Supporting Information for

How Interpretable Machine Learning Can Benefit Process Understanding in the Geosciences

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

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

* Corresponding author. Email: sjiang@bgc-jena.mpg.de

Contents of this file

Introduction

Texts S1-S4 provide detailed examples and case studies from the existing literature that illustrate the general rules and theoretical steps involved in the workflow of using interpretable machine learning (IML) presented in Section 3 of the main text. Specifically,

- Text S1 supplements Section 3.1: Translating geoscientific research questions into IML tasks
- Text S2 supplements Section 3.2: Preparing and preprocessing data
- Text S3 supplements Section 3.3: Training and validating ML models
- Text S4 supplements Section 3.5: Distilling interpretation results into geoscientific understanding

Text S1. Examples of translating geoscientific research questions into IML tasks

The types of relationships typically explored with IML involve identifying key influencing factors and their contributions (i.e., understanding how a specific outcome can be individually attributed to each factor) and deciphering dependencies and conditional effects (i.e., determining how multiple factors collectively or interactively affect a particular outcome). Thus, it is essential to identify the specific outcomes (Y) that need to be explained and to determine the factors or variables (X) that are hypothesized to influence those outcomes.

For instance, Jiang *et al.* (2022b) aimed to identify possible mechanisms behind river flooding by examining the predictive contribution of meteorological drivers, where the formulated IML task was to quantify the relationship between extreme runoff events (Y), and precipitation and temperature over the past 180 days (X). In another study focused on identifying meteorological patterns critical for predicting extreme precipitation events (Davenport & Diffenbaugh, 2021), the selected predictors (X) were sea level pressure and 500-hPa geopotential height anomalies, and the outcome of interest (Y) in this case was a binary variable indicating extreme versus non-extreme precipitation. Similarly, for ecological studies such as investigating which environmental factors may affect the distribution of certain species (e.g., African elephant) (Ryo *et al.*, 2020), the potential predictors (X) in the IML included several land use and climatic factors, while the dependent variable (Y) was the geographic distribution of the species. Moreover, following the functional capability of the IML in examining how X contributes to Y, a broader range of potential research questions can be considered, which may include identifying the primary factors relevant to wildfire occurrence in a given region (Kondylatos *et al.*, 2022), or understanding how urban development may affect stream water quality (Wang *et al.*, 2021). Possible questions that focus on feature dependencies and interactions (i.e., the interactive effects of X on Y) include how atmospheric chemical processes and meteorological factors together may influence ozone formation (Xu *et al.*, 2023) or at what soil water content does vapor pressure deficit have the greatest effect on plant water stress (Wang *et al.*, 2022). The question of critical thresholds in systems can also sometimes be translated into IML tasks by identifying inflection points in the contribution of X relative to its value. For example, Chakraborty *et al.* (2021) used IML to explore the non-linear hydroclimatic dependencies and interactions underlying hydrological droughts, uncovering a critical temperature point beyond which groundwater depletion occurs despite increased average precipitation.

Text S2. Examples of data preparation and preprocessing

In addition to following the general principles of data preparation for ML models, attention must be taken to ensure consistency between the data and the underlying processes so that they are represented at appropriate and relevant temporal and spatial scales. For example, in Jiang *et al.* (2022a), overly large or small catchments were excluded from the dataset to account for potential heterogeneity and to ensure a match between the spatial resolution of the meteorological data and the size of the catchment. Likewise, in Li *et al.* (2022), who investigated changes in global vegetation sensitivity to soil moisture, growing seasons were identified across experiments and hydroclimatic zones to ensure temporal consistency in the datasets.

During data pre-processing, depending on the research question, it may be necessary to remove seasonality and long-term trends from time series data. For example, when dealing with leaf area index (LAI) products from different satellite sources where discrepancies are common, removing trends is important to lessen common trends caused by exogenous factors such as CO_2 and biases resulting from multi-sensor drifts in satellite instruments (e.g., Li *et al.*, 2022). Similarly, an observed systematic increase in geopotential height values in recent years, attributed to tropospheric warming, necessitated the removal of such trends when using this variable to interpret anomalies in extreme precipitation events (e.g., Davenport & Diffenbaugh, 2021). This would allow the model to focus on spatially non-uniform changes in geopotential height, rather than the overall homogeneously elevated values.

When the target variable is a relative value (e.g., when using a binary value to encode extreme and non-extreme precipitation), it is important to note that if this relative value is calculated, for instance, per grid cell, the causes of the target variable may now vary in space. Therefore, the same combinations of predictor values could now correspond to different target values depending on the location. This could lead to poor ML model performance and/or misleading IML results, especially if the distribution of the target values is very spatially heterogeneous.

Text S3. Examples and details of ML model training and validation

It is important to recognize the applicability and limitations inherent in different ML models. Each model has its own set of strengths and weaknesses, which can make some models more appropriate for certain types of geoscience questions and data than others. For instance, the conceptual analogy between the catchment memory effect and the recurrent cells of the long short-term memory (LSTM) network (Lees *et al.*, 2022), has led to the prevalence of LSTM in hydrological studies using basin-scale time series (e.g., Jiang *et al.*, 2022a; Kratzert *et al.*, 2019). The ability of convolutional neural networks (CNNs) to handle multidimensional array data makes them particularly suitable for spatiotemporal climate and weather data (e.g., Ham *et al.*, 2019). Tree-based regression models are versatile for tabular data (Grinsztajn *et al.*, 2022), but are usually limited in their ability to predict values beyond the range encountered in their training data. In contrast, generalized additive models or neural networks may be more successful at extrapolating trends. Moreover, in spatial regionalization tasks, it is particularly important to not only exploit tabular spatial data, but also to account for spatial proximity and autocorrelation by combining with geostatistical Kriging or Gaussian process models (Saha *et al.*, 2021).

When training ML models, data splitting is a fundamental step to ensure that the model can learn a generalizable relationship. For robustness and reliability of model interpretation, it is necessary to ensure that the split training and testing sets can accurately represent the true distribution of the data (de Burgh-Day & Leeuwenburg, 2023), yet do not overlap in the sequential or spatial information they contain. For example, when dealing with datasets that span decadal timescales, sequential splitting may inadvertently lead to distributional shifts between subsets in the presence of climate change (Lopez-Gomez *et al.*, 2023). It should be noted, however, that applying a random shuffle strategy to data with sequential dependencies may introduce a risk of data leakage due to potential information overlap in the input data (Sweet *et al.*, 2023). Similar issues arise when learning environmental relationships from spatial data, where adjacent data locations may be autocorrelated, and random splitting into training and test data may not adequately represent the typical prediction distances

encountered in model application (Brenning, 2022; Meyer & Pebesma, 2022). Moreover, in scenarios involving unbalanced datasets, such as those encountered in classification tasks with disproportionate frequencies of categories (e.g., rare extreme events), it may be helpful to employ a stratified data splitting strategy to preserve the distribution observed in the full dataset (e.g., Davenport & Diffenbaugh, 2021; McGovern *et al.*, 2019) or to use custom loss functions that emphasize more extreme events (e.g., Lopez-Gomez *et al.*, 2023). In addition to data splitting, ML models usually involve multiple hyperparameters that should be tuned to optimize learning behavior, often using techniques such as grid search or random search (Bischl *et al.*, 2023). Cross-validation is an essential step in assessing the generalizability of a model to independent datasets. It involves repeatedly partitioning the data into different subsets, using one subset for training and the rest for validation, to ensure that every data point has been used for both training and testing. Data splitting or cross-validation for hyperparameter tuning and model assessment should be performed in a nested fashion to avoid information leakage from tuning into the validation step (Schratz *et al.*, 2019). Throughout the model training process, additionally, it is usually necessary to implement strategies (e.g., regularization and early stopping) to prevent the model from exploiting certain patterns or shortcuts in the training data to overfit (Ying, 2019).

Text S4. Details on the examples of distilling interpretation results into geoscientific understanding

Figure 2b-d in the main text illustrates the form of interpretation results based on three types of data typical in the geosciences, i.e., spatial data, multivariate time series, and tabular data, from the literature (Davenport & Diffenbaugh, 2021; Jiang *et al.*, 2022a; Wang *et al.*, 2022). For spatial data, interpretation methods are often used to produce heatmaps of feature importance over the spatial input domain (e.g., the pixel relevance map in Figure 2b). For instance, the pixel relevance map for the daily sea level pressure (SLP) anomaly highlights specific areas that influence the model's prediction of the occurrence of extreme precipitation circulation patterns (EPCPs) over the Midwest on a given day. The example suggests the sensitivity of the EPCP to the location and presence of strong, negative SLP anomalies (Davenport & Diffenbaugh, 2021). In multivariate time series, interpretation methods can reveal how input variables contribute to a particular prediction over time by assigning feature importance values. In the illustrated case of predicting a specific streamflow peak, it is shown that past precipitation events collectively contribute more than recent precipitation, as evidenced by a higher sum of integrated gradient values than other periods or variables (not shown here). This pattern often implies that soil moisture, elevated by antecedent precipitation, plays a key role in predicting flood events other than recent precipitation events (Jiang *et al.*, 2022a). With tabular data, where dimensions are comparatively fewer, interpretations can be more straightforward. The example shows how variables such as soil water content (SWC) and vapor pressure deficit (VPD) positively influence the model's prediction of gross primary production (GPP) in a given half hour for a specific FLUXNET site, while air temperature (TA), incoming shortwave radiation (RAD), and CO_2 exhibit negative effects. The role of each variable can be presented in a quantifiable way that illustrates how it moves the predicted value of GPP from the expected model output over the background dataset to the model output for that prediction (Wang *et al.*, 2022).

Figure 2e-g in the main text presents aggregated interpretation results, corresponding to those in Figure 2b-d, using various strategies. In the study by Davenport & Diffenbaugh (2021), for example, composite maps were created for EPCP and non-EPCP days to elucidate overall circulation regions and patterns that drive extreme event predictions. Figure 2e shows the composite map of SLP anomalies during EPCP days, highlighting the CNN model's focus on circulation features over the Midwest. In contrast, non-EPCP days exhibit no clear spatial coherence or a particular region of high relevance (not shown here). Alternatively, clustering patterns of feature importance across different instances or scenarios that segment the data into groups is useful for identifying common underlying mechanisms or processes. For instance, Jiang *et al.* (2022a) applied cluster analysis to the feature importance values (as exemplified in Figure 2c) for all annual maximum discharge events across European basins and identified three major contribution patterns. The cluster depicted in Figure 2f is characterized by high importance of antecedent precipitation, with a spatial pattern in the proportion of events clustered in this category in individual basins.

Moreover, it can be informative to examine how a feature's contribution to model predictions changes with its value and the value of other variables. For example, the bee swarm plots shown in Figure 2g provide a dense summary of each feature's impact on model output, where each dot corresponds to a site-half-hourly sample in the study (Wang *et al.*, 2022). The position of each dot along the x-axis indicates the impact of a variable on the predicted GPP for that sample, with the color indicating the original feature value. The figure shows the important role of VPD, RAD, and SWC in GPP sensitivities. Specifically, high VPD negatively induces larger variations in GPP dynamics with long left tails, while the contribution of SWC shows a relatively balanced pattern across its range. The dependence plot in Figure 2g further reveals how the model's predictions depend on interactions between multiple features. As shown in the plot, the same VPD value can have different effects on GPP, which is relevant to SWC values (indicated by the color). This example illustrates a significant coupling and interactive effect between SWC and VPD, which is valuable for understanding the influence of climatic drivers on water stress-vegetation relationships.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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