The authors' replies to the referees comments can be found *in italic* below each comment.

REPLIES TO REFEREE # 2:

This manuscript argues that Mt. Pinatubo's 1991 eruption had little to no impact on continental surface temperatures, and hence observed surface warming in midlatitude winters was due to natural variability. The implication is that similar conclusions may hold for other eruptions. Indeed, there has been perhaps excessive focus on explaining and simulating the observed Pinatubo response, without sufficient regard to the inherent role of natural variability.

This manuscript contributes to ongoing discussion by frankly addressing the issue of natural variability, and its novel employment of a coupled atmosphere-ocean-chemistry model for volcanic simulation is a useful addition to the literature, even if having only 13 members for that model leads to difficulties with statistical significance. After tempering the overall claims and investigating a lower-stratospheric pathway as detailed below, this manuscript is suitable for publication.

1. General comments:

1. The text is too quick to dismiss temperature reconstructions. Despite the inherent uncertainties of temperature reconstruction, the key is that averaging over several centuries reveals a statistically significant pattern of winter warming, apparently even stronger for the subsequent winter (Fischer et al., Geophys. Res. Lett. 2007), which would be highly coincidental if volcanic eruptions were unrelated.

Thanks for the suggestion: we now cite Fischer et al (2007), with the appropriate caveats.

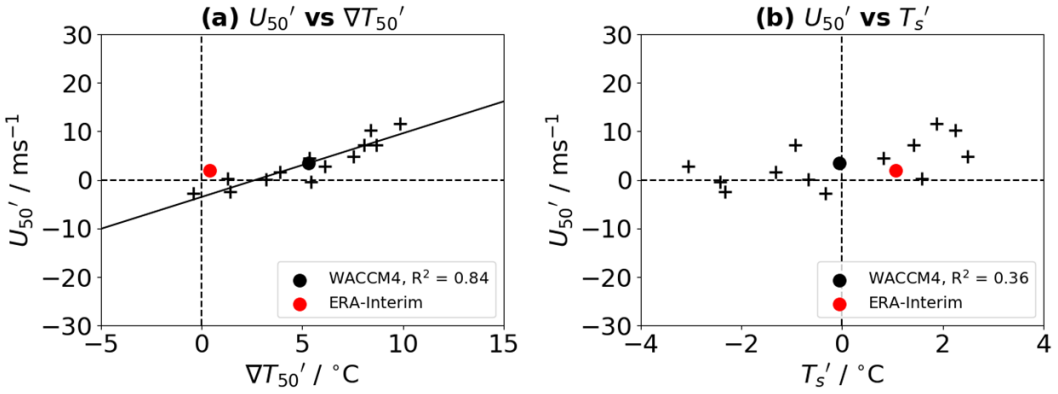
2. P5 L19 and elsewhere describes the ensemble sizes as "large" (13, 42, and 50 members). However, P3 L12 mentions that Bittner et al. (2016) needed 60 members for 95% confidence in the stratospheric response, and other comparable examples are mentioned. Thus "large ensemble" seems incorrect a priori, so the difficulties throughout in achieving significance should not be surprising.

The term "large ensemble" is widely used in the literature to refer to the both the CAM5-LE and CanESM2 datasets. We noted in the manuscript that our relatively modest 13-member WACCM ensemble might not be best described as "large": however, it is the largest ensemble of chemistry-climate model integrations studied to date. This is why we use the term here.

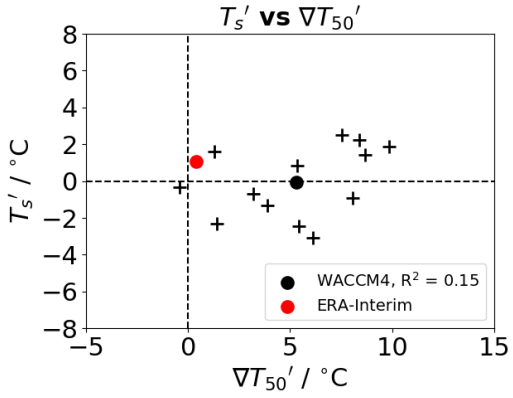
As for the lack of a surface temperature response: we find none in our 13-member WACCM ensemble, and neither did Bittner et al (2016) in their larger 100-member ensemble (see figure 6.4 of this thesis). And neither did Driscoll et al (2012) using many CMIP5 model runs. So the "difficulties in achieving significance" are very robust result across the literature, and have little to do with the relatively small size of our WACCM ensemble.

3. To address the underlying mechanism(s), the manuscript uses a 13-member ensemble with an improved stratosphere to argue against a pathway between stratospheric vortex perturbations and surface temperature perturbations. However, the discussion does not address the possibility of a lower-stratospheric pathway, which from Fig. 7 seems plausible. This could be important if the impact on the vortex does not have the same vertical structure as natural variability. Addressing this would be straightforward by repeating the analysis of Fig. 5 at lower levels such as 50 hPa.

Following the referee's suggestion, we have recomputed Fig. 5 using U at 50 hPa (see below). It is very similar to the one using U at 10 hPa, which we included in the paper. The mean surface temperature response at the surface is zero (the black dot in the right panel), with 7 members showing cooling and 6 members showing warming.



Furthermore, we combine the two scatter plots into one, to directly show that there is no connection between the temperature gradient at 50 hPa (which is affected by volcanic aerosols) and the Eurasian surface temperature (which is not). This is shown below.



We hope this will suffice to convince the referee that the wind and temperature-gradient anomalies in the stratosphere (whether upper or lower) are not the cause of the surface temperature anomalies. This is the key point of our paper: there is no mechanism to be explained. Internal atmospheric variability suffices to produce the surface anomalies.

2. Specific comments:

P1 L17: "is likely to be very small" is unclear – it was likely very small based on these model results?

Thank you for the suggestion. We have added "in our models" to clarify the sentence.

P2 L1: "short lived" is relative; P6 L12 cites an e-folding time of 12 months, longer than most natural modes of variability.

We are using the expression "short lived" as it refers to a forcing of the climate system. The volcanic forcing is short-lived compared to anthropogenic forcings (e.g. the multi-decadal increase in carbon dioxide) or the solar forcing (whether the 11-year solar cycle or longer fluctuations of the solar constant).

P2 L16–25: The suggestive tone is biased. In the "widely believed" (P1 L1) viewpoint of winter warming, it should not be "remarkable" that the subsequent winter after Pinatubo "happened to be" warm, nor is it "highly perplexing" and "difficult to reconcile" that historical warming is not exactly correlated with eruption magnitude. Rather, the question is whether eruptions of a given strength can induce a statistically and physically significant winter warming. Missing here is Fischer et al. (Geophys. Res. Lett. 2007), which should be cited here as the most recent (known to this reviewer) post-eruption temperature reconstruction.

As mentioned above, we have now cited the Fischer et al (2007) paper, with the appropriate caveats. As for the biased tone: we are simply saying that one expects surface cooling from a strong volcanic eruption, so any surface warming is surprising to us. Perhaps we are too naive, but naivete is not bias: we are simply offering our viewpoint.

P3 L5: An implicit assumption of this mechanism is that the balanced acceleration lies in the vortex region, which is not necessarily the case. Bittner et al. (J. Geophys. Res. Atmos. 2016) discusses this.

That may be the case, but we are simply summarizing the "widely believed" viewpoint as stated, e.g. by Robock (Science, 2002), where one can read: The polar vortex is strengthened by lower stratosphere warming at low latitudes, which is caused by absorption of solar and terrestrial radiation by the volcanic aerosol cloud.

P3 L13: "tiny" is relative; perhaps relate this to annular mode, standard deviation of lowpass-filtered winds, or similar. See also comment for P9 L13–15.

We have rephrased the sentence.

P3 L19: Stenchikov et al. (2002) had only 4 ensemble members, and argued for a reduction in planetary wave activity, contrary to what Graf et al. (2007) said about a single eruption. Perhaps this paragraph should conclude that the mechanism is not demonstrated by these single-model, small-ensemble studies. Remove "clearly", "abundantly clear," etc. for P3 L21, P7 L24, P8 L5, P9 L34, P10 L11, P10 L29, P11 19, P12 L22. The word doesn't enhance the argument and may come across as proof by intimidation.

We have removed most instances of the word "clearly" where suggested by the reviewer.

P4 L11–14: should also cite Barnes et al. (J. Clim. 2016), which does find a significant response. Thus even when experimental design and intermodel spread are controlled for, the result can still vary!

Thank for the suggestion. However, Barnes et al. (J. Clim. 2016) find that the CMIP5 model response of the circulation in the NH projects very poorly on the annular mode (which contradicts with the originally proposed mechanism); also they do not explicitly examine Eurasian temperatures in wintertime.

P5 L30–P6 L7: a common limitation, here and in many other studies, is that it is unknown whether or not the response is linear. (Perhaps a threshold magnitude of forcing is necessary, or stronger forcing induces feedbacks.) This limitation should be stated, as the conclusions for these Pinatubo-sized eruptions may not hold for smaller or larger eruptions.

We agree with the reviewer. We have added a sentence to that effect.

P7 L23: ensemble averaging reduces, but does not eliminate, internal variability.

Thanks for noting this. We have corrected the sentence.

P7 L25: in F3, the larger ensembles have slight windows of significance. An estimate (even via simple bootstrapping) of the necessary number of ensemble members to achieve continental-scale significance would be very helpful for this discussion and for future studies.

The areas of significance are non existent for WACCM and minuscule for the two low-top models. Also, recall that Bittner (2015) reported no significant warming over Eurasia, even with a 100 member ensemble. If hundreds of members are needed to produce a significant warming, the suggested bootstrapping calculation is an academic exercise.

P7 L35: internal variability is not superimposed to any forced response – it may very well be non-linear (i.e., the higher moments of the underlying probability distribution functions may change).

Following Deser et al (2012), we decompose the anomalies in any one realization as a sum of a forced response (defined as the ensemble mean anomaly) and the internal variability (the difference). Hence our use of the word "superimposed". This procedure is standard, we think, in all papers that have analyzed large ensembles of model simulations.

P7 L11 and F4: rather than a box-and-whisker plot, a plot of the 3 probability distribution functions is preferable here in my opinion, so that the reader can compare the distributions.

The whisker plots are PDFs. They show the mean, the percentile ranges and the full extent of each ensemble. Also, in Fig 4 the y-axis is identical: the reader can immediately and quantitatively compare these three distributions.

P8 L25–26: the other two models may not have as accurate a representation of the stratosphere, but do they give similar results? If so, they should be included. If not, why is the subsequent comparison with Bittner et al. (2016) (which similarly has a non-interactive stratosphere) valid but not with the other two?

The CAM-LE and CanESM are "low-top models": as such they do not simulate the observed stratospheric variability, e.g. Stratospheric Sudden Warming events. They are, therefore, inappropriate for examining the stratospheric pathway. This is why we focus our discussion on the WACCM ensemble in the paper.

Nonetheless, following the referee's suggestion, we now have added to the supplementary material the equivalent of Fig. 5 for the other two modes. The results are similar, and the very fact these low top models give similar results to WACCM is, of itself, a demonstration that the stratospheric pathway is not needed to explain the surface warming anomalies.

P9 L13–15: it is not appropriate to compare weekly SSW variability with volcanic forcing as their timescales are well-separated. The appropriate comparison would be some- thing like variability of DJF average, which is approximately 10 m/s at 10 hPa and 6 m/s at 50 hPa, more comparable to the 3.5 m/s reported here.

We politely disagree. The point we are making is that even SSWs, which are huge disruptions of the stratospheric circulation, are often incapable of reaching the surface. Hence, it is difficult to imagine how a 1-2 m/s wind anomaly from Mt. Pinatubo would result in strong Eurasian warming. That amplitude is too small, and the accompanying surface signal is swamped by the internal variability.

P9 L16–17 and F5: It might be helpful to add a third scatter plot of a lower comparison point like ∇T_{50} and T_s, which may correlate better than 10 hPa, if the perturbation is comparatively larger than natural variability.

This suggestion has been addressed above, in the "General Comments" section.

P9 L23–34: examining two individual ensemble members does not offer any insight into the mechanism, especially since the manuscript already argues that natural variability is large. This paragraph and the corresponding F6 should be removed.

We politely disagree. There is much recent literature on understanding internal climate variability using large ensembles, and showing two individual members of the ensemble is the simplest and most immediate way to visually convey the importance of internal variability, which often overwhelms the forced response. This was done, for instance, in the papers below, just to cite a few examples.

- *Deser, C., R. Knutti, S. Solomon, and A. S. Phillips: Communication of the role of natural variability in future North American climate. Nat. Clim. Change (2012)*
- *Deser, C., et al.: Projecting North American Climate over the next 50 years: Uncertainty due to internal variability. J. Climate (2014)*
- *Deser, C., J. W. Hurrell and A. S. Phillips: The Role of the North Atlantic Oscillation in European Climate Projections. Clim. Dyn. (2017)*

P10 L7: even if 10 hPa is "canonical," it may be the wrong level for finding the mechanism. Repeating the same analysis at a lower level, such as 50 hPa, would either strengthen the current null-hypothesis argument or provide new insight into the mechanism's vertical extent.

This suggestion has been addressed above, in the "General Comments" section.

P10 L14–16: "quite likely" and "would have emerged" are purely speculative and should be removed, as the null hypothesis was not rejected by the significance test. Instead, a simple bootstrapping estimate of the requisite number of samples to achieve significance may again be helpful.

Thanks. Bittner et al (2016) showed that a 100-member ensemble yields a significant vortex response. This is what we are referring to. We have rephrased that sentence.

P10 L17–24: again, the possibility of a lower stratospheric pathway should, and could easily, be addressed here with the existing methodology.

This suggestion has been addressed above, in the "General Comments" section.

P11 L28–30: are their low model tops thus an indirect argument for a lower stratospheric pathway?

To the contrary! Poor vertical resolution in the stratosphere and a low model top suppress stratospheric variability, giving the false impression that the forced response is dominant. As models have added more levels and raised the top over the years, the forced surface warming has disappeared (in the CMPI3 and CMIP5 models).

P10 L25 to P13 L4: the conclusions should be updated following any relevant changes made as a result of these comments.

There have been no relevant changes we are aware of.

Technical comments:

P8 L6: "stonger" should be "stronger"

Fixed. Thank you.

P8 L11: "Fig. 5" should be "Fig. 4"

Fixed. Thank you.

P9 L12: "need" should be "needed"

Fixed. Thank you.

F4: "nbox" should be "box"

Fixed. Thank you.

F5: "R_2 " in legend of subplot (a) should be "R^2 "

Fixed. Thank you.

REPLIES TO REFEREE # 3:

The paper deals with a longstanding issue of the inability of climate models to reproduce the high-latitude near-surface winter warming following the major low-latitude volcanic eruptions. I appreciate the authors have risen this issue again using a set of the "new generation" models. I believe the reviving this issue is useful but I cannot completely agree with some interpretations and methodology the authors use in this study.

1. General comments:

To test the mechanism based on the troposphere-stratosphere dynamic interaction, the authors conducted the Pinatubo case study focusing on the first winter after the June 1991 volcanic explosion in the Philippines. However, the choice of the case-study is unfortunate as in the winter of 1991/92 the positive AO was not forced by the "stratospheric" mechanism. In observations, the polar vortex was weak and asymmetric with the wave number 2 prevailing. So, it is pointless to analyze this response to prove or disprove the stratosphere/troposphere dynamic interaction mechanism. Stenchikov et al. (2004) indicated that the easterly QBO phase in winter of 1991/92 weakened the polar vortex, and winter of 1992/93 with a westerly QBO phase provides a better case-study to test the "stratospheric" mechanism.

As the title of the paper makes clear, our goal is to understand what happened over the NH continents in the winter following the Pinatubo eruption, and to reconcile models and observations. Pinatubo is the largest, most recent, best observed, low-latitude eruption: as such it needs to be understood before any other, much older and poorly observed eruptions. In fact, it is routinely used as the "poster child" for the impact volcanic aerosols on NH continental temperatures, e.g. Robock (Science, 2002).

The prevailing narrative has been that the "models are missing something" as they show no surface warming after averaging many runs (of the same or of different models). Our paper argues that this an erroneous interpretation. The model average shows no warming because there is no significant warming response. *It's that simple.*

The referee suggests that Pinatubo is a bad choice to test the "stratospheric mechanism" because the positive AO was not forced by that mechanism after that eruption. We agree: is was not not forced by the "stratospheric" mechanism and, in fact, it was not forced by any other mechanism. It was not forced at all. It was just variability. This is what our analysis of the three large ensembles very clearly demonstrates.

As for the possible role of the QBO. First, recent studies (Bittner et al 2016, Robock and Zambri 2016) are agreed that only the first winter after the eruption should be averaged: that is when the aerosol presence is largest. Second, the simple fact that the QBO phase can wipe out the volcanic signal confirms our claim: that internal variability (of which the QBO is one aspect) overwhelms any forced response (should one exist). With large ensembles we can actually quantify the forced response, and we find that is is small and confined to the stratosphere. This is what Bittner et al (2016) also found.

As it is correctly stated in P8, L22-24, the "stratospheric" mechanism involves two steps: strengthening of the stratospheric polar vortex and downward propagation of the signal. The proof of the latter portion of the mechanism did not come directly from "volcanic" studies, as volcanic eruptions are rear and provide insufficient statistics, but from climatological studies of Baldwin and Dunkerton (1999). As mentioned by Stenchikov et al. (2006) the strengthening of the polar vortex caused by the equatorial lower stratospheric warming due to aerosol-induced heating, is robust in the models, but the models fail to reproduce the downward transport. So, to disprove this "stratospheric" mechanism the authors have to deal with the climatological analysis as well.

> *The strengthening of the polar vortex was* **not** *observed the first winter after the Pinatubo eruption, and yet surface warming* **was** *observed. Therefore that surface warming could not have been caused by a stronger polar vortex (see Fig. 8).*

It is not surprising that some of the model ensemble members could produce a "winter warming" pattern. It is more important how frequently this pattern appears and what mechanism causes it. Models have to produce this pattern more frequently to be consistent with the climatological studies that show a statistically significant positive AO pattern after compositing multiple equatorial eruptions. The conclusion that the up-to-date models could perfectly reproduce the winter warming based on the fact that some ensemble members capture it, is not supported.

> *The main finding of our paper is that the models do not produce the winter warming more frequently after the Pinatubo eruption, because the ensemble average is zero. Warming happens half of the time, as a simple consequence of internal variability.*
>
> *We did not conclude that "models could perfectly reproduce the winter warming based on the fact that some ensemble members capture it". Our key finding, after analyzing three large ensembles, is that the observed warming falls well within the distribution of the model members. From this we conclude that the models capture the observations.*

2. Specific comments:

P3, L13-15: A vertical propagation of the planetary waves is a threshold process as suggested by Charney and Drazin (1961), so small change of the wind could qualitatively change the planetary wave reflection coefficient.

> *This is a linear result from highly idealized theoretical studies. Whether and how it may be relevant in practice remains to be demonstrated.*

P4, L26-27: AO response is an atmospheric effect. Why increasing of model complexity should matter to answer the question that the Stratosphere-Troposphere Interaction is real? E.g., if ozone additional radiative effect matters, this has to be specifically shown.

> *The literature on the impact of stratospheric ozone the annular mode is quite large. The referee might wish to consult, e.g. Thompson et al, (Nature Geoscience, 2011).*

P5, L23: The chosen models are inconsistent in reproducing the aerosol forcing. In Figure 2 the aerosol forcing in the models differs by 50%. It would be useful to mention what was the observed forcing to compare with.

There is nothing peculiar about the models we have analyzed. They are of the same kind as those analyzed in Driscoll et al (2012) or Robock & Zambri (2016). In Fig. 2 we show the ERA-Interim temperature time series: one can see that WACCM and CAM5-LE simulate excessive warming, a common bias, as noted by Driscoll et al (2012).

P5, L30-33: The winter of 1991/92 after the Pinatubo eruption is a wrong choice (see Figure 5). A "composite" approach has to be considered to obtain statistically significant anomalies in observations.

See our reply to this in the General Comments section above as to why Pinatubo is not the wrong choice. As for adopting a "composite" approach: we have done so, by averaging all members of each ensemble. What we find, in agreement with Driscoll et al (2012) and Bittner (2015), that the surface response is not statistically significant. There is no reason to expect, a priori, that the surface anomaly should be significant, unless one thinks that internal variability is small, which is not the case (see Fig 3).

P6, L2: This is incorrect. The eruption of Mt Agung of 1963 developed an aerosol equatorial reservoir that caused warming of the equatorial lower stratosphere and enhanced equator-pole temperature gradient in the lower stratosphere. The re-distribution of aerosols between the hemispheres is not directly relevant.

We politely disagree: the hemispheric distribution is likely to matter. In any event, we have not analyzed the Mt Agung eruption, so the whole point in nugatory.

P6, L8: The first winter is a wrong choice.

We politely disagree: the aerosols forcing in the stratosphere is largest in the first winter.

P6, L12: Volcanic aerosols remain in the equatorial reservoir in the second winter after the eruption that is why the effect is seen in the second winter as well.

We politely disagree. As already pointed out, others have also concluded that only the first winter should be analyzed (Bittner et al 2019, Robock and Zambri 2016). The amount of volcanic aerosols in the winter 1992-93 was only a small fraction the one in 1991-92: see, for instance, Figure 3 of Stenchikov et al (JGR, 1991). It make no sense to average a large and a small forcing: that simply washes out the signal.

P6, L18: Driscoll et all. (2012) adopted this methodology from Stenchikov et al. (2006).

Thank you for pointing this out. We have now added the Stenchikov et al (2006) paper.

P7, L8–9: The shortwave (SW) radiative forcing in three chosen models differs by 50%. There is much more differences in SW and Longwave (LW) aerosol absorption.

Intermodel differences are not uncommon, and we have noted then.

P7, L13–15: The models three times overestimate the equatorial lower stratospheric heating caused by volcanic aerosols. This is the main forcing of the stratosphere-troposphere dynamic interaction. There is something wrong here.

Yes, we are agreed. This is a well know bias in many of the current-generation models. However, the reviewer will agree that the models we have analyzed are no more biased than the ones in Stenchikov et al (2006), Driscoll et al (2012), Robock & Zambri (2016) and many other studies.

More importantly: this model biases makes our argument stronger! Even with an overestimated equatorial lower stratospheric heating caused by volcanic aerosols, our models (and those of the other recent studies) show no statistically significant surface warming. Had the models not overestimated the stratospheric aerosol heating the surface signal would be even smaller. We have noted this in the revised version of the paper (on page 7, lines 17–19).

P7, L20: "With this IN mind"

Thank you. We have corrected this.

P7, L32–35: I think the correct question to ask is whether models are able to correctly reproduce the probability distributions of the Arctic oscillation (AO) responses to volcanic forcing. But for this one has to extract multiple cases from observations and to construct the observed probability distribution. It is not doable with only one post-eruption season considered.

We politely disagree. We believe that one can indeed use "only one post-eruption season" provided on has large esembles of runs available. This is what we did.

P9, L10: Planetary wave reflection is the threshold process (Charney and Drazin, 1961) when small changes matter.

We do not question the Charney and Drazin (1961) result, which is based on small-amplitude linear theory in a highly idealized configuration. What we question is how relevant that particular threshold behavior is to the problem at hand. The claim that a mere 1-2 m/s acceleration of the polar vortex from volcanic aerosol heating has a major impact on planetary wave propagation is purely speculative and, to the best of our knowledge, has yet to be demonstrated. For instance, one would first need to show that the climatological conditions are found to be very near the wave propagation threshold, so that a tiny wind perturbation is able to make the system cross that threshold. Then one would have to show that the linear approximations are actually valid. And so on. We are not aware of studies which have carefully performed such work.

P9, L14: The wind variability coming from SSW is not relevant to the process. As soon as polar vortex zonal wind weakens below the threshold, planetary waves can propagate upward nonlinearly weakening the polar vortex. So, the amplitude of wind changes below the threshold, no matter how large it is, does not count. Sampling has to focus on the strong vortex cases for this purpose.

See the answer to our previous point, and also the answer to the other referee.

P10, L17–18: Exactly, the winter of 1991/92 is not suitable to study the forced stratosphere-troposphere dynamic interaction, as the positive phase of AO in the troposphere was caused by a different mechanism.

This point has been addressed above, in the "General Comments" section.

P11, L2: You mean surface cooling/warming, not in the lower stratosphere. Please clarify.

Yes, at the surface. Thanks for pointing out this ambiguity. We have correct the text.

P11, L5-8: You have to explain why do we see a positive AO anomaly climatologically after multiple volcanic eruptions. If this would be extremely rear events as in the models, then a positive AO anomaly would not be seen in observations.

First: we do not think we need to explain "why we see a positive AO anomaly climatologically after multiple volcanic eruptions". Our paper is about Mt Pinatubo and that anomaly was absent in the winter following the 1991 eruption, demonstrating ipso facto *that it could not possibly have been responsible for the observed NH surface warming.*

Second: the evidence for "a positive AO anomaly climatologically after multiple volcanic eruptions" is not terribly robust (see, e.g. Wunderlich & Mitchell, ACP 2017).

P11, L15–23: The strengthening of the polar vortex caused by the volcanic aerosols heating in the lower equatorial stratosphere is robust. This is the threshold process, so a weak strengthening matters. And it is unfair to apply wind variability in SSW to scale the increase in maximum wind.

We have addressed this comment in several locations in the discussion above.

P11, L25–35: The downward propagation mechanism was proved using climatological analysis (Baldwin and Dunkerton, 1999) and has to be challenged on this basis.

We are not sure what the referee means. We are not challenging the fact that SSWs can affect the annular modes and produce surface anomalies. That result is robust. We are questioning the claims of the early papers on the NH warming following volcanic eruptions. Those papers reported significant surface responses because the models used were flawed (they lacked vertical resolution and stratospheric variability). The fact that significant surface responses are not seen in the more recent studies (which are based on much better models) supports our interpretation.

P12, L25–30: ENSO definitely could affect surface temperatures, although Volcano-ENSO interaction is highly nonlinear. At least contribution of ENSO variability in the volcanic signal has to be removed properly, which was never done in this analysis. Another important mode of variability is QBO that was not considered and reported in this study. QBO plays an important role in stratospheric wave propagation and could directly affect polar vortex and shape the stratosphere-troposphere dynamic interaction.

We agree with the referee: ENSO and the QBO do affect stratospheric wave propagation. We also hope the referee agrees with us: that ENSO and the QBO are part of the internal variability of the climate system. Therefore, if their influence needs to be removed in order to detect any putative surface response to the volcanic aerosols (should it be detectable at all) it means that the volcanic surface influence in the NH is clearly masked by natural variability. This is the key point of our paper.