Referee 2:

Bilbao et al have written a description of experiments run with different volcanic forcing data sets. It is important to be able to quickly generate new volcanic forcings for decadal predictions should a major eruption occur. Therefore, it is great that this work has been done and it is written up in a clear and exhaustive manner. I only have one major comment.

EVA_H has some problems with Agung. For people who might be using EVA_H in the future, can any lessons be learned from how the Agung forcing was produced? See minor comments below. It feels a little like this issue is ignored. If not for Agung and a weak ElChichon peak, EVA_H might be a clear choice over EVA. Can its output be improved? It is clearly capable of hemispherically asymmetrical forcings.

Reply: We thank the reviewer for the positive feedback and valuable comments.

We agree that the possibility to specify a hemispherical asymmetry in EVA but not EVA_H is an important limitation. This issue actually recently came up for the production of the CMIP7 stratospheric aerosol dataset. As a quick fix, we simply spreaded the injection of Agung from the tropic to SH mid-latitudes, but implementing an EVA_style asymmetry factor is an important development required. Other EVA_H improvements have been made for CMIP7 and will be documented in the associated papers, but these were brought forward long after we conducted this work. For the peak forcing of El Chichón, we note that both the SO_2 emission that were used in the model and the observations used in GloSSAC are subject to high uncertainty, so we would recommend caution in assessing which model is better there.

Technical comments:

L28 aerosols are transported sediment to the troposphere - I don't understand this

Reply: Rephrased.

L108-110 Why use a 95% range for maps, but a 90% range for time series? Not that it matters, as long as you say what you are doing, but it might be better to be consistent.

Reply: The statistical significance of the differences (DCPP-A minus DCPP-C, i.e. volcano minus no volcano) for both time series and fields is computed in the same way. We create a distribution of 10-member mean differences by bootstrapping with replacement, and produce 1000 draws to determine if the 2.5%-97.5% range includes zero. That is consistent with a 95% confidence interval.

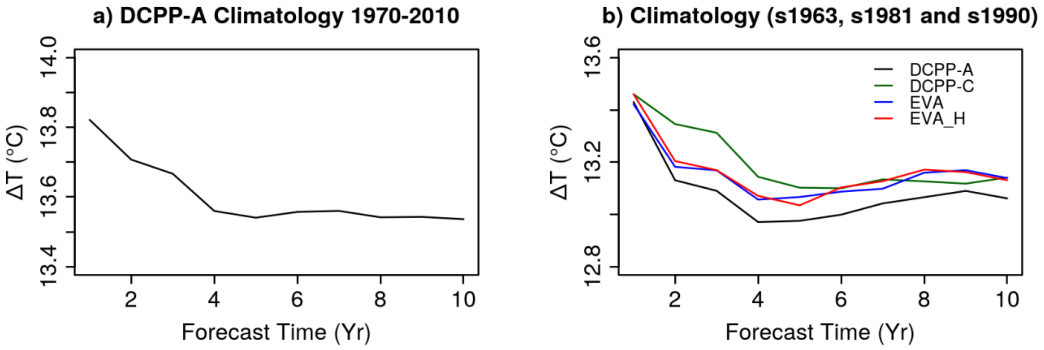
Additionally, for the timeseries, the intra-ensemble spread is indicated by showing the 10th and the 90th percentiles. This has been clarified in the text.

L112 If the hindcast drift is state dependent then the drift will be different for those forecasts that contain volcanoes compared to those that do not. Is this worth mentioning?

Reply: Like in previous papers (e.g. Hermanson et al., 2020; Bilbao et al., 2023; Sospedra-Alfonso et al., 2024), we assume that the forecast drift that develops in the

DCPP-A and DCPP-C predictions is the same. Both are initialised from the same initial conditions and have the same radiative forcings, except for the volcanic forcing.

While the precise impact of the volcanic forcing on the forecast drift cannot be determined using three start dates only, comparing the annual global mean temperature climatology (1970-2010) in DCPP-A with the mean of mean of the three start dates s1963, s1981 and s1990 for each prediction experiment (Support Figure 1) shows a similar evolution of the forecast drift, characterised by cooling in the first 4 years. Despite the differences among the experiments, which are due to volcanic forcing and a small contribution from climate noise, the evolution of the drift is similar, which indicates that the impact on the drift is probably small.



Support Figure 1. a) DCPP-A climatology for the period of 1970-2010 using all start dates. b) Climatology computed using only the s1963, s1981 and s1990 for the hindcast experiments.

This is consistent with Wu et al. (2023), who produced two hindcasts with and without volcanic forcings and show that the SST drift evolution in the central-eastern tropical Pacific is the same in both with a small offset difference associated with the volcanic forcing (see the Supplementary Figure S9A&B), as the one we can see in Support Figure 1b when comparing EVA/EVA_H with DCPP-A.

L120 Is that not the 90% range? (95-5=90)

Reply: Yes, it has been corrected.

L136 It would also be interesting to discuss why EVA peaks earlier and has a higher peak than EVA_H consistently over the three eruptions.

Reply: The two models were calibrated against different datasets (as detailed in section 2.1), both for emission and aerosol optical properties, and with EVA calibrated against Pinatubo only. This resulted in a longer aerosol production timescale explaining the later peak in EVA_H. It also resulted in greater stratospheric aerosol optical depth (SAOD) for the same mass of SO_2 in EVA compared to EVA_H, mostly because EVA_H accounts for both Pinatubo and Cerro Hudson emissions.

L157 Repetition of line 143.

Reply: Removed.

L164 Specify that you are talking specifically about Agung here

Reply: Corrected.

L175 Why did EVA and EVA_H produce such different latitudinal structures for Agung? Could something be done to EVA_H input to improve the outcome?

Reply: Following from the reply at the beginning of the document, while EVA can account for the latitudinal asymmetry with a parameter (indicating the degree of asymmetry in the transport of aerosol to the Northern Hemisphere and Southern Hemisphere), this was not implemented in EVA_H. The hemispheric asymmetry parameter can be estimated for past volcanic eruptions based on observations, however in an operational context the hemispheric asymmetry of a volcanic eruption cannot be anticipated and remains a source of uncertainty.

L181 Again what happened with EVA_H? Why does it not do a descending peak altitude for Agung as it does for the other two eruptions?

Reply: First, we note that although the height of the peak SAOD does not change much for Agung, there is still a shift towards lower altitudes of the extinction values. However, this descend is indeed less clear than for El Chichón and Pinatubo, and the reason is simply the lower SO_2 injection height (18 and 20 km for the two phases) whereas for Pinatubo and El Chichón, we used 25 and 28 km respectively. This means that the injection is already relatively close to the tropopause for Agung. We clarified this in the manuscript.

L313 This is wrong, DCPP-C is a "better prediction", please rephrase as it is not clear what is meant here.

Reply: This sentence refers to the Pinatubo eruption in the different experiments (Figure 10l–o). It highlights that in the DCPP-C experiment (Figure 10m), the observed SST in the tropical Pacific lies within the ensemble, unlike in the other experiments. However, we note that this agreement is probably not for the right reasons, as none of the experiments simulate the observed El Niño. Rather, the DCPP-C experiment is warmer due to the absence of volcanic forcing. This evidences the limitations of predicting tropical variability on these 2-5 year timescales.

Line 334 "absence of […] preferential transport to one hemisphere" - this needs to be expanded on. How does this fit with the asymmetrical forcing shown in Figure 2h? EVA_H can clearly do something. This should perhaps be described here or in the introduction.

Reply: The asymmetry (slightly greater AOD values in the NH with respect to the SH) of the EVA_H forcing for the eruption of El Chichón (figure 2h) is consistent with the asymmetry of the eruption of Pinatubo (figure 2i). Similarly, this latitudinal structure of AOD is also present for the eruption of Agung, when in reality it should be more strongly weighted to the SH.

L343 "weaker as it was calibrated accounting for the Cerro Hudson" I thought it didn't have Cerro Hudson and then this was added in an extra experiment? Please clarify.

Reply: EVA_H was not run with Cerro Hudson in our primary experiment, however this eruption (like any other eruption in the MSVOLSO2L4 SO_2 emission dataset) was accounted

for when calibrating. This means that without accounting for the Cerro Hudson eruption, EVA_H is not expected to reproduce the post-Pinatubo SAOD as well, in particular with a weaker forcing produced. In addition to this important caveat, there is a difference in the observationally based dataset used in the calibration of these models. As mentioned in section 2.1, EVA was calibrated with the Chemistry-Climate Model Initiative (CCMI, Eyring and Lamarque, 2012) satellite dataset for the 1991 Pinatubo eruption, while for EVA_H used the more recent and improved Global Space-based Stratospheric Aerosol Climatology (GloSSAC). Further details on the calibration can be found in Toohey et al. (2016) and Aubry et al. (2020).

Captions: Many of them have "volc - no volc", while this is probably widely understandable, it is perhaps worth writing it out fully for clarity in case the reader is not familiar with the language used in the community.

Reply: Changed to ‘hindcasts with volcanic forcing minus hindcasts without volcanic forcing’.