Characterizing climate pathways using feature importance on echo state networks

Running Title
Climate Pathways via Echo State Networks

Authors
Katherine Goode[1]
Daniel Ries[1]
Kellie McClernon[1]

Affiliations
[1] *Sandia National Laboratories*
1515 Eubank SE
Albuquerque, NM 87123

Acknowledgements
The authors thank Lyndsay Shand, Gabriel Huerta, and J. Derek Tucker for their thoughtful suggestions during development stages. Additionally, we thank Gabriel Huerta for his careful read through and feedback on the paper.

Funding Statement
This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Conflict of Interest
The authors have no conflict of interest to declare.

Data Availability Statement
All data and code associated with this manuscript are available at https://github.com/sandialabs/listenr/tree/main/examples/CoDA-paper.

Received XX XX XXXX; | Revised XX XX XXXX; | Accepted XX XX XXXX

DOI: xxx/xxxx

RESEARCH ARTICLE

Characterizing climate pathways using feature importance on echo state networks

Katherine Goode[1]* | Daniel Ries[1] | Kellie McClernon[1]

[1]Sandia National Laboratories, United States

Correspondence
*Katherine Goode, Sandia National Laboratories, Albuquerque, NM. Email: kjgoode@sandia.gov

Abstract

The 2022 National Defense Strategy of the United States listed climate change as a serious threat to national security. Climate intervention methods, such as stratospheric aerosol injection, have been proposed as mitigation strategies, but the downstream effects of such actions on a complex climate system are not well understood. The development of algorithmic techniques for quantifying relationships between source and impact variables related to a climate event (i.e., a climate pathway) would help inform policy decisions. Data-driven deep learning models have become powerful tools for modeling highly nonlinear relationships and may provide a route to characterize climate variable relationships. In this paper, we explore the use of an echo state network (ESN) for characterizing climate pathways. ESNs are a computationally efficient neural network variation designed for temporal data, and recent work proposes ESNs as a useful tool for forecasting spatio-temporal climate data. However, ESNs are non-interpretable black-box models along with other neural networks. The lack of model transparency poses a hurdle for understanding variable relationships. We address this issue by developing feature importance methods for ESNs in the context of spatio-temporal data to quantify variable relationships captured by the model. We conduct a simulation study to assess and compare the feature importance techniques, and we demonstrate the approach on reanalysis climate data. In the climate application, we consider a time period that includes the 1991 volcanic eruption of Mount Pinatubo. This event was a significant stratospheric aerosol injection, which acts as a proxy for an anthropogenic stratospheric aerosol injection. We are able to use the proposed approach to characterize relationships between pathway variables associated with this event that agree with relationships previously identified by climate scientists.

KEYWORDS:
explainable machine learning, interpretability, black-box models, spatio-temporal data, climate security, climate interventions, stratospheric aerosol injections

1 | INTRODUCTION

Climate change poses a serious threat to national security, as acknowledged in the 2022 National Defense Strategy of the United States [1]. This threat is made more complicated by the possibility of artificial climate modifications. For example, strategies such as stratospheric aerosol injections, marine cloud brightening, and cirrus cloud thinning have been

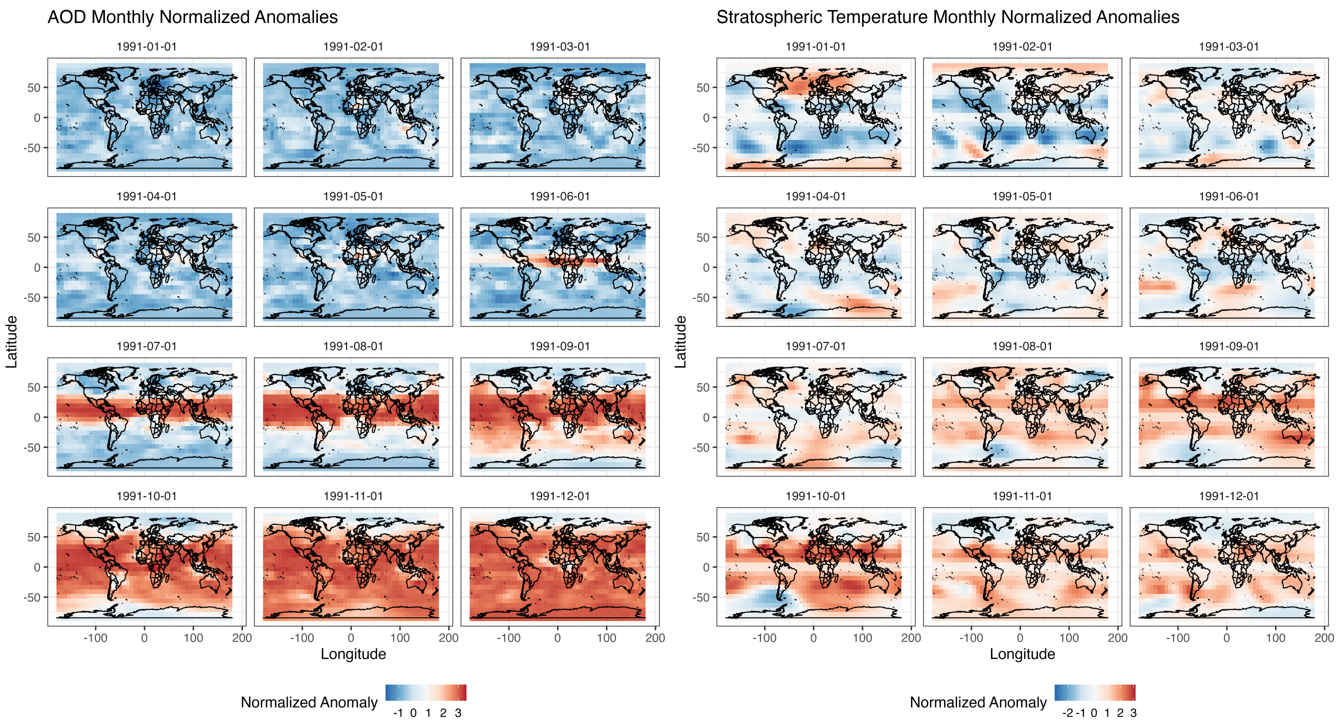

FIGURE 1 MERRA-2 monthly normalized anomalies in 1991 showing the effect of the Mount Pinatubo eruption in June on AOD and stratospheric temperature.

proposed for solar climate interventions [10]. Weather modification strategies have already been implemented on regional scales such as the cloud-seeding array project ‘Sky River’ in China, which was intended to control rainfall over the Tibetan Plateau [50]. While these modifications are meant as mitigation strategies for climate change, the downstream effects of such actions on a complex climate system are not well understood. The development of algorithmic methods for quantifying (i.e., characterizing) the relationships between a climate event source and its impacts could assist policy makers with high consequence decisions.

There are no real world large scale anthropogenic stratospheric aerosol injections to study. Instead, we consider the 1991 volcanic eruption of Mount Pinatubo in the Philippines, which acts as a proxy for an artificial stratospheric aerosol injection. This event has been frequently studied by climate scientists, so the relationships between the source and climate impacts are well understood. For example, the eruption released a massive injection of sulfur dioxide (SO_2; 18-19 Tg) into the atmosphere [15], which led to increases in aerosol optical depth (AOD; a vertically integrated measure of aerosols in the air going from surface to stratosphere) [46, 15]. Ultimately, the increase in AOD resulted in stratospheric temperatures at pressure levels of 30 to 50 mb rising between 2.5 to 3.5 degrees centigrade compared to the 20-year means [23]. Figure 1 shows heatmaps of monthly normalized anomalies (i.e., scaled deviations from monthly spatial averages; see Equation (30)) of AOD and stratospheric temperature in 1991 generated using Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) data [12]. The eruption of Mount Pinatubo occurred in June of 1991, and these visualizations show the clear effects of the eruptions: above average AOD and stratospheric temperature values occurring in July through December.

In the Mount Pinatubo example, SO_2 is the source variable in the climate pathway, and stratospheric temperature is the impact variable. AOD acts as an intermediate variable in the pathway. Our objective is to quantify relationships between climate pathway variables such as these. In this paper, we explore the use of echo state networks (ESNs) [18, 28], a machine learning algorithm, for the task of characterizing climate pathways.

ESNs are known for providing good predictions with chaotic systems [2], and recent work demonstrates the abilities of ESNs with long-lead forecasts on spatio-temporal climate data [32, 33]. ESNs are also computationally efficient models in comparison to recurrent neural networks [4], their sibling machine learning model for temporal data, and other current statistical methods for spatio-temporal forecasting [33]. The efficiency of ESNs is due to many parameters in the model being randomly sampled from distributions instead of estimated using gradient back-propagation as is done with recurrent neural networks. The predictive performance and computational efficiency of ESNs make them appealing models for working with large climate datasets. However, there is a clear obstacle to overcome in order to use ESNs to characterize climate pathways: lack of model *interpretability*.

While there have been different definitions proposed in the literature for what makes a model *interpretable* [9, 26, 45, 39], in this paper, we define *interpretable* as follows.

Definition 1 (Interpretable)**.** A model is *interpretable* if it is possible to assign meaning to the model's parameters in the context of the application, which provides insight into how the model inputs relate to the model outputs.

As an example, consider a linear model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1.$$

We can interpret the coefficient $\hat{\beta}_1$ as the amount the response variable $\hat{y}$ increases for a one unit increase in the predictor variable x_1. With an ESN model, it is not possible to assign meaning to the model parameters due to the complicated nonlinear transformation applied to the input variables. Along with ESNs, many machine learning models including neural networks and random forests are classified as *black-box* models due to their complex algorithms that result in this lack of interpretability. Regardless, black-box machine learning models continue to be implemented in practice due to successful demonstrations of their predictive capabilities and their data driven approach to extracting patterns in complicated applications. An approach to remedy the lack of interpretability is through *explainability*.

The research area of *explainable machine learning* has grown rapidly since 2015 [36]. The objective of explainable machine learning is to understand how black-box models make predictions. This objective is of particular importance in high-consequence application spaces such as the medical sciences, forensics science, and national security. Since it is not possible to directly interpret black-box model parameters, many explainability approaches apply post-hoc techniques to infer how the model inputs relate to model outputs. Similar to the term *interpretable*, various definitions have been proposed for *explainable*. We say a model is *explainable* if it satisfies the following definition.

Definition 2 (Explainable)**.** A model is *explainable* if it is possible to implement post hoc investigations on a trained model that infer how the model inputs relate to the model outputs.

A range of explainability techniques have been proposed including interpretable surrogate models (e.g. LIME [43]), counterfactual examples (e.g., [38]), and model structure visualizations (e.g., [48]). A common explainability method is feature importance (FI). FI aims to quantify the effect of an input variable on a model's predictions. Various techniques have been proposed for computing FI. One example is permutation feature importance (PFI). The concept of PFI was originally introduced by Breiman [7] as a FI technique for random forests and was later generalized to be model-agnostic [11]. The idea with PFI is to randomly permute an input variable, while the other variables remain fixed at their observed values, and quantify how the model prediction performance is affected. Inputs that lead to the largest decrease in model performance are considered the most "important".

In this paper, we approach explainability for an ESN trained on spatio-temporal data by developing two FI techniques. Both methods approach the computation of FI by adjusting an input and quantifying how the model performance is affected similar to PFI. In fact, the first method adapts the concept of PFI to work with spatio-temporal data. PFI has been previously adapted to work with temporal data [47], but to our knowledge, no work has adapted the technique for spatio-temporal. We will refer to this approach as *spatio-temporal permutation feature importance (stPFI)*. Instead of permuting the values, our second approach sets the values of an input variable to zero. This essentially "turns off" the input. A similar idea was proposed as a FI technique for ESNs in Arrieta et al. [4] under the name of *pixel absence effect*. In this paper, we extend the methodology to work with spatio-temporal data, and we refer to this method as *spatio-temporal zeroed feature importance (stZFI)*. We compare these approaches on simulated spatio-temporal datasets and demonstrate how the proposed methods characterize climate pathways with the Mount Pinatubo example.

The remainder of the paper is organized as follows. Section 2 provides the details of a single layer ESN in the context of a climate pathway scenario with spatio-temporal data and introduces the proposed techniques for computing spatio-temporal FI on ESNs. Section 3 describes a simulation study implemented to compare the techniques of stPFI and stZFI for ESNs. In Section 4, the approach is applied to characterize climate variable relationships with the Mount Pinatubo example. Finally, Section 5 describes our conclusions and avenues for future research.

2 | METHODOLOGY

We consider the scenario where a spatio-temporal process that contains a known (or presumed) impact of a climate event,

$$\mathbf{Z}_{Y,t} = \left(Z_{Y,t}(\mathbf{s}_1), Z_{Y,t}(\mathbf{s}_2), ..., Z_{Y,t}(\mathbf{s}_N)\right)', \quad (1)$$

is observed at a discrete set of spatial locations $\{\mathbf{s}_i \in \mathcal{D} \subset \mathbb{R}^2; i = 1, ..., N\}$ over times $t = 1, ..., T$. We are interested in quantifying the relationship between this impacted spatio-temporal variable and source and/or intermediate variables pathway variables that are also observed as spatio-temporal processes:

$$\mathbf{Z}_{k,t} = \left(Z_{k,t}(\mathbf{s}_1), Z_{k,t}(\mathbf{s}_2), ..., Z_{k,t}(\mathbf{s}_N)\right)', \tag{2}$$

$k = 1, ..., K$. We assume that these K processes are observed at the same locations and times as $\mathbf{Z}_{Y,t}$, but it is possible for this assumption to be relaxed.

We investigate the relationship between $\mathbf{Z}_{Y,t}$ and $\mathbf{Z}_{1,t}, ..., \mathbf{Z}_{K,t}$ with a two-step process:

1. First, we *model the variable relationships* by training an ESN on times $t = 1, ..., T$ to forecast $\mathbf{Z}_{Y,t}$ using $\mathbf{Z}_{1,t-\tau}, ..., \mathbf{Z}_{K,t-\tau}$ as inputs to the model, where $\tau \in \mathbb{N}$ is the forecast lead time.

2. Next, we *quantify the relationships* between the input variables and the forecasted variables using FI. For the set of forecasts at time t, we compute the importance of input variables at time t over a block of times $\{(t-\tau), (t-\tau) - 1, ..., (t-\tau) - b + 1\}$, where $b \in \mathbb{N}$ is the number of times in the block.

In the rest of this section, we provide the details under this scenario of a single layer ESN and the two proposed methods for computing spatio-temporal FI. Note that much of the notation used to define the ESN is borrowed from or influenced by McDermott and Wikle [33]. We implemented the methodologies in an R package called *listenr* [14]. The code included in *listenr* for fitting an ESN model is adapted from code provided in Wikle et al. [51].

2.1 | Single Layer Echo State Network

For each spatio-temporal process, the spatial dimensions are reduced using basis functions such that for $k = 1, ..., K$,

$$\mathbf{Z}_{Y,t} \approx \mathbf{\Phi}_Y \mathbf{y}_t \quad \text{and} \quad \mathbf{Z}_{k,t} \approx \mathbf{\Phi}_k \mathbf{x}_{k,t}, \tag{3}$$

where $\mathbf{\Phi}_Y$ is an $N \times Q$ matrix of spatial basis functions and $\mathbf{\Phi}_k$ is an $N \times P_k$ matrix of spatial basis functions. $\mathbf{y}_t$ and $\mathbf{x}_{k,t}$ are vectors of length Q and P_k, respectively, which contain the basis expansion coefficients. Q and P_k are user selected and are typically chosen to be much smaller than N. In this paper, we use principal components for the basis functions.

We create a matrix of response variables, $\mathbf{Y}$, with Q rows and T columns, where column t contains the vector of basis functions $\mathbf{y}_t$. Let $\mathbf{y}_t$ represent column t of $\mathbf{Y}$. Then let $\mathbf{X}$ be a matrix of predictor variables with P rows and T columns, and let $\mathbf{x}_t$ represent column t in $\mathbf{X}$ such that $\mathbf{x}_t = [\mathbf{x}'_{1,t}, ..., \mathbf{x}'_{K,t}]'$. Note that $P = \sum_{k=1}^{K} P_k$.

A single layer ESN consists of two levels:

$$\textit{Output stage:} \quad \mathbf{y}_t = \mathbf{V}\mathbf{h}_t + \epsilon_t \tag{4}$$

$$\textit{Hidden stage:} \quad \mathbf{h}_t = g_h\left(\frac{\nu}{|\lambda_w|}\mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\tilde{\mathbf{x}}_{t-\tau}\right). \tag{5}$$

The input variables enter the model in the hidden stage through $\tilde{\mathbf{x}}_{t-\tau}$, which is referred to as the *embedding vector* and is defined as

$$\tilde{\mathbf{x}}_{t-\tau} = \left[\mathbf{x}'_{t-\tau}, \mathbf{x}'_{t-\tau-\tau^*}, ..., \mathbf{x}'_{t-\tau-m\tau^*}\right]'. \tag{6}$$

τ^* and m are the *embedding vector lag* and *length*, respectively, which are pre-specified to determine the number of lagged inputs that are "emphasized" when computing each hidden stage. The original formulations of ESNs did not include embedding vectors (only $\mathbf{x}_t$) [18, 28], but we elect to include it in our formulation since McDermott and Wikle [33] found that an embedding vector improved spatio-temporal forecasting. Additionally, in our analyses, we always use lagged inputs, so we write the ESN model with $\tilde{\mathbf{x}}_{t-\tau}$, but other ESN model formulations [18, 28, 33] specify the embedding vector in the hidden stage to occur at time t (i.e., $\tilde{\mathbf{x}}_t$).

As McDermott and Wikle [33] point out, the hidden stage acts as "nonlinear stochastic transformation of the input vectors". The parameter matrices of $\mathbf{W}$ and $\mathbf{U}$ are referred to as *reservoir weight matrices* with dimensions of $n_h \times n_h$ and $n_h \times P(m+1)$, respectively, where n_h is the number of hidden units selected to include in the model. As a result, $\mathbf{h}_t$ is a vector of length n_h containing the *hidden units*.

The elements of $\mathbf{W}$ and $\mathbf{U}$ are randomly sampled from distributions as follows:

$$\mathbf{W}[h, c_w] = \gamma^w_{h,c_w}\text{Unif}(-a_w, a_w) + (1 - \gamma^w_{h,c_w})\delta_0, \tag{7}$$

$$\mathbf{U}[h, c_u] = \gamma^u_{h,c_u}\text{Unif}(-a_u, a_u) + (1 - \gamma^u_{h,c_u})\delta_0, \tag{8}$$

where $\mathbf{W}[h, c_w]$ represents the element row h and column c_w of $\mathbf{W}$, and similarly, $\mathbf{U}[h, c_u]$ represents the element in row h and column c_u of $\mathbf{U}$. $\gamma^w_{h,c_w} \sim Bern(\pi_w)$, $\gamma^u_{h,c_u} \sim Bern(\pi_u)$, and δ_0 is a Dirac function. The values of a_w, a_u, π_w, and π_u are pre-specified and set to small values. a_w and a_u are selected to prevent overfitting, and π_w and π_u are used to create sparse matrices.

The additional elements in the hidden stage are defined as follows:

- $\nu \in [0, 1]$ is a pre-specified scaling parameter that helps control the amount of memory in the system,
- λ_w is the spectral radius of $\mathbf{W}$, and
- g_h is a nonlinear activation function. Our implementation of an ESN uses a hyperbolic tangent function.

In the output stage, $\mathbf{V}$ is a $Q \times n_h$ parameter matrix of coefficients estimated using a ridge regression with a penalty parameter of λ_r, and $\boldsymbol{\epsilon}_t \sim Gaussian\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}\right)$. Note that the only parameters estimated in the model are $\mathbf{V}$ and σ_ϵ^2. All other parameters are randomly sampled or pre-specified, which results in the computational efficiency of the ESN model. See Lukoševičius [27] for an in depth discussion of practical recommendations for ESNs including specifying tuning parameters.

Note that there are multiple locations where regularization occurs in the model. The first place is in the basis decomposition, which captures the spatial trends but reduces the dimensions and removes noise. The second place is in the output stage with the penalty parameter in the ridge regression, which drives coefficients in $\mathbf{V}$ towards 0 when estimated. The third place regularization occurs in is the reservoir weight matrices. Both the sparsity that is induced in the matrices and the randomness in the generation of the matrices act as regularization mechanisms. These steps all help to prevent the ESN from over-fitting the in-sample data.

It is possible to extend this ESN model by adding terms to the output stage to account for more complicated relationships between $\mathbf{h}_t$ and $\mathbf{y}_t$. For example, in this paper, we incorporate a quadratic term in the output stage to mimic the *quadratic echo state network* (QESN) described in McDermott and Wikle [33]:

$$\mathbf{y}_t = \mathbf{V}_1 \mathbf{h}_t + \mathbf{V}_2 \mathbf{h}_t^2 + \boldsymbol{\epsilon}_t. \tag{9}$$

2.2 | ESN Feature Importance

Both methods that we develop for computing spatio-temporal FI for ESN models quantify "importance" through a similar concept: Determine how much model performance is affected after "adjusting" inputs at times(s) of interest in some manner. If the model performance decreases, it suggests that the input at the specified time(s) are used by the model for prediction. If the adjustment has no or little effect on the model performance, it suggests that the input at the specified time(s) are not used by the model. The larger the decrease in model performance when an input is adjusted, the larger the "importance" of the input.

With spatio-temporal data, there are various perspectives that we could consider when computing FI (e.g., blocks of time, space, or a combination). In this paper, we focus on the importance of one input spatio-temporal variable, $\mathbf{Z}_{k,t}$, over a block of times, $\{t, t-1, ..., t-b+1\}$, $b \in \mathbb{N}$, on the forecasts of the spatio-temporal response variable, $\mathbf{Z}_{Y,t+\tau}$, at time $t+\tau$, averaged over locations. However, the methodology presented here could be extended to other perspectives.

Since the ESN is trained using vectors of basis expansion coefficients, $\mathbf{x}_t$ and $\mathbf{y}_{t+\tau}$, instead of the vectors $\mathbf{Z}_{k,t}$ and $\mathbf{Z}_{Y,t+\tau}$ on the original spatial scale, we define the FI in terms of $\mathbf{x}_t$ and $\mathbf{y}_t$. Recall that in our implementations, these vectors contain principal components. We will later discuss how to use back-transformations to obtain FI on the original spatial scale. Further, note that the time indices on $\mathbf{x}$ and $\mathbf{y}$ are altered from the previous section for notational convenience.

First, let $f(\mathbf{x}_t, \mathbf{x}_{t-1}, ..., \mathbf{x}_1) = \hat{\mathbf{y}}_{t+\tau}$ represent the vector of forecasts from a trained ESN, f, at time $t+\tau$ given $\mathbf{x}_t, \mathbf{x}_{t-1}, ..., \mathbf{x}_1$. Note that while not explicitly stated, f is a function of all model parameters including $\hat{\mathbf{V}}$ and $\hat{\sigma}_\epsilon^2$. We then let

$$\mathcal{I}_{t,t+\tau}^{(k,b)} \tag{10}$$

denote the FI on the trained ESN model f for

- spatio-temporal input variable k
- over the block of times $\{t, t-1, ..., t-b+1\}$
- on the forecasts of the spatio-temporal response variable at time $t+\tau$.

We compute the FI $\mathcal{I}_{t,t+\tau}^{(k,b)}$ as follows:

1. Obtain forecasts $f(\mathbf{x}_t, \mathbf{x}_{t-1}, ..., \mathbf{x}_1) = \hat{\mathbf{y}}_{t+\tau}$ at time $t+\tau$.
2. Let $\mathcal{M}$ be a model prediction performance metric comparing observed to predicted values with the constraint that smaller values indicated better model performance (e.g., root mean squared error). Compute the performance metric on the trained model f at time $t+\tau$ as:
$$\mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}\right). \tag{11}$$
3. Generate *adjusted* forecasts using one of the following two methods:
   (a) *Permutation (stPFI)*: For replicate $r = 1, 2, ..., R$, randomly permute the values within each vector $\mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}, ..., \mathbf{x}_{k,t-b+1}$. Replace the corresponding observed values within $\mathbf{x}_t, \mathbf{x}_{t-1}, ..., \mathbf{x}_{t-b+1}$ with the permuted versions. Let the versions of $\mathbf{x}_t, \mathbf{x}_{t-1}, ..., \mathbf{x}_{t-b+1}$ containing the permuted values associated with variable k and replicate r be denoted as
$$\mathbf{x}_t^{(k,r)}, \mathbf{x}_{t-1}^{(k,r)}, ..., \mathbf{x}_{t-b+1}^{(k,r)}, \tag{12}$$
   respectively. Then obtain forecasts at time $t+\tau$ as
$$f\left(\mathbf{x}_t^{(k,r)}, \mathbf{x}_{t-1}^{(k,r)}, ...\mathbf{x}_{t-b+1}^{(k,r)}, \mathbf{x}_{t-b}, ..., \mathbf{x}_1\right) = \hat{\mathbf{y}}_{t+\tau}^{(k,b,r)}. \tag{13}$$
   The R replications are implemented to account for variability among permutations.
   (b) *Zeroing (stZFI)*: Replace the vectors of $\mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}, ..., \mathbf{x}_{k,t-b+1}$ within $\mathbf{x}_t, \mathbf{x}_{t-1}, ..., \mathbf{x}_{t-b+1}$

	$x_{1,t,1}$	…	x_{1,t,P_1}	$x_{2,t,1}$	…	x_{2,t,P_2}	…	$x_{K,t,1}$	…	x_{K,t,P_K}
$t = 1$										
$t = 2$										
$t = 3$										
$t = 4$										
$t = 5$										
…										
$t = T$										

	$y_{1,t}$	…	$y_{Q,t}$
$t = 1$			
$t = 2$			
$t = 3$			
$t = 4$			
$t = 5$			
…			
$t = T$			

FIGURE 2 Schematic showing an example of values in $\mathbf{X}'$ (left) and $\mathbf{Y}'$ (right) associated with the computation of FI $\mathcal{I}_{4,5}^{(1,3)}$.

with zeros. Let the versions of $\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-b+1}$ containing the inserted zeros associated with variable k be denoted as

$$\mathbf{x}_t^{(k)}, \mathbf{x}_{t-1}^{(k)}, \ldots, \mathbf{x}_{t-b+1}^{(k)}, \tag{14}$$

respectively. Then obtain forecasts at time $t+\tau$ as

$$f\left(\mathbf{x}_t^{(k)}, \mathbf{x}_{t-1}^{(k)}, \ldots \mathbf{x}_{t-b+1}^{(k)}, \mathbf{x}_{t-b}, \ldots, \mathbf{x}_1\right) = \hat{\mathbf{y}}_{t+\tau}^{(k,b)}. \tag{15}$$

Note that no replications are needed to account for variability with zeroing.

4. Compute the prediction performance metric on the forecasts obtained by inputting the adjusted predictions into the trained model f. That is, with stPFI compute

$$\mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}^{(k,b,r)}\right), \tag{16}$$

for $r = 1, \ldots, R$, and with stZFI compute

$$\mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}^{(k,b)}\right). \tag{17}$$

5. Finally, either compute stPFI at time $t+\tau$ as the average change in model prediction performance when inputs $\mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}, \ldots, \mathbf{x}_{k,t-b+1}$ are permuted:

$$\mathcal{I}_{t,t+\tau}^{(k,b)} = \left[\frac{1}{R}\sum_{r=1}^{R}\mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}^{(k,b,r)}\right)\right] - \mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}\right), \tag{18}$$

or stZFI at time $t+\tau$ as the change in model prediction performance when inputs $\mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}, \ldots, \mathbf{x}_{k,t-b+1}$ are set to 0:

$$\mathcal{I}_{t,t+\tau}^{(k,b)} = \mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}^{(k,b)}\right) - \mathcal{M}\left(\mathbf{y}_{t+\tau}, \hat{\mathbf{y}}_{t+\tau}\right). \tag{19}$$

As an example, let the metric $\mathcal{M}$ used to quantify the model predictive performance be the root mean squared error (RMSE). Then FI is calculated as

$$\mathcal{I}_{t,t+\tau}^{(k,b)} = \left[\frac{1}{R}\sum_{r=1}^{R} Q^{-1/2}\left\|\mathbf{y}_{t+\tau} - \hat{\mathbf{y}}_{t+\tau}^{(k,b,r)}\right\|\right] - Q^{-1/2}\left\|\mathbf{y}_{t+\tau} - \hat{\mathbf{y}}_{t+\tau}\right\|, \tag{20}$$

where $\|\cdot\|$ represents the Euclidean norm, and recall that Q is the length of $\mathbf{y}_t$ (i.e., the number of principal components retained for model training). In the case of zeroing, FI computed with RMSE reduces to

$$\mathcal{I}_{t,t+\tau}^{(k,b)} = Q^{-1/2}\left(\left\|\mathbf{y}_{t+\tau} - \hat{\mathbf{y}}_{t+\tau}^{(k,b)}\right\| - \left\|\mathbf{y}_{t+\tau} - \hat{\mathbf{y}}_{t+\tau}\right\|\right). \tag{21}$$

While FI is defined here in terms of the principal component transformed variables, we can include a back-transformation to the spatial scale as a part of the metric $\mathcal{M}$. This will allow the interpretation of the FI values to be on a more meaningful scale. Examples of where the back-transformation is included in the performance metric function are further described and implemented in Sections 3 and 4.

Figure 2 provides an example schematic display of the values in the input and output matrices associated with the computation of stPFI and stZFI. The transpose of the input matrix, $\mathbf{X}'$, is shown on the left where each row corresponds to $\mathbf{x}_t'$. Recall that $\mathbf{x}_t' = [\mathbf{x}_{1,t}', \ldots, \mathbf{x}_{K,t}']$, and let the elements of the vector $\mathbf{x}_{k,t}'$ be defined as $\mathbf{x}_{k,t}' = [x_{k,t,1}, x_{k,t,2}, \ldots, x_{k,t,P_k}]$. The matrix on the right is the transpose of the output matrix, $\mathbf{Y}'$, where each row corresponds to $\mathbf{y}_t'$. Let the elements of $\mathbf{y}_t'$ be defined as $\mathbf{y}_t = [y_{1,t}, y_{2,t}, \ldots, y_{Q,t}]'$. This example depicts the computation of

$\mathcal{I}_{4,5}^{(1,3)}$: the importance of $\mathbf{x}_{1,t}$ during the block of times $\{4, 3, 2\}$ (i.e., $\mathbf{x}_{1,4}, \mathbf{x}_{1,3}$, and $\mathbf{x}_{1,2}$) on the forecasts of $\mathbf{y}_t$ at time 5 (i.e., $\mathbf{y}_5$).

The colored cells in $\mathbf{X}'$ highlight the block of values that will be permuted/zeroed when computing stPFI/stZFI. Note that if permutation is used for computing FI, then the permutation is implemented within each row. The colored cells in $\mathbf{Y}'$ are the values that will be used in the computation of the performance metric to understand the effect of the adjustment of the input values.

3 | SIMULATION STUDY

To assess the behavior of stZFI and stPFI, we conduct a simulation study with spatio-temporal data. The goals of this simulation are to (1) assess stZFI/stPFI on features with known differing impacts on the response and (2) determine how stZFI/stPFI is affected by varying degrees of noise in the simulated data.

3.1 | Data Generating Mechanism

The data are generated on a lattice grid region of $[0, 1] \times [0, 1]$ at equally spaced locations. For $k = 1, 2$, let $\mathbf{Z}_{k,t} = (Z_{k,t}(\mathbf{s}_1), Z_{k,t}(\mathbf{s}_2), ..., Z_{k,t}(\mathbf{s}_N))'$ denote two spatially and temporally varying covariates. These covariates are simulated according to:

$$\mathbf{Z}_{k,t} = \mu_{k,t} + \rho_z \mathbf{Z}_{k,t-1} + \boldsymbol{\eta}_{k,t}, \tag{22}$$

$$\boldsymbol{\eta}_{k,t} \sim N\left(\mathbf{0}_N, \Sigma(\phi_z, \sigma_z^2)\right), \tag{23}$$

for $t = 2, 3, ..., T$, where the initial state is $\mathbf{Z}_{k,1} \sim N\left(\mu_{k,1_N}, \Sigma(\phi_z, \sigma_z^2)\right)$. The mean functions $\mu_{k,t}$ are given by:

$$\mu_{1,t} = \frac{1}{\sqrt{2\pi 6}} e^{-\frac{(t-20)^2}{2\times 6^2}}, \tag{24}$$

$$\mu_{2,t} = \frac{1}{\sqrt{2\pi 6}} e^{-\frac{(t-45)^2}{2\times 6^2}}, \tag{25}$$

for $t = 1, 2, ..., T$, which result in the mean values of the covariates varying over time and peaking at $t = 20$ and $t = 45$, respectively. The covariance function Σ is defined with a squared exponential kernel:

$$\Sigma\left(\phi, \sigma^2\right) = \sigma^2 e^{-\frac{||s_i - s_j||^2}{2\phi^2}}. \tag{26}$$

A response is simulated by:

$$Z_{Y,t}(\mathbf{s}_i) = Z_{2,t}(\mathbf{s}_i)\beta + \delta_t(\mathbf{s}_i) + \epsilon_t(\mathbf{s}_i), \tag{27}$$

where $\epsilon_t(\mathbf{s}_i) \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, $\forall\ t, i$ with $t = 1, ..., T$ and $i = 1, ..., N$. The spatio-temporal random effect $\delta_t(\mathbf{s}_i)$ is generated the same way as the covariates, letting $\boldsymbol{\delta}_t = (\delta_t(\mathbf{s}_1), \delta_t(\mathbf{s}_2), ..., \delta_t(\mathbf{s}_N))'$:

$$\boldsymbol{\delta}_t = \rho_\delta \boldsymbol{\delta}_{t-1} + \boldsymbol{\xi}_t, \tag{28}$$

$$\boldsymbol{\xi}_t \sim N\left(\mathbf{0}_N, \Sigma(\phi_\delta, \sigma_\delta^2)\right), \tag{29}$$

for $t = 2, 3, ..., T$ with initial condition $\boldsymbol{\delta}_1 \sim N\left(\mathbf{0}_N, \Sigma(\phi_\delta, \sigma_\delta^2)\right)$ with the same covariance function, Σ, as in Equation (26).

For this study, we set $\beta = 1$. Notice in Equation (27) that the first covariate $\mathbf{Z}_{1,t}$ has no effect on the response $\mathbf{Z}_{Y,t}$. Therefore, its importance should be close to zero $\forall\ t$, while importance for $\mathbf{Z}_{2,t}$ should change over time as its mean values change.

We assess the effect of noise in the data on FI by considering changes in the variance parameters of the covariates, random effect and white noise terms $(\sigma_z, \sigma_\delta, \sigma_\epsilon)$. When generating data, we set each variance parameter to either a low variability value of 0.2 or a high variability value of 4. Additionally, we adjust the block size when computing FI, where we consider block sizes of $b = 1, 2, 3$. We also consider changes in the spatial and correlation structures (i.e., $\phi_z, \phi_\delta, \rho_z, \rho_\delta$), but these results are presented in the supplemental material since their effect on FI is relatively minor.

Fifty data sets are created for each combination of parameters, and FI results are averaged over those 50. The number of time points, T, is set to 70, and the number of spatial locations, N, is set to 100, with 10 unique values in both spatial directions. Figure 3 shows an example of the spatio-temporal simulated response for one set of parameters. Figure 4 shows examples of two spatially averaged data sets: one with minimum variability and one with maximum variability.

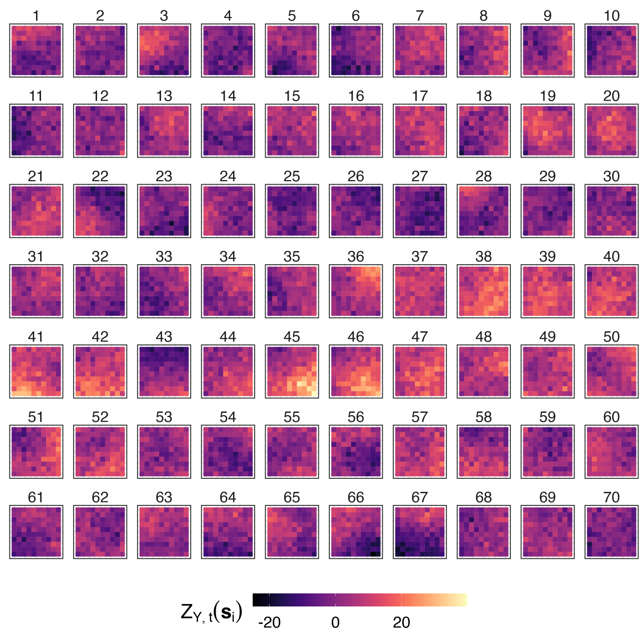

FIGURE 3 Example of a simulated spatio-temporal data set according to Equation (27).

3.2 | Models and Feature Importance

We train an ESN to predict $\mathbf{Z}_{Y,t}$ given the covariate values at a time lag of $\tau = 1$: $\mathbf{Z}_{1,t-1}$ and $\mathbf{Z}_{2,t-1}$. We first standardize the response and covariates at each location by removing the sample mean across time by location and dividing by the standard deviation across time by location. Then, following the ESN set up in Section 2.1, we perform PCA on the standardized versions of $\mathbf{Z}_{1,t}$, $\mathbf{Z}_{2,t}$, and $\mathbf{Z}_{Y,t}$, $\forall\ t$. For each time t, the first

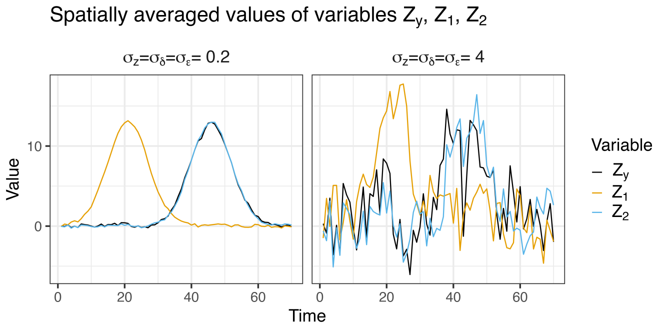

FIGURE 4 Examples of two spatially averaged synthetic data sets: one with minimum variability and one with maximum variability.

five principal components from $\mathbf{Z}_{Y,t}$ make up the output vector $\mathbf{y}_t$, as in Equation (4). The vector of inputs at time t, $\mathbf{x}_t$, is constructed using the first five principal components from $\mathbf{Z}_{1,t}$ concatenated with the first five principal components from $\mathbf{Z}_{2,t}$, which is then used to construct the embedding vector $\tilde{\mathbf{x}}_t$ in Equation (5).

The tuning parameters for the ESN are set to $\tau^* = 1, m = 1, a_w = 0.1, a_u = 0.1, \pi_w = 0.1, \pi_u = 0.1, \nu = 0.35, \lambda_r = 0.1$, and $n_h = 50$, and the ESN is trained using all times ($t = 1, ..., 70$). Since $\tau = 1, \tau^* = 1$, and $m = 1$, we are able to obtain forecasts for times $t = 3, ..., 70$.

For each model, both stPFI and stZFI are computed for the two covariates (i.e., $k = 1, 2$) and the block sizes of $b = 1, 2, 3$. The number of replications for stPFI, R, is set to 10. For the performance metric, $\mathcal{M}$, the predicted values of $\hat{\mathbf{y}}_t$ are first back-transformed to the standardized spatial scale and then RMSE is computed. That is,

$$\mathcal{M}\left(\mathbf{y}_t, \hat{\mathbf{y}}_t\right) = 100^{-1/2} \left\|\mathbf{Z}_{Y,t} - \mathbf{\Phi}_Y \hat{\mathbf{y}}_t\right\|.$$

3.3 | Simulation Results

For ease of presentation, we will refer to $\mathbf{Z}_Y$, $\mathbf{Z}_1$, and $\mathbf{Z}_2$ generically as representing the response variable, first covariate, and second covariate, respectively. Figure 5 compares stPFI and stZFI in the scenario with the maximum white noise ($\sigma_\epsilon = 4$). For all values of σ_Z and σ_δ and block sizes, both stPFI and stZFI pick up on the importance of $\mathbf{Z}_2$ at the correct times, with peak importance around its mode at $t = 45$. Changes in the variability and block size affect the FI values, but the signals are clear in the mean FI. When the block size is low, there are more fluctuations in importance values that appear to be noise, especially when variability is high. The increased block size appears to remediate this issue because it helps reduce the autocorrelation information available to the ESN.

For $\mathbf{Z}_1$, there are cases where both stPFI and stZFI indicate that $\mathbf{Z}_1$ is important when $\mathbf{Z}_1$ has its mode near $t = 20$, but this importance is spurious since $\mathbf{Z}_1$ has no effect on the response. This is true for different levels of variability and for different block sizes. stZFI appears to have less of an issue with this, especially at a block size of three. We include a discussion on a probable cause for the detection of this spurious relationship and possible next steps to reduce this occurrence in Section 5.

Figure 6 shows a closer view of the effect of block size on stZFI. This figure highlights that not only is stZFI smoother when the block size is increased from one to three, but the signal is also more pronounced. The increase in magnitude should be expected since as the block size increases, more of the feature times are set to zero, so the difference in RMSEs should increase, at least until the autocorrelation is removed. Thus, even with significant noise, stZFI with a large enough block-size clearly captures the importance of covariate $\mathbf{Z}_2$ while mostly not indicating any spurious importance of $\mathbf{Z}_1$.

This simulation study builds confidence in the FI approach presented in Section 2.2. Results show that stZFI is able to correctly identify the importance of $\mathbf{Z}_2$ while being relatively unaffected by $\mathbf{Z}_1$ when the block size parameter is sufficiently large, and these results translate across varying noise levels. stPFI has similarly strong performance identifying $\mathbf{Z}_2$ but is more susceptible to detecting importance in a variable that has no direct impact on the response. Additional figures of simulation results are provided in the supplemental material. We also note that this simulated data example has a response variable that is linearly related to the covariates. In future work, it would be beneficial to consider data simulated with a more complex non-linear relationship.

4 | CLIMATE DATA APPLICATION

On June 12, 1991, Mount Pinatubo erupted in the Philippines. The eruption released 18-19 Tg of SO_2 into the atmosphere [15], which had a profound impact on the climate. The massive injection of aerosols into the atmosphere led to increases in AOD [46, 15], which in turn led to changes in stratospheric temperatures and surface temperatures [46, 25] (and references therein).

When aerosols enter the atmosphere, they either scatter sunlight, which leads to atmospheric cooling, or absorb the sunlight, which leads to warming [25, 40]. Due to the relationship between atmospheric aerosols and temperatures, the artificial injection of aerosols has been discussed as a potential mitigation to the current climate change trends. This was foreshadowed by Kiehl and Briegleb [19] in 1993 who noted that summer sulfate aerosol forcings offset greenhouse forcings in the eastern US and central Europe. However, there is great uncertainty in how such an intervention would affect the broader climate system.

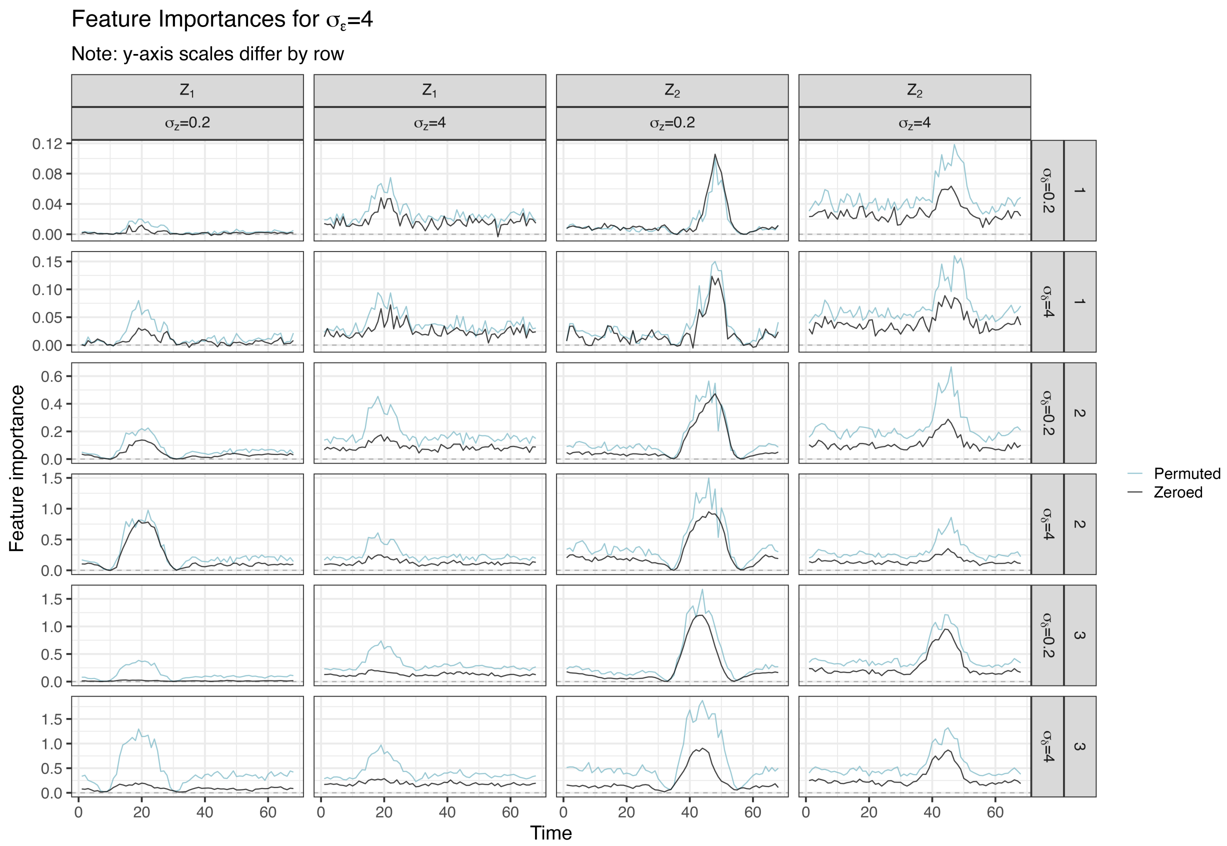

FIGURE 5 Results from simulation study comparing stPFI and stZFI, with high level of white noise ($\sigma_\epsilon = 4$). First two columns show FI for $\mathbf{Z}_1$, second two show FI for $\mathbf{Z}_2$. Varying combinations of σ_δ and block size are given in the rows. True data generating mechanism is given by Equation (27).

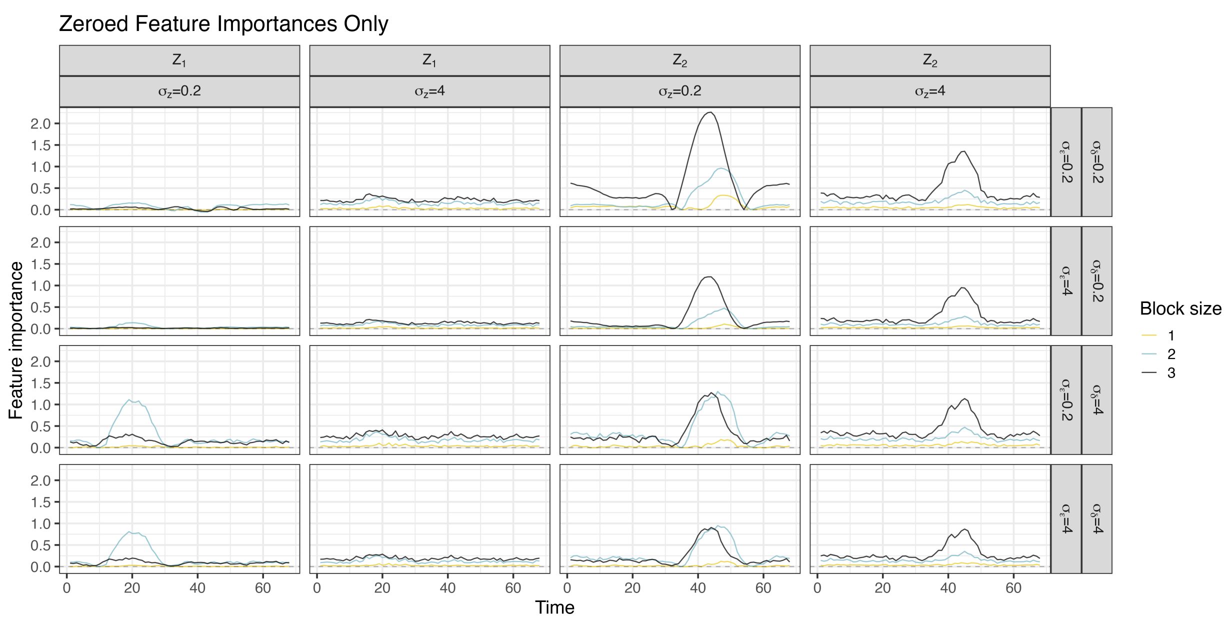

FIGURE 6 Results from simulation study comparing number of blocks for stZFI. First two columns show FI for $\mathbf{Z}_1$, second two show FI for $\mathbf{Z}_2$. Varying combinations of σ_δ and σ_ϵ are given in the rows. True data generating mechanism is given by Equation (27).

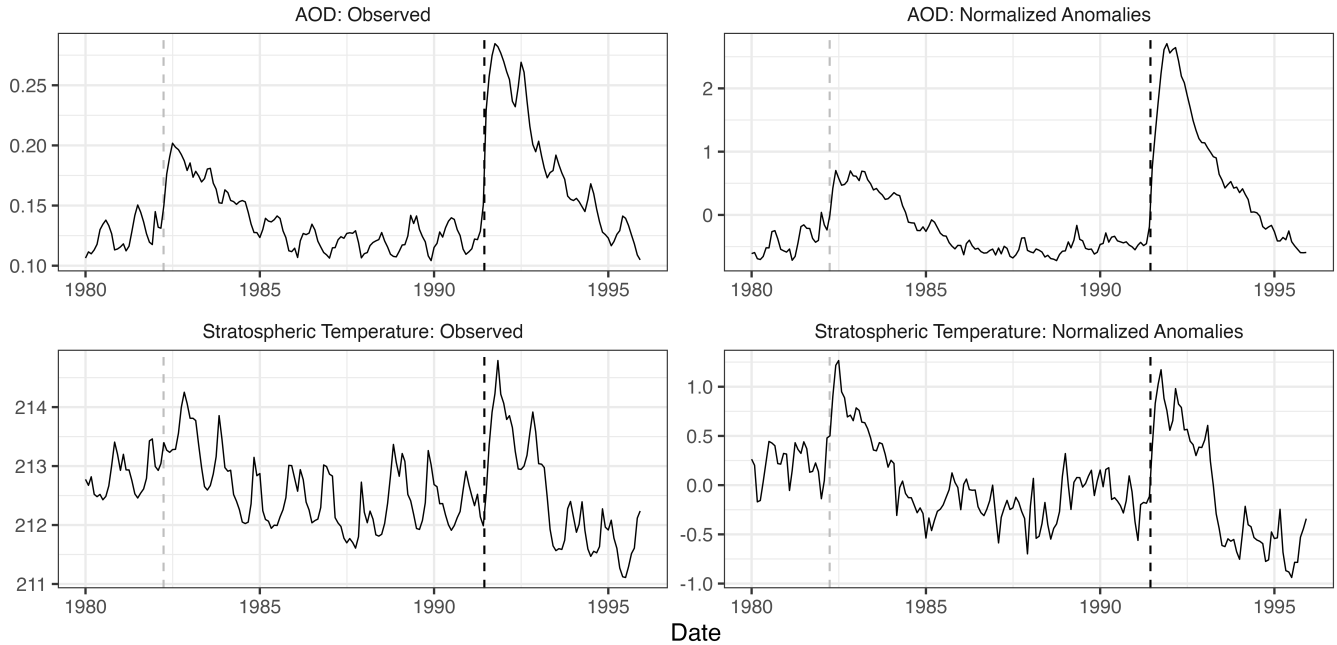

FIGURE 7 MERRA-2 global weighted averages of stratospheric temperature and AOD for each month from 1980 through 1995. Normalized anomalies remove average monthly effect by spatial location. The gray vertical dashed line indicates the eruption of El Chichón. The black vertical dashed line indicates the eruption of Mount Pinatubo.

We aim to explore the effects of anthropogenic forcings of aerosols into the atmosphere with an ESN and feature importance to quantify the impacts on the climate system. We use the 1991 Mount Pinatubo eruption as a proxy for an anthropogenic injection of aerosols. We focus on the relationship between the climate pathway variables of AOD and stratospheric temperatures, but future work could explore the inclusion of additional pathway variables such as SO_2 and surface temperatures.

4.1 | Data and Model

We analyze this event using the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) [12]. The monthly values from 1980-1995 for stratospheric temperature at 50 mb [34] and vertically integrated AOD [35] were manually downloaded from the Goddard Earth Sciences Data and Information Services Center [41]. Detailed information on AOD from MERRA-2 is provided in Randles et al. [42].

The data for AOD and stratospheric temperature are on a 24×48 equally spaced latitude and longitude lattice (Figure 1). We consider the time range of 1980-1995 because it provides climate information before the eruption of Mount Pinatubo and includes a second climate event: the 1982 volcanic eruption of El Chichón in southeast Mexico. The aerosol injection from El Chichón was smaller than Mount Pinatubo with 7.5 Tg of SO_2 injected into the atmosphere [21].

We model the relationship between AOD and stratospheric temperature by treating temporally lagged values of AOD and temperature as inputs to the model and predict stratosphere temperature one month ahead. That is, let $\mathbf{s}_i,\ i = 1, 2, ..., N$ be unique latitude/longitude locations. In a slight change of notation, let $Z_{k,mth,yr}(\mathbf{s}_i)$ be the raw, observed value of variable k at location $\mathbf{s}_i$, month mth, and year yr. Let $k = 1$ denote AOD and $k = 2$ denote stratospheric temperature. The model output of stratospheric temperature one month ahead is represented by $Z_{Y,mth+1,yr}(\mathbf{s}_i)$.

Temperature and, to a lesser extent, AOD exhibit strong seasonality. We remove the seasonality through preprocessing by computing monthly *normalized anomalies*. This computation subtracts the mean monthly effect and divides by the standard deviation of the monthly effect. That is, the normalized anomalies are calculated according to:

$$Z^{anom}_{k,mth,yr}(\mathbf{s}_i) = \frac{Z_{k,mth,yr}(\mathbf{s}_i) - \bar{Z}_{k,mth,\cdot}(\mathbf{s}_i)}{sd(Z_{k,mth,\cdot}(\mathbf{s}_i))}, \tag{30}$$

where $\bar{Z}_{k,mth,\cdot}(\mathbf{s}_i)$ is the average temperature at location $\mathbf{s}_i$ during month mth, for variable k, and $sd(Z_{k,mth,\cdot}(\mathbf{s}_i))$ is the standard deviation of temperatures at location $\mathbf{s}_i$ during month mth. Normalized anomalies for $Z_{Y,mth+1,yr}(\mathbf{s}_i)$ are equivalent to the normalized anomalies of $Z_{2,mth,yr}(\mathbf{s}_i)$.

Figure 7 shows global monthly weighted averages of stratospheric temperatures and aerosol optical depth (AOD) from 1980 through 1995. The averages in the left column are computed from the observed values, and the averages in the right column are computed from the normalized anomalies. The effects of Mount Pinatubo and El Chichón are clear in both variables: increases in AOD result in increases in stratospheric temperature immediately following the eruptions. This trend is expected due to the reflection of the sun's energy [23].

The weights used to compute the averages in Figure 7 are included to mitigate the effect of polar regions. Since the data are on an equally spaced lattice, there are relatively more locations towards the poles compared to the equator. The poles tend to be more variable temperature-wise, so the weighting helps to avoid over accounting for activity in the poles. Taking the square root of the cosine of the latitude has been shown to be a good way of weighting latitudes [17]. That is, the weight associated with location $\mathbf{s_i}$ is

$$w_{\mathbf{s}_i} = \sqrt{cos\left(latitude(\mathbf{s}_i) \times \frac{\pi}{180}\right)}, \tag{31}$$

where $latitude(\mathbf{s}_i)$ returns the latitude of location $\mathbf{s}_i$ in degrees.

Since there is randomness in the $\mathbf{W}$ and $\mathbf{U}$ matrices in an ESN, we train 25 ESNs on the MERRA-2 data with all other elements of the models held constant. The models are trained using data from all years (1980-1995). The model inputs and outputs are principal components of AOD and stratospheric temperature normalized anomalies. Here, we use the first five principal components from both variables as an example where the spatial dimensions are greatly reduced (i.e., 1,152 locations reduced to 5 principal components). The ESN embedding vector is constructed using $m = 5$, $\tau = 1$, $\tau^* = 1$. This results in an ESN that predicts stratospheric temperatures one month ahead using AOD and stratospheric temperatures from the previous six months.

The other ESN hyperparameters are set to $a_w = 0.1$, $a_u = 0.1$, $\pi_w = 0.1$, $\pi_u = 0.1$, $\nu = 0.2$, $\lambda_r = 5$, and $n_h = 1000$. These values are chosen based on predictive performance in a hyperparameter search when the years of 1994 and 1995 are held out as test data. Details on the selection of the hyperparameter values are included in the supplemental material along with a predictive performance comparison to ESNs trained with 10 principal components. A key finding in the hyperparameter analysis is that the hyperparameters with the biggest effect on predictive performance for this application are a_u, ν, n_h, and λ_r. The other hyperparameters have little to no effect.

4.2 | Model Fit

In order to believe that the FI captures meaningful dynamics in the data, it is important that the model fits the data well (i.e., the model captures meaningful dynamics in the data). We further ensure the model with the selected hyperparameters fits the data well by implementing a time series blocked training/test split evaluation. That is, an ESN with the selected hyperparameters up to a certain year. Then the model is used to predict values for the remaining years. This process is repeated for various train/test splits. For each train/test split, 25 ESNs are fit again to account for variability due to the randomness in the $\mathbf{W}$ and $\mathbf{U}$ matrices.

The prediction performance of the models is assessed using weighted root mean squared errors computed at a time and average over all spatial locations. As with the computation of the weighted averages in Figure 7 , the weighting mitigates the effect of poles. Let $\hat{Z}^{anom}_{Y,mth,yr}(\mathbf{s}_i)$ be the model's prediction of stratospheric temperature at location $\mathbf{s}_i$, month mth, year yr, which is back-transformed from the principal component scale. Then the weighted RMSE is calculated by:

$$RSE_{mth,yr}(\mathbf{s}_i) = \sqrt{\left(Z^{anom}_{Y,mth,yr}(\mathbf{s}_i) - \hat{Z}^{anom}_{Y,mth,yr}(\mathbf{s}_i)\right)^2},$$

$$Weighted\ RMSE_{mth,yr} = \frac{\sum_{i=1}^{N} w_{\mathbf{s}_i} RSE_{mth,yr}(\mathbf{s}_i)}{\sum_{i=1}^{N} w_{\mathbf{s}_i}},$$

where $w_{\mathbf{s}_i}$ is computed as in Equation 31.

Figure 8 shows the weighted RMSEs from the time series blocked training/test splits. The splits are created such that the models are trained through even numbered years, and the rows in the visualization identify which years the models are trained through. Results including models trained through odd numbered years are included in the supplemental material. The Mount Pinatubo eruption is denoted by the vertical dashed line. Overall, there is relatively little variability in RMSEs across ESNs within a time point. The most variability due to the random matrices is seen in the test RMSEs in the years immediately after the Mount Pinatubo eruption.

The results from the training/test splits show that when the model is trained with limited data (e.g., trained through 1982 and 1984), RMSEs on the test set are large throughout. When enough training data are used (e.g., trained through 1988 and 1990), test RMSEs before the Mount Pinatubo eruption are relatively similar to the training RMSEs. However, in these cases, the test RMSEs in the years following the eruption of Mount Pinatubo are poor. This is unsurprising since the eruption is a change in the climate mechanism. When the ESNs are trained through 1992 or beyond, test RMSEs closely resemble training RMSEs. These results suggest that the ESNs with the selected set of hyperparameters fit the data well when the training data contain values through at least 1992. Therefore, this provides evidence that the models trained on data from 1980-1995 fit the data well.

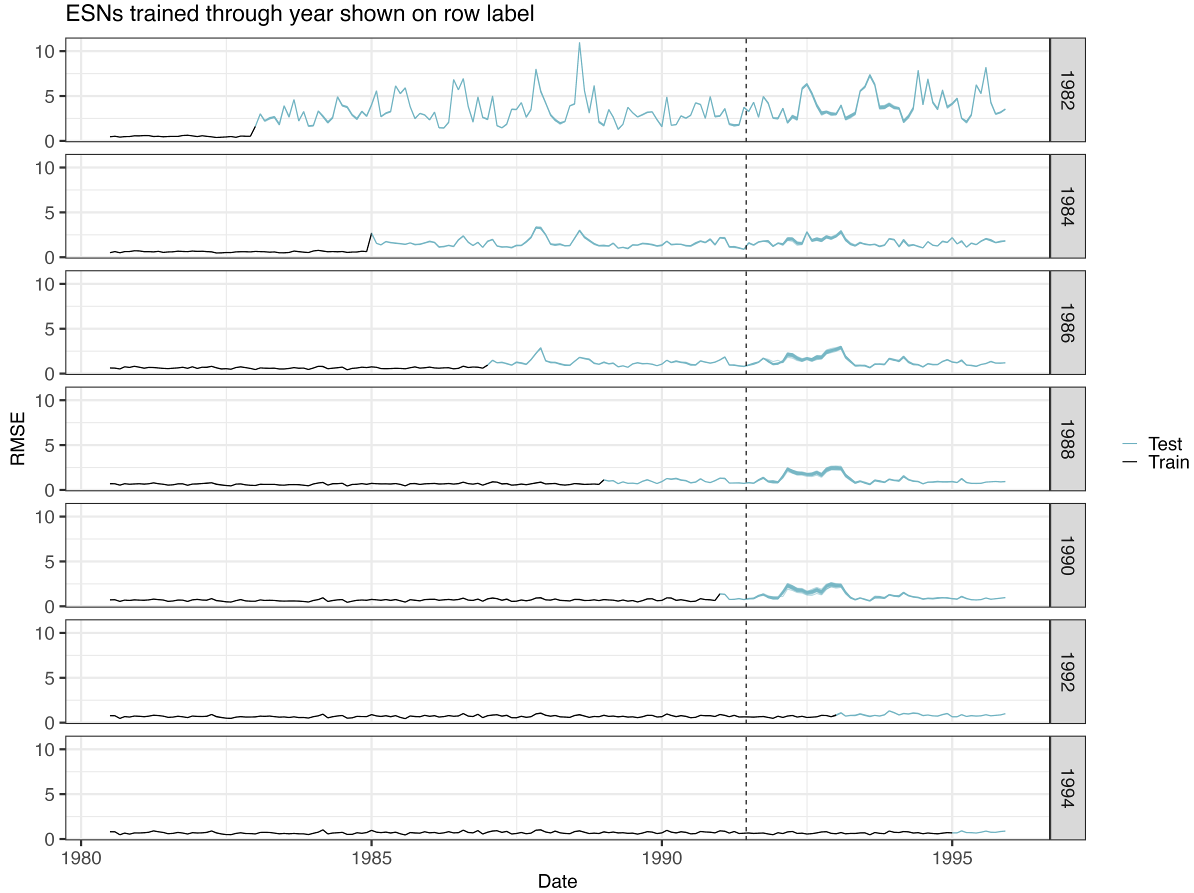

FIGURE 8 Weighted RMSEs from the ESN blocked training/test splits on MERRA-2 data. Each row represents an additional two years of training data. Mount Pinatubo eruption is denoted by the vertical dashed line.

4.3 | Feature Importance

FI is calculated on the full training data since we are ultimately interested in variable relationships, not forecasting. We opt to use the weighted RMSE from Equation 31 as the metric for FI for the same reasons as the time series blocked model assessment. We compute FI with a block size of six to correspond with our specified value of m that highlights the past six months in the embedding vector. A comparison of FI computed with different block sizes is included in the supplemental material, which shows that the FI magnitudes tend to increase as the block size increases. Lastly, we assess the sensitivity of stPFI and stZFI to the random matrices of $\mathbf{W}$ and $\mathbf{U}$ in the ESN by computing both FIs on the 25 ESNs trained with different random matrices.

Figure 9 shows stPFI and stZFI on the MERRA-2 data computed with a block size of six. The solid black line represents the average FI at a time across the 25 ESNs. The grey lines represent the FI values from individual ESNs. There is more variability in stPFI within a time across models than stZFI. This trend is likely due to the random permutation of values from stPFI. However, overall, there is relatively little variability in FI at a time across ESNs. Thus, the trends in FI are not sensitive to the randomness in the ESNs for this example. An additional analysis of the sensitivity of FI to the random matrices for ESNs fit with different tuning parameters is included in the supplemental material. In that case study, a model with less hidden units has more FI variability across individual ESNs, but the trend across individual ESNs is still clear and consistent with the trend in Figure 9 .

The two vertical dashed lines in Figure 9 show the eruptions of El Chichón and Mount Pinatubo. The effects of Mount Pinatubo are clear for AOD, as both stPFI and stZFI see a large spike in importance. These spikes indicate that the importance of the previous six months of global AOD values for forecasting global stratospheric temperatures one month ahead, on average, greatly increases following the eruption of Mount Pinatubo. Interestingly, the spike appears to decrease more quickly with stZFI than stPFI. For lagged stratospheric temperatures, there also is an increase in importance after the Mount Pinatubo eruption, but the change is not as dramatic

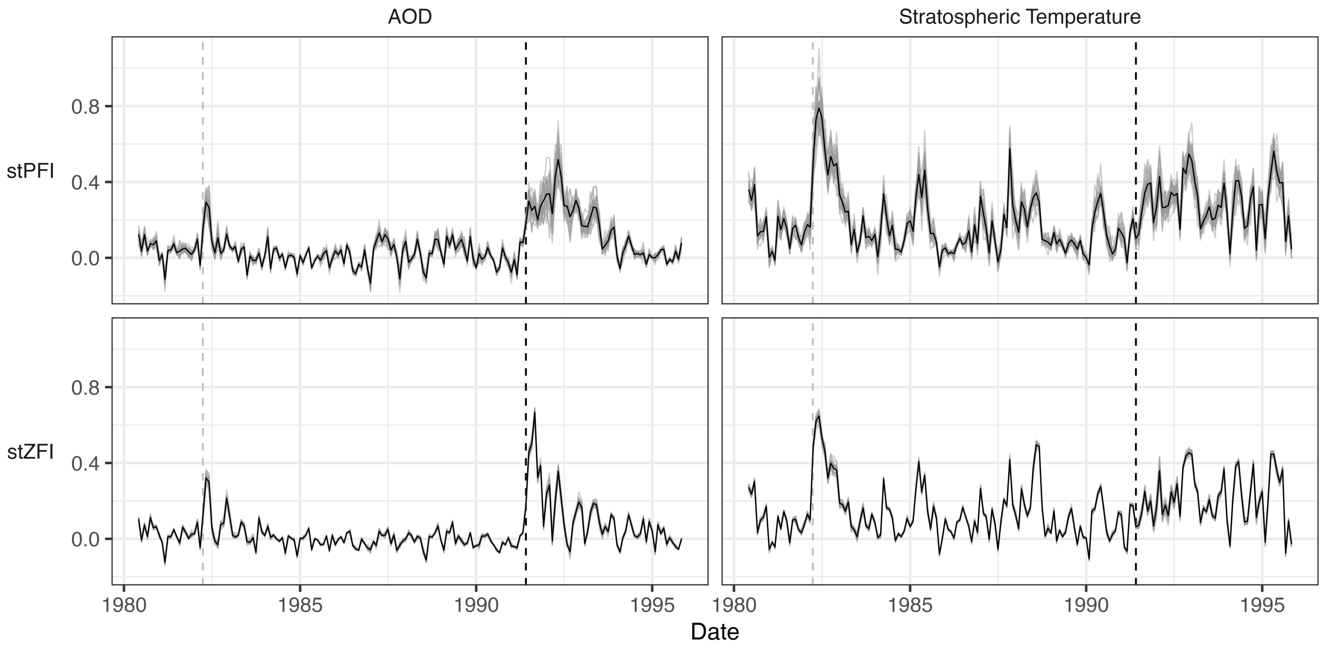

FIGURE 9 Feature importances on MERRA-2 data with a block size of 6. The gray vertical dashed line indicates the eruption of El Chichón, and the black vertical dashed line indicates the eruption of Mount Pinatubo.

as with AOD. This result suggests that standard temperature fluctuations are not sufficient for explaining the changes in temperature. These two pieces provide evidence, but not proof, that the impact on temperature due to the volcanic eruption and its subsequent injection of aerosols can be traced through AOD.

stZFI and stPFI also capture the effect from El Chichón in AOD and stratospheric temperature. However, temperature stands out as more important than AOD for El Chichón, which is different than the FI trends around the Mount Pinatubo eruption. This difference could be the result of the smaller aerosol injection from El Chichón. Although the effects of El Chichón and Mount Pinatubo have been well studied, this proof of concept showcases the methodology's ability to recapture known effects.

5 | CONCLUSIONS AND DISCUSSION

In this paper, we propose the use of ESNs for characterizing climate pathways (i.e., quantifying relationships between climate variables related to a climate event). We do this by modeling spatio-temporal climate pathway variables associated with a climate event using an ESN and quantifying the variable relationships using FI.

ESNs are a computationally efficient model that are able to capture patterns in complex systems, which makes them a desirable tool for applications with the complex climate system containing large quantities of data. In order to provide transparency to the black-box ESN, we develop two FI techniques (stPFI and stZFI) for spatio-temporal data that are applied to the ESN in order to quantify the variable relationships captured by the ESN. Both FI techniques approach the quantification of variable importance by adjusting (permuting or setting to zero) a block of times associated with a climate variable of interest and measure how this adjustment affects the model forecast performance at a specified time. By visualizing the resulting FIs, we depict how the importance of input variables on the forecast variable changes over time and compare the importance values to other input variables.

We demonstrate our approach on MERRA-2 reanalysis climate data that include two volcanic eruption events (El Chichón in 1982 and Mount Pinatubo in 1991). The eruptions act as proxies of anthropogenic stratospheric aerosol injections. We consider the relationships between the pathway variables of AOD and stratospheric temperature. The ESN FI results show that the importance of AOD on forecasting stratospheric temperature greatly increases after both eruptions, which provides support for AOD being a part of the climate pathway that leads to the effects seen in stratospheric temperature. These results agree with previous climate science research indicating that the increase in AOD due to the eruptions led to an effect on the temperature, which supports the capabilities of the method.

Future work in this area could take many forms such as a deeper analysis of the Mount Pinatubo pathway application, technical extensions to the proposed methods, and the comparison of methods to other machine learning models and explainability methods. We start by addressing extended analyses of the Mount Pinatubo application. Additional relevant variables associated with the eruption of Mount Pinatubo include SO_2 and surface temperature. These variables could be included as more inputs to the model, or in some scenarios, it may be of interest to consider the joint forecasting of multiple variables (e.g., surface and stratospheric temperatures). When computing FI, it may also be meaningful to consider the importance of groups of climate variables (e.g., AOD and SO_2). By grouping variables when computing FI, the results could be interpreted as a joint effect. This grouping approach could be especially useful with highly correlated pathway variables. More broadly, an application of the methodology to other volcanic eruptions would be beneficial for the validation of the method's ability to capture important variable connections.

We next consider technical extensions. In regards to ESNs, we consider a single-layer ESN in this paper, but the methodology could easily be extended to other ESN variations. Previous work develops ESNs with multiple layers for capturing temporal trends on different time scales, ensembles of ESNs, and Bayesian implementations [33]. We allude to the idea of an ensemble ESN in this paper by fitting multiple ESNs with different random matrices in the MERRA2 application. A next step would be to aggregate the predictions over the individual ESNs in order to obtain ensembled predictions and feature importances. An ensemble of ESNs would also provide a manner in which to quantify uncertainty on predictions, as addressed by [33], and potentially, feature importance.

In regards to FI, we compute FI as an average over locations, so that it reduces to a measure of importance over time. Another option would be to adjust the computation of FI to a measure of importance that is associated with each location (or specified regions such as latitudinal bands) at a time/block of times. This approach would allow for identification of how importance not only changes over time but also over space.

Further development of FI techniques for spatio-temporal data could also include steps that better account for correlation in the data. In the simulation study, both stPFI and stZFI pick up on the variable known to be related to the response ($\mathbf{Z}_2$). However, stPFI always detects the spurious variable ($\mathbf{Z}_1$). stZFI is less likely to pick up on the spurious variable when applied with larger block sizes, but and in some cases, stZFI still incorrectly identifies this relationship. These results are likely due to correlation between the covariates. It has been suggested that permutation based FI methods produce biased results when correlation is present and not accounted for [16]. This bias is due to the permutation leading to observations that occur outside of the observed training data, which forces the model to extrapolate and inflates the variable importance. It seems reasonable that this same phenomenon could occur when 'zeroing' the data but to a lesser extent. This could explain why stZFI is less affected than stPFI. Hooker et al. [16] suggest overcoming this issue by either retraining the model on the permuted data or developing a conditional FI. These ideas would be interesting to explore in the case of spatio-temporal FI, where the existence of correlation is essentially guaranteed.

The last avenue we touch on for future work is the comparison of the methodology to other machine learning techniques. As in many fields, analyses using machine learning are more commonly appearing in the climate literature. Previous work modeling climate data with deep learning includes convolutional neural networks (CNNs) (e.g., Mamalakis et al. [31]) and Bayesian neural networks (BNNs) (e.g., Clare et al. [8]). A comparison between ESNs and other neural networks on the MERRA-2 application would provide insight from several perspectives. ESNs offer improved computational efficiency over other neural networks. Comparing the predictive performance on the MERRA-2 data bewteen ESNs and other neural networks would help to determine if predictive accuracy is sacrificed for computational efficiency with ESNs. Even if predictive accuracy is sacrificed, it is possible that ESNs capture trends in the data that are similar to the trends captured by the more complex models. A better understanding of the trade-offs between ESNs and other neural networks would help to provide insight into which models are suited for different scenarios based on the analysis objectives.

From the explainability perspective, there have been many techniques proposed for computing feature importance including the commonly used techniques of LIME [43] and SHapley Additive exPlanation (SHAP) value [29]. The development of methodology and tools for adapting these techniques for ESNs is in early stages. However, recent examples include the development of Layer-wise Relevance Propagation (LRP) for ESNs by Landt-Hayen et al. [24], the application of LIME to ESNs in a medical application by Bouazizi and Ltifi [6], and the application of SHAP to ESNs in an aeronautics application by Baptista et al. [5].

A comparison of explainability techniques would be highly beneficial. If various methods produce similar explanations, it provides support that the explanations are correctly capturing the trends used by a model for prediction. If the methods provide different explanations, it raises concerns about the trustworthiness of the explanations and encourages further assessment of the explanation techniques. As the number of proposed techniques in the explainability literature increase, papers on the cautions and unexpected behaviors from explanation techniques including LIME and SHAP also increase (e.g., [3], [13], [22], [37], [20], and [44]). The authors highlight that explanations often disagree across different methods or different settings from the same method. Therefore, assessments should be done to understand which methods best capture how a model is using data for prediction. This is a challenging problem since the ground truth relationships captured by black-box models are, by definition, unknown. An example of how some researchers compared explanation techniques for CNNs in the climate context is provided by Mamalakis et al. [30]. Further, [49] provide a literature survey on techniques proposed for the evaluation of explanations.

As the possibility of the implementation of climate mitigation strategies becomes more of a reality, the importance of the development of algorithmic tools for understanding how such actions could affect other aspects of the climate increases. We approach this task by quantifying pathway variable relationships using FI computed on an ESN trained over a time period surrounding the event of interest. However, just as the ESN is an approximation to the workings of the climate system, FI is an approximation to the workings of the ESN. Further development of explainability techniques for spatio-temporal data that provide different perspectives on black-box models or the development of machine learning models for spatio-temporal data with directly interpretable parameters could lend more credibility to the use of machine learning models in such high-stakes applications.

ACKNOWLEDGEMENTS

The authors thank Lyndsay Shand, Gabriel Huerta, and J. Derek Tucker for their thoughtful suggestions during development stages. Additionally, we thank Gabriel Huerta for his careful read through and feedback on the paper.

FUNDING STATEMENT

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SANDXXXXXX.

SUPPORTING INFORMATION

All data and code associated with this manuscript are available at https://github.com/sandialabs/listenr/tree/main/examples/CoDA-paper. Additional supporting information can be found online in the Supporting Information section at the end of this article.

6 | Bibliography

[1] 2022: *2022 National Defense Strategy, Nuclear Posture Review, and Missile Defense Review*.
URL https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF

[2] Alao, O., P. Y. Lu, and M. Soljačić, 2021: Discovering Dynamical Parameters by Interpreting Echo State Networks. *NeurIPS 2021 AI for Science Workshop*.
URL https://openreview.net/forum?id=coaSxusdBLX

[3] Alvarez-Melis, D. and T. S. Jaakkola, 2018: On the Robustness of Interpretability Methods. *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*.
URL https://arxiv.org/abs/1806.08049

[4] Arrieta, A. B., S. Gil-Lopez, I. Laña, M. N. Bilbao, and J. D. Ser, 2022: On the post-hoc explainability of deep echo state networks for time series forecasting, image and video classification. *Neural Computing and Applications*, **34**, no. 13, 10257–10277, doi:10.1007/s00521-021-06359-y.

[5] Baptista, M. L., K. Goebel, and E. M. Henriques, 2022: Relation between prognostics predictor evaluation metrics and local interpretability shap values. *Artificial Intelligence*, **306**, 103667.

[6] Bouazizi, S. and H. Ltifi, 2023: Novel diversified echo state network for improved accuracy and explainability of eeg-based stroke prediction. *Information Systems*, 102317.

[7] Breiman, L., 2001: Random Forests. *Machine Learning*, **45**, no. 1, 5–32, doi:10.1023/a:1010933404324.

[8] Clare, M. C. A., M. Sonnewald, R. Lguensat, J. Deshayes, and V. Balaji, 2022: Explainable Artificial Intelligence for Bayesian Neural Networks: Toward Trustworthy Predictions of Ocean Dynamics. *Journal of Advances in Modeling Earth Systems*, **14**, no. 11, doi:10.1029/2022ms003162.

[9] Doshi-Velez, F. and B. Kim, 2017: Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[10] Eastham, S., S. Doherty, D. Keith, J. H. Richter, and L. Xia, 2021: Improving models for solar climate intervention research. *EOS*, **109**, doi:10.1029/2021eo156087. URL https://eos.org/science-updates/improving-models-for-solar-climate-intervention-research

[11] Fisher, A., C. Rudin, and F. Dominici, 2019: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, **20**, 1–81.
URL http://jmlr.org/papers/v20/18-760.html

[12] Gelaro, R., W. McCarty, M. J. Suarez, R. Todling, A. Molod, L. Takacs, C. A. Randles, A. Darmenov, M. Bosilovich, R. Reichle, K. Wargan, L. Coy, R. Cullather, C. Draper, S. Akella, V. Buchard, A. Conaty, A. M. da Silva, W. Gu, G.-K. Kim, R. Koster, R. Lucchesi, D. Merkova, J. E. Nielsen, G. Partyka, S. Pawson, W. Putman, M. Rienecker, S. D. Schubert, M. Sienkiewicz, and B. Zhao, 2017: The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of Climate*, **30**, 5419–5454, doi:10.1175/JCLI-D-16-0758.1.

[13] Goode, K. and H. Hofmann, 2021: Visual diagnostics of an explainer model: Tools for the assessment of lime explanations. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **14**, no. 2, 185–200.

[14] Goode, K., D. Ries, and K. McClernon, 2024: *listenr: Explainability for Echo State Networks*. R package version 0.4.0.
URL https://github.com/sandialabs/listenr

[15] Guo, S., G. J. Bluth, W. I. Rose, M. Watson, and A. Prata, 2004: Re-evaluation of so_2 release of the 15 june 1991 pinatubo eruption using ultraviolet and infrared satellite sensors. *Geochemistry, Geophysics, Geosystems*, **5**, 1–31, doi:10.1029/2003GC000654.

[16] Hooker, G., L. Mentch, and S. Zhou, 2021: Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, **31**, 1–16.

[17] Huth, R., 2006: The effect of various methodological options on the detection of leading modes of sea level pressure variability. *Tellus A: Dynamic Meteorology and Oceanography*, **58**, no. 1, 121–130, doi:10.1111/j.1600-0870.2006.00158.x.
URL https://doi.org/10.1111/j.1600-0870.2006.00158.x

[18] Jaeger, H., 2001: The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, **148**, no. 34, 13.

[19] Kiehl, J. T. and B. P. Briegleb, 1993: The relative roles of sulfate aerosols and greenhouse gases in climate forcing. *Science*, **260**, no. 5106, 311–314, doi:10.1126/science.260.5106.311.
URL https://www.science.org/doi/abs/10.1126/science.260.5106.311

[20] Krishna, S., T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, 2022: The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

[21] Krueger, A., N. Krotkov, and S. Carn, 2008: El chichon: The genesis of volcanic sulfur dioxide monitoring from space. *Journal of Volcanology and Geothermal Research*, **175**, no. 4, 408–414, doi:https://doi.org/10.1016/j.jvolgeores.2008.02.026, the 25th Anniversary of the El Chichòn's Eruption.
URL https://www.sciencedirect.com/science/article/pii/S0377027308001625

[22] Kumar, I. E., S. Venkatasubramanian, C. Scheidegger, and S. Friedler, 2020: Problems with shapley-value-based explanations as feature importance measures. *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., PMLR, volume 119 of *Proceedings of Machine Learning Research*, 5491–5500.
URL https://proceedings.mlr.press/v119/kumar20e.html

[23] Labitzke, K. and M. McCormick, 1992: Stratospheric temperature increases due to pinatubo aerosols. *Geophysical Research Letters*, **19**, 207–210, doi:10.1029/91GL02940.

[24] Landt-Hayen, M., P. Kröger, M. Claus, and W. Rath, 2022: Layer-wise relevance propagation for echo state networks applied to earth system variability. *Signal, Image Processing and Embedded Systems Trends*, D. C. Wyld, ed., ARRAY(0x55588c8d8680), number 20 in Computer Science & Information Technology (CS & IT): Conference Proceedings, 115–130.
URL https://doi.org/10.5121/csit.2022.122008

[25] Liepert, B. and I. Tegen, 2002: Multidecadal solar radiation trends in the united states and germany and direct tropospheric aerosol forcing. *Journal of Geophysical Research: Atmospheres*, **107**, no. D12, AAC 7–1–AAC

7–15, doi:https://doi.org/10.1029/2001JD000760.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001JD000760

[26] Lipton, Z. C., 2018: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, **16**, no. 3, 31–57.

[27] Lukoševičius, M., 2012: Neural Networks: Tricks of the Trade - A Practical Guide to Applying Echo State Networks. *Lecture Notes in Computer Science*, 659–686, doi:10.1007/978-3-642-35289-8_36.

[28] Lukoševičius, M. and H. Jaeger, 2009: Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, **3**, no. 3, 127–149, doi:https://doi.org/10.1016/j.cosrev.2009.03.005.
URL https://www.sciencedirect.com/science/article/pii/S1574013709000173

[29] Lundberg, S. M. and S.-I. Lee, 2017: A unified approach to interpreting model predictions. *Advances in neural information processing systems*, **30**.

[30] Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2022: Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artificial Intelligence for the Earth Systems*, **1**, no. 4, e220012.

[31] Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2022: Explainable Artificial Intelligence in Meteorology and Climate Science: Model Fine-Tuning, Calibrating Trust and Learning New Science. *Lecture Notes in Computer Science*, 315–339, doi:10.1007/978-3-031-04083-2_16.

[32] McDermott, P. L. and C. K. Wikle, 2017: An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. *Stat*, **6**, no. 1, 315–330, doi:https://doi.org/10.1002/sta4.160.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.160

[33] — 2019: Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. *Environmetrics*, **30**, no. 3, doi:10.1002/env.2553.

[34] Modeling, G., G. E. S. D. Assimilation Office (GMAO), and I. S. C. DISC), 2015: *MERRA-2 instM_3d_asm_Np: 3d, monthly mean, instantaneous, pressure-level, assimilation, assimilated meteorological fields v5.12.4*. https://disc.gsfc.nasa.gov/datasets/M2IMNPASM_5.12.4/summary, accessed May 2022.

[35] — 2015: *MERRA-2 tavgM_2d_aer_Nx: 2d, monthly mean, time-averaged, single-level, assimilation, aerosol diagnostics v5.12.4*. https://disc.gsfc.nasa.gov/datasets/M2TMNXAER_5.12.4/summary, accessed April 2022.

[36] Molnar, C., G. Casalicchio, and B. Bischl, 2021: Interpretable machine learning–a brief history, state-of-the-art and challenges. *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings*, Springer, 417–431.

[37] Molnar, C., G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl, 2022: *General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models*, Springer International Publishing, Cham. 39–68.
URL https://doi.org/10.1007/978-3-031-04083-2_4

[38] Mothilal, R. K., A. Sharma, and C. Tan, 2020: Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 607–617.

[39] Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, 2019: Definitions, methods, and applications in interpretable machine learning. *PNAS*, **116**, 22071–22080, doi:10.1073/pnas.1900654116.

[40] Myhre, G., C. E. Myhre, B. H. Samset, and T. Storelvmo, 2013: Aerosols and their relation to global climate and climate security. *Nature Education Knowledge*, **4**, no. 5.

[41] NASA, 2023: *Ges disc*. https://disc.gsfc.nasa.gov/.

[42] Randles, C. A., A. M. da Silva, P. Colarco, A. Darmenov, R. Govindaraju, A. Smirnov, B. Holben, R. Ferrare, J. Hair, Y. Shinozuka, and C. Flynn, 2017: The merra-2 aerosol reanalysis, 1980 onward. part i: System description and data assimilation evaluation. *Journal of Climate*, **30**, 6823–6850, doi:10.1175/JCLI-D-16-0609.1.

[43] Ribeiro, M. T., S. Singh, and C. Guestrin, 2016: " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

[44] Roy, S., G. Laberge, B. Roy, F. Khomh, A. Nikanjam, and S. Mondal, 2022: Why don't xai techniques agree? characterizing the disagreements between post-hoc

explanations of defect predictions. *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 444–448.
URL https://api.semanticscholar.org/CorpusID:254903647

[45] Rudin, C., 2019: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, **1**, no. 5, 206–215.

[46] Sato, M., J. E. Hansen, M. P. McCormick, and J. B. Pollack, 1993: Stratospheric aerosol optical depths, 1850-1990. *Journal of Geophysical Research: Atmospheres*, **98**, no. D12, 22987–22994, doi:https://doi.org/10.1029/93JD02553.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/93JD02553

[47] Sood, A. and M. Craven, 2021: Feature Importance Explanations for Temporal Black-Box Models. *arXiv*, doi:10.48550/arxiv.2102.11934.

[48] Urbanek, S., 2008: Visualizing Trees and Forests. *Handbook of Data Visualization*, C.-h. Chen, W. Härdle, and A. Unwin, Eds., Springer-Verlag, Berlin, Germany, volume 3, 243–266.
URL https://haralick.org/DV/Handbook_of_Data_Visualization.pdf

[49] Vilone, G. and L. Longo, 2021: Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, **76**, 89–106.

[50] Watts, J., 2020: China plans rapid expansion of ‘weather modification‘ efforts.
URL https://www.theguardian.com/world/2020/dec/03/china-vows-to-boost-weather-modification-capabilities

[51] Wikle, C. K., A. Zammit-Mangion, and N. Cressie, 2019: *Spatio-Temporal Statistics With R*. CRC Press, Boca Raton, FL.

How to cite this article: Goode K., D. Ries K. McClernon, 2023, Characterizing climate pathways using feature importance on echo state networks, *Stat Anal Data Min: The ASA Data Sci Journal*, *volume, number and page*.