We thank the reviewer for their very supportive comment and for the suggestions given to improve our submission.
Reviewer comments in blue italic, our response in black normal.

Figure 2: It would be useful to specify which of the triangles are identified in which cores. Could you split up by NH, Tr, SH – so three lines of triangles that can be linked to the different hemisphere records?

This figure is used to visualize all individual records (at annual resolution) that have been used in the three cited reconstructions. An event-by-event attribution for each of the 18 ice cores split up between 3 hemispheric locations for the close to 100 eruptions would make this figure unreadable. Figure 7 of this manuscript and Figure 3 of the companion paper (https://gmd.copernicus.org/articles/19/3725/2026/) both show CMIP7 upper tropospheric and stratospheric volcanic SO2 emission dataset together with estimated (or known) latitudes.

Line 138 : Where does this equation come from?

This equation is an arbitrary of the mid-tropopshere which has now been acknowledged.

Line 140: What is the error on height estimates roughly?

These errors are typically 10-40%, which is now acknowledged at this place in the manuscript.

Line 149: Are volcanic emissions currently included in the tropospheric forcing? It should be possible to find this out so it would be good to specify if there are instances of double counting.

CMIP provides no recommended dataset for tropospheric volcanic emission, and it is thus not straightforward to answer this question. However, as discussed a few lines below, any double-counting would have negligible consequences.

Line 176: Unclear what this means – what is "estimated at -50%" refering to?

We clarified the formulation used: "In CMIP7, we used the lower-bound value (50% of the best-estimate value) for volcanic sulfate deposition over Greenland for all eruptions (N=9) in the 20th century of the Sigl et al. (2015) dataset."

Line 193-194: Is there a bias then in the period after 1900 that is based purely on Sigl 2015 and doesn't have GISP2? Please clarify if this is an issue or not.

The GISP2 record is geographically close and has comparable snow accumulation rates compared to the other two records (NEEM, NGRIP) so we expect but a minimal bias for those 9 events that are based on the latter two ice cores. This is consistent with a systemic analysis of volcanic sulfate deposition for these three sites across the Common Era (see Figure S4 in Toohey & Sigl 2017).

Line 197: is there any consideration for delay between eruption and deposition?

In this section we only summarize the source data. In this original publication (Sigl et al., 2015) we did not estimate the specific timing of volcanic eruptions but only provide the year in which sulfur started to rise. Assuming a delay of about 6 months between an eruption in the tropics and beginning of deposition across the polar ice sheets (e.g. see Marshall et al., 2018 on the example of Tambora 1815) some of the past eruptions may have occurred in the year before sulfur rises in the ice-core records. However, our deposition-eruption matching (section III, lines 263-338) describes how we attribute the volcanic signals to known eruptions (for which in many cases the dates are known as well.). A seperate section (IV.3 – Attribution of missing eruption dates, lines 420-435) describes how we provide dates for the small number of unmatched eruptions.

Line 207: Unclear here – what is the instead refering to? What was done in other cases or before?

We clarified this sentence: "When specific eruptions date and latitude are provided in eVolv2k for events with a proposed match, we thus ignore these dates and latitudes and instead use the year and latitudinal band derived from ice cores in Toohey and Sigl (2017), before attribution to a historical eruption.". Note that all cases with differing attribution are provided in Table 2 (in addition to all attribution being documented in the dataset metadata).

Line 217: How robust is the assumption of constant accumulation rate at this site between seasons? How much does seasonal precipitation vary seasonally here?

D4 was a short firn core drilled in a few weeks so there is no long-term monitoring of snowfall or snow accumulation available from there. Long-term monitoring of snowfall and its variability has been conducted at other drill sites such as Summit (Castellani et al., 2015), shorter assessments occurred at NGRIP and NEEM. All show that snowfall is continuous year-round with only a muted seasonal cycle in accumulation (with maximum in summer and minimum in winter).

Line 224: what about unattributed eruptions? What is the difference? Is the scaling different?

Unattributed eruptions with the same deposition and latitude would have the same sulfur injection. The point we want to make by comparing three historic eruptions of known magnitudes and unknown sulfur injection with three eruptions in the satellite era of similar magnitude was to show that the range of estimated SO2 injections is comparable to similar-sized eruptions for which we have precise observations.

Lines 227-231: This is unclear what is meant. It needs to be clearer what was published in Feng paper, and what is different here. And what entails a re-assessment. And when the eruption data was a historical date v an ice core date, and the implications of that on the timing (e.g. see question about deposition timing above). These few sentences were very difficult to follow.

We reformulated this sentence to improve clarity, but did not provide more details on the re-assesment there. The aim of these lines is to clarify that we purely used ice-core data from D4i (and eVolv2k) and ignore any source parameter attribution that this study did. Section 3 is dedicated to explaining how we did eruption matching, including Table 2 showing every single eruption matches for which CMIP7 and D4i/eVolv2k eruption match differ with a short rationale. The dataset itself systematically contains rationale for match (and match confidence flag) which we cannot provide for every single eruption in the paper.

Line 250: I think it would be clearer if you instead went backwards in time and found the first point outside the fit.

We went forward because going backward would have led to exclusion of time periods where the fit still looks very reasonable with most data points within prediction uncertainty between the first point outside the fit going backward and the first point outside the uncertainty going backward. For example, the ~1950-1980 period for VEI 3.

Line 257: Maybe include a table of return rates for various VEI and how it compares to previous estimates (and include Brown et al. 2014; Rougier et al., 2016 as well as the Papale papers?).

Given the very minor influence and role played by these parameters in our dataset, we have decided to not include a further table. Agreement with e.g. Papale et al. 2018 is excellent for both return period (unused in our dataset) and time from which significant under-recording occur (used in our dataset for VEI 3 and 4 to provide a match confidence flag) for VEI 3-5. It differs meaningfully for VEI6 which we acknowledge more explicitly:
"Both our suggested time threshold for significant under-recording and eruption return period are in good agreement with Papale (2018) for VEI 3-5, with older time threshold and lower eruption rate for VEI6, which are well recorded for the full Common Era anyway."

Figure 3: It seems like the method to determine when underrecording starts is a bit flawed for the higher VEI values. Consider the VEI5 plot – there is a huge change in slope around 1600. The key thing to spot is the slope, not whether a single point falls outside the regression uncertainty. Similar with VEI6 the change in slope looks to occur around 0 BC. There have been multiple papers that address this (some cited above and in the paper) so could you use their techniques instead?

We agree that these techniques could be used instead and have added a sentence recognizing this:
"Improved consistency with methods used in e.g. Papale et al. (2018) or Rougier et al. (2016) to define the time threshold for significant under-recording should be considered in future dataset versions."
We have not changed the dataset itself because it is already in use for CMIP7 Fast Track, but note that the only parameter that would be affected by differing estimation of the year from which significant under-recording occur are match confidence flags for VEI 3 and VEI 4 eruptions, as it is clear that at global scale, VEI >4 are well recorded over the period considered (1750 onwards).

Line 280: Here the difference between eruption and deposition age is explicitly stated. What is used though? And is it assumed to be the same for all eruptions? What is the uncertainty on that offset and how is that taken into consideration?
What is used to attribute confidence flag is specific in the following paragraph, e.g. flag 3 can be given if the year of deposition is the same year as the eruption match or the year after. There is no assumption it is the same for all eruptions because no eruption date relies on an assumption of a specific time lag, except for the fact that default eruption dates (when unknown) are set to Jan 1 of the calendar year of deposition to account for a lag (see response to comment on line 421 below).

Line 293: "no other VEI ≥ 5 eruption is a low confidence match" – not sure what is meant by this?

We meant "no other VEI ≥ 5 eruption is a potential match"

Figure 4: What are the two high confidence eruptions that fall outside the large uncertainty window? Can these be labelled and is there a reason why that could help us to understand when this prediction might not work well?

We do not understand the comment as these eruptions are labelled. Our initial submission also offers the following explanation: "We excluded the tropical northern hemisphere eruptions of El Chichón (1982, three phases) and Santa Maria (1902) before fitting. These eruptions are clear outliers with α=-1, but this is unsurprising as anthropogenic sulfate complicates the detection of volcanic sulfate deposition peaks in Greenland in the 20th century.".

Line 421-422 – This is confusing how it works if there is a long lag between eruption date and deposition as above. Is that considered at all?

We are unsure what the reviewers exactly mean. A default eruption date at the very beginning of the calendar year accounts for the fact that if deposition is detected within a calendar year, the eruption is more likely to be early that year (or possibly the year before) because of the lag between eruption and deposition start (e.g. 6 months lag for Tambora, see Marshall et al. 2018).

Lines 426- 429: This is confusing because the month is specified here, so why was it assumed a month if month was known?

We clarified as below and it is indeed correct that the typo in the code is not on an assumed month. One line of code supposed to ensure the month calculated from the fractional year was always between 1 and 12 instead resulted in the month always being 12.
"For D4i events with no attributed eruption or unknown eruption date, the month was aimed to be set to the month when deposition was recorded as the measurements have resolution on the order of month. However, a code typo resulted in these unknown eruption dates to be set to December 1."

Line 478: Is mean justified? Normally distributed? What is the range, standard deviation?

There are only four such eruptions so this cannot be assessed. The range is now provided in the manuscript (0.2-7 Tg SO2). The median would be 1.73 Tg SO2.

Line 527: A discussion of why the petrological estimates are so far off of the ice core estimates is warranted.

We added the following sentence: "The discrepancy between these estimates are unsurprising given their uncertainties (34% for ice-core derived estimate, >15% for the petrological estimate, both quoted at 1-σ) and the fact that petrological estimates reflect the total sulfur released, not the sulfur injected high-enough to be deposited in Greenland. "

Line 537: the numbers seem the wrong way around as tephra volume is high then low whereas sulfur load is low then high

Excellent catch, after checking, we accidentally swapped both phases because of an inconsistency between chronological vs reverse chronological order when reading the ice-core and IVESPA data. This is now acknowledged in the manuscript both in this section and in the summary. Not the impacts will be minor given the negligible mass difference, but this mistake will be corrected in future dataset versions. We verified that the same issue does not affect Laki mass partitioning.

Line 545- 550 : It would be useful to specify how they got their S estimates given the large discrepancy between the totals.

This is now specified: "The total masses differ because Niemeier et al. (2019) use a SO2 mass derived from an aerosol optical depth reconstruction, whereas we use an ice-core derived mass."