https://doi.org/10.1038/s41561-025-01639-x

# Explainability can foster trust in artificial intelligence in geoscience

Jesper Sören Dramsch, Monique M. Kuglitsch, Miguel-Ángel Fernández-Torres, Andrea Toreti, Rustem Arif Albayrak, Lorenzo Nava, Saman Ghaffarian, Ximeng Cheng, Jackie Ma, Wojciech Samek, Rudy Venguswamy, Anirudh Koul, Raghavan Muthuregunathan & Arthur Hrast Essenfelder



Uptake of explainable artificial intelligence (XAI) methods in geoscience is currently limited. We argue that such methods that reveal the decision processes of AI models can foster trust in their results and facilitate the broader adoption of AI.

Artificial intelligence (AI) offers unparalleled opportunities for analysing multidimensional data and solving complex and nonlinear problems in geoscience[1–3]. However, as the complexity and potentially the predictive skill of an AI model increases, its interpretability — the ability to understand the model and its predictions from a physical perspective — may decrease[3,4]. In critical situations, such as scenarios caused by natural hazards, the resulting lack of understanding of how a model works and consequent lack of trust in its results can become a barrier to its implementation[5]. Here we argue that explainable AI (XAI) methods, which enhance the human-comprehensible understanding and interpretation of opaque 'black-box' AI models, can build trust in AI model results and encourage greater adoption of AI methods in geoscience[6].

## Benefits of XAI

Trust is crucial to the adoption of AI[1,7]. Thus some researchers advocate for inherently interpretable AI models; in other words, models that provide their own explanations[7,8]. Others, however, prefer to retain the predictive capabilities of deep neural networks — models able to capture highly complex and nonlinear patterns in data but with limited interpretability — and to circumvent black-box issues through XAI methods, which provide "an explanation to the user that justifies its recommendation, decision, or action"[9]. These methods can provide insight into an AI system, identifying issues related to data or the model. For example, XAI can detect spurious correlations in training data and otherwise imperceptible perturbations to remote sensing images[10]. In this sense, XAI can be regarded as a magnifying lens, enabling the human expert to analyse data through the 'eyes' of the model so that the dominant prediction strategies — and any undesired behaviours — can be understood[11]. Another benefit of XAI is that it can highlight linkages between input variables and model predictions, which may motivate further research[3,12] and support an enhanced understanding of features as well as spatiotemporal processes. For example, researchers have used XAI on an inventory of landslide data to understand why AI models classify slopes as susceptible (or not) to failure and to gain insight into failure mechanisms[13]. XAI has also been applied to time series of a meteorological drought index to determine the importance of climatic variables such as precipitation for meteorological drought prediction[14]. In the latter example, the results aligned with physical model interpretations, emphasizing the need to include specific climatic variables as predictors in the model. Figure 1 demonstrates the possible benefits of XAI across different dimensions, using natural hazards as an example domain.

## Uptake of XAI in geoscience

Given these benefits, we were curious to see how the geoscience community is applying XAI. To acquire an overview, we extracted geoscience-related articles from a corpus of 2.3 million arXiv abstracts published between 2007 and 2022. We found that while references to AI and XAI increase with time, considerably fewer papers reference XAI (6.1%) than AI (25.5%), with the relative proportion generally remaining constant, and that those mentioning XAI are mostly in the fields of geoinformatics (including remote sensing) and geophysics (including seismology and volcanology) (Box 1).

To further explore the use of XAI, we focused on a specific area of geoscience, natural hazards, for which we had access to use cases curated by the International Telecommunication Union/World Meteorological Organization/UN Environment Focus Group on AI for Natural Disaster Management[2]. These use cases exemplify how AI can be used to detect, monitor, forecast, and communicate (for example, via hazard maps and early warning systems) various types of natural hazards. We surveyed the researchers of these use cases and found that motivations for applying XAI were consistent with the benefits detailed above: some use cases aimed to achieve trust, some hoped to acquire insight into data or AI issues or to make them more efficient, and most applied XAI to make discoveries about the underlying processes. Of those use cases that did not apply XAI, many acknowledged the value of XAI for lending interpretability to and ensuring the proper functioning of AI models, and revealing underlying processes. Furthermore, almost all respondents indicated plans to apply it at a later stage, but had not done so in part because of the effort, time, and resources that XAI requires.

## Challenges and solutions to increasing XAI adoption

Based on our literature review and researcher survey, we suggest that unless an AI end user (for example, those paying for an AI-based operational forecast) explicitly demands explainability, researchers may be tempted to forgo this step to avoid investing effort and already scarce time and resources. When AI is purely used for academic research, it mainly falls on the funding agencies and the scientific community to insist on this additional step.

Another challenge relates to the relevance, accuracy, and reliability of existing XAI methods for geoscience applications. Traditionally, most XAI methods have been applied to image data, which are

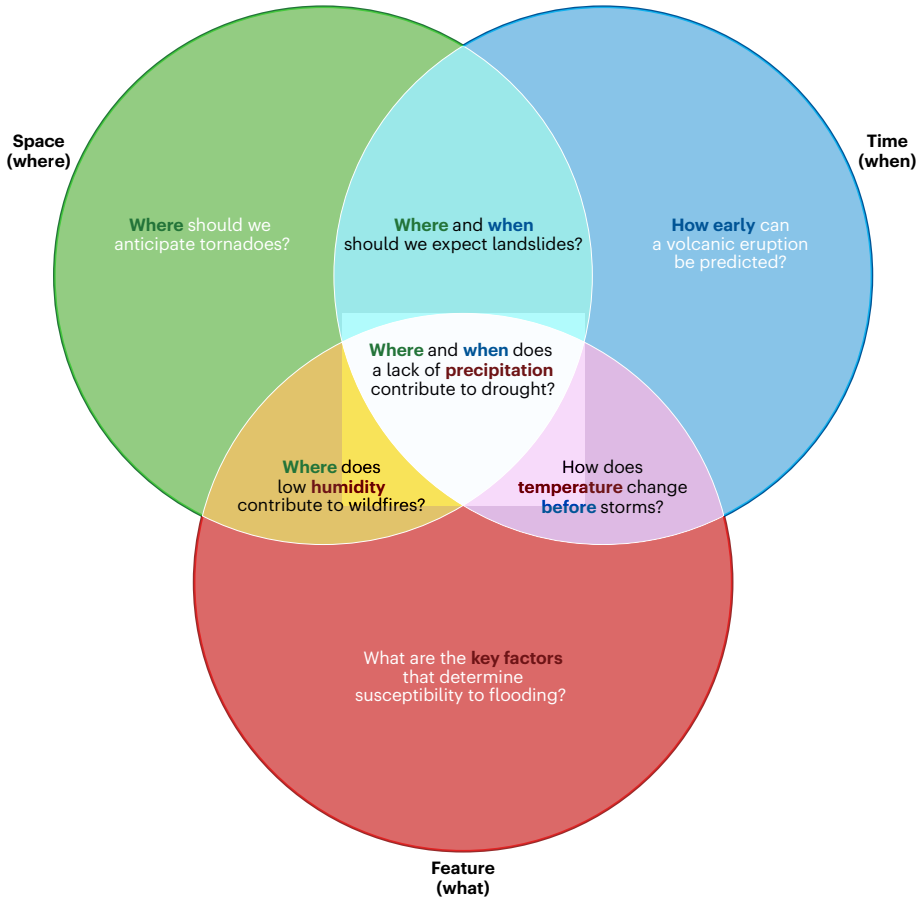



**Fig. 1 | The value of explainable artificial intelligence (XAI).** Possible benefits of XAI for natural hazard applications include gaining insights into input variables and model predictions over time, space, and feature type.

relatively simple. However, geospatial data have specific characteristics (for example, spatiotemporal dependence[2]) and XAI requirements (for example, object- and concept-level explanations[15]). Fortunately, the emergence of new techniques specifically tailored for temporal data creates new opportunities in, for example, seasonal and decadal climate forecasting models.

Traditional XAI is also often still not sufficiently understandable by non-specialists. For instance, knowing that a specific pixel in an image is relevant for the prediction does not provide any insight into the model's internal representation and inference process, and makes it very hard to interpret the model behaviour in terms of physical concepts and phenomena. Recently developed methods, such as concept relevance propagation, close this gap and provide more abstract, human-understandable explanations by combining perspectives of both the data (that is, what information is relevant) and model (that is, how it is represented and processed).

Overall, XAI methods have been shown to be suitable for addressing many geoscience inquiries, offering valuable insights into intricate models and data. To overcome challenges in uptake and facilitate the adoption of XAI, we make four recommendations.

**Demand.** If funding a project, reviewing a paper, or intending to deploy an AI system, stakeholders and end users should explore the applicability of interpretable or explainable models. Such approaches may also help research meet requirements relating to transparency.

**Resources.** XAI users should understand how different methods function, what they can provide for explanations, and where they have limitations, rather than applying them unquestioned. Once an XAI method is selected, code libraries can facilitate their application, but associated literature and metadata should be carefully reviewed, because some common libraries have shortcomings when applied to specific data or analyses. Additionally, researchers can quantitatively evaluate the performance of their black-box AI models on benchmark datasets.

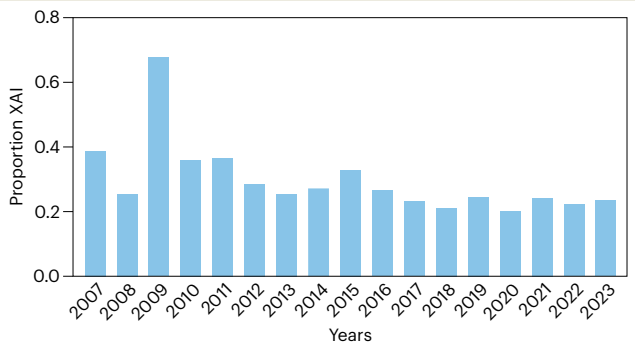
**Partnerships.** International efforts, such as the United Nations Global Initiative on Resilience to Natural Hazards through AI Solutions and the European Union-funded Climate Intelligence project, bring together geoscience and AI experts and encourage sharing of insights.

**Integration.** Streamlined workflows are crucial for the standardization and interoperability of AI in the domain of natural hazards and disasters, as well as many other branches of the geosciences. To achieve trust, such workflows must provide human-comprehensible understanding, which can be achieved by integrating XAI into them.

## BOX 1

# Geoscience papers referencing AI and XAI from 2007 to 2022

To compare the number of geoscience papers per year referencing AI and XAI, first, a search index based on the Annoy Index (https://github.com/spotify/annoy; a version of approximate nearest neighbours) coupled with Scientific Paper Embeddings using Citation-informed TransformERs was used to identify any articles referencing thirty common geoscience fields: atmospheric science, meteorology, climate science, palaeoclimatology, biogeochemistry, geobiology, geochemistry, geoinformatics, remote sensing, geology, geomagnetism, palaeomagnetism, geomorphology, glaciology, hydrology, limnology, mineralogy, mineral physics, oceanography, palaeoceanography, natural hazards, natural disasters, palaeontology, petrology, planetary science, geophysics, seismology, volcanology, space physics, and tectonics. Among these articles, 12,429 abstracts were sampled based on a reverse-keyword match per field and the full manuscripts were downloaded for further analysis. We investigated what proportion of these geoscience papers reference XAI, how this compares across geoscience disciplines, what XAI methods are most commonly referenced (and how this changes with time), and how the growth in AI through time compares with the growth in XAI. To do so, we applied the same tool to search these 12,429 manuscripts for expressions commonly associated with XAI: interpretability or explainability or explainable AI or XAI or AI model inspection or AI model interpretation or AI model visualization. Then, we clustered the geoscience fields by topic — atmospheric science or meteorology or climate science or palaeoclimatology; biogeochemistry or geobiology or geochemistry; geoinformatics or remote sensing; geology; geomagnetism or palaeomagnetism; geomorphology; glaciology; hydrology or limnology; mineralogy or mineral physics; oceanography or palaeoceanography; natural hazards or natural disasters; paleontology; petrology; planetary science; geophysics or seismology or volcanology; space physics; or tectonics — to identify those clusters of geoscience fields most commonly applying XAI. In the next step, we searched for common XAI methods[6]. Finally, we calculated the percent of articles per annum referencing artificial intelligence versus expressions commonly associated with XAI (as described earlier in this paragraph).




It is our hope that given the considerable opportunities presented by XAI – to improve underlying datasets and AI models, identify physical relationships that are captured by data, and build trust among end users, which can be lost to the detriment of progress – explainability will become part of the standard protocol in applying AI for geoscience.

**Jesper Sören Dramsch[1] ✉, Monique M. Kuglitsch[2], Miguel-Ángel Fernández-Torres[3], Andrea Toreti[4], Rustem Arif Albayrak[5,6], Lorenzo Nava[7,8,9], Saman Ghaffarian[10], Ximeng Cheng[2], Jackie Ma[2], Wojciech Samek[2,11,12], Rudy Venguswamy[13], Anirudh Koul[13], Raghavan Muthuregunathan[14] & Arthur Hrast Essenfelder[4]**

[1]European Centre for Medium-Range Weather Forecasts, Bonn, Germany. [2]Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany. [3]Image Processing Laboratory (IPL), Universitat de València (UV), Paterna, (València), Spain. [4]European Commission Joint Research Centre, Ispra, Italy. [5]NASA Goddard Space Flight Center, Greenbelt, MD, USA. [6]University of Maryland, Baltimore County, MD, USA. [7]Machine Intelligence and Slope Stability Laboratory, Department of Geosciences, University of Padova, Padova, Italy. [8]Department of Earth Sciences, University of Cambridge, Cambridge, UK. [9]Department of Geography, University of Cambridge, Cambridge, UK. [10]Department of Risk and Disaster Reduction, University College London, London, UK. [11]Technical University of Berlin, Berlin, Germany. [12]BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany. [13]Pinterest, San Francisco, CA, USA. [14]LinkedIn, Mountain View, CA, USA.
✉e-mail: jesper@dramsch.net

Published online: 5 February 2025

## References

1. Dramsch, J. S. *Adv. Geophys.* **61**, 1–55 (2020).
2. Kuglitsch, M. M. et al. *Environ. Res. Lett.* **18**, 093004 (2023).
3. Mamalakis, A., Ebert-Uphoff, I. & Barnes, E. in *xxAI – Beyond Explainable AI* Vol. 13200 (eds Holzinger, A. et al.) 315–339 (Springer, 2022).
4. Fleming, S. W., Watson, J. R., Ellenson, A., Cannon, A. J. & Vesselinov, V. C. *Nat. Geosci.* **14**, 878–880 (2021).
5. Gevaert, C. M. *Int. J. Appl. Earth Obs. Geoinf.* **112**, 102869 (2022).
6. Ghaffarian, S., Taghikhah, F. R. & Maier, H. R. *Int. J. Disaster Risk Reduct.* **98**, 104123 (2023).
7. Toms, B. A., Barnes, E. A. & Ebert-Uphoff, I. *J. Adv. Model. Earth Syst.* **12**, e2019MS002002 (2020).
8. Rudin, C. *Nat. Mach. Intell.* **1**, 206–215 (2019).
9. Gunning, D. & Aha, D. W. *AI Magazine* **40**, 44–58 (2019).
10. Czaja, W., Fendley, N., Pekala, M., Ratto, C., & Wang, I. J. Adversarial examples in remote sensing. In *Proc. 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* 408–411 (Association for Computing Machinery, 2018).
11. Lapuschkin, S. et al. *Nat. Commun.* **10**, 1096 (2019).
12. Roscher, R., Bohn, B., Duarte, M. F. & Garcke, J. *IEEE Access* **8**, 42200–42216 (2020).
13. Dahal, A. & Lombardo, L. *Comput. Geosci.* **176**, 105364 (2023).
14. Dikshit, A. & Pradhan, B. *Sci. Total Environ.* **801**, 149797 (2021).
15. Cheng, X. et al. in *Handbook of Geospatial Artificial Intelligence* (eds Gao, S. et al.) Ch. 9 (CRC Press, 2023).

## Acknowledgements

We are appreciative of Chelsea Shu's first investigations of the natural language processing tool for exploring the scientific literature and the collaboration and insight provided by use case proponents of the Focus Group on AI for Natural Disaster Management (https://www.itu.int/en/ITU-T/focusgroups/ai4ndm/Pages/default.aspx).

## Competing interests

The authors declare no competing interests.