



# Using feature importance as an exploratory data analysis tool on Earth system models

Daniel Ries, Katherine Goode, Kellie McClernon, and Benjamin Hillman

Sandia National Laboratories, Albuquerque, NM, United States of America

**Correspondence:** Daniel Ries (dries@sandia.gov)

Received: 10 July 2024 – Discussion started: 23 September 2024

Revised: 16 December 2024 – Accepted: 2 January 2025 – Published: 24 February 2025

**Abstract.** Machine learning (ML) models are commonly used to generate predictions, but these models can also support the discovery of new science. Generating accurate predictions necessitates that a model captures the structure of the underlying data. If the structure is properly extracted, ML could be a useful exploratory and evidential tool. In this paper, we present a case study that demonstrates the use of ML for exploratory data analysis (EDA) in the climate space. We apply the ML explainability method of spatiotemporal zeroed feature importance (stZFI) to understand how climate-variable associations evolve over space and time. Our analyses focus on data from ensembles of Earth system models (ESMs) which provide data on different climate states and conditions. We elect to work with ESM ensembles since they allow us to compare feature importance across alternative scenarios not available with observed data. The ensembles also account for natural variability so that we can distinguish between signal and noise due to natural climate variability when computing feature importance. The use of perturbed initial condition ensembles introduces variability mimicking the natural variability in the atmosphere; thus the signals emerging using feature importance (FI) can be evaluated against the natural variability in the climate system. For our analyses, we consider the 1991 volcanic eruption of Mount Pinatubo, which was a large stratospheric aerosol injection. We explore the climate pathway associated with the eruption from aerosols to radiation to temperature at both the near-surface and stratospheric levels. In addition to applying the method to data generated from two different ESMs, we apply stZFI to reanalysis data to compare the associations identified by stZFI. We show how stZFI tracks the importance of aerosol optical depth over time on forecasting temperatures. This case study illustrates usefulness of an ML tool (stZFI) for EDA on a well-studied climate exemplar.

## 1 Introduction

Climate science questions are often studied using ensembles of Earth system models (ESMs). Since we cannot conduct global controlled-climate experiments to understand cause and effect, ESMs allow climate scientists to explore the effects of different climate conditions on the climate system. However, ESMs generate large quantities of data (considering the number of ensemble members, spatial resolution, temporal resolution, etc.) which can be difficult to process and understand. Therefore, methods that summarize and identify trends are valuable for working with data from ESMs. Exploratory data analysis (EDA) is the general approach to exploring, analyzing, and summarizing patterns in data. EDA includes the computation of simple summary statistics such as means, standard deviations, and correlations, along with data visualizations. These approaches provide a high-level view of trends but can overlook important details. More sophisticated EDA techniques allow scientists and practitioners to understand detailed trends in the data, which promotes the ability to draw conclusions and propose new hypotheses. Our objective in this paper is to present a case study showing the utility of a new EDA technique that leverages the data-driven modeling approach of machine learning (ML) and ESMs to gain insights into climate problems.

ESMs provide a mathematical representation of the complex and chaotic nature of Earth's climate. When these models are run with different initial states, parameter values, external forcings, and so on, the models produce an ensemble of simulations that represent different possible climate scenarios. The ensemble members provide an estimate of model and natural variability given a particular forcing and enable investigation of possible climate outcomes that are not re-

alized in the observational record. From a ML perspective, these ensemble members represent a set of temporal “replicates”, where the true underlying relationships are known. In contrast, observational climate data only provide a single instance in space–time. Observed data are also limited to events in the past, which may not include all events that researchers and policymakers are interested in studying. For example, no major stratospheric aerosol injection experiment has been conducted in the field, but we are still interested in the impact that such a scenario would produce. ESMs provide a way to understand variable relationships not seen in observational data and capture natural climate variability.

There is a rich history of using ML and statistical models for analyses with ensembles of ESMs. Examples of these analyses include Tebaldi et al. (2005) and Smith et al. (2009), who built statistical models to quantify the uncertainty in replicates from different ESMs. Going in the other direction, many approaches have been developed which use observational data to calibrate climate model replicates (e.g., Reichler and Kim, 2008; Armour et al., 2013, and Baker et al., 2016). More recently, ML models are used to provide insight into climate processes by quantifying relationships between ESM variables (e.g., Hart et al., 2023; de Burgh-Day and Leeuwenburg, 2023). McClernon et al. (2024) considered the assessment of ML models used in such situations and developed a cross-validation procedure using ESM ensemble members to obtain a true “replicate hold-out” set to assess the commonly used “repeated hold-out” cross-validation process for time series data (Cerqueira et al., 2020). Replicate hold-out uses many independently generated full time series from the same ESM for the same period, with one series chosen as the training set and another as the testing set, while repeated hold-out splits single timer series into training and testing sets. Notably for replicate hold-out, the train and test set cover the same time span, whereas in repeated hold-out, the test set always covers the future relative to the training set.

While ML models are commonly used for prediction applications, their ability goes beyond simple prediction. Data-driven models are capable of finding new patterns and verifying known relationships. As Toms et al. (2020) point out, “the ultimate objective of using a neural network can also be the interpretation of what the network has learned rather than the output itself”. However, many ML models are black-box algorithms whose mathematical formulas are too complex for interpreting variable relationships captured by the model. Explainability methods applied to ML models provide a link from the predictive power of the ML model to an understanding of the underlying processes. Goode et al. (2024) defined a model as being “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.

The climate science community has recently recognized the utility of explainable ML methods. To find variables that best discover model errors in an ESM, Silva et al. (2022)

use an explainable ML method, namely the SHapley Additive exPlanation (SHAP) values. Toms et al. (2020) use backward optimization and layer-wise relevance propagation to discover scientifically meaningful connections with respect to an El Niño–Southern Oscillation (ENSO) phase detection and prediction. McGovern et al. (2019) provide an overview of potential explainability for ML applied to meteorology. Clare et al. (2022) apply Bayesian neural networks with layer-wise relevance propagation (LRP) and SHAP values to better characterize and quantify ocean circulation dynamics. On the ESMs ARISE-SAI, Mamalakis et al. (2023) explore the impacts of stratospheric aerosol injections on different variables with the explainability method of Deep SHAP (Lundberg and Lee, 2017).

Explainability techniques possess the potential for providing insight into patterns in data captured by a black-box model, but research has also identified pitfalls with current methods (e.g., Rudin, 2019; Hooker et al., 2021; Ancona et al., 2018). Mamalakis et al. (2022) compare different convolutional neural network explainability methods by utilizing synthetic data so the “true explanations” are known a priori. Their analysis highlights the strengths and weaknesses of various methods, and they conclude “that no optimal method exists for all prediction settings”. They recommend applying and comparing results from various explainability methods, while more rigorous assessments of explainability techniques are needed. We believe the case study in this paper will contribute to this body of understanding.

In this paper, we present a case study that demonstrates the explainability technique of spatiotemporal zeroed feature importance (stZFI; Goode et al., 2024) as an EDA tool for a climate problem that leverages Earth system model (ESM) ensembles. Goode et al. (2024) developed stZFI for echo state networks (ESNs), which are computationally fast, yet powerful, ML models for spatiotemporal data. stZFI measures the relative gain in predictive performance for each input, or predictor, variable over time. This allows users to see how “important” input variables are for the predictive ability of the ML model and provides insight into the dynamic nature of the relationships. We use stZFI to explore relationships between pathway variables associated with a stratospheric aerosol injection climate event. We apply stZFI to ensemble members of ESMs, which allows us to measure uncertainties in the variable importances that effectively account for variability. Our analyses are intended to showcase the applicability of stZFI as an exploratory and evidential tool for climate-related problems.

## 1.1 Motivating application

Stratospheric aerosol injection (SAI) is being studied as a possible way of mitigating climate change (Irvine et al., 2016), but there is concern over its potential side effects (MacMartin et al., 2016; McCormack et al., 2016). Although there is an abundance of computer model experiments look-

ing at SAI (Ferraro et al., 2015; Banerjee et al., 2021; Bednarz et al., 2022), we are unaware of any physical SAI experiments. In lieu of SAI experiments, the 1991 eruption of Mount Pinatubo provides a natural exemplar of a large SAI event. The eruption released 18–19 Tg of sulfur dioxide into the atmosphere, causing changes to aerosol optical depth (AOD), transporting partially through the Brewer–Dobson circulation (Butchart, 2014), and consequently changes to stratospheric temperatures (Sato et al., 1993; Guo et al., 2004). The increase in AOD scatters short-wave radiation (Twomey, 1991) and absorbs and re-emits long-wave radiation (Zhou and Savijärvi, 2014). The increase in short-wave scattering tends to cool the Earth’s surface by reflecting more incoming solar radiation, while the increase in long-wave absorption tends to warm the lower stratosphere. As a consequence of Mount Pinatubo’s eruption, temperatures at pressure levels of 30 (T030) to 50 mb (T050) rose between 2.5 and 3.5°C (Labitzke and McCormick, 1992), while temperatures at the surface (T2M) decreased by 0.5°C (Parker et al., 1996). Figure 8 shows how T050 and T2M propagate over this time.

We purposely examine the well-studied eruption of Mount Pinatubo since our goal is to demonstrate the usefulness of stZFI as an EDA tool. By picking a well-known event and phenomenon, we can compare relationships identified by stZFI with previously identified relationships in the scientific literature. An agreement in identified relationships could provide confidence in the proposed approach. We apply stZFI to data generated from a simplified ESM and a fully coupled (i.e., allowing for interactions between atmosphere, land, and ocean) ESM. We additionally consider one reanalysis dataset (i.e., the combination of observed and model data) to demonstrate the ability of stZFI to find interesting relationships and quantify how they evolve over time. When used as an EDA tool, these analyses show how stZFI can be a useful way to understand the complex relationships in the data.

The remainder of this article is structured as follows. Section 2 introduces the ML model and explainability method used in this paper, the ensemble echo state network (EESN) and stZFI, respectively. Section 3 introduces the datasets, namely HSW-V v1.0 (Hollowed et al., 2024b), the Energy Exascale Earth System Model (E3SM) (Rasch et al., 2019; Golaz et al., 2022), and the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) (Gelaro et al., 2017), reanalysis. This section additionally quantifies the variable relationship using stZFI for each dataset. Section 4 makes comparisons between E3SM and MERRA-2 results. Finally, Sect. 5 discusses results, conclusions, and future directions.

## 2 Data model

This section reviews the EESN and stZFI approaches used to measure climate-variable relationships. The EESN and stZFI

assume data are centered and scaled prior to model training to improve model performance and make interpretations of importances easier. The data throughout this section are assumed to be centered and scaled according to a preprocessing procedure of the modeler’s choice.

### 2.1 Ensemble echo state network

Echo state networks (ESNs) (Jaeger, 2001; Lukoševičius and Jaeger, 2009) are known to provide good predictions for chaotic systems (Alao et al., 2021). ESNs are also computationally efficient in comparison to recurrent neural networks, which represent their sibling ML model for temporal data. The ESN applied to the spatiotemporal climate data was first explored by McDermott and Wikle (2017) and improved upon in McDermott and Wikle (2019). We follow the notation of Goode et al. (2024) and McClernon et al. (2024), since it allows for an easier presentation of feature importance (FI) in the next section. Let

$$\mathbf{Z}_{Y,t} = (Z_{Y,t}(s_1), Z_{Y,t}(s_2), \dots, Z_{Y,t}(s_N))' \quad (1)$$

be the vector of preprocessed responses at locations  $\{s_i \in \mathcal{D} \subset \mathbb{R}^2; i = 1, \dots, N\}$  over times  $t = 1, \dots, T$ . The preprocessed model inputs are also spatiotemporal processes, represented as

$$\mathbf{Z}_{k,t} = (Z_{k,t}(s_1), Z_{k,t}(s_2), \dots, Z_{k,t}(s_N))' \quad (2)$$

for  $k = 1, \dots, K$ . The locations of all variables are assumed to be the same, which is expected for climate model simulations. To reduce spatial dimensionality, we use empirical orthogonal functions (EOFs) to decompose the variable anomalies such that

$$\mathbf{Z}_{Y,t} \approx \Phi_Y \mathbf{y}_t, \quad (3)$$

$$\mathbf{Z}_{k,t} \approx \Phi_k \mathbf{x}_{k,t} \quad (4)$$

for  $k = 1, \dots, K$ , where  $\Phi_Y$  is an  $N \times Q$  matrix of EOFs corresponding to  $\mathbf{Z}_{Y,t}$ , and  $\Phi_k$  is an  $N \times P_k$  matrix of EOFs corresponding to  $\mathbf{Z}_{k,t}$ .  $\mathbf{y}_t$  and  $\mathbf{x}_{k,t}$  are vectors of the lengths  $Q$  and  $P_k$ , respectively, which are the scores from the EOF decomposition.  $Q$  and  $P_k$  are user-chosen hyperparameters corresponding to the number of EOFs for the output and  $k$ th input, respectively. These values can be chosen using hyperparameter tuning or considering computational complexity. Without any loss of generality, we will assume  $P_1 = P_2 = \dots = P_k = P$  throughout.

McDermott and Wikle (2019) introduced the EESN as a way to quantify uncertainty and improve predictions by averaging over different initializations of the reservoir. The

EESN is given by the following:

$$\text{Output stage: } \mathbf{y}_t = \mathbf{V}^{(r)} \mathbf{h}_t + \boldsymbol{\epsilon}_t^{(r)}, \quad (5)$$

$$\text{Hidden stage: } \mathbf{h}_t = g_h \left( \frac{\nu}{|\lambda_w|} \mathbf{W}^{(r)} \mathbf{h}_{t-\tau} + \mathbf{U}^{(r)} \tilde{\mathbf{x}}_{t-\tau} \right), \quad (6)$$

$$\text{Regression error: } \boldsymbol{\epsilon}_t^{(r)} \sim \text{Gaussian}(\mathbf{0}, \sigma_\epsilon^{2(r)} \mathbf{I}), \quad (7)$$

where  $r = 1, 2, \dots, R$  represents the EESN ensemble member. We omit the quadratic term included by McDermott and Wikle (2019). During our initial tuning of the models, we found increasing the size of the EESN was more beneficial than including additional terms, although this could be treated as part of model selection. Input variables are included in the embedding vector,  $\tilde{\mathbf{x}}_{t-\tau}$ , which is defined by

$$\tilde{\mathbf{x}}_{t-\tau} = [\mathbf{x}'_{t-\tau}, \mathbf{x}'_{t-\tau-\tau^*}, \dots, \mathbf{x}'_{t-\tau-m\tau^*}]'. \quad (8)$$

$\tau$  is the forecast period (i.e., how many steps ahead in time the EESN will make predictions) and should be chosen based on the goals of the modeler.  $\tau^*$  is the embedding vector lag, and  $m$  is the number of embedding lags. Both are pre-specified and can be selected either during hyperparameter tuning or based on subject matter expertise.

$\mathbf{h}_t$  contains the  $n_h$  hidden units which include information on the inputs beyond the immediate past, where  $n_h$  is a tuning parameter. The matrices  $\mathbf{W}^{(r)}$  and  $\mathbf{U}^{(r)}$  contain the reservoir weights with dimensions of  $n_h \times n_h$  and  $n_h \times P(m+1)$ , respectively.  $\mathbf{W}^{(r)}$  and  $\mathbf{U}^{(r)}$  are not estimated but rather are randomly sampled  $R$  times from their respective distributions as follows:

$$\mathbf{W}^{(r)}[h, c_w] = \gamma_{h,c_w}^w \text{Unif}(-a_w, a_w) + (1 - \gamma_{h,c_w}^w) \delta_0, \quad (9)$$

$$\mathbf{U}^{(r)}[h, c_u] = \gamma_{h,c_u}^u \text{Unif}(-a_u, a_u) + (1 - \gamma_{h,c_u}^u) \delta_0, \quad (10)$$

where  $\mathbf{W}^{(r)}[h, c_w]$  represents the element row  $h$  and column  $c_w$  of  $\mathbf{W}^{(r)}$ , and similarly,  $\mathbf{U}^{(r)}[h, c_u]$  represents the element in row  $h$  and column  $c_u$  of  $\mathbf{U}^{(r)}$ .  $\gamma_{h,c_w}^w \sim \text{Bern}(\pi_w)$ ,  $\gamma_{h,c_u}^u \sim \text{Bern}(\pi_u)$ , and  $\delta_0$  is a Dirac function. Sampling multiple times from reservoir distributions allows us to have a distribution of predictions over which to average and calculate uncertainty.  $a_w, a_u, \pi_w$ , and  $\pi_u$  serve as regularization hyperparameters to prevent overfitting.  $\nu$  is a value in  $[0, 1]$  that helps control the amount of memory in the system through  $\mathbf{h}_t$ .  $\lambda_w$  is the spectral radius of  $\mathbf{W}$ .  $g_h$  is a nonlinear activation function for which we use a hyperbolic tangent function. The only parameters estimated in the model are contained in the matrix  $\mathbf{V}^{(r)}$  and the error term  $\sigma_\epsilon^{2(r)}$ .  $\mathbf{V}^{(r)}$  is a  $Q \times n_h$  parameter matrix of coefficients estimated using a ridge regression with a penalty parameter of  $\lambda_r$ . This ridge regression adds another layer of regularization to the model to prevent overfitting. Lukoševičius (2012) and Goode et al. (2024) provide recommendations and results for tuning ESNs.

## 2.2 Feature importance

Goode et al. (2024) introduced stZFI as a feature importance (FI) metric for assessing variable importance and its evolution over time for spatiotemporal data. Feature importance is a quantitative measure of how important an input variable is for accurately predicting an output variable at a particular time. stZFI provides a quantitative measure of importance for an input variable over time by measuring the increase in a predictive metric when the variable is removed at each time. Goode et al. (2024) computed stZFI for individual ESNs. We adjust the approach for an ensemble of ESNs. In particular, we compute stZFI using the ensemble prediction (i.e., the average of the predictions produced by each ensemble member in the EESN). This is in contrast to an approach that computes stZFI for each member of the EESN and average the stZFI results across ensembles, which is less in line with how EESNs would be used in practice to obtain the “final” prediction.

### 2.2.1 stZFI global metric

stZFI measures the importance of input variables over a block of times,  $\{t, t-1, \dots, t-b+1\}$ ,  $b \in \mathbb{N}$ , on the forecasts of the spatiotemporal output variable, at time  $t + \tau$ , averaged over locations. To simplify the notation, and without a loss of generality, we assume  $\tau^* = 1$ . Let  $f^{(r)}(\mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{x}_1) = \hat{\mathbf{y}}_{t+\tau}^{(r)}$  represent the vector of forecasts from the  $r$ th ensemble member of the EESN, at time  $t + \tau$ , given  $\mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{x}_1$ , and let  $\bar{\mathbf{y}}_{t+\tau} = \frac{1}{R} \sum_{r=1}^R \hat{\mathbf{y}}_{t+\tau}^{(r)}$  be the aggregated forecast from the EESN. Define the root mean squared error (RMSE) on the spatial scale as

$$\text{RMSE}_{t+\tau} = Q^{-1/2} \|\mathbf{Z}_{Y,t+\tau} - \Phi_Y \bar{\mathbf{y}}_{t+\tau}\|. \quad (11)$$

The procedure for stZFI also computes an “zeroed” RMSE,  $\text{RMSE}_{t+\tau}^{*(k)}$ , that captures model predictive performance when zeroing all EOF scores for an input feature  $k$ . First, replace the vectors of  $\mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}, \dots, \mathbf{x}_{k,t-b+1}$  within  $\mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-b+1}$  with zeros. Denote these as  $\mathbf{x}_t^{(k)}, \mathbf{x}_{t-1}^{(k)}, \dots, \mathbf{x}_{t-b+1}^{(k)}$ , respectively. Next, compute forecasts using the zeroed inputs as

$$f^{(r)}(\mathbf{x}_t^{(k)}, \mathbf{x}_{t-1}^{(k)}, \dots, \mathbf{x}_{t-b+1}^{(k)}, \mathbf{x}_{t-b}, \dots, \mathbf{x}_1) = \hat{\mathbf{y}}_{t+\tau}^{(r,k,b)}. \quad (12)$$

Similarly, let  $\bar{\mathbf{y}}_{t+\tau}^{(k,b)} = \frac{1}{R} \sum_{r=1}^R \hat{\mathbf{y}}_{t+\tau}^{(r,k,b)}$  be the aggregate forecast from the EESN when variable  $k$  is zeroed at time  $t$  with block size  $b$ . The RMSE after zeroing relevant variable  $k$  EOF scores on the spatial scale is

$$\text{RMSE}_{t+\tau}^{*(k)} = Q^{-1/2} \|\mathbf{Z}_{Y,t+\tau} - \Phi_Y \bar{\mathbf{y}}_{t+\tau}^{(k,b)}\|. \quad (13)$$

The stZFI metric for forecasting time  $t + \tau$ , for variable  $k$ , and for block size  $b$  is computed as

$$\mathcal{I}_{t,t+\tau}^{(k,b)} = \text{RMSE}_{t+\tau}^{*(k)} - \text{RMSE}_{t+\tau}. \quad (14)$$



Larger values of the stZFI mean variable  $k$  are relatively more important to the model for making predictions and therefore important for describing the pathway from input to output. Values of stZFI near zero imply that the variable has little impact on predictions and, therefore, does not show a strong association with the output variable. The values of the feature importance metric itself are reductions in predictive RMSE and can be interpreted as such. For example, a feature importance of  $\mathcal{I}_{t,t+\tau}^{(k,b)} = 2$  for variable  $k$  at time  $t$  means the RMSE increases by two units with variable  $k$  at time  $t$  removed from the model.

### 2.2.2 stZFI regional metric

The stZFI metric in Eq. (14) is a “global” metric; it measures the impact an input variable has on an output variable on a globally averaged scale. The metric can be decomposed regionally by calculating the contributions to  $\mathcal{I}_{t,t+\tau}^{(k,b)}$  by spatial regions such as latitude bands. Regional contributions to stZFI provide the ability to quantify the impact of a global input variable on a regional output variable. In this paper, we only consider latitude bands for regional contributions to stZFI, so we incorporate this in our notation, but more generally, other spatial regions could be considered.

Let the regional feature importance metric be represented by

$$\mathcal{I}_{t,t+\tau}^{(k,b)}[\text{lat}] = \text{RMSE}_{t+\tau}^{*(k)}[\text{lat}] - \text{RMSE}_{t+\tau}[\text{lat}], \quad (15)$$

where lat represents all locations  $s_i$  in the latitude band or lat for output variables and where lat indicates the average measure across all locations  $s_i$  in the defined latitude band. We focus on latitudinal bands since they account for the most variation in surface and stratospheric temperatures. The first RMSE on the right-hand side of Eq. (15) is the RMSE for the latitude band, lat, when the variable  $k$  is zeroed globally as follows:

$$\text{RMSE}_{t+\tau}^{*(k)}[\text{lat}] = Q^{-1/2} \|\mathbf{Z}_{Y,t+\tau}[\text{lat}] - \Phi_Y \bar{\mathbf{y}}_{t+\tau}^{(k,b)}[\text{lat}]\|. \quad (16)$$

The second RMSE in Eq. (15) is the RMSE for all locations  $s_i$  with latitudes equal to lat, much like the RMSE in Eq. (11), except it only considers data with latitude equal to lat as follows:

$$\text{RMSE}_{t+\tau}[\text{lat}] = Q^{-1/2} \|\mathbf{Z}_{Y,t+\tau}[\text{lat}] - \Phi_Y \bar{\mathbf{y}}_{t+\tau}[\text{lat}]\|. \quad (17)$$

## 3 Stratospheric aerosol injection applications

We apply stZFI to data from two climate models (HSW-V v1.0 and E3SM; Sect. 3.1 and 3.2, respectively) and one reanalysis dataset (MERRA-2; Sect. 3.3). We consider these three data sources in order to compare the behavior of feature importance across different techniques for data acquisition associated with the same SAI climate event. HSW-V v1.0 is a simplified climate model, while E3SM is a fully

coupled model. HSW-V v1.0 and E3SM have counterfactual runs which allow us to compute stZFI when no major injection of aerosols occurs. MERRA-2 gives us a representation of the observed climate and allows us to compare reanalysis values to values generated by ESMs. Details on each dataset, data preprocessing, and FI results are presented in this section.

First, standardization is performed to ensure all input variables in the EESN are on the same scale, such that feature importances are comparable between variables. Specifics for each dataset are described within each data subsection. For ease of illustration and comparison, we trained EESNs on all datasets with the same  $\tau$ ,  $\tau^*$ ,  $m$ , and  $b$  values. We use  $\tau = 1$  since we are interested in relatively short-lead forecasting, and we set  $m = 3$  and  $\tau^* = 1$  as an example of a model where the emphasis is placed on the past quarter of a year for a prediction. For the stZFI block size, we elect to use  $b = 4$  since, with HSW-V v1.0, E3SM, and MERRA-2 data for the Pinatubo eruption application, we have found that block sizes larger than this often do not change results much, potentially indicating the sufficient removal of auto-correlation in the importance. Several other values were kept constant across datasets. We set  $R = 10$  and use the first 20 EOFs from each climate variable for training the EESN. We will use  $R = 10$  ensemble members for all EESNs to balance the computational complexity of stZFI and predictive performance. We use 20 EOFs for all variables in this paper to keep comparisons simple and fair. We select 20 EOFs for computational convenience, but this value could also be tuned as part of hyperparameter optimization. The remaining EESN hyperparameters were optimized using a grid search. The procedure was implemented separately for each dataset. The data were split into training and test sets, and the hyperparameter set giving the lowest test set predictive performance was used to compute the feature importances presented in this section. Details on the EESN hyperparameter selection and tuning process is in Appendix A. A predictive assessment of the EESNs for each dataset with their best-performing hyperparameters is provided in Appendix B.

### 3.1 HSW-V v1.0

We first consider a simplified ESM with a single stratospheric injection of aerosols referred to as HSW-V v1.0 (Hollowed et al., 2024b). The primary use for this simplified climate model is to verify that stZFI finds relationships built into the ESM that are less likely to be entangled with higher-order effects. This makes it easier to verify how the method behaves in an intuitive and predictable scenario. Held and Suarez (1994) created an idealized forcing without topography and seasonality; this is combined with the modified temperature equation of Williamson et al. (1998), which allows modeling stratospheric temperatures that are necessary when considering an SAI. HSW-V v1.0 further modifies the temperature equation by making adjustments based on the observed

aerosols at a given pressure level. HSW-V v1.0 runs approximately 168 times faster than E3SM (McClernon et al., 2024), making it useful for initial model evaluations.

The aerosol injection is meant to resemble Mount Pinatubo's eruption in size, space, and time. Model outputs are remapped to a  $2^\circ \times 2^\circ$  structured latitude/longitude grid with 72 vertical levels up to 0.1 mb/ $\sim$  60 km. The temporal resolution of the output is 48 h. The simulations begin at day 0 and run for 1200 d, with the injection on day 179. There is no seasonality in the model, and the background radiation is prescribed to be in balance. Aerosols come only from the single Pinatubo-like injection of the sulfate precursor and volcanic ash, meaning that AOD is fully driven by the prescribed injection. Surface and stratospheric temperatures are then parameterized through AOD. We build two models predicting temperature for HSW-V v1.0.

- *HSW-V v1.0 stratosphere model.* It predicts T050 (temperature at 50 mb), given the AOD and T050.
- *HSW-V v1.0 surface model.* It predicts T1000 (temperature at 1000 mb), given the AOD and T1000.

All input variables are time-lagged. An ensemble of simulations with the HSW-V v1.0 configuration is used here, with five ensemble members (each with perturbed initial conditions) with Pinatubo forcing and a single counterfactual (without Pinatubo) simulation. We note that McClernon et al. (2024) also explored fitting an EESN on the HSW-V v1.0 data to forecast temperatures, but the model in McClernon et al. (2024) used AOD, T050, and T1000 to predict T1000. We make a distinction here from McClernon et al. (2024) in that we focus on distinct climatic pathways for the surface and stratosphere in an attempt to isolate the affects.

### 3.1.1 Normalized anomalies

Let  $Z_{k,t}^O(s_i)$  be the measured value for variable  $k = 1, 2$ , at time  $t = 1, 2, \dots, T$ , and for location  $s_i$ ,  $i = 1, 2, \dots, N$ . The HSW-V v1.0 normalized anomaly for each HSW-V v1.0 ensemble member for variable  $k$ , time  $t$ , and location  $s_i$ , denoted  $Z_{k,t}(s_i)$ , is calculated by

$$Z_{k,t}(s_i) = \frac{Z_{k,t}^O(s_i) - \overline{Z}_k^{\text{CF}}(s_i)}{\text{sd}(\mathbf{Z}_k^{\text{CF}}(s_i))}, \quad (18)$$

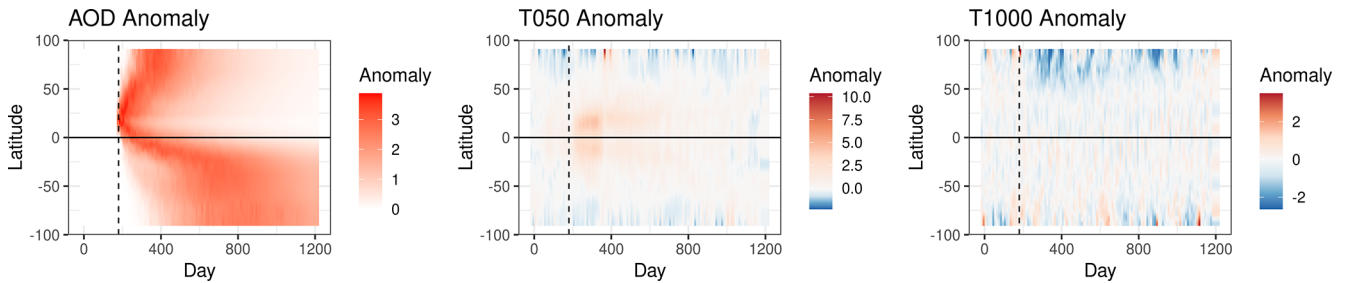
where  $\overline{Z}_k^{\text{CF}}(s_i)$  and  $\text{sd}(\mathbf{Z}_k^{\text{CF}}(s_i))$  are the mean and standard deviation, respectively, computed from the counterfactual run across all times for variable  $k$  at location  $s_i$ . Normalized anomalies for the temperature response,  $\mathbf{Z}_{Y,t}(s_i)$ , are calculated similarly. For HSW-V v1.0,  $k = 1$  refers to AOD, and  $k = 2$  refers to T050 or T1000, depending on which model is being discussed. The counterfactual in HSW-V v1.0 has AOD equal to zero for all times and locations, so we set  $\overline{Z}_1^{\text{CF}}(s_i) = 0 \forall i$ , and instead of  $\text{sd}(\mathbf{Z}_1^{\text{CF}}(s_i))$ , we calculate  $\text{sd}(\mathbf{Z}_1^O(s_i))$  as the standard deviation across measured AOD.

Because AOD is zero for all times and locations for the counterfactual, we replace it with random realizations from a standard Gaussian distribution (mean of 0; standard deviation of 1) to correspond to normalized data.

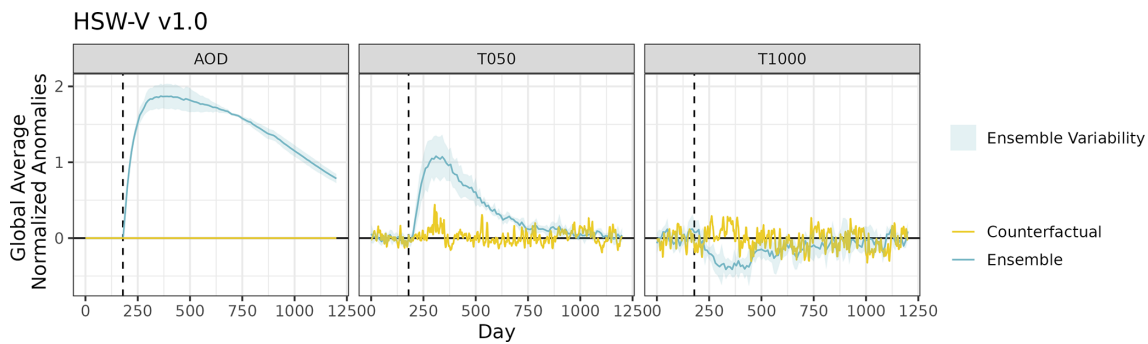
The injection of aerosols at time  $t = 179$  and how it propagates in space and time is seen in Figs. 1 and 2. The injection and spread of aerosols is due, in part, to the Brewer–Dobson circulation and is clear in latitude and time. The shaded regions in Fig. 2, and throughout the remainder of the paper, represent  $\pm 1$  standard deviation between model ensemble members. T050 and T1000 are directly related to AOD by construction in the HSW-V v1.0 simulations, with increases in AOD contributing to the long-wave heating of the upper levels and short-wave cooling by increased scattering of incoming solar radiation in the lower levels. This radiative heating is parameterized in HSW-V v1.0 by introducing temperature tendency terms parameterized by AOD, since HSW-V v1.0 does not include an explicit radiative transfer model (Hollowed et al., 2024b). Increased stratospheric AOD results in a positive temperature anomaly in T050 due to thermal absorption and a negative anomaly in T1000 due to increased scattering/reflection of short-wave radiation. AOD is advected by the stratosphere circulation faster in the Northern Hemisphere due to the fact that the injection occurs in the Northern Hemisphere, and the mean stratosphere circulation tends to be poleward. In the counterfactual, there is a small spike in T050 around day 270 that is unrelated to an aerosol injection (counterfactual AOD is zero). This spike is due to normal variation. Had there been more than one counterfactual, then taking the average would likely have smoothed over this effect.

### 3.1.2 Feature importance

After the hyperparameter selection, the HSW-V v1.0 models were trained using all 1200 d, and stZFI was computed. Figure 3 shows results from applying stZFI to HSW-V v1.0 (using the methodology described in Sect. 2.2.1). The top plot shows stZFI for the model predicting T050. The bottom plot shows stZFI for the model predicting T1000. The vertical dashed line denotes the injection time. Note that the y-axis scale is different for the two plots. The importance of AOD increases sharply for the T050 at the time of injection and slowly decays thereafter, as expected. Time-lagged T050 follows a similar trend. The increased importance for AOD when forecasting T1000 is less pronounced though present. The importance for AOD is higher for the T050 model compared to T1000 model, and the decay of importance is steeper from its peak for the T050 model compared to the T1000 model. This makes sense since T1000 is noisier than T050, and the true relationship between AOD and T050 is stronger than AOD with T1000. Thus, the FI results agree with our expectation that the impact AOD has on T1000 is less pronounced compared to T050.



**Figure 1.** Latitudinal means over time for the HSW-V v1.0 ensemble 1 normalized anomalies. Vertical dashed lines on day 179 denote the aerosol injection.



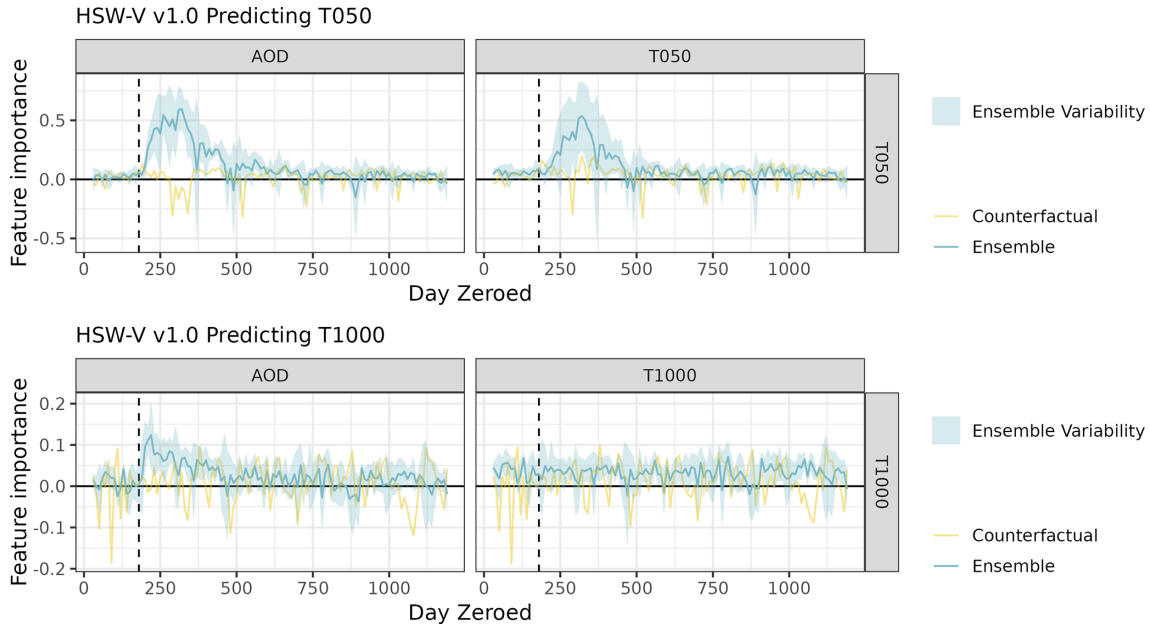
**Figure 2.** Globally averaged normalized anomalies for HSW-V v1.0 for the aerosol injection ensemble member and counterfactual. The solid line is the mean across ensemble members. Vertical dashed lines on day 179 denote the aerosol injection. Shaded regions represent  $\pm 1$  standard deviation of the ensemble variability. There is only one counterfactual run.

The counterfactual run allows us to consider how stZFI responds when there is no aerosol injection. The yellow lines in Fig. 3 show stZFI for the counterfactual of HSW-V v1.0. Feature importance for AOD is relatively flat for both models when considering the variation over time, which provides evidence that the peaks in stZFI for the runs with an aerosol injection are due to the EESN making use of the increase in aerosols for predicting temperatures. For the T050 model counterfactual, there is a small decline in AOD importance after the injection. This likely corresponds to the small spike in T050 previously identified in Fig. 2 that is known to be unrelated to the aerosol injection. Negative feature importance implies that the inclusion of the feature in question makes predictions worse than if it had not been included. However, small periods of negative stZFI are not a concern because it is a spatiotemporal metric, so it is not unreasonable to expect some time or spatial periods to not be helpful for prediction.

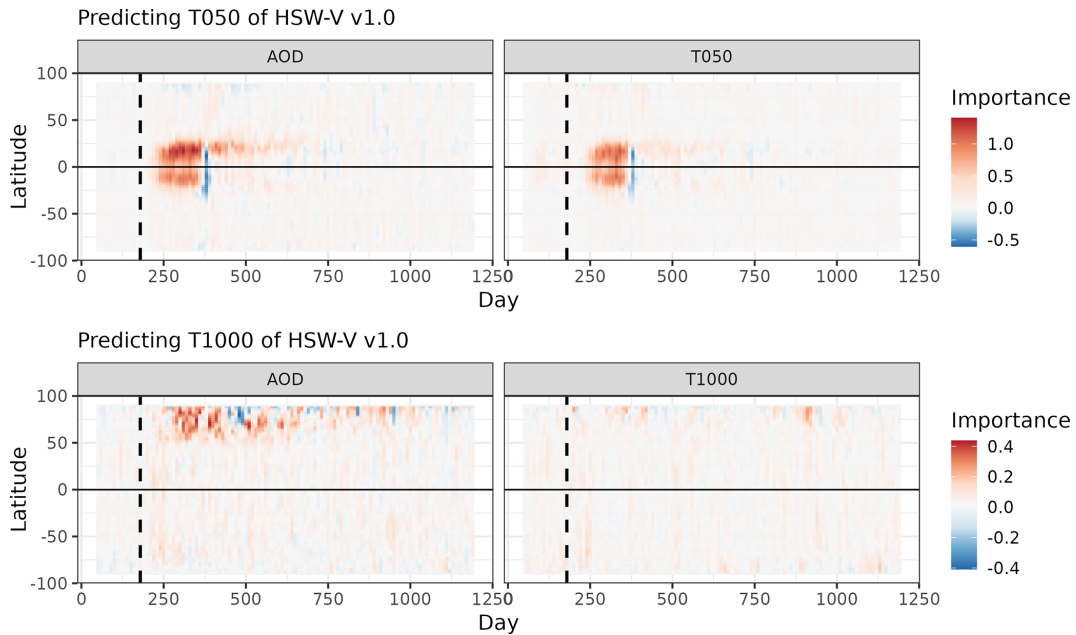
Figure 4 shows the latitudinal contributions to stZFI for the two predictive models on HSW-V v1.0 (as described in Sect. 2.2.2). The T050 model shows the impact of AOD on T050 after the aerosol injection around the Equator. The importance of AOD lasts longer in the Northern Hemisphere than the Southern Hemisphere. This is interesting since in Fig. 1 the AOD anomalies are higher for a longer period of time in the Southern Hemisphere. Thus, even though the Southern Hemisphere sees higher-AOD anomalies for longer

than the Northern Hemisphere, they are not identified as important for predicting T050 at the same location. Lagged T050 are important around the Equator after the eruption, which matches the increased anomalies in Fig. 1. For the model predicting T1000, AOD is most important at the high northern latitudes. This is also where we see the largest negative anomalies in T1000 in Fig. 1 (however, recall that the importance does not indicate a sign of a relationship).

By applying stZFI to HSW-V v1.0 ensemble members, we were able to assess the importance of each variable to the model in terms of predictive ability. For example, when predicting T050, we identified that AOD quickly increases in importance after the aerosol injection and then slowly decays, and stZFI is near zero when AOD is near zero. From an EDA perspective, this suggests that an SAI event is at least associated with changes in T050. Insights such as this could inspire additional hypotheses to explore and additional settings on the ESM to run. For example, given that we see this effect for a Pinatubo-like eruption, how would the impact differ if we changed the latitude of the aerosol injection? This simple example is merely a proof of concept since the injection of aerosols was the only change in the system. Next, we consider the fully coupled E3SM, where the relationships resulting from Mount Pinatubo are not as obvious.



**Figure 3.** stZFI for HSW-V v1.0. Vertical dashed lines denote the injection time. Shaded regions represent  $\pm 1$  standard deviation of the ensemble variability.



**Figure 4.** Latitudinal contributions to stZFI for models fit to HSW-V v1.0. Vertical dashed lines on day 179 denote the aerosol injection.

### 3.2 E3SM

E3SM is a fully coupled state-of-the-science ESM capable of simulation and prediction created by the United States Department of Energy and its national laboratories (Rasch et al., 2019). E3SM is a full physics model with active model components consisting of atmosphere, land, ocean, sea ice, and river. The data utilized in our numerical ex-

periments were generated by running simulations using an enhanced fork (available at <https://github.com/sandialabs/CLDERA-E3SM>, last access: 20 September 2024) of version 2 of E3SM (E3SMv2) (Golaz et al., 2022). In this new implementation, aerosol microphysics were modified to prognostically simulate stratospheric volcanic aerosols and are known as E3SMv2–Stratospheric Prognostic Aerosols (E3SMv2–

SPA) (Golaz et al., 2022). Data are mapped to a  $1^\circ \times 1^\circ$  structured latitude/longitude grid for 72 vertical levels.

The Mount Pinatubo eruption in the model occurs on 15 June 1991 at  $15.14167^\circ$  N and  $120.35000^\circ$  E. The magnitude is 10 Tg of the  $\text{SO}_2$  spread evenly over 6 h at an altitude of 18–20 km. For our analyses, we consider data on the monthly timescale. Five ensemble members were generated from the model. Each ensemble member was initialized with perturbed initial states beginning on 1 January 1985 to ensure that, by 15 June 1991, all ensemble members are dynamically independent (Brown et al., 2024). Dynamics arise from the seasonal heating imbalance and additional forcings beyond Mount Pinatubo that change global radiation balance. In addition, there is a positive trending background imbalance due to anthropogenic emissions of greenhouse gases. There are three aerosol precursor gases and seven aerosol species from natural and anthropogenic sources which vary seasonally. The 2 m surface temperature depends on the solar heating rate, which is affected by AOD, cloud cover, surface albedo, and ocean state. Counterfactuals with the Mount Pinatubo eruption removed were generated for each of the five ensemble members.

Again, we build two EESN models for modeling temperature pathways with the E3SM ensembles.

- *E3SM stratosphere model*. It predicts T050, given the AOD, long-wave radiative flux net top of atmosphere (LWTUP), and T050.
- *E3SM surface model*. It predicts T2M (2 m surface temperature), given the AOD, short-wave radiative flux clear sky (SWGDNCLR), and T2M.

All input variables are time-lagged. These variables form the structure explained in McCormick et al. (1995), where the Mount Pinatubo eruption injected aerosols which warmed the stratosphere with upwelling radiation and a cooling of the surface.

### 3.2.1 Normalized anomalies

Normalized anomalies for E3SM are calculated slightly differently than HSW-V v1.0 since E3SM has seasonality. The normalized anomaly for each E3SM ensemble member for variable  $k$ , time  $t$ , and location  $s_i$  is calculated by

$$Z_{k,t}(s_i) = \frac{Z_{k,t}^O(s_i) - \bar{Z}_{k,\text{month}(t)}^{\text{CF}}(s_i)}{\text{sd}(Z_{k,\text{month}(t)}^{\text{CF}}(s_i))}, \quad (19)$$

where  $\bar{Z}_{k,\text{month}(t)}^{\text{CF}}(s_i)$  and  $\text{sd}(Z_{k,\text{month}(t)}^{\text{CF}}(s_i))$  are the mean and standard deviation, respectively, computed from an ensemble member's corresponding counterfactual run across all data in the month to which time  $t$  belongs (i.e.,  $\text{month}(t)$  returns the month of time  $t$ ) for the variable  $k$  and location  $s_i$ . Normalized anomalies for the temperature response,  $Z_{Y,t}(s_i)$ , are calculated similarly. For E3SM,  $k = 1$  refers

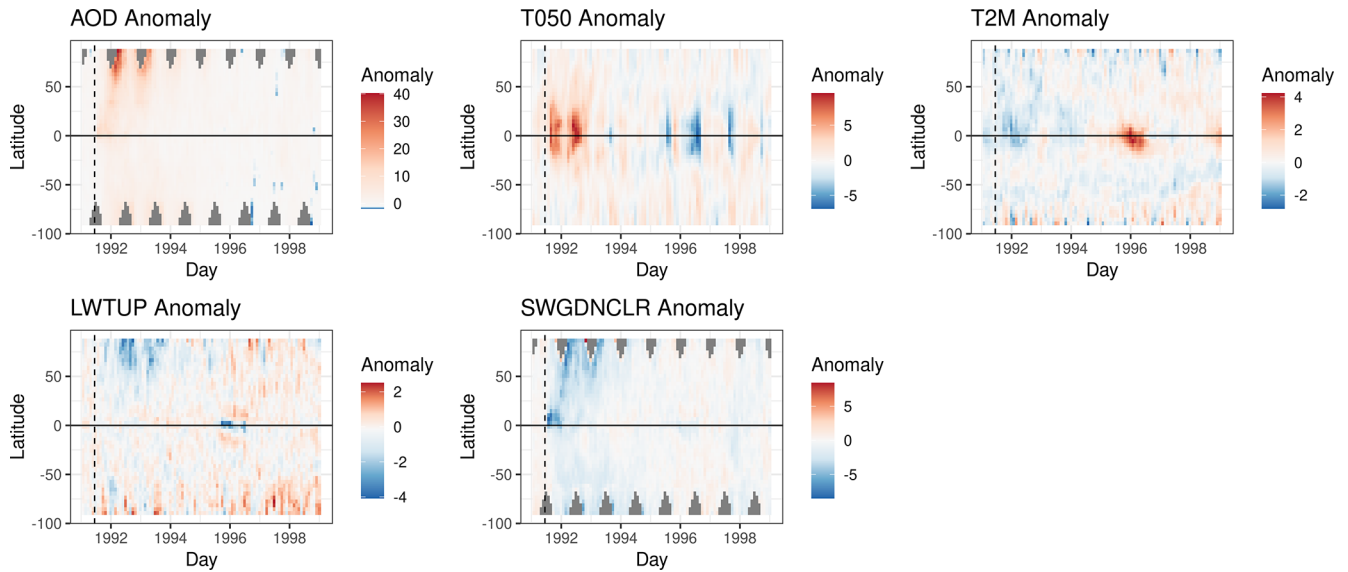
to AOD,  $k = 2$  refers to the long-wave radiation net top of stratosphere (LWTUP) for the T050 stratospheric model or incoming radiation at surface without clouds (SWGDNCLR) for the surface (T2M) model, and  $k = 3$  refers to the respective temperature for each model. Figure 5 shows the latitudinal means of normalized anomalies over time from a single ensemble member from E3SM. Due to the Pinatubo eruption, we see large positive anomalies in AOD and T050, with T050 largely focused around the Equator in the years following the eruption. We also see smaller negative anomalies for T2M around the Equator and north of the Equator. The large positive anomaly in T2M around 1996 is due to the spike in temperatures for ensemble member 1 (this heatmap is for ensemble member 1 only). Short-wave radiation also has negative anomalies after Pinatubo, particularly in the Northern Hemisphere, likely relating to the negative anomalies for T2M.

Figure 6 shows globally averaged normalized anomalies for E3SM. AOD has the largest spike relative to the counterfactuals, while T2M sees the smallest change. The impact of Mount Pinatubo is clear on the radiation measurements and T050. For the counterfactual case, there is still a small spike in AOD at the end of 1991, along with small impacts to radiation and temperature, even with Mount Pinatubo removed. The cause of this signal could be the volcanic eruption of Cerro Hudson on 8 August 1991, which was smaller than Mount Pinatubo (Miles et al., 2017).

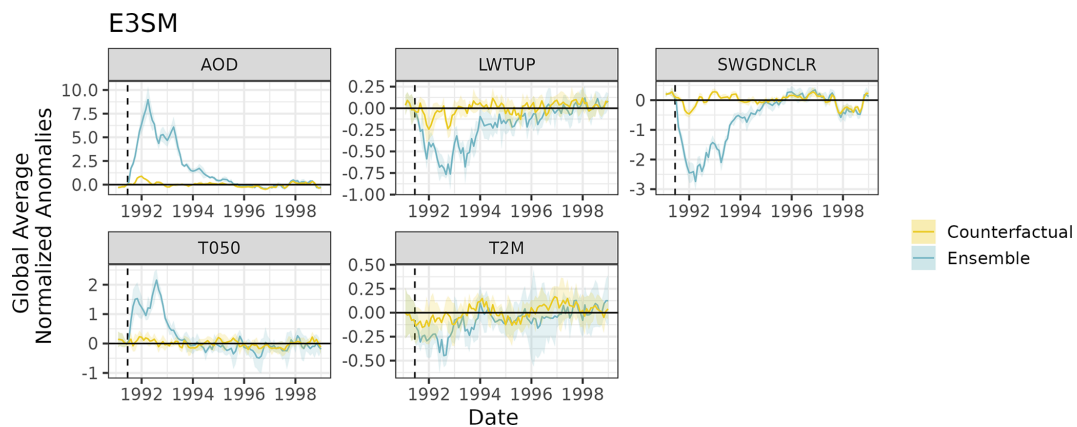
### 3.2.2 Feature importance

As with the HSW-V v1.0 data, we performed a hyperparameter optimization for the E3SM data. Details on the hyperparameter used and tuning is in Appendix A. After the hyperparameter optimization, we used data from 1991–1998 to train the EESN to E3SM data and compute stZFI. Since E3SM is a high-fidelity climate model, the data it produces will be a more realistic representation of reality than HSW-V v1.0. The pathways stZFI needs to quantify will be more complex and involve multiple variables.

Figure 7 shows stZFI for the E3SM ensemble and their counterfactuals. The vertical dashed line represents Mount Pinatubo's eruption. Although the temperature pathway in E3SM is not as direct as HSW-V v1.0, which contains interactions and confounding variables, the feature importance results tell the same story. For the E3SM stratospheric (T050) and surface (T2M) models, the AOD's feature importance immediately spikes at the eruption, and is relatively large compared to other variables, and then tapers off as time progresses. The importance of LWTUP remains relatively flat. Importance of SWGDNCLR does see a small increase after the eruption, although it is within the bounds of the counterfactual stZFI. Although we expected the radiation to have larger stZFI values, it could be because the radiation has a lower signal compared to AOD, and its impact is largely captured through AOD.



**Figure 5.** Latitudinal means over time for E3SM ensemble 1 normalized anomalies. Vertical dashed lines denote the 15 June 1991 Mount Pinatubo eruption. Gray in AOD and SWGDNCLR plots are NA (not available) values.



**Figure 6.** Globally averaged normalized anomalies for the E3SM ensemble and counterfactuals. Note the y axis is different for each plot. The shaded area represents the  $\pm 1$  standard deviation of the ensemble variability. Vertical dashed lines denote the 15 June 1991 Mount Pinatubo eruption.

Much like HSW-V v1.0, T050 is important for predicting itself after the eruption, while T2M is not. There is a spike in T2M feature importance around 1996, but this is largely because one of the ensemble members has a dramatic increase in T2M at that time (as seen by the large variation in Fig. 6). More than five runs of the ESM ensemble would be necessary to determine if this spike is truly anomalous or part of a larger trend. There are no major trends in stZFI for the counterfactuals, but all variables retain some degree of importance since the variables on the respective pathways affect T050 and T2M with or without an eruption. A broader assessment of stZFI robustness using E3SM is provided in Appendix C.

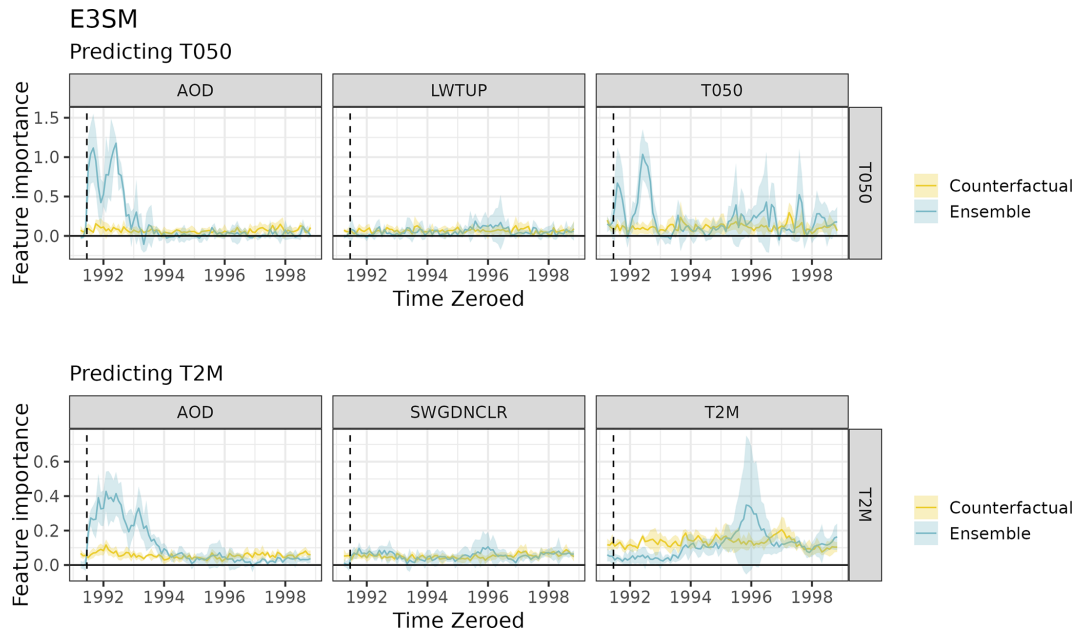
Much like the HSW-V v1.0 case, we are able to extract relevant variable relationships from a purely data-driven ML

model. As an EDA tool, this gives us an assessment of the relationships in the data. Unlike simple summaries such as means and correlations, the relationships found here are measured over time and account for complex, nonlinear associations. The latitudinal contributions to stZFI for the E3SM are deferred until Sect. 4.

### 3.3 MERRA-2 application

The last two subsections applied stZFI to ESM simulations. Now we turn to verifying that the feature importance results from the ESMs are consistent with results from observed data products. To do this, we use the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) (Gelaro et al., 2017), reanalysis as our observa-





**Figure 7.** stZFI for E3SM ensemble and counterfactuals. The shaded region denotes  $\pm 1$  standard deviation of the ensemble variability. Vertical dashed lines denote the 15 June 1991 Mount Pinatubo eruption.

tional data. As with the E3SM models, we build two EESN models for modeling temperature pathways with MERRA-2.

- *MERRA-2 stratosphere (T050) model.* It predicts T050, given the AOD, long-wave radiative flux net top of atmosphere (LWTUP), and T050.
- *MERRA-2 surface (T2M) model.* It predicts T2M (2 m surface temperature), given the AOD, short-wave radiative flux clear sky (SWGDNCLR), and T2M.

All input variables are time-lagged. Vertically integrated AOD is taken from the variable TOTEXTTAU (GMAO, 2015a). We consider the years 1991–1998 using monthly data, which provide climate information before and after the eruption of Mount Pinatubo. The spatial resolution is  $1^\circ \times 1^\circ$  on a structured latitude/longitude grid.

### 3.3.1 Normalized anomalies

Since MERRA-2 does not have a counterfactual, normalized anomalies are calculated differently for MERRA-2 than for HSW-V v1.0 and E3SM. The normalized anomaly for each MERRA-2 variable  $k$  at time  $t$  and location  $s_i$  is calculated by

$$Z_{k,t}(s_i) = \frac{Z_{k,t}^O(s_i) - \bar{Z}_{k,\text{month}(t)}(s_i)}{\text{sd}(Z_{k,\text{month}(t)}(s_i))}, \quad (20)$$

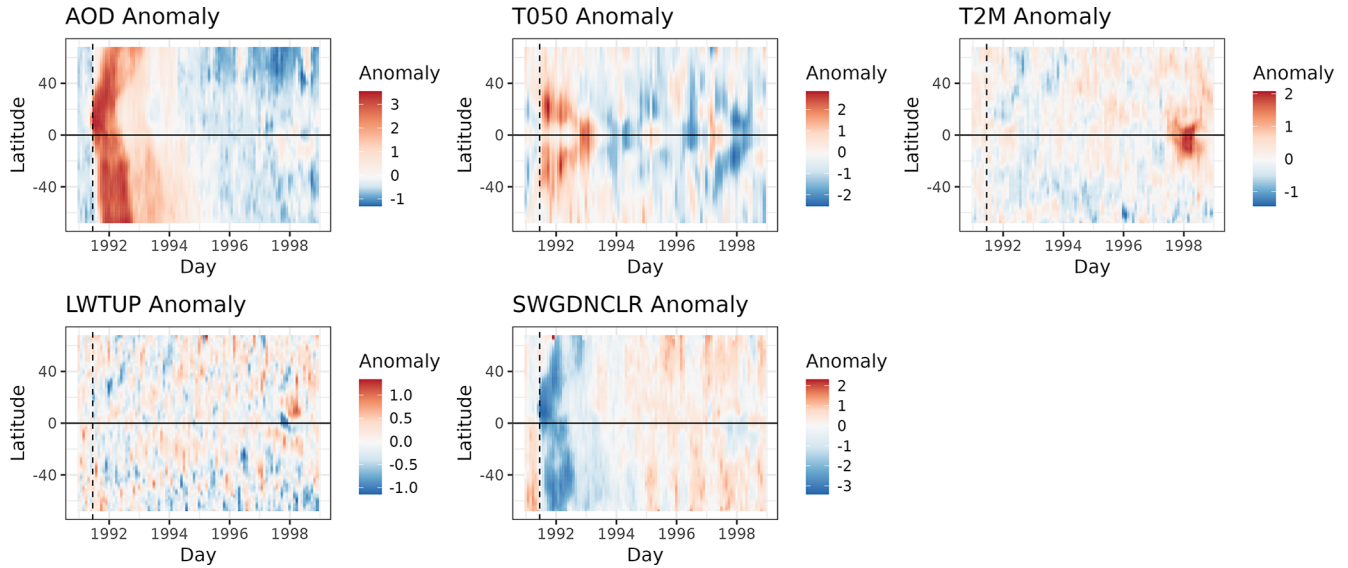
where  $\bar{Z}_{k,\text{month}(t)}(s_i)$  and  $\text{sd}(Z_{k,\text{month}(t)}(s_i))$  are the mean and standard deviation, respectively, across all data from 1991–1998 for the month in which time  $t$  belongs

(i.e.,  $\text{month}(t)$  returns the month of time  $t$ ) and for variable  $k$  at location  $s_i$ ,  $i = 1, \dots, N$ . Similar to E3SM,  $k = 1$  refers to AOD;  $k = 2$  refers to radiative flux (LWTUP for T050; SWGDNCLR for T2M); and  $k = 3$  refers to temperature, T050, or T2M, depending on which model is being discussed. Figure 8 shows the latitudinal means of normalized anomalies over time for MERRA-2. Figure 9 shows globally averaged normalized anomalies for MERRA-2. AOD and short-wave radiation have the largest relative spikes post-Pinatubo, while T2M and long-wave radiation see the smallest change. The impact of Mount Pinatubo is clear on the short-wave radiation measurements and T050. The large anomaly for T2M around 1998 is likely due to the 1997–1998 ENSO, which was the largest recorded at that time, and led to a significant increase in globally averaged temperature (Wang and Weisberg, 2000).

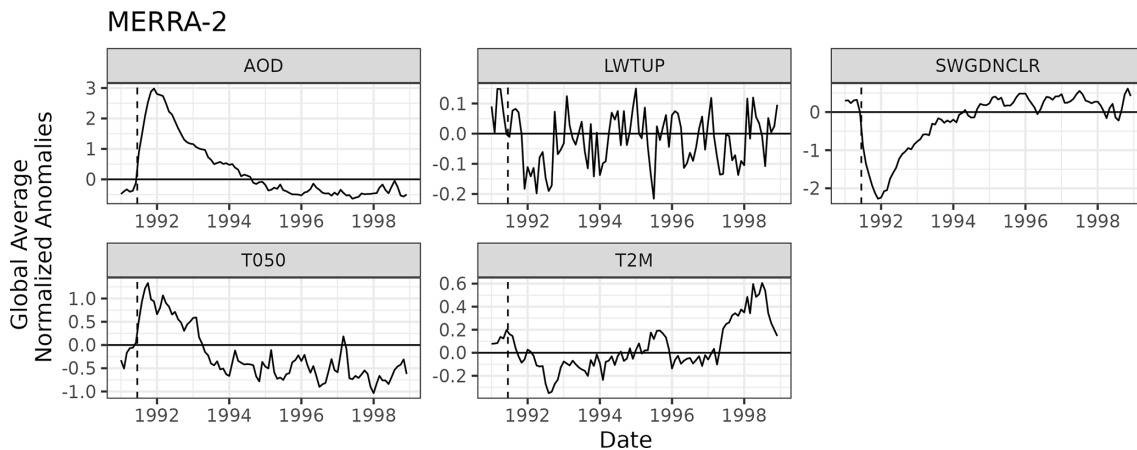
### 3.3.2 Feature importance

After hyperparameter optimization, we used data from 1991–1998 to train the EESNs on MERRA-2 data and compute stZFI. Figure 10 shows stZFI for the MERRA-2 data, where the vertical dashed black line denotes Mount Pinatubo’s eruption. There is a clear signal in the stZFI for AOD immediately after the eruption for both models, although the signal is clearer in the T050 model. Short-wave radiation also has a clear increase in importance after the eruption for the T2M model corresponding with the decrease seen in Fig. 9. The importance for T050 predicting itself is more noisy, although it is elevated immediately after the Pinatubo eruption. The importance for T2M predicting itself shows a steady in-





**Figure 8.** Latitudinal means over time for MERRA-2 normalized anomalies. Vertical dashed lines denote the 15 June 1991 Mount Pinatubo eruption.



**Figure 9.** Globally averaged normalized anomalies for MERRA-2. Note that the y axis is different for each plot. Vertical dashed lines denote the 15 June 1991 Mount Pinatubo eruption.

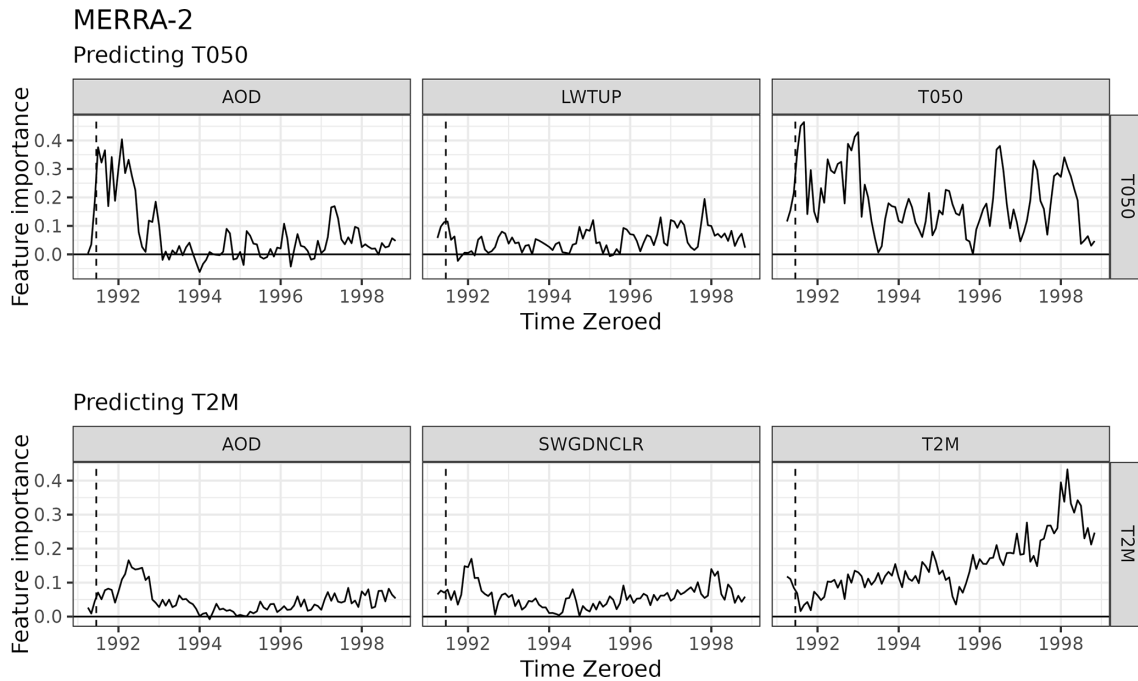
crease from 1991–1996, with a slight dip in 1995. This could potentially be due to an increase in the auto-correlation of T2M post-Pinatubo. The importance of long-wave radiation is relatively flat. Similar to E3SM, we believe the feature importance for radiation is largely flat due to a lower signal-to-noise ratio compared to AOD, coupled with its correlation with AOD.

#### 4 Comparing stZFI from E3SM to MERRA-2

The models for E3SM and MERRA-2 in Sect. 3.2 and 3.3, respectively, are trained on the same time frame and spatial scale. We cannot compare observations to HSW-V v1.0 since it is a notional lower-fidelity model. Because the data are

standardized in different ways, we will avoid exact quantitative comparisons and make a qualitative comparison.

When considering effects averaged over the entire globe, this can hide regional effects, so we consider the latitudinal contributions to stZFI to explore its importance in space and time. Figure 11 shows the regional contributions to stZFI for models predicting T050 using E3SM and MERRA-2. These contributions show the relative importance of each of the three variables for predicting T050 by latitude and over time, thus providing a spatiotemporal feature importance. E3SM shows a relatively uniform importance for AOD between 30° N and –20° S from July 1991 to November 1993, with a slight decline in the winter of 1992. MERRA-2 shows the importance of AOD in the regions from 10 to 30° N and –30 to –10° S beginning in July 1991 and ending in the



**Figure 10.** stZFI for MERRA-2. Vertical dashed lines denote the 15 June 1991 Mount Pinatubo eruption.

late summer of 1992. The FI for MERRA-2 appears to be more drawn out for T050; this could be due to model misspecification in E3SM that does not fully account for the effects post-Pinatubo. E3SM and MERRA-2 largely agree on the importance of time-lagged T050 for predicting T050, as both show high importance near the Equator at similar times. Trends for long-wave radiation are more difficult to assess, but it does appear that E3SM and MERRA-2 have higher variability in FI near the Equator, while towards the poles FI tends to remain more consistent. There is negative importance for long-wave radiation immediately after Pinatubo in E3SM and MERRA-2, indicating that its presence hurts the model. This likely due to long-wave radiation being only weakly correlated with T050 and combined with a relatively high noise compared to the signal post-Pinatubo, as shown in Fig. 6. There is also a negative spike for AOD for E3SM in mid-1993, which could be due to AOD levels converging on those of the counterfactual (Fig. 6). Note that E3SM importances appear to be “smoother” since they are averaged over five ensemble members, whereas MERRA-2 is a single dataset.

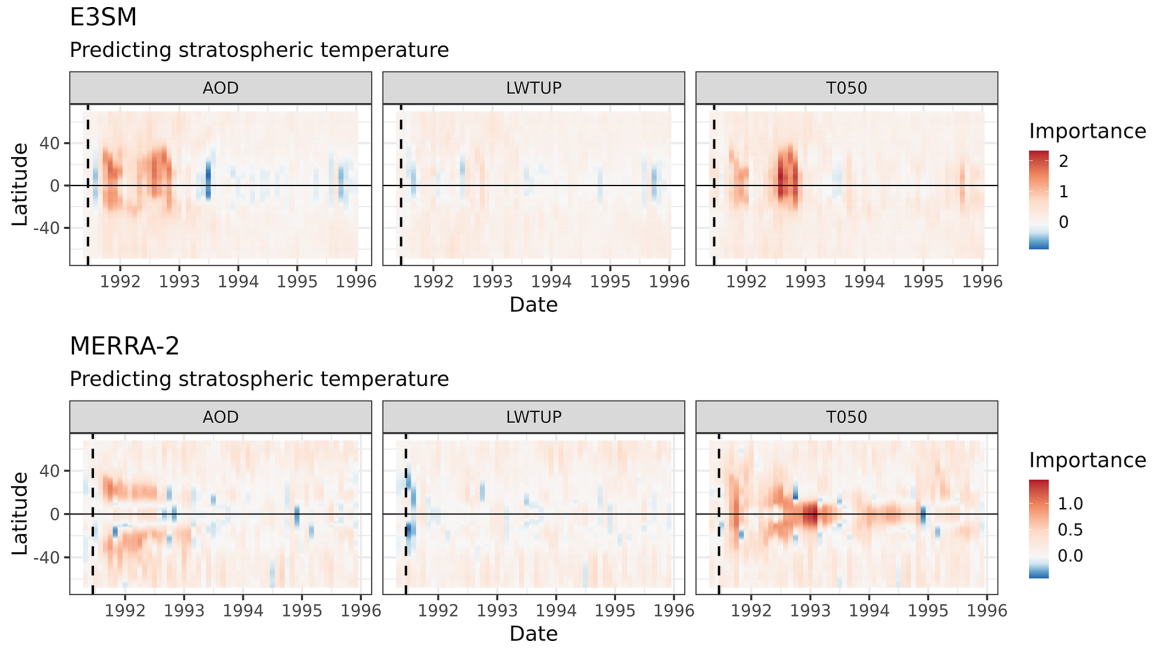
Figure 12 shows the regional contributions to stZFI for models predicting T2M using E3SM and MERRA-2. E3SM sees the importance for AOD mostly just north of the Equator post-Pinatubo, while MERRA-2 sees the importance further from the Equator in both directions. The importance of short-wave radiation is spread across all latitudes for E3SM, while MERRA-2 has a higher importance further from the Equator. There are not clear trends for importance of T2M, except during later years. E3SM puts a high importance on T2M on

the Equator at the end of 1995, while MERRA-2 shows the importance at the end of 1994 just north of the Equator.

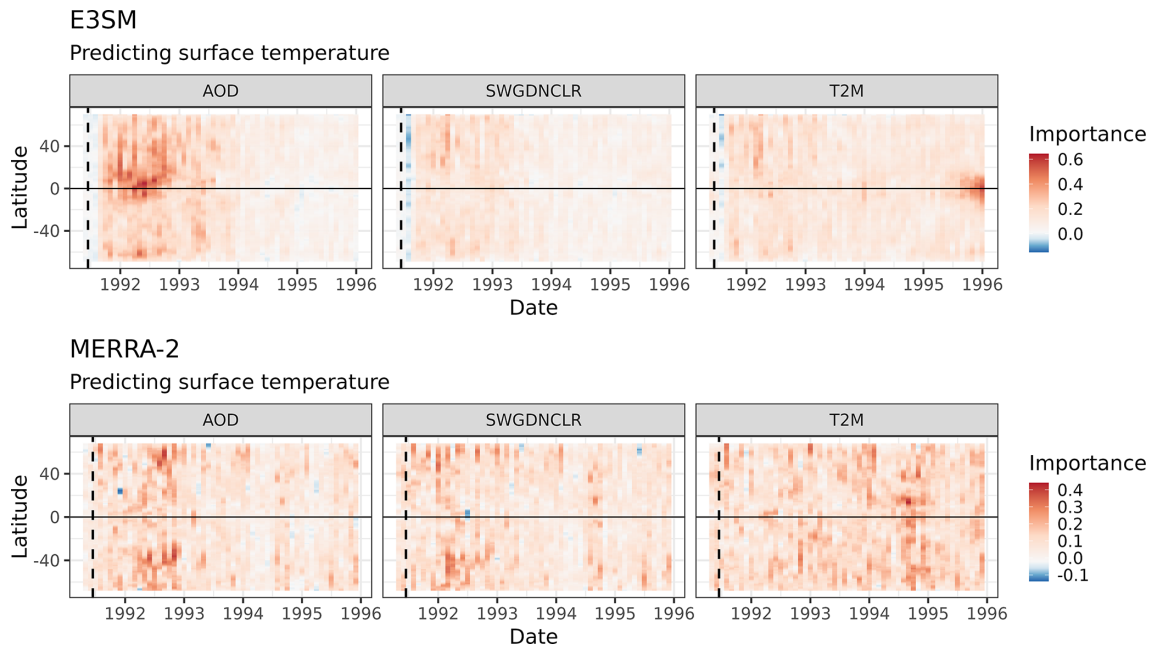
## 5 Discussion

ESMs provide rich information about the physical state of the climate and its variations. ML and other data-driven models offer one path to taking advantage of the vast amounts of data produced by ESMs to advance the understanding of climate systems. In addition to using ML for predictive reasons, explainability methods allow the ability to discover and quantify patterns in data via ML models for prediction. In this article, we provided an example of how an explainable ML technique, stZFI, can be used as an EDA tool for climate applications to understand how variable relationships evolve over space and time. We demonstrated stZFI via a case study that explored climate-variable relationships associated with a natural exemplar of an SAI event, namely the 1991 volcanic eruption of Mount Pinatubo. We chose this event since it is well-studied and well-documented, which helps future users understand how stZFI could be used as an EDA tool. We leveraged ESMs to study how stZFI quantifies variable relationships with datasets that are generated with known relationships. Furthermore, we compared stZFI computed from ESM-generated data to stZFI computed from reanalysis data to determine if the results were consistent.

We considered two climate pathways previously identified in the literature that are associated with the Mount Pinatubo eruption, namely (1) aerosols to long-wave radiative flux to



**Figure 11.** Latitudinal contributions to stZFI for E3SM and MERRA-2 for models predicting T050. Note that the importance scales are different for E3SM and MERRA-2. Vertical dashed lines denote the 15 June 1991 Mount Pinatubo eruption.



**Figure 12.** Latitudinal contributions to stZFI for E3SM and MERRA-2 for models predicting T2M. Note that the importance scales are different for E3SM and MERRA-2. Vertical dashed lines denote the 15 June 1991 Mount Pinatubo eruption.

stratospheric temperature changes and (2) aerosols to short-wave radiative flux to surface temperature changes. We applied stZFI to EESNs to conduct an EDA with an interest in understanding how the three pathway variables are related to the changes in temperature over time after an SAI. We studied these pathways using three data sources: a simpli-

fied ESM with only the single forcing of aerosols (HSW-V v1.0), a fully coupled ESM (E3SM), and a reanalysis dataset (MERRA-2). For all models and data sources, the relationships identified by stZFI were relatively consistent.

- Aerosols had the most consistent FI results. In all cases, there was a clear increase in stZFI for predicting tem-

peratures immediately after the SAI, which decreases over time.

- Radiative flux variables associated with E3SM and MERRA-2 had relatively similar FI trends. For long-wave radiative flux, there was no clear trend in FI with values close to 0 across all times when predicting stratospheric temperatures. For short-wave radiative flux, there was a slight increase in FI after the SAI when predicting surface temperatures.
- Temperature FI values agreed between HSW-V v1.0 and E3SM but differed from MERRA-2 results. With stratospheric temperatures, the HSW-V v1.0 and E3SM results showed a clear increase in FI after the SAI, but the MERRA-2 results showed a noisy possible increase in FI. With surface temperatures, the HSW-V v1.0 and E3SM results showed no FI trends, but the MERRA-2 results showed a steadily increasing trend in FI.

The consistency in FI results for AOD across data sources provides evidence of these variable relationships being a part of the underlying mechanism. It is likely that AOD has the most consistent FI results due to the strong global signal-to-noise ratio of AOD after the eruption in all data sources. That radiation and temperature FI do not agree could partially be due to E3SM model discrepancies. These variables are changing at least partially due to the increase in AOD, making them downstream effects of such an event. Additionally, it is possible that measurements are similar between MERRA-2 and E3SM, while the relationships that caused them could vary, manifesting itself in FI. The EESN itself is a relatively flat predictive model, meaning it will likely not be able to capture all the complex relationships that exist, especially if they do not lead to better predictions.

The stZFI results can be used to point to new hypotheses and research directions. For example, the upward trend in stZFI for T2M is unlikely due to Mount Pinatubo alone and could lead to additional research. It is possible that this upward trend is due to a combination of increasing global surface temperatures and a strong ENSO event from May 1997 to May 1998. Another example suggested by the latitudinal contribution plots is the question of how the latitude of an SAI event will affect its impacts. It also could help find areas where climate models do not match observational data. stZFI shows the variables a model is using, and when, in order to predict. Therefore, discrepancies between a climate model and observational data could point modelers to relationships that an ESM is not currently capturing.

In addition to using the SAI case study to demonstrate the ability of stZFI as an EDA tool, the analyses in this article contribute towards an increased understanding and confidence that the stZFI will return an accurate and reliable result. ESMs played a key role in this process, since they provide a specified cause with a known outcome, with an AOD being a major driver in temperature changes both at the sur-

face and in the stratosphere. In particular, this importance largely came from equatorial regions, leaning slightly to the Northern Hemisphere. The ESMs also allowed us to examine results from counterfactual runs from which the SAI is removed. When we considered EESNs trained on the counterfactuals, we found no FI patterns associated with the SAI. This result suggests that the FI trends that appear when SAI is included in the ESM runs are associated with SAI and not some other phenomenon. With observational data only, there would be no way to know for sure whether feature importance produced the correct effect, since this effect would not be known.

However, our analyses serve only as a case study for the assessment of stZFI. A more comprehensive evaluation of the method should be performed to better understand its strengths and limitations. For example, the literature on applying explainable ML methods to climate applications has been growing recently. A comparison of stZFI to other methods would be useful for understanding when stZFI is preferable to other methods. Additionally, the current methodology for stZFI does not fully account for the correlation between input variables. Previous research suggests that ignoring the correlation between variables can result in biased feature importance (Hooker et al., 2021). Further studies could be done to assess the affect of correlation on stZFI, and the methodology could be adjusted to better account for correlations. This future work would allow users to make stronger conclusions using stZFI without worrying about biases due to correlation.

Another direction for future research is developing a tool for ML EDA that is able to account for more complicated variable relationship structures. stZFI already provides an advantage over simple summary statistics such as means and correlations since it is applied to EESNs, which are flexible and not constrained to be linear or even monotonic. However, an EESN assumes a simple input–output model structure. We know that the climate pathways, such as the ones we explore in this article, are more complicated (for example, the temperature pathway exploring the effects from an SAI event and changes in radiation to changes in temperature happens across multiple mechanisms). For example, the changes could vary, depending on major climate cycles like ENSO, or such an event could possibly affect ENSO, making it difficult to disentangle causes and effects. The EESN treats it as a simple prediction problem with temperature as the output and all other variables as inputs. A method that allows for more structure in the inputs, including interactions, could result in a more useful and representative explainability metric.

In this article, we presented stZFI as a tool for exploratory analyses. stZFI provides insights into climate events by quantifying variable relationships over space and time, which provides some insight into the underlying mechanistic relationships. Exploratory analyses are an important aspect of science, where new discoveries are made and hypotheses are generated. An additional objective in the climate science

community is attribution (Hegerl et al., 2010; Bindoff et al., 2013). Although showing attribution is a multi-step problem, we believe stZFI could be used to provide an initial step towards making attribution claims. Regardless, we hope stZFI inspires ideas for how ML could be used for attribution in the climate space.

### Appendix A: EESN hyperparameter details

This section provides details on the EESN hyperparameter tuning and selection for the models applied to HSW-V v1.0, E3SM, and MERRA-2. We select hyperparameters by performing a hyperparameter search over the following grid of values:  $n_h = \{25, 50, 100, 200\}$ ,  $U_{width} = \{0.1, 0.5\}$ ,  $W_{width} = \{0.1, 0.5\}$ ,  $U_\pi = \{0.1, 0.5\}$ ,  $W_\pi = \{0.1, 0.5\}$ ,  $\nu = \{0.1, 0.5\}$ ,  $\lambda_r = \{0.5, 5, 50\}$ , where data are split into training and testing sets. HSW-V v1.0 used days 0–800 as training, E3SM used dates from 1 January 1991 to 31 December 1994 as training, and MERRA-2 used dates from 1 January 1991 to 31 December 1994 as training. The prediction metric being optimized for was the root mean squared error (RMSE). The remaining times in each dataset were used for testing. We opted to use 20 EOFs for each variable for each of the datasets for consistency. For HSW-V v1.0, 20 EOFs represents 98 %, 82 %, and 60 % of the variation in AOD, T050, and T1000, respectively. For E3SM, 20 EOFs represents 93 %, 92 %, and 73 % of the variation in AOD, T050, and T1000, respectively. For MERRA-2, 20 EOFs represents 88 %, 87 %, and 56 % of the variation in AOD, T050, and T1000, respectively. We acknowledge that the number of EOFs could differ by model and by variable within a model. Table A1 shows the optimal hyperparameters for each model for each dataset based on the hyperparameter search.

**Table A1.** Hyperparameters used for EESN models based on lowest test set RMSEs in the hyperparameter search.

Data	Model	$n_h$	$U_{width}$	$W_{width}$	$U_\pi$	$W_\pi$	$\nu$	$\lambda_r = 5$
HSW-V v1.0	T050	200	0.5	0.1	0.1	0.5	0.1	50
HSW-V v1.0	T1000	100	0.1	0.1	0.5	0.5	0.1	50
E3SM	T050	50	0.1	0.1	0.5	0.5	0.1	5
E3SM	T2M	200	0.1	0.1	0.1	0.1	0.1	5
MERRA-2	T050	200	0.1	0.1	0.5	0.5	0.1	50
MERRA-2	T2M	200	0.1	0.1	0.1	0.1	0.1	5

### Appendix B: Predictive performance of EESN on HSW-V v1.0 and E3SM

Figure B1 shows the predictive performance of the EESN over time for both HSW-V v1.0 models. The three rows in each plot show globally weighted RMSEs using different training sets; for example, the first row corresponds to training using data using times 1–200 and then testing on data from 201–1200. Weights are calculated by taking the square

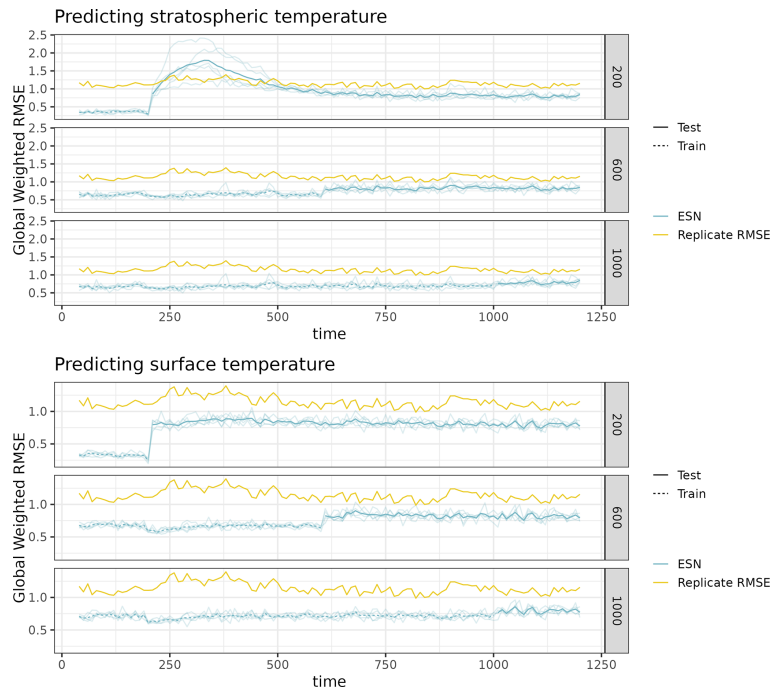
root of the cosine latitude (Huth, 2006). Thus, the weight associated with location  $s_i$  is

$$w_{s_i} = \sqrt{\cos\left(\text{latitude}(s_i) \times \frac{\pi}{180}\right)}, \quad (\text{B1})$$

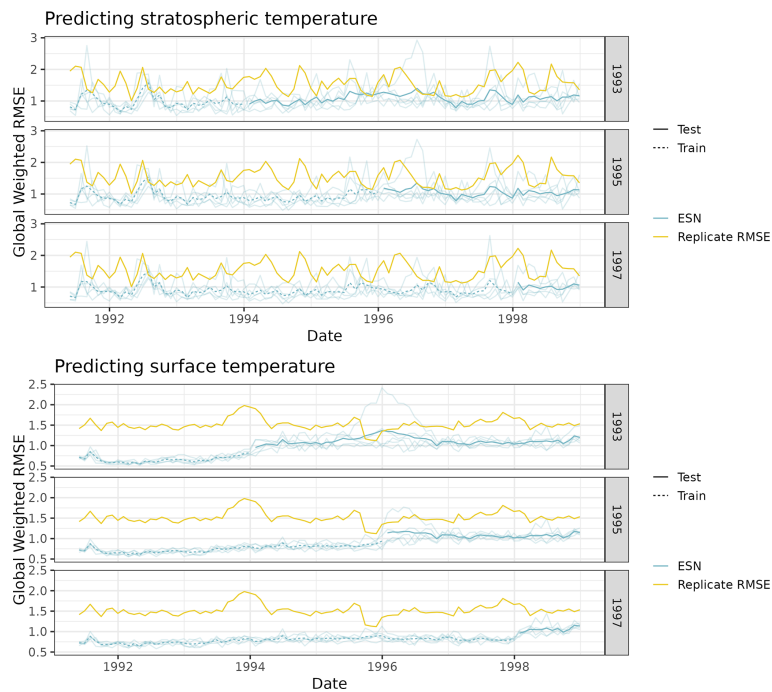
where  $\text{latitude}(s_i)$  returns the latitude of location  $s_i$  in degrees. The EESN RMSE is compared to replicate RMSE, which uses four E3SM ensemble members' values as predictions for the remaining member and then averages over the RMSEs of the four members. This process is repeated five times (each ensemble member is predicted using the other four), and the replicate RMSE is the average over those five results. All averages are calculated on a month-by-month basis. The EESN's RMSEs are typically lower than the replicate RMSE, showing that the EESN has a lower prediction error than the natural variability in the climate system itself. This shows that the EESN is providing predictive ability beyond that due to the ensemble variation, which, when combined with the hyperparameter optimization, provides credibility to stZFI computed from this EESN. Figure B1 shows that the EESN is able to capture trends better than that due to the ensemble variability when given enough training data.

Figure B2 shows the predictive performance of the EESN over time on E3SM using the globally weighted root mean squared error (RMSE). The three rows in each plot show globally weighted RMSEs using different training sets; for example, the first row corresponds to training using data from 1991–1993 and then testing on data from 1994–1998. This shows that the EESN is able to capture trends better than that due to the ensemble variability when given enough training data.

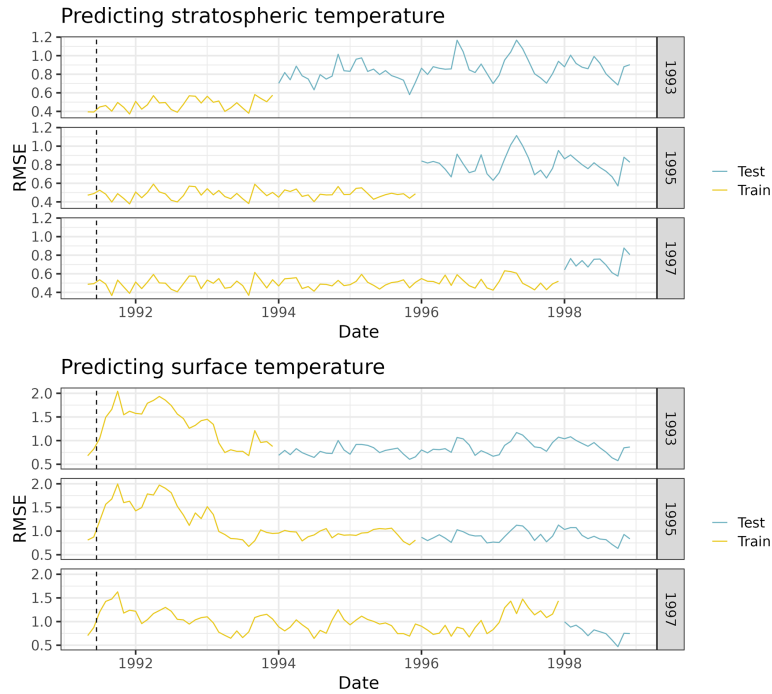
Figure B3 shows the predictive performance of the EESN over time on MERRA-2 using RMSE. The three rows in each plot show globally weighted RMSEs using different training sets; for example, the first row corresponds to training using data from 1991–1993 and then testing on data from 1994–1998. This shows that the EESN is able to capture trends better than that due to the ensemble variability when given enough training data.



**Figure B1.** Time series cross-validation globally weighted average RMSE for both HSW-V v1.0 models. Models are trained through the time shown in the row label. Bold blue lines are the average RMSE for an EESN, and light blue lines are the individual ensemble members’ RMSEs. The yellow lines represent the replicate RMSE used as a baseline comparison.



**Figure B2.** Time series cross-validation globally weighted average RMSE for both E3SM models. Models are trained through the row label year. Bold blue lines are the average RMSE for an EESN, and light blue lines are the individual ensemble members’ RMSEs. The yellow lines represent the replicate RMSE used as a baseline comparison.



**Figure B3.** Time series cross-validation global RMSE for both MERRA-2 models. Models are trained through the year in the row label.

### Appendix C: Assessing stZFI robustness on E3SM

Assessing the robustness is important for understanding the behavior of a method. Here we present several checks (non-exhaustive) to help illustrate how stZFI behaves under different model specifications. All EESNs in this section were trained with the same hyperparameters. For the T050 model,  $n_h = 50$ ,  $\nu = 0.1$ ,  $U_{\text{width}} = 0.1$ ,  $W_{\text{width}} = 0.1$ ,  $U_{\pi} = 0.5$ ,  $W_{\pi} = 0.5$ , and  $\lambda_r = 5$ . For the T2M model  $n_h = 200$ ,  $\nu = 0.1$ ,  $U_{\text{width}} = 0.1$ ,  $W_{\text{width}} = 0.1$ ,  $U_{\pi} = 0.1$ ,  $W_{\pi} = 0.1$ , and  $\lambda_r = 5$ . These are the optimized hyperparameters used for the E3SM model in the main paper.

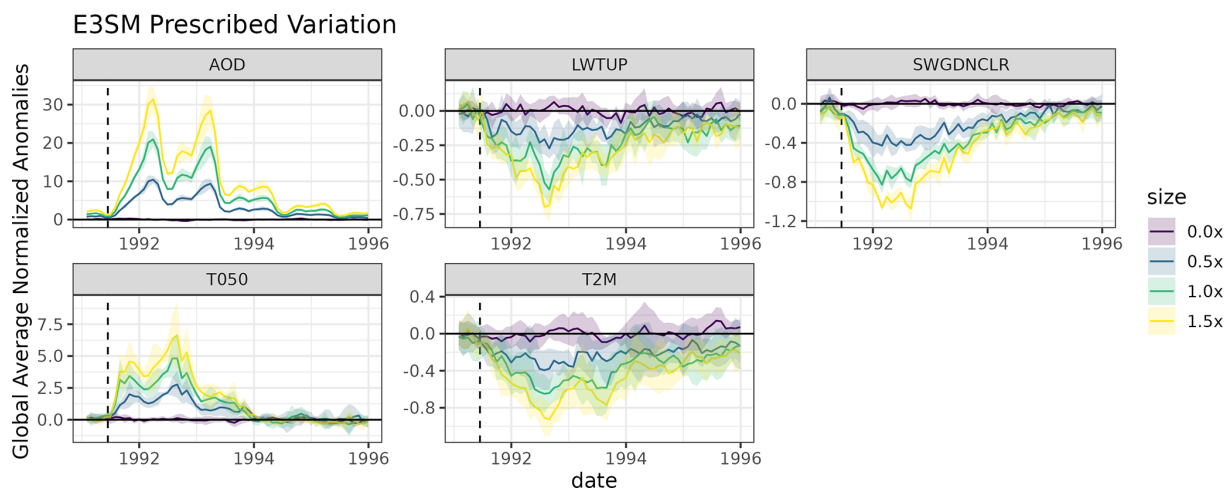
#### C1 Prescribed variation

We consider the impacts of eruptions smaller and larger than Mount Pinatubo to measure the gradient of the effects. These simulations are shorter, going from 1991–1995, and are initialized using historical CMIP6 ensemble. For these prescribed ensembles, we consider eruptions of  $0.0\times$ ,  $0.5\times$ ,  $1.0\times$ , and  $1.5\times$  the Mount Pinatubo eruption. The  $0.0\times$  eruption is the counterfactual and excludes the Mount Pinatubo and Cerro Hudson eruptions. Five ensemble members are generated for each eruption mass condition.

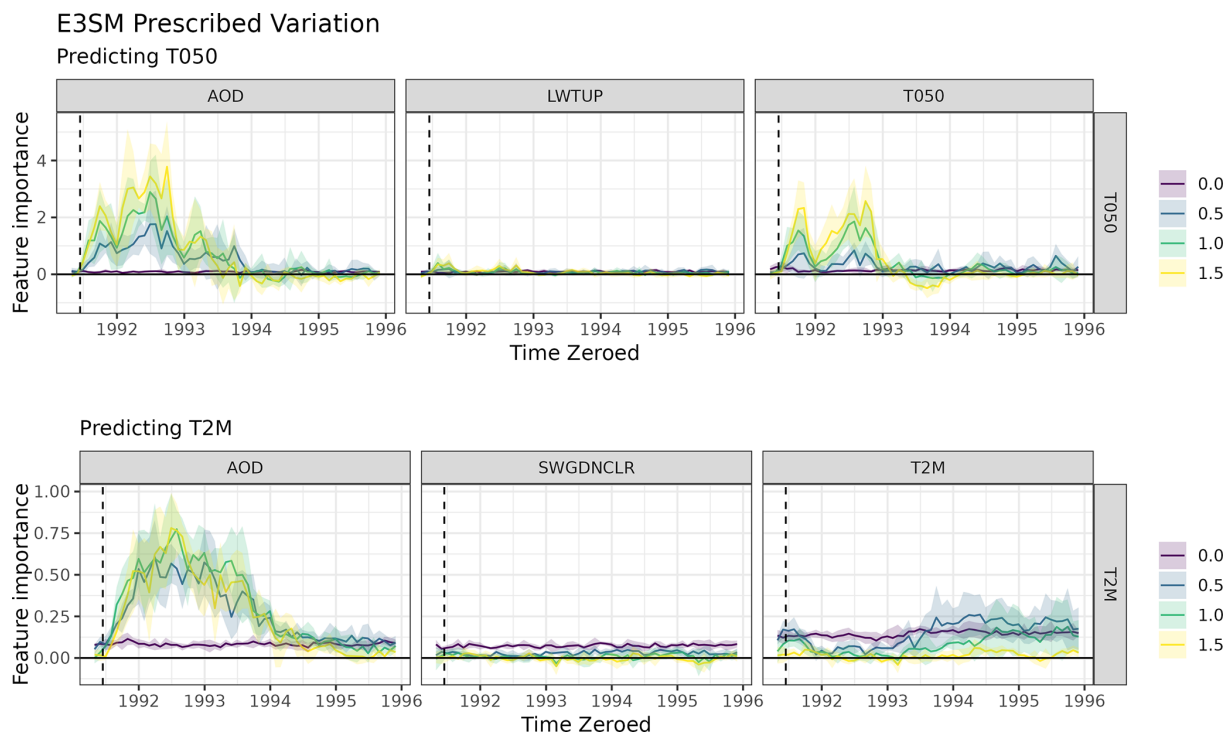
Figure C1 shows globally averaged normalized anomalies for the prescribed variation E3SM ensembles. The colors correspond to the size of the prescribed eruption relative to Mount Pinatubo (e.g.,  $0.5$  means the simulation replaces the original Mount Pinatubo eruption with an eruption half the size). There is a clear gradient in the variable value corresponding to the size of eruption for all variables. The two peaks in AOD result from the standardization process; the variation in AOD for the early months of 1993 was low relative to its mean deviation from the counterfactual.

Figure C2 shows stZFI for E3SM with prescribed eruptions. The spike in importance is relative to the magnitude of the eruption, and larger eruptions have larger spikes in the feature importance. The importance of LWTUP and SWGDNCLR is minimal, and T050 and T2M follow similar trends to the results in Fig. 7.





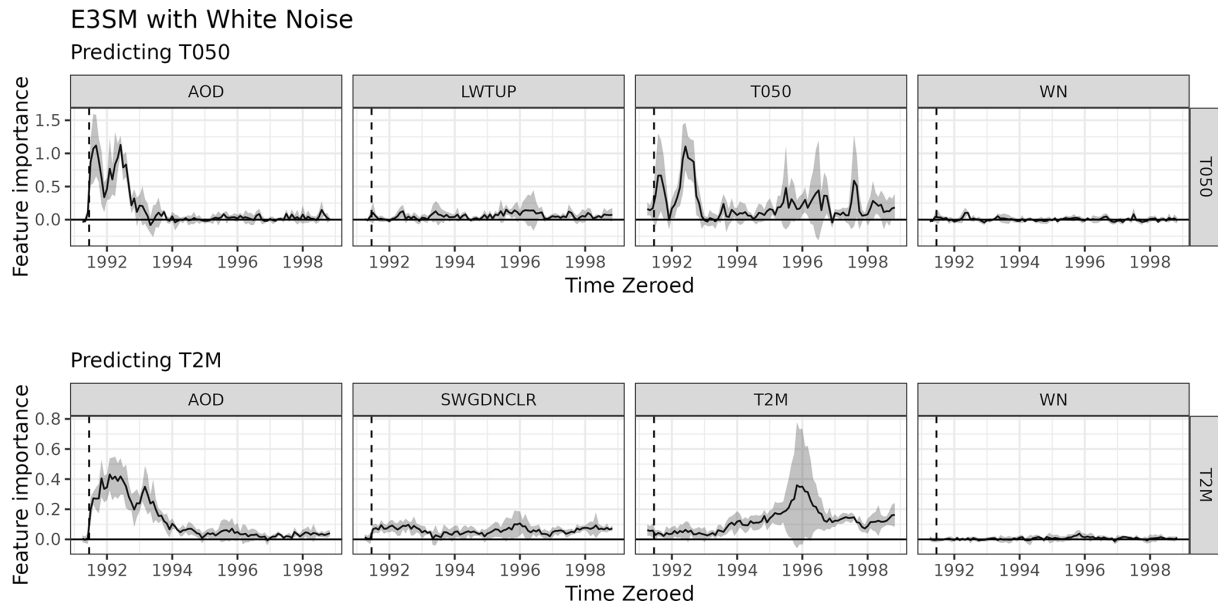
**Figure C1.** Globally averaged normalized anomalies for the prescribed variation E3SM. Note that the y-axis scale is different for each plot. The shaded area represents  $\pm 1$  standard deviation of ensemble variability.



**Figure C2.** stZFI for E3SM simulations with prescribed eruptions. The color denotes the relative size of the eruption compared to actual Pinatubo eruption. Note that the y-axis scales are different for the two rows.

## C2 Adding white noise variable

To better convince ourselves that stZFI does not pick up on unimportant signals, we fit an EESN with an additional variable that is simulated from a standard Gaussian across all time and locations; we denote this variable as white noise (WN). Figure C3 shows stZFI for the E3SM ensembles with WN added as an input for the T050 and T2M models. The feature importance for WN hovers close to zero for both models at all times. This result suggests that stZFI is not erroneously thinking it is important for predictions.

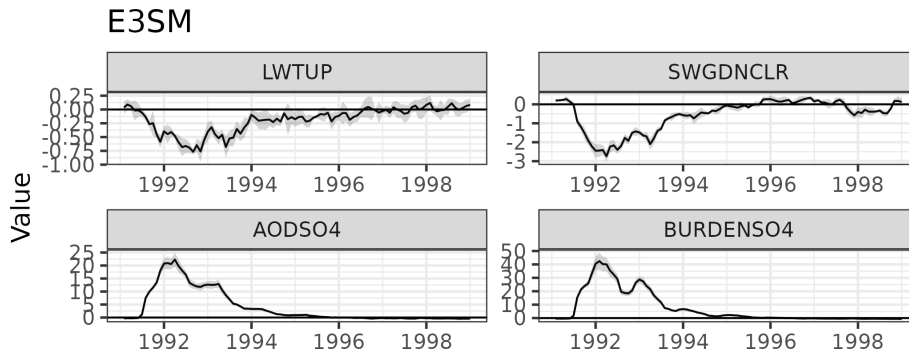


**Figure C3.** stZFI for E3SM with WN added as an input. Note that the y-axis scale differs for the two rows.

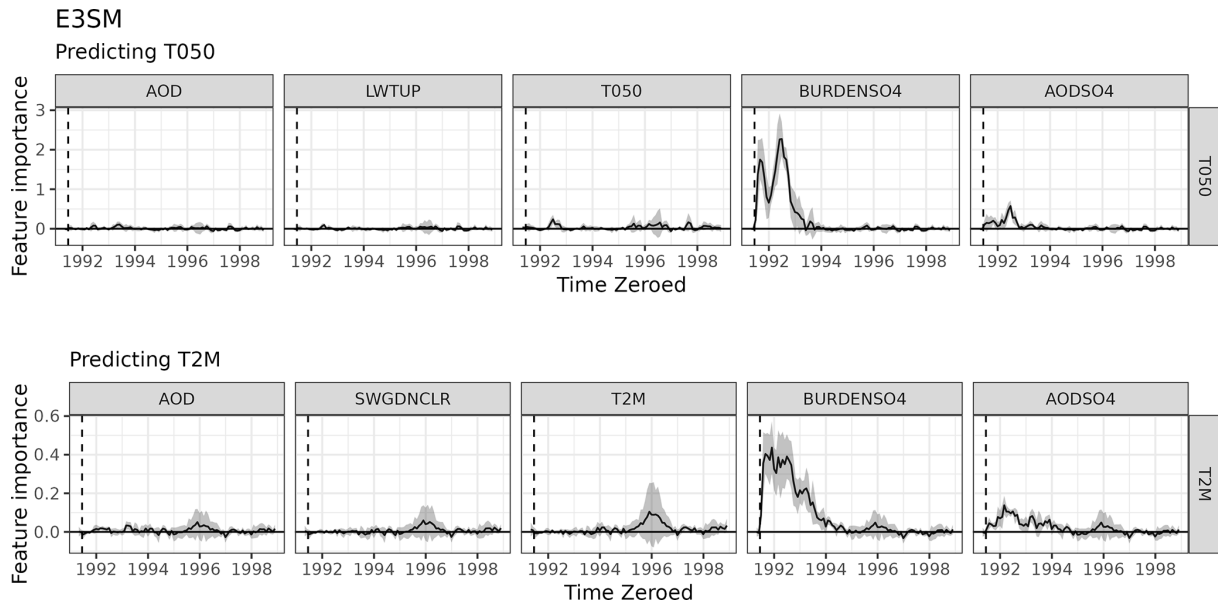
## C3 Adding higher signal-to-noise variables

Figure C4 shows globally averaged normalized anomalies for two additional variables from E3SM, namely AODSO<sub>4</sub>, which is the integrated sulfate aerosol extinction coefficient (absorption + scattering,  $m - 1$ ) at 0.55  $\mu\text{m}$  wavelengths through the entire atmosphere, and BURDENSO<sub>4</sub>, which is the column burden mass of sulfate aerosol. Both of these have a cleaner signal due to Mount Pinatubo than AOD (compare the signal-to-noise ratio of these variables in Fig. 6).

Figure C5 shows the stZFI for the E3SM ensembles with extra variables for the T050 and T2M models. Considering the relative ranking of the signal sizes seen in Fig. C4, the magnitudes of stZFI in Fig. C5 are consistent, with variables that change more after the eruption and are related to the outcome having greater feature importance. This gives evidence to the idea that the feature importance is a metric that looks at the changes in degree of association and changes in feature magnitude. Radiation FI sees a slight decrease compared to the case without the additional variables; lagged temperature sees an even bigger attenuation. This points to the additional variables having a bigger impact. This is potentially due to collinearity between the variables.



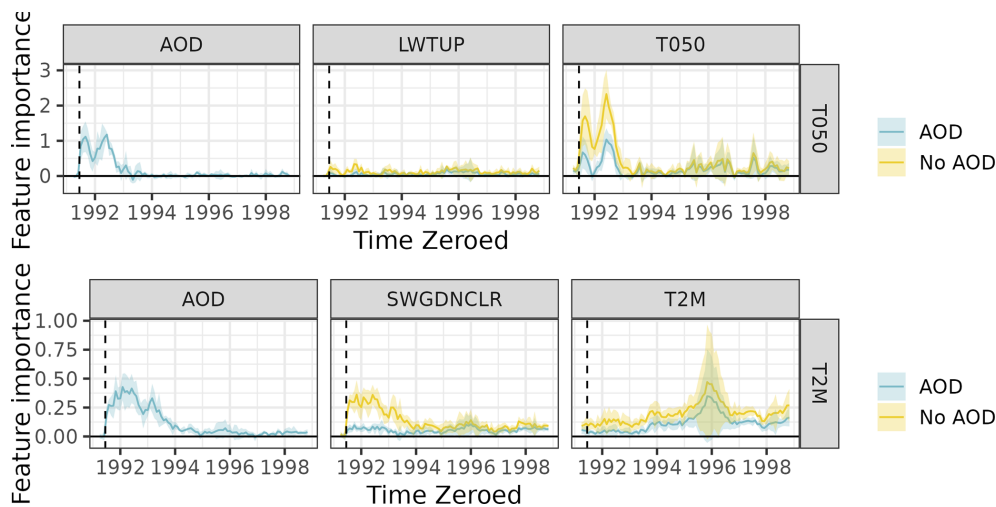
**Figure C4.** Globally averaged normalized anomalies for additional E3SM variables. Note that the y axis is different for each plot.



**Figure C5.** stZFI for EESNs fit using E3SM with additional variables that have a higher signal-to-noise ratio than AOD. Note that the y-axis scale differs for the two rows.

#### C4 Excluding AOD from EESN

Figure C6 examines stZFI for a model that does not include AOD to a model that does include AOD. This moves in the opposite direction of the previous section, where we remove an important variable instead of adding one to see the impact on FIs. This will give us an idea of the collinearity of AOD with the remaining features. Trends between the two are mostly similar, except that stZFI for the remaining variables tends to be higher in the model without AOD, likely meaning that there is collinearity between input variables, and without AOD, other variables account for that relationship.



**Figure C6.** stZFI for E3SM when AOD is not included in EESN. Note that the y-axis scale differs for the two rows.

*Code and data availability.* The HSW-V v1.0 code and data are available from Hollowed et al. (2024a, <https://doi.org/10.5281/zenodo.10524801>). E3SM data and code are available at <https://zenodo.org/records/12169924> (Ries et al., 2024). An up-to-date version of the code and the listenr package is available on GitHub at <https://github.com/sandialabs/listenr/tree/main> (last access: 20 September 2024). MERRA-2 data are publicly available (GMAO, 2015a, <https://doi.org/10.5067/FH9A0MLJPC7N>; GMAO, 2015c, <https://doi.org/10.5067/2E096JV59PK7>; GMAO, 2015b, and <https://doi.org/10.5067/OU3HJDS97300>).

*Author contributions.* DR, KG, and KM developed the methodology. DR, KG, and KM wrote the code to implement stZFI. DR ran the analyses. BH aided in the climatological interpretations. All authors helped review the paper.

*Competing interests.* The contact author has declared that none of the authors has any competing interests.

*Disclaimer.* 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.

*Publisher's note:* Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

*Acknowledgements.* This work has been supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by the 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 no. DE-NA0003525. This paper describes objective technical results and analysis. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility, under NERSC award no. BER-ERCAP0026535.

*Financial support.* This research has been supported by the Sandia National Laboratories (grant no. DE-NA0003525).

*Review statement.* This paper was edited by Rohitash Chandra and reviewed by two anonymous referees.

## References

- Alao, O., Lu, P. Y., and Soljačić, M.: Discovering Dynamical Parameters by Interpreting Echo State Networks, in: NeurIPS 2021 AI for Science Workshop, <https://openreview.net/forum?id=coaSxusdBLX> (last access: 10 July 2023), 2021.
- Ancona, M., Ceolini, E., Öztireli, C., and Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks, arXiv [preprint], <https://doi.org/10.48550/arXiv.1711.06104>, 7 March 2018.
- Armour, K. C., Bitz, C. M., and Roe, G. H.: Time-varying climate sensitivity from regional feedbacks, *J. Climate*, 26, 4518–4534, 2013.
- Baker, A. H., Hu, Y., Hammerling, D. M., Tseng, Y.-H., Xu, H., Huang, X., Bryan, F. O., and Yang, G.: Evaluating statistical consistency in the ocean model component of the Community Earth

- System Model (pyCECT v2.0), *Geosci. Model Dev.*, 9, 2391–2406, <https://doi.org/10.5194/gmd-9-2391-2016>, 2016.
- Banerjee, A., Butler, A. H., Polvani, L. M., Robock, A., Simpson, I. R., and Sun, L.: Robust winter warming over Eurasia under stratospheric sulfate geoengineering – the role of stratospheric dynamics, *Atmos. Chem. Phys.*, 21, 6985–6997, <https://doi.org/10.5194/acp-21-6985-2021>, 2021.
- Bednarz, E. M., Visionsi, D., Richter, J. H., Butler, A. H., and MacMartin, D. G.: Impact of the Latitude of Stratospheric Aerosol Injection on the Southern Annular Mode, *Geophys. Res. Lett.*, 49, e2022GL100353, <https://doi.org/10.1029/2022GL100353>, 2022.
- Bindoff, N. L., Stott, P. A., AchutaRao, K. M., Allen, M. A., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., Mokhov, I. I., Overland, J., Perlwitz, J., Sebbari, R., and Zhang, X.: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, vol. Detection and Attribution of Climate Change from Global to Regional, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, <https://www.ipcc.ch/report/ar5/wg1/> (last access: 18 February 2025), 2013.
- Brown, H. Y., Wagman, B., Bull, D., Peterson, K., Hillman, B., Liu, X., Ke, Z., and Lin, L.: Validating a microphysical prognostic stratospheric aerosol implementation in E3SMv2 using observations after the Mount Pinatubo eruption, *Geosci. Model Dev.*, 17, 5087–5121, <https://doi.org/10.5194/gmd-17-5087-2024>, 2024.
- Butchart, N.: The Brewer-Dobson circulation, *Rev. Geophys.*, 52, 157–184, <https://doi.org/10.1002/2013RG000448>, 2014.
- Cerqueira, V., Torgo, L., and Mozetič, I.: Evaluating time series forecasting models: An empirical study on performance estimation methods, *Mach. Learn.*, 109, 1997–2028, 2020.
- Clare, M. C. A., Sonnewald, M., Lguensat, R., Deshayes, J., and Balaji, V.: Explainable Artificial Intelligence for Bayesian Neural Networks: Toward Trustworthy Predictions of Ocean Dynamics, *J. Adv. Model. Earth Sy.*, 14, e2022MS003162, <https://doi.org/10.1029/2022MS003162>, 2022.
- de Burgh-Day, C. O. and Leeuwenburg, T.: Machine learning for numerical weather and climate modelling: a review, *Geosci. Model Dev.*, 16, 6433–6477, <https://doi.org/10.5194/gmd-16-6433-2023>, 2023.
- Ferraro, A. J., Charlton-Perez, A. J., and Highwood, E. J.: Stratospheric dynamics and midlatitude jets under geoengineering with space mirrors and sulfate and titania aerosols, *J. Geophys. Res.-Atmos.*, 120, 414–429, <https://doi.org/10.1002/2014JD022734>, 2015.
- Gelaro, R., McCarty, W., Suarez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Parityka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), *J. Climate*, 30, 5419–5454, <https://doi.org/10.1175/JCLI-D-16-0758.1>, 2017.
- GMAO (Global Modeling and Assimilation Office): MERRA-2 tavgM\_2d\_aer\_Nx: 2d,Monthly mean,Time-averaged,Single-Level,Assimilation,Aerosol Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], <https://doi.org/10.5067/FH9A0MLJPC7N> (last access: 26 October 2022), 2015a.
- GMAO (Global Modeling and Assimilation Office): MERRA-2 tavgM\_2d\_rad\_Nx: 2d,Monthly mean,Time-Averaged,Single-Level,Assimilation,Radiation Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], <https://doi.org/10.5067/OU3HJDS97300> (last access: 26 October 2022), 2015b.
- GMAO (Global Modeling and Assimilation Office): MERRA-2 instM\_3d\_asm\_Np: 3d,Monthly mean,Instantaneous,Pressure-Level,Assimilation,Assimilated Meteorological Fields V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], <https://doi.org/10.5067/2E096JV59PK7> (last access: 26 October 2022), 2015c.
- Golaz, J.-C., Van Roekel, L. P., Zheng, X., Roberts, A. F., Wolfe, J. D., Lin, W., Bradley, A. M., Tang, Q., Maltrud, M. E., Forsyth, R. M., Zhang, C., Zhou, T., Zhang, K., Zender, C. S., Wu, M., Wang, H., Turner, A. K., Singh, B., Richter, J. H., Qin, Y., Petersen, M. R., Mamejjanov, A., Ma, P.-L., Larson, V. E., Krishna, J., Keen, N. D., Jeffery, N., Hunke, E. C., Hannah, W. M., Guba, O., Griffin, B. M., Feng, Y., Engwirda, D., Di Vittorio, A. V., Dang, C., Conlon, L. M., Chen, C.-C.-J., Brunke, M. A., Bisht, G., Benedict, J. J., Asay-Davis, X. S., Zhang, Y., Zhang, M., Zeng, X., Xie, S., Wolfram, P. J., Vo, T., Veneziani, M., Tesfa, T. K., Sreepathi, S., Salinger, A. G., Reeves Eyre, J. E. J., Prather, M. J., Mahajan, S., Li, Q., Jones, P. W., Jacob, R. L., Huebler, G. W., Huang, X., Hillman, B. R., Harrop, B. E., Foucar, J. G., Fang, Y., Comeau, D. S., Caldwell, P. M., Bartoletti, T., Balaguru, K., Taylor, M. A., McCoy, R. B., Leung, L. R., and Bader, D. C.: The DOE E3SM Model Version 2: Overview of the Physical Model and Initial Model Evaluation, *J. Adv. Model. Earth Sy.*, 14, e2022MS003156, <https://doi.org/10.1029/2022MS003156>, 2022.
- Goode, K., Ries, D., and McClernon, K.: Characterizing climate pathways using feature importance on echo state networks, *Stat. Anal. Data Min.*, 17, e11706, <https://doi.org/10.1002/sam.11706>, 2024.
- Guo, S., Bluth, G. J., Rose, W. I., Watson, M., and Prata, A.: Re-evaluation of SO<sub>2</sub> release of the 15 June 1991 Pinatubo eruption using ultraviolet and infrared satellite sensors, *Geochem. Geophys. Geosy.*, 5, 1–31, <https://doi.org/10.1029/2003GC000654>, 2004.
- Hart, J., Gulian, M., Manickam, I., and Swiler, L. P.: Solving High-Dimensional Inverse Problems with Auxiliary Uncertainty via Operator Learning with Limited Data, *Journal of Machine Learning for Modeling and Computing*, 4, 105–133, 2023.

- Hegerl, G. C., Hoegh-Guldber, O., Casassa, G., Hoerling, M., Kovats, S., Parmesan, C., Pierce, D., and Stott, P.: Good Practice Guidance Paper on Detection and Attribution Related to Anthropogenic Climate Change, edited by: Stocker, T., Field, C., Dahe, Q., Barros, V., Plattner, G.-K., Tignor, M., Midgley, P., and Ebi, K., Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Detection and Attribution of Anthropogenic Climate Change, Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, [https://www.wcrp-climate.org/images/summer\\_school/ICTP\\_2014/documents/IPCC\\_DA\\_GoodPracticeGuidancePaper.pdf](https://www.wcrp-climate.org/images/summer_school/ICTP_2014/documents/IPCC_DA_GoodPracticeGuidancePaper.pdf) (last access: 18 February 2025), 2010.
- Held, I. M. and Suarez, M. J.: A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models, *B. Am. Meteorol. Soc.*, 75, 1825–1830, 1994.
- Hollowed, J., Jablonowski, C., Brown, H., Bull, D., Hillman, B., and Hart, J.: Code Supplement for “Localized injections of interactive volcanic aerosols and their climate impacts in a simple general circulation model”, Zenodo [code, data set], <https://doi.org/10.5281/zenodo.10524801>, 2024a.
- Hollowed, J. P., Jablonowski, C., Brown, H. Y., Hillman, B. R., Bull, D. L., and Hart, J. L.: HSW-V v1.0: localized injections of interactive volcanic aerosols and their climate impacts in a simple general circulation model, *Geosci. Model Dev.*, 17, 5913–5938, <https://doi.org/10.5194/gmd-17-5913-2024>, 2024b.
- Hooker, G., Mentch, L., and Zhou, S.: Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance, *Stat. Comput.*, 31, 1–16, 2021.
- Huth, R.: The effect of various methodological options on the detection of leading modes of sea level pressure variability, *Tellus A*, 58, 121–130, <https://doi.org/10.1111/j.1600-0870.2006.00158.x>, 2006.
- Irvine, P. J., Kravitz, B., Lawrence, M. G., and Muri, H.: An overview of the Earth system science of solar geoengineering, *WIREs Clim. Change*, 7, 815–833, <https://doi.org/10.1002/wcc.423>, 2016.
- Jaeger, H.: The “echo state” approach to analysing and training recurrent neural networks—with an erratum note, German National Research Center for Information Technology GMD Technical Report 148, Bonn, Germany, 13, 2001.
- Labitzke, K. and McCormick, M.: Stratospheric temperature increases due to Pinatubo aerosols, *Geophys. Res. Lett.*, 19, 207–210, <https://doi.org/10.1029/91GL02940>, 1992.
- Lukoševičius, M.: Neural Networks: Tricks of the Trade – A Practical Guide to Applying Echo State Networks, Lecture Notes in Computer Science, 659–686, [https://doi.org/10.1007/978-3-642-35289-8\\_36](https://doi.org/10.1007/978-3-642-35289-8_36), 2012.
- Lukoševičius, M. and Jaeger, H.: Reservoir computing approaches to recurrent neural network training, *Comput. Sci. Rev.*, 3, 127–149, <https://doi.org/10.1016/j.cosrev.2009.03.005>, 2009.
- Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, 4768–4777, Curran Associates Inc., Red Hook, NY, USA, <https://dl.acm.org/doi/10.5555/3295222.3295230> (last access: 19 February 2025), 2017.
- MacMartin, D. G., Kravitz, B., Long, J. C. S., and Rasch, P. J.: Geoengineering with stratospheric aerosols: What do we not know after a decade of research?, *Earths Future*, 4, 543–548, <https://doi.org/10.1002/2016EF000418>, 2016.
- Mamalakis, A., Barnes, E. A., and Ebert-Uphoff, I.: Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience, *Artificial Intelligence for the Earth Systems*, 1, 1–21, <https://doi.org/10.1175/AIES-D-22-0012.1>, 2022.
- Mamalakis, A., Barnes, E. A., and Hurrell, J. W.: Using Explainable Artificial Intelligence to Quantify “Climate Distinguishability” After Stratospheric Aerosol Injection, *Geophys. Res. Lett.*, 50, e2023GL106137, <https://doi.org/10.1029/2023GL106137>, 2023.
- McClernon, K., Goode, K., and Ries, D.: A comparison of model validation approaches for echo state networks using climate model replicates, *Spat. Stat.*, 59, 100813, <https://doi.org/10.1016/j.spasta.2024.100813>, 2024.
- McCormack, C. G., Born, W., Irvine, P. J., Achterberg, E. P., Amano, T., Ardron, J., Foster, P. N., Gattuso, J.-P., Hawkins, S. J., Hendy, E., Kissling, W. D., Lluich-Cota, S. E., Murphy, E. J., Ostle, N., Owens, N. J. P., Perry, R. I., Pörtner, H. O., Scholes, R. J., Schurr, F. M., Schweiger, O., Settele, J., Smith, R. K., Smith, S., Thompson, J., Tittensor, D. P., van Kleunen, M., Vivian, C., Vohland, K., Warren, R., Watkinson, A. R., Widdicombe, S., Williamson, P., Woods, E., Blackstock, J. J., and Sutherland, W. J.: Key impacts of climate engineering on biodiversity and ecosystems, with priorities for future research, *J. Integr. Environ. Sci.*, 13, 103–128, <https://doi.org/10.1080/1943815X.2016.1159578>, 2016.
- McCormick, M., Thomason, L., and Trepte, C.: Atmospheric effects of the Mt Pinatubo eruption, *Nature*, 373, 399–404, 1995.
- McDermott, P. L. and Wikle, C. K.: An ensemble quadratic echo state network for non-linear spatio-temporal forecasting, *Stat*, 6, 315–330, 2017.
- McDermott, P. L. and Wikle, C. K.: Deep echo state networks with uncertainty quantification for spatio-temporal forecasting, *Environmetrics*, 30, e2553, <https://doi.org/10.1002/env.2553>, 2019.
- McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T.: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning, *B. Am. Meteorol. Soc.*, 100, 2175–2199, <https://doi.org/10.1175/BAMS-D-18-0195.1>, 2019.
- Miles, G. M., Siddans, R., Grainger, R. G., Prata, A. J., Fisher, B., and Krotkov, N.: Retrieval of volcanic SO<sub>2</sub> from HIRS/2 using optimal estimation, *Atmos. Meas. Tech.*, 10, 2687–2702, <https://doi.org/10.5194/amt-10-2687-2017>, 2017.
- Parker, D. E., Wilson, H., Jones, P. D., Christy, J. R., and Folland, C. K.: The Impact of Mount Pinatubo on World-Wide Temperatures, *Int. J. Climatol.*, 16, 487–497, 1996.
- Rasch, P. J., Xie, S., Ma, P.-L., Lin, W., Wang, H., Tang, Q., Burrows, S. M., Caldwell, P., Zhang, K., Easter, R. C., Cameron-Smith, P., Singh, B., Wan, H., Golaz, J.-C., Harrop, B. E., Roesler, E., Bacmeister, J., Larson, V. E., Evans, K. J., Qian, Y., Taylor, M., Leung, L. R., Zhang, Y., Brent, L., Branstetter, M., Hannay, C., Mahajan, S., Mamejtanov, A., Neale, R., Richter, J. H., Yoon, J.-H., Zender, C. S., Bader, D., Flanner, M., Foucar, J. G., Jacob, R., Keen, N., Klein, S. A., Liu, X., Salinger, A., Shrivastava, M., and Yang, Y.: An Overview of the Atmospheric Component of the Energy Exascale Earth System Model, *J. Adv. Model. Earth Sy.*, 11, 2377–2411, <https://doi.org/10.1029/2019MS001629>, 2019.

- Reichler, T. and Kim, J.: How well do coupled models simulate today's climate?, *B. Am. Meteorol. Soc.*, 89, 303–312, 2008.
- Ries, D., Goode, K., McClernon, K., and Hillman, B.: Code and Data Supplement for Using feature importance as exploratory data analysis tool on earth system models, Zenodo [data set, code], <https://doi.org/10.5281/zenodo.12169924>, 2024.
- Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature machine intelligence*, 1, 206–215, 2019.
- Sato, M., Hansen, J. E., McCormick, M. P., and Pollock, J. B.: Stratospheric aerosol optical depths, 1850–1990, *J. Geophys. Res.-Atmos.*, 98, 22987–22994, <https://doi.org/10.1029/93JD02553>, 1993.
- Silva, S. J., Keller, C. A., and Hardin, J.: Using an Explainable Machine Learning Approach to Characterize Earth System Model Errors: Application of SHAP Analysis to Modeling Lightning Flash Occurrence, *J. Adv. Model. Earth Sy.*, 14, e2021MS002881, <https://doi.org/10.1029/2021MS002881>, 2022.
- Smith, R. L., Tebaldi, C., Nychka, D., and Mearns, L. O.: Bayesian modeling of uncertainty in ensembles of climate models, *J. Am. Stat. Assoc.*, 104, 97–116, 2009.
- Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O.: Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles, *J. Climate*, 18, 1524–1540, 2005.
- Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I.: Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability, *J. Adv. Model. Earth Sy.*, 12, e2019MS002002, <https://doi.org/10.1029/2019MS002002>, 2020.
- Twomey, S.: Aerosols, clouds and radiation, *Atmos. Environ. A-Gen.*, 25, 2435–2442, [https://doi.org/10.1016/0960-1686\(91\)90159-5](https://doi.org/10.1016/0960-1686(91)90159-5), 1991.
- Wang, C. and Weisberg, R. H.: The 1997–98 El Niño Evolution Relative to Previous El Niño Events, *J. Climate*, 13, 488–501, [https://doi.org/10.1175/1520-0442\(2000\)013<0488:TENOER>2.0.CO;2](https://doi.org/10.1175/1520-0442(2000)013<0488:TENOER>2.0.CO;2), 2000.
- Williamson, D. L., Olson, J. G., and Boville, B. A.: A comparison of semi-Lagrangian and Eulerian tropical climate simulations, *Mon. Weather Rev.*, 126, 1001–1012, 1998.
- Zhou, Y. and Savijärvi, H.: The effect of aerosols on long wave radiation and global warming, *Atmos. Res.*, 135–136, 102–111, <https://doi.org/10.1016/j.atmosres.2013.08.009>, 2014.