Bridging the Gap: Explainable AI for Advancing Earth System Science

Feini Huang[1,2,3], Shijie Jiang[2,3], Wei Shangguan[1*], Lu Li[1], Ye Zhang[1], Ruqing Zhang[4], Qingliang Li[5], Danxi Li[1], Yongjiu Dai[1]

[1]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies, School of Atmospheric Sciences, Sun Yat-Sen University, Zhuhai 519082, China.

[2]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena 07745, Germany.

[3]ELLIS Unit Jena, Jena 07745, Germany

[4]Shanwei Meteorological Service, Shanwei 516600, China

[5]College of Computer Science and Technology, Changchun Normal University, Changchun 130032, China.

* Corresponding author (email: shgwei@mail.sysu.edu.cn)

Abstract

Artificial Intelligence (AI) holds transformative potential for Earth System Science (ESS), yet its adoption is hindered by the interpretability gap of black-box models. Explainable AI (XAI) addresses this limitation by enhancing model transparency and enabling human-AI collaboration. This review provides an accessible introduction to XAI for ESS researchers, aiming to promote broader AI adoption by clarifying interpretability and addressing common misconceptions. We survey core XAI methodologies and their applications across ESS, highlighting their roles in model interpretation, model development, and scientific discovery. We further identify key challenges, including methodological limitations, compatibility issues, inadequate validation, and propose integrative solutions. To bridge the XAI-ESS gap, we advocate for co-designed frameworks that incorporate domain knowledge, causal inference, and human-centered interfaces. We anticipate that future advances will yield more faithful, efficient, and deeply integrated XAI tools, ultimately strengthening ESS research through reliable AI-enabled insights.

Keywords Explainable artificial intelligence, Earth system science, Interpretability, Machine learning

1. Introduction

Earth System Science (ESS) is an interdisciplinary field that seeks to understand the interactions among the atmosphere, hydrosphere, biosphere, geosphere, and human activities as a coupled, dynamic system (Schellnhuber, 1999; Steffen et al., 2020). ESS exhibits nonlinear feedback loops, multi-scale processes (from microbial interactions to global climate patterns), and emergent behaviors that defy simple mechanistic explanations (Heimann and Reichstein, 2008). Moreover, the sheer volume and heterogeneity of Earth observation data, spanning satellite remote sensing, in-situ measurements, and model simulations exacerbates difficulties in extracting actionable insights (Mahecha et al., 2020; Montero et al., 2024; Rojas et al., 2024; Vance et al., 2024). These challenges have historically constrained the depth of domain-specific knowledge, leaving critical gaps in our understanding of processes such as cloud-aerosol interactions (Li T et al., 2025), permafrost carbon feedbacks (Song C et al., 2024), and regional hydrological extremes (McMillan et al., 2025).

Artificial Intelligence (AI) and Machine/Deep Learning (ML/DL) have revolutionized ESS by enabling data-driven discovery at unprecedented scales (Reichstein et al., 2019). These techniques excel at identifying patterns in high-dimensional, noisy datasets, making them invaluable for tasks where traditional physics-based models struggle. Notable applications include AI-based weather models that surpass numerical forecasts in precipitation nowcasting (Sun T et al., 2022), high-resolution soil mapping through multi-source data fusion (Wadoux et al., 2020; Khose and Mailapalli, 2024), and automated cloud regime classification (Zhang J et al., 2018; Segal-Rozenhaimer et al., 2020). However, the critical nature of ESS applications from weather warnings to long-term climate policy precludes treating ML/DL as a black box. Interpretability is essential for ensuring model robustness, accountability, and the generation of actionable scientific insights (McGovern et al., 2023, 2024; Eyring et al., 2024; Robert Maier et al., 2024; Gilbert and Zengler 2025).

Explainable AI (XAI) has emerged as a vital discipline to enhance the interpretability of complex ML/DL models. Methods such as Class Activation Mapping (CAM), SHapley Additive exPlanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME) help elucidate the mechanisms underlying model predictions by identifying influential features, causal drivers, and underlying relationships. Within ESS, the role of XAI extends beyond model transparency: it serves to validate physical consistency, ensuring that AI-derived outputs conform to established dynamic laws (Lyu and Yong, 2025; Mallik et al., 2025). Furthermore, XAI provides mechanistic insights and aids in scientific discovery (Ham et al., 2023; Li W et al., 2022; Peng Z et al., 2024; Novielli et al., 2025).

A persistent disconnect exists between the ESS and XAI communities. XAI-ESS is a rapidly growing field, yet it remains at an early stage of development. The adoption of XAI-ESS often remains superficial, with tools applied without sufficient adaptation to domain-specific challenges. ESS scientists may lack the technical background to critically evaluate XAI methodologies. Conversely, XAI developers seldom prioritize ESS requirements, such as handling spatiotemporal dependencies and non-stationary data distributions. This misalignment introduces uncertainties: (a) overinterpreting post-hoc explanations by equating feature importance with physical causality (Aas et al., 2021; Hooker et al., 2021; Krell et al., 2025) or overlooking spatial auto-correlation artifacts (Chen C et al., 2024; Ke E et al., 2025); (b) generating explanations not readily actionable for scientific discovery (Cho and Ackom 2025); (c) obtaining unreliable insights from XAI when models exhibit instability under distribution shifts; and (d) neglecting domain constraints. Without bridging this gap, the promise of XAI in ESS risks being undermined by unverified assumptions and unreliable outputs.

In this review, our primary goal is to elucidate XAI methodologies for the broader ESS community by providing a comprehensive overview of their current applications. We aim not only to highlight their remarkable achievements in advancing ESS development, but also to explore the challenges and opportunities that lie ahead. Specifically, this paper is organized as follows:

(1) **Section 2** provides a brief technical overview of classical and common XAI methods, along with a comparative analysis of their strengths and weaknesses.

(2) **Section 3** presents a comprehensive summary of the diverse applications of XAI to different goals in the ESS, detailing how XAI helps to communicate model decisions, diagnose and improve models, and uncover scientific insights.

(3) **Section 4** summarizes the typical challenges encountered in integrating XAI into the ESS, as well as possible solutions aimed at efficient and effective use.

(4) **Section 5** examines the future outlook for XAI in the ESS, exploring expected developments, emerging trends, and potential areas of research and application that have not yet been explored.

We outline recommended reading priorities in **Table 1**, tailored to diverse backgrounds of practitioners and experts across ESS disciplines.

Table 1 Recommended Reading Priorities for Different Audiences.

ESS researchers who have different objectives for the review	Reading priorities (sections)
Newcomers considering XAI in their research	2, 3.1 and 3.5, Appendix A and B, Table S1 and S2
Forecasters who want to evaluate the quality of ML/DL forecasting by XAI	3.1, 3.2, 4.1
Policymakers and regulators who want trustworthy and responsible AI in climate policy making	3.1, 3.2, 4.1
ML/DL developers who need XAI to diagnose and debug the AI model, or even use XAI to guide AI model building	2, 3.1, 3.3, 3.5, 4
ESS modelers who want to discover valuable underlying processes in ESS from large data using ML	2, 3.1, 3.4, 3.5, 4, 5
Researchers often explaining complex phenomena (e.g., extreme hazards, climate attribution, and carbon fingerprint) with observations and mechanisms	2, 3.4, 4, 5
XAI-ESS experts interested in the challenges and future perspectives of XAI in ESS	4, 5

2. XAI Theory

2.1 Definition, Approaches and Interpretative Forms

Interpretability emphasizes the audience's ability to comprehend the rationale behind AI decisions (Biran and Cotton 2017; Gunning et al., 2019, 2021; Murdoch et al., 2019; Minh et al., 2022). We refine the definition of XAI as: "methods that provide an audience with additional insights into an AI model's internal logic, beyond the mere outputs", as explanations should cater to the diverse audiences (Barredo Arrieta et al., 2020). The XAI methods and their taxonomies reviewed here are outlined in **Table S1**, **Appendix A** and **Figure S1**. **Figure 1** illustrates several strategies (see **Section 2.2**) typically used to explain black-box ML/DL models. **Figure 2** summarizes key XAI interpretative forms exemplified in soil moisture modeling: feature importance ranks variable contributions; pixel importance visualizes critical spatial areas; sequence importance identifies pivotal time periods; feature effect analysis reveals prediction responses to inputs; rule explanation translates decisions into logical statements; and model representation techniques delve into the internal mechanisms of ML/DL models.

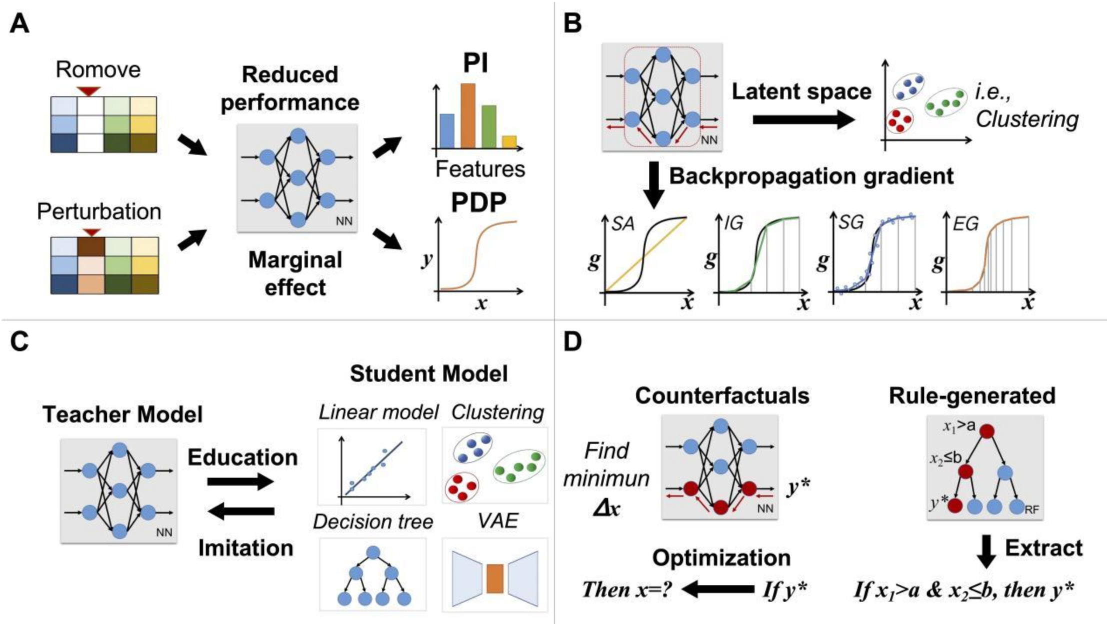

Figure 1 Approaches for explaining black-box AI models. (A) Perturbation-based methods explain models by perturbing their inputs and observing the resulting changes in the output, focusing on how inputs affect outputs. (B) Model representations that are predominantly model-specific, extracting relevant information directly from the model or data (e.g., visual attention, cell- and layer-level gradients, latent representations). This information is highly valuable but can be difficult to interpret. (C) Surrogate model methods involve constructing an inherently interpretable model to mimic the behavior of a complex ML/DL model. (D) Contrastive example methods often provide counterfactuals and rule-based explanations by comparisons. The abbreviations are: NN: Neural Networks; RF: Random Forest; PI: Permutation Importance; PDP: Partial Dependence Plots; SA: Saliency maps; IG: Integrated Gradients; SG: Smooth Gradients; EG: Expected Gradients; VAE: Variational Autoencoder.

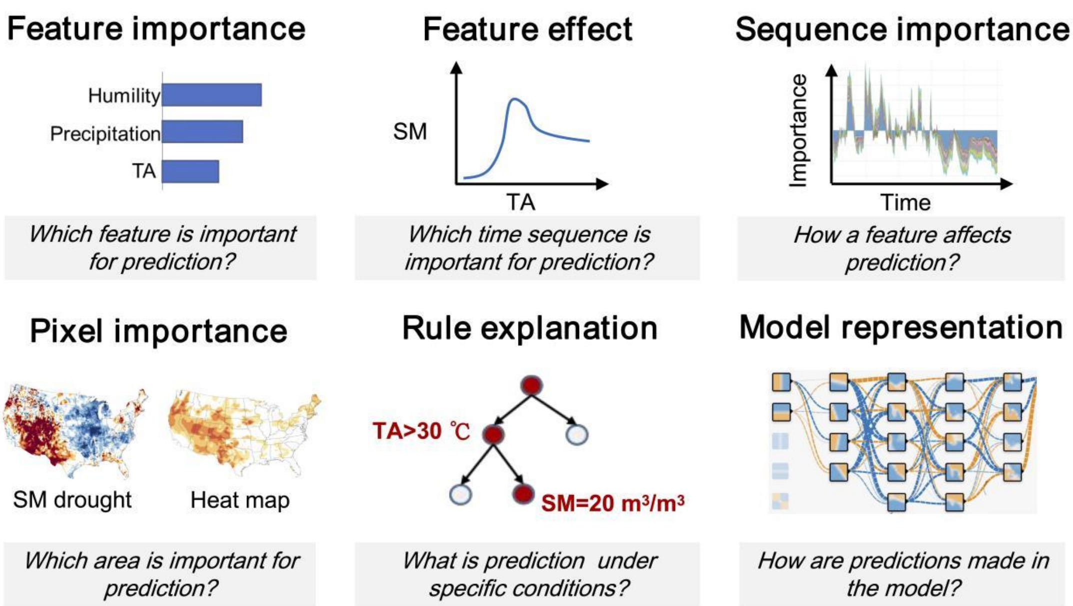

Figure 2 Interpretative forms of XAI and the questions they address. This example uses Soil Moisture (SM) prediction to demonstrate how different XAI approaches can explain the impact of Air Temperature (TA) on predictions in distinct ways.

2.2 Explaining Black-box AI Models

2.2.1 Perturbation-based Methods

These methods interpret a model by systematically modifying input features and observing the corresponding changes in model predictions with others unchanged.

(1) Feature Importance Measures

These techniques quantify each feature's contribution to a model's predictions. Permutation Importance (PI, **Fig. 1A**; Fisher et al., 2021) measures contribution by randomly permuting feature values and measuring the resulting increase in prediction error. Similarly, Tree-based Feature Importance (TFI, Breiman, 2001) generates the contribution by aggregating impurity reduction across all splits in a tree ensemble when perturbing feature values. SHapley Additive exPlanations (SHAP, Lundberg and Lee 2017) rooted in cooperative game theory, assigns an importance value to each feature for a prediction. It calculates marginal contribution of features by perturbing different subsets of inputs and averaging the effect on the output. Extensions of SHAP are tailored for more complex model structures (Covert and Lee 2021; Yang J, 2022; Chen H et al., 2019).

(2) Feature Effect Curves

These curves visualize the functional relationship between a feature and the model's prediction. Partial Dependence Plots (PDP, **Fig. 1A**; Friedman and Popescu 2001) show the average marginal effect of a feature on the prediction, while Individual Conditional Expectation (ICE, Goldstein et al., 2015) plots illustrate the prediction dependence for individual instances. Accumulated Local Effects (ALE, Apley and Zhu J 2020) plots are more reliable when features are correlated, as they compute conditional expectations to reduce bias.

2.2.2 Surrogate Models

These models use interpretable approximations to mimic the original complex model's behavior while offering transparency (**Fig. 1C**). A leading method is Local Interpretable Model-agnostic Explanations (LIME, Ribeiro et al., 2016). It assumes that, for a specific prediction, the complex behavior in a ML/DL model can be represented by an interpretable model. This model is trained on a set of artificial samples, created by perturbing the original input within its local neighborhood, along with the complex model's corresponding outputs (Guo W et al. 2018; Zafar and Khan, 2019; ElShawi et al., 2019).

2.2.3 Model Representation Methods

Model Representation (MR) methods extract and interpret latent representations of ML/DL models.

(1) Tree-Based Model Interpreters

Tools such as the Treeinterpreter package decompose prediction paths within tree ensembles (e.g., random forests, XGBoost), attributing contributions to individual features along decision routes for each instance (Saabas, 2014).

(2) Gradient-Based Attribution Methods

These techniques use backward-propagated gradients to estimate feature importance (**Fig. 1B**). The foundational saliency maps (Simonyan et al., 2014) directly visualize gradient calculus. Guiding backpropagation (Springenberg et al., 2014) modifies gradient flow to enhance visualization. Input×Gradient (Shrikumar et al., 2019) amplifies the signal by weighting gradients with input values. Integrated Gradients (IG, Sundararajan et al., 2017) accumulates gradients along a path from a baseline to the input, a concept extended by Expected Gradients through integration over distribution baselines (Erion et al., 2021). SmoothGradients averages gradients computed from noise-perturbed inputs (Smilkov et al., 2017). Occlusion sensitivity also perturbs input to evaluate the impact on predictions (Zeiler and Fergus, 2014). Layer-wise Relevance Propagation (LRP, Bach et al., 2015) redistributes prediction relevance backward through the network layers based on specific propagation rules, independent of gradient flow or neuron smoothness assumptions. Further details on these methods are provided in **Appendix B**.

(3) Attention Mechanisms and Activation-Based Visualization

Attention mechanisms dynamically weight various input elements, indicating their relative importance or relevance (Vaswani et al., 2017). Methods such as CAM, Grad-CAM (Selvaraju et al., 2017) and their variants (Patro et al., 2019; Bany and Yeasin, 2021; Ibrahim and Shafiq, 2022) leverage convolutional feature maps to produce heatmaps that highlight discriminative image regions influencing model decisions.

(4) Simplifying Latent Representations

These approaches create tractable approximations of the complex internal states of advanced models like generative adversarial networks (GANs, Bau et al., 2019; Chen X et al., 2016), diffusion models (Kwon et al., 2022; Lee S et al., 2023), and Transformers (Castangia et al., 2023; Playout et al., 2022; Schwenke et al., 2023). The approaches involve canonical correlation analysis (Burges, 2010; Raghu et al., 2017), clustering (Raghu and Schmidt, 2020), differential analysis (Motteler et al., 1995; Aires et al., 2004; Rahwan et al., 2019), linear (Lees et al., 2022) or symbolic approximations (Liu J et al., 2023).

2.2.4 Contrastive Examples

These methods explain a model's decision by showing what could have changed its outcomes. Counterfactual explanations (Chou Y L et al., 2022) identify minimal changes that would alter predictions (**Fig. 1D**). Prototypes and Criticisms (Gurumoorthy et al., 2019) reveal representative training samples. Both remain underutilized in ESS, likely due to structural data constraints and the lack of formal methodological integration.

The mentioned XAI methods are described in **Table S1**.

2.3 Evaluation of XAI

To ensure rigorous and verifiable XAI research, comparable quantitative evaluation metrics are essential. While 58% of XAI technical evaluations employ quantitative measures, 33% rely on anecdotal evidence and 22% involve human evaluations (Nauta et al., 2023). In ESS, benchmark datasets and process-based nonlinear functions often serve as ground truth for objective XAI validation (Arras et al., 2022; Mamalakis et al., 2022a,b). This review emphasizes the following five key metrics for quantitative evaluation (Melis and Jaakkola, 2018; Murdoch et al., 2019; Minh et al., 2022; Weber et al., 2023):

(1) Faithfulness: the extent to which explanations reflect the actual information in the model's decision-making (Alvarez-Melis and Jaakkola, 2018);

(2) Robustness: consistency of explanations under minor input perturbations that do not alter predictions (Huang X et al., 2020);

(3) Stability: reproducibility of explanations across similar samples (Alvarez-Melis and Jaakkola, 2018);

(4) Efficiency: computational cost required to generate explanations (Adadi and Berrada, 2018; Vilone and Longo, 2021).

(5) Localization: an explanation is measured based on its agreement with a user-defined region of interest (Arras et al., 2022).

Taking the global air temperature prediction as an example (Bommer et al., 2024; **Fig. 3**), LRP achieved the highest overall quantitative evaluation score among multiple gradient-based XAI methods, with its attribution map consistent with knowledge, most clearly highlighting the influence of the North Atlantic region. **Table S2** summarizes cross-domain case studies for XAI evaluation, while **Table S3** lists practical toolkits. Across these cases, SHAP and LRP generally demonstrate strong and reliable performance. No single method is universally optimal and evaluation remains context-dependent.

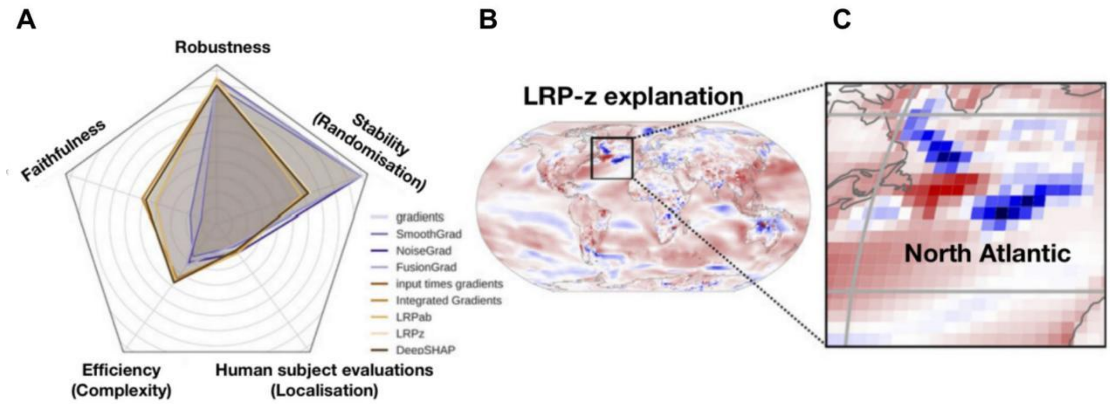

Figure 3 Evaluating XAI methods in climate science: A case study. (A) Spider plot displaying the mean skill scores of nine XAI methods. Note: The metric names have been adapted from the original source to maintain consistency with the terminology used in our text: 1. "Complexity" (original) is renamed to "Efficiency"; 2. "Randomisation" (original) is renamed to "Stability"; 3. "Localisation" (original) is renamed to "Human subject evaluation". The spider plot serves as a visual aid to identify the best-performing method, where results further from the center indicate better performance. (B) Explanation map for a decade prediction generated by the top-performing method (LRP-z) on a temperature field, with (C) a zoomed-in view over North America. Adapted from "Finding the Right XAI Method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science," by P. L. Bommer et al., 2024, arXiv preprint. [https://arxiv.org/abs/2303.00652v2]. Copyright © 2024 CC BY 4.0.

3. XAI: From Theory to ESS Application

3.1 Current State of XAI-ESS

Our analysis is based on 863 publications (review process in **Appendix C** and criteria in **Table S4**) from the Web of Science database by searching XAI-ESS related keywords (**Figure S2**), showing exponential growth since 2020 and reaching over 300 publications in 2024 (**Fig. 4B**), and the proportion of XAI papers in ML-related papers in ESS are higher than 20% in recent decades (Dramsch et al., 2025). Applications are predominantly concentrated in fields with abundant observational data, including environmental science, water resources, atmospheric science, and geoscience (**Fig. 4C**). Hydroclimatic modeling has emerged as a key focus area, with XAI increasingly used as a nonlinear statistical tool for climate attribution and prediction improvement of interpretability (**Fig. 4A**). We further analyzed the XAI and ML/DL methods to identify preferred methodological pairings (**Fig. 4D**). Perturbation-based methods, particularly SHAP, LIME, PI and PDP, are the most commonly applied alongside tree-based models. For Neural Network (NN) models such as CNN and Long Short-Term Memory (LSTM), model-representation techniques dominate.

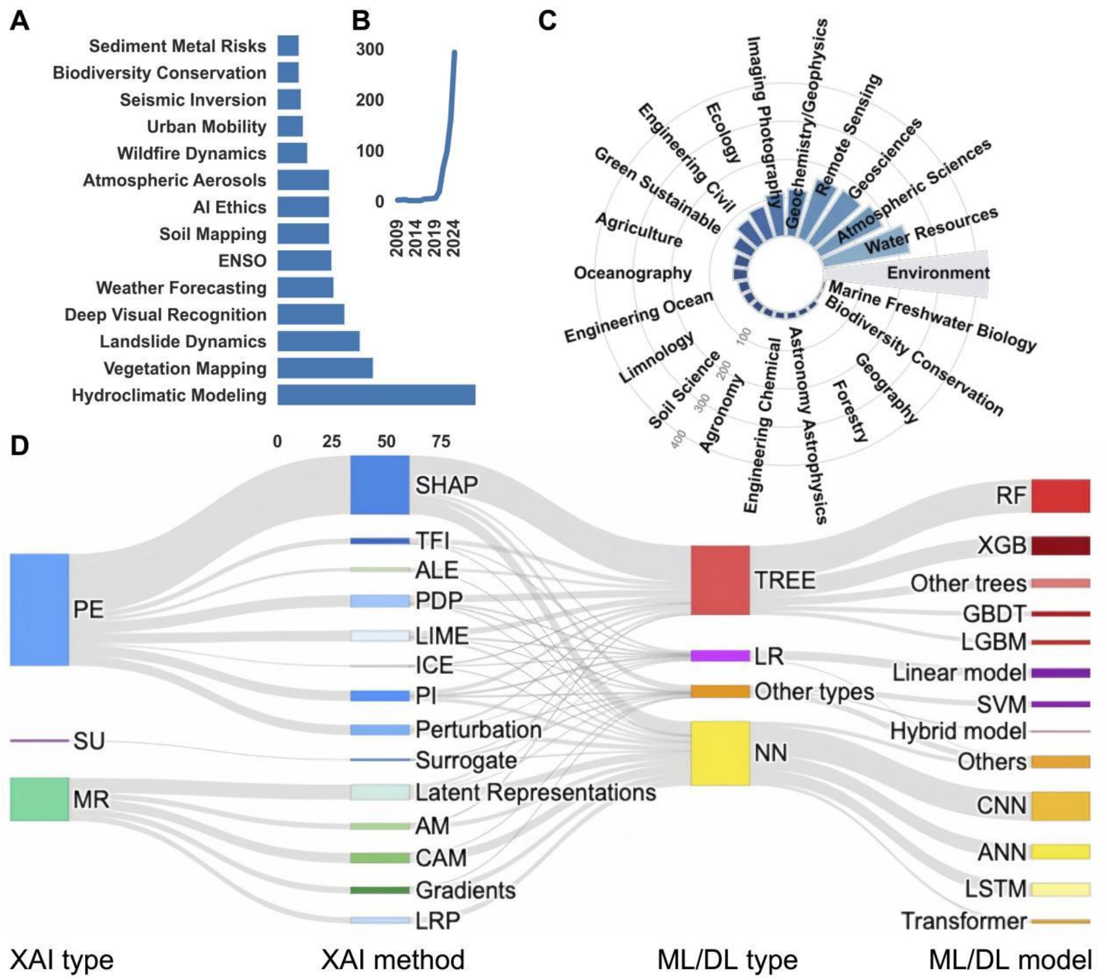

Figure 4 Current states of XAI related publications in Earth System Science (ESS) including (A) number of publications across related topics, (B) annual publication trend in recent years, (C) distribution by ESS subdomains based on Web of Science research area categories, and (D) methodological pairings of XAI types and methods, ML types and models in the selected papers. The analysis is based on 258 open-access articles from top-quartile journals in the Web of Science, all ranking within the top 80% by citation count.

The abbreviations are as follows: SU (SUrrogate models); MR (Model Representations); PE (PErturbation-based methods); TFI (Tree-based Feature Importance); SHAP (SHapley Additive exPlanations); PI (Permutation Importance); PDP (Partial Dependence Plot); LRP (Layer-Wise Relevance Propagation); LIME (Local Interpretable Model-agnostic Explanations); ICE (Individual Conditional Expectation); CAM (Class Activation Mapping); AM (Attention Mechanism); ALE (Accumulated Local Effects); LR (Logistic Regression/Linear Regression); NN (Neural Network); LGBM (Light Gradient Boosting Machine); GBDT (Gradient Boosting

Decision Tree); SVM (Support Vector Machine); CNN (Convolutional Neural Network); LSTM (Long Short-Term Memory); XGB (eXtreme Gradient Boosting); RF (Random Forest); ANN (Artificial Neural Network).

3.2 Improving Interpretability for AI Models

XAI is essential in ESS for development of transparent, accountable, and trustworthy AI systems to support high-stakes decisions (Haupt et al., 2021; Debnath et al., 2023; Camps-Valls et al., 2025a). This need is reinforced by regulatory frameworks such as the UNESCO on Ethics of AI (2021) and the EU AI Act (2021), which mandate accountability. To be actionable, explanations should be valid, generalizable, and scientifically consistent.

In short-term weather and extreme forecasts, XAI elucidates how AI models work for extreme events, including droughts (Feng P et al., 2020; Dikshit et al., 2022; Huang F et al., 2023a), floods (Yang W et al., 2020; Ekmekcioğlu et al., 2021), landslides (Al-Najjar et al., 2022), soil moisture droughts (Ye S et al., 2025), streamflow (Althoff et al., 2021; Chen M et al., 2023), hail (Gagne II et al., 2019) and earthquake warning (Fayaz and Galasso, 2024). For long-term climate science, it supports the generation of trustworthy AI-based future projections (McGovern et al., 2023, 2024; Diffenbaugh and Barnes, 2023; Evans et al., 2025), such as sea surface temperature (van Straaten et al., 2022) and climate oscillations (Schmidt et al., 2020; Gordon et al., 2021). In AI-derived Earth data products, XAI serves as a critical validation tool. It ensures the physical consistency of products for essential variables (Shangguan et al., 2017, 2023; Dueben et al., 2022; Gevaert, 2022), such as soil carbon (Wadoux and Molnar, 2022), crop types (Orynbaikyzy et al., 2020) and remote sensing classifications (Hasanpour Zaryabi et al., 2022), fostering confidence in the products for both research and policymaking.

3.3 Enhancing Efficiency of AI Building

ML/DL developers usually use XAI for feature selection, model determination, and structure design to improve the efficiency of model building.

For feature selection, XAI provides feature importance from a pre-trained model to remove the irrelevant features such as perturbance-based methods including TFI (Feng P et al., 2019; Upadhyaya et al., 2021) and PI (Ramirez et al., 2022), and attention mechanisms (Yan J et al., 2021), yielding better results (Zacharias et al., 2022; Wang J et al., 2023) and physical consistency (Carter et al., 2021) in ESS element prediction.

To determine optimal ML/DL models, developers can compare behaviors of candidate models provided by XAI with prior knowledge (Jing et al., 2023; Wu Y et al., 2023). This strategy has enhanced flood forecasting (Schmidt et al., 2020) and helped validate complex DL models against process-based simulations, yielding physically realistic insights (Hu X et al., 2021).

Many XAI methods are tailored to diagnose and improve model structure design, especially MR methods (Toms et al., 2020). MR helps detect the problems in tree nodes such as decision trees (Chen J et al., 2021) and cubist (Fu Z et al., 2022), with tools like Treeinterpreter applied to precipitation classification (Upadhyaya et al., 2021). In DL models, model representations such as Jacobian matrix help reduce bias from removal of the unimportant links in NN for atmospheric profiling (Blackwell, 2012; Maddy et al., 2021) and precipitation prediction (Shamekh et al., 2023). While XAI effectively supports model optimization and promotes physical consistency in ESS, it does not provide direct guidance for structural design or hyperparameter tuning.

3.4 Providing Scientific Insights from AI Models

ESS researchers place high expectation on XAI as a data-mining tool to gain physical insights from high-performing ML/DL models (Jiang S et al., 2024a).

3.4.1 Gaining Hydrological Insights

XAI has been increasingly employed to identify key hydrological drivers governing the dynamics of soil moisture (Ley et al., 2024), evapotranspiration (Chakraborty et al., 2021b; Zhang H et al., 2025), runoff (Althoff et al., 2021; Bai Z et al., 2022; Hao H et al., 2024; Wang H et al., 2024), streamflow (Liu J et al., 2023; Wu and Li, 2023), and groundwater (Mo Y et al., 2025), thereby helping to disentangle coupled nonlinearity and heterogeneity challenges in traditional hydrological models. Furthermore, XAI techniques can detect critical process thresholds (Wang S et al., 2022; Ding K et al., 2025; Wang H et al., 2024) and spatial heterogeneity in variable responses (Wang S et al., 2024). In this way, XAI bridges data-driven predictions with mechanistic understanding, elucidating hierarchical hydrological controls.

3.4.2 Attributing Extreme Events

XAI advances mechanistic understanding of extreme events by disentangling the complex processes within ML/DL models, such as those for droughts (Feng P et al., 2019; Saha et al., 2021; Dikshit et al., 2024), floods (Jiang S et al., 2022a, b; Xu Y et al., 2024; Feng J et al., 2024; Liu M et al., 2024; Ke E et al., 2025), wildfires (Kondylatos et al., 2022; Abdollahi and Pradhan, 2023; Bountzouklis et al., 2023), extreme precipitation (Gimeno-Sotelo et al., 2023), and heatwaves (Shu R et al., 2025). For instance, in flood analysis, XAI has been used to attribute events to abrupt shifts in environmental precursors like precipitation and temperature (Slater et al., 2024). Furthermore, interactive SHAP values have elucidated critical compounding effects (Jiang S et al., 2024), and the incorporation of remote predictors has facilitated the detection of teleconnection drivers (Lee Y et al., 2024).

3.4.3 Detecting Anthropogenic Footprint

The attribution capability of XAI varies significantly with process timescale. It effectively identifies anthropogenic drivers in rapid-response systems like point-source pollution (Liu S et al., 2025) and urban heat islands (Zumwald et al., 2021; Oukawa et al., 2022). Yet, its efficacy diminishes for climate-scale phenomena (Alam et al., 2025) like greenhouse forcing, where signals are weak and human influences accumulate nonlinearly.

XAI itself often relies on recognized climate proxies. Commonly, global mean temperature is regarded as an indicator of anthropological influence (Janssens et al., 2021). Based on this, XAI can attribute downstream effects such as abnormal biogeochemical cycles (Haaf et al., 2021; Patoine et al., 2022; Wang K et al., 2022), carbon sequestration and latent heat absorption (Berner et al., 2020), and ocean-atmosphere coupling (Sonnewald and Lguensat, 2021). The climate-scale paradigms like “dry gets drier, wet gets wetter” have been corroborated through XAI (Chakraborty et al., 2021a; Anderson and Radić, 2022).

3.4.4 Limitations and the Path Forward

While XAI can highlight influential variables or nonlinear relationships within data, it does not yield fundamental physical laws or closed-form equations. Its explanatory power remains constrained by the underlying data and model structure. It can expose correlations and interactions but cannot independently validate causality. This poses critical challenges for both geoscientists and climate policymakers including spurious correlations and poor generalization to unseen data. This limitation necessitates close collaboration between ML/DL developers and ESS experts to ensure that XAI-driven discoveries are grounded in domain physics rather than algorithmic artifacts. Ultimately, although XAI is a powerful tool for addressing complex ESS challenges, it works best when combined with traditional experimentation and theoretical approaches to build system-level understanding.

3.5 Practical Guidelines for Implementing XAI in ESS

We propose an integrated framework that embeds explanation within the iterative modeling process (**Fig. 5A**). The framework begins with clear objective definition followed by routine data processing, feature selection, model training and validation. Candidate XAI methods are then assessed by quantitative metrics (**Section 2.3**) or prior knowledge to optimize the outcomes. Within this common workflow, different objectives emphasize different steps: interpretability focuses on the reliability of explanations as the final output; efficiency leverages XAI as a diagnostic tool to streamline feature selection, data usage, and model design; and scientific understanding prioritizes embedding physical priors to enable causal insight beyond correlations. As an example, Huang F et al. (2023b) applied this framework to soil moisture prediction. Based on feature selection (**Fig. 5B**) and well trained model (**Fig. 5C**), joint XAI analysis revealed the dominant control of soil temperature consistent with thermodynamic principles (**Fig. 5D**).

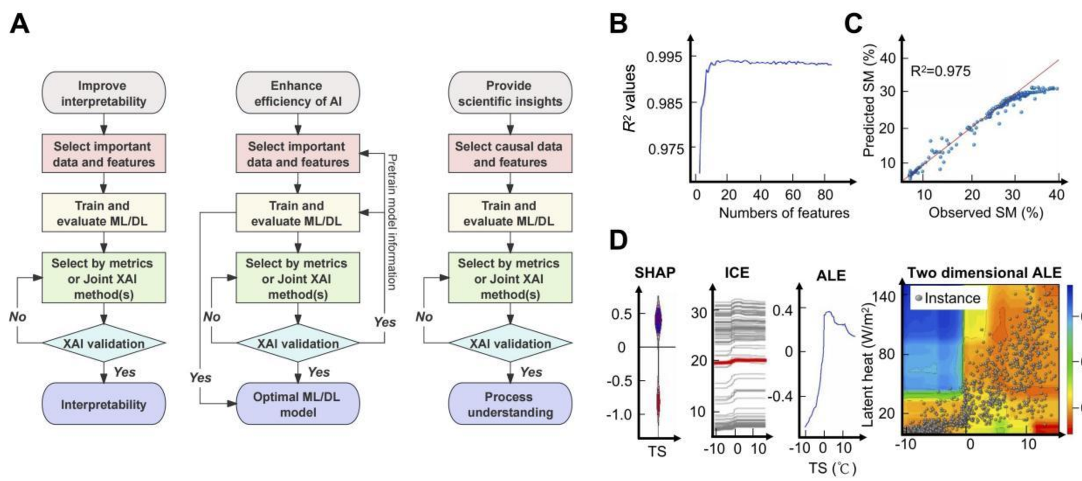

Figure 5 Practical framework for XAI application in ESS and a soil moisture (SM) case study. (A) Schematic of the iterative XAI explanation and validation pipeline. (B-D) Results from the case study: (B) feature selection, (C) model performance, and (D) joint XAI-derived explanation for soil temperature (TS) including SHAP, ICE, ALE and two-dimensional ALE of TS and latent heat. (B-D) were modified from Huang F et al., 2023b. Copyright © 2023 Elsevier.

4. Challenges and Possible Solutions

4.1 Inherent XAI Limitations

4.1.1 Methodological Limitations of XAI

XAI limitations stem from the inherent constraints in their design affecting their reliability, and no method is universally optimal.

Surrogate model-based methods such as LIME rely on strong assumptions, including local linearity and feature independence assumptions, which are rarely satisfied in real-world ESS applications. Perturbing one variable can unintentionally affect others, as observed in soil moisture prediction tasks (**Fig. 5D**; Huang F et al., 2023b). It has been found that the LIME explanation was more sensitive to correlated features than SHAP for small sample modeling (Huang F et al., 2023a). This effect becomes less pronounced when large sample sizes are used (Krell et al., 2025). Perturbation-based methods often assume Gaussian noise (Ivanovs et al., 2021) inconsistent with real-world distribution. Surrogate model and perturbation-based methods usually face computational challenges in high-dimensional settings.

Gradient-based methods often introduce noisy or unstable explanations (Saleem et al., 2022) or bring extra randomness to explanations especially in large networks (Huang F et al., 2024). Model representation methods like latent space reductions tend to yield less interpretable representations (Mylonas et al., 2024). Perturbation- and surrogate-based methods are generally more computationally expensive than intrinsic approaches. For instance, LRP takes 1-2 s to generate a heatmap for a single prediction which is far lower than the time taken by LIME (22 s) and SHAP (108 s).

4.1.2 Difficulties in Selecting an Optimal XAI Method by Metrics

When various XAI techniques are employed to interpret black-box ML/DL models, they often yield divergent explanations. This "Rashomon effect" occurs when differences in how these methods process and represent ML/DL model (Başağaoğlu et al., 2022; Huang F et al., 2023c). While different strategies have been proposed to evaluate XAI methods (see **Section 2.3**), they are typically applied in isolation and require additional computational resources. Their conclusions are often context-dependent, which limits their generalizability.

4.1.3 Lack of Benchmark for XAI Evaluation

The absence of benchmarks for evaluating explanations further complicates XAI selection. Prior knowledge is commonly used as anecdotal evidence or "ground truth". Mamalakis et al. (2022b) introduced the ClimateNet dataset, which provides expert-annotated maps to assess whether XAI techniques capture physically meaningful patterns or not. Alternatively, information derived from multiple Earth System Models (ESM) and other process-based models has been used as a reference for evaluating explanation quality, leveraging the collective physical knowledge embedded in these models to support XAI validation (Mamalakis et al., 2022a).

4.1.4 Solutions

We advocate the development of integrated interpretation frameworks that employ multiple XAI techniques to jointly analyze the same model or samples, thereby combining or weighting complementary perspectives to better reflect internal model behavior (Gibson et al., 2021; Ghada et al., 2022; Van Straaten et al., 2022; Ham et al., 2023). For example, Huang F et al. (2023a) combined PI, SHAP, LIME, ALE, PDP, and ICE to jointly explain soil moisture drought patterns, thereby enhancing the reliability and faithfulness of the explanation to the ML/DL model.

Alternatively, developing integrated XAI methods that merge ante-hoc and post-hoc approaches, combines perturbation, surrogate, and model representation techniques, or links global and local explanations to provide a comprehensive understanding of the model.

By contrast, interpretable DL models (i.e., diffusion model and graph neural networks) offer a potential pathway forward to address the problems of model complexity (O'Loughlin et al., 2025). By explicitly encoding relationships within their architectures, these models facilitate visualization, comparison, and integration of prior physical knowledge.

4.2 Compatibility Issues in XAI-ESS

4.2.1 XAI Design vs. ESS Community Requirements

Despite the various interpretative forms of XAI discussed in **Section 2.1**, they do not adequately meet the needs of ESS, largely due to a disconnect between the design priorities of XAI practitioners and the practical needs of the ESS community (Gevaert, 2022). Stakeholders including weather forecasting professionals, policymakers and regulators, ML/DL developers, AI-derived product users and geoscientists, require explanations that extend beyond the commonly used simplistic feature attributions, encompassing the dynamic behaviors and states of complex Earth system processes.

Although weather forecasting professionals intend to use feature attributions to prediction tasks, current explanatory formats fall short of enabling their comprehension. Policymakers and regulators prefer global and counterfactual explanations to support decision-making in future scenarios. Current XAI can provide limited guidance for diagnosing the AI building, especially for structure building.

4.2.2 Static Explanations vs. Dynamic, Interdependent Systems

The conceptual gap between XAI assumptions such as feature independence, model stationarity, and local linearity and the nature of ESS, which exhibit strong feature interdependence, system dynamism, and global nonlinearity is fundamental for further applications. Nonstationarity in ESS underscores the time-varying and inherently unpredictable nature of the behavior, whereas most XAI relies on static model assumptions. As a result, XAI struggles to provide reliable interpretations for extrapolation, particularly in unseen regimes, or distributional shifts beyond the training domain.

Spatial-temporal dependencies may hinder the robustness of many XAI methods. Perturbation-based methods (e.g., PI and LIME), which assess feature importance by independently perturbing inputs and monitoring output changes, are fundamentally compromised in such contexts. This often produces physically implausible samples and leads to attribution dispersion, where importance is spread across numerous spatiotemporal nodes rather than capturing coherent system-level responses or cascading processes (e.g., floods or landslides).

Moreover, most off-the-shelf XAI methods focus on unstructured data, whereas ESS data are highly structured, characterized by confounding effects, strong inter-variable correlations, heterogeneous distribution (e.g., Poisson-distributed precipitation and sinusoidal solar shortwave radiation). This mismatch introduces bias into the process understanding derived from XAI (**Figure S3**), as inter-variable correlations lead to the uncertainty in explanation has been found (Jiang S et al., 2022a, b). Various preprocessing strategies such as variance inflation factors (He F et al., 2025), Pearson coefficient (Díaz-Vallejo et al., 2024) and z-score (Cui X et al., 2021) are used to mitigate these effects. However, such approaches may remove the correlation within the model, the underlying data structure remains, meaning that perturbations to one variable can still induce nonlinear changes in others.

4.2.3 Causality vs. XAI

Geoscientists seek to extract causal relationships from observation-driven ML/DL models to complement existing knowledge systems (Irrgang et al., 2021). Recent studies use XAI to identify “drivers” in a causally meaningful manner within predictive models. However, such attempts often conflate correlation with causation. Even highly accurate ML/DL models do not inherently meet the requirements for causation without rigorous causal validation. This would be related to the spurious correlations brought by potential confounding variables. Whether a model is built on causality or correlation fundamentally determines its capacity to infer causal mechanisms

(Camps-Valls, 2025b).

Current literature suggests two pathways to address this issue: XAI for causality and causality for XAI (Carloni et al., 2025). The former uses XAI to generate scientific hypotheses, which are subsequently validated through experimental methods. The latter derives explanations from causal models such as structural causal models. Both approaches hold promise for applications in ESS.

In practice, many existing studies employing XAI to identify drivers, implicitly follow XAI for causality.

Some researchers have integrated prior knowledge into hybrid modeling frameworks in ESS to mitigate confounding biases during explanation (Althoff et al., 2021; Li W et al., 2024; Hu X et al., 2021). Nevertheless, the absence of robust validation remains an issue. Experimental approaches such as free-air CO_2 enrichment could potentially offer means of verification. Conversely, deriving explanations from causal models requires carefully delineating the scope of study to better control for confounders and obtain targeted causal explanations.

4.2.4 End-to-End Modelling vs. Process-Based Modelling

A fundamental challenge arises because XAI typically explains end-to-end black-box models, which capture complex mappings but remain difficult to interpret in terms of intermediary processes and variable interactions. In contrast, process-based models in ESS explicitly represent mechanistic relationships and intermediate states. Consequently, XAI-derived explanations often expressed in terms of input-output attributions which fail to align, both in form and cognitive relevance, with the process-oriented understanding required for diagnostic analysis and hypothesis testing in geoscientific research (Li X and Guo Y L, 2025). Geoscientists often underutilize the function of XAI and some data-driven equation discovery techniques (i.e., symbolic regression). The core of the issue lies in the fact that ML/DL and XAI are designed without considering process-oriented ESS.

4.2.5 Solutions

(1) Object-aware XAI Tasks

XAI should explicitly account for domain-relevant objects (e.g., cyclones, ecosystems, watersheds) moving beyond explanations based solely on potentially causal features. Object-aware explanations can better capture system organization and dynamics, improving their relevance for scientific interpretation.

(2) Process-informed Modeling

Incorporating physical knowledge into modeling can move predictions closer to the underlying "ground truth" embedded in structured ESS data, enabling XAI to reveal causally meaningful insights that are scientifically understandable (Chen M et al., 2023). One promising strategy is to decompose the input-output mapping into known and unknown subprocesses (Shen C et al., 2023). The known processes are modeled by the process-based model and the unknown processes are modeled by the explained ML/DL models. Causality within a subprocess may be easier to verify when focusing on simpler subsystems that exclude confounding factors from more complex systems, which would be beneficial for ESS modelers to integrate the structural data (Li, X., et al., 2023). We propose a novel causal physics-informed hybrid XAI framework (see **Appendix D**; **Figure S4**). This approach emphasizes process representation rather than data distribution alone, using physics to model known mechanisms and ML/DL to approximate unresolved processes.

(3) Human-centered XAI Design for ESS

To address diverse ESS needs, human-centered XAI design is crucial. Guidelines originally developed for medical XAI (Chen H et al., 2022) can be adapted to ESS. Recent studies highlighted accountability risks from biased ML/DL in climate action (Debnath et al., 2023). Throughout XAI-ESS development, human-centered principles ensure usability and accountability (Clement et al., 2023). A structured workflow (**Fig. 6**) may include: (a) analyzing user needs via surveys or scorecards (Hoffman et al., 2023); (b) co-design with ESS stakeholders to integrate domain expertise (e.g., physically interpretable models for forecasters (Kim et al., 2023), traceable systems for policymakers); (c) standardized workflows to streamline XAI adoption for end-users; (d) validation to test XAI empirically to refine ESS applications. Cross-disciplinary alignment of explanation goals is key. A concrete step is to develop open-source libraries that seamlessly integrate process-based models with XAI techniques.

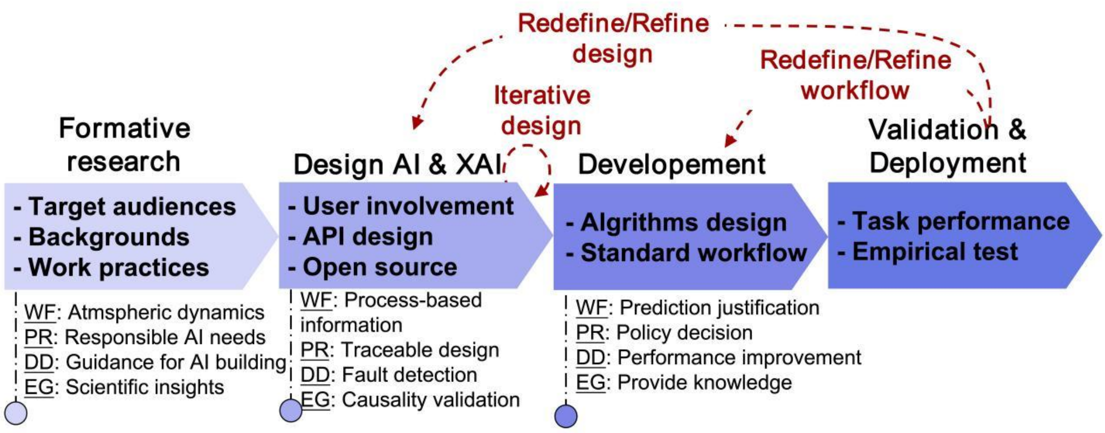

Figure 6 Operational solutions for human-centered XAI design for ESS. The abbreviations WF, PR, DD, and EG are weather forecasters, policymakers and regulators, diagnosis and debugging, and Earth process-based modelers and geoscientists, respectively. And API denotes the Application Programming Interface.

4.3 Uncertainty Assessment

Quantifying uncertainty in XAI is vital for reliable decision-making in high-stakes domains such as weather forecasting and climate policy, where unreliable explanations may lead to severe societal impacts or misleading scientific conclusions. XAI-specific uncertainties stemming from their designs and inherent assumptions compromise robustness (Thuy and Benoit, 2024). Addressing these uncertainties is paramount to ensure that XAI-derived insights are scientifically sound and actionable. One practical approach is to quantify the variability of explanations by repeatedly applying XAI methods under identical conditions and conducting convergence analyses to assess explanation stability.

5. Future Perspectives

5.1 Faithful, Efficient, and Integrated XAI

In the future, we envision a more faithful and computationally efficient integrated XAI framework that combines multiple techniques as shown in **Fig. 1** (Belaid et al., 2023). Such a framework aims to overcome the inherent limitations of individual XAI methods by unifying their complementary strengths into a coherent approach. For example, "glocal XAI", an integrated method, offers both channel-level information and feature importance, providing deep insights into a model's representations, reasoning processes, and fine-grained decision-making (Achtibat et al., 2023). Similarly, attention mechanisms that synthesize global-local perspectives (Li L et al., 2018) and spatial channel information (Song C H et al., 2022) also show considerable potential in computer vision applications.

Future XAI techniques should overcome the limitations of individual methods by seamlessly integrating them into a unified method. We posit that model-specific XAI methods are better suited for advanced DL models, integrated XAI methods should offer greater flexibility to allow comparisons of interpretability across a variety of distinct architectures (Theissler et al., 2022).

There is a pressing need for lightweight XAI solutions to keep pace with the rapid advances in foundation models. Cutting-edge models such as Pangu-Weather (Bi K et al., 2023), FuXi (Chen L et al., 2023) and GraphCast (Lam et al., 2023) have set new performance benchmarks in the ESS community. The computational cost of traditional XAI methods like SHAP severely limits their use with large models. Future solutions should favor built-in interpretability or targeted input perturbation, which efficiently probes models with strategic exemplars, avoiding interpretability biases introduced by model distillation.

Further, we need to consolidate faithfulness, robustness, stability, and other key metrics to generate a standardized metric for XAI evaluation. Currently, it is challenging to compare different methods in a consistent manner (Mi J et al., 2024). Moreover, those metrics often rely on assumptions such as local consistency for localisation, stability under perturbations. They often conflict with spatiotemporal teleconnections, nonlinear threshold dynamics and tipping point behavior in ESS. For example, while robustness metrics assume explanation stability under small input perturbations, the "butterfly effect" prevalent in atmospheric systems may invalidate this assumption at the system level. These small changes in initial conditions can produce nonlinear cascading effects that are difficult to capture or attribute reliably using XAI alone. Therefore, a crucial first step should be the development and evaluation of an XAI method specifically tailored to the nuances of the ESS domain.

5.2 Expanding the Roles of XAI in ESS

Here, we argue that the role of XAI in ESS should be further expanded. We elaborate on this from three progressive stages of ESS: recognition, description, and action. In **Fig. 7**, we mark the roles already achieved by XAI (in gray) and potential future roles (in red). Our framework conceptualizes ESS through three progressive stages: (a) recognition, which adopts a pragmatic approach to derive insights from real-world data simulations to inform decision-making, thereby advancing cognitive understanding of ESS dynamics; (b) description, which systematically characterizes processes and data patterns to elucidate the system's governing mechanisms; (c) action, which translates cognitive and descriptive knowledge into targeted interventions for ecosystem conservation or transformation.

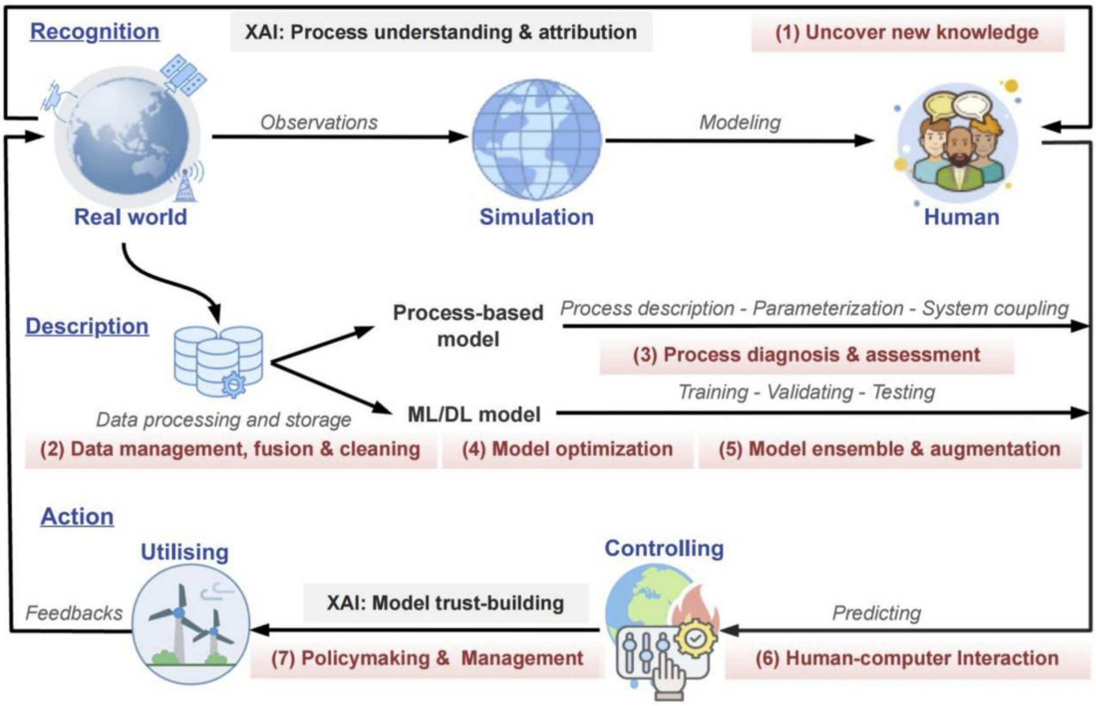

Figure 7 A hierarchical framework for Earth System Science (ESS) integration, spanning from fundamental process understanding to actionable applications. The roles already achieved by XAI (in gray) and potential future roles (in red) are marked.

(1) Recognition: Uncover New Knowledge

XAI facilitates the extraction of high-level relationships and processes from complex, low-level data that may be difficult for humans to discern. As a result, it holds great potential for uncovering new insights and discovering new knowledge across multiple disciplines. For instance, XAI has played an instrumental role in revealing unknown representations and effects in drug discovery (McCloskey et al., 2019; Ishida et al., 2019). GNNExplainer, an XAI method coupled with graph convolutional networks, optimizes mutual information between elements and serves as a powerful tool in this area. Its chemically meaningful representations align well with the intuitive understanding of chemists (Ying Z et al., 2019). This success could inspire ESS to use XAI to explore previously overlooked subprocesses within ESM.

In cases where certain ESS processes are difficult to parameterize, the successful application of material dynamics modeling using XAI offer encouraging evidence (Zhong X et al., 2022). For example, the combined use of feature importance, conceptual explanation, and testing with concept activation vector methods enabled accurate descriptions of process dynamics in scalable perovskite solar cells using video datasets (Klein et al., 2024). Moreover, inherently transparent ML/DL models have proven effective in representing a variety of chemical processes (Gallegos et al., 2024). Looking ahead, XAI presents a promising alternative approach to exploring new knowledge and representations in ESS, owing to its ability to illuminate hidden patterns and complex interactions that often inaccessible to traditional modeling and analysis.

(2) Description: Data Management, Fusion and Cleaning with XAI

The role of XAI-ESS can be expanded to a broader scope, where XAI can be used throughout the entire lifecycle of data modeling. For instance, in data management, an innovative concept known as "edge intelligence" has emerged, which involves generating edge-level data to reconstruct data representations and reduce data volume (Sinha and Vashisht, 2023). The use of XAI ensures that this process remains physically consistent and interpretable. This capacity, in turn, facilitates efficient management and storage of large datasets such as those from the Coupled Model Intercomparison Project Phase 6 (CMIP6), high-resolution reanalysis and remote sensing products.

In terms of data fusion, XAI has been used to effectively merge multi-source remote sensing products (Lohit et al., 2019; Xie Q et al., 2019; Taskin et al., 2024). The application of XAI can be further extended to fuse a wider variety of data sources, including remote sensing, reanalysis, in-situ measurements, and process-generated data, due to its prowess in identifying and extracting similar patterns from complex data sets.

In addition, XAI can also contribute significantly to data cleaning. Prototypical relevance propagation, which leverages concepts such as prototypes and LRP, can adeptly identify artifact samples and separate them from clean ones (Gautam et al., 2023). In the context of medical image analysis, XAI methods effectively selects the most informative samples for active learning settings, outperforming other selection criteria by requiring less data and fewer iterations to achieve comparable or superior performance (Mahapatra et al., 2021). Such techniques could be highly valuable for mitigating the adverse effects of noisy sensor data, for example, during FLUXNET data preprocessing.

(3) Description: Process Diagnosis and Assessment

ESM are instrumental in forecasting the Earth's future state based on our understanding of ESS processes. However, despite their significance, ESM in the latest CMIP6 exhibit inconsistencies across a wide range of process representations (Hsu and Dirmeyer, 2023; Fu W et al., 2022). To analyze these biases stemming from parametric uncertainty and external drivers, data-driven reconstruction of physical models namely emulation is essential (Lian X et al., 2018).

In this context, XAI serves as a powerful visualization and diagnostic tool that can discern distinct patterns within AI-based reconstruction models (Li X et al., 2023). For instance, XAI can reveal disparities in physical representations among ESM, such as those related to Arctic Ocean acidification (Krasting et al., 2022).

Alternatively, a complementary and more sophisticated approach to identifying varying patterns in ESM involves training an ML/DL model using annual-mean temperature or precipitation maps as inputs to predict the corresponding year of each map (Barnes et al., 2020; Labe and Barnes, 2021). By interpreting the ML/DL model, the underlying information encapsulated in ESM is revealed, effectively bypassing the noise inherent to these models. This method allows researchers to gain insights into the fundamental dynamics driving ESM behavior without being confounded by extraneous internal variability.

(4) Description: Model Optimization

Methods such as CAM and canonical correlation analysis can refine intermediate features representations and adjust a model's internal structure, thereby fine-tuning performance (Schiller et al., 2019; Anders et al., 2022). Explanations from pre-trained model can also strengthen loss functions (Ross et al., 2017), augment gradients (Narteni et al., 2025), or even enhance the model architecture itself (Yeom et al., 2021). Novel XAI has also been used to optimize hyperparameters (Ibrahim and Shafiq, 2022) and accelerate the learning process (Mukhamediev et al., 2022). Despite these advances, there is often a delay in the adoption of the latest XAI developments in

ESS applications. It is essential for ESS practitioners to keep pace with the rapid progress in XAI to ensure that the application of XAI-ESS remains up-to-date and achieve their full potential.

(5) Description: Model Ensemble and Augmentation

Some advanced XAI methods aim to enhance ML/DL model performance by employing strategies not widely adopted in ESS contexts, such as data augmentation, feature augmentation, loss augmentation, gradient augmentation, and model augmentation (Weber et al., 2023). Whether the explanations generated by XAI exhibit physical consistency could also serve as one of the evaluation criteria for weighting models within ensemble framework.

(6) Action: Human-Computer Interaction

Interactive XAI systems play a key role in promoting human-computer interaction by allowing users to interrogate the system to gain a deeper understanding of their decision-making mechanisms. For example, an XAI-based interface system designed for weather forecasting models allows users to assess overall reliability, evaluate the consistency of explanations with domain knowledge, and provide feedback. As citizen science continues to emerge as an important social activity, with individuals using smartphones and other devices to contribute to scientific research, ML/DL models can help process the immense amount of data collected, while XAI enables participants to understand the scientific findings they are contributing to, thereby increasing their engagement in learning.

Moreover, integrating XAI with virtual reality or augmented reality technologies could create a powerful visualization tool that facilitates a comprehensive understanding of ESS. On another front, interactive XAI systems should provide open interfaces that allow human intervention in ML/DL modeling to ensure fairness and accountability to the public. In particular, there are several functional visual XAI methods, such as LSTMVis (Strobelt et al., 2017) that offer interactive graphs and heatmaps showcasing relevant words and associated subsets of hidden states. In the future, interactive XAI is expected to be a critical function as an interface that facilitates communication between humans and machine-generated decisions (Naiseh et al., 2023), thereby bridging the gap between technology and user understanding in a more intuitive and collaborative manner.

(7) Action: Policymaking and Management

Given the impressive predictive capabilities of XAI in areas such as climate change prediction and attribution (**Section 3.2**), it is increasingly important to consider whether policy guidance can be derived from its results. First, the promotion of trustworthy ML/DL systems, particularly in the weather, water, and climate sectors, has become an urgent and pressing concern (Reidmiller et al., 2017). In the absence of empirical evidence on model interpretability by end-users, deontological ethics dictate that ML/DL developers have a moral obligation to create XAI solutions that provide users with sufficient information about the inner workings of the model to preserve their decision-making autonomy (McGovern et al., 2022).

Second, addressing accountability risks arising from potentially biased ML and DL systems is critical when they are used to inform climate policy (Debnath et al., 2023). We argue that if robust evidence from XAI can be validated through experiments or additional evidence, they are able to inform the policymakers to make decisions on climate change mitigation. Moreover, in domains such as water and carbon management, regional-scale interventions and experiments based on hypotheses generated by XAI become increasingly feasible. As an example, Schoenke et al. (2021) developed a conceptual platform in agriculture that visualizes data flows to provide interpretable insights and actionable recommendations for both farmers and regulators. This showcases how XAI can directly influence resource management strategies and inform policy formulation in various domains.

6. Summary

While AI has become integral to ESS, its full potential remains constrained by the interpretability gap associated with black-box models. XAI promises to unlock this potential, but our review concludes that a fundamental mismatch between generic XAI designs and domain-specific ESS requirements continues to hinder its practical utility. Transforming XAI from a technical novelty into a core ESS tool requires a coordinated shift in development priorities. Based on our analysis, we propose the following actionable framework for the community:

(1) Develop ESS-specific XAI method and benchmarks. To overcome inherent methodological limitations, we advocate for a dual strategy: (a) the systematic use of multiple XAI techniques, together with the development of inherently interpretable DL models to ensure robust explanations; and (b) the establishment of a consortium to define community-standard benchmarks for assessing XAI faithfulness and robustness on canonical ESS tasks. These benchmarks should prioritize physical plausibility alongside numerical accuracy.

(2) Prioritize human-centered, hybrid explainability. Purely technical explanations are insufficient; they should be translated into scientifically meaningful insights. Our findings strongly advocate for the widespread adoption of co-design frameworks, where ESS domain experts, ML/DL developers, and end-users collaborate to build interpretability into models from the outset. Furthermore, we identify the integration of physics-inform and causal inference techniques with data-driven XAI as the most promising path toward generating trustworthy and mechanistic insights. A concrete initial step is to develop open-source libraries that seamlessly integrate process-based models with leading XAI techniques.

(3) Mandate uncertainty quantification for explanations. An explanation without an associated measure of uncertainty is of limited scientific value. We urge that uncertainty quantification should be treated as a non-negotiable component of XAI in ESS. This requires developing new methods to propagate uncertainty through explanation generation process, as well as effective visual strategies to communicate this uncertainty to scientists and policymakers.

The future trajectory of ESS is inextricably linked to advances of trustworthy AI, necessitating a focused effort on next-generation XAI. We foresee two critical frontiers: first, the creation of lightweight, inherently interpretable models that seamlessly integrate physical principles; and second, the development of robust diagnostic tools for model auditing. By prioritizing these avenues, XAI will transition from a passive explanatory tool to an active scientific discovery, catalyzing progress in key areas including knowledge extraction from complex datasets, hybrid models development, reconciliation of ESM discrepancies, and design of reliable decision-support systems for policy.

Acknowledgements

The study was supported by the National Natural Science Foundation of China (Grant Nos. 42375144, 4227515 & 42205149), Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China, Guangdong Basic and Applied Basic Research Foundation (2021B0301030007) and China Meteorological Administration Youth Innovation Team (CMA2024QN01). F.H. acknowledges a PhD scholarship from the China Scholarship

Council (202306380183).

Conflict of Interest

The authors declare that they have no conflict of interest.

Supporting Information: The supporting information is available online at http://earth.scichina.com and http://link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

Aas K, Jullum M, Løland A. 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif Intell, 298: 103502. doi: 10.1016/j.artint.2021.103502

Abdollahi A, Pradhan B. 2023. Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model. Sci Total Environ, 879: 163004. doi: 10.1016/j.scitotenv.2023.163004

Achtibat R, Dreyer M, Eisenbraun I, et al. 2023. From attribution maps to human-understandable explanations through Concept Relevance Propagation. Nat Mach Intell, 5: 1006–1019. doi: 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. doi: 10.1109/ACCESS.2018.2870052

Aires F, Prigent C, Rossow W B. 2004. Neural network uncertainty assessment using Bayesian statistics: A remote sensing application. Neural Comput, 16(11): 2415–2458. doi: 10.1162/0899766041941925

Alam G M I, Arfin Tanim S, Sarker S K, et al. 2025. Deep learning model based prediction of vehicle CO_2 emissions with eXplainable AI integration for sustainable environment. Sci Rep, 15: 3655. doi: 10.1038/s41598-025-87233-y

Al-Najjar H A H, Pradhan B, Beydoun G, et al. 2022. A novel method using explainable artificial intelligence (XAI)-based Shapley Additive Explanations for spatial landslide prediction using Time-Series SAR dataset. Gondwana Res, 164: 185 – 203. doi: 10.1016/j.gr.2022.08.004

Althoff D, Bazame H C, Nascimento J G. 2021. Untangling hybrid hydrological models with explainable artificial intelligence. H_2Open J, 4: 13–28. doi: 10.2166/h2oj.2021.066

Alvarez-Melis D, Jaakkola T S. 2018. Towards robust interpretability with self-explaining neural networks. In: Bengio S, Wallach H, Larochelle H, et al., eds. Advances in Neural Information Processing Systems 31. Red Hook: Curran Associates, Inc. 7786–7795. doi:10.5555/3327345.3327498.

Anders C J, Weber L, Neumann D, et al. 2022. Finding and removing Clever Hans: Using explanation methods to debug and improve deep models. Inf Fusion, 77: 261–295. doi: 10.1016/j.inffus.2021.07.015

Anderson S, Radić V. 2022. Evaluation and interpretation of convolutional long short-term memory networks for regional hydrological modelling. Hydrol Earth Syst Sci, 26: 795–825. doi: 10.5194/hess-26-795-2022

Apley D W, Zhu J. 2020. Visualizing the effects of predictor variables in black box supervised learning models. JJ R Stat Soc Series B Stat Methodol, 82(4): 1059 – 1086. doi: 10.1111/rssb.12377

Arras L, Osman A, Samek W. 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. Inf Fusion, 81: 14–40. doi: 10.1016/j.inffus.2021.11.008

Bach S, Binder A, Montavon G, et al. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE, 10: e0130140. doi: 10.1371/journal.pone.0130140

Bai Z, Liu Q, Liu Y. 2022. Groundwater potential mapping in Hubei region of China using machine learning, ensemble learning, deep learning and AutoML methods. Nat Resour Res, 31: 2549–2569. doi: 10.1007/s11053-022-10100-4

Muhammad M B, Yeasin M. 2021. Eigen-CAM: Visual explanations for deep convolutional neural networks. SN Comput Sci, 2: 47. doi: 10.1007/s42979-021-00449-3

Barnes E A, Toms B, Hurrell J W, et al. 2020. Indicator patterns of forced change learned by an artificial neural network. J Adv Model Earth Syst, 12:e2020MS002195. doi: 10.1029/2020MS002195

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion, 58: 82–115. doi: 10.1016/j.inffus.2019.12.012

Bau D, Zhu J Y, Strobelt H, et al. 2019. Visualizing and understanding generative adversarial networks (extended abstract). arXiv doi:10.48550/arXiv.1901.09887

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

Belaid M K, Bornemann R, Rabus M, et al. 2023. Compare-xAI: Toward unifying functional testing methods for post-hoc XAI algorithms into a multi-dimensional benchmark. In: Longo L, ed, Explainable Artificial Intelligence: First World Conference, xAI 2023, Lisbon, Portugal, July 26 – 28, 2023, Proceedings, Part II. Cham: Springer . 88–109

Berner L T, Massey R, Jantz P, et al. 2020. Summer warming explains widespread but not uniform greening in the Arctic tundra biome. Nat Commun, 11: 4621. doi: 10.1038/s41467-020-18479-5

Bi K, Xie L, Zhang H, et al. 2023. Accurate medium-range global weather forecasting with 3D neural networks. Nature, 619: 533–538. doi: 10.1038/s41586-023-06185-3

Biran O, and Cotton C. 2017. Explanation and justification in machine learning: A survey. In: IJCAI-17 Workshop on Explainable AI (XAI), Melbourne: IJCAI. 8-13.

Blackwell W J. 2012. Neural network Jacobian analysis for high-resolution profiling of the atmosphere. EURASIP J Adv Signal Process, 2012: 71. doi: 10.1186/1687-6180-2012-71

Bommer P L, Kretschmer M, Hedström A, et al. 2024. Finding the right XAI method—A guide for the evaluation and ranking of explainable AI methods in climate science. Artif Intell Earth Syst, 3: e230074. doi: 10.1175/AIES-D-23-0074.1

Bountzouklis C, Fox D M, Bernardino E D. 2023. Predicting wildfire ignition causes in Southern France using explainable artificial intelligence (XAI) methods. Environ Res Lett, 18: 044038. doi: 10.1088/1748-9326/acc8ee

Breiman L. 2001. Random forests. Mach Learn, 45: 5–32. doi: 10.1023/A:1010933404324

Burges C J C. 2010. Dimension reduction: A guided tour. Found Trends Mach Learn, 2: 275–365. doi: 10.1561/2200000002.

Camps-Valls G, Fernández-Torres M Á, Cohrs K H, et al. 2025. Artificial intelligence for modeling and understanding extreme weather and climate events. Nat Commun, 16: 1919. doi: 10.1038/s41467-025-56573-8

Camps-Valls G. 2025. Explaining what, exactly? A critical appraisal of why XAI fails science. Medium. https://medium.com/@gcampsvalls/explaining-what-exactly-a-critical-appraisal-of-why-xai-fails-science-bc0e821c8531

Carloni G, Berti A, Colantonio S. 2025. The role of causality in explainable artificial intelligence. WIREs Data Min Knowl Discov, 15(2). doi: 10.1002/widm.70015

Carter E, Herrera D A, Steinschneider S. 2021. Feature engineering for subseasonal-to-seasonal warm-Season precipitation forecasts in the midwestern United States: Toward a unifying hypothesis of anomalous warm-season hydroclimatic circulation. J Climate, 34: 8291-8318. doi: 10.1175/JCLI-D-20-0264.1

Castangia M, Grajales L M, Aliberti A, et al. 2023. Transformer neural networks for interpretable flood forecasting. Environ Modell Softw, 160: 105581. doi: 10.1016/j.envsoft.2022.105581

Chakraborty D, Başağaoğlu H, Gutierrez L, et al. 2021. Explainable AI reveals new hydroclimatic insights for ecosystem-centric groundwater management. Environ Res Lett, 16: 114024. doi: 10.1088/1748-9326/ac2fde

Chakraborty D, Başağaoğlu H, Winterle J. 2021. Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. Expert Syst Appl, 170: 114498. doi: 10.1016/j.eswa.2020.114498

Chen C, Liu Y, Li Y, et al. 2024. Explainable artificial intelligence framework for urban global digital elevation model correction based on the SHapley additive explanation-random forest algorithm considering spatial heterogeneity and factor optimization. Int J Appl Earth Obs Geoinf, 129: 103843. doi: 10.1016/j.jag.2024.103843

Chen H, Gomez C, Huang C M, et al. 2022. Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. npj Digit Med, 5: 156. doi: 10.1038/s41746-022-00699-2

Chen H, Lundberg S, Lee S-I. 2019. Explaining models by propagating Shapley values of local components. arXiv. doi: 10.48550/arXiv.1911.11888

Chen M, Qian Z, Boers N, et al. 2023. Iterative integration of deep learning in hybrid Earth surface system modelling. Nat Rev Earth Environ, 4(8): 568–581. doi: 10.1038/s43017-023-00452-7

Chen J, Zhang H, Fan M, et al. 2021. Machine-learning-based prediction and key factor identification of the organic carbon in riverine floodplain soils with intensive agricultural practices. JJ Soils Sediments, 21: 2896–2907. doi: 10.1007/s11368-021-02987-y

Chen L, Zhong X, Zhang F, et al. 2023. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. npj Clim Atmos Sci, 6: 190. doi: 10.1038/s41612-023-00512-1

Chen S, Huang J, Huang J-C, 2023. Improving daily streamflow simulations for data-scarce watersheds using the coupled SWAT-LSTM approach. JJ Hydrol, 622: 129734. doi: 10.1016/j.jhydrol.2023.129734

Chen X, Duan Y, Houthooft R, et al. 2016. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: Lee D D, Sugiyama M, Luxburg U V, et al., eds. Advances in Neural Information Processing Systems 29. Red Hook: Curran Associates, Inc. 2172–2180. doi:10.5555/3157382.3157494

Cho H and Ackom E. 2025. Artificial intelligence (AI)-driven approach to climate action and sustainable development. Nat Commun, 16: 1228. doi: 10.1038/s41467-024-53956-1

Chou Y L, Moreira C, Bruza P, et al.. 2022. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. Inf Fusion, 81: 59–83. doi: 10.1016/j.inffus.2021.11.003

Clark K, Khandelwal U, Levy O, et al. 2019. What does BERT look at? An analysis of BERT's attention. In: Linzen T, Chrupała G, Belinkov Y, et al., eds. Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Florence: Association for Computational Linguistics. 276–286. doi: 10.18653/v1/W19-4828

Clement T, Kemmerzell N, Abdelaal M, et al. 2023. XAIR: A systematic metareview of explainable AI (XAI) aligned to the software development process. Mach Learn Knowl Extr, 5: 78–108. doi: 10.3390/make5010006

Covert I, Lee S-I. 2021. Improving KernelSHAP: Practical Shapley value estimation via linear regression. arXiv. doi: 10.48550/arXiv.2012.01536

Cui X, Zhou F, Ciais P, et al. 2021. Global mapping of crop-specific emission factors highlights hotspots of nitrous oxide mitigation. Nat Food, 2: 886–893. doi: 10.1038/s43016-021-00384-9

Debnath R, Creutzig F, Sovacool B K, et al. 2023. Harnessing human and machine intelligence for planetary-level climate action. npj Clim Action, 2: 20. doi: 10.1038/s44168-023-00056-3

Díaz-Vallejo M, Peña-Peniche A, Mota-Vargas C, et al. 2024. Analyses of the variable selection using correlation methods: An approach to the importance of statistical inferences in the modelling process. Ecol Model,, 498: 110893. doi: 10.1016/j.ecolmodel.2024.110893

Diffenbaugh N S, Barnes E A. 2023. Data-driven predictions of the time remaining until critical global warming thresholds are reached. Proc Natl Acad Sci USA, 120: e2207183120. doi: 10.1073/pnas.2207183120

Dikshit A, Pradhan B, Santosh M. 2022. Artificial neural networks in drought prediction in the 21st century—A scientometric analysis. Appl Soft Comput, 114: 108080. doi: 10.1016/j.asoc.2021.108080

Dikshit A, Pradhan B, Matin S S, et al. 2024. Artificial Intelligence: A new era for spatial modelling and interpreting climate-induced hazard assessment. Geosci Front, 15(4): 101815. doi: 10.1016/j.gsf.2024.101815

Ding K, Zhao X, Cheng J, et al. 2025. GRACE/ML-based analysis of the spatiotemporal variations of groundwater storage in Africa. J Hydrol, 647: 132336. doi: 10.1016/j.jhydrol.2024.132336

Dramsch J S, Kuglitsch M M, Fernández-Torres M Á, et al. 2025. Explainability can foster trust in artificial intelligence in geoscience. Nat Geosci, 18: 112–114. doi: 10.1038/s41561-025-01639-x

Dueben P D, Schultz M G, Chantry M, et al. 2022. Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook. Artif Intell Earth Syst, 1(3): e210002. doi: 10.1175/AIES-D-21-0002.1

Ekmekcioğlu Ö, Koc K, Özger M. 2021. District based flood risk assessment in Istanbul using fuzzy analytical hierarchy process. Stoch Environ Res Risk Assess, 35: 617–637. doi: 10.1007/s00477-020-01924-8

Elshawi R, Al-Mallah MH, Sakr S. 2019. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak, 19: 146. doi: 10.1186/s12911-019-0874-0

Erion G, Janizek J D, Sturmfels P, et al. 2021. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nat Mach Intell, 3(7): 620–631. doi: 10.1038/s42256-021-00343-w

European Commission Directorate-General for Communications Networks Content and Technology. 2021. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Evans M E, Adler P B, Angert A L, et al. 2025. Reconsidering space-for-time substitution in climate change ecology. Nat Clim Change, 15(8): 809 – 812. doi: 10.1038/s41558-025-02392-0

Eyring V, Collins W D, Gentine P, et al. 2024. Pushing the frontiers in climate modelling and analysis with machine learning. Nat Clim Chang, 14: 916–928. doi: 10.1038/s41558-024-02095-y

Fayaz J, Galasso C. 2024. Interpretability and spatial efficacy of a deep-learning-based on-site early warning framework using explainable artificial intelligence and geographically weighted random forests. Geosci Front, 15: 101839. doi: 10.1016/j.gsf.2024.101839

Feng J, Li J, Xu C Y, et al. 2024. Viewing soil moisture flash drought onset mechanism and their changes through XAI lens: A case study in Eastern China. Water Resour Res, 60: e2023WR036297. doi: 10.1029/2023WR036297

Feng P, Wang B, Liu D L, et al. 2019. Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in south-eastern Australia. Agr Syst, 173: 303–316. doi: 10.1016/j.agsy.2019.03.015

Feng P, Wang B, Luo J J, et al. 2020. Using large-scale climate drivers to forecast meteorological drought condition in growing season across the Australian wheatbelt. Sci Total Environ, 724: 138162. doi: 10.1016/j.scitotenv.2020.138162

Fisher A, Rudin C, Dominici F. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res, 20(177): 1–81. doi: 10.5555/3455716.3455788

Friedman J H, Popescu B E. 2008. Predictive learning via rule ensembles. Ann Appl Stat, 2(3): 916-954. doi: 10.1214/07-AOAS148

Fu W, Moore J K, Primeau F, et al. 2022. Evaluation of ocean biogeochemistry and carbon cycling in CMIP Earth System Models with the International Ocean Model Benchmarking (IOMB) software system. J Geophys Res Oceans, 127: e2022JC018965. doi: 10.1029/2022JC018965

Fu Z, Hu L, Chen Z, et al. 2020. Estimating spatial and temporal variation in ocean surface pCO_2 in the Gulf of Mexico using remote sensing and machine learning techniques. Sci Total Environ, 745: 140965. doi: 10.1016/j.scitotenv.2020.140965

Gagne D J, Haupt S E, Nychka D W, et al. 2019. Interpretable deep learning for spatial analysis of

severe hailstorms. Mon Weather Rev, 147: 2827–2845. doi: 10.1175/MWR-D-18-0316.1
Gallegos M, Vassilev-Galind V, Poltavsky I, et al. 2024. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. Nat Commun, 15: 4345. doi: 10.1038/s41467-024-48567-9
Gautam S, Höhne M M C, Hansen S, et al. 2023. This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. Pattern Recognit, 136: 109172. doi: 10.1016/j.patcog.2022.109172
Gevaert C M. 2022. Explainable AI for earth observation: A review including societal and regulatory perspectives. Int J Appl Earth Obs Geoinf, 112: 102869. doi: 10.1016/j.jag.2022.102869
Ghada W, Casellas E, Herbinger J, et al. 2022. Stratiform and convective rain classification using machine learning models and micro rain radar. Remote Sens, 14(18): 4563. doi: 10.3390/rs14184563
Gibson P B, Chapman W E, Altinok A, et al. 2021. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. Commun Earth Environ, 2: 159. doi: 10.1038/s43247-021-00225-4
Gilbert J A, Zengler K. 2025. The role of artificial intelligence in microbial sciences to support climate resilience. Nat Rev Microbiol, 23: 333–334. doi: 10.1038/s41579-025-01178-7
Gimeno-Sotelo L, Bevacqua E, Gimeno L. 2023. Combinations of drivers that most favor the occurrence of daily precipitation extremes. Atmos Res, 294: 106959. doi: 10.1016/j.atmosres.2023.106959
Goldstein A, Kapelner A, Bleich J, et al. 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Statist, 24: 44–65. doi: 10.1080/10618600.2014.907095
Gordon E M, Barnes E A, Hurrell J W. 2021. Oceanic harbingers of Pacific Decadal Oscillation predictability in CESM2 detected by neural networks. Geophys Res Lett, 48: e2021GL095392. doi: 10.1029/2021GL095392
Gunning D, Stefik M, Choi J, et al. 2019. XAI—Explainable artificial intelligence. Sci Robot, 4(37): eaay7120. doi: 10.1126/scirobotics.aay7120
Gunning D, Vorm E, Wang J Y, et al. 2021. DARPA 's explainable AI (XAI) program: A retrospective. Appl AI Lett, 2(4): e61. doi: 10.1002/ail2.61
Guo W, Mu D, Xu J, et al. 2018. LEMNA: Explaining deep learning based security applications. In: Lie D, Mannan M, Backes M, Wang X, et al., eds. CCS '18: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York: Association for Computing Machinery. 364–379. doi: 10.1145/3243734.3243792
Gurumoorthy K S, Dhurandhar A, Cecchi G, et al. 2019. Efficient data representation by selecting prototypes with importance weights. In: Taniar D, Wang H, Li J, et al., eds. 2019 IEEE International Conference on Data Mining (ICDM). Piscataway, NJ: IEEE. 262–271. doi: 10.1109/ICDM.2019.00036
Haaf D, Six J, Doetterl S. 2021. Global patterns of geo-ecological controls on the response of soil respiration to warming. Nat Clim Chang, 11: 623–627. doi: 10.1038/s41558-021-01068-9
Ham Y G, Kim J H, Min S K, et al. 2023. Anthropogenic fingerprints in daily precipitation revealed by deep learning. Nature, 622: 301–307. doi: 10.1038/s41586-023-06474-x
Hao H, Hao Y, Li Z, et al. 2024. Insight into glacio-hydrologicalprocesses using explainable

machine-learning (XAI) models. J Hydrol, 634: 131047. doi: 10.1016/j.jhydrol.2024.131047

Hasanpour Zaryabi E, Moradi L, Kalantar B, et al. 2022. Unboxing the black box of attention mechanisms in remote sensing big data using XAI. Remote Sens, 14(24): 6254. doi: 10.3390/rs14246254

Haupt S E, Chapman W, Adams S V, et al. 2021. Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop. Phil Trans R Soc A, 379(2194): 20200091. doi: 10.1098/rsta.2020.0091

He F, Liu S, Mo X, et al. 2025. Interpretable flash flood susceptibility mapping in Yarlung Tsangpo River Basin using H_2O Auto-ML. Sci Rep, 15: 1702. doi: 10.1038/s41598-024-84655-y

Heimann M, Reichstein M. 2008. Terrestrial ecosystem carbon dynamics and climate feedbacks. Nature, 451: 289–292. doi: 10.1038/nature06591

Hoffman R R, Jalaeian M, Tate C, et al. 2023. Evaluating machine-generated explanations: a “Scorecard” method for XAI measurement science. Front Comput Sci, 5: 1114806. doi: 10.3389/fcomp.2023.1114806

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. Stat Comput, 31: 82. doi: 10.1007/s11222-021-10057-z

Hsu H, Dirmeyer P A. 2023. Soil moisture-evaporation coupling shifts into new gears under increasing CO_2. Nat Commun, 14: 1162. doi: 10.1038/s41467-023-36794-5

Hu X, Shi L, Lin G, et al. 2021. Comparison of physical-based, data-driven and hybrid modeling approaches for evapotranspiration estimation. J Hydrol, 601: 126592. doi: 10.1016/j.jhydrol.2021.126592

Huang F, Zhang Y, Zhang Y, et al. 2023. Towards interpreting machine learning models for predicting soil moisture droughts. Environ Res Lett, 18: 074002. doi: 10.1088/1748-9326/acdbe0

Huang F, Shangguan W, Li Q, et al. 2023. Beyond prediction: An integrated post-hoc approach to interpret complex model in hydrometeorology. Environ Modell Softw, 167: 105762. doi: 10.1016/j.envsoft.2023.105762

Huang F, Zhang Y, Zhang Y, et al. 2023. Interpreting Conv-LSTM for spatio-temporal soil moisture prediction in China. Agriculture, 13(5): 971. doi: 10.3390/agriculture13050971

Huang F, Shangguan W, Jiang S. 2024. Identifying potential drivers of land-atmosphere coupling variation under climate change by explainable artificial intelligence. In: EGU General Assembly 2024. Vienna: European Geosciences Union. EGU24-7202.

Huang X, Kroening D, Ruan W, et al. 2020. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and Interpretability. Comput Sci Rev, 37: 100270. doi: 10.1016/j.cosrev.2020.100270

Ibrahim R, Shafiq M O. 2022. Augmented score-CAM: High resolution visual interpretations for deep neural networks. Knowl-Based Syst, 252: 109287. doi: 10.1016/j.knosys.2022.109287

Irrgang C, Boers N, Sonnewald M, et al. 2021. Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. Nat Mach Intell, 3(8): 667 – 674. doi: 10.1038/s42256-021-00374-3

Ishida S, Terayama K, Kojima R, et al. 2019. Prediction and interpretable visualization of retrosynthetic reactions using graph convolutional networks. J Chem Inf Model, 59(12):

5026–5033. doi: 10.1021/acs.jcim.9b00538

Ivanovs M, Kadikis R, Ozols K. 2021. Perturbation-based methods for explaining deep neural networks: A survey. Pattern Recognit Lett, 150: 228–234. doi: 10.1016/j.patrec.2021.06.030

Janssens M, Vilà ‐ Guerau De Arellano J, Scheffer M, et al. 2021. Cloud patterns in the trades have four interpretable dimensions. Geophys Res Lett, 48(5): e2020GL091001. doi: 10.1029/2020GL091001

Jiang S, Bevacqua E, Zscheischler J. 2022. River flooding mechanisms and their changes in Europe revealed by explainable machine learning. Hydrol Earth Syst Sci, 26: 6339–6359. doi: 10.5194/hess-26-6339-2022

Jiang S, Sweet L, Blougouras G, et al. 2024. How interpretable machine learning can benefit process understanding in the geosciences. Earths Future, 12: e2024EF004540. doi: 10.1029/2024EF004540

Jiang S, Tarasova L, Yu G, et al. 2024. Compounding effects in flood drivers challenge estimates of extreme river floods. Sci Adv, 10(13): eadl4005. doi: 10.1126/sciadv.adl4005

Jiang S, Zheng Y, Wang C, et al. 2022. Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments. Water Resour Res, 58(1): e2021WR030185. doi: 10.1029/2021WR030185

Jing H, He X, Tian Y, et al. 2023. Comparison and interpretation of data-driven models for simulating site-specific human-impacted groundwater dynamics in the North China Plain. J Hydrol, 616: 128751. doi: 10.1016/j.jhydrol.2022.128751

Ke E, Zhao J, Zhao Y. 2025. Investigating the influence of nonlinear spatial heterogeneity in urban flooding factors using geographic explainable artificial intelligence. J Hydrol, 648: 132398. doi: 10.1016/j.jhydrol.2024.132398

Khose S B, Mailapalli D R. 2024. Spatial mapping of soil moisture content using very-high resolution UAV-based multispectral image analytics. Smart Agric Technol, 8: 100467. doi: 10.1016/j.atech.2024.100467

Kim S, Choi Junho, Choi Y, et al. 2023. Explainable AI-based interface system for weather forecasting model. In: Stephanidis C, Kurosu M, Degen H, et al., eds. HCI International 2023 – Late Breaking Papers. Berlin, Heidelberg: Springer. 101–119. doi: 10.1007/978-3-031-48057-7_7

Klein L, Ziegler S, Laufer F, et al. 2024. Discovering process dynamics for scalable perovskite solar cell manufacturing with explainable AI. Adv Mater, 36: 2307160. doi: 10.1002/adma.202307160

Kondylatos S, Prapas I, Ronco M, et al. 2022. Wildfire danger prediction and understanding with deep learning. Geophys Res Lett, 49: e2022GL099368. doi: 10.1029/2022GL099368

Krasting J P, De Palma M, Sonnewald M, et al. 2022. Regional sensitivity patterns of Arctic Ocean acidification revealed with machine learning. Commun Earth Environ, 3: 91. doi: 10.1038/s43247-022-00419-4

Krell E, Mamalakis A, King S A, et al. 2025. The influence of correlated features on neural network attribution methods in geoscience. Environ Data Sci, 4: e29. doi: 10.1017/eds.2025.19

Kwon M, Jeong J, Uh Y. 2022. Diffusion models already have a semantic latent space. arXiv. doi: 10.48550/arXiv.2210.10960

Labe Z M, Barnes E A. 2021. Detecting climate signals using explainable AI with single-forcing large ensembles. J Adv Model Earth Syst, 13: e2021MS002464. doi: 10.1029/2021MS002464

Lam R, Sanchez-Gonzalez A, Willson M, et al. 2023. Learning skillful medium-range global weather forecasting. Science, 382(6677): 1416 – 1421. doi: 10.1126/science.adi2336

Lee S, Lee G, Kim H, et al. 2023. Sequential Data Generation with Groupwise Diffusion Process. arXiv. doi: 10.48550/arXiv.2310.01400

Lee Y, Cho D, Im J, et al. 2024. Unveiling teleconnection drivers for heatwave prediction in South Korea using explainable artificial intelligence. npj Clim Atmos Sci, 7: 176. doi: 10.1038/s41612-024-00722-1

Lees T, Reece S, Kratzert F, et al. 2022. Hydrological concept formation inside long short-term memory (LSTM) networks. Hydrol Earth Syst Sci, 26(12): 3079 – 3101. doi: 10.5194/hess-26-3079-2022

Ley A, Bormann H, Casper M. 2024. Linking explainable artificial intelligence and soil moisture dynamics in a machine learning streamflow model. Hydrol Res, 55: 613–627. doi: 10.2166/nh.2024.003

Li L, Tang S, Zhang Y, et al. 2018. GLA: Global–local attention for image description. IEEE Trans Multimedia, 20: 726–737. doi: 10.1109/TMM.2017.2751140

Li T, Yu Y, Wang X, et al. 2025. A review of aerosol-cloud interactions: Mechanisms, climate effects, and observation methods. Atmos Res, 325: 108267. doi: 10.1016/j.atmosres.2025.108267

Li W, Liu C, Xu Y, et al. 2024. An interpretable hybrid deep learning model for flood forecasting based on Transformer and LSTM. J Hydrol Reg Stud, 54: 101873. doi: 10.1016/j.ejrh.2024.101873

Li W, Reichstein M, O S, et al. 2023. Contrasting drought propagation into the terrestrial water cycle between dry and wet regions. Earths Future, 11(7): e2022EF003441. doi: 10.1029/2022ef003441

Li X, Feng M, Ran Y, et al. 2023. Big Data in Earth system science and progress towards a digital twin. Nat Rev Earth Environ, 4(5): 319–332. doi: 10.1038/s43017-023-00409-w

Li X, Guo Y. 2025. Paradigm shifts from data-intensive science to robot scientists. Sci Bull, 70(1): 14−18. doi: 10.1016/j.scib.2024.09.029

Lian X, Piao S, Huntingford C, et al. 2018. Partitioning global land evapotranspiration using CMIP5 models constrained by observations. Nat Clim Change, 8(7): 640 – 646. doi: 10.1038/s41558-018-0207-9

Liu J, Huang W, Li H, et al. 2023. SLAFusion: Attention fusion based on SAX and LSTM for dangerous driving behavior detection. Inf Sci, 640: 119063. doi: 10.1016/j.ins.2023.119063

Liu J, Ren K, Ming T, et al. 2023. Investigating the effects of local weather, streamflow lag, and global climate information on 1-month-ahead streamflow forecasting by using XGBoost and SHAP: two case studies involving the contiguous USA. Acta Geophys, 71: 905–925. doi: 10.1007/s11600-022-00928-y

Liu M, Trugman A T, Peñuelas J, et al. 2024. Climate-driven disturbances amplify forest drought sensitivity. Nat Clim Change, 14: 746–752. doi: 10.1038/s41558-024-02022-1

Liu, Q, Gui, D, Zhang L, et al. 2022. Simulation of regional groundwater levels in arid regions using interpretable machine learning models. Sci Total Environ, 831: 154902. doi: 10.1016/j.scitotenv.2022.154902

Liu S, Xu J, Wang R, et al. 2025. Investigating the causal effects of anthropogenic factors on urban streams and lakes water quality by integrating causal inference with interpretable machine learning. J Clean Prod, 488: 144626. doi: 10.1016/j.jclepro.2024.144626

Liu Y, Duffy K, Dy J G, et al. 2023. Explainable deep learning for insights in El Niño and river flows. Nat Commun, 14: 339. doi: 10.1038/s41467-023-35968-5

Lohit S, Liu D, Mansour H, et al. 2019. Unrolled projected gradient descent for multi-spectral image fusion. In: Paliwal K K, Sanei S, eds. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE. 7720–7724. doi: 10.1109/ICASSP.2019.8683445

Lundberg S, Lee S I. 2017. A unified approach to interpreting model predictions. arXiv. https://arxiv.org/abs/1705.07874.

Lyu Y, Yong B. 2025. Using an explainable machine learning approach to produce high-resolution hourly precipitation estimates for a typical data-deficiency basin. J Geophys Res Mach Learn Comput, 2: e2024JH000489. doi: 10.1029/2024JH000489

Maddy E S, Boukabara S A. 2021. MIIDAPS-AI: An explainable machine-learning algorithm for infrared and microwave remote sensing and data assimilation preprocessing - Application to LEO and GEO sensors. IEEE J Sel Top Appl Earth Obs Remote Sens, 14: 8566 – 8576. doi: 10.1109/JSTARS.2021.3104389

Mahapatra D, Poellinger A, Shao L, et al. 2021. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE Trans Med Imaging, 40(10): 2548–2562. doi: 10.1109/TMI.2021.3061724

Mahecha M D, Gans F, Brandt G, et al. 2020. Earth system data cubes unravel global multivariate dynamics. Earth Syst Dynam, 11: 201–234. doi: 10.5194/esd-11-201-2020

Maity R, Srivastava A, Sarkar S, et al. 2024. Revolutionizing the future of hydrological science: Impact of machine learning and deep learning amidst emerging explainable AI and transfer learning. Appl Comput Geosci, 24: 100206. doi: 10.1016/j.acags.2024.100206

Mallik S, Chakraborty A, Podder K, et al. 2025. Enhancing soil moisture prediction with explainable AI: Integrating IoT and multi-sensor remote sensing data through soft computing. Appl Soft Comput, 180: 113406. doi: 10.1016/j.asoc.2025.113406

Mamalakis A, Barnes E A, Ebert-Uphoff I. 2022. Investigating the fidelity of rxplainable artificial intelligence methods for applications of convolutional neural networks in geoscience. Artif Intell Earth Syst, 1: e220012. doi: 10.1175/AIES-D-22-0012.1

Mamalakis A, Ebert-Uphoff I, Barnes E A. 2022. Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. Environ Data Sci, 1: e8. doi: 10.1017/eds.2022.7

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. Artif Intell Earth Syst, 2: e220058. doi: 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: Holzinger A, Goebel R, Fong R, et al., eds. xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers. Cham: Springer International Publishing. 315 – 339. doi: 10.1007/978-3-031-04083-2_16

Martin Z K, Barnes E A, Maloney E. 2022. Using simple, explainable neural networks to predict the Madden-Julian Oscillation. J Adv Model Earth Syst, 14(7): e2021MS002774. doi:

10.1029/2021MS002774

McCloskey K, Taly A, Monti F, et al. 2019. Using attribution to decode binding mechanism in neural network models for chemistry. Proc Natl Acad Sci USA, 116(24): 11624-11629. doi: 10.1073/pnas.1820657116

McGovern A, Gagne D J, Wirz C D, et al. 2023. Trustworthy artificial intelligence for environmental sciences: An innovative approach for summer school. Bull Am Meteorol Soc, 104: E1222–E1231. doi: 10.1175/BAMS-D-22-0225.1

McGovern A, Demuth J, Bostrom A, et al. 2024. The value of convergence research for developing trustworthy AI for weather, climate, and ocean hazards. npj Nat Hazards, 1: 13. doi: 10.1038/s44304-024-00014-x

McMillan H, Araki R, Bolotin L, et al. 2025. Global patterns in observed hydrologic processes. Nat Water, 3: 497–506. doi: 10.1038/s44221-025-00407-w

Melis D A, Jaakkola T. 2018. Towards robust interpretability with self-Explaining neural networks. arXiv. doi: 10.48550/arXiv.1806.07538

Mi J X, Jiang X, Luo L, et al. 2024. Toward explainable artificial intelligence: A survey and overview on their intrinsic properties. Neurocomputing, 563: 126919. doi: 10.1016/j.neucom.2023.126919

Minh D, Wang H X, Li Y F, et al. 2022. Explainable artificial intelligence: A comprehensive review. Artif Intell Rev, 55(5): 3503–3568. doi: 10.1007/s10462-021-10088-y

Mo Y, Xu J, Zhu S, et al. 2025. Spatial heterogeneity of groundwater depths in coastal cities and their responses to multiple factors interactions by interpretable machine learning models. Geosci Front, 16: 102033. doi: 10.1016/j.gsf.2025.102033

Montero D, Kraemer G, Anghelea A, et al. 2024. Earth System Data Cubes: Avenues for advancing Earth system research. Environ Data Sci, 3: e27. doi: 10.1017/eds.2024.22

Motteler H E, Strow L L, McMillin L, et al. 1995. Comparison of neural networks and regression based methods for temperature retrievals. Appl Opt, 34(24): 5390-5397. doi: 10.1364/AO.34.005390

Mukhamediev R I, Popova Y, Kuchin Y, et al. 2022. Review of Artificial Intelligence and Machine Learning Technologies: Classification, restrictions, opportunities and challenges. Mathematics, 10(15): 2552. doi: 10.3390/math10152552

Murdoch W J, Singh C, Kumbier K, et al. 2019. Interpretable machine learning: Definitions, methods, and applications. Proc Natl Acad Sci USA, 116: 22071–22080. doi: 10.1073/pnas.1900654116

Mylonas N, Mollas I, Bassiliades N, et al. 2024. Exploring local interpretability in dimensionality reduction: Analysis and use cases. Expert Syst Appl, 252: 124074. doi: 10.1016/j.eswa.2024.124074

Naiseh M, Al-Thani D, Jiang N, et al. 2023. How the different explanation classes impact trust calibration: The case of clinical decision support systems. Int J Hum Comput Stud, 169: 102941. doi: 10.1016/j.ijhcs.2022.102941

Narteni S, Orani V, Ferrari E, et al. 2025. Explainable evaluation of generative adversarial networks for wearables data augmentation. Eng Appl Artif Intelle, 145: 110133. doi: 10.1016/j.engappai.2025.110133

Nauta M, Trienes J, Pathak S, et al. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Comput Surv, 55(13s): 1 –

42. doi: 10.1145/3583558

Novielli P, Magarelli M, Romano D, et al. 2025. Leveraging explainable AI to predict soil respiration sensitivity and its drivers for climate change mitigation. Sci Rep, 15: 12527. doi: 10.1038/s41598-025-96216-y

O'Loughlin R J, Li D, Neale R, et al. 2025. Moving beyond post hoc explainable artificial intelligence: A perspective paper on lessons learned from dynamical climate modeling. Geosci Model Dev, 18: 787–802. doi: 10.5194/gmd-18-787-2025

Orynbaikyzy A, Gessner U, Mack B, et al. 2020. Crop type classification using fusion of Sentinel-1 and Sentinel-2 data: Assessing the impact of feature selection, optical data availability, and parcel sizes on the accuracies. Remote Sens, 12: 2779. doi: 10.3390/rs12172779

Oukawa G Y, Krecl P, Targino A C. 2022. Fine-scale modeling of the urban heat island: A comparison of multiple linear regression and random forest approaches. Sci Total Environ, 815: 152836. doi: 10.1016/j.scitotenv.2021.152836

Pan X, Chen D, Pan B, et al. 2025. Evolution and prospects of Earth system models: Challenges and opportunities. Earth-Sci Rev, 260: 104986. doi: 10.1016/j.earscirev.2024.104986

Patoine G, Eisenhauer N, Cesarz S, et al. 2022. Drivers and trends of global soil microbial carbon over two decades. Nat Commun, 13: 4195. doi: 10.1038/s41467-022-31833-z

Patro B, Lunayach M, Patel S, et al. 2019. U-CAM: Visual explanation using uncertainty based class activation maps. In: IEEE/CVF International Conference on Computer Vision (ICCV) Piscataway: IEEE. 7444–7453. doi: 10.1109/ICCV.2019.00754

Peng Z, Zhang B, Wang D, et al. 2024. Application of machine learning in atmospheric pollution research: A state-of-art review. Sci Total Environ, 910: 168588. doi: 10.1016/j.scitotenv.2023.168588

Playout C, Duval R, Boucher M C, et al. 2022. Focused Attention in Transformers for interpretable classification of retinal images. Med Image Anal, 82: 102608. doi: 10.1016/j.media.2022.102608

Raghu M, Schmidt E. 2020. A survey of deep learning for scientific discovery. arXiv doi: 10.48550/arXiv.2003.11755

Raghu M, Gilmer J, Yosinski J, et al. 2017. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In: Guyon I, Luxburg U V, Bengio S, et al., eds. Advances in Neural Information Processing Systems 30. New York: Curran Associates Inc. 6076–6085. doi: 10.48550/arXiv.1706.05806

Rahwan I, Cebrian M, Obradovich N, et al. 2019. Machine behaviour. Nature, 568(7753): 477–486. doi: 10.1038/s41586-019-1138-y

Ramirez S G, Hales R C, Williams G P, et al. 2022. Extending SC-PDSI-PM with neural network regression using GLDAS data and Permutation Feature Importance. Environ Modell Softw, 157: 105475. doi: 10.1016/j.envsoft.2022.105475

Reichstein M, Camps-Valls G, Stevens B, et al. 2019. Deep learning and process understanding for data-driven Earth system science. Nature, 566: 195–204. doi: 10.1038/s41586-019-0912-1

Reidmiller D R, Avery C W, Easterling D R, et al. 2017. Impacts, risks, and adaptation in the United States: Fourth national climate assessment, volume II. Washington DC: U.S. Global Change Research Program. 1515

Reunanen, J. 2003. Overfitting in making comparisons between variable selection methods. J

Mach Learn Res, 3: 1371–1382. https://dl.acm.org/doi/doi: 10.5555/944919.944978

Ribeiro M T, Singh S, Guestrin C. 2016. "Why Should I Trust You?": Explaining the predictions of any classifier. In: Krishnapuram B, Shah M, Smola A J, et al., eds. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM. 1135 – 1144. doi:10.1145/2939672.2939778

Robert Maier H, Rosa Taghikhah F, Nabavi E, et al. 2024. How much X is in XAI: Responsible use of “Explainable” artificial intelligence in hydrology and water resources. J Hydrol X, 25: 100185. doi: 10.1016/j.hydroa.2024.100185

Rojas C A, Padrão P, Fuentes J, et al. 2024. Combining multi-satellite remote and in-situ sensing for unmanned underwater vehicle state estimation. Ocean Eng, 310: 118708. doi: 10.1016/j.oceaneng.2024.118708

Ross A, Li Z, Perezhogin P, et al. 2023. Benchmarking of machine learning ocean subgrid parameterizations in an idealized model. J Adv Model Earth Syst, 15. doi: 10.1029/2022MS003258

Saha S, Gogoi P, Gayen A, et al. 2021. Constructing the machine learning techniques based spatial drought vulnerability index in Karnataka state of India. J Clean Prod, 314: 128073. doi: 10.1016/j.jclepro.2021.128073

Saabas A. 2014. Interpreting random forests. Diving into data. https://blog.datadive.net/interpreting-random-forests/

Saleem R, Yuan B, Kurugollu F, et al. 2022. Explaining deep neural networks: A survey on the global interpretation methods. Neurocomputing, 513: 165–180. doi: 10.1016/j.neucom.2022.09.129

Schellnhuber H J. 1999. ‘Earth system’ analysis and the second Copernican revolution. Nature, 402: C19–C23. doi: 10.1038/35011515

Schiller D, Huber T, Lingenfelser F, et al. 2019. Relevance-based feature masking: Improving neural network based whale classification through explainable artificial intelligence. In: Kubín G, Kacic Z, eds. Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019). Baixas: ISCA. 2798–2802. doi:10.21437/Interspeech.2019-2707

Schmidt L, Heße F, Attinger S, et al. 2020. Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany. Water Resour Res, 56(5). doi: 10.1029/2019WR025924

Schoenke J, Aschenbruck N, Interdonato R, et al. 2021. Gaia-agstream: An explainable AI platform for mining complex data streams in agriculture. In: Boumerdassi S, Ghogho M, Renault É, eds. Artificial Intelligence in Agriculture. Cham: Springer. 71 – 83. doi: 10.1007/978-3-030-88259-4_6

Schwenke L, Bloemheuvel S, Atzmueller M. 2023. Identifying informative nodes in attributed spatial sensor networks using attention for symbolic abstraction in a GNN-based modeling approach. In: Proceedings of the International FLAIRS Conference(Vol. 36).. Gainesville: Florida OJ. 1–7. doi: 10.32473/flairs.36.133109

Segal-Rozenhaimer M, Li A, Das K, et al. 2020. Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN). Remote Sens Environ, 237: 111446. doi: 10.1016/j.rse.2019.111446

Selvaraju R R, Cogswell M, Das A, et al. 2020. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis, 128: 336–359. doi: 10.1007/s11263-019-01228-7

Shamekh S, Lamb K D, Huang Y, et al. 2023. Implicit learning of convective organization explains precipitation stochasticity. Proc Natl Acad Sci USA, 120: e2216158120. doi: 10.1073/pnas.2216158120

Shangguan W, Hengl T, Mendes De Jesus J, et al. 2017. Mapping the global depth to bedrock for land surface modeling. J Adv Model Earth Syst, 9: 65–88. doi: 10.1002/2016MS000686

Shangguan W, Xiong Z, Nourani V, et al. 2023. A 1 km global carbon flux dataset using in situ measurements and deep learning. Forests, 14: 913. doi: 10.3390/f14050913

Shen C, Appling A P, Gentine P, et al. 2023. Differentiable modelling to unify machine learning and physical models for Geosciences. Nat Rev Earth Environ, 4(8): 552 – 567. doi: 10.1038/s43017-023-00450-9

Shrikumar A, Greenside P, Kundaje A. 2019. Learning important features through propagating activation differences. arXiv. doi:10.48550/arXiv.1704.02685

Shu R, Wu H, Gao Y, et al. 2025. Advanced forecasts of global extreme marine heatwaves through a physics-guided data-driven approach. Environ Res Lett, 20: 044030. doi: 10.1088/1748-9326/adbddd

Simonyan K, Vedaldi A, Zisserman A. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv. doi: 10.48550/arXiv.1312.6034

Sinha S, Vashisht P. 2023. Explainable data fusion on edge: Challenges and opportunities. In: Hassanien A E, Gupta D, Singh A K, et al., eds. Explainable Edge AI: A Futuristic Computing Perspective. Cham: Springer. 117 – 138. doi: 10.1007/978-3-031-18292-1_8

Slater L, Coxon G, Brunner M, et al. 2024. Spatial sensitivity of river flooding to changes in climate and land cover through explainable AI. Earths Future, 12: e2023EF004035. doi: 10.1029/2023EF004035

Smilkov D, Thorat N, Kim B, et al. 2017. SmoothGrad: Removing noise by adding noise. arXiv. doi: 10.48550/arXiv.1706.03825

Song C, Rousseau A N, Song Y, et al 2024. Research progress and perspectives on ecological processes and carbon feedback in permafrost wetlands under changing climate conditions. Fundam Res. doi: 10.1016/j.fmre.2024.05.002

Song C H, Han H J, Avrithis Y. 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Los Alamitos: IEEE. 488 – 497. doi: 10.1109/WACV51458.2022.00051

Sonnewald M, Lguensat R. 2021. Revealing the impact of global heating on North Atlantic circulation using transparent machine learning. J Adv Model Earth Syst, 13(8): e2021MS002496. doi: 10.1029/2021MS002496

Springenberg J T, Dosovitskiy A, Brox T, et al. 2014. Striving for simplicity: The all convolutional net. arXiv. doi:10.48550/arXiv.1412.6806

Steffen W, Richardson K, Rockström J, et al. 2020. The emergence and evolution of Earth System Science. Nat Rev Earth Environ, 1: 54–63. doi: 10.1038/s43017-019-0005-6

Strobelt H, Gehrmann S, Pfister H, et al. 2017. LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE Trans Vis Comput Graph, 24(1): 667-676. doi: 10.1109/TVCG.2017.2744158

Sun T, Sun J, Chen Y, et al. 2022. Improving short-term precipitation forecasting with radar data assimilation and a multiscale hybrid ensemble–variational strategy. Mon Weather Rev, 150: 2357–2377. doi: 10.1175/MWR-D-21-0325.1

Sundararajan M, Taly A, Yan Q. 2017. Axiomatic attribution for deep networks. arXiv. doi: 10.48550/arXiv.1703.01365

Taskin G, Aptoula E, Ertürk A. 2024. Chapter 7 - Explainable AI for Earth observation: Current methods, open challenges, and opportunities. In: Prasad S, Chanussot J, Li J, eds. Advances in Machine Learning and Image Analysis for GeoAI. Amsterdam: Elsevier. 115–152. doi: 10.1016/B978-0-44-319077-3.00012-2

Tavanaei A. 2021. Embedded encoder-decoder in convolutional networks towards explainable AI. arXiv. doi:10.48550/arXiv.2007.06712

Theissler A, Thomas M, Burch M, et al. 2022. ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices. Knowl-Based Syst, 247: 108651. doi: 10.1016/j.knosys.2022.108651

Thuy A, Benoit D F. 2024. Explainability through uncertainty: Trustworthy decision-making with neural networks. Eur J Oper Res, 317: 330–340. doi: 10.1016/j.ejor.2023.09.009

Toms B A, Barnes E A, Ebert-Uphoff I. 2020. Physically interpretable neural networks for the geosciences: Applications to Earth system variability. J Adv Model Earth Syst, 12: e2019MS002002. doi: 10.1029/2019MS002002

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Ullah I, Rios A, Gala V, et al. 2022. Explaining deep learning models for tabular data using layer-wise relevance propagation. Appl Sci, 12(1): 136. doi: 10.3390/app12010136

Upadhyaya S A, Kirstetter P, Kuligowski R J, et al. 2021. Classifying precipitation from GEO satellite observations: Diagnostic model. Q J R Meteorol Soc, 147(739): 3318–3334. doi: 10.1002/qj.4130

Vance T C, Huang T, Butler K A. 2024. Big data in Earth science: Emerging practice and promise. Science, 383(6688): eadh9607. doi: 10.1126/science.adh9607

van Straaten C, Whan K, Coumou D, et al. 2022. Using explainable machine learning forecasts to discover subseasonal drivers of high smmer temperatures in Western and Central Europe. Mon Weather Rev, 150: 1115–1134. doi: 10.1175/MWR-D-21-0201.1

Vaswani A, Shazeer N, Parmar N, et al. 2017. Attention Is All You Need. arXiv. doi:10.48550/arXiv.1706.03762

Vilone G, Longo L. 2020. Explainable Artificial Intelligence: a Systematic Review. arXiv. doi:10.48550/arXiv.2006.00093

Wadoux A M J C, Minasny B, McBratney A B, 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Sci Rev, 210: 103359. doi: 10.1016/j.earscirev.2020.103359

Wadoux A M J C, Molnar C. 2022. Beyond prediction: methods for interpreting complex models of soil variation. Geoderma, 422: 115953. doi: 10.1016/j.geoderma.2022.115953

Wang H, Li T, Wang G, et al. 2024. Significant spatiotemporal changes in atmospheric particulate mercury pollution in China: Insights from meta-analysis and machine-learning. SSci Total Environ, 955: 177184. doi: 10.1016/j.scitotenv.2024.177184

Wang K, Bastos A, Ciais P, et al. 2022. Regional and seasonal partitioning of water and temperature controls on global land carbon uptake variability. Nat Commun, 13(1): 3469. doi: 10.1038/s41467-022-31175-w

Wang S, Liu Y, Wang W, et al. 2024. Interpretable machine learning guided by physical mechanisms reveals drivers of runoff under dynamic land use changes. J Environ Manage, 367: 121978. doi: 10.1016/j.jenvman.2024.121978

Wang S, Peng H, Hu Q, et al. 2022. Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method. J Hydrol Reg Stud, 42: 101139. doi: 10.1016/j.ejrh.2022.101139

Wang Y, Shi L, Hu Y, et al. 2023. A comprehensive study of deep learning for soil moisture prediction. Hydrol Earth Syst Sci, 2023: 1-38. doi: 10.5194/hess-28-917-2024

Weber P, Carl K V, Hinz O. 2023. Applications of explainable artificial intelligence in finance—A systematic review of finance, information systems, and computer science literature. Manag Rev Q, 1-41. doi: 10.1007/s11301-023-00320-0

Wu J, Wang Z, Dong J, et al. 2023. Robust runoff prediction with explainable artificial intelligence and meteorological variables from deep learning ensemble model. Water Resour Res, 59: e2023WR035676. doi: 10.1029/2023WR035676

Wu Y, Li N. 2023. Nonlinear control of climate, hydrology, and topography on streamflow response through the use of interpretable machine learning across the contiguous United States. J Water Clim Change, 14: 4084–4098. doi: 10.2166/wcc.2023.279

Xie Q, Zhou M, Zhao Q, et al. 2019. Multispectral and hyperspectral image fusion by MS/HS fusion net. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos: IEEE. 1585–1594. doi:10.1109/CVPR.2019.00168

Xu Y, Lin K, Hu C, et al. 2024. Uncovering the dynamic drivers of floods through interpretable deep learning. Earths Future, 12: e2024EF004751. doi: 10.1029/2024EF004751

Yan J, Xu T, Yu Y, et al. 2021. Rainfall forecast model based on the TabNet model. Water, 13: 1272. doi: 10.3390/w13091272

Yang J. 2022. Fast treeshap: Accelerating shap value computation for trees. arXiv. doi:10.48550/arXiv.2109.09847

Yang W, Yang H, Yang D. 2020. Classifying floods by quantifying driver contributions in the Eastern Monsoon Region of China. J Hydrol, 585: 124767. doi: 10.1016/j.jhydrol.2020.124767

Ye S, Chai Y, Li J, et al. 2025. Explainable transfer learning for subsurface soil moisture prediction. J Hydrol, 661: 133473. doi: 10.1016/j.jhydrol.2025.133473

Yeom S K, Seegerer P, Lapuschkin S, et al. 2021. Pruning by explaining: A novel criterion for deep neural network pruning. Pattern Recognit, 115: 107899. doi: 10.1016/j.patcog.2021.107899

Ying Z, Bourgeois D, You J, et al. 2019. GNNExplainer: Generating explanations for graph neural networks. In: Wallach H, Larochelle H, Beygelzimer A, et al., eds. Advances in Neural Information Processing Systems 32. New York: Curran Associates, Inc. 9240 – 9251. doi:10.5555/3454287.3455073

Zacharias J, Von Zahn M, Chen J,et al. 2022. Designing a feature selection method based on explainable artificial intelligence. Electron Mark, 32(4): 2159–2184. doi: 10.1007/s12525-022-00608-1

Zafar MR, Khan NM. 2019. DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. arXiv. doi:10.48550/arXiv.1906.10263

Zeiler M D, Fergus R. 2014. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, et al., eds. Computer Vision – ECCV 2014. Cham: Springer. 818–833. doi:10.1007/978-3-319-10590-1_53

Zhang H, Wang G, Li S, et al. 2025. Understanding evapotranspiration driving mechanisms in China with explainable machine learning algorithms. Int J Climatol, 45: e8774. doi: 10.1002/joc.8774

Zhang J, Liu P, Zhang F, et al. 2018. CloudNet: Ground-based cloud classification with deep convolutional neural network. Geophys Res Lett, 45: 8665–8672. doi: 10.1029/2018GL077787

Zhong X, Gallagher B, Liu S, et al. 2022. Explainable machine learning in materials science. npj Comput Mater, 8(1): 204. doi: 10.1038/s41524-022-00884-7

Zumwald M, Baumberger C, Bresch D N, et al. 2021. Assessing the representational accuracy of data-driven models: The case of the effect of urban green infrastructure on temperature. Environ Modell Softw, 141: 105048. doi: 10.1016/j.envsoft.2021.105048