Response to the comments from Chaochao Gao.

We thank the reviewer for taking the time to consider this long manuscript in details. Detailed replies are given in the following.

Reply to specific comments:

1. 'Section 2.2, please explain briefly the choice of different filter length, i.e., 45 yr for the Antarctic cores and NGRIP while 181yr for GISP2 & NEEM.'

   The filter length choice is empirical, depending on the depth resolution of the sulfate records for the individual cores. The criteria are to ensure that the sulfate background is smoothed for non-volcanic high-frequency spikes and at the same time preserves the abrupt changes across climate transitions to the highest degree possible.

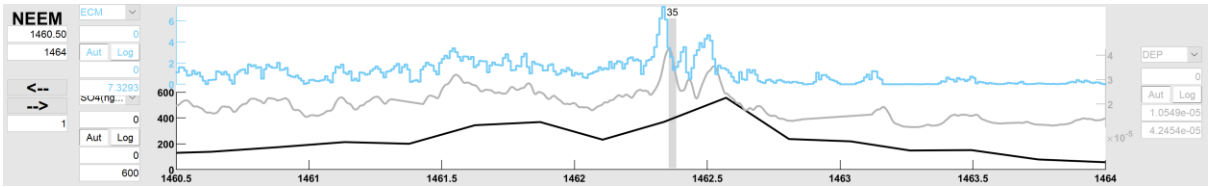
2. 'Also in section 2.2, it is a bit confusing about which parameter was used to measure the volcanic signals, the depth or the yr. Based on the text, the background was filtered by window length indicated by year, while the duration of the volcanic signal was indicated by depth.'

   We use both the original depth scale and an interpolated time scale. The filtering and the event duration are using the time scale, but in order to preserve the maximum resolution we do the integration of the sulfate spike using the original dataset on a depth scale.

3. 'Section 2.4 please provide some description of how the manual correction was performed. For example, what are the resolutions for the ECM and DEP records? How many or what percentage of the signals have been corrected. It would also be great to give an example of how the correction was done'.

   A manual correction was done for multiple volcanic sulfate peaks that are merged into one peak by separating neighbouring peaks according to the corresponding high-resolution DEP or ECM peaks. We added some explanation in the text line 229-233: 'As the depth resolution of the sulfate records for the GISP2 and NEEM cores is relatively low (Fig. S2 (a)), adjacent acidity peaks may be merged into falsely large acidity spikes. We made a manual correction for this effect by comparing to the corresponding higher-resolution ECM (NEEM in 10 mm and GISP2 in 1-5 mm resolution) and DEP (NEEM in 5mm resolution) records of the same core and split falsely large peaks according to the associated ECM or DEP peaks for the top 50 largest events and removed the peaks for smaller events. The specific correction for individual volcanic signals is indicated in Table S3. Twenty volcanic events were corrected for NEEM and 14 volcanic signals were corrected for the GISP2 core.

   One example of this correction is shown in the following figure. The volcanic sulfate peak in NEEM at 1461.3-1461.6 m depth includes three ECM and DEP peaks. We split this suflate peak into three volcanic sulfate signals and assign the sulfate deposition values according to the proportion of DEP or ECM peak areas. This approach assumes that sulfate is the dominant acid contributing to the electric signals.



4. 'A sulfate deposition of 20 kg km^{-2} corresponds to half the Greenland deposition from the 1815 AD Tambora eruption, thus refers to quite large events in terms of total sulfur injections into the atmosphere." Please explain briefly how was the 20 kg km^{-2} (and also the 10 kg km^{-2} for Antarctic) cut-off line estimated. And was the 40 kg km^{-2} 1815 AD Tambora deposition corresponding to the average deposition from the three Greenland cores?'

   The 20 kg km^{-2} and 10 kg km^{-2} cut-offs were applied because it becomes difficult to distinguish the sulfate background variability from volcanic eruptions for smaller events. In Greenland the variability of the background is much higher than that in Antarctica, therefore the cut-off needs to be higher. Another reason for applying the cut-offs is to obtain a dataset that is consistent through the whole

investigate period. Without using a cut-off to identify volcanic events we would detect more smaller events in the most recent part of the records where the temporal resolution is higher. By making a conservative cutoff for the entire profile we homogenize the dataset. The volcanic sulfate deposition 40 kg km^{-2} for Tambora 1815 AD is that obtained by Sigl et al., 2015 and is the average of NEEM-2011-S1 and NGRIP.

5. 'Section 2.6 Please provide more details on the SVM model. For example, what are the requirements, the pro. and cons of the model in this particular application. What validation had been done on the model performance? For example, taken one of the 21 eruption signals used to train the model out from the analysis, could the model accurately simulate its location?'

   More description for this model was added to the text in Section 2.6 and the caption of Fig. S4. 'The volcanic sulfate deposition in Greenland and Antarctica shows a distribution pattern related to the latitudinal band of the eruption site (Fig. 1) (Marshall et al., 2019). To estimate the latitudinal band of bipolar volcanic eruptions of unknown origin, we applied the Support Vector Machine (SVM) classification model of Hastie et al. (2009) and Vapnik (1998). The classification model is based on a kernel function generation and logistic regression. The model was trained using 21 eruptions for which the eruption site is known from tephra deposits in the ice (Table S6). The input values for each eruption to the model are the average Greenland sulfate deposition, the average Antarctic sulfate deposition and the latitudinal band (above 40°N, 40°N-40°S, or below 40°S) of the eruption site. The cross-validation used for tuning the algorithm is 10-fold partition for each evolution. For an optimal classification, a maximum-margin hyperplane was used to separate two classes. The kernel scale and box constraints were chosen for the model and a Bayesian optimization was used to optimize the above two parameters to yield the best classification model (Fig. S4) (more detailed descriptions are in the Hastie et al. (2009), page 17). The bipolar eruptions of unknown origin were predicted into two latitudinal bands – above 40°N (NHHL) and below 40°N (LL or SH) (Table S5) based on the trained model. Due to the low number of known volcanoes erupted in the high latitudes of the Southern Hemisphere, the method does not allow unambiguous identification of eruptions potentially located in this region.'.

   The caption of Figure S4 is now: '(a) The samples (trained + predicted) are classified by latitudinal bands: above 40°N (NHHL) in red '+', below 40°N (LL or SS) in green '*'. The support vectors, that are shown as circles close to the hyperplane, are applied to tune the hyperparameters. (b, c) Bayesian optimization of the model with two parameters (kernel scale and box constraint) yields the best classification model.'.

6. 'Ln 305-307, the comparison between IC and CFA records in Fig S2e needs further demonstration. For example, what is the exact meaning of "very large uncertainties"? What is the implication of the uncertainties on the interpretation of the "face value"?'

   In line 305-307 and Fig. S3 (e), 'the very large uncertainty' estimation is based on comparing the same volcanic sulfate peak in NGRIP ice core as derived by the CFA and IC analytical methods. For this comparison, we can exclude uncertainties related to the ice flow model (layer thinning), as they are obtained from the same core. The average sulfate deposition measured by CFA is around 20% higher than that obtained by IC. This difference may be caused by the different analytical techniques and by the different sample resolutions between CFA and IC.

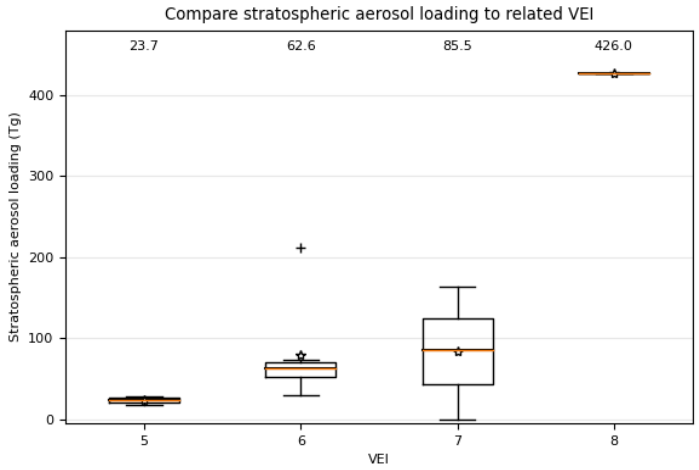
   'Section 3.3 Please explain why borrow the bipolar eruptions from the previous studies, rather than estimate a list using the results from this study.'

   This may be a question related to section 4.3. We compare the magnitude estimates of the glacial eruptions to already published values for volcanic eruptions of the most recent 2500 years in order to set them in a historical context. We are not recalculating the magnitudes of the eruptions of the last 2500 years because the applied methods are similar and because we do not apply the exact same sulfate records for the last glacial as for the historical period. For example, the high-resolution NGRIP CFA dataset only covers the glacial period and is not available for the Holocene.

7. 'Ln 532-534 In my understanding, the VEI list is a discrete (i.e., it is not a complete but continuously evolving) reconstruction of historical volcanism based on geological investigation. So, I am not sure it is appropriate to directly compare the event frequency from the ice-core-based reconstruction (which is assumed to be continuous) with that from geological investigation.'

Indeed, the VEI scale is based on the volume of ejected magma combined with other parameters such as the height of ejected ash cloud (Pyle, D., 2000) and not including the volcanic sulfur emission strength. However, we want to have an impression of the magnitude and frequency of volcanic sulfur emissions from the ice-core-based reconstruction. So, we chose the well-known eruptions (Table S6) and compared their volcanic sulfur emission strength (the stratospheric aerosol loading) to the VEI scale as shown in the following figure. The figure shows that to first order there exists a positive relationship between the VEI scale and the volcanic sulfur strength.

Caption of the below figure: Comparison of stratospheric aerosol loading to the VEI of the same volcanic events. The red line is the median value (values are shown in the upper part) and stars represent the average values. The box lines represent 25 and 75 percentiles. 5 percent outliers are indicated with '+'.



We have added the following discussion in section 4.3 'The above comparison rests on the observation that there exists a positive relationship between the volume of ejected magma and sulfur emission gas for a volcanic eruption.

8. Section 4.4 Please explain why do some events have forcing estimation in a range (for example, the #2 largest signal has forcing ranging from 17.8 to 176.5 W/m2), while others have finite forcing estimation (for example, the #3 largest signal has a forcing estimation of 82.8 W/m2)? If it was due to the number of ice cores available for signal extraction, this should be clarified.

   Yes, that is correct, when we provide a radiative forcing range this means that the eruptions have been detected in several cores. We now state in Table 2 'Number of ice cores' refers to the number of ice cores in Greenland and Antarctica in which the volcanic sulfate signal has been detected.'. This clarification is also added in the section 4.4. 'The largest eruptions of the last glacial period and early Holocene are listed by the average climate forcing in Table 2'.

9. Is there any reason why the authors choose Tambora & Samalas for the magnitude comparison of #3 and #4 events, respectively?

   The comparison to Tambora 1815 AD is because it is a well known and well-studied tropical VEI-7 eruption, and Samalas 1257 AD is the largest tropical volcanic eruption for which the source is known from the last 2500 years.

10. Please add captions for the supplementary figures.

    The captions have been added to the supplementary figures.

    Reference:

    Pyle, D. Sizes of volcanic eruptions in: Encyclopedia of Volcanoes (ed. Sigurdsson, H.), 263–269 (Academic, San Diego, 2000).