We would like to thank the reviewers for their valuable contributions. We have compiled a list of all changes below, followed by a point-by-point response to the reviewers‘ comments. Please find our responses in blue and the reviewers‘ comments in black.

List of all changes:

DATA

- In response to the reviewers‘ concerns regarding tuning to volcanic forcing only, we have revisited our tuning and have changed our tuning process to tuning to HadCM3 all forcings simulations. This has not changed any results, only small changes in the fast and the slow response of the response model.

MANUSCRIPT:

- L2: Rephrased the sentence.
- L44: Changed citations.
- L53: Changed wording to „solar activity“.
- L82: Specified wording: simulation error to structural and tuning error
- L110: Rephrased original sentence in response to the reviewers‘ remarks
- L116: Rephrased original sentence in response to the reviewers‘ remarks
- L130: Rephrased original sentence in response to the reviewers‘ remarks
- L139: Added explanation in response to colleague’s comment
- L144: Changed wording slightly to allow for better readability
- L150: Added sentence for clarification in response to colleague’s comment
- L174: Changed „volcanic forcing only“ to „all forced“, to reflect the revised methodology.
- L206: Changed notation from j to t, to make clear that this is the timestamp of the associated residual.
- L228: Added more detail about the reconstruction methods of the N-TREND reconstructions.
- L248: Changed wording in response to reviewers‘ comments.
- L250: Changed wording in an attempt to improve readability.
- L259, 264: Changed wording in response to reviewers‘ comments.
- L361: Added note in response to collaborator’s suggestion
- L440: Changed wording to allow for more concise structure, following reviewers‘ comments.
- L444, 457, 471: Introduced bullet points to improve structure of the conclusions.
- L461: Changed wording to allow for better readability.
- L474: Added reference in response to colleague’s suggestion.
- L483: Added sentence to improve the structure of the section, as suggested by reviewers
- L417: Changed wording to allow for better readability.

FIGURES

- Fig. 1: Added how anomalies were taken to caption.
- Fig. 2: Updated with the new data (following re-tuning to all forcings rather than volcanic only).
- Fig. 3: Updated with the new data (following re-tuning to all forcings rather than volcanic only). Caption: added reference to different colours in response to a collaborator’s suggestion.
- Fig. 4: Updated with the new data (following re-tuning to all forcings rather than volcanic only).
- Fig. 5: Updated with the new data (following re-tuning to all forcings rather than volcanic only).

SUPPLEMENTARY INFORMATION

- All figures in the Supplementary Information were updated with the newly generated data.

Reviewer #1:

Review of: 'The effect of uncertainties in natural forcing records on simulated temperature during the last Millennium'; https://doi.org/10.5194/egusphere-2022-1039

This paper presents an interesting assessment of historical forcing reconstructions in light of paleoclimate reconstructions over the last millennium.

I only have one potentially major comment. At lines 112-114 it appears that the method for adding dating error to the volcanic forcing timeseries appears to treat each date independently. Wouldn't every incorrectly dated eruption have down-core dating impacts? If this is the case, then simply perturbing each year independently is not the appropriate process to use but rather the authors need to be using the 'BAM' model for chronological errors (Comboul et al. 2014, https://doi.org/10.5194/cp-10-825-2014) or something very similar. If the years can indeed be treated independently, then the authors need to clearly explain why this is so in the paper.

While the ice core records from which the volcanic sulfur estimates are derived are layer counted, age-markers are also used to absolutely date points in the chronology. For the time period 1257 CE to present, ice-core chronologies were constrained by numerous historic eruptions and large sulfate peaks, while dates before 1257 were constrained by isotopic anomalies at 775 CE, tying ice cores to tree ring chronologies (Sigl et al., 2015), as well as 3 absolutely dated volcanic events (536 CE, 626 CE and 929 CE). Between age-markers, the dating uncertainty may vary somewhat with the temporal distance from the age markers, as pointed out by the reviewer. However, output from the annual-layer method of Sigl et al. (2015)

suggests an absolute dating uncertainty of better than +-2 years over the past millennium. Some portion of this uncertainty comes from the temporal lag between eruption and deposition to the ice sheets, which does not depend on the age of the event. Given these inputs, we use +-2 years as a conservative (i.e., potentially too large) estimate of the dating uncertainty for all unidentified events rather than adjusting based on lag from a fixed dating point, so as not to overconstrain our results.

Minor comments:

l.244: Should be 'band-pass' instead of 'passband'?

The term passband refers to the frequency band that is allowed to pass through the filter and is frequently used in signal processing. However, to avoid confusion we have changed to “we use a bandpass filter between 50 and 300 years. “.

l.436: 'ahmwhatelse uncertainty [if space]' ???

We are mortified this has slipped our attention and have corrected it!

Figures: I think the figures are very nicely made!

Thank you very much!

Reviewer #2:

Summary: The study explores the consequences of uncertain in the external forcing for the simulated temperature over the past millennium. The authors force a simple climate model with a large ensemble of reconstructions of the external forcing (solar and volcanic) that span the range of uncertainty, and compare the simulated temperature with proxy-based reconstructions.

The main conclusion is that the temperature simulated using small variations of solar forcing better agree with temperature reconstructions

Recommendation: The manuscript is well written and the conclusion is important for the design of paleo simulations with GCMs. I have a few suggestions that the authors may want to consider, most of them related to clarify some technical aspects of the study, and on the structure of the Conclusion section

1) When constructing the volcanic forcing ensemble, it is not totally clear if the 1-sigma volcanic uncertainty is unique to each eruption or is it an average value across all eruptions. Related to this , are the gaussian-distributed uncertainties (or z-scores thereof) added to the central estimate individually for each eruption or as a single time series. Form other description in the text it seems to me that the realizations of the errors for each eruptions are uncorrelated, but it would be helpful if this could be explicitly stated.

line 110 ‘For all eruptions, we perturbed the VSSI amount by a normally distributed random variable of mean zero and standard deviation of the reported VSSI uncertainty’

For all eruptions or for each eruption separately ?

Yes, definitely for each individual eruption, thanks for pointing this out. We have changed this to “For each individual eruption, we perturbed the VSSI amount by a normally distributed random variable of mean zero and standard deviation of the reported VSSI uncertainty for that eruption.”

2) line 46 Despite these latest advances, substantial uncertainties remain in the reconstruction of volcanic forcing from ice core records regarding e.g. timing, magnitude, injection height and latitude of eruptions et al., 2006; Gao et al., 2008; Schmidt et al., 2012a; Crowley and Unterman, 2013; Stoffel et al., 2015; Schneider et al., 2017;Stevenson et al., 2017; Marshall et al., 2021)

The sentence looks a bit strange, since most of the reference are a decade old

Thank you for this observation. We fully agree and have excluded the old references from the sentence.

3) line 48 ‘Solar forcing is primarily driven by photospheric magnetism, leading to varying numbers of sunspots and faculae concentrations on the solar surface, which modulate the total solar irradiance (TSI)

However, prior to the telescopic era, the reconstruction of solar variation is based mainly on cosmogenic isotopes deposited in polar ice cores and tree-rings, of which sunspot numbers can be estimated by applying a chain of physics-based models.’

In my understanding, the paragraph seems to me a bit unclear or misleading. Sun spots are regions of reduced luminosity. The fact that periods with higher numbers of sun spots display higher TSI is because the occurrence of sun spots is correlated with faculae, which display a higher luminosity and have a stronger impact . The link between both is however nonl-inear. . Thus it is not the sun spot number that is actually reconsttructed by physics-based models but directly the TSI.

Yes, this sentence is slightly misleading- thank you for pointing this out. We have corrected it based on your suggestions and replaced it with the more generic wording: “However, prior to the telescopic era, the reconstruction of solar variation is based mainly on cosmogenic isotopes deposited in polar ice cores and tree-rings, of which solar activity can be estimated by applying a chain of physics-based models.”

3) Figure 1 displays the volcanic forcing but the figures also shows slightly positive values. I guess these are anomalies, as for solar forcing

Thanks for pointing this out. We have added this information to the caption, now reading: “Timeseries of natural forcing records, as anomalies over the whole period …”

4) line 110 normal distribution of volcanic uncertainty.

Perhaps this is not terribly important, but a gaussian assumption would lead in some small eruptions to positive values of the volcanic forcing, which is unrealistic.

True! In such cases we set the VSSI to zero, a point we have added to the description of the randomization method a few lines above.

5) line 115 ..’ This procedure was iterated to produce 1000 different timeseries of VSSI, each an equally probable version of past volcanic activity given the estimated values and uncertainties listed in eVolv2k. For each eruption, the eVolv2k-ENS members produce a distribution of potential VSSI amount and timing, with the original default eVolv2k values at the peak of the distribution,representing the estimated most probable value.’I am a bit confused by this paragraph. I would say that the ensemble is a sampling of the underlying probability distribution, but I do not think that each member is equally probable. The probability of a number in the case of continuous distribution is not really defined , only the probability density. Also, the part of the sentence stating that the original eVolv2k represents the most probable value is in contradiction with the statement that each member is equally probable.

Thank you for this useful comment. We definitely agree that with continuous distributions, the probability of any particular number is infinitesimally small, so it would be more accurate to phrase things in terms of probability density. Regarding our statement regarding the equal probability of each member of the ensemble, our reasoning had two parts. Firstly, if we take analogous situations sampling from discrete distributions, under uniformly distributed sampling, we know that the probability of each unique sample is the same, e.g., when flipping a coin 5 times, the probability of HHHHH is the same as HTTHT. Now, if we move to sampling from a normal distribution, it is true that any single sample will have a larger or smaller probability (whether the sample happens to come from the center or the tails of the distribution), but our assumption was that if one takes a large enough set of samples, the probabilities of the different sets of samples should converge and be approximately equal. We assumed that our number of eruptions was large enough for this approximate convergence of probabilities. That being said, the “equal probability” is not an important part of the procedure, and to avoid inaccuracies or the need for some proof, we have modified the statement to refer to each ensemble member as a “possible version of past volcanic activity…”. Finally, we believe that the original eVolv2k does represent the most probable value *for each eruption individually* (as stated)–it represents the peak in the pdf. On the other hand, the time series of the eVolv2k VSSIs is not more likely than any of the randomly perturbed time series–it would be the result of running an random number generator and getting 0 for each eruption, which is no more or less likely than any of the other specific sets of random numbers.

Modified text:

This procedure was iterated to produce 1000 different time series of VSSI, each a possible version of past volcanic activity given the estimated values and uncertainties listed in eVolv2k. For each individual eruption, the eVolv2k-ENS members produce a distribution of potential VSSI amount and timing. The original default eVolv2k values are found at the peak of the distribution, representing the estimated most probable value for each individual eruption.’

6) line 121 ‘The volcanic forcing ensemble therefore represents a best estimate of the range of possible volcanic#

In which sense ‘best estimate’ ? I would say it is just a sample from the distribution. A second sampled can be drawn and this will be different form the first. both cannot be the best estimate.

Thanks for pointing this out, “best estimate” can take different meanings and this is unclear here. We have removed “best” to simplify the statement to “The volcanic forcing ensemble therefore represents an estimate of the range of possible volcanic…”

7) line 255 For N = 20 years all models are consistent within the lower and upper quartile of the population, showing that most of the models roughly agree on the extent of decadal variability

Which population ? It cannot be the population of models, as only 50% would be within the lower and upper quartile.

Thanks for this great question. This has perhaps not been phrased in the best possible way. We have rephrased it in the main text, and explained in depth below:

Here, I am referring to the population of the root sum of square of 20 year slices of the control runs. This is shown in figure S11a. Each violinplot includes the data for an individual model. If we compare the individual models to the distribution of all data (ALL), we find that the median of all but one individual models agrees with the lower and upper quantile of ALL. Or, as described in the text, that the lower and upper quartile of each individual model's distribution agrees with the lower and upper quantile of the complete distribution. This suggests that the variability between the different models may be an artefact of internal variability, and given that every control run simulates a different scenario of internal variability we do expect a degree of variability across the different models, but to have an agreement around the mean value.

8) line 435 'In this study, we have, for the first time, estimated the effects of both volcanic and solar forcing uncertainty on simulated temperature, with volcanic forcing uncertainty including magnitude, timing and ahmwhatelse uncertainty'

Indeed life is full with uncertainty but we should not despair

Thank you for your humorous take on this mishap. The sentence has been edited!

9) The discussion and conclusion section is rather comprehensive. I would suggest to include some structure into it, for instance by highlighting one conclusion as a bullet point followed by the discussion related to it.

We appreciate that this section has gotten a little bit lengthy, and thank the reviewer for their suggestion on how to address this. We have revisited the section, and while we did not find much scope to shorten it, we found that using bullet points to highlight the separate conclusions have significantly improved the readability and the flow of this section. We have also more clearly separated the conclusions from the discussion. We hope this addresses the reviewer's concerns.

10) Why was the impulse response model fitted to a volcanic-only simulation instead of a full forcing simulation? This choice seems not totally logical, and ma raise the suspicion that the better agreement of the low-amplitude solar forcing with the proxies may be rooted in this choice ? If the solar forcing has a minor importance, the tuning would not be very different. Perhaps the difference of the model parameters could be included in the ms.

This is a very good point. While we have explained the reasoning of our strategy in line 167 ("We use the HadCM3 volcanic only simulations as the target for tuning to ensure an optimal choice of the fast response, which we found to be the most critical parameter for simulating the the pre-industrial millennium."), we acknowledge that this argument relies on the findings of previous studies, that volcanic forcing is the dominant driver of variability. Thus, we understand that readers may conclude that perhaps there is a certain degree of circularity in this strategy.

However, we found from previous simulations tuned to all forcings, rather than volcanic only, that there is no difference in results between these two tuning methods.

Thus, in order to make sure that our study stands up to the highest levels of scrutiny, we have decided to rerun our simulations with the tuning parameters found by tuning the all forced response model to the all forced simulations in HadCM3. As mentioned before, this has not made a difference to any results, but we hope will increase the credibility of our conclusions.