导读：
人工智能正重塑地球系统科学（ESS），但“黑箱”特性制约其深度应用。文章系统综述可解释人工智能在 ESS 中的方法与应用，揭示其与领域需求间的结构性脱节，提出融合领域知识、因果推断与人机协同的协同设计框架。展望未来，发展深度耦合 ESS 机理的高可信可解释人工智能体系，推动科学发现与可信预测。

加强融合：可解释人工智能推动地球系统科学发展

黄菲妮[1,2,3]，蒋世杰[2,3]，上官微[1*]，李璐[1]，张叶[1]，张如清[4]，李清亮[5]，李丹曦[1]，戴永久[1]

1 南方海洋科学与工程广东省实验室（珠海），广东省气候变化与自然灾害研究重点实验室，中山大学大气科学学院，珠海 519082
2 Department of Biogeochemical Integration, MaxPlanck Institute for Biogeochemistry, Jena 07745, Germany
3 ELLIS Unit Jena, Jena 07745, Germany
4 汕尾市气象局，汕尾 516600
5 长春师范大学计算机科学与技术学院，长春 130032

*通讯作者，E-mail：shgwei@mail.sysu.edu.cn

摘要

人工智能（AI）在地球系统科学（ESS）展现出变革性潜力，但其“黑箱”特性带来的可解释性不足，仍是制约其广泛应用的关键瓶颈。可解释人工智能通过增强模型透明度、揭示决策逻辑并促进人机协同，为破解这一难题提供了有效路径。本文面向 ESS 工作者，以清晰、实用的方式系统介绍可解释人工智能的核心理念与方法体系，澄清常见误区，并梳理其在 ESS 典型场景中的应用实践。本文重点阐述了可解释人工智能如何赋能模型解释、算法优化与科学发现，并深入剖析当前面临的三大挑战和相应解决思路：方法本身的局限性、可解释人工智能与 ESS 协调性不足，以及验证体系欠缺。为弥合可解释人工智能与 ESS 之间的认知与实践鸿沟，本文倡导构建“领域知识—因果推理—人本设计”三位一体的协同框架。展望未来，发展更可靠、高效且深度嵌入模型的可解释人工智能工具，将显著提升科学研究的可信赖性与洞察力，从利用 AI 持续拓展 ESS 的边界。

关键词 可解释人工智能，地球系统科学，可解释性，机器学习

1 引言

地球系统科学（ESS，Earth System Science）作为一门高度交叉的研究领域，致力于从耦合动态系统的角度，揭示大气圈、水圈、生物圈、岩石圈与人类活动之间的相互作用机制（Schellnhuber，1999；Steffen 等，2020）。该系统不仅涉及复杂的非线性反馈和多尺度过程，还涵盖从微观生态相互作用到全球气候格局的过程，还常常涌现出难以通过简单机理模型解释的整体行为特征（Heimann 和 Reichstein，2008）。此外，随着卫星遥感、原位观测与数值模拟等技术的广泛使用，ESS 数据在规模持续增长的同时，也呈现出高度的多源异构性，这进一步增加了从中提取可靠科学知识并支撑有效决策的难度（Mahecha 等，2020；Montero 等，2024；Rojas 等，2024；Vance 等，2024）。历史上，上述挑战制约了对若干关

键过程的深入理解，如云 – 气溶胶相互作用（Li T 等，2025）、永久冻土碳反馈（Song C 等，2024）及区域极端水文事件（McMillan 等，2025）。

人工智能（AI，Artificial Intelligence），尤其是机器学习与深度学习（ML/DL，Machine Learning/Deep Learning）技术，正以前所未有的方式推动 ESS 向数据驱动的新范式转型（Reichstein 等，2019）。这类方法擅长从高维、混杂噪声的复杂数据集中识别潜在过程，相较于传统基于物理的模型，它们展现出了重要的价值。典型应用包括：在短临降水预报中显著优于传统数值方法的 AI 天气模型（Sun T 等，2022）；基于多源数据融合的高分辨率土壤属性制图（Wadoux 等，2020；Khose 和 Mailapalli，2024）；以及自动化云型分类系统（Zhang J 等，2018；Segal-Rozenhaimer 等，2020）。然而，必须指出，从天气预警到长期气候政策制定等 ESS 的 AI 应用场景往往具有高度的社会敏感性与决策重要性。因此，将 ML/DL 模型简单视为“黑箱”使用存在重大风险。在此背景下，模型的可解释性成为确保其预测结果具备稳健性、可追责性，并能真正转化为可行动科学见解的关键前提（McGovern 等，2023，2024；Eyring 等，2024；Robert Maier 等，2024；Gilbert 和 Zengler，2025）。

可解释人工智能（XAI，eXplainable Artificial Intelligence）已成为提升复杂机器学习与深度学习模型可解释性的关键研究方向。典型方法如类激活映射（CAM，Class Activation Mapping）、SHapley 加性解释（SHAP，SHapley Additive exPlanations）与局部可解释模型无关解释（LIME，Local Interpretable Model-agnostic Explanations）等，能够识别关键特征、推断潜在因果驱动因素及变量间关系，从而揭示模型预测的内在逻辑。在 ESS 中，XAI 不仅有助于提高模型透明度，还可用于验证物理一致性，即确保人工智能的推断结果符合已知的 ESS 规律（Lyu 和 Yong，2025；Mallik 等，2025）。此外，XAI 所提供的机制性见解，亦具有辅助发现新科学现象的潜力 (Ham 等, 2023；Li W 等, 2022；Peng Z 等, 2024；Novielli 等，2025）。

然而，当前 ESS 领域与 XAI 研究之间仍存在显著的认知与实践鸿沟。尽管 XAI 在 ESS 中的应用迅速兴起，但其整体发展尚处初级阶段。现有研究往往机械套用 XAI 方法，未能充分结合 ESS 特有的科学问题与数据结构特点。与此同时，许多 ESS 学者缺乏系统评估 XAI 方法的技术基础，而 XAI 开发者则常忽略 ESS 数据中时空依赖性、非平稳分布等复杂特性。这种双向理解不足，在实际应用中引入多种不确定性及风险，主要包括：（1）过度解读事后解释，误将特征重要性与真实物理因果混淆（Aas 等，2021；Hooker 等，2021；Krell 等，2025），或忽视空间自相关等因素可能导致的解释假象（Chen C 等，2024；KeE 等，2025）；（2）所生成解释与科学机制脱节，难以有效支撑理论构建与发现（Cho 和 Ackom，2025）；（3）数据分布偏移时模型稳定性不足，致使 XAI 解释可靠性下降；（4）缺乏对领域先验知识及物理约束的融合。若不能有效弥合此鸿沟，XAI 在 ESS 中的长远发展将受限于未经验证的假设与可信度不足的输出结果。

本综述旨在系统梳理当前 XAI 方法在 ESS 中的应用现状，为 ESS 研究社群提供清晰的方法论参考。在重点总结 XAI 推动 ESS 发展已取得的关键进展的同时，本文也深入剖析了该领域当前面临的核心挑战与未来机遇。为便于不同背景的读者高效获取信息，我们在网络版附表 S1 中根据不同研究需求与专业基础，提供了相应的章节阅读建议。

2 XAI 理论基础

2.1 定义、解释方法与形式

可解释性强调使受众能够理解人工智能决策背后的基本原理（Biran 和 Cotton，2017；Gunning 等，2019，2021；Murdoch 等，2019；Minh 等，2022）。因此，解释需满足不同背景受众的需求（Barredo Arrieta 等，2020）。在此基础上，我们将 XAI 进一步定义为：“为不同受众提供模型输出之外、深入理解人工智能模型内部逻辑的方法体系”。本综述涉及的

XAI 方法及其分类详见网络版在线附图 S1、附表 S2 及附录 A。图 1 展示了阐释机器学习与深度学习黑箱模型的几种典型策略（详见第 2.2 节），图 2 则以土壤湿度建模为例，系统总结了 XAI 的主要解释形式：特征重要性用于评估变量贡献度；像素重要性可视化关键空间区域；序列重要性识别关键时间阶段；特征效应分析揭示预测结果对输入变量的响应规律；规则解释将复杂决策转化为逻辑陈述；模型表征技术则深入探究机器学习与深度学习模型的内部工作机制。

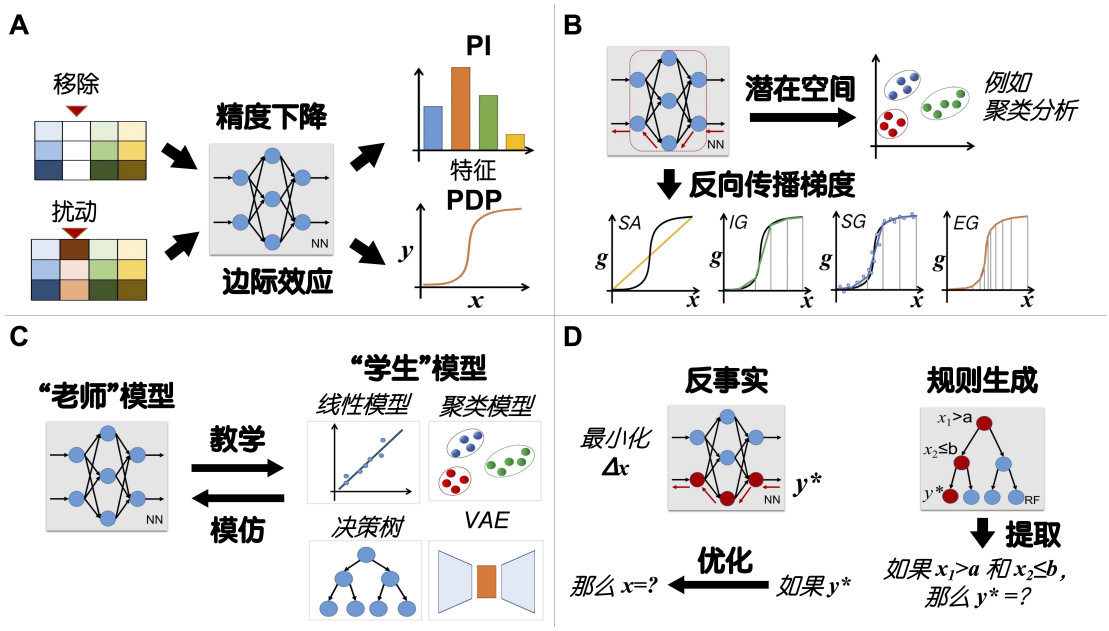

图 1 阐释黑箱 AI 模型的主要策略。（A）基于扰动的方法通过对输入进行干预并观测输出变化来解释模型，侧重于分析输入对输出的影响；（B）模型表征方法通常适用于特定模型，直接从模型结构或数据中提取相关信息（如视觉注意力、单元与层级梯度、潜在表征等），这些信息具有重要价值但解读难度较高；（C）替代模型方法通过构建一个本身可解释的模型来近似模拟复杂机器学习与深度学习模型的行为；（D）对比示例方法通常通过比较提供反事实解释和基于规则的说明。图中缩写：NN：神经网络；RF：随机森林；PI：置换重要性；PDP：偏依赖图；SA：显著图；IG：积分梯度；SG：平滑梯度；EG：期望梯度；VAE：变分自编码器。

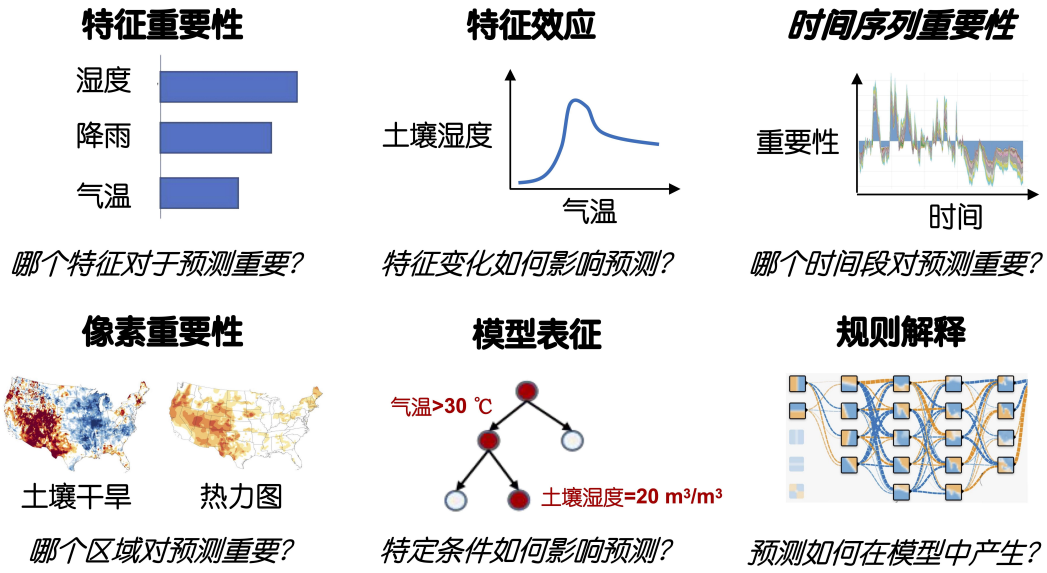

图 2 XAI 的解释形式及其解决的问题。本示例以土壤湿度预测为例，展示了不同 XAI 方法在解释输入变量气温对预测结果影响的具体应用方式。

2.2 黑箱 AI 模型的解释方法

2.2.1 基于扰动的方法

此类方法通过系统性地修改输入特征，并观察模型预测的相应变化（同时保持其他特征不变）来解释模型行为。

（1）特征重要性度量

这类方法旨在量化各输入特征对模型预测的贡献程度。其中，置换重要性（Permutation Importance，图 1A；Fisher 等，2021）通过随机扰动某一特征的值，并度量因此导致的模型预测性能下降来评估其特征重要性。基于树的特征重要性（Tree-based Feature Importance；Breiman，2001）则在树集成模型中，通过汇总各节点在分裂时因使用该特征所减少的不纯度总和来计算其贡献。SHAP（Lundberg 和 Lee，2017）基于合作博弈论，为每个特征分配一个反映其边际贡献的重要性值；该方法通过遍历特征子集，计算某一特征加入前后模型输出的平均差异来实现。近年来，SHAP 已衍生出多种针对复杂模型结构的优化变体（Covert 和 Lee，2021；Yang J，2022；Chen H 等，2019）。

（2）特征效应曲线

这类可视化工具用于刻画特征与模型预测之间的函数依赖关系。偏依赖图（Partial Dependence Plot， PDP；图 1A；Friedman 和 Popescu，2001）展示某一特征对预测结果的平均边际效应。个体条件期望图 (Individual Conditional Expectation, ICE; Goldstein 等, 2015) 则进一步描绘单个样本的预测随该特征变化的轨迹。当特征间存在较强相关性时，累积局部效应图（Accumulated Local Effects，ALE；Apley 和 Zhu，2020）通过计算条件期望来校正相关性引起的偏倚，其结果通常更为稳健。

2.2.2 替代模型方法

此类方法通过可解释的近似模型来模拟原始复杂模型的行为，从简单模型中提供解释，从而提升模型透明性 (图 1C)。其中一种主流方法是局部可解释模型无关解释 (LIME, Ribeiro 等，2016）。该方法基于局部线性的假设，对于某个特定预测，ML/DL 模型中复杂的局部行为可以通过一个可解释的模型来近似表示。该方法通过一组人工生成的样本来进行局部训练，这些样本通过在原始输入的局部邻域内进行扰动生成，从而得到复杂模型的局部解释

（Guo W 等，2018；Zafar 和 Khan，2019；ElShawi 等，2019）。

2.2.3 模型表征方法

模型表征方法旨在提取并解释 ML/DL 模型的内部结构和信息。

（1）基于树的模型解释方法

Treeinterpreter 等工具（Saabas，2014），可分解树集成模型（如随机森林、XGBoost）内的决策路径，将预测贡献归因于每个实例在决策过程中涉及的特征。

（2）基于梯度的归因方法

这类方法通常用于层级结构的神经网络，利用反向传播梯度估计特征重要性（图 1B）。基础方法如梯度显著图（Simonyan 等，2014）直接对梯度进行可视化。引导反向传播（Springenberg 等，2014）通过修正梯度传播路径以增强可视化效果。Input × Gradient (Shrikumar 等, 2019) 将输入值与梯度相乘, 以放大重要信号的贡献。积分梯度 (Sundararajan 等，2017）通过沿输入路径对梯度积分，从设定基线逐步累加贡献，其扩展方法期望梯度（Erion 等，2021）则通过对数据分布调整基线进行积分，计算梯度的期望值。平滑梯度（Smilkov 等，2017）通过对加入噪声的输入多次计算梯度并取平均，以增强稳定性。遮挡敏感性分析（Zeiler 和 Fergus，2014）则通过遮挡部分输入并观察预测变化，评估该区域对输出的影响。层间相关性传播（LRP，Layer-wise Relevance Propagation；Bach 等，2015）基于特定传播规则，将输出层的相关性逐层反向分配至输入，该方法不依赖于梯度或平滑性假设。上述方法的进一步说明参见附录 B。

（3）注意力机制与基于激活神经元的可视化方法

注意力机制可以动态地为不同输入元素分配权重，从而挖掘其相对重要性或关联性（Vaswani 等，2017）。诸如 CAM 和梯度加权类激活映射（Grad-CAM）（Selvaraju 等，2017）及其变体（Patro 等，2019；Bany 和 Yeasin，2021；Ibrahim 和 Shafiq，2022）等方法，利用卷积特征图生成热力图，突出显示对模型决策具有区分性的图像区域。

（4）隐含层信息的简化表达

该类方法旨在为生成对抗网络（GAN，Bau 等，2019；ChenX 等，2016）、扩散模型（Kwon 等，2022；Lee S 等，2023）以及 Transformer 模型（Castangia 等，2023；Playout 等，2022；Schwenke 等，2023）等复杂模型的高维内部状态构建易于理解的近似表示。所采用的技术包括典型相关分析 (Burges, 2010; Raghu 等, 2017)、聚类分析 (Raghu 和 Schmidt, 2020）、差异分析（Motteler 等，1995；Aires 等，2004；Rahwan 等，2019），以及线性近似（Lees 等，2022）或符号近似（Liu J 等，2023）等方法。

2.2.4 反事实和典型案例解释方法

此类方法通过展示可能改变模型决策结果的条件来解释模型的预测行为。反事实解释（Chou Y 等，2022）旨在识别能够改变预测结果的最小输入变化（图 1D）。原型与批评样本方法（Gurumoorthy 等，2019）则用于揭示具有代表性的训练样本。目前这两类方法在 ESS 领域的应用仍显不足，主要原因可能包括 ESS 数据特有的时空结构与复杂性，以及缺乏针对该领域的系统化解释框架。

上述 XAI 方法的详细说明汇总于网络版在线附表 S2 中。

2.3 XAI 评估方法

为确保 XAI 在 ESS 中的应用具备严谨性与可验证性，建立可比较、可量化的评估指标至关重要。据 Nauta 等（2023）统计，尽管有 58%的 XAI 技术评估采用定量指标，仍有 33%依赖领域真实证据，22%采用人工评估策略。在 ESS 中，基准数据集与基于物理过程的非线性函数常被用作 XAI 客观验证的“真实基准”（Arras 等，2022；Mamalakis 等，2022a，b）。本文重点关注以下五项核心定量评估指标 (Melis 和 Jaakkola，2018；Murdoch 等，2019；

Minh 等，2022；Weber 等，2023）：

(1) 忠实性：解释反映模型实际决策机制的程度（Alvarez-Melis 和 Jaakkola，2018）。

(2) 鲁棒性：在未改变模型预测结果的轻微输入扰动下，解释保持一致性的能力（Huang X 等，2020）。

(3) 稳定性：对相似样本生成解释结果的可重现性（Alvarez-Melis 和 Jaakkola，2018）。

(4) 高效性：生成解释所需的计算成本（Adadi 和 Berrada，2018；Vilone 和 Longo，2021）。

(5) 定位性：解释与预定义的重要区域的一致性程度（Arras 等，2022）。

以全球气温预测为例（Bommer 等，2024；图 3），在多种基于梯度的 XAI 方法中，LRP 取得了最高的综合定量评估得分，其归因图与领域知识相符，最为清晰地突出了北大西洋区域的影响。网络版附表 S3 总结了跨领域的 XAI 评估案例研究，网络版附表 S4 则列出了 XAI 评估工具包。综合来看，SHAP 与 LRP 方法普遍表现出较强且可靠的性能。然而，没有一种方法在所有场景下都是最优的，评估效果仍高度依赖于具体应用背景。

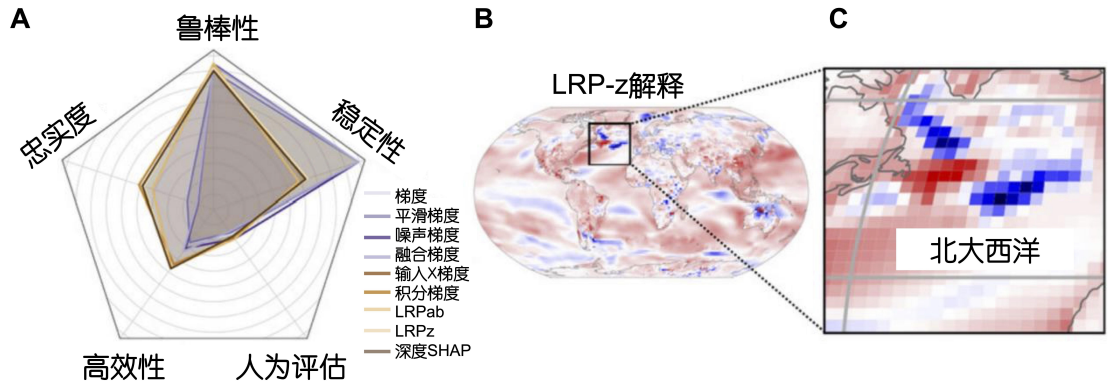

图 3 气候科学中 XAI 方法评估的案例研究。（A）九种 XAI 方法的平均技能得分。注：为与本文术语保持一致，图中部分指标名称已做调整："复杂度"（原始名称）调整为"高效性"；"随机化"（原始名称）调整为"稳定性"；"定位性"（原始名称）调整为"人为评估"。该图用于直观识别性能最优的方法，图中结果离中心越远表示性能越好。（B）性能最佳的方法（LRP-z）生成的十年期气温场解释图，（C）北美区域的局部放大视图。资料来源：改编自 Bommer 等人（2024）Finding the Right XAI Method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science，预印版 https://arxiv.org/abs/2303.00652v2，许可证为 CC BY 4.0。

3 XAI：从理论走向 ESS 应用

3.1XAI 在 ESS 领域的研究现状

本研究通过 Web of Science 数据库，以 XAI 与 ESS 相关关键词进行检索（附图 S2），共获得 863 篇文献（文献筛选流程详见附录 C，具体标准见网络版附表 S5），并基于此对研究现状进行分析。

分析结果表明，该领域研究自 2020 年起呈现指数级增长态势，至 2024 年发文量已超过 300 篇（图 4B）。近十年来，XAI 相关论文在 ESS 领域机器学习相关研究中的占比已超过 20%（Dramsch 等，2025）。应用研究主要集中在观测数据相对丰富的方向，如环境科学、水资源、大气科学和地球科学等（图 4C）。其中，水文气候建模已成为 XAI 应用的关键焦点领域（图 4A），XAI 常被用作进行气候归因研究和提升预测模型可解释性的核心工具。

我们进一步分析了 XAI 方法与各类机器学习/深度学习模型的搭配使用模式，以识别常见的技术组合（图 4D）。结果显示，基于扰动的方法（如 SHAP、LIME、置换重要性、偏依赖图）常与基于树的模型（如随机森林、XGBoost）结合使用。对于卷积神经网络与长短期记忆网络等神经网络模型，模型表征技术（如梯度归因、类激活映射）则占据主导地位。

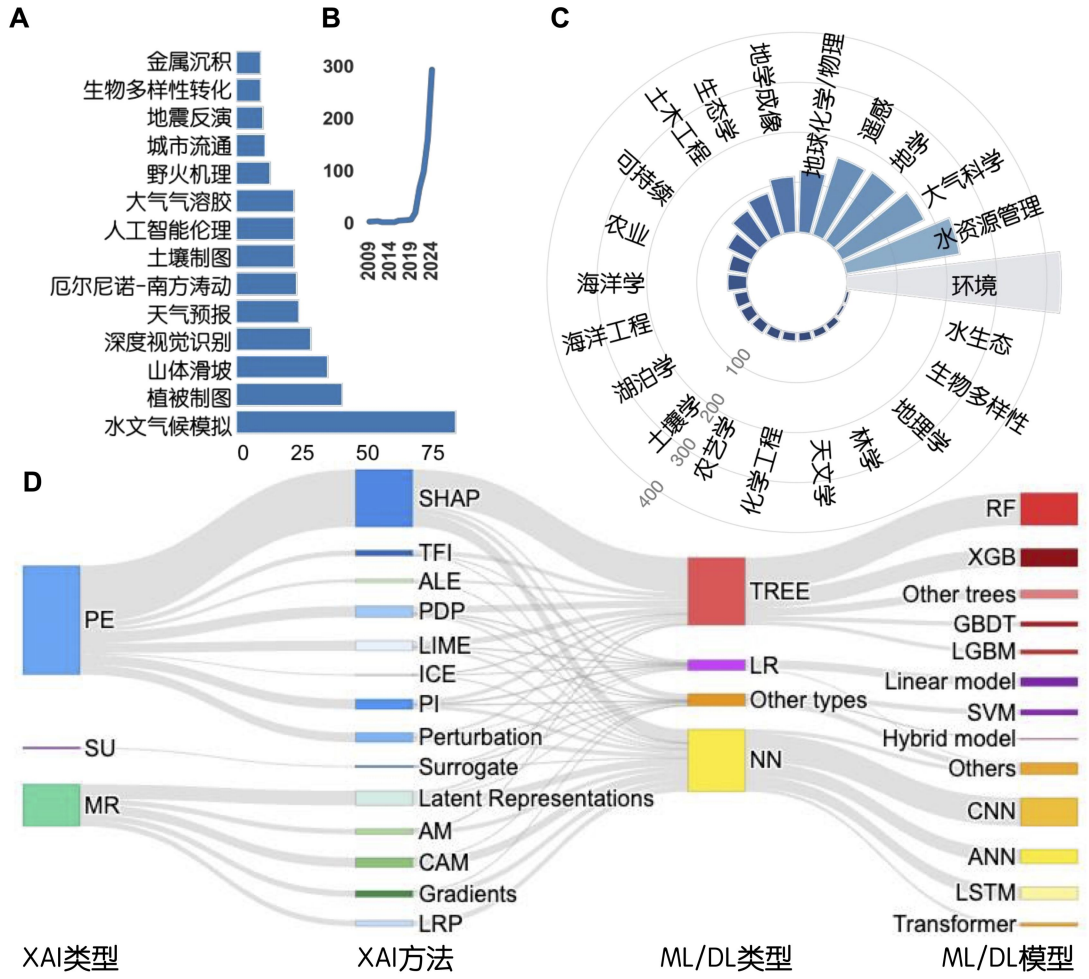

图 4 地球系统科学中 XAI 相关研究现状。（A）各相关主题的文献发表数量；（B）近年来的年度发表趋势；（C）基于 Web of Science 学科分类的 ESS 子领域分布；（D）所选文献中 XAI 类型与方法、ML 类型与模型的方法学结合应用情况。该分析基于 Web of Science 数据库中来自顶尖期刊（JCR 一区，前 25%）的 258 篇开放获取文章，且所有文章的引用量均排名在前 80%。图中缩写如下：PE：基于扰动的方法；SU：替代模型方法；MR：模型表征方法；SHAP：SHapley 加性解释；TFI：基于树的特征重要性；ALE：累积局部效应；PDP：偏依赖图；LIME：局部可解释模型无关解释；ICE：个体条件期望；PI：置换重要性；Perturbation：扰动变量；Surrogate：替代模型；Latent Representations：隐含层信息表达；AM：注意力机制；CAM：类激活映射；Gradients：梯度显示；LRP：层间相关性传播；TREE：基于树的模型；LR：逻辑回归/线性回归；Other types：其他类型；NN：神经网络；RF：随机森林；XGB：极限梯度提升；Other trees：其他基于树的模型；GBDT：梯度提升决策树；LGBM：轻量级梯度提升机；Linear model：线性模型；SVM：支持向量机；Hybrid model：混合模型；Others：其他模型；CNN：卷积神经网络；ANN：人工神经网络；LSTM：长短期记忆网络。

3.2 提升 AI 模型的可解释性

在 ESS 中, 建立透明、可问责、可信赖的人工智能系统对于支撑风险决策至关重要 (Haupt 等，2021；Debnath 等，2023；Camps-Valls 等，2025a）。这一需求得到了联合国教科文组织《人工智能伦理倡议书》（2021）以及欧盟《人工智能法案》（2021）等监管框架的强化，这些法规明确要求人工智能系统需具备可问责性。为确保解释具备实际指导意义，其本身应满足有稳健、可泛化、满足科学规律的要求。

在短期天气与极端事件预报方面, XAI 阐明了 AI 模型针对干旱 (Feng P 等, 2020; Dikshit 等，2022；Huang F 等，2023a）、洪水（Yang W 等，2020；Ekmekcioğlu 等，2021）、滑坡（Al-Najjar 等，2022）、土壤湿度干旱（Ye S 等，2025）、径流（Althoff 等，2021；Chen M 等，2023）、冰雹（Gagne II 等，2019）及地震预警（Fayaz 和 Galasso，2024）等极端事件的预测机制。在长期气候科学研究中，XAI 也有助于 AI 生成的未来情景预测的可信度评价（McGovern 等，2023，2024；Diffenbaugh 和 Barnes，2023；Evans 等，2025），例如海表温度（van Straaten 等，2022）和气候振荡（Schmidt 等，2020；Gordon 等，2021）。在 AI 生成的地球数据产品领域，XAI 发挥着关键的验证作用。它能够确保关键变量（如土壤碳储量（Wadoux 和 Molnar，2022）、作物类型（Orynbaikyzy 等，2020）和遥感地物分类（Hasanpour Zaryabi 等，2022））的生成产品符合物理规律（Shangguan 等，2017，2023；Dueben 等，2022；Gevaert，2022），从而增强科研与决策对使用这些产品的信心。

3.3 提升 AI 建模效率

ML/DL 开发者常借助 XAI 技术进行特征筛选、模型选择和结构设计，以提升模型构建效率。

在特征选择方面，XAI 可用于从预训练模型中提取特征重要性信息，帮助剔除无关或冗余特征。例如，基于扰动的方法，例如 TFI（Feng P 等，2019；Upadhyaya 等，2021）、PI（Ramirez 等，2022）以及 AM（Yan J 等，2021）等技术，已在 ESS 相关变量预测中取得优于传统筛选方法的效果，（Zacharias 等，2022；Wang J 等，2023），同时更好地保持了预测结果的物理一致性（Carter 等，2021）。

为确定最优的 ML/DL 模型，开发人员借助 XAI 提供候选模型的行为解释，并结合领域先验知识进行比较分析（Jing 等，2023；Wu Y 等，2023）。该策略已在洪水预报中得到有效应用（Schmidt 等，2020），并有助于验证复杂深度学习模型相对于过程模拟的合理性，从而获得更具物理现实性的见解（Hu X 等，2021）。

许多 XAI 方法专门针对模型结构设计的诊断与优化，尤其是模型表征方法（Toms 等，2020）。模型表征技术能够发现决策树（Chen J 等，2021）和 cubist 模型（FuZ 等，2022）等树节点结构中存在的问题，例如 Treeinterpreter 工具已应用于降水分类任务的性能提升（Upadhyaya 等，2021）。在深度学习模型中，雅可比矩阵等模型表征技术有助于减少因移除神经网络中不重要的连接而带来的偏差，这已应用于大气廓线反演（Blackwell，2012；Maddy 等，2021）和降水预测（Shamekh 等，2023）。尽管 XAI 能有效支持模型优化并促进 ESS 建模中的物理一致性，其目前仍无法直接指导具体的模型结构设计或超参数调优，该过程仍需依赖领域经验与系统的实验验证。

3.4 从 AI 模型中获取科学见解

ESS 研究人员寄望于将 XAI 作为一种数据挖掘工具，从高性能机器学习与深度学习模型中提取物理机制层面的新见解（Jiang S 等，2024a）。

3.4.1 解析水文机制

XAI 被用于识别控制土壤湿度（Ley 等，2024）、蒸散发（Chakraborty 等，2021b；Zhang

H 等，2025）、径流（Althoff 等，2021；Bai Z 等，2022；Hao H 等，2024；Wang H 等，2024）、河道流量（Liu J 等，2023；Wu 和 Li，2023）及地下水（Mo Y 等，2025）等动态过程的关键驱动因子，有助于解决传统水文模型中复杂的非线性耦合与空间异质性难题。此外，XAI 技术还能识别关键过程阈值（Wang S 等，2022；Ding K 等，2025；Wang H 等，2024）以及变量响应的空间异质性（Wang S 等，2024）。通过这种方式，XAI 架起了数据驱动预测与机理理解之间的桥梁，阐明水文过程的控制机制。

3.4.2 极端事件归因

XAI 通过解释 ML/DL 模型内部的复杂过程，提升了对干旱（Feng P 等，2019；Saha 等，2021；Dikshit 等，2024）、洪水（Jiang S 等，2022a，b；Xu Y 等，2024；Feng J 等，2024；Liu M 等，2024；Ke E 等，2025）、野火（Kondylatos 等，2022；Abdollahi 和 Pradhan，2023；Bountzouklis 等，2023）、极端降水（Gimeno-Sotelo 等，2023）和热浪（Shu R 等，2025）等极端事件的机理认识。例如，在洪水分析中，XAI 已被用于将极端事件归因于降水、气温等环境先兆条件的突变（Slater 等，2024）。交互式 SHAP 值进一步揭示了关键的复合效应（Jiang S 等，2024b），而引入远程预测因子则促进了对遥相关驱动机制的识别（Lee Y 等，2024）。

3.4.3 检测人类活动足迹

XAI 的归因能力随过程时间尺度的不同而有显著差异。它能够有效识别点源污染（Liu S 等，2025）和城市热岛效应（Zumwald 等，2021；Oukawa 等，2022）等快速响应系统中的人类活动驱动因子。然而，对于气候尺度现象（如温室气体强迫），由于信号微弱且人类影响以非线性方式累积，XAI 的效能则有所下降（Alam 等，2025）。

通常，XAI 基于公认的气候代用指标（如全球平均温度）进行归因，并可进一步揭示一系列下游效应（Janssens 等，2021），如异常的生物地球化学循环（Haaf 等，2021；Patoine 等，2022；Wang K 等，2022）、碳汇与潜热吸收（Berner 等，2020）以及海气耦合（Sonnewald 和 Lguensat，2021）。类似“干旱区趋于更干，湿润区趋于更湿”等气候尺度的现象，已通过 XAI 得到了验证。

3.4.4 局限性与未来发展方向

尽管 XAI 能够从数据中揭示重要影响的变量或非线性关系，但其本身无法推导出基本物理定律或解析形式的方程。XAI 的解释能力仍受限于底层数据质量及模型结构，它可以反映变量之间的相关性和相互作用，却无法独立验证因果关系。这一局限性为 ESS 研究者和气候政策制定者带来了关键挑战，例如虚假相关以及对未参与模型训练的数据的泛化能力不足。因此，亟需加强 ML/DL 开发者与 ESS 领域专家之间的紧密合作，以确保基于 XAI 的发现植根于领域物理机制，而非“算法伪影”。总体而言，尽管 XAI 是应对复杂 ESS 问题的有力工具，但只有与传统实验研究和理论方法相结合，才能构建对系统层面过程的深入理解。

3.5 ESS 中 XAI 的实践指南

我们提出了一个将解释机制嵌入迭代建模过程的集成框架（图 5A）。该框架始于研究目标的明确，随后依次开展常规的数据处理、特征选择、模型训练与验证。在此基础上，通过定量评价指标或结合领域先验知识，对候选 XAI 方法进行评估与筛选，以优化解释结果。

在这一通用工作流程中，不同研究目标侧重的环节有所不同：

（1）以提升可解释性为目的时，重点在于输出可靠且可复现的解释结果；

（2）以提升 AI 建模效率为目的时，XAI 作为诊断工具用于优化特征选择、数据利用和模型结构设计；

（3）以获取科学见解为目的时，则应先引入物理先验约束，从而获取超越相关性的因果认识。

我们以 Huang 等（2023b）的研究为例，展示了该框架在土壤湿度预测任务中的应用。在完成特征选择（图 5B）与模型充分训练（图 5C）的基础上，联合多种 XAI 分析揭示了土壤温度的主导控制作用（图 5D），该结果与土壤热力原理基本一致。

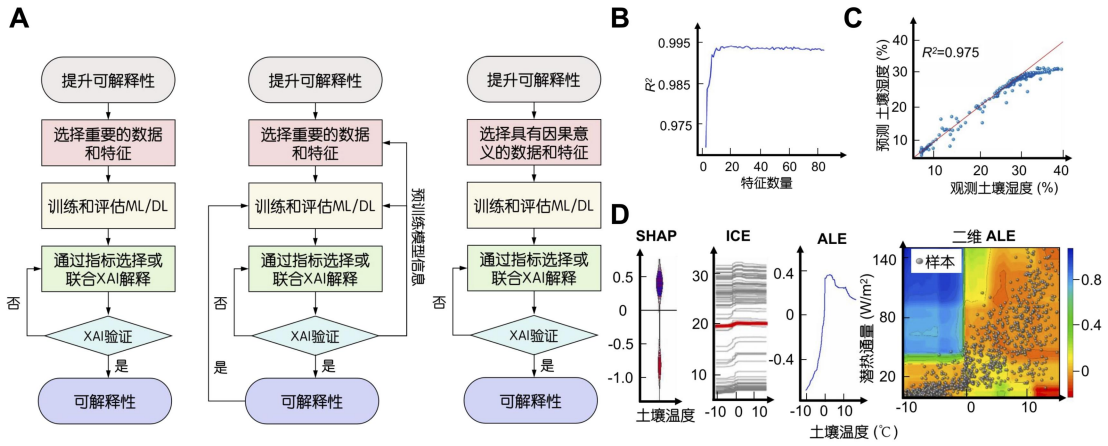

图 5 XAI 在 ESS 中的应用实践框架及土壤湿度预测案例研究。（A）迭代式 XAI 解释与验证流程示意图；（B）–（D）案例研究对应各阶段输出结果：（B）特征选择阶段筛选出的关键变量，（C）模型训练与验证阶段的性能评估曲线，（D）基于多方法联合 XAI 分析得到的土壤温度主导作用解释，其中包括 SHAP 贡献图、个体条件期望（ICE）曲线、累积局部效应（ALE）图，以及土壤温度与潜热通量的二维 ALE 交互分析。（B）–（D）根据 Huang 等（2023b）研究改编绘制。

4 挑战与可能的解决方案

4.1XAI 固有的局限性

4.1.1XAI 方法设计层面的局限性

XAI 方法在设计中存在固有的局限性，直接影响了解释结果的可靠性，目前尚不存在适用于所有场景的通用最优方法。

基于替代模型的方法（如 LIME）通常依赖于局部线性与特征独立性等强假设，而这些条件在实际 ESS 问题中往往难以满足。当变量间存在相关性时，扰动单一变量可能引起其他变量的同步变化，这一现象已在土壤湿度预测任务中得到验证 (图 5D; Huang 等, 2023b)。研究表明，在小样本建模情况下，LIME 对相关特征的敏感性高于 SHAP (Huang 等, 2023a)；而在大样本条件下，该问题可在一定程度上被忽略（Krell 等，2025）。此外，此类方法通常假设扰动噪声服从高斯分布（Ivanovs 等，2021），与真实数据分布常有偏差，且在高维特征空间中面临显著的计算负担。

基于梯度的方法所产生的解释往往存在噪声大、不稳定的问题（Saleem 等，2022），尤其在大规模神经网络中更为明显，甚至可能引入额外的随机性（Huang 等，2024）。模型表征类方法（如降维表示隐含层信息）虽然能够提取模型内部结构信息，但其结果的可理解性通常有所下降（Mylonas 等，2024）。总体而言，基于扰动和替代模型的方法在计算成本上普遍高于模型相关的解释方法。以单次预测解释生成时间为例，LRP 通常需 1 – 2 秒，而 LIME 与 SHAP 分别需约 22 秒和 108 秒。

4.1.2 难以通过评估指标选择最优 XAI 方法

当采用多种 XAI 技术解释同一模型时，常会得到差异显著的结论，这种现象被称为“罗生门效应”。其根源在于不同方法在信息处理与表征形式上存在本质差异（Başağaoğlu 等，

2022；Huang 等，2023c）。尽管已有多种评估策略（见第 2.3 节），但它们通常孤立使用、依赖额外计算资源，且结论常具有较强的情境依赖性，限制了其普适性与推广性。

4.1.3 缺乏统一的 XAI 评估基准

缺乏广泛认可的评估基准进一步增加了 XAI 方法选择的难度。现有研究多依赖领域先验知识作为“真实基准”进行轶事验证。例如，Mamalakis 等（2022b）引入 ClimateNet 数据集，其中包含专家标注的“大气河流”，用于检验 XAI 是否捕捉到具有物理意义的模式。此外，来自多个地球系统模型及其他过程模型的衍生信息也被用作评估参考，借助其中蕴含的集体物理知识支持 XAI 验证（Mamalakis 等，2022a）。

4.1.4 可能的解决方案

为应对上述挑战，建议从以下方向探索解决路径：

（1）发展集成化解释框架

鼓励同时采用多种 XAI 技术对同一模型或样本进行联合分析，通过综合或加权不同方法提供的互补视角，更全面地刻画模型行为（Gibson 等，2021；Ghada 等，2022；Van Straaten 等，2022；Ham 等，2023）。例如，Huang 等（2023a）通过联合使用 PI、SHAP、LIME、ALE、PDP 和 ICE 对土壤湿度干旱格局进行分析，显著提升了解释的可靠性和对模型的忠实度。

（2）构建集成式 XAI 方法

推动将事前可解释与事后可解释方法结合，融合扰动型、替代模型型与模型表征型技术，或联结全局与局部解释，形成对模型行为的系统性理解。

（3）探索可解释的深度学习架构

可解释的深度学习模型（如扩散模型、图神经网络等）为应对模型复杂性提供了新的可能。这类模型不仅能提供局部解释，还可支持全局层面的行为剖析（O’Loughlin 等，2025）。通过在结构中显式编码变量间关系，有助于将先验物理知识融入解释过程，支持可视化、对比分析与系统集成。

4.2XAI 与 ESS 的协调性挑战

4.2.1XAI 设计目标与 ESS 领域需求之间的不适配

尽管第 2.1 节系统梳理了 XAI 的多种解释形式，但这些方法在很大程度上仍未充分满足 ESS 研究的核心需求。这种不足主要源于 XAI 方法的设计重点与 ESS 实际应用场景之间的结构性脱节（Gevaert，2022）。ESS 相关利益相关方，包括天气预报业务人员、政策制定者与监管机构、ML/DL 开发者、AI 衍生产品使用者以及地学科学家，他们所期望的解释已超越传统的简化特征归因，而更关注复杂地球系统过程的动态演化与状态依赖性。

尽管天气预报员已尝试将特征归因方法的解释引入预报系统，但现有解释形式在促进其对模型行为的深入理解方面仍显不足。与此同时，政策制定者和监管机构更倾向于获取全局性解释与反事实分析，以支持未来情景下的决策制定。然而，当前 XAI 在指导 AI 模型诊断（尤其是模型结构设计与改进）方面仍具有明显局限。

4.2.2 静态解释难以满足动态的、相互影响的系统要求

XAI 的假设（如特征独立性、模型平稳性和局部线性）与 ESS 的本质（表现为强烈的特征相关性、系统动态性和全局非线性）之间存在根本性的概念鸿沟，这制约了其在 ESS 中的进一步应用。ESS 的非平稳性强调系统的时变性及不可预测的行为，而大多数 XAI 方法依赖于静态模型假设，因此在面临外推预测（尤其在不熟悉或超出训练数据分布的情形下）时，往往难以提供可靠解释。

时空依赖性可能削弱许多 XAI 方法的鲁棒性。基于扰动的方法（如 PI 和 LIME）通过独立扰动输入并监测输出变化来评估特征重要性，在这种情境下，常常会产生不符合物理规

律的样本。在模型归因时，XAI 方法会将重要性分散到不同的时空节点上，而非捕捉连续的响应或串联过程（如洪水或滑坡）。

此外，大多数现成的 XAI 方法聚焦于非结构化数据，而 ESS 数据则是高度结构化的，其特征包括混杂效应、强烈的变量间相关性、异质性分布（例如泊松分布的降水、正弦变化的太阳短波辐射）。这种不匹配会导致从 XAI 中获得的关于过程的理解存在偏差（见附图 S3），因为变量间的相关性已被发现会引发解释的不确定性（Jiang S 等，2022a，b）。研究人员采用各种预处理策略（如方差膨胀因子（He F 等，2025）、皮尔逊系数 Díaz-Vallejo 等，2024）、Z-Score（Cui X 等，2021）等）试图缓解数据结构化带来的影响，然而此类方法虽可移除模型内数据的相关性，基础的数据结构依然存在，扰动一个变量仍可能引发其他变量的非线性变化。

4.2.3 因果关系与 XAI

地学科学家旨在从观测驱动的 ML/DL 模型中提取因果关系，以补充现有知识体系（Irrgang 等，2021）。近期研究尝试利用 XAI 在预测模型中以具有因果暗示的方式识别“驱动因子”。然而，这类实验常常将相关性与因果关系混为一谈。即使是高精度的 ML/DL 模型，在没有经过严格因果验证的情况下，也未必满足因果推断的要求。这可能涉及潜在混杂变量带来的虚假相关性。模型是基于因果关系还是相关性构建，从根本上决定了其推断因果机制的能力（Camps-Valls，2025b）。

当前文献提出了两条解决路径：面向因果的 XAI 与基于因果的 XAI (Carloni 等, 2025)。前者使用 XAI 生成科学假设，随后通过实验方法进行因果关系验证；后者则从结构因果模型等因果模型中推导解释。这两种方法在 ESS 中都有应用前景。在实践中，许多利用 XAI 识别驱动因子的现有研究，实际上隐性地遵循了“面向因果的 XAI”路径。

一些研究者已将先验知识整合到 ESS 的混合建模框架中，以在解释过程中减轻混杂因素带来的偏差（Althoff 等，2021；Li W 等，2024；Hu X 等，2021）。然而，缺乏稳健的验证仍然是一个问题。例如，自由大气 CO_2 富集等实验方法可能提供验证途径。相反，从因果模型中推导解释则需要仔细界定研究范围，以更好地控制混杂因素，获得有针对性的因果解释。

4.2.4 端到端建模与基于过程的建模

一个根本性挑战在于，XAI 通常解释的是端到端的黑箱模型，这些模型捕获了复杂的映射关系，但其中间过程和变量相互作用仍难以捉摸。相比之下，ESS 中的基于过程的模型则明确表征了机理关系和中间状态。因此，XAI 推导出的解释通常以输入-输出归因的形式表达，无论从形式还是认知相关性上，都难以与地学研究中诊断分析和假设检验所需的、面向过程的理解相匹配（Li X 和 Guo Y，2025）。地学科学家常常忽视了 XAI 以及某些数据驱动的方程发现技术（如符号回归）的功能。问题的核心在于，ML/DL 和 XAI 在设计时并未充分考虑面向过程的 ESS 需求。

4.2.5 可能的解决方案

（1）面向对象的 XAI 任务

XAI 方法应显式引入领域相关的“对象”概念（如气旋、生态系统、流域等），而非仅停留在对潜在因果特征的归因层面。面向对象的解释能够更好地刻画地球系统的组织结构与动态演化过程，从而显著提升解释结果在科学分析中的可用性和可靠性。

（2）融入过程知识的建模框架

将物理过程知识显式引入建模过程，有助于使预测结果更接近结构化 ESS 数据中所蕴含的“真实机理”，从而使 XAI 能够揭示具有因果意义且易于科学理解的关键信息 (Chen M 等, 2023)。一种具有前景的策略是将复杂的输入–输出映射分解为已知与未知的子过程 (Shen

C 等，2023），已知过程由基于过程的模型加以刻画，而未知过程则由可解释的 ML/DL 模型进行描述。

在相对简化的子系统中开展分析，有助于降低复杂系统中混杂因素的干扰，使子过程层面的因果关系更易于验证，这对 ESS 建模人员整合结构化观测数据尤为有利（Li X 等，2023）。在此基础上，我们提出了一种因果物理约束的混合 XAI 框架（见附录 D；附图 S4）。该框架强调以过程表征为核心，而非仅依赖数据分布：利用物理模型刻画已知机制，同时借助 ML/DL 模型逼近尚未解析的过程，从而实现对复杂地球系统的因果解释。

（3）以用户为中心的 XAI 设计

鉴于 ESS 应用场景与用户群体的高度多样性，采用以用户为中心的 XAI 设计理念至关重要。最初为医学领域 XAI 提出的设计准则（Chen H 等，2022）可在适当调整后推广至 ESS。近期研究指出，在气候行动中使用存在偏差的 ML/DL 模型可能引发问责风险（Debnath 等，2023）。因此，在 XAI–ESS 的整个开发与应用过程中，引入以人为中心的原则，有助于提升系统的可用性、透明性与责任可追溯性（Clement 等，2023）。一个结构化的实施流程（图 6）可包括以下关键步骤：

（1）通过问卷调查或评分卡分析用户需求（Hoffman 等，2023）；

（2）与 ESS 利益相关方协同设计，将领域知识系统性地融入模型与解释中（例如，为天气预报人员构建具有物理可解释性的模型（Kim 等，2023），为政策制定者提供可追责的决策支持系统）；

（3）构建标准化工作流程，以降低终端用户采用 XAI 的技术门槛；

（4）通过经验验证对 XAI 进行测试与迭代，不断优化其在 ESS 中的应用效果。

在这一过程中，不同学科间对解释目标的协调至关重要。一个切实可行的方向是开发开源软件库，实现基于过程模型与 XAI 技术的无缝链接，从而促进跨学科合作与方法推广。

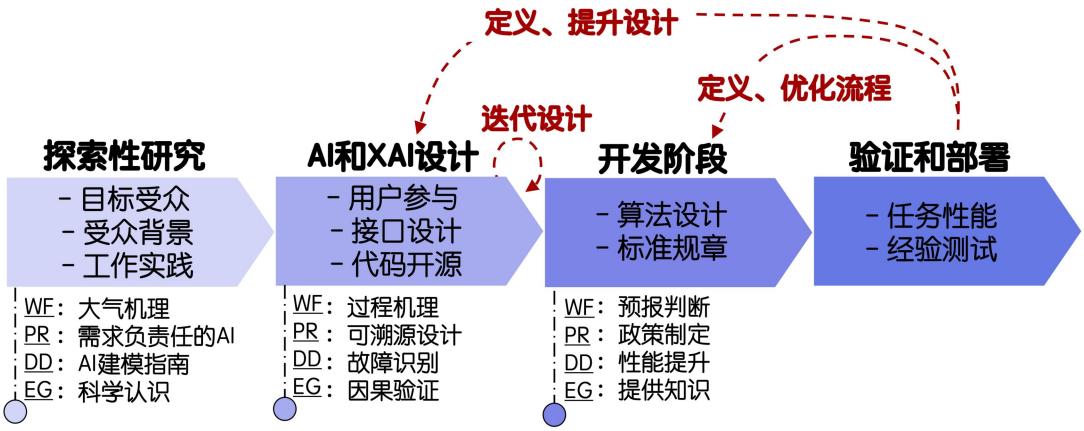

图 6 面向 ESS 的以人为中心 XAI 设计的操作性解决方案。其中，WF、PR、DD 和 EG 分别代表天气预报业务人员、政策制定者与监管机构、模型诊断与调试，以及地球过程模型研究人员与地学科学家。

4.3 不确定性评估

在天气预报和气候政策等高风险应用领域，对 XAI 不确定性的量化，对于支撑可靠决策来说至关重要。不稳定或不可靠的解释结果可能导致严重的社会影响，或引发误导性的科学结论。XAI 方法自身的不确定性主要源于其方法设计的机制和假设，这些因素会削弱解释

结果的稳健性（Thuy 和 Benoit，2024）。因此，系统性地识别和评估此类不确定性，是确保基于 XAI 的认知具有科学可信度和可行性的关键前提。

一种切实可行的途径是通过重复实施 XAI 方法来量化解释结果的变异性，来评价不确定性，即在相同条件下多次生成解释，并通过收敛性分析评估其稳定性与一致性，从而为解释结果的不确定性提供定量约束。

5 未来展望

5.1 可信、高效与集成的 XAI

展望未来，我们期待构建一种更具可信度且计算效率更高的集成化 XAI 框架，将图 1 所示的多种解释技术有机融合（Belaid 等，2023）。该框架旨在通过整合不同 XAI 方法的互补优势，克服单一方法在假设、适用范围和稳定性方面的固有限制。例如，作为一种集成式方法，“glocalXAI”能够同时提供通道层级信息与特征重要性评估，在揭示模型表征方式、推理过程以及细粒度决策机制方面具有显著潜力（Achtibat 等，2023）。类似地，融合全局和局部视角的注意力机制（Li L 等，2018）以及结合空间通道信息的注意力方法（Song C 等，2022）已在计算机视觉领域展现出良好的应用前景。

未来的 XAI 技术应突破单一方法的局限，实现多种解释策略的无缝整合。我们认为，用于特定模型的 XAI 方法更适用于结构复杂的先进深度学习模型；而对于相对简单的模型，集成的 XAI 方法应提供更高的灵活性，以支持在不同模型架构之间开展可解释性比较（Theissler 等，2022）。

随着基础模型的快速发展，对轻量级 XAI 方案的需求愈发迫切。诸如“盘古”（Bi K 等，2023）、“伏羲”（Chen L 等，2023）和 GraphCast（Lam 等，2023）等前沿模型已在 ESS 领域树立了新的性能标杆。然而，传统 XAI 方法（如 SHAP）所需的高昂计算成本，严重限制了其在大规模模型中的应用。未来的解决方案应更多依赖模型相关的 XAI 方法设计或有针对性的输入扰动策略，或者通过选取具有代表性的样本高效探测模型行为，从而避免模型蒸馏带来的额外偏差。

此外，有必要整合忠实性、鲁棒性、稳定性等关键指标，构建统一的 XAI 评价标准。目前，评价不同 XAI 方法仍然是十分挑战的（Mi J 等，2024）。更重要的是，现有评价指标往往依赖于诸如局部一致性或扰动稳定性等假设，而在 ESS 的时空遥相关、非线性阈值过程及临界转变等现象出现时，这些假设通常不成立。例如，稳健的 XAI 方法通常要求在微小输入扰动下给出一致解释，但大气系统中普遍存在的“蝴蝶效应”可能使这一假设不再成立，初始条件的微小变化即可引发难以预测的系统响应，而这类效应往往难以仅凭 XAI 手段加以识别。因此，未来工作的关键起点之一，是研发并评估专门面向 ESS 领域特性的 XAI 方法，以更好地适配地球系统的复杂动力学特征。

5.2 拓展 XAI 在 ESS 中的作用

本文认为，XAI 在 ESS 中的作用仍有必要进一步拓展。我们从 ESS 发展的三个阶段（认知、描述和行动）对这一问题进行系统阐述。如图 7 所示，我们以灰色标注 XAI 已经实现的功能，以红色标注其潜在的未来角色。该框架将 ESS 概念化为三个层次递进的阶段：（1）认知阶段：通过对真实世界数据与模拟结果的分析，提炼有助于决策的信息，从而提升对 ESS 动力学行为的认知理解；（2）描述阶段：系统刻画过程特征与数据模式，以揭示系统的控制机制；（3）行动阶段：将认知与描述层面的知识转化为针对生态系统保护或具体干预措施。

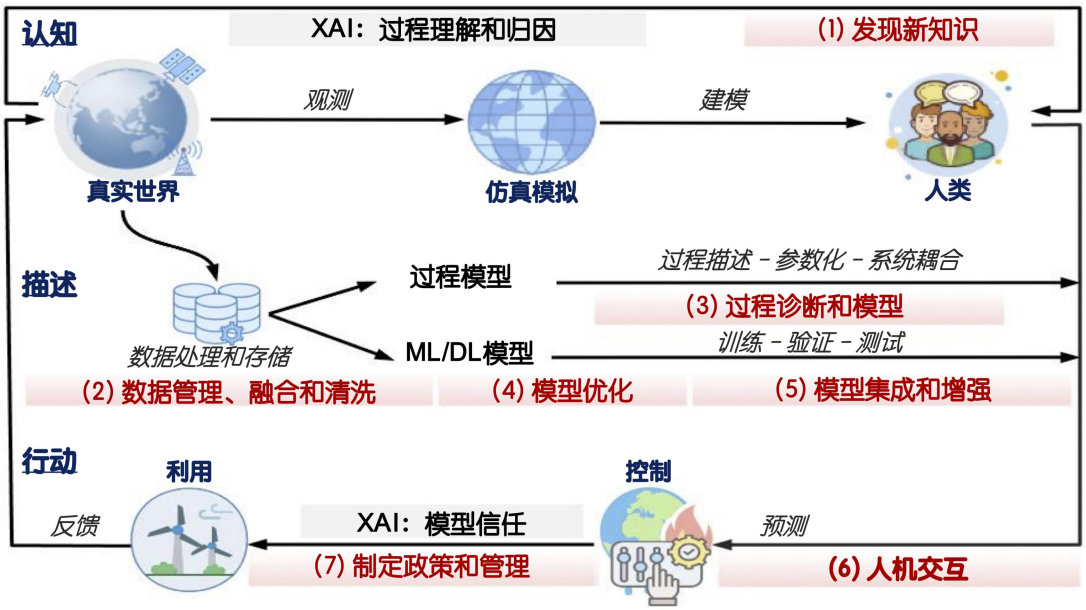

图 7 地球系统科学的分层集成框架：从基础的过程认知到可行的实际应用。图中灰色方框表示 XAI 已实现的功能，红色方框表示 XAI 潜在的拓展方向。

（1）认知阶段：发掘新知识

XAI 能够从复杂、低阶的数据中提取高阶的关系与过程信息，而这些信息往往难以通过人工分析直接识别。因此，XAI 在跨学科新知识发现方面展现出巨大潜力。例如，在药物研发领域，XAI 在揭示未知表征形式和效应方面已发挥了关键作用（McCloskey 等，2019；Ishida 等，2019）。其中，GNNExplainer 作为一种与图卷积网络相结合的 XAI 方法，通过优化图结构要素之间的互信息，为模型解释提供了有力工具，其所生成的化学上可解释的表征与化学家的直觉认知高度一致（Ying Z 等，2019）。这一成功经验为 ESS 利用 XAI 探索 ESM 中尚未充分关注的子过程提供了重要启示。

在部分 ESS 过程难以参数化的情形下，基于 XAI 的材料动力学建模已取得令人鼓舞的成果（Zhong X 等，2022）。例如，通过联合使用特征重要性分析、概念性解释以及概念激活向量方法，研究人员成功利用视频数据对可扩展钙钛矿太阳能电池中的过程动力学进行了精确刻画（Klein 等，2024）。此外，本身可解释的 ML/DL 模型也已被证明能够有效表征多类化学过程（Gallegos 等，2024）。展望未来，XAI 有望成为 ESS 中探索新知识和新表征的重要替代路径，其优势在于能够揭示传统建模和分析方法难以发现的潜在过程模式与复杂相互作用。

（2）描述阶段：基于 XAI 的数据管理、融合与清洗

XAI 在 ESS 中的应用可扩展至更广泛的范围，使其贯穿数据建模的全生命周期。在数据管理方面，一种被称为“边际智能（Edge Intelligence）”的新理念逐渐兴起，其核心在于通过生成边际级别的数据来重构数据表征并减少数据体量（Sinha 和 Vashisht，2023）。引入 XAI 有助于确保这一过程在物理上保持一致性和可解释性，从而促进对 CMIP（Coupled Model Intercomparison Project）、多分辨率再分析数据以及遥感产品等大规模数据集的高效管理与存储。

在数据融合方面，XAI 已被成功用于多源遥感产品的融合分析（Lohit 等，2019；Xie Q 等，2019；Taskin 等，2024）。鉴于 XAI 在复杂数据中识别与提取相似模式方面的优势，其应用有望进一步扩展至融合更为多样化的数据源，包括遥感数据、再分析资料、原位观测

以及基于过程模型生成的数据。

此外，XAI 在数据清洗方面同样具有重要潜力。原型相关性传播方法通过结合“原型”概念与 LRP 技术，能够有效识别并分离伪影样本与高质量样本（Gautam 等，2023）。在医学影像分析中，XAI 方法已被证明能够在主动学习框架下筛选最具信息量的样本，在更少的数据量和迭代次数下获得与甚至优于其他样本选择策略的性能（Mahapatra 等，2021）。类似技术在 ESS 中同样具有重要价值，例如可用于减轻传感器噪声对数据质量的影响，如在 FLUXNET 数据预处理过程中。

（3）描述阶段：过程诊断与评估

地球系统模式（Earth System Model，ESM）在预测未来地球状态方面发挥着核心作用，其基础是对 ESS 关键过程的认知。然而，即便在最新的 CMIP6 计划中，不同 ESM 在多种过程表征上仍表现出显著不一致性（Hsu 和 Dirmeyer，2023；Fu W 等，2022）。基于数据驱动的物理模型重建（即模式模拟器或“仿真器”）可以分析由参数不确定性和外部强迫引起的系统性偏差，已成为不可或缺的手段（Lian X 等，2018）。

在这一背景下，XAI 可作为强有力的可视化与诊断工具，用于识别基于 AI 的重建模型中所蕴含的不同模式特征（Li X 等，2023）。例如，XAI 能够揭示不同 ESM 在物理表征方面的差异，如北冰洋酸化过程中的模型分歧（Krasting 等，2022）。

另一种复杂的路径可以更稳健地揭示 ESM 所编码的关键信息，通过训练 ML/DL 模型以年平均气温或降水分布图为输入来预测其对应的年份，可以一定程度上在归因时避免模式内部噪声的干扰（Barnes 等，2020；Labe 和 Barnes，2021）。该方法使研究者能够在不受内部变率显著影响的情况下，更清晰地识别驱动 ESM 行为的基础动力学机制。

（4）描述阶段：模型优化

类激活映射和典型相关分析等方法，可用于提取中间层表征的信息来调整模型内部结构，从而实现性能的精细优化（Schiller 等，2019；Anders 等，2022）。来自预训练模型的解释结果还可用于增强损失函数（Ross 等，2017）、放大或引导梯度信息（Narteni 等，2025），或者直接改进模型架构本身（Yeom 等，2021）。近年来，新型 XAI 方法也被用于超参数优化（Ibrahim 和 Shafiq，2022）以及加速模型学习过程（Mukhamediev 等，2022）。尽管这些进展不断涌现，最新 XAI 技术在 ESS 应用往往存在一定滞后。因此，ESS 研究人员有必要持续跟进 XAI 的快速发展，以确保 XAI–ESS 的应用始终保持前沿性并充分释放其潜力。

（5）描述阶段：模型集成与增强

部分前沿 XAI 方法通过数据增强、特征增强、损失增强、梯度增强以及模型结构增强等方式来提升 ML/DL 模型性能（Weber 等，2023），但这些策略尚未在在 ESS 领域得到广泛应用。一般来说，XAI 所生成解释是否具备物理一致性，这可以作为模型集成中分配模型权重的重要评价依据之一，从而提高集成模型在科学合理性与预测性能上的综合表现。

（6）行动阶段：人机交互

交互式 XAI 系统通过允许用户主动查询和探索模型决策机制，在促进人机交互方面发挥着关键作用。例如，面向天气预报模型设计的 XAI 交互界面，使用户能够评估模型整体可靠性、检验解释结果与领域知识的一致性，并向系统反馈改进建议。随着公众参与科学逐渐发展为重要的社会活动，民众可以通过智能手机等设备参与科学数据采集，ML/DL 模型用于处理海量观测数据，而 XAI 则帮助参与者理解其所贡献数据背后的科学发现，从而增强其学习参与度与信任感。

此外，将 XAI 与虚拟现实或增强现实技术相结合，有望构建强大的可视化工具，促进对 ESS 的整体性理解。在另一方面，交互式 XAI 系统还应提供开放接口，允许人在 ML/DL 建模过程中进行干预，以保障系统的公平性与公共问责性。现有多种功能性可视化 XAI 方法（如 LSTMVis；Strobelt 等，2017）已能够通过交互式图的形式展示相关变量及其模型内

部的部分隐藏状态。展望未来，交互式 XAI 有望成为连接人类认知与机器决策的重要接口，促进更直观、协同的人机沟通（Naiseh 等，2023）。

（7）行动阶段：政策制定与管理

鉴于 XAI 在气候变化预测与归因等方面已展现出卓越的预测能力（见第 3.2 节），有必要进一步探讨其结果能否为政策制定提供直接指导。首先，在天气、水资源和气候等领域，构建可信赖的 ML/DL 系统已成为紧迫需求（Reidmiller 等，2017）。在缺乏终端用户对模型可解释性进行实证评估的情况下，基于必要的伦理原则，ML/DL 开发者需要提供 XAI 方案，使用户能够充分了解模型的内部工作机制，从而维护其决策自主性（McGovern 等，2022）。

其次，当气候政策依赖于 ML/DL 时，识别并应对潜在偏差所带来的问责风险尤为关键（Debnath 等，2023）。我们认为，如果 XAI 所揭示的关键证据能够通过实验或其他独立证据加以验证，则其结果有望为政策制定者在气候变化减缓与适应方面的决策提供科学依据。此外，在水资源管理和碳管理等领域，基于 XAI 生成假设所开展的区域尺度干预和实验正变得日益可行。例如，Schoenke 等（2021）在农业领域构建了一个概念性平台，通过可视化数据流，为农民和监管机构提供可解释的洞见与可操作的建议，展示了 XAI 在资源管理策略制定和政策形成中的直接影响力。

6 总结与结论

尽管 XAI 已成为 ESS 研究中不可或缺的工具，但其潜力仍受制于黑箱模型在可解释性方面的不足。本文综述表明，尽管 XAI 有望释放这一潜力，然而当前通用的 XAI 方法设计与 ESS 领域特定需求之间存在显著不匹配，这严重阻碍了其实际应用。为使 XAI 从一项新兴技术真正转变为 ESS 研究的核心工具，亟需在技术发展方向和优先事项上进行系统性协调与优化。

基于本文的综合分析，我们提出以下三项可行框架，供学界与业界共同推进：

（1）开发面向 ESS 的 XAI 方法与基准

为克服 XAI 的固有方法学局限性，我们倡导将两种的策略相结合：（a）系统性地使用多种 XAI 技术，同时开发本身可解释的深度学习模型，以确保解释结果的稳健性；（b）加强学科间合作，制定 ESS 任务中评估 XAI 忠实性与鲁棒性的标准化基准。该类基准应在预报数值精度的基础上，优先考虑物理合理性，以保证科学解释的可靠性。

（2）以人为中心的多维可解释性应成为 XAI 优先发展方向

单纯的技术性解释不足以满足 ESS 的科学需求，解释结果必须转化为具有科学意义的认知洞见。我们的研究强调，合作设计框架的广泛采用至关重要，即 ESS 领域专家、ML/DL 开发者与终端用户开展合作，从模型开发之初便考虑嵌入解释机制。此外，我们认为，将物理信息驱动和因果推断技术与数据驱动的 XAI 相结合，是生成可信且物理合理的过程性认知的很有前景路径。关键的第一步是建立开源库，能够将基于过程的模型与先进 XAI 技术无缝结合。

（3）将不确定性量化纳入 XAI 标准流程

缺乏不确定性度量的解释，其科学价值有限。我们强调，不确定性量化应成为 XAI 在 ESS 中应用不可或缺的步骤。这不仅要求开发新的方法，将不确定性有效地传播至解释生成流程，还需要设计科学可视化策略，将不确定性直观传达给科研人员与政策制定者。

ESS 的未来发展与可信赖 AI 的进展紧密相关，推动下一代 XAI 的研究显得尤为迫切。我们总结两大关键前沿方向：一是开发轻量化、本身可解释的模型，能够无缝融合物理知识；二是开发稳健的诊断工具，用于模型提升与验证。通过优先推进上述方向，XAI 有望从被动的解释工具转变为主动的科学发现平台，推动关键领域的进展，其中包括从复杂数据集中提取知识、混合模型的开发与优化、整合 ESM 模型间差异、为政策制定设计可靠的决策支持

系统。这一转变将使 XAI 在 ESS 中不仅是理解模型的工具，更成为推动科学认知与实践创新的核心驱动力。

项目资助

国家自然科学基金项目（42375144、4227515 和 42205149）、教育部基础学科交叉突破计划项目、广东省基础与应用基础研究基金项目（2021B0301030007）、中国气象局青年创新团队项目（CMA2024QN01）和国家留学基金委项目（202306380183）资助。

补充材料

补充材料为作者提供的原始数据，作者对其学术质量和内容负责，详见网络版(http://earthcn.scichina.com)。

参考文献

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.

684 doi: 10.5194/hess-26-795-2022
685 Apley D W, Zhu J. 2020. Visualizing the effects of predictor variables in black box supervised
686 learning models. JJ R Stat Soc Series B Stat Methodol, 82(4): 1059 – 1086. doi:
687 10.1111/rssb.12377
688 Arras L, Osman A, Samek W. 2022. CLEVR-XAI: A benchmark dataset for the ground truth
689 evaluation of neural network explanations. Inf Fusion, 81: 14–40. doi:
690 10.1016/j.inffus.2021.11.008
691 Bach S, Binder A, Montavon G, et al. 2015. On pixel-wise explanations for non-linear classifier
692 decisions by layer-wise relevance propagation. PLoS ONE, 10: e0130140. doi:
693 10.1371/journal.pone.0130140
694 Bai Z, Liu Q, Liu Y. 2022. Groundwater potential mapping in Hubei region of China using
695 machine learning, ensemble learning, deep learning and AutoML methods. Nat Resour Res, 31:
696 2549–2569. doi: 10.1007/s11053-022-10100-4
697 Muhammad M B, Yeasin M. 2021. Eigen-CAM: Visual explanations for deep convolutional
698 neural networks. SN Comput Sci, 2: 47. doi: 10.1007/s42979-021-00449-3
699 Barnes E A, Toms B, Hurrell J W, et al. 2020. Indicator patterns of forced change learned by an
700 artificial neural network. J Adv Model Earth Syst, 12:e2020MS002195. doi:
701 10.1029/2020MS002195
702 Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. 2020. Explainable artificial intelligence
703 (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion,
704 58: 82–115. doi: 10.1016/j.inffus.2019.12.012
705 Bau D, Zhu J Y, Strobelt H, et al. 2019. Visualizing and understanding generative adversarial
706 networks (extended abstract). arXiv doi:10.48550/arXiv.1901.09887
707 Başağaoğlu H, Chakraborty D, Lago C D, et al. 2022. A review on interpretable and explainable
708 artificial intelligence in hydroclimatic applications. Water, 14(8): 1230. doi:
709 10.3390/w14081230
710 Belaid M K, Bornemann R, Rabus M, et al. 2023. Compare-xAI: Toward unifying functional
711 testing methods for post-hoc XAI algorithms into a multi-dimensional benchmark. In: Longo
712 L, ed, Explainable Artificial Intelligence: First World Conference, xAI 2023, Lisbon, Portugal,
713 July 26 – 28, 2023, Proceedings, Part II. Cham: Springer . 88–109
714 Berner L T, Massey R, Jantz P, et al. 2020. Summer warming explains widespread but not
715 uniform greening in the Arctic tundra biome. Nat Commun, 11: 4621. doi:
716 10.1038/s41467-020-18479-5
717 Bi K, Xie L, Zhang H, et al. 2023. Accurate medium-range global weather forecasting with 3D
718 neural networks. Nature, 619: 533–538. doi: 10.1038/s41586-023-06185-3
719 Biran O, and Cotton C. 2017. Explanation and justification in machine learning: A survey. In:
720 IJCAI-17 Workshop on Explainable AI (XAI), Melbourne: IJCAI. 8-13.
721 Blackwell W J. 2012. Neural network Jacobian analysis for high-resolution profiling of the
722 atmosphere. EURASIP J Adv Signal Process, 2012: 71. doi: 10.1186/1687-6180-2012-71
723 Bommer P L, Kretschmer M, Hedström A, et al. 2024. Finding the right XAI method—A guide
724 for the evaluation and ranking of explainable AI methods in climate science. Artif Intell Earth
725 Syst, 3: e230074. doi: 10.1175/AIES-D-23-0074.1
726 Bountzouklis C, Fox D M, Bernardino E D. 2023. Predicting wildfire ignition causes in Southern
727 France using explainable artificial intelligence (XAI) methods. Environ Res Lett, 18: 044038.

728 doi: 10.1088/1748-9326/acc8ee
729 Breiman L. 2001. Random forests. Mach Learn, 45: 5–32. doi: 10.1023/A:1010933404324
730 Burges C J C. 2010. Dimension reduction: A guided tour. Found Trends Mach Learn, 2: 275–365.
731 doi: 10.1561/2200000002.
732 Camps-Valls G, Fernández-Torres M Á, Cohrs K H, et al. 2025. Artificial intelligence for
733 modeling and understanding extreme weather and climate events. Nat Commun, 16: 1919. doi:
734 10.1038/s41467-025-56573-8
735 Camps-Valls G. 2025. Explaining what, exactly? A critical appraisal of why XAI fails science.
736 Medium.
737 https://medium.com/@gcampsvalls/explaining-what-exactly-a-critical-appraisal-of-why-xai-fa
738 ils-science-bc0e821c8531
739 Carloni G, Berti A, Colantonio S. 2025. The role of causality in explainable artificial intelligence.
740 WIREs Data Min Knowl Discov, 15(2). doi: 10.1002/widm.70015
741 Carter E, Herrera D A, Steinschneider S. 2021. Feature engineering for subseasonal-to-seasonal
742 warm-Season precipitation forecasts in the midwestern United States: Toward a unifying
743 hypothesis of anomalous warm-season hydroclimatic circulation. J Climate, 34: 8291-8318.
744 doi: 10.1175/JCLI-D-20-0264.1
745 Castangia M, Grajales L M, Aliberti A, et al. 2023. Transformer neural networks for interpretable
746 flood forecasting. Environ Modell Softw, 160: 105581. doi: 10.1016/j.envsoft.2022.105581
747 Chakraborty D, Başağaoğlu H, Gutierrez L, et al. 2021. Explainable AI reveals new
748 hydroclimatic insights for ecosystem-centric groundwater management. Environ Res Lett, 16:
749 114024. doi: 10.1088/1748-9326/ac2fde
750 Chakraborty D, Başağaoğlu H, Winterle J. 2021. Interpretable vs. noninterpretable machine
751 learning models for data-driven hydro-climatological process modeling. Expert Syst Appl, 170:
752 114498. doi: 10.1016/j.eswa.2020.114498
753 Chen C, Liu Y, Li Y, et al. 2024. Explainable artificial intelligence framework for urban global
754 digital elevation model correction based on the SHapley additive explanation-random forest
755 algorithm considering spatial heterogeneity and factor optimization. Int J Appl Earth Obs
756 Geoinf, 129: 103843. doi: 10.1016/j.jag.2024.103843
757 Chen H, Gomez C, Huang C M, et al. 2022. Explainable medical imaging AI needs
758 human-centered design: Guidelines and evidence from a systematic review. npj Digit Med, 5:
759 156. doi: 10.1038/s41746-022-00699-2
760 Chen H, Lundberg S, Lee S-I. 2019. Explaining models by propagating Shapley values of local
761 components. arXiv. doi: 10.48550/arXiv.1911.11888
762 Chen M, Qian Z, Boers N, et al. 2023. Iterative integration of deep learning in hybrid Earth
763 surface system modelling. Nat Rev Earth Environ, 4(8): 568–581. doi:
764 10.1038/s43017-023-00452-7
765 Chen J, Zhang H, Fan M, et al. 2021. Machine-learning-based prediction and key factor
766 identification of the organic carbon in riverine floodplain soils with intensive agricultural
767 practices. JJ Soils Sediments, 21: 2896–2907. doi: 10.1007/s11368-021-02987-y
768 Chen L, Zhong X, Zhang F, et al. 2023. FuXi: A cascade machine learning forecasting system for
769 15-day global weather forecast. npj Clim Atmos Sci, 6: 190. doi:
770 10.1038/s41612-023-00512-1
771 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