Using an Interpretable Machine Learning Approach to Characterize Earth System Model Errors: Application of SHAP Analysis to Modeling Lightning Flash Occurrence

Sam J Silva[1], Christoph A Keller[2,3], Joseph Hardin[1,4]

[1]Pacific Northwest National Laboratory, Richland, WA, USA

[2]Universities Space Research Association, Columbus, MD, USA

[3]Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA

[4]ClimateAi, Inc. San Francisco, CA, USA

Corresponding author: Sam J Silva (sam.silva@pnnl.gov)

Key Points:

- Errors in simulated lightning flash occurrence are learned through a machine learning classification approach
- An interpretable machine learning technique is used to explore the drivers of the simulation errors
- This error prediction system indicates that errors are strongly related to convective processes and the characteristics of the land surface

Abstract

Computational models of the Earth System are critical tools for modern scientific inquiry. Efforts toward evaluating and improving errors in representations of physical and chemical processes in these large computational systems are commonly stymied by highly nonlinear and complex error behavior. Recent work has shown that these errors can be effectively predicted using modern Artificial Intelligence (A.I.) techniques. In this work, we go beyond these previous studies to apply an interpretable A.I. technique to not only predict model errors but also move toward understanding the underlying reasons for successful error prediction. We use XGBoost classification trees and SHapley Additive exPlanations (SHAP) analysis to explore the errors in the prediction of lightning occurrence in the NASA GEOS model, a widely used Earth System Model. This interpretable error prediction system can effectively predict the model error and indicates that the errors are strongly related to convective processes and the characteristics of the land surface.

Plain Language Summary

Computer models of the Earth are very important tools in the modern Earth scientist's toolkit. Understanding when and why these models are wrong is a major challenge facing the scientific community. Work published in the last few years has shown that you can actually predict when these models are wrong using artificial intelligence. We build on that work by applying existing fancy mathematical tools to these artificial intelligence methods to understand why these these computer models are wrong. We demonstrate this approach to predictions of lightning in a model created by NASA, and find that the lightning in the model is wrong in ways that are strongly related to convection in the atmosphere and the aspects of the land surface.

1 Introduction

Computational models are a key component of modern scientific efforts throughout the Earth System Sciences. These models have become sufficiently complex as to include representations of a wide array of important physical and chemical processes in the atmosphere, oceans, and land (IPCC, 2021). As model complexity has grown, so has their applicability and utility for answering policy relevant questions ranging from short term forecasting through climate prediction. Central to the efforts to improve these models is accurate assessment and diagnosis of errors throughout the representations of these physical and chemical processes. Traditionally, error assessment approaches usually combine computational and statistical tools with expert judgement to uncover the error behavior in these Earth System Models.

Recent work applying techniques from the artificial intelligence literature has shown that, in certain cases, errors in these models can be predicted using machine learning methods. For example, Rasp & Lerch (2018) use neural networks to correct errors in numerical weather forecasts to better predict surface temperature across Germany. Keller et al. (2021) use boosted regression trees to predict and adjust for the errors in simulating the chemical composition of the atmosphere. While these previous studies use machine learning techniques to predict model error with respect to observed quantities, similar approaches have been applied to predict the error of simplified models with more complex theoretical baselines (e.g. Silva, et al., 2021).

In concurrent research, the application of so-called "interpretable AI" techniques to research problems in the Earth System Sciences has shown great promise in the ability to evaluate how and why various machine learning techniques make a given prediction. Barnes et al. (2020) demonstrate how neural networks, when combined with interpretable A.I. techniques can be used to discover indicator patterns of change in the climate system. Toms et al. (2020) further explored how two different methods, layerwise relevance propagation and backward optimization, can be used to glean scientifically relevant information from neural network predictions of variability in the Earth System. Stirnberg et al. (2021) use an alternative method, SHapley Additive exPlanations (SHAP) applied to boosted regression trees, to quantify the importance of various meteorological drivers on particulate matter concentrations.

We build upon these studies and integrate interpretable A.I. techniques with machine learning predictions of errors in Earth System Models, with the ultimate goal of improving the representation of physical and chemical processes. As a demonstration of the methodology described in this work, we apply SHAP analysis to boosted classification trees to characterize errors in lightning flash occurrence in the NASA Goddard Earth Observing System (GEOS) model. We find that these errors are strongly related to convective processes and the characteristics of the land surface. In principle, the technique we describe in this work generalizes to any error prediction that can be framed as a classification task in the Earth System Sciences.

2. Methodological Approach

This work centers around first developing a machine learning predictor of the error in a model system, followed by interrogating that machine learning predictor using interpretable A.I. techniques. Here, we specifically use boosted classification trees from the XGBoost software library and SHAP regression values for interpretability. Our methodological approach is summarized visually in Figure 1, and described in more detail in the following section. This methodology is predicated on the assumption that if a machine learning system provides a high-quality skillful prediction of the error in a given Earth System model, probing the behavior of the machine learning system can yield insight into the behavior of the Earth System model errors.

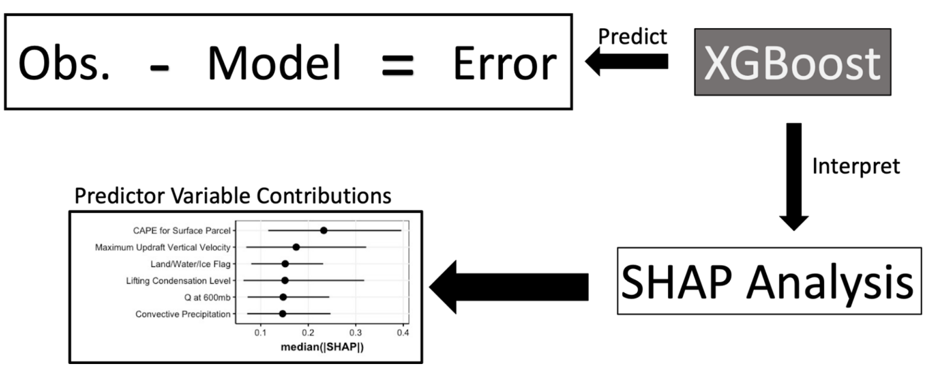

Figure 1. A visual schematic of the methodological approach used in this work.

2.1 Error Prediction

For the purposes of this work, we define the model error as simply the residual of the model prediction with respect to a true value. Stated mathematically:

Equation 1) $x_{true} = x_{pred} + \varepsilon$

where x_{true} is the true value, x_{pred} is the prediction, and ε is the error term. While the direct calculation of the error term, ε, is relatively simple, in Earth System Model applications this term can vary as a highly complex function of the model state and structure.

2.2 SHAP Interpretability Analysis

There are a multitude of high-fidelity machine learning interpretability techniques available and applied throughout the sciences (e.g. Barnes et al., 2020; Murdoch et al., 2019; Rasp & Thuerey, 2021, Molina et al. 2021). Here we use SHapley Additive exPlanations (SHAP) regression values (Lundberg et al., 2018, 2020), as they are relatively uncomplicated to interpret and have fast implementations associated with many popular machine learning techniques (including the XGBoost machine learning technique we use in this work).

Analysis of interpretability through SHAP regression values aims to evaluate the contribution of input variables (often called "input features") to the predictions made by a machine learning predictor model. The contribution of that input feature to a prediction is calculated mathematically through the construction of a so-called "explanation model". The explanation model evaluates the predictions of a machine learning system as the sum of the contributions of each input feature and the mean predicted value. Mathematically the explanation model can be stated as:

Equation 2) $y = \bar{y} + \sum_i \varphi_i$

where y is an individual prediction, $\bar{y}$ is the average predicted value across all predictions, and φ_i is the contribution of input feature i to the prediction (also known as the "SHAP regression value" or "SHAP value"). Input variables with larger magnitude SHAP values are interpreted as contributing more to a specific prediction than those with a smaller magnitude SHAP values. For a given case, positive SHAP values indicate a specific feature contributes toward increasing the final predicted value y, and negative SHAP values indicate a contribution toward decreasing the prediction. These SHAP values, i, are calculated following a game theoretic approach to assess prediction contributions (e.g. Štrumbelj and Kononenko, 2014), and have been extended to the machine learning literature in Lundberg et al. (2018, 2020).

Explicitly calculating SHAP values can be prohibitively computationally expensive (e.g. Aas et al., 2020). As such, there are a variety of fast implementations available which approximate SHAP values, optimized for a given machine learning technique (e.g. Chen & Guestrin, 2016). In short, these techniques calculate SHAP values through sampling the predictions of a given model by replacing some model input values with random values from that input distribution. The results of those predictions are weighted as described in Lundberg et al. (2018) and the linear model shown in equation 2 is derived. A more detailed description of the SHAP calculation process and other interpretability metrics can be found in Lundberg et al. (2018, 2020) and Molnar (2019).

The SHAP framework has several key desirable properties, including that the sum of the contributions accurately reproduces the predicted value, and that the contributions of input features that are not present in the machine learning model are assigned values of 0. Additionally, SHAP is a model agnostic technique and can be applied to a wide class of machine learning prediction models. Despite these key advantages, there are several potential deficiencies to the application of SHAP analysis to error characterization in the Earth System Sciences. Principle among these are that SHAP analysis cannot directly yield causal insights and that the direct calculation of SHAP values is very computationally expensive. As such, care must be taken to properly interpret the SHAP values resulting from any particular analysis. It is important to note that to improve computational performance, common implementations of SHAP analysis in existing machine learning libraries (e.g. Chen & Guestrin, 2016) contain assumptions about the data distributions which are not always valid in applications in the Earth System Sciences (e.g., feature independence, Aas et al., 2020).

3. Lightning Occurrence Case Study

As a demonstration of the methodology outlined in Section 2, we evaluate the errors in the lightning occurrence parameterization in the NASA GEOS model using observations from the Geostationary Lightning Mapper (GLM) onboard the GOES-16 satellite as the ground truth. Lightning is a natural hazard in the Earth System with important interactions with biomass burning, atmospheric chemistry, and climate (Schumann and Huntrieser, 2007). Despite its importance, the representation of lightning occurrence in atmospheric models remains a key challenge (Finney et al., 2018; Liu & Yang, 2020; Murray, 2016).

3.1 Dataset Description

Model predicted lightning occurrences (flash rates) were generated using the NASA GEOS ESM, a General Circulation Model (GCM) and Data Assimilation System (DAS) consisting of a suite of model components that can be flexibly connected via the Earth System Modeling Framework (ESMF, Hill et al., 2004) and Modeling Analysis and Prediction Layer (MAPL, Suarez et al., 2007). Here, we use GEOS version 5 (Jason-3_5) with the finite-volume dynamical core of Putman and Lin (2007) at a cube-sphere c90 horizontal grid (approximately 1x1 degrees horizontal resolution) and 72 hybrid-eta levels from the surface to 0.01 hPa. Using the GEOS 'replay' feature (Orbe et al., 2017), the model simulation is nudged toward the pre-computed meteorological analysis fields obtained from the MERRA-2 reanalysis (Gelaro et al., 2017). Convection, which is a key driver of lightning, is parameterized using a combination of the Grell-Freitas mass-flux scheme for deep convection (Freitas et al., 2018) and the Park and Bretherton parameterization for shallow convection (Park and Bretherton, 2009).

The parametrization of lightning used in this work follows the unconstrained cloud top height (CTH) approach described in Murray et al. (2012). Briefly, the parameterization calculates the occurrence of lightning at a given time using a fifth- and second-power function of cloud top height over continents and oceans, respectively, following Price and Rind (1992, 1993, 1994). The cloud top height is defined as the altitude where the upward convective mass flux - as calculated by the GEOS convection code - becomes zero. Lightning is restricted to convective columns that span the full temperature range from 0 ˚C to -40 ˚C (Williams, 1985). Additionally, simulated lightning cannot occur over regions with snow or ice at the surface or regions without any clouds.

While the CTH parameterization has a long development history and is widely used, several other lightning parameterizations exist that are based on different input variables and functional fits. These include parameterizations based on updraft mass flux (Allen and Pickering, 2002), convective precipitation (Meijer et al., 2001) or cloud ice flux (Finney et al., 2014). While these parameterizations are not explicitly tested in this study, we include the input variables for these parameterizations in our machine learning model (see below) to probe a possible relationship between errors in the CTH parameterization and these quantities.

We used the GEOS ESM system to produce hourly-averaged lightning flash rates covering the year 2018, and evaluate this parameterization using lightning flash observations from the Geostationary Lightning Mapper (GLM) on board the GOES-16 satellite (Koshak et al., 2018; Rudlosky et al., 2019). GOES-16 is the first of the latest generation of geostationary weather satellites operated by NASA and the National Oceanic and Atmospheric Administration (NOAA). It covers the GOES East position at 75 deg W, providing a continuous view centered on the Americas. GLM on board of GOES-16 is the first operational geostationary lightning mapper, offering continuous detection of lightning with a spatial resolution of ~10km. It detects and locates lightning within its field-of-view using a single-channel, near-infrared (777.4 nm) optical transient detector with a framerate of 2ms. Here, we use the GOES-16 GLM Level 2 lightning flash product available on the NOAA CLASS data portal (https://www.avl.class.noaa.gov/saa/products/search?datatype_family=GRGLMPROD), which combines individual lightning events that are combined spatially and temporally (GOES-R Algorithm Working Group and GOES-R Series Program, 2018).

We develop a predictor for the error in lighting occurrence predicted by the GEOS ESM by encoding the flash rates in both the GEOS model and the GLM observations as a binary prediction: 0 if there was no flash, 1 if there was a flash. This is mapped on to one of three values for the error prediction as described in Equation 1: 1 if the parameterization predicts no flash where there was an observed flash, 0 if the model predictions are consistent with the observations, and -1 if the model predicts a flash where there was not an observed flash. These three values are then treated as three different classes for a multi-label classification prediction task using the XGBoost library. In order to predict any lightning at all, the NASA GEOS parameterization requires the presence of clouds and a lack of ice at the surface. We pre-filter these trivial cases where the parameterization will never predict a flash to focus our analysis on circumstances where the entirety of the parameterization can be assessed. As lightning is a relatively infrequent event, the datasets here are highly imbalanced with respect to flash occurrence. To treat this dataset imbalance, we downsample such that parity is reached in the flash and no flash cases in the observations.

3.2 Machine Learning Model Training

We develop a gradient boosted classification tree using the XBoost machine learning library (Chen & Guestrin, 2016) to predict the error of the NASA GEOS model parameterization of lightning prediction relative to corresponding GLM lightning observations. The XGBoost model is trained to predict whether the GEOS model accurately predicts the occurrence of lightning for a given set of model conditions (input features). Gradient boosted classification trees are a type of machine learning classification model, wherein a number of small tree-based models are trained to predict a categorical variable. Here, that categorical variable is the error as defined in Equation 1. The possible categories are: -1, when the NASA GEOS model underestimates lightning occurrence, 0, when the NASA GEOS model correctly predicts lightning occurrence, and 1, when the NASA GEOS model over predicts lightning occurrence. After the first tree is trained, each new tree is iteratively trained to predict the residuals of the previous tree. This residual prediction through addition of trees continues for either a specified number of iterations or until satisfactory or convergent predictive skill has been achieved with respect to some particular criteria. In this multilabel classification task, the XGBoost machine learning model predicts a probability that a given set of input features will lead to any of the three classes (underestimation/correct prediction/overestimation). The class with the highest probability is selected for the final classification. The gradient boosted tree implementation in the XGBoost library has been applied widely across applications in the Earth System Sciences (e.g. Batunacun et al., 2021; Ivatt & Evans, 2020; Keller et al., 2021; Silva, et al., 2020) and has computationally efficient open-source implementations in a variety of commonly used programming languages, including the calculation of SHAP values.

The input values to the XGBoost classifier are summarized in Table 1, consisting of a variety of diagnostics related to atmospheric physics and dynamics as well as the land surface. These parameters were chosen based on the characteristics of the CTH parameterization used in GEOS, as well as other parameters thought to be important for lightning and commonly used in other lightning parameterizations, such as cloud ice, vertical updraft velocity, or convective precipitation (Meijer et al., 2001; Allen and Pickering, 2002; Finney et al., 2016). A sampling height of 440 hPa is chosen for 3-dimensional quantities, following the approach by Finney et al. (2018).

Variable Name	Units
Temperature	K
Eastward Wind Component (U)	$m\ s^{-1}$
Northward Wind Component (V)	$m\ s^{-1}$
Specific Humidity	$kg\ kg^{-1}$
Q at 600mb	$kg\ kg^{-1}$
Grid Box Mass Fraction of Cloud Ice Water	$kg\ kg^{-1}$
Mass Fraction of Convective Cloud Ice Water	$kg\ kg^{-1}$
In-Cloud Cloud Ice	$kg\ kg^{-1}$
Ice Water Content	$kg\ m^{-3}$
Sedimentation Loss of Cloud Ice	$kg\ m^{-2}\ s^{-1}$
Ice Water Path	$kg\ m^{-2}$
Level of Free Convection	m
Lifting Condensation Level	m
Height of Cloud Base Layer	m
Inhibition for Surface Parcel	$J\ kg^{-1}$
Buoyancy of Surface Parcel	$m\ s^{-2}$
Vertical Pressure Velocity	$Pa\ s^{-1}$
Updraft Vertical Velocity	$hPa\ s^{-1}$
Pressure at Convective Cloud Top	Pa
Pressure at Convective Cloud Base	Pa
Total Cloud Area Fraction	-
CAPE for Surface Parcel	$J\ kg^{-1}$
Convective Precipitation	$kg\ m^{-2}\ s^{-1}$
Land/Water/Ice Flag	-
Latitude	Degrees East
Longitude	Degrees North

Table 1. Variables used as inputs to the machine learning predictors in this work. Unless otherwise stated, all meteorological variables are taken at 440 hPa.

To treat spatial and temporal autocorrelation in the dataset, we reserve the months of February, May, August, and November as the “test set” for the machine learning method trained here. All other months are used for machine learning model development, with 10% of that data used as a validation set for hyperparameter tuning. All results here are shown for the test set only.

We explore the hyperparameter optimization space using a grid search technique for both the maximum depth and learning rate hyperparameters associated with the XGBoost framework. The maximum tree depth hyperparameter is searched as 2^x, where x ranges in integer steps from 2 to 12, and the learning rate is searched within the range of 0.1 to 0.9 in steps of 0.1. Other fixed hyperparameters associated with the XGBoost framework include a maximum of 1000 boosting iterations and early stopping set to 25 iterations. All other hyperparameters are maintained at package default values (Chen & Guestrin, 2016). Results of the hyperparameter search are summarized in Figure 2. A learning rate of 0.4 and a maximum depth of 14 minimized the classification error on the validation set, and is what we used for the final model trained in this work. We optimize the XGBoost classifier using a multiclass softmax prediction and the package default cross-entropy loss function. The softmax prediction predicts a value between 0 and 1 for each class that can be interpreted as the probability that a given prediction belongs to a specific class. The maximum class probability is taken as the final class prediction.

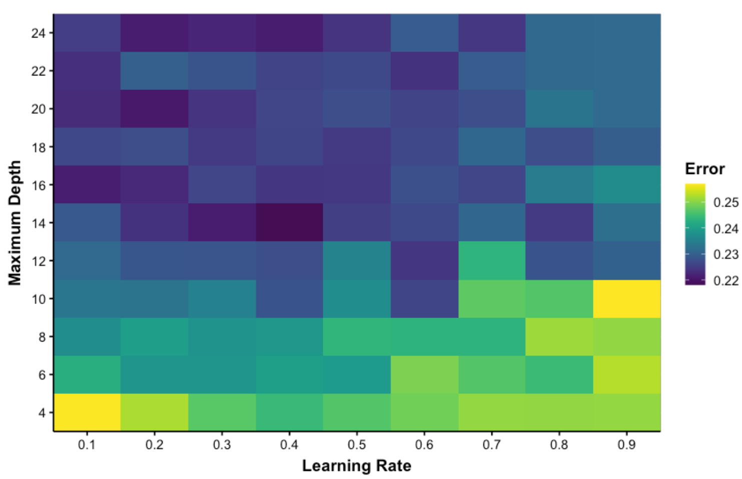

Figure 2. Validation classification error as a function of the learning rate and maximum depth hyperparameters.

The overall machine learning classifier accuracy is 75% across all data available in the test set. This is considerably higher than a random baseline accuracy of 33%. On a per class basis, the true positive rate tends to be higher for classes that are more represented in the dataset at 69%, 79%, and 58% for the underestimation, correct prediction, and overestimation classes, respectively. The prevalence of these classes are 37%, 61%, and 1%, respectively. As a best practice we evaluate the use of XGBoost for this prediction task with a far simpler benchmark that is nonetheless more advanced than a random baseline. Here we use multiple linear regression treating the values as a regression and optimized using ordinary least squares. We find that the linear model has an overall accuracy of 69%, with performance heavily biased on a per-class basis. The linear model per-class accuracy is 43%, 87%, and 0.07% for the underestimation, correct prediction, and overestimation classes, respectively. This poor per-class performance of the linear baseline further motivates the use of a more advanced prediction technique, such as the XGBoost method applied here.

3.3 Error Characterization

We explore the error term learned by the XGBoost classifier through computing the SHAP values for all prediction cases. We first explore the average SHAP values across all predictions, and then investigate individual predictions and their dependence on the input variable distributions. As stated in Section 2, this work assumes that the behavior and interpretation of the skillful XGBoost classifier can provide information on the error in the actual NASA GEOS lightning parameterization.

Figure 3 summarizes the median SHAP value magnitude for all prediction cases and input variables, with the interquartile range across ensemble members shown as the line ranges. For a given prediction and input variable, larger SHAP value magnitudes correspond to a larger contribution from that variable to that prediction. Following this, the average magnitude of the SHAP values across all predictions is commonly interpreted as a metric of variable importance (Molnar, 2019). Variables that have larger associated SHAP values are ranked as more important for the prediction task as they have larger average contributions to the predictions. In the case of lightning occurrence as simulated by the GEOS model, the most important variables for predicting the error of the CTH lightning scheme are the CAPE (Convective Available Potential Energy) for a surface parcel, the convective updraft velocity, the Land/Water/Ice flag, the lifting condensation level, the specific humidity at 600mb, and the convective precipitation. This is consistent with the lightning scheme errors varying with meteorological conditions across the observational domain, and the known importance of convective processes and the land surface type in influencing lightning formation (Murray et al., 2012).

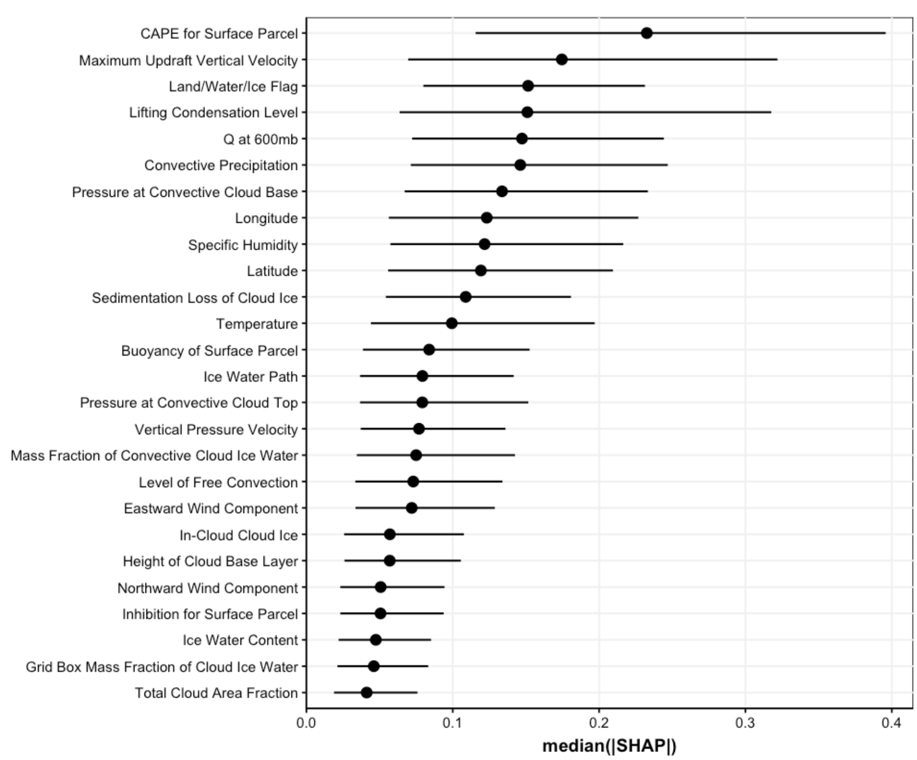

Figure 3. The median SHAP magnitude across all prediction cases, selecting for the predicted class for each case. Line ranges represent the upper and lower quantile (25^{th} to 75^{th} percentile) across the distribution, and points represent the median average absolute SHAP value.

We further investigate the error behavior through comparison of the SHAP value with the original value of the input feature, visualized through so-called “SHAP dependence plots”. This comparison can help illuminate the relationship between the value of an input feature and its contribution to a given prediction case, potentially highlighting model biases in certain input feature regimes. We explore in detail three important variables as identified in Figure 3 as illustrative examples: CAPE for a surface parcel, the maximum updraft vertical velocity, and the convective precipitation.

Figure 4 shows the SHAP dependence plot for CAPE, the highest importance ranked variable in Figure 3. As with the analysis of the median SHAP value magnitudes, values closer to zero in the SHAP dependence plots are indicative of a smaller contribution to the prediction of the error by the XGBoost classifier. Positive values (shown in red) indicate a contribution toward predicting a given class, negative values (shown in blue) indicate a contribution toward not predicting a given class, and values with interquartile ranges that cross zero (shown in black) indicate little contribution to a prediction case. In general, across the three prediction cases, symmetries are common. Regimes that are strongly predictive of one class (e.g. underestimation) are commonly predictive against the other classes (e.g. correct predictions). For CAPE, there are two dominant regimes in the SHAP dependence figure. Very low model CAPE values contribute toward predicting the underestimation class in the dataset, whereas higher values contribute toward predicting away from the underestimation class (e.g. either the correct prediction class or the overestimation class). From the earth system model lightning prediction perspective, lower simulated CAPE values are associated with lightning prediction underestimation, and higher values are likely not associated with driving that underestimation. SHAP values near zero at approximately 1000 J $kg^{-1}$are consistent with a CAPE regime that does not necessarily imply anything about the model behavior. While these regimes can indicate potential drivers of earth system model error behavior, it is important to note that causality cannot be determined through the SHAP analysis presented here.

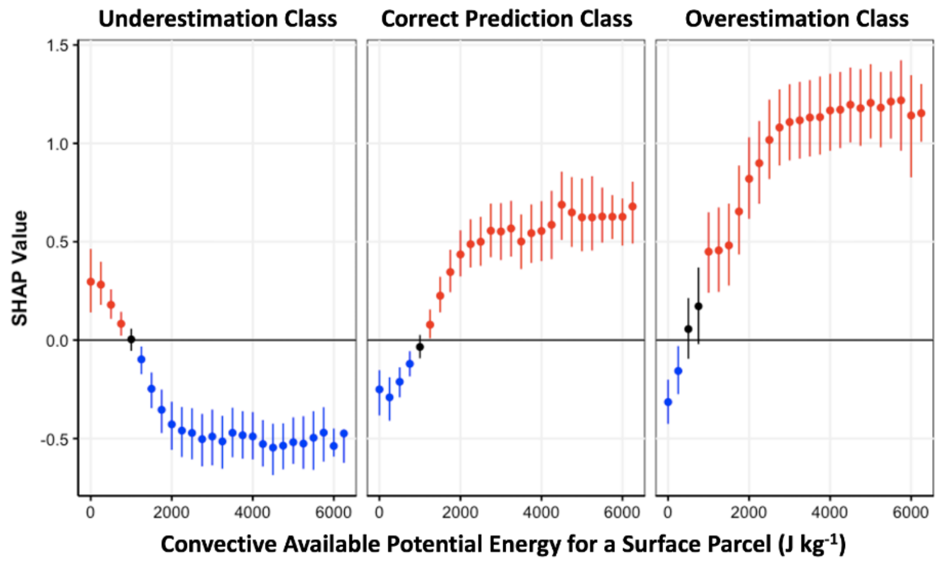

Figure 4. The SHAP dependence plot for convective available potential energy (CAPE) for a surface parcel. Line ranges represent the upper and lower quantile (25th to 75th percentile), and points represent the median SHAP value. Data are binned in 250 J kg^{-1} size bins. Colors represent the sign of the quantile range, positive as red, negative as blue, and crossing zero as black.

The second highest ranked variable is the convective updraft velocity, and the associated SHAP dependence plot is shown in Figure 5. In contrast to the CAPE SHAP dependence figures, convective updraft velocity is treated as nearly a binary variable in the XGBoost classifier. For cases where the convective updraft velocity is identically zero, the variable is a very strong predictor that the model is not underestimating lightning occurrence, where the SHAP value of ~-5.0 is among the largest magnitudes in the entire dataset. Additionally, while convective updraft velocities less than zero contribute very little to the correct prediction class, they strongly reduce the likelihood of predicting the overestimation class.

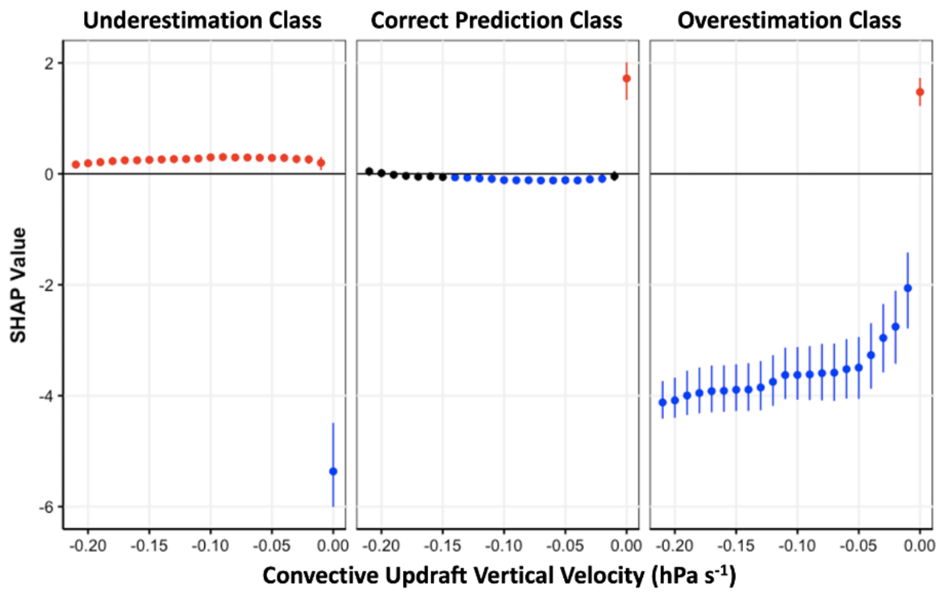

Figure 5. The SHAP dependence plot for the convective updraft velocity. Line ranges represent the upper and lower quantile (25th to 75th percentile), and points represent the median SHAP value. Data are binned in 0.01 hPa s^{-1} size bins. Colors represent the sign of the quantile range, positive as red, negative as blue, and crossing zero as black.

While the general SHAP dependence behavior of both the correct prediction and overestimation prediction classes are similar in Figures 4 and 5, this is not always the case. This is illustrated in Figure 6, which shows the SHAP dependence plot for convective precipitation. In this case, very low convective precipitation values correspond to an increased prediction of the underestimation class, and higher precipitation values lead to a decrease in the prediction of that class. The reverse is true for the correct prediction class, where very low precipitation values lead to a reduction in that predicted value, whereas higher values increase the prediction toward the correct prediction class. For low-mid range convective precipitation values, the overestimation class follows the correct prediction class. At ~2e-4 kg m^{-2} s^{-1}, convective precipitation ceases to contribute toward the overestimation task, and actually trends toward contributing away from predicting that class.

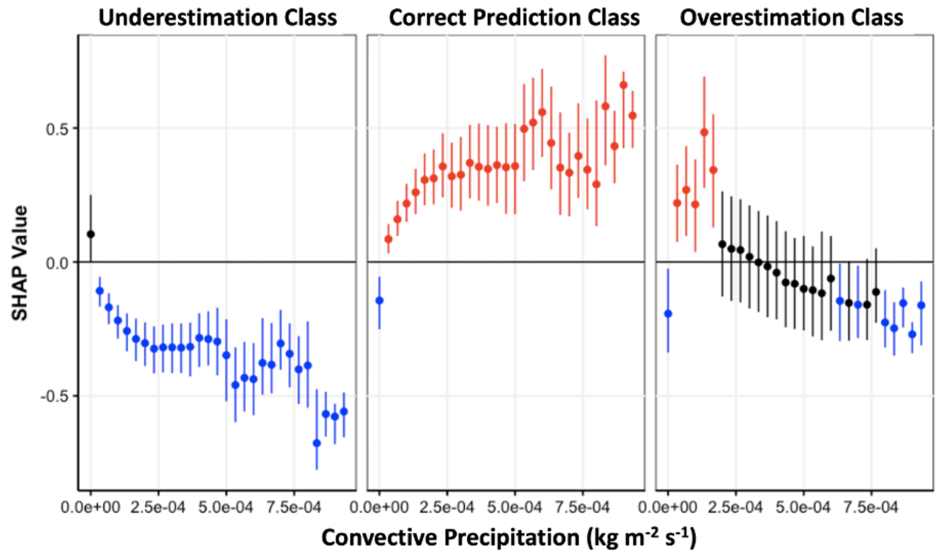

Figure 6. The SHAP dependence plot for the convective precipitation. Line ranges represent the upper and lower quantile (25th to 75th percentile), and points represent the median SHAP value. Data are binned in 3x10^{-4} kg m^{-2} s^{-1} size bins. Colors represent the sign of the quantile range, positive as red, negative as blue, and crossing zero as black.

Taken as a whole, the results from this SHAP dependence analysis highlight several key regimes that can be used to guide interpretation of errors in the NASA GEOS lightning occurrence parameterization. Consistent with previous literature, convective processes have a large influence on predictions of lightning model errors (e.g. Murray et al., 2012). More specifically, model input variable regimes with substantial influences on classification predictions include low CAPE values (below ~1000 J kg^{-1}), and identically zero convective updraft velocity. Additionally, high convective precipitation values are strongly predictive of the correct prediction class, and negatively predictive of both error classes (over- and under-prediction). These results allow for data-driven hypothesis generation regarding improving representations of lightning formation in the NASA GEOS model, in particular that changes in the computational representation of convective processes will likely have a strong influence on the errors in the lightning prediction scheme.

3.4. Input Variable Dependence

As stated previously, one potential disadvantage of the SHAP calculation implementation in the XGBoost library is that the algorithm assumes independence across input features. This assumption is violated in many applications in the Earth system sciences, including the application in this work. This ultimately calls into question the validity of the results presented in this work. To address this, we evaluate the SHAP values calculated from the XGBoost library against an approach that does more directly account for dependent input variables described in Aas et al. (2020).

For machine learning tasks with large datasets with many input features (such as the one used here) the SHAP value calculation approach in Aas et al. (2020) is many orders of magnitude more computationally expensive than the Lundberg et al. (2018) method, and does not currently support multi-label classification. This ultimately makes it impossible to apply the Aas et al. (2020) method to the entire dataset used in this work. However, we can compare the two methods on a smaller representative example problem to get a sense for the potential cost associated with assuming independence across the input variables. We trained a binary classification problem using the top six most important variables as identified through the Lunberg et al. (2019) approach (see section 3.3) on 2575 observations from 9 days of data in June, focusing on classifying a case as either "underestimation" or "correct prediction". The dataset was class balanced and the same hyperparameters were used as in the larger classification task in Section 3.3. The overestimation cases were removed and only constituted ~1% of the data (61 cases). Overall prediction accuracy for this small subset example was ~64%. We then compare the SHAP values calculated from the Aas et al. (2020) method, and the implementation in the XGBoost software package. A comparison of all calculated SHAP values is summarized in Figure 7 below. In general we find a strong correlation between the two calculated SHAP values (R = 0.82), though the absolute magnitudes of the SHAP values differ. This is consistent with different attribution calculation methodologies, but overall similar attribution interpretations.

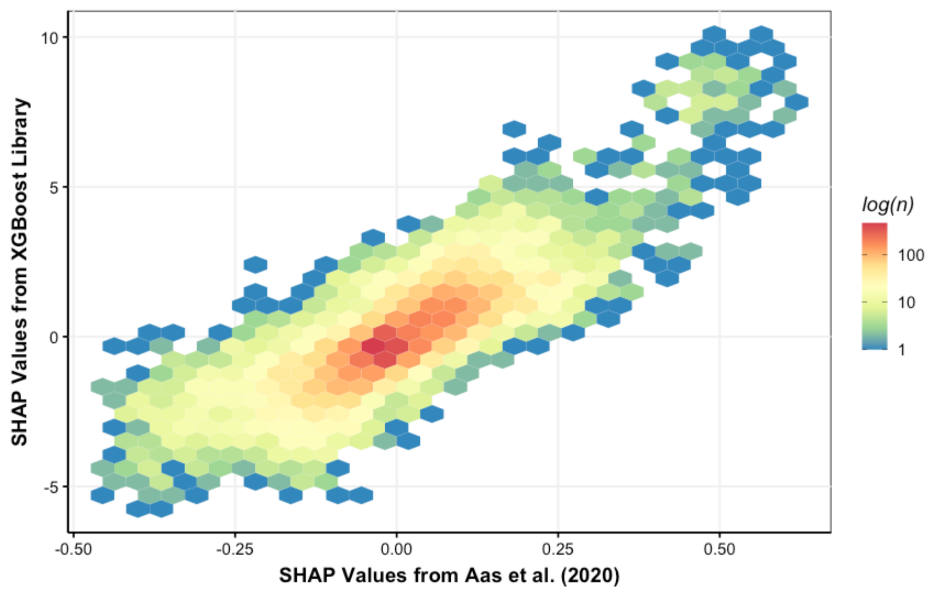

Figure 7. Hexbin comparison of the SHAP Value predictions from the XGBoost library and the Aas et al. (2020) method. Colors represent the log of the number of cases in a given hexagon.

Spearman rank correlations across individual cases (i.e. evaluating if both methods produce the same variable ranking of the six input features) are additionally high, with a median value of 0.77. A histogram of the rank correlations across prediction cases is shown in Figure 8. Additionally, the final median(|SHAP|) comparisons between the two methods show the same final variable rankings. This lends confidence to the application of the SHAP value calculations from the XGBoost library for this use case. It is important to note that a comparison of this sort is likely necessary for all applications of SHAP analysis when input variables are dependent. The quality performance in this work is not a guarantee of algorithm skill for all applications in the Earth Sciences.

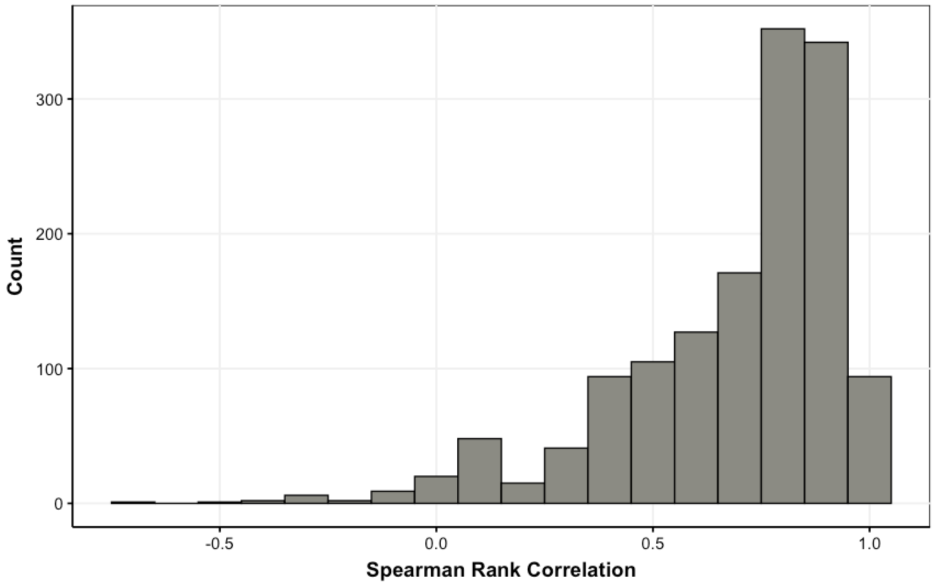

Figure 8. A histogram of the Spearman rank correlation between the prediction case-specific SHAP Value predictions from the XGBoost library and the Aas et al. (2020) method for all test dataset cases on the subset of data described in section 3.4. Histogram bins have a width of 0.1.

4. Summary and Implications for Earth System Model Development

Here we describe an application of interpretable artificial intelligence methods for the characterization of errors in computational models of the Earth System. This application operates in a two-step process, where first the model errors are learned through a widely used machine learning classification technique, followed by the use of SHAP value analysis for interpretability. This ultimately results in a domain agnostic technique for the characterization and exploration of model errors as a function of related parameters.

From an Earth System Model development perspective, we can use this analysis technique to guide efforts toward model improvement and as a data-driven approach to inform hypothesis generation. We demonstrate this approach through investigation of the lightning occurrence parameterization in the NASA GEOS model. The convective available potential energy is on average assigned the most credit for predicting the error in the lightning parameterization, with the highest average magnitude SHAP values. Additional important input variable regimes include very low convective updraft velocities and high convective precipitation values. These results are consistent with issues surrounding capturing convective processes being important drivers of model biases. Other important variables include the local Land/Water/Ice flag, the lifting condensation level, and the specific humidity at 600mb. On aggregate these variable importances are consistent with the importance of convective processes and land-surface heterogeneities in influencing errors in the lightning parameterization. From the results presented here, we can hypothesize that changes to the representation of convective processes in the NASA GEOS model will likely have a substantial impact on the errors in the model prediction of lightning occurrence.

As modern Earth System Models grow in complexity, approaches for the characterization and diagnosis of process-level errors which complement existing efforts are highly valuable.

Techniques from the machine learning and data analytics research literature can be particularly useful in this regard, as they are ideal tools to exploit the massive volumes of data currently generated by modern computational earth system science.

Acknowledgements

The authors declare no conflict of interest. A portion of the research described in this manuscript was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL), a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. SJS is grateful for the support of the Linus Pauling Distinguished Postdoctoral Fellowship program. The Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract DE-AC05-76RL01830. We are grateful to Bill Koshak (NASA Marshall Space Flight Center) for giving us early access to GLM data.

Data Availability

The code used to train and interpret the XGBoost classifier is available here: https://github.com/samjsilva91/lightningSHAP. The code, trained xgboost classifier, and the training/testing dataset is also available here: https://doi.org/10.5281/zenodo.5593568.

References

Aas, K., Jullum, M., & Løland, A. (2020). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *ArXiv:1903.10464 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1903.10464

Allen, D. J. and Pickering, K. E.: Evaluation of lightning flash rate parametrizations for use in a global chemical transport model, J. Geophys. Res., 107, 4711, doi:10.1029/2002JD002066, 2002.

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020). Indicator Patterns of Forced Change Learned by an Artificial Neural Network. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002195. https://doi.org/10.1029/2020MS002195

Batunacun, Wieland, R., Lakes, T., & Nendel, C. (2021). Using Shapley additive explanations to interpret extreme gradient boosting predictions of grassland degradation in Xilingol, China. *Geoscientific Model Development*, *14*(3), 1493–1510. https://doi.org/10.5194/gmd-14-1493-2021

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., et al. (2018). Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. *Agricultural and Forest Meteorology*, *263*, 225–241. https://doi.org/10.1016/j.agrformet.2018.08.019

Finney, D. L. et al. Using cloud ice flux to parametrise large-scale lightning.Atmos. Chem. Phys. 14, 12665–12682 (2014).

Finney, D. L., Doherty, R. M., Wild, O., Stevenson, D. S., MacKenzie, I. A., & Blyth, A. M. (2018). A projected decrease in lightning under climate change. *Nature Climate Change*, *8*(3), 210–213. https://doi.org/10.1038/s41558-018-0072-6

Freitas, S. R., Grell, G. A., Molod, A., Thompson, M. A., Putman, W. M., Santos e Silva, C. M., & Souza, E. P. (2018). Assessing the Grell-Freitas convection parameterization in the NASA GEOS modeling system. *Journal of Advances in Modeling Earth Systems*, 10, 1266–1289. https://doi.org/10.1029/2017MS001251

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., & Zhao, B. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), Journal of Climate, 30(14), 5419-5454.

GOES-R Algorithm Working Group and GOES-R Series Program, (2018): NOAA GOES-R Series Geostationary Lightning Mapper (GLM) Level 2 Lightning Detection: Events, Groups, and Flashes. [indicate subset used]. NOAA National Centers for Environmental Information. doi:10.7289/V5KH0KK6. [access date]. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C01527

Ivatt, P. D., & Evans, M. J. (2020). Improving the prediction of an atmospheric chemistry transport model using gradient-boosted regression trees. *Atmospheric Chemistry and Physics*, *20*(13), 8063–8082. https://doi.org/10.5194/acp-20-8063-2020

Keller, C. A., Knowland, K. E., Duncan, B. N., Liu, J., Anderson, D. C., Das, S., et al. (2021). Description of the NASA GEOS Composition Forecast Modeling System GEOS-CF v1.0. *Journal of Advances in Modeling Earth Systems*, *13*(4), e2020MS002413. https://doi.org/10.1029/2020MS002413

Koshak, W., D. Mach, M. Bateman, P. Armstrong, and K. Virts (2018). GOES-16 GLM level 2 data full validation data quality: Product performance guide for data users. NOAA, 16 pp., https://www.ncdc.noaa.gov/sites/default/files/attachments/GOES16_GLM_FullValidation_ProductPerformanceGuide.pdf.

Liu, M., & Yang, L. (2020). Human-caused fires release more carbon than lightning-caused fires in the conterminous United States. *Environmental Research Letters*, *16*(1), 014013. https://doi.org/10.1088/1748-9326/abcbbc

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. *ArXiv:1802.03888 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1802.03888

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. The American Statistician, 32(1), 12–16. https://doi.org/10.2307/2683468

Meijer, E., van Velthoven, P., Brunner, D., Huntrieser, H., and Kelder, H.: Improvement and evaluation of the parametrisation of nitrogen oxide production by lightning, Phys. Chem. Earth C, 26, 577–583, doi:10.1016/S1464-1917(01)00050-2, 2001.

Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

Molod, A., Takacs, L., Suarez, M., and Bacmeister, J.: Development of the GEOS-5 atmospheric general circulation model: evolution from MERRA to MERRA2, Geosci. Model Dev., 8, 1339–1356, https://doi.org/10.5194/gmd-8-1339-2015, 2015.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 201900654. https://doi.org/10.1073/pnas.1900654116

Murray, L. T., D. J. Jacob, J. A. Logan, R. C. Hudman, and W. J. Koshak (2012), Optimized regional and interannual variability of lightning in a global chemical transport model constrained by LIS/OTD satellite data, J. Geophys. Res., 117, D20307, doi:10.1029/2012JD017934.ß

Murray, L. T. (2016). Lightning NOxand Impacts on Air Quality. *Current Pollution Reports*, *2*(2), 115–133. https://doi.org/10.1007/s40726-016-0031-7

Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, *11*, 169–198. https://doi.org/10.1613/jair.614

Price, C., and D. Rind (1992), A simple lightning parameterization for calculating global lightning distributions, J. Geophys. Res., 97, 9919–9933, doi:10.1029/92JD00719.

Price, C., and D. Rind (1993), What determines the cloud-to-ground lightning fraction in thunderstorms, Geophys. Res. Lett., 20(6), 463–466, doi:10.1029/93GL00226.

Price, C., and D. Rind (1994), Modeling global lightning distributions in a general-circulation model, Mon. Weather Rev., 122(8), 1930–1939, doi:10.1175/1520-0493.

Rasp, S., & Lerch, S. (2018). Neural Networks for Postprocessing Ensemble Weather Forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1

Rasp, S., & Thuerey, N. (2021). Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002405. https://doi.org/10.1029/2020MS002405

Rudlosky, S. D., S. J. Goodman, K. S. Virts, and E. C. Bruning (2019). Initial geostationary lightning mapper observations. Geophys. Res. Lett., 46, 1097–1104, https://doi.org/10.1029/2018GL081052.

Schumann, U. and Huntrieser, H.: The global lightning-induced nitrogen oxidessource, Atmospheric Chemistry and Physics, 7, 3823–3907, https://doi.org/10.5194/acp-7-3823-2007, URL https://www.atmos-chem-phys.net/7/3823/2007/, 2007.

Silva, S. J., Ridley, D. A., & Heald, C. L. (2020). Exploring the Constraints on Simulated Aerosol Sources and Transport Across the North Atlantic With Island-Based Sun Photometers. *Earth and Space Science*, *7*(11), e2020EA001392. https://doi.org/10.1029/2020EA001392

Silva, S. J., Ma, P.-L., Hardin, J. C., & Rothenberg, D. (2020). Physically Regularized Machine Learning Emulators of Aerosol Activation. *Geoscientific Model Development Discussions*, 1–19. https://doi.org/10.5194/gmd-2020-393

Stirnberg, R., Cermak, J., Kotthaus, S., Haeffelin, M., Andersen, H., Fuchs, J., et al. (2021). Meteorology-driven variability of air pollution (PM_1) revealed with explainable machine learning. *Atmospheric Chemistry and Physics*, *21*(5), 3919–3948. https://doi.org/10.5194/acp-21-3919-2021

Štrumbelj, E., Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst* 41, 647–665 (2014)

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. https://doi.org/10.1029/2019MS002002

Williams, E. (1985), Large-scale charge separation in thunderclouds,J. Geophys. Res., 90, 6013–6025, doi:10.1029/JD090iD04p06013.

Xiao, Q., Chang, H. H., Geng, G., & Liu, Y. (2018). An Ensemble Machine-Learning Model To Predict Historical PM2.5 Concentrations in China from Satellite Data. *Environmental Science & Technology*, *52*(22), 13260–13269. https://doi.org/10.1021/acs.est.8b02917

Zhang, C., & Ma, Y. (Eds.). (2012). *Ensemble Machine Learning*. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-9326-7