models

Decision tree and ensemble of trees are another important family of inherently interpretable models. Decision trees partition the data space into rectilinear regions, with splits chosen to maximize information gain. The tree structure shows how predictions are made based on input variables that meet certain conditions. Ensembles of decision trees, such as random forests, improve accuracy while retaining some model transparency through tools like variable importance scores. For example, one study explores the use of random forests method for improving convective initiation predictions (Mecikalski et al., 2015). It discusses the methodology for calculating feature importance in random forests. This process involves replacing each predictor variable with a randomized resampling and measuring the impact of this replacement on the random forests trees' prediction accuracy. Variables that significantly degrade accuracy when randomized are deemed the most important. The most critical findings reveal that two measures each of Convective Inhibition (CIN) and Convective Available Potential Energy (CAPE) are the most important predictors. Another study focuses on the use of random forests for predicting severe weather hazards (Loken et al., 2022). It compares two methods of creating random forests based prediction for next-day severe weather using simulated data from the High Resolution Ensemble Forecast (HREF) system. In both models, storm variables are identified as the most crucial, followed by index and environment variables. Additionally, they use a Python module tree interpreter to evaluate variable importance and the relationships acquired by the random forests, and find that the model focuses on different features when predicting different hazards in a physically meaningful way. In addition to random forests, other tree-based machine learning models such as AdaBoost are also used in weather forecast, like predicting the evolution of Mesoscale Convective Systems (MCS) into tropical cyclones (Zhang et al., 2019). The AdaBoost classifier is built on environmental predictors and MCS properties known to influence tropical cyclones genesis. The model is used to quantify the contribution of each critical predictor to these classifiers' performance, contributing to the uncovering of new aspects of the tropical cyclones genesis.

5.3 Attention-based Interpretable Models

Beyond linear and tree-based models, DNNs via attention mechanism is another important inherently interpretable model family. Some work explores making complex neural networks more interpretable by incorporating transparent model components such as attention layers. Adding interpretability modules increases understandability without sacrificing too much prediction performance. For example, one study discusses an innovative deep learning architecture for high-resolution numerical weather forecasting (Tekin et al., 2021). This architecture utilizes Convolutional Long Short-Term Memory (ConvLSTM) and CNN units, integrated with an encoder-decoder structure. The model's interpretability and performance are enhanced by integrating the attention and context matching mechanism. The experiments conducted on the ERA5 hourly dataset demonstrate significant improvements in capturing spatial and temporal correlations for temperature forecasting, outperforming baseline models like ConvLSTM and U-Net. Another study presents a novel deep learning model, Spatial Feature Attention Long Short-Term Memory (SFA-LSTM), designed for accurate temperature forecasting (Suleman & Shridevi, 2022). Using an encoder-decoder architecture with LSTM layers, this model efficiently captures both spatial and temporal relationships among various meteorological variables when applied to temperature forecast. The spatial feature attention mechanism within the model enhances interpretability, allowing it to accurately forecast temperature changes by understanding the mutual influence of input weather variables. The model's performance, verified through domain knowledge, shows high accuracy and interpretability, especially in predicting temperature changes in weather forecast.

6 Research Challenges

Despite the current research progress in applying explainability to meteorological applications, there are also some research challenges remaining.

6.1 Mechanistic Interpretability

Recent advances in explainability have significantly impacted meteorological research, offering new insights and predictive capabilities. However, a critical limitation in the current explainability research is the heavy reliance on feature attributions. These methods, while useful in identifying which features in the data are most influential in the model's predictions, often fall short in providing deeper understanding of the underlying mechanisms. This superficial level of interpretation can be especially problematic in meteorology, where the causal relationships and dynamic interactions are complex and critical for accurate predictions and understanding. Feature attributions can highlight influential factors but without elucidating how these factors interact and contribute to the overall system dynamics. For instance, a model might identify water vapor and vertical velocity as key features in predicting rainfall, but it does not explain the mechanistic relationship between these variables and how they lead to precipitation. This gap hinders the ability of meteorologists to fully trust and understand AI predictions, which is crucial for high-stakes decision-making and advancing scientific knowledge.

Mechanistic interpretability, on the other hand, aims to address these shortcomings by providing insights into the 'why' and 'how' behind AI predictions (Olah et al., 2020; Conmy et al., 2023; Zhao et al., 2024). It seeks to uncover the underlying causal relationships and principles that drive the model's output, aligning more closely with scientific discovery and reasoning. In meteorology, this means not just identifying important features, but understanding how these features interact in atmospheric and related systems (e.g., ocean, land and biosphere) to produce specific weather events or patterns. The need for mechanistic interpretability in meteorology is two-fold. Firstly, it improves the credibility and trustworthiness of AI models by aligning their functioning with known scientific principles and theories. Secondly, it contributes to scientific discovery by potentially uncovering new insights and relationships within meteorological data that are not apparent through traditional analysis.

6.2 Interpretability Evaluation

Despite advances in developing explanation techniques, evaluating their utility and faithfulness remains an open challenge, largely due to the unavailability of ground-truth explanations. Although some studies have assessed different explainability methods applied to climate science, these researches are mainly based on existing benchmarks and evaluation methods not standardized for the characteristics of climate data (Mamalakis et al., 2022; Bommer et al., 2023). Without an objective standardized measure of the quality of the explanation, it is extremely difficult to rigorously assess different methods and ensure that they provide meaningful insights aligned with reality. Therefore, the development of standardized benchmarks and metrics is critical for systematically evaluating explanation fidelity. Possible solutions involve designing proxy ground truth and pseudo ground-truth for the interpreted results (Li et al., 2023b). This requires carefully designed datasets with known explanatory factors, allowing explanation accuracy to be measured against ground-truth explanations. However, crafting appropriate benchmarks is highly complex for meteorological data, given intricate dynamical interdependencies. Collaboration with domain experts is essential to validate that learned relationships are scientifically sound.

6.3 Making Use of Interpretability

Explainable AI techniques provide valuable insights into the rationale behind the predictions of complex machine learning models. However, the full potential of interpretability lies in utilizing those insights to further improve model performance. The integration of explainability analysis into model development workflows can enable developers to continuously monitor feature attributions and model behaviors. This allows identifying areas where the model violates known constraints or relies on faulty correlations, guiding iterative refinement steps (Fig. 5).

In particular, explainability can be used to analyze whether data-driven deep learning models make predictions that violate established physical principles or constraints. Since data-driven models are not based on encoding physics-based equations, they may learn spurious correlations that produce forecasts conflicting with the laws of atmospheric sciences. By attributing predictions to input features and visualizing the learned

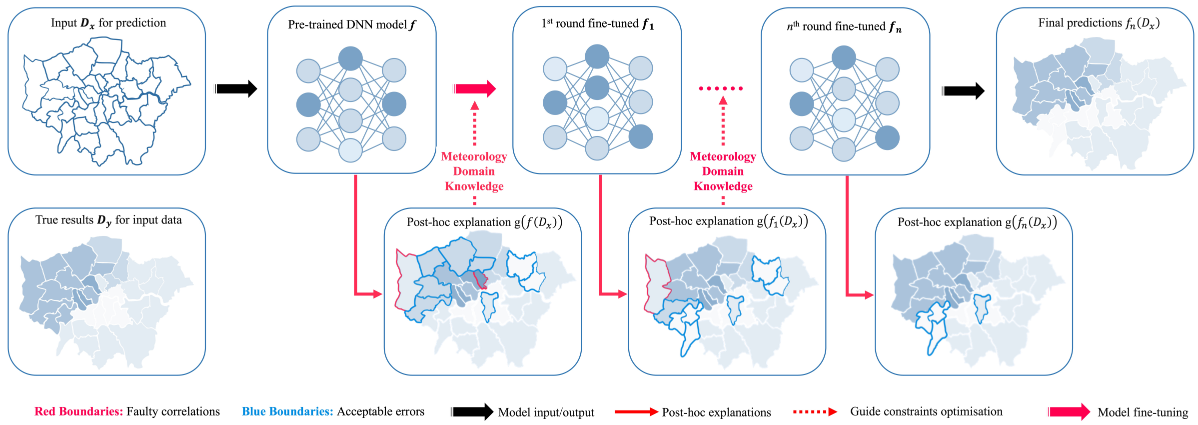


Figure 5: An example of using explainability to guide iterative physical consistency improvement of a machine learning model (DNN model) until no faulty corrections (red boundaries) are detected.

representations, developers can identify unreasonable logic and relationships in the model. Once violations of physical consistency are detected through explanations, constraints can be introduced into model training and architectures. This includes adding loss penalty terms based on physical principles, using physics-infused neural networks, or representing conservation properties. Imposing appropriate physical constraints helps prevent models from making unreal predictions while still leveraging the flexibility of machine learning. This will eventually shift deep learning models closer to a hybrid data-driven and physics-based paradigm.

6.4 Interpretability for Large Foundataion Models

Recently, large foundation models have demonstrated impressive capabilities in various meteorological tasks such as weather forecasting, climate analysis, and understanding atmospheric processes (Chen et al., 2023c). These models like ClimaX (Nguyen et al., 2023), GraphCast (Lam et al., 2022), Fengwu (Chen et al., 2023a), Fuxi (Chen et al., 2023b), OceanGPT (Bi et al., 2023b) leverage massive training datasets and billions of parameters to learn complex patterns and relationships. Nowadays, these large models are increasingly being deployed in weather and climate forecasting and decision support systems. However, their large scale and complexity introduce significant challenges for interpretability. Key challenges include faithfully localizing decision factors across billions of parameters, explaining how heterogeneous data modalities are integrated, and providing human-understandable explanations beyond low-level attributions. Potential solutions involve developing sparse attribution methods to isolate critical components, and using higher-level concept-based explanations based on atmospheric science. Besides, we can integrate causal reasoning techniques for mechanistic interpretations and create interactive interfaces for explanation at different granularities.

7 Conclusions

This survey presents a comprehensive review of interpretable machine learning techniques applied to weather and climate prediction. First, we provide an overview of how machine learning techniques are being applied to weather and climate prediction tasks. Then, we categorize the explainability methods into two main paradigms: post-hoc explanation approaches that interpret pre-trained models, and inherently interpretable model architectures designed from scratch. We analyze representative post-hoc techniques like SHAP, LRP, Grad-CAM, and LIME, highlighting how they uncover novel meteorological relationships captured by machine learning models. We also examine self-explainable model families such as linear models, tree ensembles, and attention-based neural networks that enhance transparency. Lastly, we have also summarized the open challenges around developing customized mechanistic interpretability methods to provide a more in-depth understanding of these methodology prediction models, evaluating explanation utility, leveraging insights obtained from explainability to improve the consistency of ML models with physical mechanisms, and developing effective explanation method for large foundation models.

References

Ismail Alaoui Abdellaoui and Siamak Mehrkanoon. Deep multi-stations weather forecasting: explainable recurrent convolutional neural networks. *arXiv preprint arXiv:2009.11239*, 2020.

Shreya Agrawal, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey. Machine learning for precipitation nowcasting from radar images. *arXiv preprint arXiv:1912.12132*, 2019.

Jeffrey Anderson, Tim Hoar, Kevin Raeder, Hui Liu, Nancy Collins, Ryan Torn, and Avelino Avellano. The data assimilation research testbed: A community facility. *Bulletin of the American Meteorological Society*, 90(9):1283–1296, 2009.

Troy Arcomano, Istvan Szunyogh, Jaideep Pathak, Alexander Wikner, Brian R Hunt, and Edward Ott. A machine learning-based global atmospheric forecast model. *Geophysical Research Letters*, 47(9): e2020GL087776, 2020.

Rossella Arcucci, Jiangcheng Zhu, Shuang Hu, and Yi-Ke Guo. Deep data assimilation: integrating deep learning with data assimilation. *Applied Sciences*, 11(3):1114, 2021.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Georgy Ayzel, Tobias Scheffer, and Maik Heistermann. Rainnet v1. 0: a convolutional neural network for radar-based precipitation nowcasting. *Geoscientific Model Development*, 13(6):2631–2644, 2020.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Elizabeth A Barnes, Benjamin Toms, James W Hurrell, Imme Ebert-Uphoff, Chuck Anderson, and David Anderson. Indicator patterns of forced change learned by an artificial neural network. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002195, 2020.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023a.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. Oceangpt: A large language model for ocean science tasks. *arXiv preprint arXiv:2310.02031*, 2023b.

Bogdan Bochenek and Zbigniew Ustrnul. Machine learning in weather prediction and climate analyses—applications and perspectives. *Atmosphere*, 13(2):180, 2022.

Nicola Bodini, Julie K Lundquist, and Mike Optis. Can machine learning improve the model representation of turbulent kinetic energy dissipation rate in the boundary layer for complex terrain? *Geoscientific Model Development*, 13(9):4271–4285, 2020.

Philine Bommer, Marlene Kretschmer, Anna Hedström, Dilyara Bareeva, and Marina M-C Höhne. Finding the right xai method–a guide for the evaluation and ranking of explainable ai methods in climate science. *arXiv preprint arXiv:2303.00652*, 2023.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023a.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *arXiv preprint arXiv:2306.12873*, 2023b.

Shengchao Chen, Guodong Long, Jing Jiang, Dikai Liu, and Chengqi Zhang. Foundation models for weather and climate data understanding: A comprehensive survey. *arXiv preprint arXiv:2312.03014*, 2023c.

Sandy Chkeir, Aikaterini Anesiadou, Alessandra Mascitelli, and Riccardo Biondi. Nowcasting extreme rain and extreme wind speed with machine learning techniques applied to different input datasets. *Atmospheric Research*, 282:106548, 2023.

Rosangela Saher Cintra and Haroldo F de Campos Velho. Data assimilation by artificial neural networks for an atmospheric general circulation model. *Advanced applications for artificial neural networks*, 265, 2018.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*, 2023.

Christopher Daly. Guidelines for assessing the suitability of spatial climate data sets. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 26(6):707–721, 2006.

Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, and Xia Hu. A unified taylor framework for revisiting attribution methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11462–11469, 2021.

Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, Ziwei Yang, Zheyang Li, and Quanshi Zhang. Unifying fourteen post-hoc attribution methods with taylor interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.

Peijun Du, Xuyu Bai, Kun Tan, Zhaohui Xue, Alim Samat, Junshi Xia, Erzhu Li, Hongjun Su, and Wei Liu. Advances of four machine learning methods for spatial data handling: A review. *Journal of Geovisualization and Spatial Analysis*, 4:1–25, 2020.

Debashree Dutta and Sankar K Pal. Interpretation of black box for short-term predictions of pre-monsoon cumulonimbus cloud events over kolkata. *Journal of Data, Information and Management*, 4(2):167–183, 2022.

Imme Ebert-Uphoff and Kyle Hilburn. Evaluation, tuning and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society*, pp. 1–47, 2020.

Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Rob Carver, Marcin Andrychowicz, Jason Hickey, et al. Deep learning for twelve hour precipitation forecasts. *Nature communications*, 13(1):1–10, 2022.

Loris Foresti, Ioannis V Sideris, Daniele Nerini, Lea Beusch, and Urs Germann. Using a 10-year radar archive for nowcasting precipitation growth and decay: A probabilistic machine learning approach. *Weather and Forecasting*, 34(5):1547–1569, 2019.

Weijie Fu, Meng Wang, Mengnan Du, Ninghao Liu, Shijie Hao, and Xia Hu. Differentiated explanation of deep neural networks with skewed distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2909–2922, 2021.

David John Gagne II, Sue Ellen Haupt, Douglas W Nychka, and Gregory Thompson. Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, 147(8):2827–2845, 2019.

Vittorio A Gensini, Cody Converse, Walker S Ashley, and Mateusz Taszarek. Machine learning classification of significant tornadoes and hail in the united states using era5 proximity soundings. *Weather and Forecasting*, 36(6):2143–2160, 2021.

Wael Ghada, Enric Casellas, Julia Herbinger, Albert Garcia-Benadí, Ludwig Bothmann, Nicole Estrella, Joan Bech, and Annette Menzel. Stratiform and convective rain classification using machine learning models and micro rain radar. *Remote Sensing*, 14(18):4563, 2022.

Peter B Gibson, William E Chapman, Alphan Altinok, Luca Delle Monache, Michael J DeFlorio, and Duane E Waliser. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Communications Earth & Environment*, 2(1):159, 2021.

Filippo Giorgi and Linda O Mearns. Approaches to the simulation of regional climate change: a review. *Reviews of geophysics*, 29(2):191–216, 1991.

Jose González-Abad, Jorge Baño-Medina, and José Manuel Gutiérrez. Using explainability to inform statistical downscaling based on deep learning beyond standard validation approaches. *Journal of Advances in Modeling Earth Systems*, 15(11):e2023MS003641, 2023.

Sarah M Griffin, Anthony Wimmers, and Christopher S Velden. 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.

Shengnan Guo, Youfang Lin, Huaiyu Wan, Xiucheng Li, and Gao Cong. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5415–5428, 2021.

Nils Gustafsson, Tijana Janjić, Christoph Schraff, Daniel Leuenberger, Martin Weissmann, Hendrik Reich, Pierre Brousseau, Thibaut Montmerle, Eric Wattrelot, Antonín Bučánek, et al. Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Quarterly Journal of the Royal Meteorological Society*, 144(713):1218–1256, 2018.

Lei Han, Mingxuan Chen, Kangkai Chen, Haonan Chen, Yanbiao Zhang, Bing Lu, Linye Song, and Rui Qin. A deep learning method for bias correction of ecmwf 24–240 h forecasts. *Advances in Atmospheric Sciences*, 38(9):1444–1459, 2021.

Fabrício P Härter and Haroldo Fraga de Campos Velho. Data assimilation procedure by recurrent neural network. *Engineering Applications of Computational Fluid Mechanics*, 6(2), 2012.

Xinlei He, Yanping Li, Shaomin Liu, Tongren Xu, Fei Chen, Zhenhua Li, Zhe Zhang, Rui Liu, Lisheng Song, Ziwei Xu, et al. Improving predictions of land-atmosphere interactions based on a hybrid data assimilation and machine learning method. *Hydrology and Earth System Sciences Discussions*, 2022:1–33, 2022.

Gregory R Herman and Russ S Schumacher. “dendrology” in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Monthly Weather Review*, 146(6):1785–1812, 2018.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Maiki Higa, Shinya Tanahara, Yoshitaka Adachi, Natsumi Ishiki, Shin Nakama, Hiroyuki Yamada, Kosuke Ito, Asanobu Kitamoto, and Ryota Miyata. Domain knowledge integration into deep learning for typhoon intensity classification. *Scientific reports*, 11(1):12972, 2021.

Kyle A Hilburn, Imme Ebert-Uphoff, and Steven D Miller. 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.

Weiming Hu, Mohammadvaghef Ghazvinian, William E Chapman, Agniv Sengupta, Fred Martin Ralph, and Luca Delle Monache. Deep learning forecast uncertainty for precipitation over the western united states. *Monthly Weather Review*, 151(6):1367–1385, 2023.

Jaykumar Kakkad, Jaspal Jannu, Kartik Sharma, Charu Aggarwal, and Sourav Medya. A survey on explainability of graph neural networks. *arXiv preprint arXiv:2306.01958*, 2023.

Eugenia Kalnay, Masao Kanamitsu, Robert Kistler, William Collins, Dennis Deaven, Lev Gandin, Mark Iredell, Suranjana Saha, Glenn White, John Woollen, et al. The ncep/ncar 40-year reanalysis project. In *Renewable energy*, pp. Vol1_146–Vol1_194. Routledge, 2018.

Karthik Kashinath, M Mustafa, Adrian Albert, JL Wu, C Jiang, Soheil Esmaeilzadeh, Kamyar Azizzadenesheli, R Wang, A Chattopadhyay, A Singh, et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194):20200093, 2021.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, James Lottes, Stephan Rasp, Peter Düben, Milan Klöwer, et al. Neural general circulation models. *arXiv preprint arXiv:2311.07222*, 2023.

Zachary M Labe, Nathaniel Johnson, and Thomas L Delworth. Changes in united states summer temperatures revealed by explainable neural networks. *Authorea Preprints*, 2023.

Valliappa Lakshmanan, Christopher Karstens, John Krause, Kim Elmore, Alexander Ryzhkov, and Samantha Berkseth. Which polarimetric variables are important for weather/no-weather discrimination? *Journal of Atmospheric and Oceanic Technology*, 32(6):1209–1223, 2015.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Seung Hoon Lee, Yeon Ah Yoon, Jin Hyeong Jung, Tai-Woo Chang, Yong Soo Kim, et al. A machine learning model for predicting silica concentrations through time series analysis of mining data. *Journal of Korean Society for Quality Management*, 48(3):511–520, 2020.

Stefanie Legler and Tijana Janjić. Combining data assimilation and machine learning to estimate parameters of a convective-scale model. *Quarterly Journal of the Royal Meteorological Society*, 148(743):860–874, 2022.

Jussi Leinonen, Daniele Nerini, and Alexis Berne. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223, 2020.

Jussi Leinonen, Ulrich Hamann, Ioannis V Sideris, and Urs Germann. Thunderstorm nowcasting with deep learning: A multi-hazard data fusion model. *Geophysical Research Letters*, 50(8):e2022GL101626, 2023.

Min Li, Yi Yang, Zhaoshuang He, Xinbo Guo, Ruisheng Zhang, and Bingqing Huang. A wind speed forecasting model based on multi-objective algorithm and interpretability learning. *Energy*, 269:126778, 2023a.

Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. M4: A unified xai benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023b.

Yang Li, Yubao Liu, Yueqin Shi, Baojun Chen, Fanhui Zeng, Zhaoyang Huo, and Hang Fan. Probabilistic convective initiation nowcasting using himawari-8 ahi with explainable deep learning models. *Monthly Weather Review*, 2023c.

Qi Liu, Xiao Lou, Zhongwei Yan, Yajie Qi, Yuchao Jin, Shuang Yu, Xiaoliang Yang, Deming Zhao, and Jiangjiang Xia. Deep-learning post-processing of short-term station precipitation based on nwp forecasts. *Atmospheric Research*, 295:107032, 2023.

Eric D Loken, Adam J Clark, and Amy McGovern. Comparing and interpreting differently designed random forests for next-day severe weather hazard prediction. *Weather and Forecasting*, 37(6):871–899, 2022.

Zhiying Lu, Xudong Ding, Qin Yan, and Jianlin Guo. Regional forecast of heavy precipitation and interpretability based on td-vae. In *2021 40th Chinese Control Conference (CCC)*, pp. 7260–7265. IEEE, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Xingxing Ma, Hongnian Liu, Qiushi Dong, Qizhi Chen, and Ninghao Cai. Statistical post-processing of multiple meteorological elements using the multimodel integration embedded method. *Atmospheric Research*, 301:107269, 2024.

Antonios Mamalakis, Elizabeth A Barnes, and Imme Ebert-Uphoff. 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.

Amy McGovern, Ryan Lagerquist, David John Gagne, G Eli Jergensen, Kimberly L Elmore, Cameron R Homeyer, and Travis Smith. Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, 100(11):2175–2199, 2019.

John R Mecikalski, John K Williams, Christopher P Jewett, David Ahijevych, Anita LeRoy, and John R Walker. Probabilistic 0–1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *Journal of Applied Meteorology and Climatology*, 54(5): 1039–1059, 2015.

Maria J Molina, David John Gagne, and Andreas F Prein. A benchmark to test generalization capabilities of deep learning methods to classify severe convective storms in a changing climate. *Earth and Space Science*, 8(9):e2020EA001490, 2021.

Franco Molteni, Roberto Buizza, Tim N Palmer, and Thomas Petroliagis. The ecmwf ensemble prediction system: Methodology and validation. *Quarterly journal of the royal meteorological society*, 122(529): 73–119, 1996.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65: 211–222, 2017.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Baoxiang Pan, Kuolin Hsu, Amir AghaKouchak, and Soroosh Sorooshian. Improving precipitation estimation using convolutional neural network. *Water Resources Research*, 55(3):2301–2321, 2019.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Timo Ewalds, Andrew El-Kadi, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, 2023.

QiFeng Qian and XiaoJing Jia. Seasonal forecast of winter precipitation over china using machine learning models. *Atmospheric Research*, 294:106961, 2023.

Umamaheswari Rajasekaran, GK Sriram, A Malini, and Vandana Sharma. Hybrid explainable srnn-lstm architecture for irradiance, temperature and wind speed forecasting. 2023.

Neelesh Rampal, Peter B Gibson, Abha Sood, Stephen Stuart, Nicolas C Fauchereau, Chris Brandolino, Ben Noll, and Tristan Meyers. High-resolution downscaling with interpretable deep learning: Rainfall extremes over new zealand. *Weather and Climate Extremes*, 38:100525, 2022.

Stephan Rasp and Sebastian Lerch. Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900, 2018.

Stephan Rasp, Michael S Pritchard, and Pierre Gentine. Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39):9684–9689, 2018.

Xiaoli Ren, Xiaoyong Li, Kaijun Ren, Junqiang Song, Zichen Xu, Kefeng Deng, and Xiang Wang. Deep learning-based weather prediction: a survey. *Big Data Research*, 23:100178, 2021.

Mathieu Renault and Siamak Mehrkanoon. Sar-unet: Small attention residual unet for explainable nowcasting tasks. *arXiv preprint arXiv:2303.06663*, 2023.

MH Retsch, C Jakob, and MS Singh. Identifying relations between deep convection and the large-scale atmosphere using explainable artificial intelligence. *Journal of Geophysical Research: Atmospheres*, 127 (3):e2021JD035388, 2022.

Eloy Reulen and Siamak Mehrkanoon. Ga-smaat-gnet: Generative adversarial small attention gnet for extreme precipitation nowcasting. *arXiv preprint arXiv:2401.09881*, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016a.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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, 2016b.

Lewis F Richardson. *Weather prediction by numerical process*. University Press, 1922.

DA Sachindra, Khandakar Ahmed, Md Mamunur Rashid, S Shahid, and BJC Perera. Statistical downscaling of precipitation using machine learning techniques. *Atmospheric research*, 212:240–258, 2018.

Sebastian Scher and Gabriele Messori. Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, 144(717):2830–2841, 2018.

Axel Seifert and Stephan Rasp. Potential and limitations of machine learning for modeling warm-rain cloud microphysical processes. *Journal of Advances in Modeling Earth Systems*, 12(12):e2020MS002301, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Stephen A Shield and Adam L Houston. Diagnosing supercell environments: A machine learning approach. *Weather and Forecasting*, 37(5):771–785, 2022.

Sam J Silva, Christoph A Keller, and Joseph Hardin. Using an explainable machine learning approach to characterize earth system model errors: Application of shap analysis to modeling lightning flash occurrence. *Journal of Advances in Modeling Earth Systems*, 14(4):e2021MS002881, 2022.

Maike Sonnewald and Redouane Lguensat. Revealing the impact of global heating on north atlantic circulation using transparent machine learning. *Journal of Advances in Modeling Earth Systems*, 13(8): e2021MS002496, 2021.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Selim Furkan Tekin, Oguzhan Karaahmetoglu, Fatih Ilhan, Ismail Balaban, and Suleyman Serdar Kozat. Spatio-temporal weather forecasting and attention mechanism on convolutional lstms. *arXiv preprint arXiv:2102.00696*, 4, 2021.

Ngoan Thanh Trieu, Bernard Pottier, Vincent Rodin, and Hiep Xuan Huynh. Interpretable machine learning for meteorological data. In *2021 The 5th International Conference on Machine Learning and Soft Computing*, pp. 11–17, 2021.

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

Benjamin A Toms, Elizabeth A Barnes, and James W Hurrell. Assessing decadal predictability in an earth-system model using explainable neural networks. *Geophysical Research Letters*, 48(12):e2021GL093842, 2021a.

Benjamin A Toms, Karthik Kashinath, Da Yang, et al. Testing the reliability of interpretable neural networks in geoscience using the madden–julian oscillation. *Geoscientific Model Development*, 14(7):4495–4508, 2021b.

Julio J Valdés and Antonio Pou. A machine learning-explainable ai approach to tropospheric dynamics analysis using water vapor meteosat images. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8. IEEE, 2021.

Chong Wang and Xiaofeng Li. A deep learning model for estimating tropical cyclone wind radius from geostationary satellite infrared imagery. *Monthly Weather Review*, 151(2):403–417, 2023.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Yaqiang Wang. Meteoinfo: Gis software for meteorological data visualization and analysis. *Meteorological Applications*, 21(2):360–368, 2014.

YaQiang Wang. An open source software suite for multi-dimensional meteorological data computation and visualisation. *J. Open Res. Softw*, 7(1):21, 2019.

Yueya Wang, Xiaoming Shi, Lili Lei, and Jimmy Chi-Hung Fung. Deep learning augmented data assimilation: Reconstructing missing information with convolutional autoencoders. *Monthly Weather Review*, 150(8): 1977–1991, 2022.

Peter AG Watson. Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. *Journal of Advances in Modeling Earth Systems*, 11(5):1402–1417, 2019.

Jonathan A Weyn, Dale R Durran, and Rich Caruana. Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, 11(8):2680–2693, 2019.

Jonathan A Weyn, Dale R Durran, and Rich Caruana. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109, 2020.

Pin Wu, Xuting Chang, Wenyan Yuan, Junwu Sun, Wenjie Zhang, Rossella Arcucci, and Yike Guo. Fast data assimilation (fda): Data assimilation by machine learning for faster optimize model state. *Journal of Computational Science*, 51:101323, 2021.

Haoyi Xiong, Xiaofei Zhang, Jiamin Chen, Xinhao Sun, Yuchen Li, Zeyi Sun, Mengnan Du, et al. Towards explainable artificial intelligence (xai): A data mining perspective. *arXiv preprint arXiv:2401.04374*, 2024.

Ruyi Yang, Jianli Mu, Shudong Wang, and Lijuan Wang. Hourly rolling correction of precipitation forecast via convolutional and long short-term memory networks. *Atmospheric Science Letters*, 23(10):e1100, 2022.

Tingzhao Yu and Ruyi Yang. Temporal dynamic network with learnable coupled adjacent matrix for wind forecasting. *IEEE Geoscience and Remote Sensing Letters*, 2023.

Tingzhao Yu, Qiuming Kuang, and Ruyi Yang. Atmconvgru for weather forecasting. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

Tingzhao Yu, Ruyi Yang, Yan Huang, Jinbing Gao, and Qiuming Kuang. Terrain-guided flatten memory network for deep spatial wind downscaling. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:9468–9481, 2022.

Chang-Jiang Zhang, Jing Zeng, Hui-Yuan Wang, Lei-Ming Ma, and Hai Chu. Correction model for rainfall forecasts using the lstm with multiple meteorological factors. *Meteorological Applications*, 27(1):e1852, 2020.

Tao Zhang, Wuyin Lin, Yanluan Lin, Minghua Zhang, Haiyang Yu, Kathy Cao, and Wei Xue. Prediction of tropical cyclone genesis from mesoscale convective systems using machine learning. *Weather and Forecasting*, 34(4):1035–1049, 2019.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619(7970):526–532, 2023.

Haiyan Zhao, Fan Yang, Himabindu Lakkaraju, and Mengnan Du. Opening the black box of large language models: Two views on holistic interpretability. *arXiv preprint arXiv:2402.10688*, 2024.

Kanghui Zhou, Yongguang Zheng, Wansheng Dong, and Tingbo Wang. A deep learning network for cloud-to-ground lightning nowcasting with multisource data. *Journal of Atmospheric and Oceanic Technology*, 37(5):927–942, 2020.

Jing-Yi Zhuo and Zhe-Min Tan. Physics-augmented deep learning to improve tropical cyclone intensity and size estimation from satellite imagery. *Monthly Weather Review*, 149(7):2097–2113, 2021.