Answers to the reviewer comments for the paper: "Assimilation of volcanic sulfur dioxide products from IASI and TROPOMI into the chemical transport model MOCAGE: case study of the 2021 La Soufrière Saint-Vincent eruption"

In the remainder of this document, we provide our full response (in blue) to the reviewers' comments (in black), which details what we did to address their concerns. We include in this document (in italics) some original passages of the article (in green), with modifications and improvements (in red). The following nomenclature has been adopted for the figures and tables: Fig.X are the figures in the initial version of the article, FigRev.x/TabRev.x are the figures/tables in the new version of the article and FigRep.x the plots used exclusively in this document.

1 General comments

This new study focuses on the assimilation of volcanic sulphur dioxide (SO_2) data from the TROPOMI and IASI satellite instruments into the MOCAGE chemical transport model, using the 2021 La Soufrière Saint-Vincent eruption as a case study. The research highlights the importance of integrating data from different satellite sensors, exploiting their complementary capabilities, to improve real-time atmospheric monitoring and forecasting of volcanic SO_2 plumes. The assimilation of combined observations significantly improves the accuracy of SO_2 plume predictions compared to using individual data sets, and also captures secondary transformations such as the conversion of SO_2 to sulphate aerosols. The study highlights the potential benefits of similar multi-sensor approaches for volcanic hazard monitoring and operational aviation safety.

The study is scientifically sound. The draft paper is well written and mostly clear and concise. I would like to recommend the paper for publication in AMT, subject to the following specific comments and technical corrections given below.

Thank you for your positive evaluation of our study and for your constructive feedback. We are pleased that you find our work scientifically sound, well-written, and clear. The manuscript has been improved as a result and we are grateful for your contributions.

2 Specific comments

l204: "The background error covariance is spread on many vertical levels and on many meshgrids thanks to the correlation matrix." Please revisit the sentence and try to be a bit more specific, e.g. what are the actual correlation lengths imposed in the correlation matrix?

We added a paragraph in the article which describes the correlation matrix and gives information about values we used in our study.

$\mathcal{H}$ is the observation operator used to obtain the model data in the observation space. Before running an assimilation experiment, a full description of R and B matrices is required. The background error covariance is spread in space thanks to the correlation matrix described in [El +20]. This matrix contains both horizontal and vertical components.

The horizontal correlation $C^h_{m,n}$ between two points m and n is defined as follows:

$$C^h_{m,n} = exp[\frac{-d}{2(L_x + L_y)}] \quad (1)$$

where d is the distance between the points m and n, L_x and L_y are the longitude and latitude length scales in kilometers. In our study, the longitude and latitude length scales are equal to one meshgrid (1°).

In kilometers, length scales become:

$$L_x = L_y = 2R_e.sin(\frac{\pi}{360}) \quad (2)$$

R_e is the Earth's radius (6371.22 km).

The vertical correlation $C^v_{i,j}$ between two pressure levels (p_i and p_j) is defined as follows:

$$C^v_{i,j} = exp[-100.[log(\frac{p_i}{p_j})]^2] \quad (3)$$

In our study, the values of the vertical correlation between two consecutive levels are set to 1.

l359: I was wondering why the authors used the Probability of Detection (POD) but not the False Alarm Rate (FAR) and the Critical Success Index (CSI) to evaluate the model results? The POD is a useful metric, but it has limitations, for example if the model is too dispersive and largely overestimates the SO_2 plume concentrations, it will produce many hits and a high POD, but overall may not have good predictive quality. Could you please comment on this? Or even better, try to add FAR and CSI estimates?

Yes, we agree POD has limitations, in particular to see if the model is too dispersive. We added in the article the CSI and FAR computed against OMI observations for analyses and forecasts. Nevertheless, we kept a figure with the POD computed when one instrument is assimilated (Fig.6/FigRev.7). It is important to notice that we found a bug in our script used to compute POD. We fixed it and changed the impacted figures in the article in Fig.6/FigRev.7 on the first line of the FigRev.8. and on the first line of the FigRev.11. CSI and FAR metrics are plotted on FigRev.8 and FigRev.11.

In the "Impact of the assimilation on the detection of SO_2 threshold exceedances" section:

To assess the accuracy of the model in simulating SO_2 total columns, a threshold-based analysis was implemented. The goal was to determine the number of instances where both the observations and the model successfully identified SO_2 total columns above certain thresholds (labelled as Hits), the instances where the observations exceeded these thresholds but the model failed to detect them (labelled as Misses), the instances where the model exceeds these thresholds but the observations do not reach these thresholds (labelled FalseAlarms) as well as the instances where both the observations and the model successfully identified SO_2 total columns under certain thresholds (labelled as CorrectRejections). Using these numbers, we defined three metrics.

The first one is the Probability of Detection (POD), a ratio that ranges from 0 to 1. The POD is calculated by dividing the number of Hits by the sum of Hits and Misses for a given threshold. A POD score of 1 indicates a perfect detection by the model, meaning that all observed instances above the threshold were correctly simulated. On the other hand, a POD of 0 signifies that none of the observed SO_2 total columns above the threshold were detected by the model. The POD is computed with the following equation:

$$POD = \frac{Hits}{Hits + Misses} \quad (4)$$

The second one is the Critical Success Index (CSI), a ratio that ranges from 0 to 1. The CSI is calculated by dividing the number of Hits by the sum of Hits, Misses and False Alarms for a given threshold. A CSI score of 1 indicates a perfect detection by the model, meaning that all observed instances above the threshold were correctly simulated. On the other hand, a POD of 0 signifies that

none of the observed SO_2 total columns above the threshold were detected by the model. The CSI is computed with the following equation:

$$CSI = \frac{Hits}{Hits + Misses + FalseAlarms} \tag{5}$$

The last one is the False Alarm Rate (FAR), a ratio that ranges from 0 to 1. The FAR is calculated by dividing the number of False Alarms by the sum of False Alarms and Correct Rejections for a given threshold. A FAR score of 0 indicates that there is only Correct Rejection instances. On the contrary, a FAR score of 1 indicates that there is only False Alarm instances. The FAR is calculated with the following equation:

$$FAR = \frac{CorrectRejections}{CorrectRejections + FalseAlarms} \tag{6}$$

To study these metrics, the notations in the table TabRev.2 are adopted. For POD metrics, times when there are no hits nor misses events are shown by a dot and times when there is no hits bur misses event are represented by a cross. For CSI metrics, times when there is no hits, no misses and no false alarm are shown by a dot, times when there are no hits and no false alarm but misses events occur are represented by a cross. If there is no hits event but if there are misses and false alarms events, a start is plotted. For FAR metrics, a dot is plotted when there are no false alarm and no correct rejections events. A cross is plotted when there are no false alarm event but correct rejections events.

	POD		CSI			FAR	
	Hits	Misses	Hits	Misses	FalseAlarms	FalseAlarms	CorrectRejections
●	0	0	0	0	0	0	0
×	0	> 0	0	> 0	0	0	> 0
★			0	> 0	> 0		

TabRev.2: Symbols used in plots according the studied metric and the number of hits, misses, false alarms and correct rejection events.

Fig.6/FigRev.7 shows the Probability of Detection computed for 1 DU and 5 DU thresholds against TROPOMI and IASI observations. Dots represent times when there is no observation. Crosses represent the moments when simulated SO_2 total columns are under a threshold whereas some observations exceeds this threshold. Against TROPOMI instrument, POD values are generally better in the experiments in which TROPOMI observations have been assimilated. In these experiments, POD values exceed 0.75. The POD values are over 0.75 until 12^{th} April in iasi_assim experiment except on 9^{th} April. POD values decrease at the end of the study period. For the 5 DU threshold, POD values are slightly higher in joint_assim experiment, especially on 10^{th} and 11^{th} April, when around 100 TROPOMI observations exceed 5 DU. On 12^{th} April, POD values are around 0.4 in the experiments in which TROPOMI instrument is assimilated. No SO_2 total column higher than 5 DU is simulated for this date in iasi_assim experiment. Between the 13^{th} and the 15^{th} April and on 9^{th} April, no SO_2 total column above 5 DU is simulated in MOCAGE. For these days, between 1 and 9 observations above 5 DU are measured by TROPOMI.

POD values, computed for a 1 DU threshold and with IASI observations, exceed 0.75 in experiments in which IASI instruments are assimilated. No SO_2 total column is simulated with the TROPOMI assimilation until 10^{th} April because TROPOMI overpasses the plume after IASI. In the morning of 9^{th} April, no observation above 1 DU is detected by IASI. From 11^{th} April, POD values vary between 0.3 and 0.8 in the tropomi_assim experiment. In this experiment, POD values are often higher in the morning. In the afternoon of 9^{th} April, only one observation above 5 DU is measured by IASI. At this location, the total column is under 5 DU. For this threshold and compared to tropomi_assim experiment, the probability to simulate high SO_2 total columns increases thanks to the IASI assimilation. Despite numerous observations above 5 DU, many events are missed on 10^{th} April with a POD reaching nearly 0.4 in the morning and 0.5 in the afternoon. Most of simulated SO_2 total columns are between 1 DU and 5 DU. The maximum of POD is obtained on 11^{th} April after the assimilation of many observations measured by IASI exceeding 5 DU on 10^{th} and on 11^{th} April.

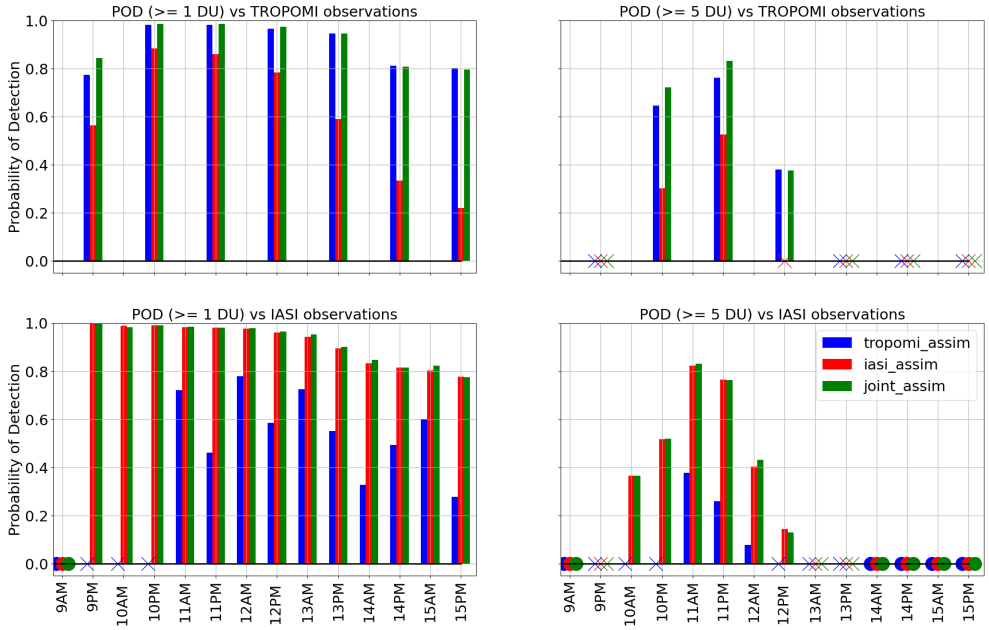
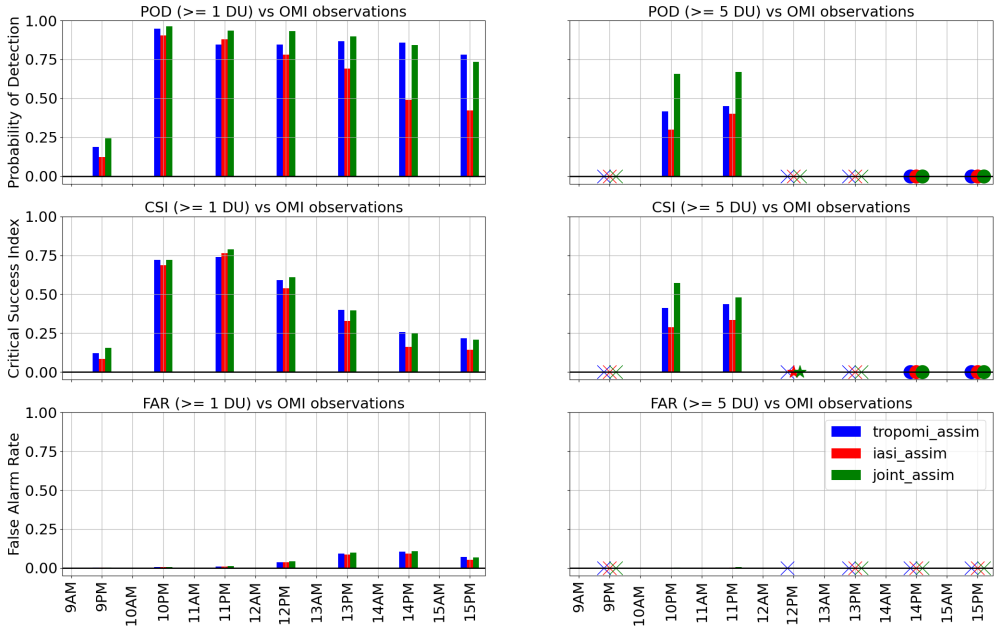


Fig.6/FigRev.7: Probability of detection for 1 and 5 DU thresholds for the three experiments: tropomi_assim in blue, iasi_assim in red and joint_assim in green. Dots represent times when there is no observation. Crosses represent the moments when there are only misses events.



FigRev.8: Probability of detection (first line), Critical Success Index (second line) and False Alarm Rate (last line) for 1 and 5 DU thresholds for the three experiments: tropomi_assim in blue, iasi_assim in red and joint_assim in green. The meaning of the symbols is described in the Table **??**.

The first line of the FigRev.8 shows POD computed against OMI observations for 1 DU and 5 DU thresholds. On this line, dots represent times when there are no hits and no misses events. Crosses represent the moments when simulated SO_2 total columns are under a threshold whereas some observations exceed this threshold. The POD values computed with a 1 DU threshold are often consistent between the tropomi_assim and the joint_assim experiments. POD values are slightly better in joint_assim experiment between 9^{th} to 13^{th} April. For the 5 DU threshold, POD value is greater with the joint_assim experiment on 10^{th} and on 11^{th} April. On 12^{th} and 13^{th} April, no SO_2 total column above 5 DU is modelled by MOCAGE whereas OMI measured observations above this threshold. Elsewhere in the study period, no observation greater than 5 DU is measured by OMI instrument and modelled by MOCAGE.

The second line of the FigRev.8 shows CSI computed against OMI observations for 1 DU and 5 DU thresholds. As for POD, the CSI values computed with a 1 DU threshold are often consistent between the tropomi_assim and the joint_assim experiments. CSI values are slightly better in joint_assim experiment on 9^{th}, on 11^{th} and on 12^{th} April. On 10^{th} and on 11^{th}, CSI values are around 0.75 whereas POD values are around 0.9. It means that there are few false alarms events during this both days. On the contrary, from the 13^{th} April, CSI values are much lower than POD values. It means that the plume in MOCAGE becomes too large, leading to a lot of false alarms events. For the 5 DU threshold, CSI values are better in the joint_assim experiment on 10^{th} April, and to a lesser extent on 11^{th} April. From 14^{th} April, no observations higher than 5 DU are measured by OMI instruments and modelled by MOCAGE. On 9^{th}, on 12^{th} with tropomi_assim experiment and on 13^{th} April, no observations greater than 5 DU and observed by OMI instrument but some values above 5 DU are simulated by MOCAGE. On 12^{th} April, there are misses and false alarms events in iasi_assim and joint_assim experiments.

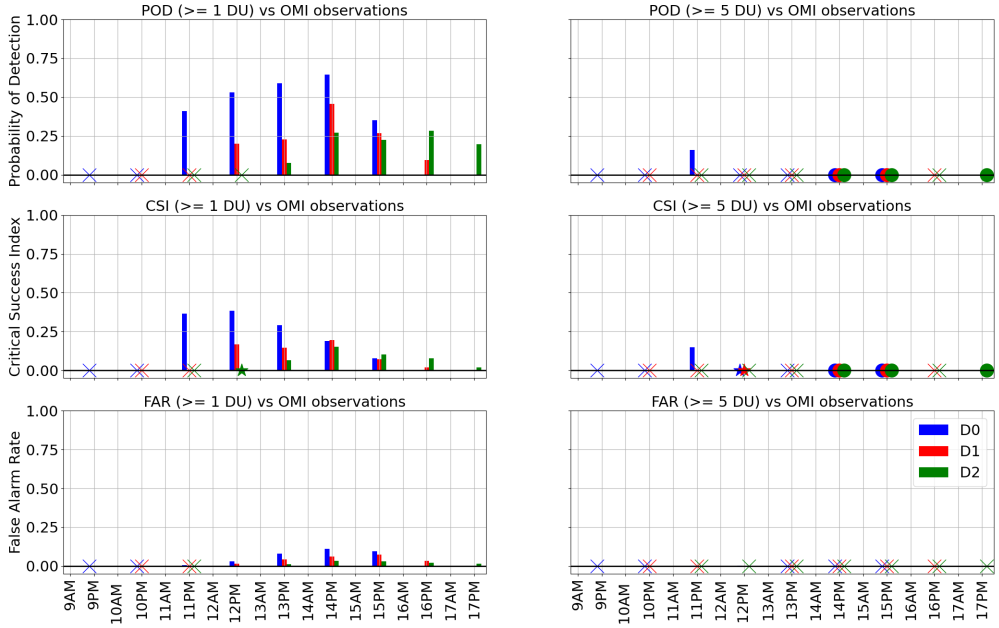
The third line of the Fig. FigRev.8 shows FAR computed against OMI observations for 1 DU and 5 DU thresholds. The FAR values computed with a 1 DU threshold are similar between experiments. Up to 11^{th} April, FAR values are approximatively equal to 0. This shows that the number of correct rejections events is larger than the number of false alarms events. The FAR values increases from 12^{th} April meaning that the plume is too large in the model. The FAR values computed with the 5 DU threshold are always close to 0. On 9^{th}, on 12^{th} in tropomi_assim experiment and from 13^{th} April, there are only correct rejections events.

In the "Impact of assimilation on forecasts" section:

Generally speaking, the number of meshes exceeding 1 DU or 5 DU decreases with the forecast period. This was not observed on 9^{th} and 10^{th} April, when no mesh exceeded 1 DU. The D0, D1 and D2 forecasts show the presence of a plume from 11^{th}, 12^{th} and 13^{th} April 2021. These forecasts are initialised by the output of the assimilation of 9^{th} April, i.e. before the beginning of the eruption. On 11^{th} April, there were around 200 grid cells where the model exceeded 1 DU for the D0 forecast. This number corresponds to the minimum number of grid cells where the TROPOMI and OMI observations reach 1 DU and is below the number of grid cells where the median of the IASI observations reaches 1 DU. On 12^{th} April, the plume forecast with D0 was larger, with around 500 meshes exceeding 1 DU, corresponding to the number of meshes where the 25^{th} quantile of the TROPOMI observations reached 1 DU and also where the median of the OMI observations reached this threshold. After this date, the number of occurrences of the total column in SO_2 exceeding 1 DU increases and becomes greater than the number of grid cells where the IASI and OMI observations reach 1 DU. However, this number remains smaller than the maximum number of meshes calculated using TROPOMI observations. The number of points where the total column reaches 1 DU decreases with the forecast term range. Nevertheless, this number always exceeds the number of grid cells where the OMI and IASI observations reach 1 DU from 14 April onwards. Regarding the 5 DU threshold, no grid cell exceeds this threshold for the D2 forecast. The D1 forecast shows a low number on 12^{th} April when the total columns reach 5 DU. However, this number is similar to the number of grid cells where the median of TROPOMI and IASI observations reaches 5 DU. In addition, for this day, the number of points where the model reaches 5 DU is similar between the D0 and D1 forecasts. For 11^{th} April, the number of occurrences of a total column greater than or equal to 5 DU is low in the model compared with the UV instruments. This number is within the range of grid cells where IASI observations exceed 5 DU.

FigRev.11 represents the POD (on the first line), the CSI (on the second line) and the FAR (on

the third line) metrics calculated by comparing the observations measured by OMI and the forecasts at several time steps for 1 DU and 5 DU thresholds. The first forecast is launched on 9^{th} April. Consequently, there is no D1 forecast available for this day and no D2 forecast available on on 9^{th} and on 10^{th} April. The last forecast is launched on 15^{th} April so there is no D0 forecast available on 16^{th} and on 17^{th} April. There is no D1 forecast available on 17^{th} April. For POD and CSI metrics computed for the 1 DU threshold, best values are found in the D0 forecast except on 14^{th} and on 15^{th} April for CSi metric. On 14^{th} April CSI are similar between D0 and D1 forecast and on 15^{th} April, the CSI is slightly better in D2 forecast. Compared to POD values, CSI values are much lower especially with D0 forecasts on 13^{th} and on 14^{th} April. It means that there are a lot of location where MOCAGE wrongly simulated total columns stronger than 1 DU. This finding is strengthened by the presence of highest FAR values during these days. POD and CSI metrics computed with a 5 DU threshold show similar values. These metrics are 0 except on 11^{th} April with values around 0.2 for both POD and CSI metrics. On 12^{th} April both misses and false alarms events occur with the D0 and D1 forecasts. Elsewhere, no SO_2 total columns stronger than 5 DU are modelled and no observations stronger than this threshold are observed by the OMI instrument on 14^{th}, on 15^{th} and on 17^{th} April. With the 5 Du threshold, there are only correct rejections events.



FigRev.11: POD on the first line, CSI on the second line and FAR on the last line for 1 DU and 5 DU thresholds for the D0 forecast in blue, the D1 forecast in red and the D2 forecast in green. The meaning of the symbols is described in the Tabrev.2.

l486: It would be good if the text was a bit more specific about how the background error covariances are defined.

To choose the background error covariances, we investigated the values of the background error covariances used in IFS to assimilate SO_2. Before 2022, background error standard deviation was defined as a profile containing a peak at 550 hPa, with a valaue of 5e-7 ppv. We decided to lower this specific value for the background error covariances in order to reduce the weight given to the observations.

l491: I'd like to suggest adding a few references regarding the limitations of the chemical modelling and the uncertainties of the meteorological data for the volcanic SO_2 chemistry-transport simulations, as these issues have been addressed in several previous studies.

We agree and now cite [WT22] for the uncertainties of the meteorological data and [**schumann2011airborne**] for the limitations of the chemical modelling of volcanic SO_2.

l495: Using the information on where the satellites did not actually detect SO_2 sounds very helpful, especially to reduce false alarms. In this context it would be good to know if the FAR of the MOCAGE simulations is significant.

We compute FAR for analysis outputs. In our case study, FAR does not reach 0.1 using the TROPOMI instrument. Nevertheless, we observed a slight rise from 12^{th} April. However, we computed FAR between 90°W and 40°E and between 20°S and 30°N. Nevertheless, a weak value of FAR means that the number of correct rejections events is very large compared to the number of false alarm events. With TROPOMI, there is a lot of observations. A FAR of 0.1 means that there are around 100 000 false alarms events. So, our model is definitively too dispersive. The use of observations where no SO_2 is detected would be useful to correct the size of the plume but it is a real challenge to take that into account during the assimilation, not to mention the data volume that would generate. The 3D-Var technique may also be a limitation in that case.

3 Technical corrections

l195: please combine multiple citations in a single set of parentheses
l201: "searched as a sum" → "found as a sum"
l215: "15km of high" → "15 km of altitude
l226: please use "AVK" or "Avk" consistently
l310 (and other places): please use abbreviations, "figure 4" → "Fig. 4" (see AMT manuscript composition guidelines)
l372: please correct "The model did not ??? SO_2 total columns..."
l461: "TROPOMI and IASI instruments" → "TROPOMI and IASI data"
l463: "more important amount of SO_2" → do you mean "more realistic"?
l475: "The more the forecast range term is small, the more the plume size is important". → Revise/improve sentence?

All technical corrections are taken into account. Moreover, we combined all multiple citations in a single set of parentheses. l463: Yes we do mean realistic. l475: As the forecast period increases, the size of the plume decreases.

References

[El +20] Laaziz El Amraoui et al. "Aerosol data assimilation in the MOCAGE chemical transport model during the TRAQA/ChArMEx campaign: lidar observations". In: *Atmospheric Measurement Techniques* 13.9 (2020), pp. 4645–4667.

[WT22] Helen N Webster and David J Thomson. "Using ensemble meteorological data sets to treat meteorological uncertainties in a Bayesian volcanic ash inverse modeling system: A case study, Grímsvötn 2011". In: *Journal of Geophysical Research: Atmospheres* 127.24 (2022), e2022JD036469.