Reply to Nina Iren Kristiansen (Reviewer 2)

Thanks a lot for taking the time to read the paper and giving us valuable comments. We have changed the manuscript according to the suggestions and have listed our replies and changes in blue below.

General comments

This paper presents the CAMS assimilation of volcanic SO2 satellite data, and in particular improvements made to the system by the use of layer height information retrieved from satellite, which show to improve the SO2 forecasts. The paper is interesting and presents both improvements to and current challenges with the system. The topic of the paper is highly relevant as it addresses a method which can be used to fuse models and observations and targets a particular application to volcanic clouds. The paper is well written and highly suited for publication; however, I would like the below comments to first be addressed.

I miss some discussion around the applied/assumed thickness of the SO2 plume and if/how this might affect the results. See specific comment on L260.

I am concerned that the model simulations do not directly consider the vertical averaging kernel information from the SO2 retrieval. See specific comment on L262.

I miss some further details on the TROPOMI SO2 retrieval. Other TROPOMI SO2 total column retrievals are interlinked with assumptions on the SO2 plume altitude and often different products are available based on different a priori plume altitudes (e.g., the Copernicus SP5 products). L122 mention prior SO2 vertical profile shapes but this is not mentioned again or discussed any further for the DLR TROPOMI retrieval (only for IASI on L186). Please elaborate further on which prior profiles are used in the DLR TROPOMI retrievals (both NRT and LH) and if these vary, and also how/if this affects the retrieval of the layer height.

We have added this information in Section 2.1:

The DOAS vertical column SO2 retrieval requires knowledge of a prior SO2 profile to convert the slant columns into vertical columns. Because this profile shape is generally not known at the time of the observation and it is also not know whether the observed SO2 is of volcanic origin or from pollution (or both) the TROPOMI algorithm calculates four vertical columns for different hypothetical SO2 profiles. One vertical column is provided for anthropogenic SO2 with the prior SO2 profile taken from the TM5 CTM and three for volcanic scenarios assuming the SO2 is either located in the boundary layer, in the mid-troposphere (around 7 km) or in the stratosphere (around 15 km). These volcanic prior profiles are box profiles of 1 km thickness, located at the corresponding altitudes. The NRT CAMS system uses the mid-troposphere product. TROPOMI SO2 data are provided with averaging kernels based on the prior hypothetical SO2 profiles (i.e. the 1 km box profiles centred around the assumed SO2 altitude for the volcanic columns). However, as these do not provide any real information about the altitude of the volcanic plume they are not used in the CAMS system. More information about the NRT TROPOMI SO2 retrieval can be found in the TROPOMI ATBD. For the TROPOMI data (and also the other SO2 products used in this paper) observation errors as given by the data providers are used.

We have addressed the other comments in the list of specific comments below.

Can you provide some indications as to how much more expensive (in terms of run time) the model runs are with the higher spectral resolutions used? For example, it would be very interesting to know the difference in run time for each of the experiments in Table 3.

The experiments shown in Table 3 all use the same spectral resolution (model at T511, minimisations at T159/T255) so it does not make sense to add any run time information. Compared to the operational configuration which uses T95/T159 spectral resolutions in the minimisation the numerical cost for one analysis cycle is increased by about 20-30%, with the largest increase from the second

minimisation which is about 50% more expensive when going from T159 to T255. We have added this text:
‘ The numerical cost of one analysis cycle increases by about 20-30% when the spectral resolution of the minimisation is increased in this way, with the largest increase coming from the second minimisation which is about 50% numerically more expensive.’

Specific comments
L30 - the last sentence of the abstract: It would be good to include here something about the increase in skill time scales by including the LH information. I would also include a couple more key results here; that including LH information leads to higher modelled TCSO2 values in better agreement with the satellite observations, but that plume area and burden are overestimated also when including LH data and that the reason for this overestimation is explored.

We have added this to the abstract:
Including the layer height information leads to higher modelled TCSO2 values in better agreement with the satellite observations. However, the plume area and SO2 burden are generally overestimated in the CAMS analysis also when LH data are used. The main reason for this overestimation is the coarse horizontal resolution used in the minimisations. By assimilating the SO2 layer height data the CAMS system can predict the overall location of the Raikoke SO2 plume up to 5 days in advance for about 20 days after the initial eruption which is better than what is obtained with the operational CAMS configuration (without prior knowledge of the plume height) where the forecast skill drops much more for longer forecast lead-times.

L40:” SO2 in the aircraft cabin is the biggest issue leading to respiratory problems for passengers and crew”. Respiratory problems related to SO2 depend on the SO2 concentrations/dose and air quality standards for SO2 exist. Potential problems also depend on people’s underlying health problems like asthma. It doesn’t therefore always lead to respiratory problems as this sentence seem to indicate.

We have changed the sentence to: ‘*SO2 in the aircraft cabin is the biggest issue and can lead to respiratory problems….’*

L80: You use the different terms ‘injection height’ / ‘plume height’ / ‘layer height’ but not consistently and the difference between them (if any) is not explained. Personally, I’d use injection height as above the volcano and plume/layer height for the cloud altitude away from the vent, but it might be best to keep to as few terms as possible throughout the paper.

In L80 it makes sense to keep injection height, because this is what is determined by Flemming and Inness (2013). We have modified the rest of the paper to not use injection height and only use either layer height or plume height.

L135-147 (section 2.2): I miss some details on how the retrieval of the LH is done and what it relies on besides the exact wavelength ranges used. It which cases does it work well and which not (see related comment on L416). Why does it not work well below 20 DU? Also, what does this LH mean physically? You later use the height of the modelled maximum concentration as the model equivalent, would be good to comment on this here to justify that that is appropriate.

For SO2 columns below 20 DU the error of the retrieval of the layer height gets larger, making the data less accurate and less useful. The retrieval algorithm is documented in detail in other papers that we refer to. We have added a sentence in Section 2.2 and also a reference to a new validation paper by Koukouli et al. which has just been submitted to ACP:

For low SO2 columns, high-altitude layer heights cannot be retrieved and the retrieval is biased towards low layer heights (Hedelt et al., 2018). Therefore, the use of the data in the CAMS system is restricted to values > 20 DU. More details about the retrieval algorithm can be found in Hedelt et al. (2018) and Koukouli et al. (2021). Koukouli et al. (2021) compared the S5P LH data with IASI observations for the 2019 Raikoke, the 2020 Nishinoshima and the 2021 La Soufrière-St Vincent eruptive periods and found good agreement with a mean difference of ~0.5±3km, while for the 2020 Taal eruption, a larger difference of between 3 and 4±3km was found.

L207/section 3.2: The reader needs to know quite a bit about 4DVar assimilation systems to follow this section. It would be good to expand a little in particular on those aspects which you later explore in more detail: background error covariance matrix and the minimisations. Also, observations errors are not mentioned at all, how are errors in the observations taken into account?

The treatment of the background error formulation is already described in detail in section 3.2.1.We have added more information about 4D-var in Section 3.2:
In the CAMS 4D-Var a cost function that measures the differences between the model's background fields and the observations is minimized to obtain the best possible forecast through the length of the assimilation window by adjusting the initial conditions.

We have added a sentence in Section 2.1 to state that we use the observation errors provided by the data providers.

L260: "calculate the SO2 column not between the surface and the top of the atmosphere, but between the pressure values that correspond to the bottom and the top of the retrieved volcanic SO2 layer. The depth of this layer is currently set in the FP_ILM retrieval as 2 km, which corresponds to the uncertainty of the retrieved layer height." I am a little confused about this. Does it mean you use a fixed plume thickness of 2 km to calculate the modelled total columns, i.e., that you only calculate the SO2 column between the bottom of the plume (retrieved LH – 2 km) and the retrieved LH? What if there is a much thicker plume say several km thick, then the calculation of the SO2 column loading will miss a large fraction of the SO2 in the vertical by only summing only over the LH-2km depth.

You are right, the observations assume a depth of the layer of 2 km and we use this to calculate the model equivalent in the observation operator. If the SO2 layer was deeper this would not be accounted for in our method, but as that information is also not available from the observations we use it is not possible to include it based on assimilation of the SO2 LH product alone. Additional data (lidar?) which would give vertically resolved information would be needed. Some vertical variation in the SO2 loading will be achieved if parts of the plume have different altitudes as this information will be available from the observations. As you can see in Figure 3 there is quite a spread in retrieved LH for the eruption and that information will be brought into the SO2 analysis. Apart from that, we will depend on vertical transport to modify the vertical SO2 distribution.

We have added this sentence to the paper to document the limitation: '*One limitation of this method is that the SO2 LH product gives the plume altitude with an accuracy of 2 km, but does not give a value for the lower vertical boundary of the SO2 plume, and for a thick plume part of the SO2 loading could be missed in the calculation of the model equivalent. However, as the model's background SO2 concentrations in the free troposphere are low this should not be a big issue in the column calculation. Also, some vertical variation of the SO2 loading will be achieved if parts of the plume have different altitudes, and Figure 3 shows that this is indeed the case for the Raikoke eruption.'*

L262: "This approach mimics the procedure of using averaging kernels with box profiles given for the SO2 layer.". I don't understand how this mimic the use of averaging kernels because if applying an averaging kernel sensitivity, the model data would be multiplied with a different sensitivity/ averaging kernel (AK) value at different vertical levels. Please elaborate. Ideally the satellites vertical AK profiles should be applied to the model data prior to any comparisons to the satellite data – this AK profile can vary from one satellite pixel to the next.

It mimics the use of the averaging kernels that are supplied with the volcanic SO2 data, which are supplied with the TROPOMI data and are box profiles (see our reply to reviewer 1 above) that represent typical volcanic SO2 profiles and are used in the AMF calculation to calculate the vertical columns.

We have added information about the AK in section 2.1 and changed the sentence to:
This approach mimics the procedure of using TROPOMI SO2 averaging kernels which are box profiles, but for the retrieved layer and not an assumed hypothetical volcanic SO2 profile (see TROPOMI SO2 ATBD, http://www.tropomi.eu/documents/).

L278: "The 'dip' in the TROPOMI SO2 burden after the initial peak is an artefact that results from missing observations in the TROPOMI NRT data." This 'dip' is not seen in the equivalent time series shown in the de Leeuw paper (their Fig 11) which also show TROPOMI data (different retrieval method). What is the cause of these 'missing observations?

In the NRT data we are using there is a data gap in the area of highest SO2 values (also visible in Figure 9c2) on 25 June. We do not know why. Possible cloud/ ash contamination or data being flagged because of 'unrealistically' high SO2 columns? De Leeuw et al. (2021) do not mention that they used NRT TROPOMI SO2 data, so we assume they used an offline product for which that problem might have been corrected. We have added this information in the paper:
The 'dip' in the TROPOMI SO2 burden after the initial peak is an artefact that results from missing observations in the TROPOMI NRT data on 25 June 2019 in the area of highest SO2 values (also visible in Figure 9c2 below).

L300 / Table 3: From the order of the experiments given in the table I expected first the difference between the BLexp and LHexp to be discussed, however the LH50/100/250 cases are first discussed. Perhaps guide the reader at the start of the section to say which experiments are compared first and why. Similarly, would be good there to guide the reader to say that the BLexp and LHexp will be further explored later to assess the skill timescales to see if using a more realistic height rather than the default 5 km will improve the forecasts – a key point and question for the paper.

We think this is addressed in the paper when introducing Table 3 because we already have a paragraph describing the experiments*: '…listed in Table 3. The baseline experiment (BLexp) which assimilated NRT TROPOMI TCSO2 data with the operational CAMS configuration and the layer height experiment (LHexp) which uses the FP_ILM S5P LH data with a horizontal background error correlation length of 100 km and background error standard deviation values of 0.7e-7 kg/kg are the main experiments used in this paper (Section 4.3 below) to assess if the assimilation of the SO2 LH data using a more realistic height rather than the default 5 km improves the CAMS SO2 analyses and forecasts. The other LH experiments assess the impact of using different horizontal SO2 background error correlation length scales and various SO2 background error standard deviation values.'*
We have added the green part to make it even clearer.

We can change the order of the entries in the table if the editor deems this necessary.

L415: "TROPOMI NRT lower detection limit": is this a true detection limit from the sensor/retrieval or do you mean rather than you applied a lower DU threshold (5 DU) for the NRT TROPOMI data compared to the SP ILM SO2LH retrieval data (20 DU)? Not clear to me if this is a direct 'detection limit' or more a 'chosen threshold' based on various limitations (not necessarily a detection limit). For the 5 DU threshold you mention this is applied to avoid assimilating SO2 from outgassing volcanoes which are covered by SO2 emissions in the CAMS model. Also see related question below.

This is a real detection limit. The plots of the NRT TROPOMI data show all available volcanic NRT SO2 data even though only values > 5DU are assimilated.

L416: “FP_ILM SO2LH retrieval (v3.1) does not provide reliable information for TCSO2 < 20 DU and therefore only picks up those parts of the plume that are associated with the highest SO2 load” The work ‘information’ is ambiguous. Does it mean that both the retrieved column load values and the layer height values are not reliable under 20 DU, or is it only the retrieved layer height data which is not reliable under 20 DU? Maybe to add in section 2.2.

It means the layer height retrieval has too large an error to be useful. We have added more information in Section 2.2. See reply to L135-147 above.

L425/ Figure 9: Would be useful if the figure caption could explain why the NRT TROPOMI data differ to the SO2LH TROPOMI data (i.e., DU levels used/displayed).
We already mention this in the text, but have now added in the caption of Fig 9:
‘In panels (c)-(e) all available observations are shown, illustrating that the SO2 LH product only picks up those parts of the plume that are associated with the highest SO2 load.’

L430: It is not directly explained why the SO2 burden is so much larger (2-3 Tg) for the LHexp compared with BLexp. Is it because of higher TCSO2 values as well as overestimating the plume area? 2-3 Tg is quite a lot higher than the total burden values from the satellite data.

The overestimation of the plume area in LHexp is actually less than in BLexp for >5 DU. The larger overestimation of the burden in LHexp is likely the result of differences in the background error standard deviation values and the fact that lower SO2 columns that could correct an overestimation in parts of the plume are not assimilated. We have added this sentence to the paper:
The larger overestimation of the SO2 burden in LHexp is the result of differences in the background error standard deviation values used in the experiments and of the fact that lower SO2 columns, which could correct an overestimation in parts of the plume, are not assimilated.

L575-L590: It would be good to compare these skill time scales to what was found by de Leeuw et al for the NAME model (skill for 12-17 days for the low-density (<1 DU) parts of the SO2 cloud and 2-4 days for the denser parts (>20 DU) of the SO2 cloud).

We have added a sentence to this section:
Leeuw et al. (2021), using the Met Office's Numerical Atmospheric-dispersion Modelling Environment (NAME) dispersion model, found skill timescales of 12–17 days for low density (> 1 DU) parts of the Raikoke SO2 cloud and shorter skill timescales of 2–4 days for the denser parts of the cloud (>20 DU). It is interesting to see skill timescales of similar magnitude to the ones obtained in our study even though the method is different. Leeuw et al. (2021) initialized the NAME dispersion model with eruption source parameters and then followed the evolution of the SO2 cloud, while we use data assimilation to update the location of the plume daily and provide daily SO2 forecasts with a maximum length of 5 days.

Technical comments
Figure text and labels need to be increased as on a print-out version some figures (especially figures 4, 12, 13, 16,17) are very hard or near to impossible to read.

We have improved the figures.

Figure 3: Suggest changing the colour scale as there are very few values >100 DU so hard to distinguish the dots. Is this showing values only >20 DU as you mention the retrieval is accurate only for larger DU values.

We have changed the colour scale to only show values up to 250 DU.

The de Leeuw reference should be updated to the final revised version for 2021.

Done.

The reference Prata et al. 2019 is used in the main text but is not in the reference list.

Added.

Figure 18 could be removed as there are many figures and the difference to Fig 17 is not very big so describing by words in text should be sufficient.

We have removed Figure 18 but kept the text referring to the GOME-2 result.