LESSONS LEARNED

Limitations of XAI Methods for Process-Level Understanding in the Atmospheric Sciences

SAM J. SILVA\textsuperscript{a,b} AND CHRISTOPH A. KELLER\textsuperscript{c,d}

\textsuperscript{a} Department of Earth Sciences, University of Southern California, Los Angeles, California
\textsuperscript{b} Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, California
\textsuperscript{c} Morgan State University, Baltimore, Maryland
\textsuperscript{d} NASA Global Modeling and Assimilation Office, Greenbelt, Maryland

(Manuscript received 1 June 2023, in final form 10 October 2023, accepted 7 December 2023)

ABSTRACT: Explainable artificial intelligence (XAI) methods are becoming popular tools for scientific discovery in the Earth and atmospheric sciences. While these techniques have the potential to revolutionize the scientific process, there are known limitations in their applicability that are frequently ignored. These limitations include that XAI methods explain the behavior of the AI model and not the behavior of the training dataset, and that caution should be used when these methods are applied to datasets with correlated and dependent features. Here, we explore the potential cost associated with ignoring these limitations with a simple case study from the atmospheric chemistry literature – learning the reaction rate of a bimolecular reaction. We demonstrate that dependent and highly correlated input features can lead to spurious process-level explanations. We posit that the current generation of XAI techniques should largely only be used for understanding system-level behavior and recommend caution when using XAI methods for process-level scientific discovery in the Earth and atmospheric sciences.

KEYWORDS: Atmospheric chemistry; Artificial intelligence; Decision trees; Machine learning; Model interpretation and visualization; Regression

1. Introduction

Explainable artificial intelligence (XAI) is a set of artificial intelligence (AI) techniques that can provide human-understandable explanations for the decision-making processes and predictions of AI models. In the atmospheric sciences, XAI has seen rapid adoption in conjunction with the rise in popularity of AI techniques more generally (e.g., Barnes et al. 2020; Holzinger et al. 2022; McGovern et al. 2019). As they are currently used, these XAI methods aid in scientific discovery, aim to promote trust and transparency in AI prediction systems, and broadly lead to improved understanding of the strengths and limitations of a particular AI application (e.g., Ebert-Uphoff and Hilburn 2020; Mamalakis et al. 2022; McGovern et al. 2020; Molina et al. 2021; Silva et al. 2022; Toms et al. 2020).

There are several important caveats in using these XAI methods. Principal among them is that the explanations derived tell users about the AI model and do not necessarily tell users anything useful about the dataset used to train the model (e.g., Molnar 2019; Molnar et al. 2021). In addition, many of these techniques assume that model input features are independent (e.g., Lundberg and Lee 2017). Both of these caveats are commonly violated when XAI methods are used in the Earth and atmospheric sciences. The purpose of this study is to explore the potential pitfalls in ignoring these known issues.

Variables in the atmosphere often exhibit high levels of dependence that can manifest in data as high correlation. Leveraging this correlation for predictive skill has provided a major benefit to many modern machine learning tasks, where the exact details of the relationship being studied are not as critical as accurate prediction (e.g., Geiss et al. 2022; Rasp et al. 2018; Silva et al. 2021b). However, these machine learning methods are increasingly being used for scientific discovery. In such cases, the methods are used to predict a quantity of interest on the basis of a set of these correlated inputs. The behavior of that model is then interrogated using XAI methods, the results of which are often used for hypothesis generation surrounding a particular study case (e.g., Silva et al. 2022; Wang et al. 2022).

Here, we investigate the impact of feature dependence on commonly used explainability techniques. Using predictions of atmospheric chemical reaction processes with boosted regression trees as a case study, we demonstrate that correlated and dependent features lead to spurious process-level explanations. This effect is sufficiently large as to potentially lead to incorrect scientific understanding of the underlying process. In the highlighted case here, a user would attribute chemical...
reactions to the fundamentally wrong compounds. These results are broadly applicable to the use of explainability methods with tree-based machine learning models across the Earth and atmospheric sciences and suggest caution in their interpretation.

2. Atmospheric chemistry case study

To investigate the impact of correlated and dependent features on XAI methods, we use simulated atmospheric chemistry data. These data are a useful case study because they represent a real machine learning use case with an active research community (e.g., Ivatt and Evans 2020; Keller and Evans 2019; Kelp et al. 2022; Sturm et al. 2023; Sturm and Wexler 2022), and the system represents a wide range of correlations across chemical species. The dataset we used was generated through coarse-resolution simulations from the NASA GEOS composition forecast modeling system (GEOS-CF; Keller et al. 2021), which predicts the global distribution of hundreds of chemical species using the GEOS-Chem chemistry mechanism (Bey et al. 2001). We sample a single time step of the simulated global 3D global distribution of atmospheric chemical species at approximately 2° × 2° spatial resolution (c8) with 72 vertical levels on 1230 UTC 2 June. We then subset the model output to 17 of the species contained within the SuperFast chemical mechanism (Kelp et al. 2022), and the system represents a wide range of correlations across chemical species. The variables in this dataset are known to be chemically interdependent, with dependencies based on the structure of the GEOS-CF model. This is evident in the spatial correlation plots in Fig. 1, showing a histogram of all cross correlations in the dataset (without self-correlations) and a heat map of all correlation combinations. The nonidentity correlations vary from −0.6 to 0.8, centered near a correlation of 0. This is broadly consistent with the highly coupled nature of atmospheric chemistry, with certain variables strongly dependent on others through direct chemical reactions. It is worth noting that these correlations, while high, are not extreme. Correlations of up to 0.8 are reasonably common in the atmospheric sciences.

The target prediction quantity investigated here is the bimolecular reaction rate from a particular reaction. Bimolecular reactions are generally of the following form:

\[ A + B \rightarrow C, \quad \text{rate} = k_{A,B} \times [A] \times [B], \tag{1} \]

where \( A \) and \( B \) are chemical reactants and \( C \) is a chemical product. The reaction rate (molecules cm\(^{-3}\) s\(^{-1}\)) is calculated as the product of the concentration of chemical species \( A \) and \( B \) along with the reaction-rate constant \( k_{A,B} \) (cm\(^3\) molecules\(^{-1}\) s\(^{-1}\)).

In this case study, we use the oxidation of methyl hydroperoxide (MP; CH\(_3\)OOH) by the hydroxyl radical (OH) to various product chemical species:

\[ \text{MP} + \text{OH} \rightarrow \text{products}, \quad \text{rate} = k_{\text{OH,MP}} \times [\text{OH}] \times [\text{MP}]. \tag{2} \]

This reaction is present in most chemical mechanisms and is a chemical intermediate in the oxidation of methane in the atmosphere. Both OH and MP are known to be strongly dependent on other species in the chemical system and thus have moderately high correlations with these other species (see Fig. 1). The rate constant used was generated following the SuperFast mechanism, assuming constant temperature of 300 K.

We ultimately chose this system and prediction task because of the relative simplicity of interpreting the true functional form we are trying to learn (multiplication) and the high levels of dependence across the atmospheric chemical space.

3. Boosted regression trees and explainability

We predicted the reaction rate using boosted regression trees, implemented with the extreme gradient boosting (XGBoost) software package (Chen and Guestrin 2016). Boosted regression trees, and XGBoost in particular, have been used widely across the atmospheric chemistry space and Earth and atmospheric sciences.
We investigate three widely used XAI methods applied to XGBoost models: gain, permutation feature importance, and Shapley additive explanations (SHAP) analysis. Gain and SHAP analysis methods are available directly from the XGBoost library, and we implement permutation feature importance ourselves for this work. We describe these methods in brief here. Each of them provides a quantitative measure of feature importance that is sensitive to the training dataset and feature distributions. For more detail on explainability techniques, see Molnar (2019) and McGovern et al. (2019).

Gain is an importance metric that is unique to tree-based models. It represents the improvement in accuracy brought to a model by the splits on a given variable within the tree structure. The addition of a split on a variable can (and frequently does) improve the accuracy of a model relative to if that split is not included. At a high level, that improvement of accuracy (with the inclusion of a regularization term) is what is used to calculated gain.

Permutation feature importance explains the model behavior by ranking the variables by their contribution to an accurate prediction. The accuracy of a trained model is assessed by randomly permuting input variables one at a time. The influence of that random permutation on the model predictive skill is quantified across many prediction cases. The average difference in error between permuted cases and the original dataset is used as the importance ranking for each individually permuted variable. In this way the variables ranked as “most important” are those that when permuted lead to the worst predictions. Permutation feature importance model agnostic and works on all widely used machine learning models (e.g., McGovern et al. 2019).

SHAP analysis is an XAI technique that uses methods based in game theory literature to explain individual model predictions (Lundberg et al. 2019, 2020). In contrast to gain and permutation feature importance, which provide one bulk explanation for the entire model, SHAP analysis provides information at the individual prediction level. This ability to explain individual predictions, termed “local explainability,” has made SHAP analysis a popular choice for interrogating machine learning methods in the Earth and atmospheric sciences (e.g., Batunacun et al. 2021; Wang et al. 2021). For each model prediction, SHAP values can be calculated for each input feature. These SHAP values quantify the influence of every feature toward either increasing or decreasing the final predicted quantity using a local surrogate modeling method. Higher-magnitude SHAP values indicate that a variable has a larger influence on the resulting prediction. Averaging the magnitude of all SHAP values leads to a bulk, or “global,” importance ranking that can be compared with those from gain and permutation feature importance. The detailed derivation and mathematical calculation of SHAP values for machine learning models can be found in Lundberg et al. (2019, 2020). As with permutation feature importance, SHAP analysis is model agnostic, works on all widely used machine learning models, and the implementation in XGBoost is optimized for computational efficiency (Lundberg and Lee 2017).

4. Model training and XAI experiments

We trained a boosted regression tree to predict the reaction rate following Eq. (2) from chemical species concentrations. To explore the impact of correlated input features, we selected all species from the SuperFast mechanism with correlations greater than 0.5 with either MP or OH as input features to the prediction model. The final species feature list was (CH₃O, CH₄, CO, H₂O, H₂O₂, HO₂, MO₂, MP, OH). We log-transformed the target prediction quantity (reaction rate) for improved predictive skill. The dataset contains 922,896 samples and was randomly split into training, validation, and test sets with ratios of 0.85/0.1/0.05. Model hyperparameters selected include setting the maximum depth of a tree (“max_depth”) to 12 and the learning rate (“eta”) to 0.1. We trained the model for 50 boosting iterations using a mean square error cost function. All other hyperparameters were kept at the package default values. Results discussed further in this paper are solely based on the test dataset.

We evaluated the impact of variable correlation on the prediction of atmospheric chemical reaction rates through comparison with a random baseline. In this random baseline, the relationships between all variables are broken through randomly reordering each variable in the original dataset. This was achieved by an individual row-wise permutation of each input feature (here, chemical species) by a randomly shifted index. This breaks all correlation and dependence in the dataset but importantly preserves the individual variable distributions (Gaussian, lognormal, etc.). From there, the reaction rate is recomputed and used as a new target for training.

We posit that the difference in the results of the XAI methods between the original and randomized baseline will inform us as to the potential scale of the issue associated with ignoring feature correlations in XAI analyses.

5. Results and discussion

We trained an XGBoost model to predict the MP + OH reaction rate on both the original dataset and the randomized baseline. Overall performance is very good for both cases. Predictions of the original dataset had a relative mean square error of 0.008 and a correlation coefficient of 0.999. The randomized dataset showed even better predictive skill, with a mean square error on the test dataset of 0.006 and a correlation coefficient of 0.999. Scatterplots of the respective model’s predictive skill are shown in Fig. 2 (residual plots are shown in Fig. S1 in the online supplemental material). This very high skill for both models is not particularly surprising given the reasonable triviality in learning multiplication with high-capacity machine learning models.

We then ran the three XAI methods described above to generate global feature importance rankings for this prediction task. Results are summarized in Fig. 3 for both the original and randomized datasets. The most striking result is the distinction between the importance rankings in the original dataset and the randomized baseline. A particular input feature, the compound HO₂, is shown to be important for all three methods in the original dataset despite having no
explicit connection to the target quantity described in Eq. (2). In the randomized baseline experiments, only OH and MP are ranked as important. All other variables, which are not used in the direct calculation, are correctly given trivially low importance rankings.

The results shown in Fig. 3 demonstrate the distinction between process- and system-level explanations. HO2 is an oxidant and is tightly coupled to the system of chemical reactions in the atmosphere and does correlate well with the rate (Seinfeld and Pandis 2016). It is correct to state that HO2 has an important influence on the reaction rate predicted here through the influence that HO2 has on the overall atmospheric chemical system. However, HO2 is not important for the specific instantaneous process of the reaction between OH and MP, as predicted here. In this way, these XAI methods are only useful as system-level explanations given this particular correlated input distribution.

In the random baseline scenario, the importance rankings for MP and OH differ. This difference exists even though in Eq. (2) the reaction rate is calculated while mathematically weighing both variables equally through their product. This difference in importance demonstrates how many XAI methods depend on the distribution (variance, etc.) of an individual variable and how differences in distributions manifest in the final explainability result. If both variables were randomly sampled from the same distribution (e.g., Gaussian) with the same variance and mean, their importance would be identical.

The spurious results from the overall importance rankings are echoed in so-called local explanations from the SHAP analysis. Local explanations provide explanations of individual model predictions. This can be very useful to study the model behavior at a high level of detail. Using the local SHAP analysis method, we tracked the number of times that both OH and MP appeared in the top three most important variables for a given prediction as assessed by the highest-magnitude SHAP values. In ~91% of individual prediction explanations, either OH or MP was not present in the top three most important variables for a given prediction as assessed by the highest-magnitude SHAP values. In the random baseline, SHAP analysis assigns the top two most important variables to OH and MP in nearly all cases.

In this case study, knowing the true functional form of the target variable makes it easy to identify the model trained on the original dataset as untrustworthy from a process-level

---

**Fig. 2.** Performance scatterplots of the log10 of the reaction-rate (molecules cm^{-3} s^{-1}) predictions for this case study for (left) the original dataset and (right) the randomized baseline.

**Fig. 3.** Average importance across the three methods. The original dataset is shown in black, and the randomized baseline is shown in red. Differences between the original dataset and randomized baseline importance values are highlighted with gray lines. Values very close to zero have both the original and randomized points overplotted.
perspective. The challenge arises if we imagine this scenario as one in which we did not know the functional form of the reaction rate a priori. If we were using this technique for scientific discovery, the results shown in Fig. 3 for the correlated distribution could very easily lead a reasonable scientist astray. It is reasonable to assume that an unknown reaction in the atmosphere is dependent on HO₂ as a reactant. HO₂ is a very common reactant in the atmosphere, participating in many reactions in current models of atmospheric chemistry (Silva et al. 2021a). If the drivers of the reaction rate were unknown and we used XAI techniques to try to discover them at a process level, variable dependence would lead to an incorrect answer.

It is important to note that the issue demonstrated here cannot be fixed through more advanced XAI techniques currently available. In the correlated dataset example, the model trained on the original dataset makes high-quality predictions using information from HO₂. The issue is that HO₂ is only important at a system level as an indicator of active oxidation chemistry. HO₂ is not important for the specific reaction process—it is not part of the reaction calculation. The consequences of this are that any correct post hoc XAI technique will be fundamentally out of line with the scientific process being studied. This would even be true if we were to use certain methods that aim to help address issues of feature dependence in XAI techniques. These include updated but computationally expensive calculations of SHAP (e.g., Aas et al. 2020) and block permutation feature importance (e.g., McGovern et al. 2019). These spurious results are directly a consequence of XAI techniques providing explanations for AI models and not necessarily the training datasets. The incorrect explanation is due to the XGBoost model behavior and not the XAI method chosen.

6. Implications for XAI in atmospheric sciences

We explored the impact of feature dependence on XAI methods in the Earth and atmospheric sciences. Using a simple but realistic example from atmospheric chemistry, we demonstrated that correlated data lead to incorrect process-level explanations with boosted regression trees. These incorrect explanations exist despite very high model accuracy, highlighting the challenge that high-quality predictions and scientifically reasonable explanations are insufficient evidence that a machine learning model is correctly representing the processes being studied.

This work reinforces that XAI techniques inform us about AI model behavior and that good predictive skill of an AI model is not a sufficient criterion for assuming that the model correctly represents the data-generation properties of the training dataset. Instead of learning specific processes, under atmospherically relevant correlations the model is learning systemwide behavior. This does not render XAI methods useless or invalidate the predictive skill of the model, and it in fact directly reflects the properties of the training dataset. However, this result does demonstrate that these XAI techniques should not be used on their own for scientific discovery and that instead these techniques should be used as part of a suite of analytical approaches to ensure robust process-level scientific discovery (physical modeling, causal inference, etc.). We believe that these results are broadly applicable across the Earth and atmospheric sciences and suggest caution in using and interpreting the results of XAI methods.

Acknowledgments. Author Silva acknowledges NASA Grant 80NSSC23K0025 for supporting this work. Resources supporting the GEOS model simulations were provided by the NASA Center for Climate Simulation at the Goddard Space Flight Center (https://www.nccs.nasa.gov/services/discover).

Data availability statement. All data and analysis code are publicly available online (https://doi.org/10.5281/zenodo.10520095).

REFERENCES


