Earth's Future

RESEARCH ARTICLE
10.1029/2024EF004540

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

How Interpretable Machine Learning Can Benefit Process Understanding in the Geosciences

Shijie Jiang[1,2], Lily-belle Sweet[3,4], Georgios Blougouras[1,2,5], Alexander Brenning[2,5], Wantong Li[1], Markus Reichstein[1,2], Joachim Denzler[2,6], Wei Shangguan[7], Guo Yu[8], Feini Huang[1,2,7], and Jakob Zscheischler[3,4,9]

[1]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany, [2]ELLIS Unit Jena, Jena, Germany, [3]Department of Compound Environmental Risks, Helmholtz Centre for Environmental Research—UFZ, Leipzig, Germany, [4]Faculty of Environmental Sciences, Technische Universität Dresden, Dresden, Germany, [5]Department of Geography, Friedrich Schiller University Jena, Jena, Germany, [6]Computer Vision Group, Friedrich Schiller University Jena, Jena, Germany, [7]School of Atmospheric Sciences, Sun Yat–Sen University, Zhuhai, China, [8]Division of Hydrologic Sciences, Desert Research Institute, Las Vegas, NV, USA, [9]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Leipzig, Germany

Abstract Interpretable Machine Learning (IML) has rapidly advanced in recent years, offering new opportunities to improve our understanding of the complex Earth system. IML goes beyond conventional machine learning by not only making predictions but also seeking to elucidate the reasoning behind those predictions. The combination of predictive power and enhanced transparency makes IML a promising approach for uncovering relationships in data that may be overlooked by traditional analysis. Despite its potential, the broader implications for the field have yet to be fully appreciated. Meanwhile, the rapid proliferation of IML, still in its early stages, has been accompanied by instances of careless application. In response to these challenges, this paper focuses on how IML can effectively and appropriately aid geoscientists in advancing process understanding—areas that are often underexplored in more technical discussions of IML. Specifically, we identify pragmatic application scenarios for IML in typical geoscientific studies, such as quantifying relationships in specific contexts, generating hypotheses about potential mechanisms, and evaluating process-based models. Moreover, we present a general and practical workflow for using IML to address specific research questions. In particular, we identify several critical and common pitfalls in the use of IML that can lead to misleading conclusions, and propose corresponding good practices. Our goal is to facilitate a broader, yet more careful and thoughtful integration of IML into Earth science research, positioning it as a valuable data science tool capable of enhancing our current understanding of the Earth system.

Plain Language Summary Artificial Intelligence is a rapidly advancing field, in which Interpretable Machine Learning (IML) is seen as having the potential to significantly improve our understanding of Earth's complex environmental systems. IML goes beyond the predictive power of machine learning models, focusing instead on uncovering the relationships within the data that are revealed by the model's learning process. However, there is still a lack of straightforward, practical domain-specific guidelines for geoscientists that facilitate both broader and more careful application in the field. In this paper, we aim to demonstrate the real-world benefits of IML in typical geoscientific analysis. We provide a clear, step-by-step workflow that shows how IML can be used to address specific questions. We also point out some common pitfalls in using IML and offer solutions to avoid them. Our goal is to make IML more accessible and useful to a wider range of geoscientists, and we believe that IML, if used properly and thoughtfully, can become an essential and valuable tool to advance our understanding of complex Earth systems.

Key Points:

- We demonstrate the broader relevance of Interpretable Machine Learning (IML) to most geoscientists and underexplored opportunities for its use
- We describe a workflow for the effective use of IML while cautioning against potential and common pitfalls
- We suggest good practices for its adoption and advocate for more careful application to ensure reliable and robust insights for the field

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Jiang,
sjiang@bgc-jena.mpg.de

Citation:

Jiang, S., Sweet, L.-b., Blougouras, G., Brenning, A., Li, W., Reichstein, M., et al. (2024). How interpretable machine learning can benefit process understanding in the geosciences. *Earth's Future*, *12*, e2024EF004540. https://doi.org/10.1029/2024EF004540

Received 8 FEB 2024
Accepted 14 JUN 2024

Author Contributions:

Conceptualization: Shijie Jiang, Lily-belle Sweet, Georgios Blougouras, Wantong Li, Markus Reichstein, Wei Shangguan, Guo Yu, Jakob Zscheischler
Formal analysis: Markus Reichstein
Funding acquisition: Shijie Jiang, Markus Reichstein

1. Introduction

The widespread application of machine learning (ML) in the geosciences, particularly for predictive modeling, represents a significant technological advance (e.g., Bi et al., 2023; Ham et al., 2019). While their predictive capabilities are widely acknowledged, ML methods are often considered separate from the fundamental scientific methodologies of the geosciences, typically being viewed as more practical tools for simulation and forecasting rather than integral components of scientific exploration (Nearing et al., 2021). It was hoped that ML would revolutionize scientific inquiry (H. Wang et al., 2023), but this anticipated transformation has yet to fully

Investigation: Shijie Jiang, Lily-belle Sweet, Georgios Blougouras, Alexander Brenning, Guo Yu, Feini Huang
Project administration: Shijie Jiang
Supervision: Shijie Jiang, Markus Reichstein
Validation: Lily-belle Sweet, Alexander Brenning, Markus Reichstein, Joachim Denzler, Jakob Zscheischler
Visualization: Shijie Jiang, Georgios Blougouras
Writing – original draft: Shijie Jiang, Lily-belle Sweet, Georgios Blougouras, Wantong Li, Markus Reichstein, Wei Shangguan, Guo Yu, Jakob Zscheischler
Writing – review & editing: Shijie Jiang, Lily-belle Sweet, Georgios Blougouras, Alexander Brenning, Wantong Li, Markus Reichstein, Joachim Denzler, Wei Shangguan, Guo Yu, Feini Huang, Jakob Zscheischler

manifest. One concern is that these innovations may not be fully aligned with the core scientific goals of the discipline (e.g., Birhane et al., 2023).

In recent years, the ML community has made significant progress in developing strategies to improve model interpretability, leading to the evolution of interpretable ML (IML) and explainable AI (XAI) (Gunning & Aha, 2019; Murdoch et al., 2019). Despite the differences between IML and XAI as they are used in the ML community, for example, IML focuses more on models and XAI includes a broader set of techniques to make ML more explainable (Rudin et al., 2022), we choose not to emphasize these distinctions in this paper. Our concentration is more on the broader concept and practical applications of IML, and thus most terms related to IML can be used interchangeably with XAI in the following discussions. In this paper, we approach IML from a practical perspective, focusing primarily on the use of post-hoc interpretation techniques, although inherently interpretable models are also discussed in context. These post-hoc interpretation techniques, such as Shapley additive explanations (SHAP) (Lundberg & Lee, 2017), integrated gradients (Sundararajan et al., 2017), and local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016), are intended to help users demystify the inner workings of complex, often opaque ML models. For geoscientists with ML expertise, these techniques (albeit with varying effectiveness) provide a way to demonstrate the credibility of a model by justifying its mechanisms against existing knowledge (Dwivedi et al., 2023). Here, however, we highlight the potential of IML to benefit a much broader range of geoscientists, including those who have not engaged with ML models in their research. Essentially, IML provides a new lens for exploring, interpreting, and understanding the complex relationships within geoscientific data (Toms et al., 2020). Through the process of interpreting ML models, we may gain insight into how different input features interact and influence geoscientific phenomena, including relationships that might be difficult to identify through traditional analysis (e.g., Ham et al., 2023; Jiang et al., 2024; Kraft et al., 2019).

Despite the progress made in implementing IML to improve scientific understanding and discovery in many fields (Roscher et al., 2020b), its integration into established geoscientific methodologies still requires both more diverse and careful application. On the one hand, IML is often introduced through a data science-centric lens that typically focuses on its fundamental concepts, various algorithms, and major research trajectories and trends (e.g., Adadi & Berrada, 2018). However, for geoscientists engaged with process-based models, the value of IML may not be immediately apparent, as IML is more likely to be perceived as a technical tool for justifying or debugging ML models. On the other hand, the rapid proliferation of IML has also led to instances of careless application without a thorough understanding of its limitations and underlying assumptions (Arif & MacNeil, 2022; Molnar et al., 2022; Roscher et al., 2020a).

Therefore, this paper aims to bridge this gap by highlighting both the practical benefits and good practices of using IML in geoscientific research. Our goal is not to provide an exhaustive review of IML techniques and their applications in the geosciences, which have been extensively covered in the literature, such as Gevaert (2022) on IML in Earth observation and remote sensing, Başağaoğlu et al. (2022) in hydroclimate, and Mamalakis, Ebert-Uphoff, and Barnes (2022) in meteorology and climate science. Here, we narrow our focus to the direct relevance of IML for broad geoscience purposes, particularly with respect to process understanding, an aspect that has been variously highlighted as critical in AI for Earth system science (e.g., Irrgang et al., 2021; Reichstein et al., 2019; Shen et al., 2023) but still needs more focused discussion. Specifically, we will concentrate on promising applications of IML for targeted insights for geoscientists, including non-linear quantification of relationships within data, generation of hypotheses about potential mechanisms, and evaluation of process-based models. We present a general but practical workflow for effectively integrating IML into routine research activities, from translating specific geoscience research questions into IML tasks to obtaining actionable insights from the IML models. Importantly, we identify several common pitfalls of IML applications and emphasize that careless use of IML can not only lead to potentially misleading conclusions but also undermine its credibility in the geoscience field. Ultimately, we hope to make IML more accessible and relevant to a broader range of geoscientists, enabling them to properly use these innovative tools in their scientific endeavors and opening new avenues for understanding Earth's complex systems.

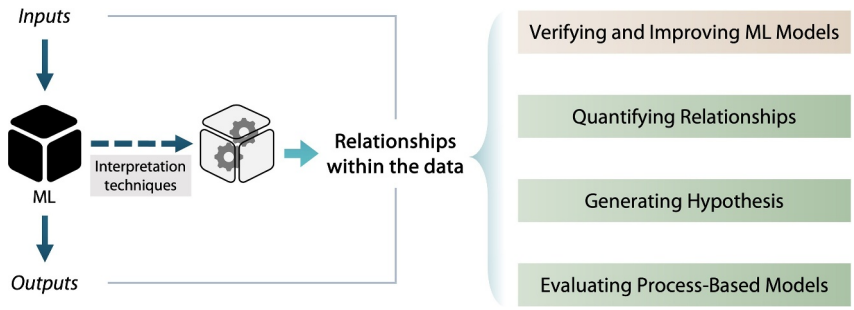

Figure 1. *The relationship between data, machine learning (ML) models, and post-hoc interpretation techniques, in the framework of interpretable ML (IML), as well as the usefulness of their results in Earth science studies*. The primary goal of using IML in this context is to uncover relationships within the data used to make predictions. Dark blue arrows represent the flow from data through opaque ML models to post-hoc interpretation techniques that make the ML models interpretable. The revealed relationships can support various aspects of geoscientific research, with green boxes indicating applications that are directly relevant to broader geoscience studies in terms of process understanding.

2. Relevance for Geosciences

2.1. Demystifying IML for Geoscientists

While there has been a surge in IML research over the past decade, the concept of deriving interpretable models from data has a longer tradition (Molnar et al., 2020). For instance, linear regression, which dates back to the early nineteenth century and has been widely used in scientific studies, can be broadly considered an early incarnation of IML. Linear regression has evolved into a variety of regression analysis tools, such as logistic regression, generalized linear models (GLMs) (Nelder & Wedderburn, 1972), and generalized additive models (GAMs) (Wood, 2017). These models are often designed based on specific distributional assumptions and a predefined limit on complexity to ensure interpretability. Similarly, models such as decision trees (Quinlan, 1986) and decision rules (Quinlan, 1987) are inherently interpretable, as their decision logic can be easily traced by examining the learned rules or structured hierarchy of decisions.

However, relying solely on the inherent interpretability of these simple models can have limitations, particularly in terms of predictive performance, as their simplicity may restrict their ability to capture arbitrarily non-linear interactions. On the other hand, complex ML models, such as deep neural networks (NNs) and boosting algorithms, often outperform simpler models in terms of accuracy, but lack inherent interpretability (Dramsch, 2020; Molnar et al., 2020). As a result, popular IML research uses post-hoc interpretation methods to explain the output of "black-box" ML models (Dwivedi et al., 2023). Particularly, the built-in measure of feature importance in random forests (RFs) was a major milestone (Breiman, 2001a). Since around 2015, the IML field has experienced significant growth, with the emergence of numerous model-agnostic explanation methods applicable to different ML model types as well as model-specific explanation methods tailored to NNs or tree ensembles (Molnar et al., 2020). These post-hoc methods (hereafter referred to as interpretation methods or interpretation techniques) do not simplify the model itself, but rather provide a window into the complex interactions and non-linear relationships that the model has captured from the data.

Generally, interpretation methods analyze the relationships learned by ML models by examining the model components or sensitivities (Figure 1). For instance, activation maps help reveal how internal representations are formed by convolutional NNs by visualizing the layer-wise activation patterns (Olah et al., 2017). In comparison, interpretation methods (e.g., SHAP, integrated gradients, and LIME) study the sensitivity aspect of ML models by perturbing the original inputs, computing the gradient of model outputs with respect to inputs, or approximating complex models with inherently interpretable models. Other than understanding the contribution of each input feature to individual predictions, interpretation methods such as partial dependence plots (Friedman, 2001) and permutation feature importance (Altmann et al., 2010) illustrate the general impact of features across the data set. Overall, compared to inherently interpretable models, post-hoc methods bring partial but functional interpretability without sacrificing the predictive power of advanced ML models.

In summary, IML is not an entirely new concept but can be regarded as a form of data analysis, or more specifically, an approach to understanding data through the lens of the data-driven models that process it. This perspective, though differing from the formal definition of IML, is useful in its pragmatism. IML essentially extends the capabilities of traditional statistical tools by providing sophisticated methods for analyzing variable relationships, which is particularly valuable in the geosciences where complex interactions and non-linear relationships are common. For readers interested in a more detailed technical understanding of the IML algorithms and methods discussed in this paper, we recommend referring to comprehensive reviews (e.g., Adadi & Berrada, 2018; Barredo Arrieta et al., 2020; Başağaoğlu et al., 2022; Gevaert, 2022; Gilpin et al., 2018; Gunning et al., 2019; Mamalakis, Ebert-Uphoff, & Barnes, 2022; Molnar et al., 2020; Murdoch et al., 2019; Roscher et al., 2020a, 2020b) that provide in-depth discussions of various interpretation techniques, their theoretical underpinnings, their implementation details, and their applications in various subfields of the geosciences.

2.2. Usefulness of IML for Geoscientists

IML offers a variety of applications in the geosciences, and its usefulness may be most apparent to geoscientists who focus on ML, primarily to justify and diagnose their models for predictive tasks (e.g., Kratzert et al., 2019; Mayer & Barnes, 2021). In this paper, however, we will not discuss such applications in depth, but rather explore how IML can be directly used to potentially enhance process understanding for the field (Figure 1).

2.2.1. Quantifying Relationships Within a Given Context

A fundamental aspect of process understanding in the geosciences is quantifying the relationships within data, including identifying which variables are most influential, understanding the nature of their influence (whether linear, non-linear, or conditional), and determining how changes in one variable might affect others. IML is directly applicable in this context, equipping geoscientists with the tools necessary to quantitatively delineate relationships within established frameworks. These relationships may be partially known, but possibly remain qualitative, conceptual, or local. For example, IML has been used to explore relationships between environmental variables and diverse phenomena, such as species distributions (Ryo et al., 2020), flooding mechanisms (Jiang, Zheng, et al., 2022), landslide generation processes (Brenning et al., 2015), and soil-vegetation coupling (W. Li et al., 2022). IML has facilitated the identification of hotspot regions where precipitation anomalies are highly sensitive to anthropogenic warming (Ham et al., 2023), or where regional temperature signals exhibit significant sensitivity to aerosol forcing (Labe & Barnes, 2021). Overall, IML allows geoscientists to refine and enhance the current scientific understanding in a quantifiable and non-linear context. However, it should be emphasized that the relationships uncovered by IML using predictive models, while potentially useful, are not inherently causal, as discussed in more detail in Section 4.2.

In general, the primary approaches to infer variable relationships in the geosciences are conventional statistical analysis and numerical experiments using process-based models. Conventional (parametric) statistical methods, which are based on solid theory and usually provide additional confidence intervals, prediction intervals, and significance tests, are best suited for confirming well-defined hypothetical relationships. In contrast, IML enhances data exploration within large, high-dimensional data sets that often contain a multitude of interacting factors and patterns that are not readily apparent through traditional statistical analysis (Breiman, 2001b). Moreover, a practical advantage of IML is its ability to provide granular interpretations for individual instances (Lundberg et al., 2020), which is important in scenarios where we need to understand specific data points, such as the potential drivers of extreme events (van Oldenborgh et al., 2021).

Numerical experiments, such as controlled experiments, scenario analyses, and sensitivity analyses using process-based models, are common in the geosciences and critical to understanding how systems respond to environmental change (e.g., O'Neill et al., 2016). However, conducting such numerical experiments can be time-consuming, limiting the number of experiments that can be realistically conducted. Controlled simulation experiments also run the risk of inadvertently disrupting natural interdependencies, such as the typically anti-correlated relationship between temperature and precipitation at interannual scales during summer (Madden & Williams, 1978). If specific variables are manipulated in isolation, these experiments may lead to artificial combinations of variables that are not physically plausible. This consideration is particularly important for understanding compound weather and climate events, where the combination of non-extreme drivers can lead to extreme impacts (Zscheischler & Seneviratne, 2017). Moreover, the effectiveness of numerical experiments often

depends on a well-established understanding of the underlying mechanisms of the systems. In cases where these mechanisms are not fully known, or where comprehensive process-based models are not available, the application of IML to observational data may be partially useful (Irrgang et al., 2021).

2.2.2. Generating Hypotheses About Mechanisms With IML

In traditional geoscientific research, hypothesis generation often follows a time-intensive path that begins with careful observation of phenomena, followed by the formulation of a hypothesis based on those observations (Sivapalan & Blöschl, 2017). This process typically involves extensive data collection, analysis, and integration of multiple data sources (often based on the researcher's intuition) to identify patterns or anomalies. While thorough, this approach can be slow and sometimes limited by the inherent biases and perceptual limitations of human analysis. This is especially challenging in the era of big Earth data, where traditional analytical methods may struggle to navigate the complexities inherent in large, diverse, and multimodal data sets (X. Li et al., 2023). In comparison, IML can quickly analyze large and complex data sets, such as multidimensional data from multiple sources (e.g., satellite imagery, sensor networks, and historical records). For instance, using a large data set of Earth observations and climate variables, Kraft et al. (2019) analyzed variable contributions to temporally lagged dependencies (i.e., memory effects) in vegetation modeling through interpretation of long short-term memory (LSTM) models. This investigation revealed some new aspects of memory effects, such as their associations with climate gradients. While IML by itself does not confirm causality, because ML models may predict the right outcome for the wrong reasons (Lapuschkin et al., 2019), the correlations, statistical dependencies, and patterns it uncovers can still be informative. For example, IML may reveal that certain variables, previously deemed unlikely to be relevant, play a significant role in predictions, or that the relevance of variables shifts in ways that defy initial expectations (Ryo et al., 2020). These findings can prompt geoscientists to reevaluate their prior assumptions, providing valuable starting points for further rigorous testing and investigation through targeted studies and experiments (Carloni et al., 2023). This ability of IML to efficiently sift through and interpret large amounts of data can accelerate the hypothesis generation process, and thus the entire research cycle. This rapid turnaround is particularly beneficial in climate research or natural hazard assessment, where timely insights can have a significant impact (van Oldenborgh et al., 2021). Furthermore, IML's ability to handle large data sets means that hypotheses can be generated and refined in real time as new data become available, keeping pace with the dynamic and evolving nature of the Earth system.

2.2.3. Evaluating Process-Based Models With IML Insights

The relationships and patterns revealed by IML also facilitate the assessment of the variability of specific factors across models, data sets, or scenarios (Reichstein et al., 2019), something that is often less emphasized. Understanding whether different models consistently reproduce the dependence structure of variables observed in real-world data would help evaluate and refine process-based models (e.g., Gnann et al., 2023). Process-based models are essential for projections of future trends, though the reliability of these models in simulating future climate events cannot be directly evaluated. Traditional model evaluation and intercomparison have largely relied on benchmarking approaches that focus on univariate comparisons, where the performance of models is assessed based on their ability to reproduce observed values of individual variables (Jägermeyr et al., 2021). However, the univariate approach may overlook compensating errors that arise from interactions among multiple variables within a system, potentially masking problems in model structure or parameterization (Touzé-Peiffer et al., 2020). C. Müller et al. (2024) emphasize the need to include analyses of functional properties in process-based model evaluation, which may reveal more about model plausibility and skill than merely comparing variables, since different model responses to drivers may offset each other in the historical evaluation period, but not in future scenarios. To this end, the use of IML to evaluate these multifaceted relationships holds promise to provide geoscientists with a tool that complements and enhances traditional evaluation techniques and moves toward pattern- and process-oriented model evaluation (Reichstein et al., 2019). By inter-comparing IML-derived relationships from models with those from observational data sets, we can uncover the consistencies and inconsistencies in their covariability, and thus identify specific aspects of the model that may require adjustment or further investigation. Such evaluations are particularly relevant for addressing the challenges of predicting extreme climate and weather events under climate change, which are often caused by complex interactions among multiple factors (Zscheischler et al., 2018). In this case, however, the challenges of out-of-distribution predictions and representational biases are considerable.

Recently, advanced data science methods such as complex networks and causal discovery algorithms have been increasingly used in climate model evaluation. For example, Feldhoff et al. (2014) applied complex networks to evaluate a regional climate model simulating multiple climate variables in South America, where the characteristics of the constructed networks were compared between the model and reanalysis data. Likewise, Nowack et al. (2020) used causal networks to evaluate coupled model intercomparison project phase 5 (CMIP5) models, focusing on their ability to simulate atmospheric dynamical interactions represented by lagged correlations between climate variables at remote locations. They found models that more accurately capture characteristic causal relationships tend to have smaller biases in their precipitation simulations. However, the potential for using IML for model evaluation, for example, in various model intercomparison projects (e.g., Eyring et al., 2016; Warszawski et al., 2014), remains largely unexplored. IML could be used to systematically compare these models to identify common or different model biases, to constrain uncertainties in climate change projections, and to provide a comprehensive overview of areas for improvement.

3. Typical Workflow of IML for Process Understanding

Having established the relevance and applicability of IML in the geosciences, this section is dedicated to outlining an actionable workflow for the effective use of IML in geoscientific research (Figure 2a). This workflow is intended as a practical guide to assist geoscientists in structuring their research questions and methodologies around IML to achieve reliable and meaningful results. Here, we focus on presenting the key stages and general principles of the workflow, exemplified by selected cases from the existing literature (Figures 2b–2g). Detailed technical discussions and more extensive examples can be found in Supporting Information S1.

In all cases, the decision to use IML, whether simple or complex, should be contextually appropriate to the specific complexity and demands of the data and research questions. Once the IML workflow has been implemented, it is advisable to compare the results with those derived from traditional analysis methods to assess the unique insights and added value that IML can bring to the study.

3.1. Translating Geoscientific Research Questions Into IML Tasks

The first and perhaps most critical step is to clearly define the research question and translate it into a task that can be effectively addressed using IML methods. Typical investigations may focus on identifying key influencing factors and their contributions, or untangling dependencies and conditional effects. For example, geoscientists may want to understand how a specific outcome (e.g., extreme weather and climate events) can be attributed to potential drivers (e.g., Davenport & Diffenbaugh, 2021; Jiang, Zheng, et al., 2022; Kondylatos et al., 2022; Ryo et al., 2020; R. Wang et al., 2021). In these scenarios, the IML task is to quantify the relationships between these events (Y) and a number of possible influencing factors (X), such as atmospheric conditions or geographic features. Beyond attribution to individual factors, IML can be used to determine how multiple factors interactively affect a particular outcome (e.g., H. Wang et al., 2022; Xu et al., 2023). The question of critical thresholds in systems can also sometimes be translated into IML tasks of identifying inflection points in the contribution of X relative to its value (e.g., Chakraborty et al., 2021). These examples and additional case studies are elaborated in Text S1 in Supporting Information S1.

At this stage, it is important to form preliminary hypotheses based on existing knowledge, literature review, or exploratory data analysis. These hypotheses can guide the selection of appropriate IML methods that address specific types of data and are consistent with the goals of the research. For instance, if the hypothesis involves exploring the complex interaction effects between variables in tabular data, the IML methods considered should be capable of explicitly quantifying these interactions. Possible approaches may include the use of tree-based models, such as RFs or extreme gradient boosting (XGBoost) (Chen & Guestrin, 2016), in conjunction with TreeSHAP (Lundberg et al., 2020), which allows the decomposition of model predictions into the contributions of feature pairs based on the structured decision paths inherent in these models. Alternatively, one could consider Explainable Boosting Machines (Lou et al., 2013), where interaction terms can be explicitly specified during model configuration and each interaction term is modeled and interpreted separately.

For readers seeking detailed guidance on method selection, numerous comprehensive reviews and studies (e.g., Barredo Arrieta et al., 2020; Bommer et al., 2024; Graziani et al., 2023; Mamalakis, Barnes, & Ebert-Uphoff, 2022; McGovern, Lagerquist, et al., 2019; Schwalbe & Finzel, 2023; Zhong et al., 2022) are available that thoroughly examine the suitability and conditions for using specific IML methods. For instance, Schwalbe and

10.1029/2024EF004540

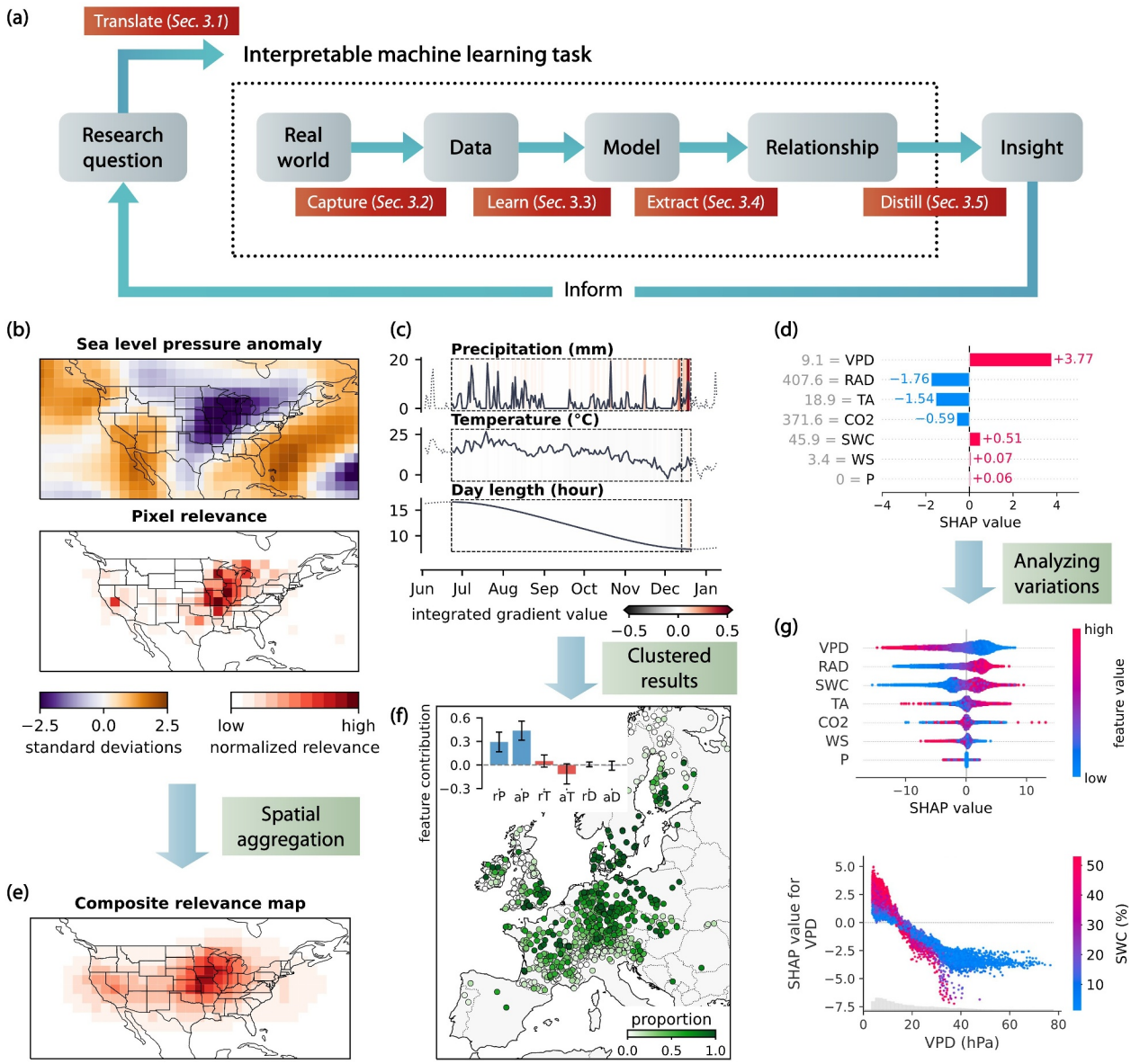

Figure 2. *Workflow and examples of applying interpretable machine learning for geoscientific process understanding*. (a) Flowchart illustrating the general workflow, where gray boxes represent objects and red boxes represent operations (explained in the corresponding subsections). (b–g) Illustrate how the algorithmic explanation results for different types of data can be translated into scientific understanding with examples from the literature (briefly explained in Section 3.5 and detailed in Text S4 in Supporting Information S1). (b, e) Are modified from Davenport and Diffenbaugh (2021), where (b) shows a sea level pressure anomaly map for a given day, with the IML-derived pixel-wise relevance indicating its contribution to the classification of the day as having large-scale extreme precipitation circulation patterns (EPCP). (e) Presents composite relevance maps for EPCP days, which aggregate the relevance maps exemplified in (b). (c, f) Are adapted from Jiang, Bevacqua, and Zscheischler (2022). (c) Shows the IML-derived feature importance of precipitation, temperature, and day length over 180 days for predicting streamflow on the subsequent day. (f) Illustrates the results of a clustering analysis applied to all feature importance values across events and basins, with the bar plot indicating the average feature contribution pattern (aP: antecedent precipitation from 180 to 7 days before the event) and the map showing the proportion of events falling into this cluster in individual basins. (d, g) Are adapted from H. Wang et al. (2022). (d) Indicates the contribution of seven variables in predicting gross primary productivity for a specific sample, as estimated by the SHAP value. The actual values of these variables are shown in gray. The top plot in (g) illustrates the relationship between feature contribution (*x*-axis) and values (color) across all variables. The bottom plot in (g) is a dependence plot of vapor pressure deficit (VPD) versus its contribution value along the soil water content (SWC) gradient in grasslands. For more information, including definitions of other abbreviations, see the respective references.

Finzel (2023) provide a structured and detailed taxonomy of IML methods that synthesizes insights from a multitude of surveys on IML techniques, metrics, and characteristics, which can assist researchers in identifying the most suitable IML methods for various domain-specific explanation use cases. Additionally, Bommer

et al. (2024) discuss metrics for evaluating different IML methods in the context of climate science. They highlight key considerations in selecting an appropriate IML method and propose a framework using evaluation metrics to support the selection of an appropriate IML method for a specific research task.

3.2. Preparing and Preprocessing Data

Data preparation is a fundamental step in the IML workflow. The accuracy and reliability of IML outcomes depend heavily on the collection of appropriate and comprehensive data relevant to the defined problem. Typically, variable selection is improved iteratively, guided by model evaluation and interpretation in subsequent steps. In addition to following general principles of data preparation for ML models, such as handling missing values or outliers (Zhu et al., 2023), it is necessary to ensure that the data adequately reflect the temporal and spatial scales relevant to the processes under study (Jiang, Bevacqua, & Zscheischler, 2022; W. Li et al., 2022). Importantly, data volume alone may not be sufficient for IML studies; diversity within the data is equally important (Fang et al., 2022), and different scenarios, conditions, and variations should be included. However, the sample distribution should not disproportionately favor, for instance, certain climatic zones or geographic features, a common issue in site-based observational data sets (Chu et al., 2017). In addition, as a general principle, data often require cleaning, formatting, and transformation to be used effectively (e.g., L. Yu et al., 2006), and this is no different for geoscience data. Depending on the research question, it may also be necessary to remove seasonality and long-term trends from time series data in order to focus on more specific variables of interest (e.g., Davenport & Diffenbaugh, 2021; W. Li et al., 2022) (detailed in Text S2 in Supporting Information S1).

3.3. Training and Validating ML Models

Training a ML model for process understanding may require more consideration than for purely predictive tasks. For example, the choice of an appropriate ML model should be informed by the complexity of the geoscience question and data, as well as the goals of the analysis. In general, the chosen model should be as complex as necessary to capture the essential dynamics of the data, but as simple as possible so that its interpretations can be translated into concise and actionable insights (Toms et al., 2020). The full extent of complexity is often not immediately evident, so it is wise to start with a simpler and more transparent model as a baseline and increase complexity incrementally (discussed in Section 4.5). Also, different ML models have unique strengths and are better suited to specific types of geoscience questions, and the choice of model affects how well it can handle spatial and/or temporal aspects of the data (Grinsztajn et al., 2022; Ham et al., 2019; Jiang, Bevacqua, & Zscheischler, 2022; Kraft et al., 2019; Kratzert et al., 2019; Lees et al., 2022; Saha et al., 2021) (detailed in Text S3 in Supporting Information S1).

An important consideration throughout the model building process is to prevent potential information leakage and ensure that the resulting model is generalizable and does not learn shortcuts (Schratz et al., 2019; Sweet et al., 2023). This requires careful management of the training and test data sets with consideration of the specific characteristics of the data (Bischl et al., 2023; Brenning, 2022; Davenport & Diffenbaugh, 2021; de Burgh-Day & Leeuwenburg, 2023; Lopez-Gomez et al., 2023; McGovern, Jergensen, et al., 2019; Meyer & Pebesma, 2022) (detailed in Text S3 in Supporting Information S1), and implementing strategies such as regularization and early stopping to prevent the model from exploiting certain patterns in the training data to overfit (Ying, 2019).

Careful evaluation of model performance is essential to derive meaningful interpretations in subsequent steps. While sufficient predictive accuracy is necessary, it alone does not guarantee that the model has effectively captured the underlying patterns and relationships in the data (Murdoch et al., 2019). Therefore, a comprehensive, multifaceted approach to evaluation is essential. Ideally, model performance should be tested across diverse subsets that vary in time, space, and/or feature distribution (Sweet et al., 2023). Rigorous testing helps to challenge the model, ensuring that it has not only learned specific patterns, shortcuts, or biases that may be inherent to a particular segment of the training data, but has instead developed a broad, generalizable understanding of the data (discussed in Section 4.1). Typically, fitting multiple, independent models (e.g., Jiang, Bevacqua, & Zscheischler, 2022; McGovern, Jergensen, et al., 2019) or exploring different data sets (e.g., Davenport & Diffenbaugh, 2021; Ham et al., 2019; W. Li et al., 2022) and then examining the distribution of their performance metrics can solidify the robustness of the findings. Ultimately, the success of IML in producing reliable and insightful results depends on thorough and thoughtful ML model training and validation.

3.4. Implementing Interpretations and Ensuring Robustness

The choice of an appropriate interpretation technique depends not only on its compatibility with the specific ML model, but also on the level of explanation required (e.g., explanation for individual predictions or global understanding of model behavior). In recent years, SHAP values have gained popularity for their ability to provide detailed insight into each feature's contribution to instance-level model predictions, which can be further aggregated to provide a global perspective of the data set (Lundberg et al., 2020). Other methods such as integrated gradients or expected gradients can be applied to temporal models including LSTM (e.g., Jiang, Bevacqua, & Zscheischler, 2022; Kratzert et al., 2019), while techniques such as layer-relevant propagation or occlusion sensitivity are often used for image-based models (e.g., Ham et al., 2023; Toms et al., 2020). However, choosing among the available interpretation techniques can be challenging due to the lack of ground truth for evaluation. Several metrics have been developed to evaluate the suitability and effectiveness of interpretation techniques, focusing on comparing key properties between techniques for specific research problems (e.g., Hedström et al., 2024; Nauta et al., 2023). Typical examples of these properties include faithfulness—where the high importance assigned to a feature by the interpretation technique should significantly affect the model's prediction—and robustness, which assesses the stability of the explanations against minor input variations. This evaluation is critical for making an informed decision about the most appropriate interpretation technique(s). Readers are encouraged to consult recent studies (e.g., Bommer et al., 2024; Mamalakis, Barnes, & Ebert-Uphoff, 2022) that have conducted comprehensive evaluations of different methods against various metrics tailored to the specific context of Earth science.

It should be recognized that no interpretation technique is universally optimal or suitable for all models and tasks, and results from different interpretation methods have been found to be inconsistent (Krishna et al., 2022; Mamalakis, Barnes, & Ebert-Uphoff, 2022). It is therefore advisable to use more than one method whenever possible to assess the robustness of findings. In addition, an essential consideration for some interpretation techniques is the selection of appropriate baselines or background data, which serve as reference points for understanding how different feature values shift the model output from a base value. Different baselines can lead to divergent interpretations (Mamalakis et al., 2023).

To ensure the robustness and generalizability of the interpretation results obtained, the interpretations should be confirmed as not merely artifacts of the specific data set, ML model, or interpretation technique used. For example, the major patterns of interpretation results should remain as consistent as possible under minor input data perturbations or when using independent data sets from various data sources. Therefore, validation across multiple satellite products, model-based data, or in-situ measurements is appreciated (e.g., W. Li et al., 2022). The inclusion of random variables unrelated to the target variable can also serve as a point of comparison to evaluate the importance of genuine features (e.g., Zhou & Hooker, 2021). Furthermore, the sensitivity of interpretation results to various model configurations (e.g., filter sizes in CNNs, temporal lengths in LSTM, random seeds) should be examined (Mishra et al., 2021). Note that the uncertainty arising from the above processes is an important aspect to consider when applying IML, which will be further discussed in Section 4.4.

3.5. Distilling Interpretation Results Into Geoscientific Understanding

The process of distilling meaningful geoscientific insights from interpretation results requires interpreting the revealed model behavior within the existing geoscientific context. Some interpretation methods are capable of directly describing model behavior within its operational domain by illustrating how input features affect predictions on average, or what concepts a model has generally learned to encode. These include partial dependence plots (Friedman, 2001), permutation feature importance (Altmann et al., 2010), and several emerging techniques such as concept relevance propagation (Achtibat et al., 2023), network dissection (Bau et al., 2020), and structural causal model-based feature relevance (Reimers et al., 2020). In contrast, some interpretation methods (e.g., SHAP value) focus on instance-level explanations, detailing the contribution of individual variables to specific predictions. Figures 2b–2d showcases the form of instance-level interpretation results based on three types of data typical in the geosciences (i.e., spatial data, multivariate time series, and tabular data) from the literature (Davenport & Diffenbaugh, 2021; Jiang, Bevacqua, & Zscheischler, 2022; H. Wang et al., 2022). For spatial data, for instance, the pixel relevance map in Figure 2b highlights areas that significantly influence model predictions. For multivariate time series, the interpretation assigns feature importance values over time, revealing how input variables contribute to specific predictions at each time step, as shown in Figure 2c. In the context of tabular data,

which often has fewer dimensions, interpretations tend to be more straightforward (Figure 2d), indicating how each input variable moves the output value from the model's baseline value to the actual prediction for a given instance. In addition, several other interpretation methods, such as anchor algorithms (Ribeiro et al., 2018) and counterfactual explanations (Wachter et al., 2017), can provide more problem-specific and actionable insights for decision making by identifying precise conditions for predictions or pinpointing minimal input changes that alter the outcome.

Generally, elevating these instance-level interpretations to a comprehensive understanding requires synthesizing these individual insights into a cohesive perspective. Figures 2e–2g presents aggregated interpretation results, corresponding to those in Figures 2b–2d, using various strategies. For example, methods such as composite maps (Figure 2e) and clustering of feature importance (Figure 2f) can help identify key features or common underlying mechanisms across different scenarios or instances. Moreover, investigating how a feature's contribution to model predictions changes with its value and the value of other variables can be informative. For instance, the bee swarm plots in Figure 2g provide a dense summary of each input feature's impact on model output, while the dependence plot illustrates how model predictions depend on interactions between multiple features. These examples are described in more detail in Text S4 in Supporting Information S1 and can be found in the respective literature. In addition, examining variations in feature contributions is also helpful in identifying thresholds or saturation points at which a feature value begins to have diminishing or increasingly significant effects on the predicted outcome (Chakraborty et al., 2021).

4. Common Pitfalls and Good Practices

To effectively apply IML in the geosciences, it is essential to recognize and understand common pitfalls, which are not isolated but interrelated, and to adopt good practices that ensure robust, reliable, and scientifically sound outcomes. This section aims to summarize some key considerations and practical advice on both what to avoid and how best to approach IML applications.

4.1. Model Interpretations Do Not Always Reflect Data Truths

A common pitfall in seeking insights from IML is the misconception that the model's interpretations necessarily equate to truths about the underlying data-generating process or real-world phenomena. In reality, the interpretations offered by these methods merely estimate how a specific ML model arrives at certain predictions based on inputs (Good & Hardin, 2012). Misinterpreting these as direct insights into real-world phenomena can lead to misleading or incorrect conclusions, especially if the model's learned decision rules do not match the actual underlying data relationships (Figure 3a). For example, models that are underfitted due to overly general decision rules will perform poorly on both training and test data, indicating a failure to capture the true underlying relationship. Conversely, overfitted models that learn rules too close to the training data, including noise and anomalies, may also struggle to generalize the underlying relationships. Perhaps more imperceptibly, even models that perform well on training and independent and identically distributed test data, but not on out-of-distribution data, may misrepresent the data-generating process. This scenario can occur when a model relies on superficial or spurious patterns (e.g., shortcut learning) (Geirhos et al., 2020)—for instance, classifying images based on embedded text labels rather than their actual features. In essence, ML algorithms can skillfully perform tasks based on spurious, non-physical relationships, but the true relationships may deviate from the correlations initially observed in the training data.

In the context of the geosciences, the unique spatial and temporal structures inherent in geoscientific data, such as autocorrelation, make these issues particularly critical. For instance, in large-scale ecological mapping studies, ML models are often used to characterize the relationship between local environmental conditions (e.g., climate, topography, and soil types) and targets of interest, such as vegetation reflectance properties, in order to extrapolate the targets of interest beyond the sampling locations (Ploton et al., 2020). However, it has been reported that the predictive power of ML models in the literature is often evaluated using nonspatial cross-validation, which can lead to misleadingly confident interpretations of model accuracy and reliability where autocorrelation may act as a shortcut (Stock et al., 2023). Consequently, any inference of ecological determinism based on post-hoc interpretations of these models must be approached with extreme caution (Ploton et al., 2020). In practice, rigorous validation is essential, using resampling procedures such as holdout or (repeated) cross-validation, depending on sample size. These validation procedures should reflect the structure of the prediction task, taking into account

10.1029/2024EF004540

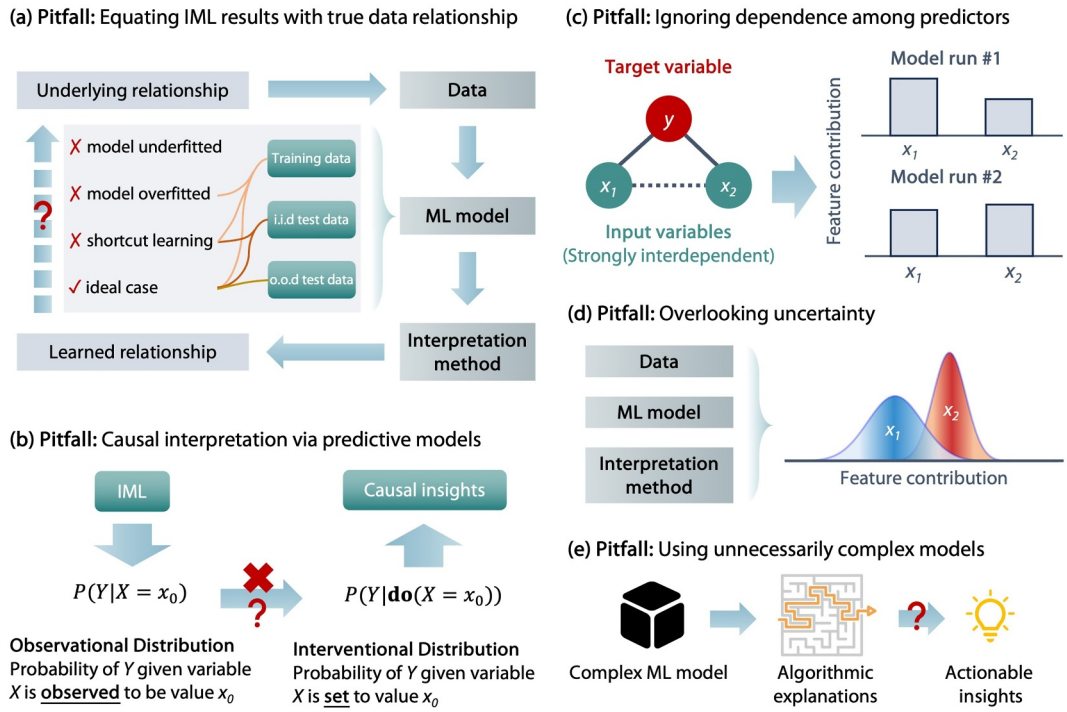

Figure 3. *Common pitfalls in geoscience interpretable machine learning (IML) applications*. (a) ML model training can result in underfitting, overfitting, shortcut learning, or successful capture of the underlying data generation process. These results can be reflected by sufficient model performance on training data, independent and identically distributed (i.i.d.) test data, and out-of-distribution (o.o.d.) test data, as indicated by the corresponding links in the diagram. (b) The difference between predictive and causal goals. The predictive model generally captures only the observational distribution of the data and cannot be equated with causal insights based on the interventional distribution. (c) Strongly interdependent input variables can lead to varying feature importance scores in different model runs, due to similar information about the target output. (d) Different methodological choices can lead to diverse insights, thereby introducing uncertainty into the interpretation process. (e) Complex models may accurately capture intricate data patterns, but the interpretations may be difficult for humans to intuitively understand, hindering the ability to gain actionable insights from the IML framework.

spatial or temporal prediction distances and out-of-sample estimation where appropriate (Brenning, 2022), since validation results as well as model interpretations will inevitably depend on the chosen resampling strategy (Meyer & Pebesma, 2022; Schratz et al., 2019; Sweet et al., 2023). Ideally, model interpretations should also be validated against out-of-sample data sets (e.g., independent data sets relevant to the study) to ensure that the insights they provide are not artifacts of the unique characteristics or biases present in the training data but reflect more general patterns and relationships. Overall, insights derived from IML should not be regarded as definitive interpretations of data truths, but rather as hypotheses that require further validation through additional analysis and experimentation.

4.2. Tendency of Causal Interpretation

In IML applications, a subtle yet significant risk is the often-unintentional misinterpretation of relationships derived from predictive models as causal in nature (Figure 3b). Standard supervised ML models are designed to exploit associations in the data rather than explicitly model causal relationships. Generally, predictive models focus primarily on understanding the observational conditional probability $p(Y|X = x_0)$ by inferring the probable values of Y when X is observed to be value x_0. Conversely, causal tasks focus on the interventional probability $p(Y|\mathbf{do}(X = x_0))$, which attempts to understand the effect of a change or intervention in X (e.g., setting it to x_0) on Y (Pearl & Mackenzie, 2018). For example, consider a flood prediction model that uses vegetation cover as one of its input variables. Such a model might perform well by exploiting the observed association between vegetation cover and certain flood processes (Calder & Aylward, 2006). However, this observed association within the predictive model does not inherently reveal the direct impact of interventions in vegetation cover (e.g., afforestation or deforestation) on flood events (Rogger et al., 2017). This is because the observed association may arise

from correlations between vegetation and climate characteristics or geomorphology, which also influence the distribution and characteristics of flood events. Moreover, when building predictive ML models, it is common to include as many explanatory variables as possible to maximize performance. However, this approach can be counterproductive when the goal is interpretation. For example, research has shown that IML methods used to identify influential variables and uncover underlying functional relationships in ecology are negatively affected by the inclusion of spurious variables (those that are correlated with, but not causally related to, the target variable) (Q. Yu et al., 2021). Therefore, when process understanding is important, it can be helpful to construct ML models using independent variables that have clear causal effects on response variables.

Typically, an important condition for a predictive model to yield a causal effect estimate is that its input variables are independent of unobserved confounders (i.e., variables that affect both the input and the model target). Otherwise, interpretations derived directly from a predictive model do not directly indicate whether a variable acts as a cause, an effect, or has no causal relationship with the target variable (Molnar et al., 2022). Despite the awareness that correlation does not imply causation, there is a tendency to interpret the results of IML methods from a causal perspective (Arif & MacNeil, 2022), especially if such an interpretation is consistent with pre-existing beliefs or theories. Recent literature suggests that predictive ML models have already been conflated with causality in ecological studies, where ML models are increasingly being misused for causal interpretations (Arif & MacNeil, 2022). When interpreting IML outputs, it is important to use language that accurately reflects the nature of these findings. For example, terms such as "associated with" may be more appropriate than "driven by" (Thapa et al., 2020). However, there remains the possibility that readers may interpret correlational statements as causal (Gershman & Ullman, 2023). Explicitly stating the limitations of the analysis and acknowledging the potential for alternative explanations or confounding factors can help readers understand the nature of the relationships presented.

In most cases, IML should not be considered a definitive source of causal knowledge. The challenge of causal discovery and inference remains an important open question in ML research (Runge et al., 2023). In general, a thorough investigation is needed to make explicit under which assumptions causal insights can be extracted from the interpretation of ML models (Janzing et al., 2020). Recently, there has been a growing interest in integrating causal inference concepts such as structural causal models, do-operators, and causal metrics into ML interpretation (e.g., Carloni et al., 2023; Reimers et al., 2020). For example, Heskes et al. (2020) proposed causal Shapley values, which extend the traditional Shapley value framework by explicitly incorporating interventional expectations to account for both direct and indirect contributions of a feature to the model's predictions. Similarly, the knockoff framework allows causal exploration with ML models by generating synthetic control variables to rigorously assess the importance of features, distinguishing between causally relevant features and correlated features (Popescu et al., 2021). In addition, innovations such as double ML (Chernozhukov et al., 2018) and causal ML (Tesch et al., 2023) are being explored in Earth science research. To robustly explore and validate causal relationships, it may be necessary to complement IML findings with additional causal inference frameworks, such as quasi-experimental approaches (Butsic et al., 2017) and time-series causal analysis (Runge et al., 2019).

4.3. Multicollinearity and Dependence Among Features

Another issue that often receives insufficient attention in IML applications is interdependence among features, where one or more features can be explained non-linearly by ML models using the other features, which is referred to as "concurvity" in some contexts (Wood, 2017). A widely known example of this is multicollinearity among input variables, where some features are strongly correlated with one another. In addition to exacerbating the risk of misattribution of causality discussed above, the problem of multicollinearity can also affect the reliability of IML results (Figure 3c). While this concern is well recognized in classical statistical analysis, for example, variance inflation factor (Mansfield & Helms, 1982), its importance seems to be less emphasized in the context of IML. This oversight may be due to the fact that ML models, even when trained on multicollinear data, are likely to retain predictive power, especially when the test data used have a similar dependence structure (Farrell et al., 2019). However, this predictive power does not negate the interpretive challenges posed by multicollinearity, especially when attempting to derive quantifiable insights for process understanding from a predictive model. This issue is particularly prevalent and critical in the geosciences, where variables often exhibit strong dependence and multicollinearity due to the interconnected nature of Earth systems. A case study in atmospheric chemistry demonstrated that correlated and dependent features can lead to spurious process-level

explanations, where chemical reactions can be wrongly attributed to fundamentally incorrect compounds (Silva & Keller, 2024).

Theoretically, ML models could arbitrarily assign importance or weight across highly correlated variables when making predictions because they carry similar information about the target variable. In this case, the importance of features may be spread across multiple features, suggesting a weak or negligible association with the response (Brenning, 2023), or it may show high variability and even directional shifts (Chan et al., 2022). Furthermore, the presence of multicollinearity can lead to unreliable interpretations, especially when using perturbation-based methods. When features are highly correlated, these perturbations can extrapolate into "uncharted" regions within the feature space that lie outside the observed joint distribution of the variables, leading to biased assessments of feature importance (Hooker et al., 2021).

A common and straightforward strategy to mitigate the effects of multicollinearity is to exclude highly correlated variables whose information may be redundant in feature selection (Katrutsa & Strijov, 2017). However, this can sometimes conflict with the goal of identifying underlying relationships based on a comprehensive set of as many relevant variables as possible, which may contain subtle but crucial information. For instance, two climate variables may be closely related, but may affect an ecological process differently under varying conditions. In this case, it is important to increase the diversity of the environment (e.g., varied climate regions, geographic conditions, and species diversity) for the variables. This increased diversity can help account for multifaceted relationships between variables, especially when certain correlations are actually dependent on other factors (Dormann et al., 2013). For example, the dependence between soil moisture and evapotranspiration is generally determined by water and energy availability, which varies with season and geographic location (Hsu & Dirmeyer, 2023). Moreover, where possible and appropriate, closely related variables may be grouped or transformed for collective or conditional interpretation, where their contributions can be considered more holistically, rather than attempting to separate the individual contributions of these variables (Brenning, 2023; Jiang et al., 2024). Krell et al. (2023) further suggest that models based on gridded geospatial data can be sensitive to the choice of grouping scheme, and thus it is beneficial to compare explanations from multiple grouping schemes for more accurate insights, as each may probe the model differently.

4.4. Uncertainty in Interpretations

Using interpretation to enhance the transparency of ML models may inadvertently create an illusion of certainty about their results. However, as highlighted earlier, these interpretations are subject to various uncertainties, including those inherent in the data, the structure and training processes of the ML model, and the specific assumptions and computations behind the interpretation methods (Figure 3d). For example, multiple distinct ML models with comparable performance may provide divergent explanations for the same set of data (i.e., model multiplicity (Breiman, 2001b) or equifinality)—how can we discern which explanation is the most accurate or valid? The different narratives offered by each model often stem from their unique approaches to processing and using the input data, including biases in feature selection (Strobl et al., 2007). Furthermore, while predictive accuracy is not typically a primary concern in the pursuit of process understanding, interpretations from poorly or unstably performing models are likely to be inherently unreliable (Murdoch et al., 2019). In many cases, applying different interpretation methods to a single model (Mamalakis, Barnes, & Ebert-Uphoff, 2022), or even applying the same interpretation method but with varying settings or hyperparameters (S. Müller et al., 2023), can lead to different results. The variance in the latter case can be largely due to the approximations used by the interpretation techniques. These approximations simplify complex mathematical models into forms that are more understandable and computationally manageable, but can vary with each computation when stochastic processes are involved. For instance, the LIME method constructs simpler, surrogate models based on perturbed samples to locally approximate the prediction function of complex models (Tulio Ribeiro et al., 2016). Consequently, the explanations provided by LIME are sensitive to changes in the number of perturbed samples (Bansal et al., 2020). Similarly, Monte Carlo integration methods are often used to approximate Shapley values, which are also subject to sampling variability (Goldwasser & Hooker, 2023; Štrumbelj & Kononenko, 2013).

For example, Hu et al. (2023) have compared 11 IML methods to gain process understanding of climate and crop interactions from crop yield prediction modeling and found divergent results among these methods. They advised that future studies should not uncritically rely on the variable importance rankings produced by a single IML method to draw definitive conclusions. In practice, it is advisable to consider approaches or strategies for

quantifying uncertainty in IML explanations, such as probabilistic and bootstrapping techniques. For example, Slack et al. (2020) proposed a Bayesian framework to generate probabilistic versions of LIME and SHAP, instead of pointwise estimates of feature importance. To account for various sources of uncertainty and enhance the robustness and reliability of interpretations, it may be beneficial to perform IML analysis repeatedly by resampling the data, using different subsets of data, varying initial random seeds in ML models, or applying multiple interpretation methods (e.g., Jiang et al., 2024; Labe & Barnes, 2021; W. Li et al., 2022). In addition, it is important to be aware of the assumptions, limitations, and potential weaknesses of the interpretation methods applied to realistic and complex geoscientific data sets. Recently, Mamalakis, Barnes, and Ebert-Uphoff (2022) developed synthetic attribution benchmark data sets specifically tailored for geoscience applications, providing a solid foundation for more falsifiable and rigorous research. Bommer et al. (2024) also introduced a suite of metrics to evaluate the effectiveness of different interpretation methods in climate research, such as robustness, faithfulness, randomization, complexity, and localization, to facilitate the selection of the most appropriate interpretation methods for both technically robust and contextually relevant applications.

4.5. Gap Between Complexity and Interpretability

The development of post-hoc interpretation methods has somewhat alleviated the long-standing trade-off between accuracy and interpretability of ML models (Murdoch et al., 2019). However, extracting scientific insights from ML models with complex structures remains a practical challenge (Figure 3e). Interpreting the internal mechanisms of complex, high-performing ML models in a human-understandable way often requires a degree of simplification that may obscure the subtle intricacies captured by the model. Moreover, even when interpretation methods accurately reflect the algorithmic functioning of ML models, the resulting explanations are not necessarily intuitive and aligned with human understanding in specialized domains (Ehsan et al., 2022). This mismatch between the computational logic of algorithms and human intuition can lead to misinterpretations, requiring thoughtful translation of algorithmic explanations into terms that are both accessible and relevant to the domain (Achtibat et al., 2023). In a study examining ozone mapping models, SHAP values revealed that the models placed more importance on geographical features such as absolute latitude and altitude than chemical factors like NOx emissions (Betancourt et al., 2022). The authors noted that this finding might appear counterintuitive, as ozone chemistry is typically expected to play a more significant role in such models. However, on the other hand, comparing these interpretations to existing knowledge can be fraught with cognitive biases that tend to reinforce existing theories or expectations and potentially overlook novel insights. For example, if a model suggests an unconventional factor as influential in climate change, it may be dismissed if it contradicts long-held beliefs, despite its potential validity. These challenges highlight the need to balance the advanced computational accuracy of complex ML models with the practical need for clear, concise, and actionable insights.

As noted previously, an iterative model building strategy is advocated, where complexity is incrementally increased and the interpretability of the model is continuously evaluated (Molnar et al., 2022). This method aims to find a sweet spot where the model achieves both high accuracy and meaningful interpretability. For example, a GAM and its geospatial extensions can serve as a gradual transition between linear models and complex ML models in this iterative process (Rudin, 2019; Wood, 2017). The additive structure of a GAM is specified prior to model fitting, allowing for the estimation of prescribed features or interaction effects. In addition to GAM implementations based on smoothing splines (Wood, 2017), tree-based GAM smoothers, such as Explainable Boosting Machines (Lou et al., 2013), can provide greater flexibility and robustness, especially in high-dimensional situations. Such additive models are often as accurate as state-of-the-art ML models (e.g., XGboost), while remaining inherently interpretable (Goetz et al., 2015).

Furthermore, as noted by Betancourt et al. (2022), the counterintuitive results for ozone attribution may arise because the purely data-driven model approach is inherently a posteriori and not process-oriented in any way, that is, scientific consistency was not enforced during the training process. These shortcomings highlight the value of hybrid (Reichstein et al., 2019) or differentiable modeling (Shen et al., 2023) strategies in Earth sciences that aim to be effective in creating inherently interpretable models, that is, models that follow a domain-specific set of constraints that make the reasoning processes understandable (Rudin et al., 2022). These strategies involve the integration of physical relationships or models into ML architectures (e.g., Jiang et al., 2020; Kraft et al., 2022; C. Wang et al., 2024). In this way, the complexity inherent in the data can be effectively managed by anchoring the models in well-established scientific principles, which helps constrain the models to plausible behaviors and thus reduces ambiguity in their explanations.

5. Conclusion and Outlook

The rapid development of AI and its subfield IML has opened new frontiers in various scientific disciplines, including the geosciences. However, amidst the rapid expansion in the use of IML, there has been both a tendency toward careless and superficial application and an underestimation of its much broader potential in the field. This study aims to address these issues, improve the accessibility and relevance of IML to a wider range of geoscientists and, more importantly, facilitate more effective and appropriate use of these innovative tools. In this paper, we specifically focus on the potential benefits of IML for process understanding in the geosciences. It is anticipated that IML will become an indispensable method for enhancing our current, often conceptual and qualitative understanding with quantifiable non-linear insights, and for generating innovative hypotheses with large data sets. In particular, IML is expected to play an important role in evaluating and revising existing process-based models. However, it is important to recognize that AI tools alone are not sufficient to drive progress in domain science, and to remain vigilant about the potential risks of scientific monocultures that AI-led science may foster (Messeri & Crockett, 2024). Rather than advocating a shift away from process-based modeling, we emphasize the complementary role of IML in addressing tasks that are challenging for traditional methods. While the current application of IML to understanding the complexities of the Earth system is in its early stages, its far-reaching implications are undeniable. We envision a future in which a broad spectrum of geoscientists benefit from the insights provided by IML, using it as an advanced analytical method in an era of abundant data to deepen our understanding of Earth's complex systems both directly and indirectly in the future.

This study presents a practical workflow with examples for geoscientists to effectively integrate IML into their research. Especially, we identify several potential pitfalls that are likely to be encountered when applying IML. We advocate cross-disciplinary collaboration between geoscientists, data scientists, and ML experts to tailor IML tools to specific geoscience needs, with a focus on causal and multifactorial process considerations, knowledge integration, and uncertainty quantification. In general, we argue for a pragmatic approach to these tools and their more thoughtful use in geoscientific research to ensure responsible knowledge production. While existing research has pioneered the use of IML, we recognize the need to be more cautious in drawing conclusions, especially in scenarios where rigorous validation is not possible. We encourage researchers to carefully evaluate the robustness of their results, taking into account the good practices we have suggested, before reporting them, with the goal of further solidifying the role of IML as a reliable and effective tool for advancing geoscientific research.

Data Availability Statement

No new data were created or analyzed in this study.

Acknowledgments
The authors acknowledge the contributions of Dr. Frances Davenport and Dr. Huan Wang for providing the data that facilitated the generation of the plots in Figure 2. This research was supported by the Carl Zeiss Foundation (Junior Research Group "Knowledge integration for spatio-temporal environmental modeling"). LS and JZ acknowledge the Helmholtz Initiative and Networking Fund (Young Investigator Group COMPOUNDX, Grant Agreement VH-NG-1537). Open Access funding enabled and organized by Projekt DEAL.

References

Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., & Lapuschkin, S. (2023). From attribution maps to human-understandable explanations through Concept Relevance Propagation. *Nature Machine Intelligence*, *5*(9), 1006–1019. https://doi.org/10.1038/s42256-023-00711-8

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/access.2018.2870052

Altmann, A., Tolosi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Arif, S., & MacNeil, M. A. (2022). Predictive models aren't for causal inference. *Ecology Letters*, *25*(8), 1741–1745. https://doi.org/10.1111/ele.14033

Bansal, N., Agarwal, C., & Nguyen, A. (2020). SAM: The sensitivity of attribution methods to hyperparameters. In *Paper presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.00870

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

Başağaoğlu, H., Chakraborty, D., Lago, C. D., Gutierrez, L., Şahinli, M. A., Giacomoni, M., et al. (2022). A review on interpretable and explainable artificial intelligence in hydroclimatic applications. *Water*, *14*(8), 1230. https://doi.org/10.3390/w14081230

Bau, D., Zhu, J. Y., Strobelt, H., Lapedriza, A., Zhou, B., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, *117*(48), 30071–30078. https://doi.org/10.1073/pnas.1907375117

Betancourt, C., Stomberg, T. T., Edrich, A.-K., Patnala, A., Schultz, M. G., Roscher, R., et al. (2022). Global, high-resolution mapping of tropospheric ozone – Explainable machine learning and impact of uncertainties. *Geoscientific Model Development*, *15*(11), 4331–4354. https://doi.org/10.5194/gmd-15-4331-2022

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, *5*(5), 277–280. https://doi.org/10.1038/s42254-023-00581-4

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., et al. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, *13*(2), e1484. https://doi.org/10.1002/widm.1484

Bommer, P. L., Kretschmer, M., Hedström, A., Bareeva, D., & Höhne, M. M. C. (2024). Finding the right XAI method — A guide for the evaluation and ranking of explainable AI methods in climate science. *Artificial Intelligence for the Earth Systems*, *3*(3), e230074. https://doi.org/10.1175/aies-d-23-0074.1

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215. https://doi.org/10.1214/ss/1009213726

Brenning, A. (2022). Spatial machine-learning model diagnostics: A model-agnostic distance-based approach. *International Journal of Geographical Information Science*, *37*(3), 584–606. https://doi.org/10.1080/13658816.2022.2131789

Brenning, A. (2023). Interpreting machine-learning models in transformed feature space with an application to remote-sensing classification. *Machine Learning*, *112*(9), 3455–3471. https://doi.org/10.1007/s10994-023-06327-8

Brenning, A., Schwinn, M., Ruiz-Páez, A. P., & Muenchow, J. (2015). Landslide susceptibility near highways is increased by 1 order of magnitude in the Andes of southern Ecuador, Loja province. *Natural Hazards and Earth System Sciences*, *15*(1), 45–57. https://doi.org/10.5194/nhess-15-45-2015

Butsic, V., Lewis, D. J., Radeloff, V. C., Baumann, M., & Kuemmerle, T. (2017). Quasi-experimental methods enable stronger inferences from observational data in ecology. *Basic and Applied Ecology*, *19*, 1–10. https://doi.org/10.1016/j.baae.2017.01.005

Calder, I. R., & Aylward, B. (2006). Forest and floods. *Water International*, *31*(1), 87–99. https://doi.org/10.1080/02508060608691918

Carloni, G., Berti, A., & Colantonio, S. (2023). The role of causality in explainable artificial intelligence. Retrieved from https://arxiv.org/abs/2309.09901

Chakraborty, D., Başağaoğlu, H., Gutierrez, L., & Mirchi, A. (2021). Explainable AI reveals new hydroclimatic insights for ecosystem-centric groundwater management. *Environmental Research Letters*, *16*(11), 114024. https://doi.org/10.1088/1748-9326/ac2fde

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics*, *10*(8), 1283. https://doi.org/10.3390/math10081283

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA*. https://doi.org/10.1145/2939672.2939785

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Chu, H., Baldocchi, D. D., John, R., Wolf, S., & Reichstein, M. (2017). Fluxes all of the time? A primer on the temporal representativeness of FLUXNET. *Journal of Geophysical Research: Biogeosciences*, *122*(2), 289–307. https://doi.org/10.1002/2016jg003576

Davenport, F. V., & Diffenbaugh, N. S. (2021). Using machine learning to analyze physical causes of climate change: A case study of U.S. Midwest extreme precipitation. *Geophysical Research Letters*, *48*(15), e2021GL093787. https://doi.org/10.1029/2021gl093787

de Burgh-Day, C. O., & Leeuwenburg, T. (2023). Machine learning for numerical weather and climate modelling: A review. *Geoscientific Model Development*, *16*(22), 6433–6477. https://doi.org/10.5194/gmd-16-6433-2023

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. In B. Moseley & L. Krischer (Eds.), *Machine learning in geosciences* (Vol. 61, pp. 1–55). Elsevier. https://doi.org/10.1016/bs.agph.2020.08.002

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., et al. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1–33. https://doi.org/10.1145/3561048

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé Iii, H., et al. (2022). Human-centered Explainable AI (HCXAI): Beyond opening the black-box of AI. In *Paper presented at the CHI Conference on Human Factors in Computing Systems Extended Abstracts, New Orleans, LA, USA*. https://doi.org/10.1145/3491101.3503727

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Fang, K., Kifer, D., Lawson, K., Feng, D., & Shen, C. (2022). The data synergy effects of time-series deep learning models in hydrology. *Water Resources Research*, *58*(4), e2021WR029583. https://doi.org/10.1029/2021wr029583

Farrell, A., Wang, G., Rush, S. A., Martin, J. A., Belant, J. L., Butler, A. B., & Godwin, D. (2019). Machine learning of large-scale spatial distributions of wild turkeys with high-dimensional environmental data. *Ecology and Evolution*, *9*(10), 5938–5949. https://doi.org/10.1002/ece3.5177

Feldhoff, J. H., Lange, S., Volkholz, J., Donges, J. F., Kurths, J., & Gerstengarbe, F.-W. (2014). Complex networks for climate model evaluation with application to statistical versus dynamical modeling of South American climate. *Climate Dynamics*, *44*(5–6), 1567–1581. https://doi.org/10.1007/s00382-014-2182-9

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Gershman, S. J., & Ullman, T. D. (2023). Causal implicatures from correlational statements. *PLoS One*, *18*(5), e0286067. https://doi.org/10.1371/journal.pone.0286067

Gevaert, C. M. (2022). Explainable AI for earth observation: A review including societal and regulatory perspectives. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102869. https://doi.org/10.1016/j.jag.2022.102869

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., Kagal, L., & IEEE. (2018). Explaining explanations: An overview of interpretability of machine learning. In *Paper presented at the 5th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA), Turin, ITALY*. https://doi.org/10.1109/dsaa.2018.00018

Gnann, S., Reinecke, R., Stein, L., Wada, Y., Thiery, W., Müller Schmied, H., et al. (2023). Functional relationships reveal differences in the water cycle representation of global water models. *Nature Water*, *1*(12), 1079–1090. https://doi.org/10.1038/s44221-023-00160-y

Goetz, J. N., Brenning, A., Petschko, H., & Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & Geosciences*, *81*, 1–11. https://doi.org/10.1016/j.cageo.2015.04.007

Goldwasser, J., & Hooker, G. (2023). Stabilizing estimates of Shapley values with control variates. Retrieved from https://arxiv.org/abs/2310.07672

Good, P. I., & Hardin, J. W. (2012). *Common errors in statistics (and how to avoid them)*. John Wiley & Sons. https://doi.org/10.1002/9781118360125

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J. P., Yordanova, K., Vered, M., et al. (2023). A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*, *56*(4), 3473–3504. https://doi.org/10.1007/s10462-022-10256-8

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? Retrieved from https://arxiv.org/abs/2207.08815

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI-Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

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

Ham, Y. G., Kim, J. H., Min, S. K., Kim, D., Li, T., Timmermann, A., & Stuecker, M. F. (2023). Anthropogenic fingerprints in daily precipitation revealed by deep learning. *Nature*, *622*(7982), 301–307. https://doi.org/10.1038/s41586-023-06474-x

Hedström, A., Weber, L., Bareeva, D., Krakowczyk, D., Motzkus, F., Samek, W., et al. (2024). Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, *24*(1), 1339–1349. https://doi.org/10.5555/3648699.3648733

Heskes, T., Bucur, I. G., Sijben, E., & Claassen, T. (2020). Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Paper presented at the Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada*. https://doi.org/10.5555/3495724.3496125

Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, *31*(6), 82. https://doi.org/10.1007/s11222-021-10057-z

Hsu, H., & Dirmeyer, P. A. (2023). Soil moisture-evaporation coupling shifts into new gears under increasing CO(2). *Nature Communications*, *14*(1), 1162. https://doi.org/10.1038/s41467-023-36794-5

Hu, T., Zhang, X., Bohrer, G., Liu, Y., Zhou, Y., Martin, J., et al. (2023). Crop yield prediction via explainable AI and interpretable machine learning: Dangers of black box models for evaluating climate change impacts on crop yield. *Agricultural and Forest Meteorology*, *336*, 109458. https://doi.org/10.1016/j.agrformet.2023.109458

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., & Saynisch-Wagner, J. (2021). Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, *3*(8), 667–674. https://doi.org/10.1038/s42256-021-00374-3

Jägermeyr, J., Muller, C., Ruane, A. C., Elliott, J., Balkovic, J., Castillo, O., et al. (2021). Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. *Nature Food*, *2*(11), 873–885. https://doi.org/10.1038/s43016-021-00400-y

Janzing, D., Minorics, L., & Bloebaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. In *Paper presented at the Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research*. https://doi.org/10.48550/arXiv.1910.13413

Jiang, S., Bevacqua, E., & Zscheischler, J. (2022). River flooding mechanisms and their changes in Europe revealed by explainable machine learning. *Hydrology and Earth System Sciences*, *26*(24), 6339–6359. https://doi.org/10.5194/hess-26-6339-2022

Jiang, S., Tarasova, L., Yu, G., & Zscheischler, J. (2024). Compounding effects in flood drivers challenge estimates of extreme river floods. *Science Advances*, *10*(13), eadl4005. https://doi.org/10.1126/sciadv.adl4005

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, *47*(13), e2020GL088229. https://doi.org/10.1029/2020gl088229

Jiang, S., Zheng, Y., Wang, C., & Babovic, V. (2022). Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments. *Water Resources Research*, *58*(1), e2021WR030185. https://doi.org/10.1029/2021wr030185

Katrutsa, A., & Strijov, V. (2017). Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems With Applications*, *76*, 1–11. https://doi.org/10.1016/j.eswa.2017.01.048

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., et al. (2022). Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, *49*(17), e2022GL099368. https://doi.org/10.1029/2022gl099368

Kraft, B., Jung, M., Körner, M., Koirala, S., & Reichstein, M. (2022). Towards hybrid modeling of the global hydrological cycle. *Hydrology and Earth System Sciences*, *26*(6), 1579–1614. https://doi.org/10.5194/hess-26-1579-2022

Kraft, B., Jung, M., Körner, M., Requena Mesa, C., Cortés, J., & Reichstein, M. (2019). Identifying dynamic memory effects on vegetation state using recurrent neural networks. *Frontiers in Big Data*, *2*, 31. https://doi.org/10.3389/fdata.2019.00031

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., & Klambauer, G. (2019). NeuralHydrology – Interpreting LSTMs in hydrology. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 347–362). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_19

Krell, E., Kamangir, H., Collins, W., King, S. A., & Tissot, P. (2023). Aggregation strategies to improve XAI for geoscience models that use correlated, high-dimensional rasters. *Environmental Data Science*, *2*, e45. https://doi.org/10.1017/eds.2023.39

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. Retrieved from https://arxiv.org/abs/2202.01602

Labe, Z. M., & Barnes, E. A. (2021). Detecting climate signals using explainable AI with single-forcing large ensembles. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2021MS002464. https://doi.org/10.1029/2021ms002464

Lapuschkin, S., Waldchen, S., Binder, A., Montavon, G., Samek, W., & Muller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), 1096. https://doi.org/10.1038/s41467-019-08987-4

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., et al. (2022). Hydrological concept formation inside long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *26*(12), 3079–3101. https://doi.org/10.5194/hess-26-3079-2022

Li, W., Migliavacca, M., Forkel, M., Denissen, J. M. C., Reichstein, M., Yang, H., et al. (2022). Widespread increasing vegetation sensitivity to soil moisture. *Nature Communications*, *13*(1), 3959. https://doi.org/10.1038/s41467-022-31667-9

Li, X., Feng, M., Ran, Y., Su, Y., Liu, F., Huang, C., et al. (2023). Big Data in Earth system science and progress towards a digital twin. *Nature Reviews Earth & Environment*, *4*(5), 319–332. https://doi.org/10.1038/s43017-023-00409-w

Lopez-Gomez, I., McGovern, A., Agrawal, S., & Hickey, J. (2023). Global extreme heat forecasting using neural weather models. *Artificial Intelligence for the Earth Systems*, *2*(1), e220035. https://doi.org/10.1175/aies-d-22-0035.1

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Paper presented at the Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA*. https://doi.org/10.1145/2487575.2487579

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Paper presented at the Advances in Neural Information Processing Systems, Long Beach, CA, USA*. https://doi.org/10.5555/3295222.3295230

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Madden, R. A., & Williams, J. (1978). The correlation between temperature and precipitation in the United States and Europe. *Monthly Weather Review*, *106*(1), 142–147. https://doi.org/10.1175/1520-0493(1978)106<0142:Tcbtap>2.0.Co;2

Mamalakis, A., Barnes, E. A., & Ebert-Uphoff, I. (2022). Investigating the fidelity of Explainable Artificial Intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, *1*(4), e220012. https://doi.org/10.1175/aies-d-22-0012.1

Mamalakis, A., Barnes, E. A., & Ebert-Uphoff, I. (2023). Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. *Artificial Intelligence for the Earth Systems*, *2*(1), e220058. https://doi.org/10.1175/aies-d-22-0058.1

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2022). Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond explainable AI* (pp. 315–339). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_16

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician*, *36*(3), 158–160. https://doi.org/10.2307/2683167

Mayer, K. J., & Barnes, E. A. (2021). Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophysical Research Letters*, *48*(10), e2020GL092092. https://doi.org/10.1029/2020gl092092

McGovern, A., Jergensen, G. E., Lagerquist, R., & Smith, T. (2019). Classifying convective storms using machine learning. *Weather and Forecasting*, *35*(2), 537–559. https://doi.org/10.1175/waf-d-19-0170.1

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

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Meyer, H., & Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, *13*(1), 2208. https://doi.org/10.1038/s41467-022-29838-9

Mishra, S., Dutta, S., Long, J., & Magazzeni, D. (2021). A survey on the robustness of feature importance and counterfactual explanations. Retrieved from https://arxiv.org/abs/2111.00358

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning - A brief history, state-of-the-art and challenges. In *Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. https://doi.org/10.1007/978-3-030-65965-3_28

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., et al. (2022). General pitfalls of model-agnostic interpretation methods for machine learning models. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond explainable AI: International Workshop, held in conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and extended papers* (pp. 39–68). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_4

Müller, C., Jägermeyr, J., Franke, J. A., Ruane, A. C., Balkovic, J., Ciais, P., et al. (2024). Substantial differences in crop yield sensitivities between models call for functionality-based model evaluation. *Earth's Future*, *12*(3), e2023EF003773. https://doi.org/10.1029/2023ef003773

Müller, S., Toborek, V., Beckh, K., Jakobs, M., Bauckhage, C., & Welke, P. (2023). An empirical evaluation of the Rashomon effect in explainable machine learning. In *Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Turin, Italy*. https://doi.org/10.1007/978-3-031-43418-1_28

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., et al. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, *55*(13s), 1–42. https://doi.org/10.1145/3583558

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*(3), e2020WR028091. https://doi.org/10.1029/2020wr028091

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384. https://doi.org/10.2307/2344614

Nowack, P., Runge, J., Eyring, V., & Haigh, J. D. (2020). Causal networks for climate model evaluation and constrained projections. *Nature Communications*, *11*(1), 1415. https://doi.org/10.1038/s41467-020-15195-y

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7. https://doi.org/10.23915/distill.00007

O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., et al. (2016). The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, *9*(9), 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (1st ed.). Basic Books Inc. https://doi.org/10.5555/3238230

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, *11*(1), 4540. https://doi.org/10.1038/s41467-020-18321-y

Popescu, O.-I., Shadaydeh, M., & Denzler, J. (2021). Counterfactual generation with knockoffs. Retrieved from https://arxiv.org/abs/2102.00951

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, *27*(3), 221–234. https://doi.org/10.1016/S0020-7373(87)80053-6

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

Reimers, C., Runge, J., & Denzler, J. (2020). Determining the relevance of features for deep neural networks. In *Paper presented at the computer vision – ECCV 2020: 16th European Conference, Glasgow, UK*. https://doi.org/10.1007/978-3-030-58574-7_20

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? In *Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA*. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Paper presented at the Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, Louisiana, USA*. https://doi.org/10.1609/aaai.v32i1.11491

Rogger, M., Agnoletti, M., Alaoui, A., Bathurst, J. C., Bodner, G., Borga, M., et al. (2017). Land use change impacts on floods at the catchment scale: Challenges and opportunities for future research. *Water Resources Research*, *53*(7), 5209–5219. https://doi.org/10.1002/2017WR020723

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020a). Explain it to me – Facing remote sensing challenges in the bio- and geosciences with explainable machine learning. In *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*. V-3-2020 (pp. 817–824). https://doi.org/10.5194/isprs-annals-V-3-2020-817-2020

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020b). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, *8*, 42200–42216. https://doi.org/10.1109/access.2020.2976199

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., Chen, C. F., Chen, Z., Huang, H. Y., Semenova, L., & Zhong, C. D. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*(none), 1–85. https://doi.org/10.1214/21-Ss133

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., et al. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*(1), 2553. https://doi.org/10.1038/s41467-019-10105-3

Runge, J., Gerhardus, A., Varando, G., Eyring, V., & Camps-Valls, G. (2023). Causal inference for time series. *Nature Reviews Earth & Environment*, *4*(7), 487–505. https://doi.org/10.1038/s43017-023-00431-y

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2020). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, *44*(2), 199–205. https://doi.org/10.1111/ecog.05360

Saha, A., Basu, S., & Datta, A. (2021). Random forests for spatially dependent data. *Journal of the American Statistical Association*, *118*(541), 665–683. https://doi.org/10.1080/01621459.2021.1950003

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, *2023*, 1–59. https://doi.org/10.1007/s10618-022-00867-8

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., et al. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, *4*(8), 552–567. https://doi.org/10.1038/s43017-023-00450-9

Silva, S. J., & Keller, C. A. (2024). Limitations of XAI methods for process-level understanding in the atmospheric sciences. *Artificial Intelligence for the Earth Systems*, *3*(1), e230045. https://doi.org/10.1175/aies-d-23-0045.1

Sivapalan, M., & Blöschl, G. (2017). The growth of hydrological understanding: Technologies, ideas, and societal needs shape the field. *Water Resources Research*, *53*(10), 8137–8146. https://doi.org/10.1002/2017wr021396

Slack, D., Hilgard, S., Singh, S., & Lakkaraju, H. (2020). Reliable post hoc explanations: Modeling uncertainty in explainability. Retrieved from https://arxiv.org/abs/2008.05030

Stock, A., Gregr, E. J., & Chan, K. M. A. (2023). Data leakage jeopardizes ecological applications of machine learning. *Nature Ecology & Evolution*, *7*(11), 1743–1745. https://doi.org/10.1038/s41559-023-02162-1

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Štrumbelj, E., & Kononenko, I. (2013). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Paper presented at the Proceedings of the 34th International Conference on Machine Learning, Sydney*. https://doi.org/10.5555/3305890.3306024

Sweet, L.-B., Müller, C., Anand, M., & Zscheischler, J. (2023). Cross-validation strategy impacts the performance and interpretation of machine learning models. *Artificial Intelligence for the Earth Systems*, *2*(4), e230026. https://doi.org/10.1175/aies-d-23-0026.1

Tesch, T., Kollet, S., & Garcke, J. (2023). Causal deep learning models for studying the Earth system. *Geoscientific Model Development*, *16*(8), 2149–2166. https://doi.org/10.5194/gmd-16-2149-2023

Thapa, D. K., Visentin, D. C., Hunt, G. E., Watson, R., & Cleary, M. (2020). Being honest with causal language in writing for publication. *Journal of Advanced Nursing*, *76*(6), 1285–1288. https://doi.org/10.1111/jan.14311

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

Touzé-Peiffer, L., Barberousse, A., & Le Treut, H. (2020). The Coupled Model Intercomparison Project: History, uses, and structural effects on climate research. *WIREs Climate Change*, *11*(4), e648. https://doi.org/10.1002/wcc.648

Tulio Ribeiro, M., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. Retrieved from https://arxiv.org/abs/1606.05386

van Oldenborgh, G. J., van der Wiel, K., Kew, S., Philip, S., Otto, F., Vautard, R., et al. (2021). Pathways and pitfalls in extreme event attribution. *Climatic Change*, *166*(1–2), 13. https://doi.org/10.1007/s10584-021-03071-7

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Retrieved from https://arxiv.org/abs/1711.00399

Wang, C., Jiang, S., Zheng, Y., Han, F., Kumar, R., Rakovec, O., & Li, S. (2024). Distributed hydrological modeling with physics-encoded deep learning: A general framework and its application in the Amazon. *Water Resources Research*, *60*(4), e2023WR036170. https://doi.org/10.1029/2023wr036170

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*(7972), 47–60. https://doi.org/10.1038/s41586-023-06221-2

Wang, H., Yan, S., Ciais, P., Wigneron, J. P., Liu, L., Li, Y., et al. (2022). Exploring complex water stress-gross primary production relationships: Impact of climatic drivers, main effects, and interactive effects. *Global Change Biology*, *28*(13), 4110–4123. https://doi.org/10.1111/gcb.16201

Wang, R., Kim, J. H., & Li, M. H. (2021). Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. *Science of the Total Environment*, *761*, 144057. https://doi.org/10.1016/j.scitotenv.2020.144057

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): Project framework. *Proceedings of the National Academy of Sciences*, *111*(9), 3228–3232. https://doi.org/10.1073/pnas.1312330110

Wood, S. N. (2017). Generalized additive models. https://doi.org/10.1201/9781315370279

Xu, H., Yu, H., Xu, B., Wang, Z., Wang, F., Wei, Y., et al. (2023). Machine learning coupled structure mining method visualizes the impact of multiple drivers on ambient ozone. *Communications Earth & Environment*, *4*(1), 265. https://doi.org/10.1038/s43247-023-00932-0

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, *1168*(2), 022C022. https://doi.org/10.1088/1742-6596/1168/2/022022

Yu, L., Wang, S., & Lai, K. K. (2006). An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering*, *18*(2), 217–230. https://doi.org/10.1109/tkde.2006.22

Yu, Q., Ji, W., Prihodko, L., Ross, C. W., Anchang, J. Y., & Hanan, N. P. (2021). Study becomes insight: Ecological learning from machine learning. *Methods in Ecology and Evolution*, *12*(11), 2117–2128. https://doi.org/10.1111/2041-210X.13686

Zhong, X., Gallagher, B., Liu, S., Kailkhura, B., Hiszpanski, A., & Han, T. Y.-J. (2022). Explainable machine learning in materials science. *Npj Computational Materials*, *8*(1), 204. https://doi.org/10.1038/s41524-022-00884-7

Zhou, Z., & Hooker, G. (2021). Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data*, *15*(2), 1–21. https://doi.org/10.1145/3429445

Zhu, J.-J., Yang, M., & Ren, Z. J. (2023). Machine learning in environmental research: Common pitfalls and best practices. *Environmental Science & Technology*, *57*(46), 17671–17689. https://doi.org/10.1021/acs.est.3c00026

Zscheischler, J., & Seneviratne, S. I. (2017). Dependence of drivers affects risks associated with compound events. *Science Advances*, *3*(6), e1700263. https://doi.org/10.1126/sciadv.1700263

Zscheischler, J., Westra, S., van den Hurk, B. J. J. M., Seneviratne, S. I., Ward, P. J., Pitman, A., et al. (2018). Future climate risk from compound events. *Nature Climate Change*, *8*(6), 469–477. https://doi.org/10.1038/s41558-018-0156-3