Summary of responses

We provide responses to reviewer 2's comments in this document. Reviewer comments are in normal text and in black. Our responses are in blue. We would like to thank both reviewers for time and expertise since their comments contributed to making this a better manuscript.

1 Reviewer 2

1.1 General comments:

Ries et al. (2024) primarily apply the spatio-temporal zeroed feature importance (stZFI), an explainable AI (XAI) tool, to investigate the relationships between various variables associated with a stratospheric aerosol injection event. Notably, this stZFI method can reveal how the feature importance of predictors evolves over time. Utilizing this approach, the authors evaluate the time-variant contributions of volcanic aerosols to the prediction of local and surface temperature. They validate the results with mutiple datasets, including both model simulations and observations, demonstrating that stZFI can identify relationships consistently accross different datasets. This article showcases the capability of stZFI as an exploratory data analysis tool in climate research with great detail and precision. However, I would recommend the authors to devote more effort to explain the stZFI results physically. Please find my comments as below.

1.2 Specific comments:

- Line 9-10: The meaning of this sentence is unclear to me. Here the authors only use feature importance to distinguish between signal related to volcanic aerosols and others, not really natural climate variability.
  The goal of using multiple ensemble members is to be able to run the analysis and have some confidence that the signals that emerge in the ensemble vs CF are real, and not due to natural variability. I.e., if we'd just used a single ensemble member and single CF simulation, we wouldn't be able to say with confidence that any difference between those simulations are real, of if you just got the result you got because you picked a lucky (or unlucky) possible climate state by chance. *The use of perturbed initial condition ensembles introduces variability mimicking the natural variability in the atmosphere, thus the signals emerging using FI can be evaluated against the natural variability in the climate system.*

- Section 1.1: It is necessary to include the possible latitudinal transport of volcanic aerosols driven by the large-scale circulation in stratosphere - the Brewer-Dobson circulation (Butchart 2014).
  We agree that the BD circulation is largely responsible for the poleward transport of Pinatubo sulfate, and we don't believe we are saying anything

to the contrary. The focus of Section 1.1 is to explain the Pinatubo eruption in terms of its magnitude and motivate its use as an exemplar problem. We believe discussion of the BD circulation in detail is beyond the scope of this paper, which is focused on the development and application of a data-driven EDA method that leverages ESMs to gain insights into climate problems. We will add the following to the first paragraph of section 1.1 to make clear the effect BD circulation has on the transport of the Pinatubo sulfates: *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)*

- Figure 1: The movement of aerosols from equator to polar regions could also be driven by the Brewer-Dobson circulation, not only just due to diffusion. Please clarify it.
  We will add the following to line 257: *The injection and spread of aerosols due, in part to the Brewer-Dobson circulation, is clear in latitude and time.*

- Section 2.2.2: Why do you only consider latitude bands for regional contributions? Is there meridional transport of volcanic aerosols? If yes, it would be interesting to give a latitude-longitude global plot showing regional feature importance when T050/T2M peaks.
  We only show latitudinal bands for regional contributions since that is where we see the most variation in temperatures, both surface and stratospheric. We will add the following to line 194: *We focus on latitudinal bands since they account for the most variation in surface and stratospheric temperatures.*

- Line 233: What's the highest height of model outputs?
  Assuming the reviewer is asking about the model top: 0.1mb / 60 km (refer to the E3SMv2 overview paper, `https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003156`). We will add the following to line 233: *Model outputs are remapped to a 2°× 2° structured latitude/longitude grid with 72 vertical levels up to 0.1mb / 60 km.*

- Figure 4: What does negative importance in three subfigures mean? When can people trust that the feature importance from stZFI is reflecting a real relationship?

  Negative feature importance implies the inclusion of the feature in question makes predictions worse than if it had not been included. Often, this will be due to overfitting. Because we are working with spatio-temporal features, it is possible and somewhat common to see feature importance go negative for a brief period of time and space such as in Figure 4. We will add the following to line 278: *Negative feature importance implies the inclusion of the feature in question makes predictions worse than if it*

had not been included. However, small periods of negative stZFI is not a concern, because it is a spatial-temporal metric so it is not unreasonable to expect some time or spatial periods to not be helpful for prediction.

- Line 439-440, Line 448-449: The T2M FI shows a large increase over 1997/98 (Figure 10, subfigure for T2M). Could the increase of FI in this period be caused by the internal variability instead of the volcanic aerosol radiative effect? For example, there is a strong El Niño event from May 1997 to May 1998 (Wang and Weisberg 2000), which could lead to higher autocorrelation in T2M.
  This assessment seems perfectly plausible. In line with a response to a similar comment from R1, we added the following to line 449: *It is possible this upward trend is due to a combination of increasing global surface temperatures and a strong El Niño event from May 1997 to May 1998 (Wang and Weisberg 2000).*