Journal of Computational and Applied Mathematics 464 (2025) 116479

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Random forest regression feature importance for climate impact pathway detection

Meredith G.L. Brown [a], Matt G. Peterson [a], Irina K. Tezaur [b,*], Kara J .Peterson [a], Diana L. Bull [a]

[a] *Sandia National Laboratories, Albuquerque, NM, USA*
[b] *Sandia National Laboratories, Livermore, CA, USA*

ARTICLE INFO

Dataset link: https://www.sandia.gov/cldera/e3sm-simulations-data/, https://github.com/sandialabs/CLDERA-E3SM

Keywords:
Random forest regression (RFR)
Feature importance
SHapley Additive exPlanation (SHAP)
Mount Pinatubo
Climate impacts
Source-impact pathways

ABSTRACT

Disturbances to the climate system, both natural and anthropogenic, have far reaching impacts that are not always easy to identify or quantify using traditional climate science analyses or causal modeling techniques. In this paper, we develop a novel technique for discovering and ranking the chain of spatio-temporal downstream impacts of a climate source, referred to herein as a source-impact pathway, using Random Forest Regression (RFR) and SHapley Additive exPlanation (SHAP) feature importances. Rather than utilizing RFR for classification or regression tasks (the most common use case for RFR), we propose a fundamentally new workflow in which we: (i) train random forest (RF) regressors on a set of spatio-temporal features of interest, (ii) calculate their pair-wise feature importances using the SHAP weights associated with those features, and (iii) translate these feature importances into a weighted pathway network (i.e., a weighted directed graph), which can be used to trace out and rank interdependencies between climate features and/or modalities. Importantly, while herein we employ RFR and SHAP feature importance in steps (i) and (ii) of our algorithm, our novel workflow is in no way tied to these approaches, which could be replaced with *any* regression and sensitivity method, respectively. We adopt a tiered verification approach to verify our new pathway identification methodology. In this approach, we apply our method to ensembles of data generated by running two increasingly complex benchmarks: (i) a set of synthetic coupled equations, and (ii) a fully coupled simulation of the 1991 eruption of Mount Pinatubo in the Philippines performed using a modified version 2 of the U.S. Department of Energy's Energy Exascale Earth System Model (E3SMv2). We find that our RFR feature importance-based approach can accurately detect known pathways of impact for both test cases.

1. Introduction

The Earth's climate is ever changing, with complex interactions between the atmosphere, solid Earth, hydrosphere, cryosphere and biosphere. Since disturbances, both natural and anthropogenic, in any of these spheres have far ranging impacts on life on Earth, it is critical to understand what we will refer to herein as "source-impact pathways": the relationships and interactions of a set of climate variables in space–time due to an external climate forcing. Methods designed to identify and quantify multiple steps in a pathway by flagging which disturbances arise could greatly enhance our ability to understand high consequence ramifications with legal, political, and national security implications [1].

* Corresponding author.
E-mail address: ikalash@sandia.gov (I.K. Tezaur).

https://doi.org/10.1016/j.cam.2024.116479
Received 26 September 2024; Received in revised form 23 December 2024
Available online 15 January 2025
0377-0427/Published by Elsevier B.V.

The most common method for assessing impacts in the climate system is fingerprinting. In this approach, spatial and/or temporal patterns are established under various disturbances (i.e., greenhouse gases, aerosol loading, etc.) and matched to observations [2,3]. Although the past few years have seen extensions of fingerprinting to regional analyses [4,5], multiple variables [6,7] and challenging problems with very small signal-to-noise ratios [8–11], the method is designed to work within a single step. There has been some recent work to develop conditional multi-step fingerprinting methods [12], but this field is still in its infancy.

While causal modeling is able to identify relationships between multiple variables [13,14], these techniques are valid for spatially-stationary signals. Moreover, the high-dimensionality of climate data often limits the scalability of these methods [13] (although there are recent developments towards spatially resolved causal methods (CaStLe) [15]). Since realistic source-impact pathways between climate variables dynamically evolve, off-the-shelf causal modeling methods cannot be applied to identify these relationships.

Recent years have also seen the emergence of deep learning-based methods for climate attribution, detection and impact analysis [16–20]. These approaches typically train a neural network (NN) on an ensemble of climate data and use the NN to make predictions or as a surrogate in an inverse attribution workflow. The primary downside of NNs is that they typically require massive amounts of data, and can be very costly to train, both in terms of developer time and computational requirements. Importantly, existing approaches do not have the capability to *discover* source-impact pathways from data; instead, they require an analyst to postulate a set of possible relationships, which are then confirmed/denied. Additionally, existing methods typically do not provide information about relative pathway strengths.

This paper presents a novel data-driven methodology for discovering and ranking source-impact pathways using random forest regression (RFR) and feature importance that can be applied to ensembles of simulated and/or observed climate data. RFR [21,22] is an ensemble learning method, typically used for classification or regression, that operates by constructing an ensemble of decision trees at training time, each of which creates a set of if-then-else rules to approximate a dataset or function [23,24]. RFR has a number of advantages over other data-driven methods, including easy offline training, efficiency, interpretability and built-in feature importance metrics (quantitative measures of how much influence a particular input variable has on the model's output prediction [25]). While RFR and feature importance are both well-known in the field of machine learning (ML), we propose herein a first-of-its-kind combination of these two concepts that enables both source-impact pathway identification and ranking, and operates by: (i) training an individual random forest (RF) regressor for each output, (ii) calculating pairwise feature importance values between the inputs and outputs across all RF regressors, and (iii) converting these feature importances into a pathway network, which can be represented as a directed graph. This "outer loop" algorithm for identifying relationships from data is the primary innovation in this work. Indeed, in steps (i) and (ii) of our algorithm, RFR and SHAP could easily be swapped with any regression and sensitivity metric, respectively.

After verifying our methodology on a manufactured problem, we deploy it on our targeted climate exemplar problem: the 1991 volcanic eruption of Mount Pinatubo in the Philippines, a stratospheric aerosol injection (SAI) event, the climate impacts of which have been well documented and studied [26–32]. This eruption, which took place on June 15, 1991, was the second largest volcanic eruption of the twentieth century [33] and injected approximately twenty million tons of sulfur dioxide (SO_2) into the stratosphere, causing surface temperatures to decrease for up to two years after the eruption due to a reduction in shortwave radiation, and stratospheric temperatures to increase due to the greenhouse effect of increased sulfates in the stratosphere. The Mount Pinatubo eruption is ideal for feature and pathway detection because of its well-known impacts governed by well-characterized stratospheric dynamics, and the large signal-to-noise ratio in climate variables such as aerosol optical depth (AOD), stratospheric temperature, surface temperature and others. We utilize as training data an ensemble of simulations of the Mount Pinatubo eruption performed using a modified version of the fully-coupled E3SM version 2 (E3SMv2) [34] known as E3SMv2-Stratospheric Prognostic Aerosols (E3SMv2-SPA) [35], augmented to simulate prognostically the evolution of aerosols in the stratosphere. We note that several alternative methods for identifying/detecting pathways within this data set have been proposed during the past 1–2 years, each with its own strengths and weaknesses. These methods are based on techniques such as changepoint detection [36,37], echo state networks (ESNs) [38–40], space–time dynamic models [41], conditional multi-step fingerprinting [12], operator neural network [19,20], and in-situ profiling [42,43].

The remainder of this paper is organized as follows. In Section 2, we describe our new RFR and feature importance-based method for source-impact pathway detection. In Section 3, we describe the two test cases used to develop/verify our method: (i) a set of synthetic coupled equations, and (ii) and the actual Mount Pinatubo eruption itself, simulated in E3SMv2-SPA. Results for each of these two benchmarks are presented in Section 4. In Section 5, we provide conclusions and describe possible directions for future work.

2. Method

The goal of our method is to use time-series data to create a directed graph comprised of nodes and edges, in which the nodes represent features of interest and the edges represent relationships between these features. These relationships are directional, weighted and have a time lag associated with them. Once a graph is fully constructed, it can be used to trace out source-impact pathways, defined as the interactions of a set of variables in space–time due to an external forcing, within the provided time-series data (in our case, climate data).

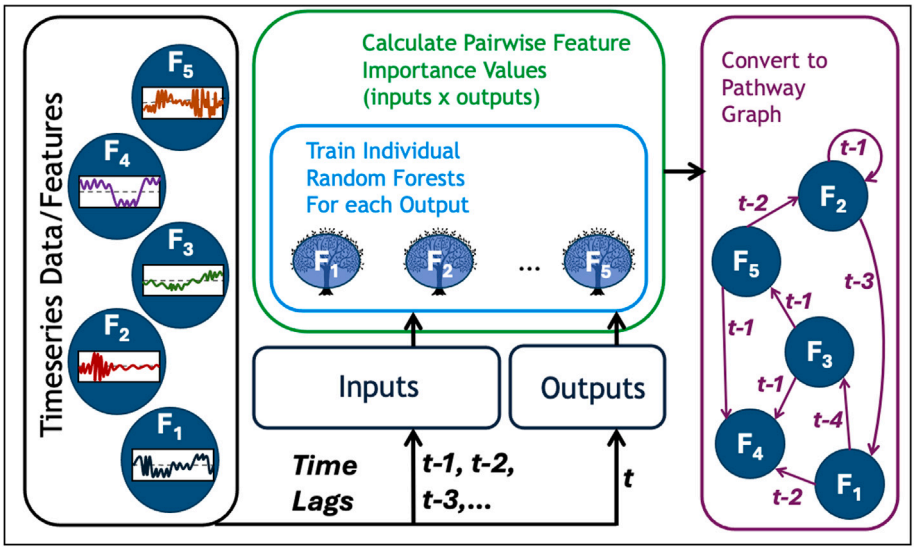

Fig. 1. Visual depiction of the RFR-based pathway construction approach described in Section 2 assuming $\boldsymbol{F} = \tilde{\boldsymbol{F}}$.

Suppose we are given a set of time-series data $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \dots \boldsymbol{F}_n\}$, where $\boldsymbol{F}_i \in \mathbb{R}^K$ and $n, K \in \mathbb{N}$. In this notation[1], the jth entry of $\boldsymbol{F}_i$, denoted by $F_i(j)$, represents the value of $\boldsymbol{F}_i$ at time j for $1 \le j \le K$. From this point forward, we will refer to $\{\boldsymbol{F}_i\}$ as the set of "features" used in our analysis. In the climate application considered herein, each feature $\boldsymbol{F}_i$ is a time-series of a variable associated with a given spatial location and outputted by a climate model, e.g., temperature, aerosol optical depth (AOD), etc. As a concrete example, suppose we are given time-series data for two variables, temperature and AOD, at five spatial locations. In this case, our features $\boldsymbol{F}_1, \dots, \boldsymbol{F}_5$ and $\boldsymbol{F}_6, \dots, \boldsymbol{F}_{10}$ represent the temperature and AOD time-series, respectively, at locations $1, \dots, 5$, so that $n = 10$. In general, if we are given m variables at N spatial locations, for $m, N \in \mathbb{N}$, the total number of features will be $n = mN$.

Our pathway detection algorithm consists of several key steps, described in detail below and summarized succinctly in Algorithm 1 and Figs. 1–2. Although the discussion herein is focused on RFR and SHAP feature importance as the primary workhorses in Algorithm 1, we emphasize that this algorithm is not tied to these approaches.

2.1. Temporal lag selection and data pre-processing

The first step in our approach is to choose a set of temporal lags $L := \{l_1, \dots, l_q\}$ (for $l_i, q \in \mathbb{N}$, where $1 \le i \le q$) to investigate. To give a concrete example, if we choose L to be $L = \{1, 3, 5\}$, the features to be investigated are $F_i(t-1)$, $F_i(t-3)$ and $F_i(t-5)$ for $1 < i < n$ and $2 < t \le K$ with $t \in \mathbb{N}$. Ideally, the choice of lags L should be informed by the dynamics of the problem and the physical processes of interest.

In developing our method, we found that it is often important to pre-process the data contained in $\boldsymbol{F} := \{\boldsymbol{F}_1, \boldsymbol{F}_2, \dots, \boldsymbol{F}_n\}$. The most common pre-processing to consider is a spatial dimension reduction of the features in $\boldsymbol{F}$. Suppose the data comprising $\boldsymbol{F}$ are given at a large number of spatial locations, so that $N \gg 1$ and hence $n \gg 1$. The simplest spatial dimension reduction approach to apply is a basic averaging of the data in space, achieved by first decomposing the spatial domain into a set of $\tilde{N} \ll N$ subdomains, e.g., by splitting the grid into latitudinal bands or Intergovernmental Panel on Climate Change (IPCC) regions [44], then computing a mean of the data in $\boldsymbol{F}$ over each subdomain. The result is a new set of data $\tilde{\boldsymbol{F}} := \{\tilde{\boldsymbol{F}}_1, \dots, \tilde{\boldsymbol{F}}_{\tilde{n}}\}$, where $\tilde{n} = \tilde{N}m$, and $\tilde{\boldsymbol{F}}_i \in \mathbb{R}^K$. We note that other pre-processing approaches can be beneficial to our analysis, e.g., Feature normalization; these are problem specific, and discussed later, in Section 4. In Algorithm 1, we use the notation $preProc(\boldsymbol{F})$ to denote any pre-processing/dimension reduction performed on the data $\boldsymbol{F}$ prior to applying RFR.

2.2. Training of individual RF regressors to predict a given feature of interest

RFR [21,22] is a supervised, data-driven technique that utilizes an ensemble of decision trees for regression. Each decision tree is trained on a subset of the training data in order to reduce overfitting; this is commonly referred to as bagging or bootstrap aggregation. Output values predicted by the decision trees are combined within RFR using a majority voting.

The second step in our workflow is to train individual RF regressors to predict feature i, $\tilde{\boldsymbol{F}}_i$ (the output), with the set of all features at all times, $\tilde{\boldsymbol{F}}$, as inputs. To do this, for each output $\tilde{\boldsymbol{F}}_i$ with $1 \le i \le \tilde{n}$, we create an input vector by concatenating all

[1] We note that our workflow allows for the $\boldsymbol{F}_i$ to have different lengths K_i, but assume that these vectors have the same length to facilitate the presentation of the method.

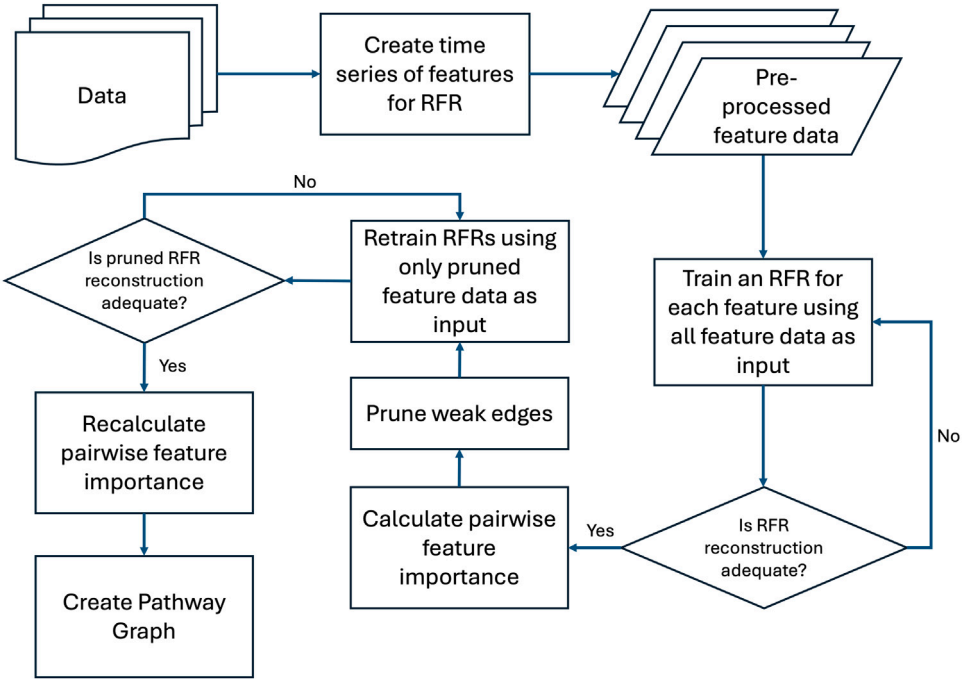

Fig. 2. Schematic of RFR and Feature importance-based source-impact pathway construction method described in Algorithm 1.

possible features at all possible times $t - l_j$ for all possible lags $l_j \in L$, and train a RF regressor to predict $\tilde{F}_i$, as described in Algorithm 1. In Algorithm 1, the subroutine $trainRFR(Inputs, Outputs)$ refers to the RF training step of our workflow. In our current implementation, we perform this step by utilizing Python's sci-kit learn package [45]. The number of decision trees and decision tree depth are controlled parameters in the RFR and hence input parameters for sci-kit learn. For the results presented herein, we used 100 trees, each with a depth of four.

2.3. Evaluation of RF regressors' goodness of fit

Once our RF regressors are trained, the next step is to assess their goodness of fit using appropriate evaluation metrics. Here, we rely on two commonly-used evaluation metrics: R^2_{adj} (the adjusted coefficient of determination) and $RMSE$ (the root mean square error), defined as

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-p-1}\right] \tag{1}$$

where

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

and

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}. \tag{3}$$

In Eqs. (1)–(3), y_i is a reference data point (usually a ground truth measurement, in our case, it is either the synthetic or climate data we are training to), $\hat{y}_i$ is the corresponding RFR prediction, $\bar{y}$ is the mean of the reference data, n is the number of samples, and p is the number of independent variables. We prefer in general R^2_{adj} over R^2, the coefficient of determination, as R^2_{adj} tends to be less sensitive to the number of inputs. We generally look for $R^2_{adj} > 0.75$ and an $RMSE$ on the order of at most 0.15. If these tolerances are not met, it may be necessary to redo the training step of the workflow with different data, pre-processing and/or lags.

2.4. Feature importance calculation

Feature importances (also referred to herein as "weights" and denoted by ω_i) are obtained for machine learning models like RFR via a post-processing, and are computed herein using Python's scikit-learn package [45]. The relative magnitudes of these feature

importances are representative of the relative strength of the relationship between any two features. In the present analysis, we utilize the SHapley Additive exPlanation (SHAP) method to calculate the feature importances/weights between each input feature and output feature $(\tilde{F}_i(t), \tilde{F}_j(t-l))$ [46]. This approach draws on Shapley values and cooperative game theory [47], which aims to evaluate how important any given player is to winning a cooperative game. We note that SHAP is just one possible feature importance measure; other commonly used feature importances include Gini [48,49], permutation [50], amongst others. We select the SHAP approach, as several recent studies have shown that SHAP values can achieve better results for assessing feature importance in machine learning models than other metrics for various problems of interest [51–54]. Additionally and importantly, SHAP proved to be a computationally efficient metric for our analysis.

2.5. Extension to data ensembles

Our algorithm description thus far has assumed that the input data consist of a single set of features $\boldsymbol{F}$. However, since climate analyses typically operate on ensembles of data, generated, for example by slightly perturbing the initial condition, there is a need to have a way of applying our workflow to ensembles of features $\{\{\boldsymbol{F}\}^r\}_{r=1}^{R}$ for $R \in \mathbb{N}$. The multiple ensemble member case is handled by repeating the RFR training process described above and in Algorithm 1 for each set of ensemble members $\{\boldsymbol{F}\}^r$ and averaging the SHAP feature importance weights across the ensemble members for each pairwise input to output. In the case ensembles are used, all goodness of fit metrics (Section 2.3) are calculated on a per ensemble basis.

2.6. Pruning of edges and construction of directed pathway graph

The final step in our procedure involves constructing a directed pathways graph from the RF regressors and feature importance information. For each pair of features $\tilde{\boldsymbol{F}}_i$ and $\tilde{\boldsymbol{F}}_j$, we draw a set of edges. The weight of each edge is given by the value of the feature importance. After calculating the mean and standard deviation of the SHAP weights across a set of ensembles (Section 2.5), we use these values to prune potential edges in our pathway graph using the following criteria:

1. We select the top four incoming edges, i.e., the four largest w_i, to each feature.
2. We prune edges with w_i that does not exceed a minimum threshold δ. In all of our numerical results (Section 4), we used a threshold of $\delta = 1.0 \times 10^{-4}$.
3. We prune edges where the standard deviation (σ) of w_i exceeds the mean weight.
4. We prune edges that are not represented in a majority of ensemble members (a discussion of how we treat an ensemble of data is given in Section 2.5).

The first two criteria enumerated above are data specific, and may need to be refined on a dataset by dataset basis.

Recall that, in our feature importance-based workflow, each RF regressor predicts one variable using all features as inputs, and one SHAP value is assigned for each incoming edge. As described in Section 2.2, our feature importance-based analysis is repeated many times, iterating through all variables of interest. While Step 1 in the procedure outlined above discusses only incoming edges in our directed pathway graph, the nodes of this graph will have both incoming and outgoing edges. The outgoing edges and their associated SHAP values can then be inferred from the incoming edges, as they were incoming edges for alternate nodes in the graph. When we combine all these edges from all the individual calculations, we obtain a graph whose nodes have both incoming and outgoing edges.

Our pruning methodology, which effectively sets all the pruned weights to $w_i = 0$, is referred to as $prune(Edges)$ in Algorithm 1. Once pruning has been completed, we again use the goodness of fit metrics described in Section 2.3 to evaluate the RFR reconstructions of the features. If reconstruction is poor, refinement of the pruning criteria may be needed.

Once the pruned RFR reconstruction has been determined to be a good fit to the original data, we translate our feature importance values w_i into a directed graph, denoted by the $createPathwaysGraph(Nodes, Edges)$ routine in Algorithm 1. Here, each node represents a feature $\tilde{\boldsymbol{F}}_i$, and each edge between node $\tilde{\boldsymbol{F}}_i$ and node $\tilde{\boldsymbol{F}}_j$ is weighted by the feature importance w_i corresponding to the relevant pair provided $w_i > 0$ (see Fig. 1, right column). To make graph depiction more manageable, temporal lags at which relationships between features have been detected are represented by numbers that are printed over each edge (see Fig. 1, right column). Once a pathway graph is fully constructed, pathways can be inferred by following the directed edges in the graph from one feature to another.

3. Data

Since we are proposing a novel data-driven approach for creating pathway graphs, it is important to take a systematic approach to method verification and validation. Herein, we adopt a tiered verification approach that applies Algorithm 1 to two increasingly-complex benchmarks, described in more detail below. First, we develop a system of synthetic coupled equations, where the relationships between the variables are explicitly given and can be controlled to test potential corner cases. Next, we apply our method to data obtained by simulating the 1991 Mount Pinatubo eruption in the Philippines using the fully coupled E3SMv2-SPA [35] (described in more detail below). While some variable relationships and pathways in the second benchmark are well-understood, making it a good verification test case, the underlying physical processes are very complex, leaving room for our model to discover relationships beyond those that are expected.

Algorithm 1 Pseudocode to create pathway graph from a single realization of spatio-temporal data. The algorithm is depicted visually as a workflow in Fig. 1. If an "Abort" is hit, it is necessary to rerun the algorithm with different features, pre-processing, temporal lags and/or pruning criteria.

Require: Set of spatio-temporal features $\boldsymbol{F} \in \mathbb{R}^{K \times n}$. ▷ Input time-series data
Specify temporal lags $L = \{l_1, ..., l_q\}$. ▷ Select temporal lags to be investigated (Section 2.1).
$\tilde{\boldsymbol{F}} = preProc(\boldsymbol{F}) \in \mathbb{R}^{K \times \tilde{n}}$. ▷ Apply pre-processing/dimension reduction to $\boldsymbol{F}$ (Section 2.1)
$Edges \leftarrow \{\}$, $Nodes \leftarrow \{\tilde{\boldsymbol{F}}_1, \tilde{\boldsymbol{F}}_2, ..., \tilde{\boldsymbol{F}}_n, \}$. ▷ Initialize Edges and Nodes arrays
for i in $\{1, ..., \tilde{n}\}$ **do**
  $Outputs = \{\tilde{\boldsymbol{F}}_i\}$, $Inputs = \{\}$. ▷ Create a set of Outputs and Inputs, and populate
  for j in $\{1, ..., \tilde{n}\}$ **do**
    for k in $\{q+1, ..., K\}$ **do**
      for l in $\{1, ..., q\}$ **do**
        Append $F_j(k-l)$ to $Inputs$.
  $RFR^i \leftarrow trainRFR(Inputs, Outputs)$ ▷ Train RFR to predict $\tilde{\boldsymbol{F}}_i$ (Section 2.2)
  Calculate R^2_{adj} and $RMSE$. ▷ Calculate goodness of fit metrics (Section 2.3)
  if R^2_{adj} sufficiently large and $RMSE$ sufficiently small **then**
    $Edges^i \leftarrow featureImportance(RFR^i)$ ▷ Find edge weights going into $\tilde{\boldsymbol{F}}_i$ (Section 2.4)
    Concatenate $Edges^i$ to $Edges$
  $prune(Edges)$ ▷ Prune out weak or irrelevant edges (Section 2.6)
  Calculate R^2_{adj} and $RMSE$. ▷ Calculate goodness of fit metrics (Section 2.3)
  if R^2_{adj} sufficiently large and $RMSE$ sufficiently small **then**
    $Graph = createPathwayGraph(Nodes, Edges)$ ▷ See Section 2.6
  else
    Abort.
  Abort.

3.1. Synthetic coupled equations

For our first benchmark, we consider a five member ensemble of data generated by simulating the following analytically-given time-series of coupled equations:

$$\begin{aligned} W_t &= 0.9W_{t-1} + \epsilon_{W_t} \\ X_t &= 0.8X_{t-1} + 0.5W_{t-1} + \epsilon_{X_t} \\ Y_t &= -0.9W_{t-1} + \epsilon_{Y_t} \\ Z_t &= 0.6X_{t-1} + 0.5Y_{t-1} + \epsilon_{Z_t}. \end{aligned} \tag{4}$$

In Equation set (4), W_t, X_t, Y_t and Z_t are the primary variables/features evaluated at time t, and ϵ_{W_t}, ϵ_{X_t}, ϵ_{Y_t} and ϵ_{Z_t} are random variables representing background noise, herein sampled separately from a uniform $[-0.5, 0.5]$ distribution. We use (4) to create a five member ensemble, each containing 750 time-values for each variable. Each time-series is initialized with a random seed based on the noise term for the variable. Different ensemble members are generated by varying the noise terms ϵ_{W_t}, ϵ_{X_t}, ϵ_{Y_t} and ϵ_{Z_t}. Since this synthetic dataset has known, clearly-encoded variable interactions, it enables us to test our method's ability to identify these known relationships before proceeding to more sophisticated test cases. Specifically, we expect the variable W_t to be most strongly dependent on W_{t-1} (defining a pathway from W_{t-1} to W_t), the variable X_t to be most strongly dependent on X_{t-1} followed by W_{t-1} (defining pathways from X_{t-1} and W_{t-1} to X_t), etc. It is noted that, since the synthetic Eqs. (4) only contain a temporal dependence, it is not necessary to perform a spatial dimension reduction (Section 2.1) prior to applying our method.

3.2. Mount Pinatubo eruption simulated using E3SMv2-SPA

Our second benchmark involves analyses on data from simulations of the 1991 eruption of Mount Pinatubo using a modified version of E3SMv2. Typically, ESMs such as E3SMv2 prescribe the location, absorption, and scattering of stratospheric aerosol from explosive volcanic eruptions (i.e., take them from an input file), rather than simulating them explicitly within the model. In order to study the impacts of the volcanic eruption using the fully-coupled E3SMv2, we extended this model to handle stratospheric aerosols prognostically, that is, to enable us to simulate the dynamic evolution of volcanic sulfate from an injection of SO_2, together with downstream climate impacts. We will refer to the resulting fully-coupled version of E3SMv2, described in more detail in [35], as E3SMv2 Stratospheric Prognostic Aerosol, or E3SMv2-SPA[2].

[2] Available at: https://github.com/sandialabs/CLDERA-E3SM.

Table 1
Mount Pinatubo temperature pathway variables.

Variable name	Long name	Units	Description
AEROD_v	Aerosol Optical Depth	–	Column-integrated aerosol optical depth (missing data in polar winter)
FLNTC	Clear-sky net longwave radiative flux at the top of atmosphere	W m^{-2}	Dominant radiative flux in the Stratosphere
T050	Temperature at 50 hPa	K	Stratospheric temperature
FSDSC	Clear-sky downward Shortwave radiative flux at the surface	W m^{-2}	Dominant radiative flux at the surface
TREFHT	Temperature at 2 m	K	Near-surface air temperature

In the stratosphere, we expect to find a pathway in which the formulation of sulfates increases aerosol optical depth (AEROD_v), which, in turn, increases longwave radiation absorption (FLNT) and leads to an increase in the stratospheric temperature at 50 hPa (T050) [55]. At the Earth's surface, we still expect to see an increase in aerosol optical depth (AEROD_v) due to the presence of sulfates; however, as a consequence, we expect to see a decrease in the amount of shortwave radiation reaching the surface (FSDS), followed by a lowering of the temperature at the surface (TREFHT) [28,56]. It is well-known that the aerosol cloud from Mount Pinatubo's June 15, 1991 eruption encircled the globe in just 22 days and filled the entire tropical belt in approximately two months, before spreading to higher latitudes [57]. This result is confirmed by our simulations of the Mount Pinatubo eruption using E3SMv2-SPA. Fig. 3 shows the stratospheric sulfate burden 1, 6, 11, 16, 21 and 31 days after the eruption. The location of Mount Pinatubo is marked with a red triangle. The global surface temperature was reduced by −0.5 °C by September 1992 [58].

In our analysis herein, we focus our attention on five variables, summarized in Table 1. The reader can observe that we have chosen to utilize the clear-sky versions of the longwave and shortwave radiation, denoted by FLNTC and FSDSC, respectively. We choose these variables because they provide a clearer signal than FLNT and FSDS.

Our analysis was based on five limited variability (LV) simulations, described in more detail in [59] and summarized briefly below. Limited variability was achieved by matching historical conditions for the Quasi-Biennial Oscillation (an alternating equatorial zonal wind pattern in the stratosphere) and the El Niño Southern Oscillation (changes in tropical Pacific sea surface temperatures). These two major modes of natural variability respectively precondition the direction of travel of the injected SO_2 gas as well as temperature and precipitation values globally [60]. The five ensemble members were generated by perturbing the temperature initial conditions on June 1, 1991 with the eruption introducing a point injection of 10 Tg SO_2 at 15.15°N and 120.35°E at an altitude of 18–20 km occurring on June 15, 1991. Simulations ran through December 31, 1998. For comparison purposes, we also had at our disposal a paired five member set of "counterfactual" ensembles, containing no volcanic eruption but using the same initial conditions as the ensembles containing the eruption. Simulations were performed using a 1° configuration of E3SMv2-SPA, which has a resolution of 110 km with 72 vertical layers for the atmosphere, and a resolution of 165 km for land. Our RF regressors were trained with daily averaged climate data. The details of how these data were pre-processed are given later, in Section 4.

4. Results

In the following sections, we present results for each tier in our tiered verification strategy. We start with the synthetic data example, where the features are time-series from the system of coupled equations given in (4) (Section 4.1). We then apply our method to an ensemble of Mount Pinatubo simulations performed using the fully-coupled E3SMv2-SPA [35] (Section 4.2).

4.1. Synthetic coupled equations

Since the coupled synthetic Eqs. (4) are relatively simple and well-behaved, and do not have a spatial dependence, no pre-processing was performed on these data. The time lags used in our analysis were $L = \{1, 2, 3, 4, 5\}$. While the "ground truth" only contains lags of 1 time step, we wanted to verify that the algorithm could identify the "correct" time lag in the presence of extra time lags.

After pre-processing the data, we begin by evaluating our RF regressors' ability to reconstruct each feature of interest, which, for the synthetic coupled equation, is each variable W, X, Y and Z. In Fig. 4, we show the comparison between the time-series of the system of synthetic equations and the RFR reconstruction for each feature. Fig. 4 subfigures (a)-(d) show the ensemble mean time series of the data, with the standard deviation shown in the shaded area. Subfigures (e)-(h) show the scatterplots between the system of equation (SOE) generated data on the y-axis and the RFR reconstructions on the x-axis, with each ensemble shown in a different color. The reader can observe that the RF regressors capture the average signals well, and that the less noisy variables (X) are better captured than the more noisy variables (W, Y, Z). These results are confirmed by Table 2, which shows the corresponding average coefficient of determination (R^2_{adj}) for each feature across all the ensemble members. The RF regressors are able to fit all the synthetic features well ($R^2_{adj} > 0.75$ and $RMSE < 0.15$).

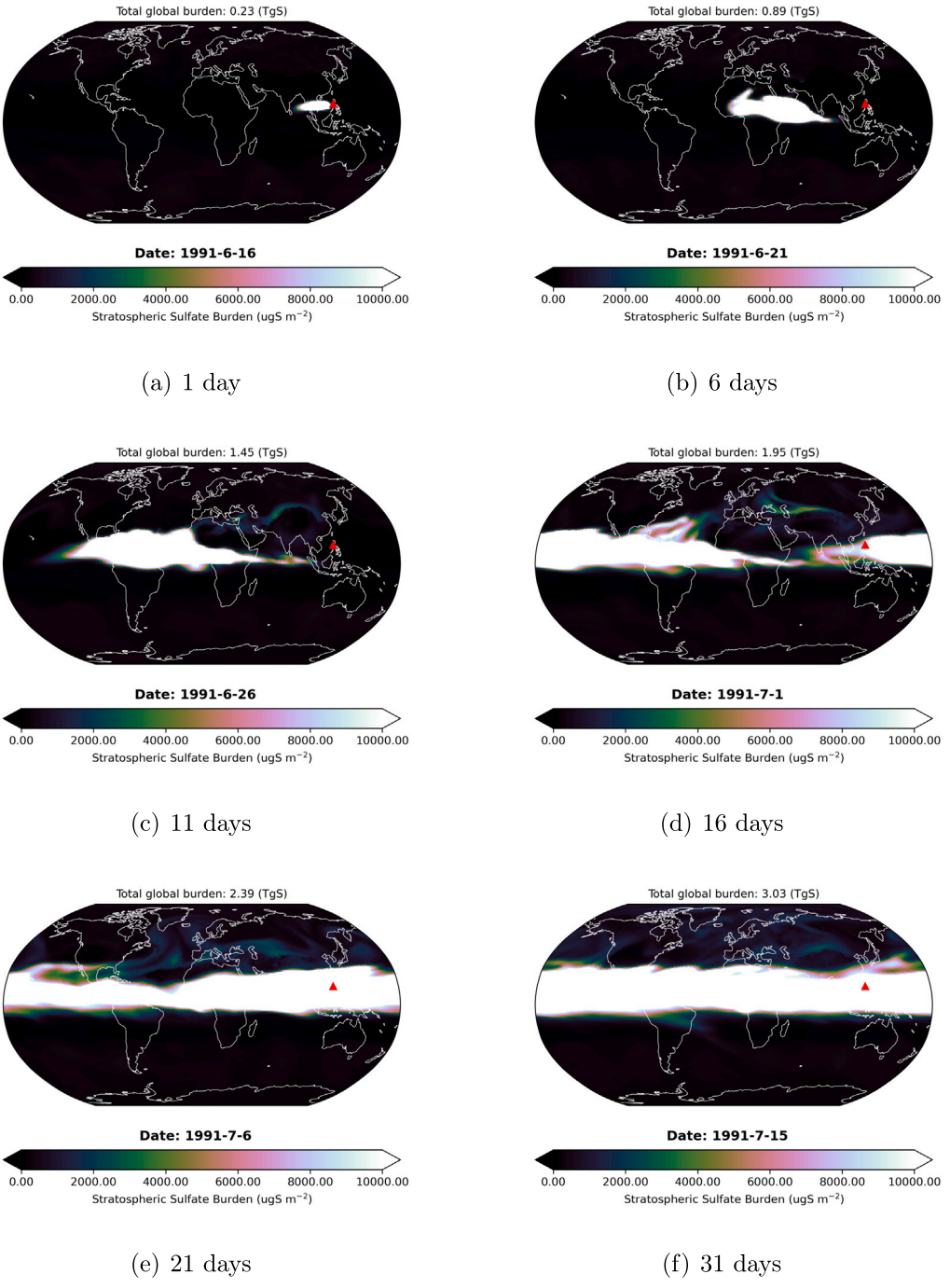

Fig. 3. The stratospheric sulfate burden at different days up to roughly one month following the Mount Pinatubo eruption. The location of Mount Pinatubo is marked with a red triangle. The sulfates have encircled the Earth by approximately 21 days post eruption (e), and have made their way into the subtropics by approximately 31 days post eruption (f).

Next, we calculate and report the feature importances for the coupled synthetic data set, which are summarized in Table 3, sorted from highest to lowest feature importance (weight). In this table, the variables in the "Target" column are the variables being predicted, W_t, X_t, Y_t and Z_t, and the variables in the "Source" column are the dependent variables W_{t-l}, X_{t-l}, Y_{t-l} and Z_{t-l}, where the value of l is given by "Lag" variable. The reader can observe that the ordering of variable dependencies in Table 3 is consistent with the known relationship encoded in the governing Eqs. (4), e.g., the variable X depends on itself and W at the previous time-step, the variable Z depends on the variables X and Y at the previous time-step, etc. It is important to recognize that

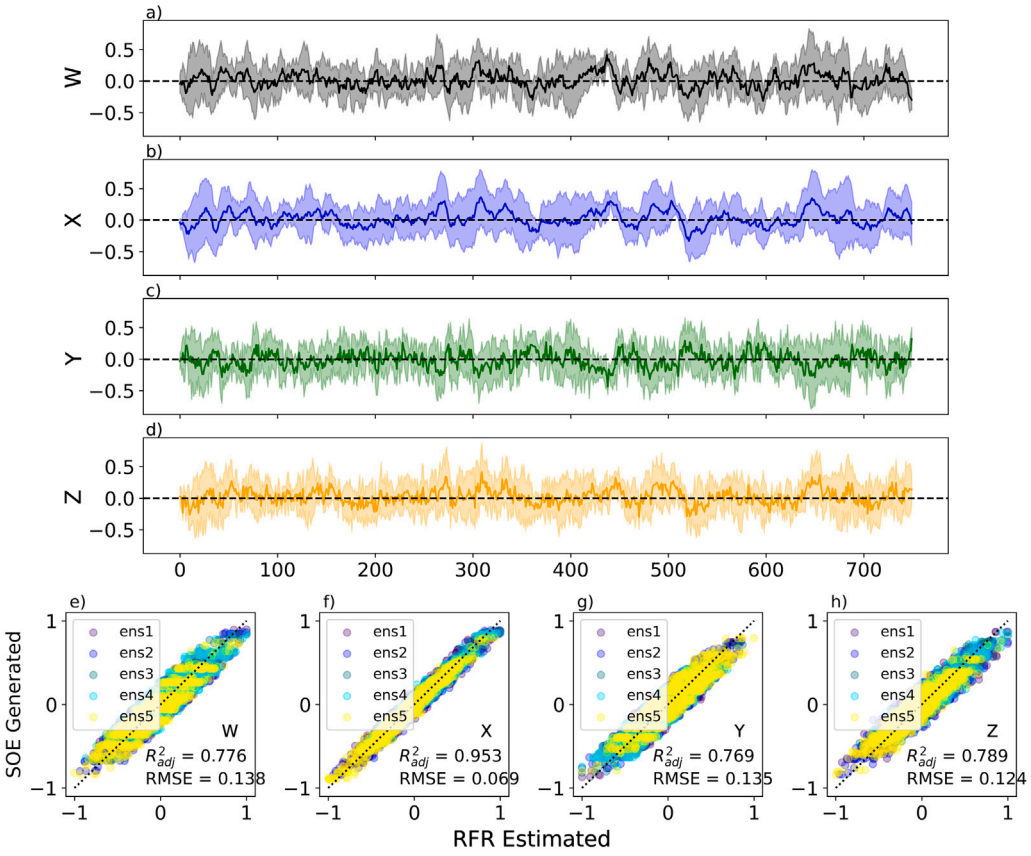

Fig. 4. Coupled synthetic equations: ensemble mean goodness of fit statistics. Subfigures (a)-(d) show the ensemble mean and standard deviation of W, X, Y, and Z respectively. Subfigures (e)-(h) show the scatterplot comparison between the SOE generated time series and their RFR reconstructions, for W, X, Y, and Z respectively.

Table 2
Coupled synthetic equations: goodness of fit statistics of the RFR reconstruction.

Variable	R^2_{adj} mean	R^2_{adj} σ	$RMSE$ mean	$RMSE$ σ
W	0.776	0.043	0.138	0.009
X	0.953	0.010	0.069	0.007
Y	0.769	0.036	0.135	0.007
Z	0.789	0.040	0.124	0.012

our RFR and feature importance approach cannot recover the actual coefficients pre-multiplying each term on the right-hand side of (4); however, the approach *is* able to detect correctly the relative strength of dependence for each variable and each time lag. Fig. 5 is a conversion of Table 3 into its corresponding pathway directed graph form. No edge pruning was performed to generate this graph. Each feature is shown in a labeled circle with arrows indicating the strength and the direction of the feature connections from source to target feature. The numbers labeling the connection arrows denote the time lag associated with the connection.

4.2. Mount Pinatubo eruption

For our Mount Pinatubo exemplar, we worked with average daily data for each variable of interest (defined earlier in Table 1). The Mount Pinatubo simulations were 2770 days long (1 June 1991 to 31 December 1998) and used time lags of $L = \{1, 6, 11, 16, \ldots, 61\}$. Since the Mount Pinatubo data are defined on climate grids having a 1° resolution spatial grid, spatial dimension reduction was needed to make our workflow and subsequent analysis/interpretation tractable. Towards this effect, for each variable of interest from each ensemble member, we reduced the spatial dimension by taking a regional average. We considered two possible regional averages: (i) an average over the entire globe, and (ii) a regional average over seven latitudinal (or zonal)

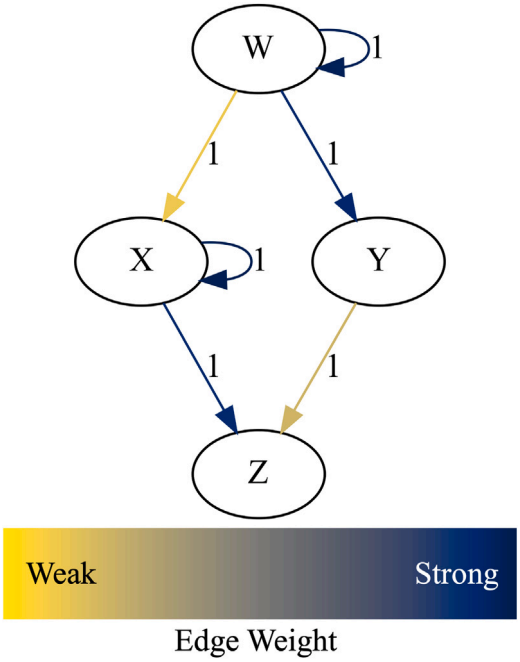

Fig. 5. Coupled synthetic equations: pathway graph. Features are shown in circles, feature connections are shown by the arrows pointing from source feature to target feature. Numbers next to the arrows are the time lags associated with the connection. The edge colors represent SHAP weights, blue indicating higher values and yellow indicating lower values.

Table 3
Coupled synthetic equations: pathway edge weights and their standard deviations.

Source	Target	Lag (time step)	SHAP Weight	Weight σ
X	X	1	0.243	0.014
X	Z	1	0.225	0.014
W	Y	1	0.229	0.010
W	W	1	0.240	0.016
Y	Z	1	0.051	0.009
W	X	1	0.027	0.004

Table 4
Latitudinal zones.

Zone	Extent
Polar North	66.5°N–90°N
Temperate North	35°N–66.5°N
Subtropical North	23.5°N–35°N
Tropical	23.5°S–23.5°N
Subtropical South	35°S–23.5°S
Temperate South	66.5°S–35°S
Polar South	90°S–66.5°S

bands, summarized in Table 4. We additionally performed some normalization of the input variables to get all the variables on a similar scale. Specifically, we subtracted from each ensemble member its corresponding paired counterfactual ensemble member, so as to isolate impacts from the eruption, and normalized the resulting differenced time-series to be between -1 and 1.

We will consider the two known pathways of interest described in Section 3.2, (i) the stratospheric warming pathway (AEROD_v → FLNTC → T050), and (ii) the surface cooling pathway (AEROD_v → FSDSC → TREFHT), which are treated independently of one another in our analysis.

4.2.1. Spatial dimension reduction via global averaging

Stratospheric Warming Pathway

We first consider the simplest spatial dimension reduction, obtained by taking a global average of each of the variables of interest. As before, we first evaluate our RF regressors' skill at reconstructing the features of interest relevant to the stratospheric warming

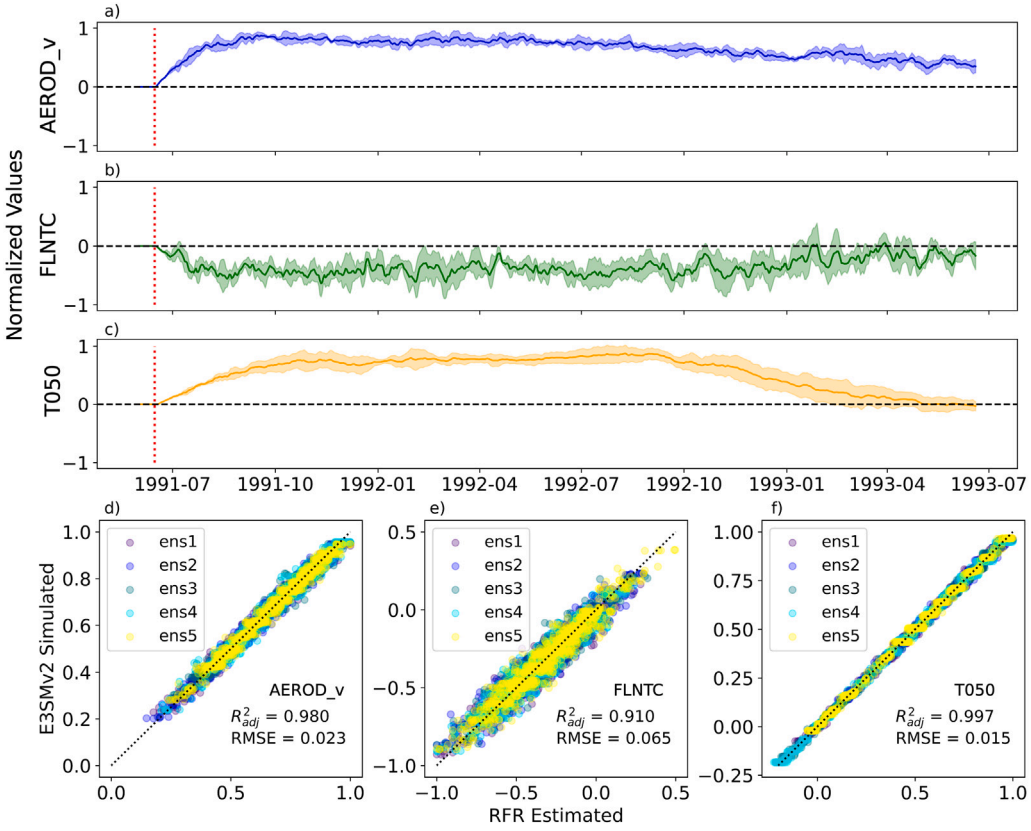

Fig. 6. Mount Pinatubo exemplar: goodness of fit statistics for the stratospheric warming pathway using globally-averaged variables. Subplots (a), (b) and (c) show the input time-series data simulated using E3SMv2-SPA for the variables AEROD_v, FLNTC and T050, respectively. Shaded areas show the ensemble standard deviation. The dark line is the ensemble mean. The dashed red line shows the date of Mount Pinatubo eruption. Subplots (d), (e), and (f) show the scatterplots between the normalized features (y-axis) and their RFR reconstructions (x-axis).

Table 5
Mount Pinatubo exemplar: goodness of fit statistics for the pruned globally-averaged data and the stratospheric warming pathway shown in Fig. 7.

Variable	R^2_{adj} mean	R^2_{adj} σ	$RMSE$ mean	$RMSE$ σ
AEROD_v	0.980	0.003	0.023	0.001
FLNTC	0.910	0.012	0.065	0.001
T050	0.997	0.001	0.015	0.001

pathway, AEROD_v, FLNTC and T050. We start by calculating the goodness of fit metrics R^2_{adj} and $RMSE$ (see Section 2.3) for the three variables defining the stratospheric warming pathway, AEROD_v → FLNTC → T050. These values are reported in Table 5. The reader can observe that R^2_{avg} is above 0.9 for all five variables with an $RMSE$ of at most 6.5%. Additionally, the standard deviation is negligible for both quantities. Corroborating these results are the images provided in Fig. 6, which shows the time-series of the three variables considered AEROD_v, FLNTC, and T050 (subplots (a), (b) and (c), respectively) and scatterplots between the normalized time-series data and their RFR reconstructions (subplots (d), (e), (f), respectively). The reader can observe that FLNTC is the noisiest field, which translates to the lowest R^2_{adj} value and the highest $RMSE$ value for the RFR reconstruction. The reader will also notice that, although the climate simulation data available ran from June 1, 1991 to December 31, 1998, in our analysis we only consider the first 750 days (approximately 2 years, or June 1, 1991 to June 20, 1993) around the eruption. This is for two reasons: (i) after approximately 2 years, the anomaly calculated by subtracting the counterfactual simulations from the fully coupled Mount Pinatubo simulations returns to zero, and (ii) when we consider the full simulation, only the autocorrelated edges survive our pruning criteria, resulting in a null pathway.

Table 6 shows the pruned pathway results for each source to target node, including the time lag, the mean and standard deviation SHAP weights, as well as the number of ensembles that contained that edge. Fig. 7 shows the stratospheric warming pathway graph obtained using globally-averaged data and corresponding to Table 6. To simplify the visualization, we exclude autocorrelated connections from this graph. The variable colors correspond to the colors used in the verification figure (Fig. 6), with AEROD_v in blue, FLNTC in green, and T050 in yellow. The edge weights are shown by the arrows connecting each node in colors ranging

Table 6
Mount Pinatubo exemplar: pathway edge weights and their standard deviations for the stratospheric warming pathway obtained using globally-averaged data windowed to 750 days.

Source	Target	Lag (days)	SHAP Weight	Weight σ	Ensembles with edge
Globe_T050	Globe_T050	1	0.2500	0.0363	5
Globe_FLNTC	Globe_FLNTC	1	0.1767	0.0136	5
Globe_AEROD_v	Globe_AEROD_v	1	0.1347	0.0074	5
Globe_T050	Globe_FLNTC	1	0.0016	0.0003	5
Globe_AEROD_v	Globe_FLNTC	6	0.0014	0.0007	5
Globe_FLNTC	Globe_AEROD_v	31	0.0004	0.0003	5
Globe_FLNTC	Globe_T050	36	0.0001	0.0001	5

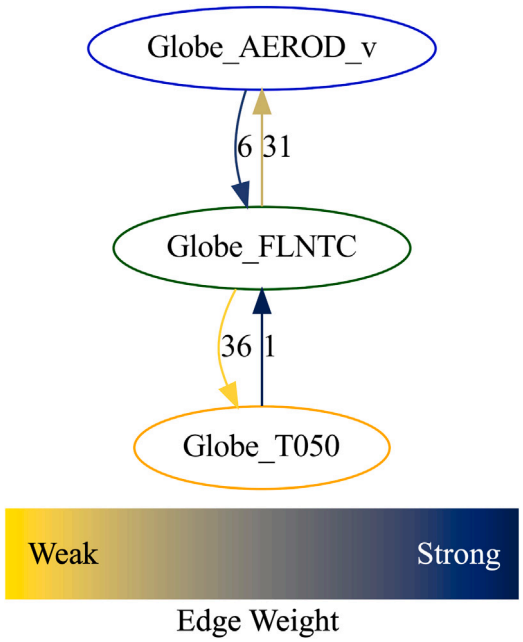

Fig. 7. Mount Pinatubo exemplar: stratospheric warming pathway obtained from globally-averaged data. . Variable colors correspond to the colors used in the verification figure (Fig. 6). Edge weights are shown in colors ranging from blue (strong edge) to yellow (weak edge) and labeled according to the lag in days.

from blue (strong edge) to yellow (weak edge) and labeled according to the lag in days. Even though our pruning criteria allowed for four incoming edges to each node (Section 2.6), to simplify the visualization, we show the top three incoming edges to each node after removing any auto-correlated edges. The reader can see a strong connection going from AEROD_v to FLNTC with a 6-day lag and a weak connection from FLNTC to T050 with a 36-day lag. These "forward" relationships are expected and demonstrate the influence of aerosols in the stratosphere on radiative forcing and temperature. While it took only approximately 22 days for the aerosols to encircle the globe (see Fig. 3), the longer 36-day lag time found in the FLNTC to T050 relationship is likely due to the fact that we are working with global averages, which can be slower to respond than regional averages or individual grid cells. The reader can observe that there are several "back" edges in Fig. 6, that is, edges where we see the opposite of the expected relationships (e.g., T050 causing changes to AEROD_v). In particular, there is a moderate back edge from FLNTC to AEROD_v with a 31-day lag, and a strong back edge from T050 to FLNTC with a 1-day lag. We attribute these back edges to missing intermediary variables in our analysis. For example, including an intermediary variable representing wind patterns could explain the T050 to FLNTC relationship: if warmer temperatures have an impact on wind patterns and therefore aerosol spread, it is conceivable for T050 to have an indirect influence on FLNTC.

Surface Cooling Pathway

Next, we turn our attention to the globally averaged surface cooling pathway: AEROD_v → FSDSC → TREFHT. The RFR reconstructions of the variables defining this pathway are shown in Fig. 8, and the goodness of fit statistics are given in Table 7. The reader can observe that TREFHT is the noisiest variable; this makes the identification of the surface cooling pathway particularly difficult. The variable fits are slightly better than they were for the stratospheric warming pathway (Table 5). We note that the R^2_{adj} and $RMSE$ values for the AEROD_v quantity are slightly different in Tables 5 and 7; this is because our RFR fit was performed for different sets of variables in these two cases.

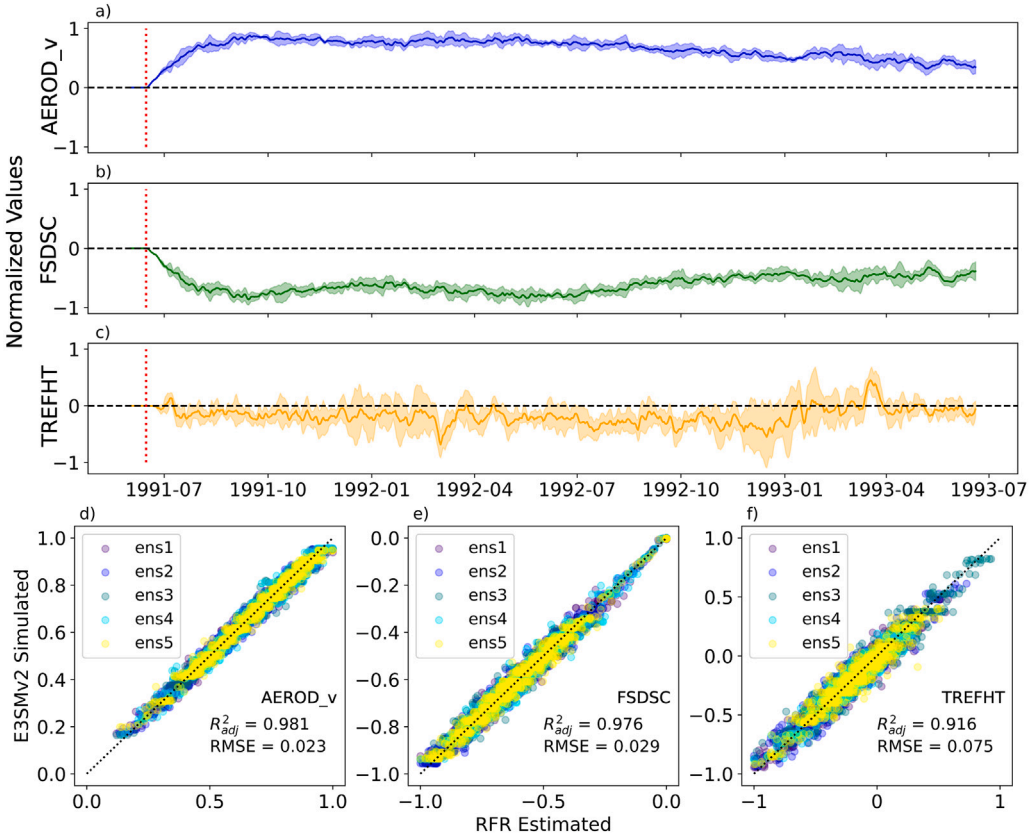

Fig. 8. Mount Pinatubo exemplar: goodness of fit statistics for the surface cooling pathway using globally-averaged variables. Subplots (a), (b) and (c) show the input time-series data simulated using E3SMv2-SPA for the variables AEROD_v, FSDSC and TREFHT, respectively. Shaded areas show the ensemble standard deviation. The dark line is the ensemble mean. The dashed red line shows the date of Mount Pinatubo eruption. Subplots (d), (e), and (f) show the scatterplots between the normalized features (y-axis) and their RFR reconstructions (x-axis).

Table 7
Mount Pinatubo exemplar: goodness of fit statistics for the pruned globally-averaged data and the surface cooling pathway shown in Fig. 9.

Variable	R^2_{adj} mean	R^2_{adj} σ	$RMSE$ mean	$RMSE$ σ
AEROD_v	0.981	0.002	0.023	0.001
FSDSC	0.976	0.003	0.029	0.001
TREFHT	0.916	0.016	0.075	0.006

Fig. 9 shows the surface cooling pathway for the variables of interest, with AEROD_v in blue, FSDSC in green, and TREFHT in yellow. As before, the edge weights are shown by the arrows, with strong edges in blue and weak edges in yellow, again excluding the autocorrelated edges. Fig. 9 shows a strong 1-day lag between AEROD_v and FSDSC, a moderate 51-day lag between FSDSC and TREFHT and a weak 1-day back edge between FSDSC and AEROD_v. While the former two "forward" relationships are expected based on what is known about the surface cooling pathway, it is difficult to corroborate the time lags associated with these variable dependencies. As before, we believe back edges are due to processes involving variables not explicitly included in our analysis (e.g., wind-related variables). Table 8 shows the ensemble mean edge strength and standard deviation between the source and target variables of the pruned surface pathway, as well as the number of ensembles that contain the edge. As with all of the pathways, the variables are strongly autocorrelated, which is expected for a time-series of Earth system variables.

4.2.2. Spatial dimension reduction via averaging across latitudinal bands

The globally averaged pathways, while informative, are admittedly sparse, making it difficult to interpret physically the lags between features. While these pathways show the expected influence of AEROD_v on radiative forcing and temperature, the time lags associated with the pathways are not always explainable, which suggests that the global averaging may be masking important relationships occurring at a regional scale. To mitigate this issue, the next step of our evaluation focuses on the two latitude bands closest to Mount Pinatubo eruption: the Tropical band (23°S to 23°N) and the Subtropical North band (23°N to 35°N).

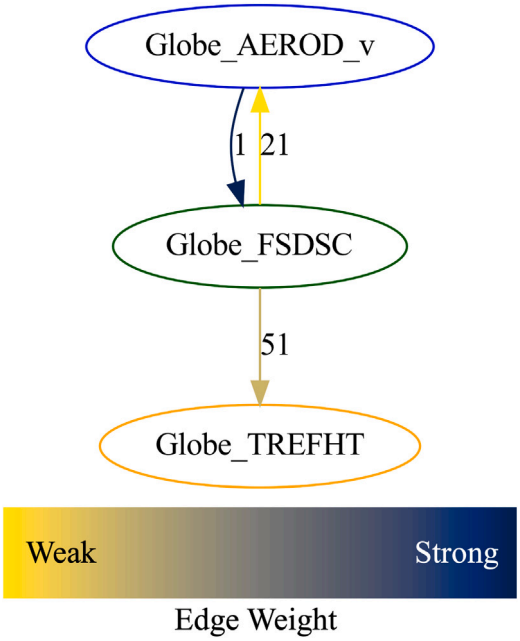

Fig. 9. Mount Pinatubo exemplar: surface cooling pathway obtained from globally-averaged data. Variable colors correspond to the colors used in the verification figure (Fig. 8). Edge weights are shown in colors ranging from blue (strong edge) to yellow (weak edge) and labeled according to the lag in days.

Table 8
Mount Pinatubo exemplar: pathway edge weights and their standard deviations for the surface cooling pathway obtained using globally-averaged data windowed to 750 days.

Source	Target	Lag (days)	SHAP Weight	Weight σ	Ensembles with edge
Globe_TREFHT	Globe_TREFHT	1	0.2066	0.0344	5
Globe_FSDSC	Globe_FSDSC	1	0.1479	0.0069	5
Globe_AEROD_v	Globe_AEROD_v	1	0.1396	0.0069	5
Globe_AEROD_v	Globe_FSDSC	1	0.0099	0.0042	5
Globe_FSDSC	Globe_TREFHT	51	0.0022	0.0011	5
Globe_FSDSC	Globe_AEROD_v	21	0.0001	0.0001	5

Table 9
Mount Pinatubo exemplar: goodness of fit statistics for the pruned zonally-averaged data and the stratospheric warming pathway.

Variable	R^2_{adj} mean	R^2_{adj} σ	$RMSE$ mean	$RMSE$ σ
Tropical_T050	0.993	0.002	0.018	0.003
Tropical_AEROD_v	0.986	0.003	0.023	0.001
SubtropN_T050	0.983	0.006	0.032	0.006
SubtropN_AEROD_v	0.902	0.022	0.052	0.002
Tropical_FLNTC	0.893	0.021	0.047	0.004
SubtropN_FLNTC	0.829	0.044	0.098	0.008

Stratospheric Warming Pathway

Table 9 shows the reconstruction statistics for the latitudinal bands in the stratospheric warming pathway. Generally, the reconstructions are better for the features in the Tropical band than the Subtropical North band. As expected, the noisy FLNTC signal is the most difficult to reconstruct with high accuracy. Table 10 shows the pruned pathway results for the two latitudinal bands. Fig. 10 shows the pathway graph inferred from Table 10, with autocorrelations removed for visualization purposes, as before. The variable colors correspond to the colors used in the verification figure (Fig. 7). The reader can observe that, while Fig. 10 is much more complex than our previous analogous figure obtained using global averages (Fig. 7), the same characteristic connections between AEROD_v, radiative flux and temperature exist. The strongest connections are those that go into the Subtropical North FLNTC from AEROD_v in both latitudinal bands and from Tropical FLNTC. In addition to intra-zonal band relationships (e.g., AEROD_v→ FLNTC), we also see some cross-band connections, many of which show a dependence of a Subtropical band

Table 10
Mount Pinatubo exemplar: pathway edge weights and their standard deviations for the stratospheric warming pathway obtained using zonally-averaged data windowed to 750 days.

Source	Target	Lag (days)	SHAP Weight	Weight σ	Ensembles with edge
Tropical_T050	Tropical_T050	1	0.1949	0.0488	5
SubtropN_T050	SubtropN_T050	1	0.1912	0.0439	5
SubtropN_FLNTC	SubtropN_FLNTC	1	0.1892	0.0255	5
Tropical_AEROD_v	Tropical_AEROD_v	1	0.1562	0.0195	5
SubtropN_AEROD_v	SubtropN_AEROD_v	1	0.1404	0.0213	5
Tropical_FLNTC	Tropical_FLNTC	1	0.1161	0.0108	5
SubtropN_AEROD_v	SubtropN_FLNTC	1	0.0038	0.0013	5
Tropical_FLNTC	SubtropN_FLNTC	1	0.0033	0.0013	5
Tropical_AEROD_v	Tropical_FLNTC	6	0.0025	0.0005	5
Tropical_AEROD_v	SubtropN_AEROD_v	6	0.0025	0.0009	5
Tropical_AEROD_v	SubtropN_FLNTC	46	0.0017	0.0005	5
Tropical_FLNTC	SubtropN_AEROD_v	1	0.0013	0.0008	5
SubtropN_T050	SubtropN_AEROD_v	46	0.0011	0.0004	5
SubtropN_AEROD_v	Tropical_FLNTC	6	0.0009	0.0005	5
SubtropN_FLNTC	Tropical_FLNTC	1	0.0009	0.0004	5
Tropical_T050	SubtropN_T050	21	0.0006	0.0004	5
Tropical_AEROD_v	SubtropN_T050	31	0.0004	0.0002	5
SubtropN_AEROD_v	Tropical_AEROD_v	21	0.0001	0.0001	5

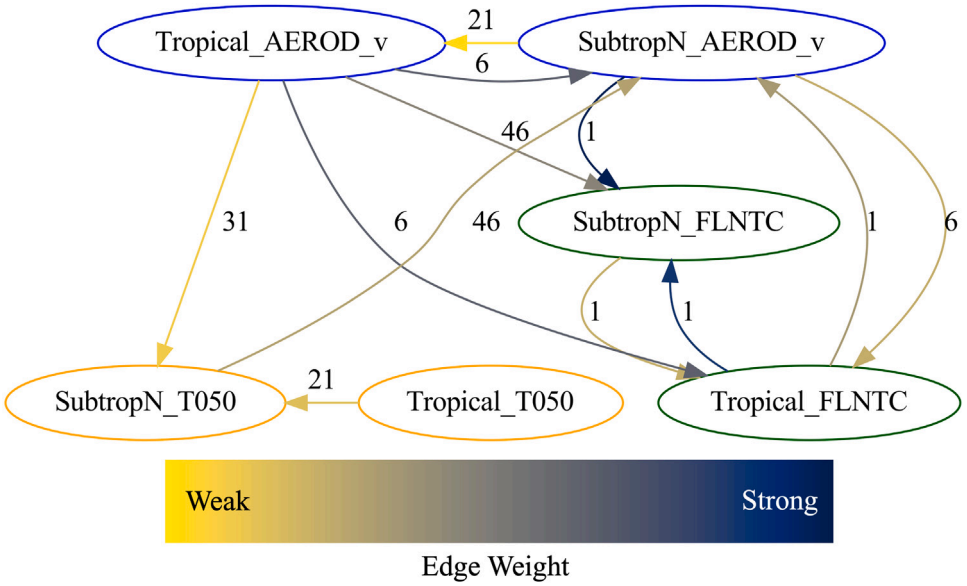

Fig. 10. Mount Pinatubo exemplar: stratospheric warming pathway obtained from zonally-averaged data. Variable colors correspond to the colors used in the verification figure (Fig. 6). Edge weights are shown in colors ranging from blue (strong edge) to yellow (weak edge) and labeled according to the lag in days.

variable on a Tropical band variable. These relationships are expected, as they demonstrates a general northward flowing direction from the Tropics to the Subtropical North, which follows the evolution of the volcanic plume that was observed after the Mount Pinatubo eruption (see Fig. 3). It is interesting to remark that, with our current set of pruning criteria, there are no connections from FLNTC to T050, and only a weak connection having a 31-day time lag from Tropical AEROD_v to Subtropical North T050. This result is surprising, given the prevalence of the FLNTC to T050 relationship in our globally-averaged analyses (Fig. 7).

Surface Cooling Pathway

When we look at the surface cooling pathway for the latitudinally averaged features, we expect to see a similar northward influence, and the connections between AEROD_v → FSDSC → TREFHT that we saw in the globally averaged pathway. Table 11 shows the reconstruction statistics for the surface cooling pathway. As in the stratospheric warming pathway, the Subtropical North variables are more difficult to reconstruct than the Tropical variables, with the Subtropical North TREFHT variable having the lowest R^2_{adj}. The pruned pathway can be inferred from Table 12 and is plotted in Fig. 11 with autocorrelations removed. Again, we see a more complex relationship between the variables than in the global surface cooling pathway graph (Fig. 9). The reader can observe the presence of the expected characteristic patterns of connections going from AEROD_v to FSDSC and to TREFHT, as well as the expected northward propagation of the aerosols and their impacts from the Tropical band to the Subtropical North band. The time

Table 11
Mount Pinatubo exemplar: goodness of fit statistics for the pruned zonally-averaged data and the surface cooling pathway.

Variable	R^2_{adj} mean	R^2_{adj} σ	$RMSE$ mean	$RMSE$ σ
Tropical_AEROD_v	0.986	0.003	0.023	0.001
Tropical_FSDSC	0.978	0.005	0.025	0.003
SubtropN_AEROD_v	0.915	0.017	0.051	0.002
Tropical_TREFHT	0.909	0.038	0.023	0.001
SubtropN_FSDSC	0.845	0.042	0.069	0.009
SubtropN_TREFHT	0.771	0.036	0.105	0.010

Table 12
Mount Pinatubo exemplar: pathway edge weights and their standard deviations for the surface cooling pathway obtained using zonally-averaged data windowed to 750 days.

Source	Target	Lag (days)	SHAP Weight	Weight σ	Ensembles with edge
SubtropN_TREFHT	SubtropN_TREFHT	1	0.1735	0.0099	5
Tropical_AEROD_v	Tropical_AEROD_v	1	0.1559	0.0197	5
SubtropN_AEROD_v	SubtropN_AEROD_v	1	0.1474	0.0200	5
SubtropN_FSDSC	SubtropN_FSDSC	1	0.1435	0.0198	5
Tropical_FSDSC	Tropical_FSDSC	1	0.1368	0.0319	5
Tropical_TREFHT	Tropical_TREFHT	1	0.0682	0.0181	5
SubtropN_FSDSC	SubtropN_TREFHT	1	0.0045	0.0010	5
Tropical_AEROD_v	Tropical_FSDSC	1	0.0038	0.0020	5
SubtropN_AEROD_v	SubtropN_FSDSC	1	0.0036	0.0012	5
Tropical_TREFHT	SubtropN_TREFHT	6	0.0029	0.0015	5
Tropical_AEROD_v	SubtropN_AEROD_v	6	0.0028	0.0018	5
SubtropN_AEROD_v	SubtropN_TREFHT	11	0.0026	0.0006	5
Tropical_AEROD_v	SubtropN_FSDSC	36	0.0020	0.0007	5
Tropical_TREFHT	SubtropN_FSDSC	41	0.0019	0.0011	5
Tropical_FSDSC	Tropical_AEROD_v	1	0.0017	0.0015	5
Tropical_TREFHT	SubtropN_AEROD_v	16	0.0016	0.0010	5
SubtropN_TREFHT	SubtropN_AEROD_v	1	0.0014	0.0007	5
SubtropN_TREFHT	Tropical_TREFHT	1	0.0011	0.0008	5
SubtropN_FSDSC	Tropical_TREFHT	31	0.0005	0.0001	5
SubtropN_TREFHT	Tropical_FSDSC	21	0.0003	0.0002	5
SubtropN_AEROD_v	Tropical_FSDSC	26	0.0003	0.0002	5
SubtropN_TREFHT	Tropical_AEROD_v	21	0.0002	0.0001	5

lags associated with many of the relationships uncovered are on the order of 21–36 days, which is consistent with the time it took for the aerosols from Mount Pinatubo to encircle the globe (Fig. 3).

5. Conclusions and future work

The primary contribution of this paper is the development of a data-driven "outer loop" algorithm for teasing out source-impact relationships in large climate datasets. Whereas traditional ML approaches employ RFR and feature importance for classification or regression, our unique combination of these two methods enables us to extract from climate data the relative influence of one feature on another, towards tracing out source-impact pathways. After developing some quantitative methods for evaluating our approach and verifying it on a set of synthetic coupled equations, we deployed the method on two known pathways that occurred as a result of the 1991 eruption of Mount Pinatubo: the stratospheric warming and surface cooling pathways. In order to make the method tractable for this complex high-dimensional exemplar, data reduction in the form of spatial averaging and temporal binning was performed.

While our approach was generally successful in finding the key "nodes" in each of the sought-after temperature pathways, it also identified some unexpected relationships. Our analysis revealed that reducing the data via a global averaging can over-smooth the data, making it more difficult for our method to pick up on the signals of interest, and complicating the expected temporal relationships between features. For both the global and zonal averaging techniques considered in this paper, we observed the presence of "back" edges in our pathway graphs, i.e., in addition to determining that FLNTC is influenced by AEROD_v, as expected, our analysis found that AEROD_v is also influenced by FLNTC. We believe that, when this happens, our method is finding connections involving variables and physical mechanisms not utilized explicitly in the pathway analysis performed, e.g., wind field information. Repeating our analysis with additional input variables to confirm this conjecture would be an interesting future research endeavor. We additionally intend to do more regional analysis, e.g., by repeating the analysis performed herein using IPCC regions [44]. Finally, for the zonal analysis, we found that our method did not always discover the expected time-lags of influence between the relevant variables/features. To try to improve on this, future work will involve more substantial windowing of the data, in particular,

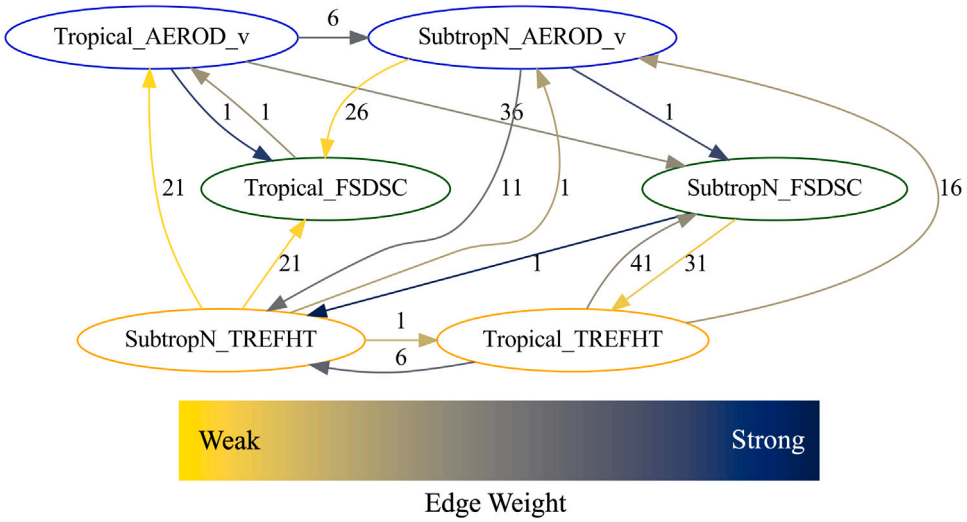

Fig. 11. Mount Pinatubo exemplar: surface cooling pathway obtained from zonally-averaged data. Variable colors correspond to the colors used in the verification figure (Fig. 8). Edge weights are shown in colors ranging from blue (strong edge) to yellow (weak edge) and labeled according to the lag in days.

using sliding windows to study the temporal system-wide dynamic impacts of a climate source.

While we have demonstrated the method in a confirmatory capacity by postulating a set of relationships and using the method to confirm/deny them, our method has the potential to *discover* source-impact pathways directly from data. Deploying the method in this *exploratory* capacity may enable the discovery of previously unknown relationships in the climate. Finally, it is important to recognize that the relationships identified by our approach are correlative rather than causal. Developing a causal analysis-based method that can pick out causal vs. correlative edges in our pathway graphs would be of tremendous interest.

Data and code availability

Data from the full E3SMv2-SPA simulation campaign including pre-industrial control, historical, and Mount Pinatubo ensembles will be hosted at Sandia National Laboratories with location and download instructions announced on https://www.sandia.gov/cldera/e3sm-simulations-data/ when available. A Python code implementing our RFR and feature importance-based approach is available at https://github.com/sandialabs/RFR-CLDERA. The E3SMv2-SPA climate code that was used to generate the data analyzed herein can be found at https://github.com/sandialabs/CLDERA-E3SM.

Acknowledgments

This material is based upon work supported by the Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories. The research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility using NERSC award BER-ERCAP0026535. The writing of this manuscript was funded in part by the third author's (Irina Tezaur's) Presidential Early Career Award for Scientists and Engineers (PECASE).

This article has been authored by an employee of National Technology & Engineering Solutions of Sandia, LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employee owns all right, title and interest in and to the article and is solely responsible for its contents. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this article or allow others to do so, for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan https://www.energy.gov/downloads/doe-public-access-plan. The authors would like to thank Hunter Brown, Benjamin Wagman, Thomas Ehrmann, and Joe Hollowed for providing invaluable subject matter expertise in stratospheric dynamics and climate modeling, which enabled the interpretations discussed within this paper. We would also like to thank Laura Swiler for helping us refine various details in our algorithm, including the choice of pruning criteria.

Data availability

Data from the full E3SMv2-SPA simulation campaign including pre-industrial control, historical, and Mount Pinatubo ensembles will be hosted at Sandia National Laboratories with location and download instructions announced on https://www.sandia.gov/cldera/e3sm-simulations-data/ when available. A Python code implementing our RFR and feature importance-based approach is

currently awaiting copyright assertion, and will be made publicly available as soon as this is possible. The E3SMv2-SPA climate code that was used to generate the data analyzed herein can be found at https://github.com/sandialabs/CLDERA-E3SM.

References

[1] M. Burger, J. Wentz, R. Horton, The law and science of climate change attribution, Columbia J. Environ. Law 45 (2020) 57, http://dx.doi.org/10.7916/cjel.v45i1.4730.

[2] K. Hasselmann, Optimal fingerprints for the detection of time-dependent climate change, J. Clim. 6 (10) (1993) 1957–1971, http://dx.doi.org/10.1175/1520-0442(1993)006<1957:OFFTDO>2.0.CO;2.

[3] B. Santer, T. Wigley, P. Jones, Correlation methods in fingerprint detection studies, Clim. Dyn. 8 (6) (1993) 265–276, http://dx.doi.org/10.1007/BF00209666.

[4] C. Bonfils, B.D. Santer, D.W. Pierce, H.G. Hidalgo, G. Bala, T. Das, T.P. Barnett, D.R. Cayan, C. Doutriaux, A.W. Wood, A. Mirin, T. Nozawa, Detection and attribution of temperature changes in the mountainous western United States, J. Clim. 21 (23) (2008) 6404–6424, http://dx.doi.org/10.1175/2008JCLI2397.1.

[5] P.A. Stott, N.P. Gillett, G.C. Hegerl, D.J. Karoly, D.A. Stone, X. Zhang, F. Zwiers, Detection and attribution of climate change: a regional perspective, Wiley Interdiscip. Rev. Clim. Change 1 (2) (2010) 192–211, http://dx.doi.org/10.1002/wcc.34.

[6] C.J.W. Bonfils, B.D. Santer, J.C. Fyfe, K. Marvel, T.J. Phillips, S.R.H. Zimmerman, Human influence on joint changes in temperature, rainfall and continental aridity, Nature Clim. Change 10 (2020) 726–731, http://dx.doi.org/10.1038/s41558-020-0821-1.

[7] K. Marvel, M. Biasutti, C. Bonfils, Fingerprints of external forcings on sahel rainfall: aerosols, greenhouse gases, and model-observation discrepancies, Environ. Res. Lett. 15 (8) (2020) 084023, http://dx.doi.org/10.1088/1748-9326/ab858e.

[8] M.R. Allen, P.A. Stott, Estimating signal amplitudes in optimal fingerprinting, part I: Theory, Clim. Dyn. 21 (5) (2003) 477–491, http://dx.doi.org/10.1007/s00382-003-0313-9.

[9] A. Ribes, S. Planton, L. Terray, Application of regularised optimal fingerprinting to attribution. Part I: method, properties and idealised analysis, Clim. Dyn. 41 (11) (2013) 2817–2836, http://dx.doi.org/10.1007/s00382-013-1735-7.

[10] R.C.J. Wills, D.S. Battisti, K.C. Armour, T. Schneider, C. Deser, Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations, J. Clim. 33 (20) (2020) 8693–8719, http://dx.doi.org/10.1175/JCLI-D-19-0855.1.

[11] M. Weylandt, L.P. Swiler, Beyond PCA: Additional dimension reduction techniques to consider in the development of climate fingerprints, J. Clim. 37 (5) (2024) 1723–1735, http://dx.doi.org/10.1175/JCLI-D-23-0267.1, URL https://journals.ametsoc.org/view/journals/clim/37/5/JCLI-D-23-0267.1.xml.

[12] C.R. Wentland, M. Weylandt, L.P. Swiler, T.S. Ehrmann, D. Bull, Conditional multi-step attribution for climate forcings, J. Clim. (2024) arXiv:2409.01396. URL https://arxiv.org/abs/2409.01396, (submitted for publication).

[13] J. Runge, S. Bathiany, E. Bollt, G. camps Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. Mahecha, J. Munoz-Mari, E. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Scholkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, J. Zscneischler, Inferring causation from time series in earth system sciences, Nature Commun. 10 (2019).

[14] P. Nowack, J. Runge, V. Eyring, J. Haigh, Causal networks for climate model evaluation and constrained projections, Nature Commun. 11 (1415) (2020).

[15] J. Nichol, M. Weylandt, M. Fricke, M. Moses, D. Bull, L. Swiler, Space-time causal discovery in climate science: A local stencil learning approachpace-time causal discovery in climate science: A local stencil learning approach, JGR Mach. Learn. Comput. (2024) http://dx.doi.org/10.22541/essoar.172253117.78663487/v1 (submitted for publication).

[16] C. Bône, G. Gastineau, S. Thiria, P. Gallinari, Detection and attribution of climate change: A deep learning and variational approach, Envir. Data Sci. 1 (2022) e27.

[17] A. Mamalakis, I. Ebert-Uphoff, E.A. Barnes, Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science, in: International Workshop on Extending Explainable AI beyond Deep Models and Classifiers, Springer, 2020, pp. 315–339.

[18] C. Buckland, R. Bailey, D. Thomas, Using artificial neural networks to predict future dryland responses to human and climate disturbances, Sci. Rep. 9 (1) (2019) 3855.

[19] J. Hart, M. Gulian, I. Manickam, L.P. Swiler, Solving high-dimensional inverse problems with auxiliary uncertainty via operator learning with limited data, J. Mach. Learn. Model. Comput. 4 (2) (2023) 105–133.

[20] J. Hart, I. Manickam, M. Gulian, L. Swiler, D. Bull, T. Ehrmann, H. Brown, B. Wagman, J. Watkins, Stratospheric aerosol source inversion: Noise, variability, and uncertainty quantification, J. Mach. Learn. Model. Comput. (2024) http://dx.doi.org/10.1615/JMachLearnModelComput.2024056144.

[21] C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[22] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[23] J. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.

[24] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

[25] G. Casalicchio, C. Molnar, B. Bischl, Visualizing the feature importance for black box models, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18, Springer, 2019, pp. 655–670.

[26] J. Hansen, A. Lacis, R. Ruedy, M. Sato, Potential climate impact of mount pinatubo eruption, Geophys. Res. Lett. 19 (2) (1992) 215–218.

[27] M. Kilian, S. Brinkop, P. Jöckel, Impact of the eruption of mt pinatubo on the chemical composition of the stratosphere, Atmos. Chem. Phys. 20 (20) (2020) 11697–11715.

[28] D. Parker, H. Wilson, P.D. Jones, J. Christy, C.K. Folland, The impact of mount pinatubo on world-wide temperatures, Int. J. Climatol.: J. R. Meteorol. Soc. 16 (5) (1996) 487–497.

[29] S. Self, J.-X. Zhao, R.E. Holasek, R.C. Torres, A.J. King, The Atmospheric Impact of the 1991 Mount Pinatubo Eruption, Tech. Rep., 1993.

[30] A. Robock, Volcanic eruptions and climate, Rev. Geophys. 38 (2) (2000) 191–219, http://dx.doi.org/10.1029/1998RG000054.

[31] C. Timmreck, Modeling the climatic effects of large explosive volcanic eruptions, Wiley Interdiscip. Rev. Clim. Change 3 (6) (2012) 545–564.

[32] L.R. Marshall, E.C. Maters, A. Schmidt, C. Timmreck, A. Robock, M. Toohey, Volcanic effects on climate: recent advances and future avenues, Bull. Volcanol. 84 (5) (2022) 54, http://dx.doi.org/10.1007/s00445-022-01559-3.

[33] C.G. Newhall, The Cataclysmic 1991 Eruption of Mount Pinatubo, Philippines, vol. 113, US Geological Survey, 1997.

[34] J.-C. Golaz, L.P. Van Roekel, X. Zheng, A.F. Roberts, J.D. Wolfe, W. Lin, A.M. Bradley, Q. Tang, M.E. Maltrud, R.M. Forsyth, et al., The DOE E3SM Model version 2: Overview of the physical model and initial model evaluation, J. Adv. Modelling Earth Syst. 14 (12) (2022) e2022MS003156.

[35] H.Y. Brown, B. Wagman, D. Bull, K. Peterson, B. Hillman, X. Liu, Z. Ke, L. Lin, Validating a microphysical prognostic stratospheric aerosol implementation in E3SMv2 using the mount pinatubo eruption, Geosci. Model Dev. 2024 (2024) 1–46.

[36] D. Yarger, D. Tucker, Detecting changepoints in globally-indexed functional time series, Environmetrics (2024) arXiv:2308.05915 (submitted for publication).

[37] S. Shi-Jun, L. Shand, B. Li, Tracing the impacts of mount pinatubo eruption on global climate using spatially-varying changepoint detection, Ann. Appl. Stat. (2024) arXiv:2409.08908.

[38] K. Goode, D. Ries, K. McClernon, Characterizing climate pathways using feature importance on echo state networks, Stat. Anal. Data Mining: ASA Data Sci. J. 17 (4) (2024) e11706, http://dx.doi.org/10.1002/sam.11706, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11706. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11706.

[39] D. Ries, K. Goode, K. McClernon, B. Hillman, Using feature importance as exploratory data analysis tool on earth system models, Geosci. Model Develop. Discuss. (2024) 1–35, http://dx.doi.org/10.5194/gmd-2024-133 (submitted for publication).

[40] K. McClernon, K. Goode, D. Ries, A comparison of model validation approaches for echo state networks using climate model replicates, Spat. Stat. 59 (2024) 100813.

[41] R. Garrett, L. Shand, G. Huerta, A multivariate space-time dynamic model for characterizing the atmospheric impacts following the Mt. Pinatubo eruption, Environmetrics (2024) arXiv:2408.13392 (submitted for publication).

[42] A. Steyer, L. Bertagna, G. Harper, J. Watkins, I. Tezaur, D. Bull, In-situ data extraction for pathway analysis in an idealized configuration of E3SM, 2024, ArXiv pre-print URL https://arxiv.org/abs/2408.04099.

[43] J. Watkins, L. Bertagna, G. Harper, A. Steyer, I. Tezaur, D. Bull, Entropy-based feature selection for capturing impacts in earth system models with extreme forcing, J. Comput. Appl. Math. (2024) arXiv:2409.18011 (in preparation).

[44] M. Iturbide, J.M. Gutiérrez, L.M. Alves, J. Bedia, R. Cerezo-Mota, E. Cimadevilla, A.S. Cofiño, A. Di Luca, S.H. Faria, I.V. Gorodetskaya, M. Hauser, S. Herrera, K. Hennessy, H.T. Hewitt, R.G. Jones, S. Krakovska, R. Manzanas, D. Martínez-Castro, G.T. Narisma, I.S. Nurhati, I. Pinto, S.I. Seneviratne, B. van den Hurk, C.S. Vera, An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets, Earth Syst. Sci. Data 12 (4) (2020) 2959–2970, http://dx.doi.org/10.5194/essd-12-2959-2020, URL https://essd.copernicus.org/articles/12/2959/2020/.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (85) (2011) 2825–2830, URL http://jmlr.org/papers/v12/pedregosa11a.html.

[46] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[47] L.S. Shapley, A value for n-person games, Contrib. Theory Games 2 (1953).

[48] S. Nembrini, I. Konig, M. Write, The revival of the gini importance? Bioinformatics 34 (21) (2018) 3711–3718, http://dx.doi.org/10.1093/bioinformatics/bty373.

[49] J.J. Nichol, M.G. Peterson, K.J. Peterson, G.M. Fricke, M.E. Moses, Machine learning feature analysis illuminates disparity between E3sm climate models and observed climate change, J. Comput. Appl. Math. 395 (2021) 113451, http://dx.doi.org/10.1016/j.cam.2021.113451, URL https://www.sciencedirect.com/science/article/pii/S0377042721000704.

[50] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (10) (2010) 1340–1347, http://dx.doi.org/10.1093/bioinformatics/btq134, arXiv:https://academic.oup.com/bioinformatics/article-pdf/26/10/1340/48851160/bioinformatics_26_10_1340.pdf.

[51] W.E. Marcílio, D.M. Eler, From explanations to feature selection: assessing SHAP values as feature selection mechanism, in: 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI, Ieee, 2020, pp. 340–347.

[52] H. Wang, Q. Liang, J.T. Hancock, T.M. Khoshgoftaar, Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods, J. Big Data 11 (1) (2024) 44.

[53] J. Dunn, L. Mingardi, Y.D. Zhuo, Comparing interpretability and explainability for feature selection, 2021, arXiv preprint arXiv:2105.05328.

[54] I.E. Kumar, S. Venkatasubramanian, C. Scheidegger, S. Friedler, Problems with Shapley-value-based explanations as feature importance measures, in: International Conference on Machine Learning, PMLR, 2020, pp. 5491–5500.

[55] K. Labitzke, M.P. McCormick, Stratospheric temperature increases due to pinatubo aerosols, Geophys. Res. Lett. 19 (2) (1992) 207–210, http://dx.doi.org/10.1029/91GL02940.

[56] B.J. Soden, R.T. Wetherald, G.L. Stenchikov, A. Robock, Global cooling after the eruption of Mount Pinatubo: A test of climate feedback by water vapor, Science 296 (5568) (2002) 727–730.

[57] L.M. Polvani, A. Banerjee, A. Schmidt, Northern hemisphere continental winter warming following the 1991 mt. pinatubo eruption: reconciling models and observations, Atmos. Chem. Phys. 19 (9) (2019) 6351–6366, http://dx.doi.org/10.5194/acp-19-6351-2019, URL https://acp.copernicus.org/articles/19/6351/2019/.

[58] E.G. Dutton, J.R. Christy, Solar radiative forcing at selected locations and evidence for global lower tropospheric cooling following the eruptions of El Chichón and Pinatubo, Geophys. Res. Lett. 19 (23) (1992) 2313–2316, http://dx.doi.org/10.1029/92GL02495, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/92GL02495, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/92GL02495.

[59] T. Ehrmann, B. Wagman, D. Bull, H. Brown, B. Hillman, K. Peterson, L. Swiler, J. Watkins, J. Hart, Identifying northern hemisphere temperature responses to the mt. pinatubo eruption through limited variability ensembles, Clim. Dyn. (2024) under review.

[60] M. Davey, A. Brookshaw, S. Ineson, The probability of the impact of ENSO on precipitation and near-surface temperature, Clim. Risk Manag. 1 (2014) 5–24, http://dx.doi.org/10.1016/j.crm.2013.12.002.