Manuscript Title: SO2, silicate clouds, but no CH4 detected in a warm Neptune

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

This study presents valuable results regarding the discovery of SO_2 in the atmosphere of the planet WASP-107b with unusual properties. The authors have included the necessary details along with a number of interesting diagrams to illustrate their points, which have made the manuscript understandable. However, to improve their work, I would recommend revising the manuscript based on the comments/suggestions below.

Major Comments/suggestions

1. Lines 68-69: I do not believe the transmission spectra from the three data reduction methods are in "excellent agreement", based on Figure 1. There are noticeable discrepancies between these spectra, in particular, at longer wavelengths. I would suggest quantifying how well these spectra agree, instead of using a word like "excellent", and briefly explaining (in the main text) why the transmission spectrum from "CASCADe" is preferred.

2. Lines 102-107: This is one of the most important parts of this study. While it has been claimed that for $T_{eq} < 1000$ K, SO_2 production below the 0.01 mbar level is halted (Tsai et al. 2023), SO_2 has been discovered in a planet with $T_{eq} = 740$ K in a more recent study, which is interesting. The authors have brought some reasons (starting from line 108) to justify their finding. However, it would be better to add a separate paragraph before these justifications and argue against the claim of Tsai et al. 2023. For example, Extended Data Figure. 10 (b) in their paper shows the average VMR between 10 and 0.01 mbar as a function of T_{eq} for some sulfur-bearing species. To produce these models, except for temperature profiles, all other planetary parameters were assumed to be consistent with WASP-39b, which are not applicable to other planets with different properties like WASP-107b. In a comprehensive study, different values of parameters, such as irradiation, intrinsic temperature, density, chemical abundances, … should be examined to reach an accurate conclusion.

3. Lines 118-121: Regarding the previous comment, the authors stated that the large atmospheric scale height of WASP-107b (which is its especial property) enables highly efficient photochemical processes to operate within $T_{eq} \sim 740$ K for this low-density planet. However, they have not examined how the molar fraction of SO_2 can change with scale height (as Tsai et al. 2023 has not either). Such an investigation may also elucidate the detection of SO_2 in WASP-107b.

4. Lines 545-549: It is stated that the observed increasing trend can be attributed to the different systematic models employed by the TEATRO (an exponential model) and Eureka! (both an exponential and polynomial models) reduction codes. However, it should be explained that how these two codes are compared with the CASCADe data reduction and how this reduction does not play any role in the increasing trend of transit depth difference.

5. Extended Data Figure 1: It should be explained why the values of 1 σ error associated with CASCADe code are discrepant relative to the values from the two other reduction methods and the simulated transit spectrum at longer wavelengths.

6. While there is a direct observation of the host star for NUV, I am not clear why the observation of another star (similar to the host star) has been used.

Minor Comments/suggestions

1. In Figures 1 and 2 (the bottom part, showing cloud and molecular contributions) as well as Extended Data Figures 1 and 2 (the bottom left panel), colors with higher contrast should be used to better distinguish them.

2. Line 20: I would suggest mentioning the wavelengths 8.69 μm and 7.35 μm after "two fundamental vibration bands of SO_2" to give the readers a general understanding about the features of interest at the beginning of the paper.

3. Lines 55-56: There is no need to include technical terms like "read-out pattern FASTR1" in the main text; but these should be introduced or defined in the Method section only. A simpler sentence may be sufficient to convey the point: "The subarray SLITLESSPRISM was used for a spectral resolution varying between 30 and 100."

4. Figure 2: In addition to molecular contribution shown in the bottom part of the figure, I would recommend marking the prominent spectral features due to H_2O, H_2S, and CO, similar to those from SO_2 and NH_3. In addition, marking or tagging the HST data points in some way would help the readers follow the figure more easily.

5. Line 418: There is no definition for the acronym "MAST".

6. Lines 468-469: The acronym FWHM must be defined where it appears first, i.e., line 461.

7. Lines 470 and 508: "band averaged light curve" or "white light curve"? It would be more appropriate to have a consistency in terminology.

8. Lines 469-471: "The orbital parameters of WASP-107b …from the TEATRO data reduction." These orbital parameters must also be mentioned in the section of TEATRO data reduction setup.

9. Lines 55, 413, 434, 486: "subarray" or "sub-array". It would be better to use the same term in the entire manuscript.

10. Lines 544-545: "mostly within 2σ", but in Extended Data Figure 7, it is mentioned "mostly within 1.5σ"… These two must be the same.

11. Lines 589-613: It is very difficult to read all the retrieval parameters and their corresponding priors in this part. I would suggest putting these parameters and priors into a table, which could make them easy to follow.

12. Suppl. Inf. Figures 5 and 6: The title axes of the corner plots are too small to be useful. One possible solution would be to plot the 1D histograms of relevant parameters (e.g., SO2 abundance) separately, in a legible figure.

13. I have found some typos and grammatical mistakes. So, the manuscript should be carefully checked and corrected if needed.

Referee #2 (Remarks to the Author):

The manuscript presents novel and interesting results of significant interest to the exoplanet community. In particular, it is the first detection of silicate clouds in an exoplanet (with the possible exception of VHS 1256 b), which have long been theoretically predicted. The detection of SO2 and constraining non-detection of CH4 are not unprecedented in the field, but they add to the scientific importance of the paper by demonstrating clear evidence of disequilibrium chemistry. I recommend that the paper be published once my concerns are addressed.

Silicate clouds:

Since the most significant result in the paper is the discovery of silicate clouds, I recommend doing further tests to establish the robustness of this result. In particular, in the ARCiS retrieval, SiO2 and MgSiO3 abundances are consistent with 0, but SiO and C abundances are not; in the petitRADTRANS retrieval, KCl and MgSiO3 abundances are consistent with 0, but SiO2 abundances are not. Which of these species is driving the detection of silicate clouds? Given the visually obvious dropoff beyond 10.5 um, it's probably one of the silicates, but which one? Are silicate clouds still preferred if a retrieval is done with only MgSiO3, only SiO, or only SiO2, without any C or KCl? In the Rayleigh regime (which <0.1 um particles would definitely be in), extinction is dominated by absorption instead of scattering, and the absorption cross-section is 8*pi^2*r^3/lambda * Im((m^2 - 1) / (m^2 + 1)) for spherical particles with complex refractive index m (Equation 7.4 of Mishchenko 2002: https://pubs.giss.nasa.gov/books/2002_Mishchenko_mi06300n/2002_Mishchenko_etal_1.pdf) It may be useful to plot, for different species, the natural log of this value vs. wavelength and compare it to the transit spectrum expressed in scale heights. They ought to have roughly the same shape at wavelengths dominated by silicate absorption (but not necessarily the same scale, if the scale height of the silicate clouds is different from the scale height from gas).

Along the same lines, more should be said about how "we mix the refractive indices of these materials using effective medium theory" (line 569) and how "we included amorphous MgSiO3, SiO2, and crystalline KCl clouds, considering them to be irregularly shaped (DHS method)". Since the inference of silicate clouds come entirely from the wavelength dependence of the refractive indices, it would be useful to have more details about how these were computed for the mixture and whether the computation is reliable.

Finally, the authors say in the main text that "the retrieval models favour a very high-altitude cloud layer composed of tiny (<0.1 um) amorphous silicate particles", but I don't see any parameter in the corner plot for either retrieval that corresponds to the particle size, and I also don't see any equation that gives the typical particle size as a function of the retrieval parameters. Since the particle size is itself an interesting physical quantity that informs atmospheric chemistry and dynamics, not to mention the cloud's absorption properties, could the authors plot the posterior of the particle size?

Data analysis:

I recommend discarding at least the first two data points in the spectrum, conservatively the first three. According to the JWST documentation (https://jwst-docs.stsci.edu/jwst-mid-infrared-instrument/miri-observing-modes/miri-time-series-observations/miri-lrs-tsos), “the dispersion profile of the LRS turns over below 4.5 μm—for a limited wavelength range around 4.5–5 μm, different wavelengths are dispersed onto the same detector pixels.” The 0.15 um bin also corresponds to very few pixels below 5 um (only 1 pixel, for 4.61 um), further decreasing the reliability of these bins. I expect this to further decrease the significance of the (very marginal) CO detection, which seems driven by the first 2-3 points.

The default non-linearity correction clearly has problems, but I have concerns about the authors’ own non-linearity correction. Suppl. Inf. Figure 1 plots pair-wise differences for different rows in column 36. What do columns 34 and 37 look like? In all the MIRI datasets I’ve seen, the differences go up with pair number in columns 34 and 37, vs. down in column 36. This phenomenon cannot be accounted for by the authors’ equation (between line 60 and 61), in which debiasing causes the loss of electrons. It is better explained by Argyriou et al 2023 (cited as reference 3): debiasing doesn’t destroy electrons, but moves them to adjacent pixels. This means it isn’t possible that “the brighter-fatter effect is not substantial in the WASP-107b data” (line 37) if debiasing is substantial, because the latter directly causes the former. I also recommend plotting a_ij,1 and tau_ik,1 as images. Is the trace visible, and does it have a systematically higher or lower value than the rest of the image? I suspect so, because where there is almost no flux, measuring a_ij,1 and tau_ik,1 is impossible.

In any case, I recommend obtaining a transmission spectrum with the default nonlinearity correction and comparing it to the one obtained with the alternate nonlinearity correction. I suspect the two will be very similar at longer wavelengths where all the interesting features are, but might be discrepant at shorter wavelengths.

Photochemical models:

Line 103: this is an interesting point, but the authors never come back to this. What is different about Tsai et al 2023’s photochemical models, such that they predict no detectable SO2 for WASP-107b while your models predict detectable SO2?

Line 668: how were these T/P profile parameters chosen? Since chemical reaction rates are highly sensitive to temperature, the authors should show that this T/P profile is justified. For example, is it consistent with the retrieved temperature at the photosphere?

The photochemical model uses the spectrum of HD 85512, which has similar FUV flux as WASP-107, but a X-ray flux 20x lower. The authors say that the X-rays do not significantly affect the model, but what about the EUV flux? If WASP-107 also has a EUV flux 20x lower than HD 85512, would the photochemistry be substantially different?

In the abstract, line 28 says “...indicate a dynamically active atmosphere with a super-solar metallicity”. Can the metallicity constraint be tightened? Extended Data Figure 3 seems to suggest that the SO2

feature is not visible until Z=6; since clouds would both block photochemistry-inducing radiation and mute spectral features, would it be fair to say that WASP-107b has Z > 6? Also, is this metallicity constraint consistent with that inferred from the retrieval?

Retrievals:

In Table 1, I recommend giving the mixing ratios (or upper limits) of all species in addition to their detection significances. Speaking of mixing ratios, based on the corner plots, they seem quite discrepant between the two retrievals for H2O (-2.3 vs. -3.8) and SO2 (-5.2 vs. -6.6). Why is this the case? For SO2, why was the value of -5.2 adopted (e.g. line 97) instead of -6.6?

Why does ARCiS infer a temperature at 1 bar of 530 K, compared to petitRADTRANS' 830 K? I see that there is a degeneracy between temperature and SO2 mixing ratio, as well as H2O mixing ratio. Could this explain the discrepancy in retrieved mixing ratios between the two codes?

I am worried about the high dimensionality of the retrievals. As a test, I suggest removing all gases that did not end up being detected (except CH4), all clouds other than a single silicate species, and all cloud particle size-related parameters (because in the Rayleigh regime, the size should trade off with number density perfectly). Do the results change significantly with this simplified retrieval?

Please mention on line 70 which data reduction was used for the retrievals. Do the results change if the other two data reductions are used instead?

Minor Comments:

Line 63: semi-major axis -> a/Rs?
Line 139-147: it would be useful to state the CH4 upper limit from the retrievals, as well as the CH4 abundance from your models, so that readers can judge the magnitude of the discrepancy. Also, unless I'm reading Extended Data Fig. 3 incorrectly, it seems that the CH4 feature is very prominent at 1x solar metallicity, but nearly disappears by 10x solar metallicity.

Line 414: the more conventional way to say it is that each integration consists of groups, and the groups contain frames. For MIRI, it doesn't matter because there's 1 frame per group.

Line 422: HST/WFG3 -> HST/WFC3
Line 445, 478: is the last group also flagged as 'do not use'?
Line 456: sum extraction or optical extraction?
Line 465, 490: how many hours is 744/250 integrations?
Line 485: what does it mean that two columns "were replicated"?
Line 489: What does "ran a sigma-clipping of 20 integrations" mean?
Line 506: is an exponential ramp also included? It's not mentioned here, but line 547 says that there's an exponential.
Line 511-512: this underestimates the transit depth uncertainties, because it doesn't include the errors caused by uncertainties in the detrending parameters. Since TEATRO is using MCMC, why not use the

chain to estimate the transit depth error (50th minus 16th percentile for the lower error, 84th minus 50th percentile for the upper error)?

Line 521: how was this custom dark reference file produced?

Line 535-536: Extended Data Fig 1 shows that for CASCADe, and for TEATRO below 11 um, the measured errors are systematically below the photon noise limit. This is not physically possible. I think part of what's happening is that the ETC includes flatfielding error in its error budget, but flatfielding error is irrelevant for time-series observations. The authors could use PandExo, which does not include flatfielding error. However, neither ETC nor PandExo knows the exact stellar spectrum of WASP-107. Rather than using PandExo, I recommend using the photon noise estimated by the data reduction pipelines. This is provided by Eureka!, and probably by the other two pipelines as well. This is still not a perfect solution because the gain is not accurately known, which (in my experience) results in overestimates of the error, but it is better than using PandExo.

Other ways of seeing the amount of correlated noise include heavily binning Extended Data Figure 6, or doing an Allan variance plot (STD vs. bin size, compared to 1/sqrt(N)).

Line 619-623: I find this explanation very confusing

Line 707: what is the difference between "line opacities" and "line absorption opacities"?

Extended Data - Table 2: The limb darkening coefficients are highly discrepant. Are you sure that Eureka! isn't reporting q1/q2 (Kipping 2013 parameters) instead of u1/u2? Also, the fixed value that TEATRO assumes for a/R_* is 11 sigma away from the Eureka! value. Could this impose an offset on the MIRI transmission spectrum?

Extended Data - Figure 3: please plot the data on top

Supplementary Info Line 214: for future work, XSPEC supports MCMC. There's no need to step through a grid of Z.

Author Rebuttals to Initial Comments:

Answer to referees' reports - WASP-107b *Nature* paper

Referee #1

This study presents valuable results regarding the discovery of SO2 in the atmosphere of the planet WASP-107b with unusual properties. The authors have included the necessary details along with a number of interesting diagrams to illustrate their points, which have made the manuscript understandable. However, to improve their work, I would recommend revising the manuscript based on the comments/suggestions below.

We thank the referee for his/her laudable words. We have addressed his/her concerns to the best of our ability and refer to our answers below.

Major Comments/suggestions

[R1_1] Lines 68-69: I do not believe the transmission spectra from the three data reduction methods are in "excellent agreement", based on Figure 1. There are noticeable discrepancies between these spectra, in particular, at longer wavelengths. I would suggest quantifying how well these spectra agree, instead of using a word like "excellent", and briefly explaining (in the main text) why the transmission spectrum from "CASCADe" is preferred.

The deviations between the three reductions are consistent with random noise as indicated by our error estimate. 5% of the 50 differences between the data points (i.e. 2 or 3 points) that we get for each reduction are between 2 and 3 sigma and the rest (95%) are below 2 sigma, according to SI Figure 8. Moreover, there is an uncertainty on the error estimate, but the three estimates follow the same general trend which is the one that we get from the ETC. We have changed that sentence in the main text and now report the agreement in terms of number of sigmas.

For presentation purposes, we have used the CASCADe reduction on Fig. 2, Table 1 and the spreadsheet that has been added to the review. We have run retrievals on the 3 transmission spectra and they all lead to the same conclusions, see SI Table 2.

[R1_2] Lines 102-107: This is one of the most important parts of this study. While it has been claimed that for Teq < 1000 K, SO2 production below the 0.01 mbar level is halted (Tsai et al. 2023), SO2 has been discovered in a planet with Teq = 740 K in a more recent study, which is interesting. The authors have brought some reasons (starting from line 108) to justify their finding. However, it would be better to add a separate paragraph before these justifications and argue against the claim of Tsai et al. 2023. For example, Extended Data Figure. 10 (b) in their paper shows the average VMR between 10 and 0.01 mbar as a function of Teq for some sulfur-bearing species. To produce these models, except for temperature profiles, all other planetary parameters were assumed to be

consistent with WASP-39b, which are not applicable to other planets with different properties like WASP-107b. In a comprehensive study, different values of parameters, such as irradiation, intrinsic temperature, density, chemical abundances, … should be examined to reach an accurate conclusion.

The authors agree that a more thorough explanation is needed to justify the difference with Tsai et al., 2023. We therefore have added a new section in the SI (Sect. 4.2). A short summary of that section was also added to the main text. The two critical parameters influencing the detectability of SO2 within the ~740K temperature regime of this low-density planet are the UV irradiation and gravity. Prior simulations were done at much higher gravities than is the case for WASP-107b. It is also important to note that the WASP-39b simulations were performed for a gravity of 1,000 cm/s2 (Tsai et al. 2023), but that the actual gravity of WASP-39b is more than a factor of 2 lower. Moreover, WASP-107b receives ~200 times less UV flux than WASP-39b. Our study reveals that the initiating pathway for SO2 formation in WASP-107b is twofold. Firstly, through H2O photodissociation in upper atmospheric layers below ~10^(-5) bar, yielding atomic H and OH radicals. Secondly, in the pressure range of ~10^(-5) - 1 bar, photolysis of various abundant molecules -- beyond just H2O -- significantly contributes to the supply of free atoms and radicals. These initiate a cascade of barrierless reactions that progressively culminate in sufficient quantities of OH radicals that can be used for oxidising SO2. Given the conditions of moderate UV irradiation and low gravity, these processes lead to detectable SO2 levels.

Another note concerns panel (b) of Extended Data Figure 10 of Tsai et al. (2023) which shows the average volume mixing ratio (VMR) between 10^(−5) – 10^(−2) bar to mark the required SO2 concentration to be detectable with WASP-39b parameters using JWST MIRI. Our models extend up to 10^(−7) bar and show that SO2 is mainly abundant between 10^(−4) and 10^(−7) bar. Hence, taking the average VMR between 10^(−5) – 10^(−2) bar is not an unequivocal indicator for the detectability (or not) of a particular molecule, here SO2.

[R1_3] Lines 118-121: Regarding the previous comment, the authors stated that the large atmospheric scale height of WASP-107b (which is its special property) enables highly efficient photochemical processes to operate within Teq ~ 740 K for this low-density planet. However, they have not examined how the molar fraction of SO2 can change with scale height (as Tsai et al. 2023 has not either). Such an investigation may also elucidate the detection of SO2 in WASP-107b.

We have investigated the effect of a larger atmospheric scale height, and hence lower gravity, in the new Sect. 4.2 of the SI, mentioned above under R1_2.

[R1_4] Lines 545-549: It is stated that the observed increasing trend can be attributed to the different systematic models employed by the TEATRO (an exponential model) and Eureka! (both an exponential and polynomial models) reduction codes. However, it should be explained how these

two codes are compared with the CASCADe data reduction and how this reduction does not play any role in the increasing trend of transit depth difference.

We wrote section 1.5 of the SI to provide a complete overview of the comparison between the three data reductions. There is no systematic bias (i.e. >1 sigma) between CASCADe/TEATRO and CASCADe/Eureka! for points shorter than 10 micron. The systematic differences are due to small differences in the fitted baseline at longer wavelengths (> 10 micron). At these wavelengths, the temporal drifts at the beginning of the time-series have a more pronounced shape. CASCADe uses a baseline based on the data itself, x and y position, FWHM and a time linear trend. TEATRO uses a linear trend, not an exponential model, and one and half hour of data is removed before fitting (this was a mistake in the initial writing and is now corrected). The difference between Eureka! and TEATRO also shows a trend at long wavelengths. We do not have strong evidence that one systematics model should be preferred over another, and using a different model in each reduction ensures that the results are not biased by the systematics model employed. Despite the different models, the conclusions from the retrievals on the 3 spectra are in agreement.

[R1_5] Extended Data Figure 1: It should be explained why the values of 1 σ error associated with CASCADe code are discrepant relative to the values from the two other reduction methods and the simulated transit spectrum at longer wavelengths.

This figure is now moved to the Supplementary Information (Fig. 7).
Four points are lower than the rest of the error estimates. The error of CASCADe is estimated based on a bootstrap analysis which gives lower estimates for bins that are at longer wavelengths where the scatter is higher. Also, for these longer wavelengths, the ETC estimation is higher than expected due to the uncertainties on the detector gain and an overestimation of the background (see also Bouwman et al., 2023). Any update to the ETC will provide a better estimation of the noise whenever we have more transit observations available to improve it.

[R1_6] While there is a direct observation of the host star for NUV, I am not clear why the observation of another star (similar to the host star) has been used.

The direct observation of the host star, from Swift-UVOT, used a photometry filter, to obtain relatively broad-band simultaneous and contemporaneous measurements. However, the atmospheric modelling of the exoplanet requires a relatively high resolution spectrum, hence the use of the closest-match available 'template' star, from the MUSCLES programme (i.e. an HST spectrum), to provide the detailed spectral shape, while the Swift photometry provides the scaling to the host star. Additionally, the photoabsorption cross sections span wavelengths from ~10 nm up to ~800 nm, so panchromatic information is needed.

Minor Comments/suggestions

[R1_7] In Figures 1 and 2 (the bottom part, showing cloud and molecular contributions) as well as Extended Data Figures 1 and 2 (the bottom left panel), colors with higher contrast should be used to better distinguish them.

We have adapted all those figures using colours with higher contrast. Note that some of the Extended Data figures are now put in the Supplementary Information to avoid fragmentation of the text.

[R1_8] Line 20: I would suggest mentioning the wavelengths 8.69 μm and 7.35 μm after "two fundamental vibration bands of SO2" to give the readers a general understanding about the features of interest at the beginning of the paper.

Suggestion has been implemented.

[R1_9] Lines 55-56: There is no need to include technical terms like "read-out pattern FASTR1" in the main text; but these should be introduced or defined in the Method section only. A simpler sentence may be sufficient to convey the point: "The subarray SLITLESSPRISM was used for a spectral resolution varying between 30 and 100."

Suggestion has been implemented.

[R1_10] Figure 2: In addition to molecular contribution shown in the bottom part of the figure, I would recommend marking the prominent spectral features due to H2O, H2S, and CO, similar to those from SO2 and NH3. In addition, marking or tagging the HST data points in some way would help the readers follow the figure more easily.

Comment has been implemented with the exception of H2S as it does not have a clear, sharp feature in the spectrum but rather several broader bands.

[R1_11] Line 418: There is no definition for the acronym "MAST".

Comment has been implemented.

[R1_12] Lines 468-469: The acronym FWHM must be defined where it appears first, i.e., line 461.

Comment has been implemented.

[R1_13] Lines 470 and 508: “band averaged light curve” or “white light curve”? It would be more appropriate to have a consistency in terminology.

We have opted for the words ‘band-averaged light curve’, since this is a better description of what it actually is.

[R1_14] Lines 469-471: “The orbital parameters of WASP-107b …from the TEATRO data reduction.” These orbital parameters must also be mentioned in the section of TEATRO data reduction setup.

In the TEATRO section, we added a reference to Extended data Table 2 where these orbital parameters are reported.

[R1_15] Lines 55, 413, 434, 486: “subarray” or “sub-array”. It would be better to use the same term in the entire manuscript.

Comment has been implemented; we have opted for the word `subarray’.

[R1_16] Lines 544-545: “mostly within 2σ”, but in Extended Data Figure 7, it is mentioned “mostly within 1.5σ”… These two must be the same.
We have changed the value to 2 sigma in the caption of that Figure and have quantified the word `mostly’ as being 95%.

[R1_17] Lines 589-613: It is very difficult to read all the retrieval parameters and their corresponding priors in this part. I would suggest putting these parameters and priors into a table, which could make them easy to follow.

We agree with this. The parameters and priors are now all collected into a table.

[R1_18] Suppl. Inf. Figures 5 and 6: The title axes of the corner plots are too small to be useful. One possible solution would be to plot the 1D histograms of relevant parameters (e.g., SO2 abundance) separately, in a legible figure.

Owing to a reshuffling of the Method Section in the Supplementary Information, these figures are now Suppl. Inf. Fig. 10 and 11.

We opt not to modify these figures, as their current format encapsulates all essential information necessary for a comprehensive comprehension of the retrieval results. Omitting portions of the figure would entail conveying only partial insights. As these figures are accessible within the online Supplementary Information section, readers have the capacity to zoom in, ensuring a detailed examination.

[R1_19] I have found some typos and grammatical mistakes. So, the manuscript should be carefully checked and corrected if needed.

We have reread the full manuscript, and hope to have corrected all typos and grammatical mistakes.

Referee #2

[R2_1] The manuscript presents novel and interesting results of significant interest to the exoplanet community. In particular, it is the first detection of silicate clouds in an exoplanet (with the possible exception of VHS 1256 b), which have long been theoretically predicted. The detection of SO2 and constraining non-detection of CH4 are not unprecedented in the field, but they add to the scientific importance of the paper by demonstrating clear evidence of disequilibrium chemistry. I recommend that the paper be published once my concerns are addressed.

We thank the referee for his/her laudable words. We have addressed his/her concerns to the best of our ability and refer to our answers below.

Silicate clouds:

[R2_2] Since the most significant result in the paper is the discovery of silicate clouds, I recommend doing further tests to establish the robustness of this result. In particular, in the ARCiS retrieval, SiO2 and MgSiO3 abundances are consistent with 0, but SiO and C abundances are not; in the petitRADTRANS retrieval, KCl and MgSiO3 abundances are consistent with 0, but SiO2 abundances are not. Which of these species is driving the detection of silicate clouds? Given the visually obvious dropoff beyond 10.5 um, it's probably one of the silicates, but which one? Are silicate clouds still preferred if a retrieval is done with only MgSiO3, only SiO, or only SiO2, without any C or KCl? In the Rayleigh regime (which <0.1 um particles would definitely be in), extinction is dominated by absorption instead of scattering, and the absorption cross-section is 8*pi^2*r^3/lambda * Im((m^2 - 1) / (m^2 + 1)) for spherical particles with complex refractive index m (Equation 7.4 of Mishchenko 2002: https://pubs.giss.nasa.gov/books/2002_Mishchenko_mi06300n/2002_Mishchenko_etal_1.pdf) It may be useful to plot, for different species, the natural log of this value vs. wavelength and compare it to the transit spectrum expressed in scale heights. They ought to have roughly the same shape at wavelengths dominated by silicate absorption (but not necessarily the same scale, if the scale height of the silicate clouds is different from the scale height from gas).

We performed retrievals with single cloud materials to determine which component dominates. We find that all silicate components (SiO, SiO2 and MgSiO3) by themselves provide a significant improvement of the fit compared to the parameterised clouds. We have added some discussion on this to the text of the paper. Also we added a figure with the silicate opacities plotted compared to the transit spectrum.

[R2_3] Along the same lines, more should be said about how "we mix the refractive indices of these materials using effective medium theory" (line 569) and how "we included amorphous MgSiO3, SiO2, and crystalline KCl clouds, considering them to be irregularly shaped (DHS method)". Since the inference of silicate clouds come entirely from the wavelength dependence of the refractive indices, it would be useful to have more details about how these were computed for the mixture and whether the computation is reliable.

The cloud refractive indices have been mixed together using a standard Bruggeman effective medium approach. This is now also described in the paper.

[R2_4] Finally, the authors say in the main text that "the retrieval models favour a very high-altitude cloud layer composed of tiny (<0.1 um) amorphous silicate particles", but I don't see any parameter in the corner plot for either retrieval that corresponds to the particle size, and I also don't see any equation that gives the typical particle size as a function of the retrieval parameters. Since the particle size is itself an interesting physical quantity that informs atmospheric chemistry and dynamics, not to mention the cloud's absorption properties, could the authors plot the posterior of the particle size?

This wording referred to the ARCiS retrievals where the particle size is a free parameter given in the corner plot as a_{cloud}. It can be seen that this parameter is constrained to be smaller than 0.1 micron (see cornerplot, Suppl. Inf. Fig. 10). For the pRT retrievals the size of the particles is computed from the Ackerman and Marley cloud formation formalism. These particles come out to be very small in the upper atmosphere and somewhat larger (micron sized) near the cloud base. We reworded this sentence to better also represent the pRT retrievals.

Data analysis:

[R2_5] I recommend discarding at least the first two data points in the spectrum, conservatively the first three. According to the JWST documentation (https://jwst-docs.stsci.edu/jwst-mid-infrared-instrument/miri-observing-modes/miri-time-series-observations/miri-lrs-tsos), "the dispersion profile of the LRS turns over below 4.5 µm—for a limited wavelength range around 4.5–5 µm, different wavelengths are dispersed onto the same detector pixels." The 0.15 um bin also corresponds to very few pixels below 5 um (only 1 pixel, for 4.61 um), further decreasing the reliability of these bins. I expect this to further decrease the significance of the (very marginal) CO detection, which seems driven by the first 2-3 points.

We confirm that the CO detection is indeed entirely based on the first 3 wavelength points (between 4.6 and 5 micron). If we remove these points from the retrieval analysis, the tentative CO detection goes away. Therefore, confirmation of this result at shorter wavelengths is required. We added this sentence to the text. Note that we do not expect the wavelengths from 4.6 micron

onwards to be strongly contaminated. The shortest wavelengths reached before the reversal of the dispersion pattern is about 3.9 micron. Any light contaminating the spectra at 4.6 microns has a wavelength well short of that, which is outside of the CO/CO2 band. Furthermore, for wavelengths much shorter than 4 microns, the photon-electron conversion efficiency drops to close to 0, so the contamination itself drops very fast the further one is from the reversal point.

[R2_6] The default non-linearity correction clearly has problems, but I have concerns about the authors' own non-linearity correction. Suppl. Inf. Figure 1 plots pair-wise differences for different rows in column 36. What do columns 34 and 37 look like? In all the MIRI datasets I've seen, the differences go up with pair number in columns 34 and 37, vs. down in column 36. This phenomenon cannot be accounted for by the authors' equation (between line 60 and 61), in which debiasing causes the loss of electrons. It is better explained by Argyriou et al 2023 (cited as reference 3): debiasing doesn't destroy electrons, but moves them to adjacent pixels. This means it isn't possible that "the brighter-fatter effect is not substantial in the WASP-107b data" (line 37) if debiasing is substantial, because the latter directly causes the former. I also recommend plotting a_ij,1 and tau_ik,1 as images. Is the trace visible, and does it have a systematically higher or lower value than the rest of the image? I suspect so, because where there is almost no flux, measuring a_ij,1 and tau_ik,1 is impossible.

In any case, I recommend obtaining a transmission spectrum with the default nonlinearity correction and comparing it to the one obtained with the alternate nonlinearity correction. I suspect the two will be very similar at longer wavelengths where all the interesting features are, but might be discrepant at shorter wavelengths.

Our first reductions were made using the default non-linearity correction. This correction was not accurate enough (SI Fig. 7), we developed our own correction and used it afterwards. Indeed, there were observable differences at the shorter wavelengths and no differences at longer wavelengths as no detector ramp is effectively linear.
The main non-linearity effects seen in the data are a combination of debiasing and electron diffusion to neighbouring pixels due to a gradient in the electric field over the detector pixels, which leads to the so called brighter-fatter-effect (Argyriou et al., 2023). The standard RSCD linearity correction was derived using extended source data. As there is no flux difference between detector pixels in this type of data, the observed non-linearity is purely a debiasing effect as there is no electrical field difference between neighbouring pixels and thus no electron diffusion. This means that for a point source, the standard CRDS calibration will under-correct the pixels in the center of the PSF and over-correct the pixels in the wings, as they will, respectively, lose or gain electrons due to diffusion.

Note that all of these effects are minor in the data of WASP-107b as the maximum signal on the detector ramps stays well away from the saturation limit of the detector. However, small effects can still be seen which warrant a custom calibration. As a matter of fact, we opted to fit a pure debiasing mode. While this is strictly speaking indeed not entirely correct, it can still very well correct the ramps in case of

WASP-107b as the electron diffusion component is comparatively small. What one effectively derives as an “effective” debiasing (i.e. response drop towards higher signal levels on the ramp), which is larger in the center of the PSF and lower in the wings compared to the CRDS correction.

We adjusted the text to reflect the explanation above. We also added 1 additional figure (SI Fig. 8) showing not just the linearisation of pixels along the spectral trace for the brightest detector column, but also across the trace, for a short wavelength detector row (high flux), showing also the detector columns in the wings of the PSF. This figure shows that our model effectively linearizes the detector ramps in contrast to the CRDS calibration file. Note that for wavelengths longer than 8 micron, no differences can be seen between our calibration and the CRDS one, as all data is effectively linear, as can be seen from the figures in section 1.4.

Photochemical models:

[R2_7] Line 103: this is an interesting point, but the authors never come back to this. What is different about Tsai et al 2023’s photochemical models, such that they predict no detectable SO2 for WASP-107b while your models predict detectable SO2?

This comment is similar to comment [R1_2]. For clarity, we copy/paste here that answer:

The authors agree that a more thorough explanation is needed to justify the difference with Tsai et al., 2023. We therefore have added a new section in the SI (Sect. 4.2). A short summary of that paragraph was also added to the main text. The two critical parameters influencing the detectability of SO2 within the ~740K temperature regime of this low-density planet are the UV irradiation and gravity. Prior simulations were done at much higher gravities than is the case for WASP-107b. Moreover, WASP-107b receives ~200 times less UV flux than WASP-39b. Our study reveals the initiating pathway for SO2 formation in WASP-107b is twofold. Firstly, through H2O photodissociation in upper atmospheric layers below ~10^(-5) bar, yielding atomic H and OH radicals. Secondly, in the pressure range of 10^(-5) - 1 bar, photolysis of various abundant molecules -- beyond just H2O -- significantly contributes to the supply of free atoms and radicals. These initiate a cascade of barrierless reactions that progressively culminate in sufficient quantities of OH radicals that can be used for oxidising SO2. Given the conditions of moderate UV irradiation and low gravity, these processes lead to detectable SO2 levels.

[R2_8] Line 668: how were these T/P profile parameters chosen? Since chemical reaction rates are highly sensitive to temperature, the authors should show that this T/P profile is justified. For example, is it consistent with the retrieved temperature at the photosphere?

The parametrization of Guillot, 2010 was chosen for its simplicity and flexibility (e.g. regarding the T_{int} parameter) despite its limitations. For example, the upper atmosphere is assumed to be isothermal, which is a simplification. However, the composition within the upper layers is dominated

by disequilibrium chemistry, reducing the role of local temperature variations. Another example is that no convection is considered, which could imply an overestimation of the temperatures in the deep atmosphere for a given T_{int} (Guillot+2010, Section 5.2). Because the goal of the forward chemistry models is not to reproduce the data, but rather show behavioural trends (e.g. with variation of metallicity, C/O, Kzz, T_{int}, …), the authors consider the adopted (P-T)-profile acceptable.

The (P-T)-profile parameters κ_{IR} (the atmospheric opacity in the IR wavelengths, i.e. the cross-section (cm2) per unit mass (g)) and γ (the ratio between the optical and IR opacity) were chosen based on Fig. 4 of Guillot, 2010. While the above parameters should vary for non-solar atmospheric compositions, the choice was made to keep these parameters fixed for all simulations to avoid any chemical variations from these temperature parameters.

The retrieved temperatures are indeed consistent with the photospheric temperature in this profile (see also the answer to R2_11).

[R2_9] The photochemical model uses the spectrum of HD 85512, which has similar FUV flux as WASP-107, but a X-ray flux 20x lower. The authors say that the X-rays do not significantly affect the model, but what about the EUV flux? If WASP-107 also has a EUV flux 20x lower than HD 85512, would the photochemistry be substantially different?

We did not alter the MUSCLES SED of HD 85512 to match the measured fluxes of WASP-107b. The main reason for this is the lack of photometric flux values outside the X-ray and NUV wavelength domain. Therefore, one would have to scale certain parts of the stellar spectrum and break the continuity of the reconstructed SED as presented in the MUSCLES survey.

To assess the impact of a different stellar SED as an input in the photochemical models, we have run a model with the SED of WASP-39; see Sect. 4.2 in the SI.

[R2_10] In the abstract, line 28 says "...indicate a dynamically active atmosphere with a super-solar metallicity". Can the metallicity constraint be tightened? Extended Data Figure 3 seems to suggest that the SO2 feature is not visible until Z=6; since clouds would both block photochemistry-inducing radiation and mute spectral features, would it be fair to say that WASP-107b has Z > 6? Also, is this metallicity constraint consistent with that inferred from the retrieval?

Indeed, the retrieval codes favour models with a higher mean molecular weight, with the 1-sigma lower boundary being six times solar. However, we refrain from putting this number explicitly in the main text (and summary paragraph) since that value (of above six times solar as deduced from Extended Data Fig. 2) is also dependent on the C/O ratio, the incident UV flux etc. We therefore still opt to use the words 'super-solar metallicity' instead of putting an explicit number here.

Retrievals:

[R2_11] In Table 1, I recommend giving the mixing ratios (or upper limits) of all species in addition to their detection significance. Speaking of mixing ratios, based on the corner plots, they seem quite discrepant between the two retrievals for H2O (-2.3 vs. -3.8) and SO2 (-5.2 vs. -6.6). Why is this the case? For SO2, why was the value of -5.2 adopted (e.g. line 97) instead of -6.6?

We have added the derived abundances now to the table. Indeed, as the referee mentions there are some differences in the absolute abundances of the molecules between the ARCiS and pRT retrievals. This comes partly from the well known degeneracy between clouds and metallicity; adding more or less cloud opacity can significantly change the absolute abundances. This is why the absolute abundances are also not very stable between various retrieval setups.

Because of these reasons we now do not talk about absolute abundances anymore in the text. We added an explanation of these issues to the text.

[R2_12] Why does ARCiS infer a temperature at 1 bar of 530 K, compared to petitRADTRANS' 830 K? I see that there is a degeneracy between temperature and SO2 mixing ratio, as well as H2O mixing ratio. Could this explain the discrepancy in retrieved mixing ratios between the two codes?

A large part of the difference here is that ARCiS retrieves a P-T profile with a gradient and a temperature at 1 bar. The observable atmosphere is significantly higher up. If we look at the confidence interval of that non-isothermal P-T structure around the region where the atmosphere is best constrained (i.e. around 1e-5 bar), the temperature retrieved by ARCiS is 724+-47 K, much closer to the value retrieved by pRT for an isothermal temperature structure.

We have now changed this also in the text and in the cornerplot (where we give the value of the profile at 1e-5 bar for ARCiS now).

[R2_13] I am worried about the high dimensionality of the retrievals. As a test, I suggest removing all gases that did not end up being detected (except CH4), all clouds other than a single silicate species, and all cloud particle size-related parameters (because in the Rayleigh regime, the size should trade off with number density perfectly). Do the results change significantly with this simplified retrieval?

We conducted this experiment exactly as the referee suggests (although we kept the size of the cloud particles in as this can influence the spectral properties when the particles are not in the Rayleigh regime). Even though some of the absolute abundances of the molecules are different (the

ARCiS abundances are for this setup closer to the ones derived with pRT), the ratios of the molecules and the cloud properties are unaffected. Thus, our conclusions are robust against this.

We have chosen not to include these results in the paper to avoid overcrowding the reader with retrieval results.

[R2_14] Please mention on line 70 which data reduction was used for the retrievals. Do the results change if the other two data reductions are used instead?

We have used the CASCADe data reduction for the results presented and have also run the most important retrievals (i.e. those directly supporting the main conclusions) with the other two reductions. The results are very robust between the three. We now added some words in the main text and the supplementary information explaining this.

Minor Comments:

[R2_15] Line 63: semi-major axis -> a/Rs?

Indeed correct. Has been changed in the text.

[R2_16] Line 139-147: it would be useful to state the CH4 upper limit from the retrievals, as well as the CH4 abundance from your models, so that readers can judge the magnitude of the discrepancy. Also, unless I'm reading Extended Data Fig. 3 incorrectly, it seems that the CH4 feature is very prominent at 1x solar metallicity, but nearly disappears by 10x solar metallicity.

Indeed, in the previous version we wrote *"At super-solar metallicity values, our models also predict a detectable CH4 feature in the JWST MIRI wavelength range at 7.8 µm."* However, this should be *"At super-solar metallicity values, our models also predict a detectable CH4 feature in the JWST MIRI wavelength range".*
Ext. Data Fig. 2 indeed shows that at Z=10 Zsun the contribution of CH4 to the 7.8 micron feature is very small and that the SO2 feature becomes the most prominent. CH4 is still present in these atmosphere models, but simply 'masked' by SO2 opacity at 7.8 um. While the SO2 opacity dominates the 7.8 micron feature, the absence of CH4 (in the MIRI wavelength range) is mainly based on data points between 5-7 micron where our forward models show a combination of H2O and CH4 opacity, while retrievals fit this wavelength region with purely H2O bands. In particular, the wavelength region around 6.3 micron is discriminating in detecting CH4 or not in the retrieval modelling. Additionally, the same super-solar metallicity synthetic spectrum (Z = 10 Zsun) shows prominent CH4 features in the HST and NIRCam wavelength range, which are not detected in the data (Kreidberg+2018 & NIRCam paper). This indicates that our forward models (with base

parameters Tint = 400k, C/O solar, and log(Kzz) =10) produce too much CH4, although not deducible from the 7.8 micron feature.

[R2_17] Line 414: the more conventional way to say it is that each integration consists of groups, and the groups contain frames. For MIRI, it doesn't matter because there's 1 frame per group.

Has been added.

[R2_18] Line 422: HST/WFG3 -> HST/WFC3

Has been corrected.

[R2_19] Line 445, 478: is the last group also flagged as 'do not use'?

Yes, it is impacted by the last frame effect (Argyriou et al., 2021). We added this to the text to make clear we drop also the last frame.

[R2_20] Line 456: sum extraction or optical extraction?

It is an aperture extraction, so a sum within an 8 pixel box. We now clearly state this in the text.

[R2_21] Line 465, 490: how many hours is 744/250 integrations?

744 integrations is 1.3 hours and 250 is 0.4 hour.

[R2_22] Line 485: what does it mean that two columns "were replicated"?

The background is a median value of several columns (7) on the left and 7 on the right of the trace (column 36). We updated the sentence in the paper.

[R2_23] Line 489: What does "ran a sigma-clipping of 20 integrations" mean?

It is a technique that uses a running median over the whole time series to detect outliers based on a 5 sigma rejection threshold. This method is used to make sure no outlier is left in the data for all spectroscopic channels as well as the white lightcurve.

[R2_24] Line 506: is an exponential ramp also included? It's not mentioned here, but line 547 says that there's an exponential.

There is no exponential, only a linear trend, and the initial decay is simply discarded. It was a mistake in the text, it has now been corrected.

[R2_25] Line 511-512: this underestimates the transit depth uncertainties, because it doesn't include the errors caused by uncertainties in the detrending parameters. Since TEATRO is using MCMC, why not use the chain to estimate the transit depth error (50th minus 16th percentile for the lower error, 84th minus 50th percentile for the upper error)?

The transit depth uncertainties computed from the MCMC chains are even smaller. We have adopted this other method to have more conservative uncertainties. We have added that information in the text. Variations of the detrending parameters would move the residuals around zero. When applying variations of about 1 sigma on these parameters, the residuals remain centred on zero (and remain that way along the time dimension), thus any offset is much smaller than the lightcurves' standard deviation. Systematics are small in this data set after the initial ramp, which we cut in TEATRO.

[R2_26] Line 521: how was this custom dark reference file produced?

It is produced by taking the standard CRDS (calibration) dark file and running a median smoothing (or running median) to remove any excess scatter in the dark estimate. This sentence has been added to the SI.

[R2_27] Line 535-536: Extended Data Fig 1 shows that for CASCADe, and for TEATRO below 11 um, the measured errors are systematically below the photon noise limit. This is not physically possible. I think part of what's happening is that the ETC includes flatfielding error in its error budget, but flatfielding error is irrelevant for time-series observations. The authors could use PandExo, which does not include flatfielding error. However, neither ETC nor PandExo knows the exact stellar spectrum of WASP-107. Rather than using PandExo, I recommend using the photon noise estimated by the data reduction pipelines. This is provided by Eureka!, and probably by the other two pipelines as well. This is still not a perfect solution because the gain is not accurately known, which (in my experience) results in overestimates of the error, but it is better than using PandExo.

Other ways of seeing the amount of correlated noise include heavily binning Extended Data Figure 6, or doing an Allan variance plot (STD vs. bin size, compared to 1/sqrt(N)).

See also our answer to [R1_5].
Four points are lower than the rest of the error estimates. The error of CASCADe is estimated based on a bootstrap analysis which gives lower estimates for bins that are at longer wavelengths where the scatter is higher. Also, for these longer wavelengths, the ETC estimation is higher than expected due to the uncertainties on the detector gain and an overestimation of the background (see also Bouwman et al., 2023). Any update to the ETC will provide a better estimation of the noise whenever we have more transit observations available to improve it.

[R2_28] Line 619-623: I find this explanation very confusing

We agree with the referee that this was confusing. We have chosen to remove this entire paragraph and switch to the formalism presented in Benneke & Seager 2013 for both pRT and ARCiS (where we already used this formalism). We managed to implement a numerically stable way of doing this also for large Bayes factors. This implies some of the significances changed slightly, but all conclusions are unaffected.

[R2_29] Line 707: what is the difference between “line opacities” and “line absorption opacities”?

We refer twice to `line absorption opacities’, and have changed the text to avoid confusion.

[R2_30] Extended Data - Table 2: The limb darkening coefficients are highly discrepant. Are you sure that Eureka! isn’t reporting q1/q2 (Kipping 2013 parameters) instead of u1/u2? Also, the fixed value that TEATRO assumes for a/R_* is 11 sigma away from the Eureka! value. Could this impose an offset on the MIRI transmission spectrum?

Limb darkening coefficients: Indeed, the coefficients reported by Eureka! are the Kipping 2013 parameters. We put the u1 and u2 parameters instead.

a/R*: In the TEATRO band-averaged light curve fit, a/R* is fixed but the impact parameter is free (which is equivalent to letting the inclination free) and the impact parameter is the quantity that shapes the lightcurve. The impact parameters from Eureka! and TEATRO agree at about 1 sigma. Thus any error in a/R* is mitigated by slightly adjusting the inclination. The transit duration is also well modelled: no deviations appear in the band-averaged lightcurve residuals around the ingress or egress. For the spectral lightcurves, we fix the impact parameter to the value derived from the band-averaged light curve (i.e. we do not use a/R* directly). Thus this discrepancy in a/R* does not play a significant role in our lightcurve fits and spectrum. Despite our slightly different assumptions, our results from retrievals remain consistent for the three reductions.

[R2_31] Extended Data - Figure 3: please plot the data on top

We have opted not to incorporate this suggestion, as our approach does not involve a direct comparison between predictions and data. This is due to the absence of cloud considerations in the forward gas-phase models. The intent of this figure is to illustrate the sensitivity of molecular features to metallicity within a cloud-free atmosphere. The inclusion of MIRI data in this figure would introduce complexity beyond the intended scope.

[R2_32] Supplementary Info Line 214: for future work, XSPEC supports MCMC. There's no need to step through a grid of Z.

Thanks for this comment.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

A. Great
B. Original and significant
C. Great
D. Great
E. Great
F. Good
G. Great
H. Great

I am happy to inform you that the corrections associated with my comments/suggestions are

satisfying and no further changes are needed.

Referee #2 (Remarks to the Author):

Thank you for your comprehensive work in addressing my comments. They greatly increase the robustness of the results and inspire confidence in their validity. I have some minor comments left, but no more major concerns.

(Incidentally, in the future, please bold the changes made to the manuscript in response to the referee comments. Without bolding, it was very hard to tell what changed.)

Remaining minor comments:

Unless I missed it, the first part of R2_16 seems unaddressed: "it would be useful to state the CH4 upper limit from the retrievals, as well as the CH4 abundance from your models, so that readers can judge the magnitude of the discrepancy."

"This wording referred to the ARCiS retrievals where the particle size is a free parameter given in the corner plot as a_{cloud}. It can be seen that this parameter is constrained to be smaller than 0.1 micron (see cornerplot, Suppl. Inf. Fig. 10)"

I don't see any a_cloud in Suppl. Inf. Fig. 10.

"As there is no flux difference between detector pixels in this type of data, the observed non-linearity is purely a debiasing effect as there is no electrical field difference between neighbouring pixels and thus no electron diffusion. This means that for a point source, the standard CRDS calibration will under-correct the pixels in the center of the PSF and over-correct the pixels in the wings, as they will, respectively, lose or gain electrons due to diffusion."

So the default non-linearity correction removes the debiasing effect, meaning there is no benefit to the custom correction if it is conceived purely as a model of debiasing. The custom correction is better thought of as a parametric model of debiasing + diffusion, and the benefit comes from the correction of diffusion, not the correction of debiasing. The custom correction fits the data better, but does that necessarily mean the a_ij,0 obtained this way, when summed across the spectral direction, is a better estimate of the true flux? It's plausible that it is, but there are plausible reasons to think otherwise. For example, if electrons are diffusing in the spatial direction, but there's little net diffusion in the spectral direction because the pixel brightnesses don't change very

much in that direction, it's possible that adding up all the electrons in the spatial direction would get you the right answer.

For this paper, I recommend calling the custom correction a parametric model of non-linearity, which consists of debiasing + diffusion, and stating that the advantage over the CRDS calibration is that the latter only corrects debiasing and not diffusion.

"What one effectively derives as an "effective" debiasing (i.e. response drop towards higher signal levels on the ramp), which is larger in the center of the PSF and lower in the wings compared to the CRDS correction."

Actually the "effective" debiasing is negative in [385,38] and [385,35], as shown in Suppl. Inf. Figure 4.

"Also, for these longer wavelengths, the ETC estimation is higher than expected due to the uncertainties on the detector gain and an overestimation of the background (see also Bouwman et al., 2023). Any update to the ETC will provide a better estimation of the noise whenever we have more transit observations available to improve it."

Importantly, the ETC estimation includes the flat fielding error, which is not relevant for transit observations. Please plot the PandExo noise estimate instead. Also, because of the limitations you pointed out, please remove the phrase "and demonstrates that the data reductions are consistent with the ETC estimation and that no major systematics are left." If the noise estimate is known to be too high, being consistent with it means there are still systematics left.

Author Rebuttals to First Revision:

Answer to referees' reports - WASP-107b *Nature* paper

Second revision

Minor Comments/suggestions

Referee #2:

[R2_1] Unless I missed it, the first part of R2_16 seems unaddressed: "it would be useful to state the CH4 upper limit from the retrievals, as well as the CH4 abundance from your models, so that readers can judge the magnitude of the discrepancy."

The CH4 upper limit value from retrievals can be found in the main text, line 99: "The upper limit of its volume mixing ratio (VMR) being a few times 10^{-6}" and the CH4 upper limit from the models varies between 10^{-4} and 10^{-6} depending on the chosen intrinsic temperatures (please refer to Extended Figure 3).

[R2_2] "his wording referred to the ARCiS retrievals where the particle size is a free parameter given in the corner plot as a_{cloud}. It can be seen that this parameter is constrained to be smaller than 0.1micron (see cornerplot, Suppl. Inf. Fig. 10)"

I don't see any a_cloud in Suppl. Inf. Fig. 10.

The a_cloud parameter posterior distribution is shown in Suppl. Inf. Fig. 10 on line 18 of the corner plot.

[R2_3] "As there is no flux difference between detector pixels in this type of data, the observed non-linearity is purely a debiasing effect as there is no electrical field difference between neighbouring pixels and thus no electron diffusion. This means that for a point source, the standard CRDS calibration will under-correct the pixels in the center of the PSF and over-correct the pixels in the wings, as they will, respectively, lose or gain electrons due to diffusion."

That is correct.

[R2_4] So the default non-linearity correction removes the debiasing effect, meaning there is no benefit to the custom correction if it is conceived purely as a model of debiasing. The custom correction is better thought of as a parametric model of debiasing + diffusion, and the benefit comes from the correction of diffusion, not the correction of debiasing. The custom correction fits the data better, but does that necessarily mean the a_ij,0 obtained this way, when summed across the spectral direction, is a better estimate of the true flux? It's plausible that it is, but there are plausible reasons to think otherwise. For example, if electrons are diffusing in the spatial direction, but there's little net diffusion in the spectral direction because the pixel brightnesses don't change very much in that direction, it's possible that adding up all the electrons in the spatial direction would get you the right answer.

We thank the referee for their thorough review of this section of the supplement to help improve the description of one of the main systematics in the LRS data. It is indeed correct, that we fit a parametric model to the combined debiassing + diffusion, resulting in an improvement compared to the standard CRDS correction, which only takes into account the debiassing. Concerning the true flux estimate, this is not a simple thing to answer. Note that the standard calibration is derived by fitting a simple polynomial to the detector ramps, which might also lead to a not 100% correct result, at least what the absolute flux calibration is concerned at different signal levels, even for extended sources where electron diffusion should not be important. Unfortunately, it is not the case that all electrons diffused out of the PSF centre will be captured in the detector pixels in the wings of the PSF. Thus a summation per detector row (assuming no or very little diffusion in the dispersion direction) will not recover the true signal, even by increasing the aperture.

Rather than the absolute flux calibration, the most important thing concerning the flux calibration of the TSO observations is the relative response of the detector. There should be no different response over the transit, i.e. at different signal levels. We are convinced that our solution indeed provides this. The best way to test this is to look at the shape of the PSF with increasing signal levels and across the transit. The photometric response of the pixels in the centre as well as in the wings of the PSF should be identical, meaning that no shape difference should be observed. Looking at figures 5 and 6 of the supplement, where we use the FWHM as a measure of the PSF shape, this is indeed the case after using our parametric model.

[R2_5] For this paper, I recommend calling the custom correction a parametric model of non-linearity, which consists of debiasing + diffusion, and stating that the advantage over the CRDS calibration is that the latter only corrects debiasing and not diffusion.

That is indeed a good suggestion to avoid confusion. We adopted it in the text.

[R2_6] “What one effectively derives as an “effective” debiasing (i.e. response drop towards higher signal levels on the ramp), which is larger in the center of the PSF and lower in the wings compared to the CRDS correction.”

Correct.

[R2_7] Actually the “effective” debiasing is negative in [385,38] and [385,35], as shown in Suppl. Inf. Figure 4.

Indeed, for the few detector rows, receiving the highest flux levels, the “effective” debiasing is starting to become negative. For most pixels, however, it will be positive or zero. As the maximum signal levels in this data set are well below the saturation level, strongly upward curved ramps can not be seen. This is in contrast to data sets where the saturation level is reached or exceeded, where a clear upward curvature of the ramps is seen.

[R2_8] “Also, for these longer wavelengths, the ETC estimation is higher than expected due to the uncertainties on the detector gain and an overestimation of the background (see also Bouwman et al., 2023). Any update to the ETC will provide a better estimation of the noise whenever we have more transit observations available to improve it.”

Importantly, the ETC estimation includes the flat fielding error, which is not relevant for transit observations. Please plot the PandExo noise estimate instead. Also, because of the limitations you pointed out, please remove the phrase “and demonstrates that the data reductions are consistent with the ETC estimation and that no major systematics are left.” If the noise estimate is known to be too high, being consistent with it means there are still systematics left.

At the time of writing the paper and our response to the referee, PandExo was based on an older version of the underlying code (Pandeia) used to calculate the performance of the JWST instruments. The problem was that the gain factors were not updated in PandExo, leading to incorrect results. Further, we checked with the TSO working group lead at STScI (Nestor Espinoza) what has been implemented in the ETC and he ensured us that for any time series modes, no flatfielding error is taken into account. As the MIRI LRS slitless sub-array is exclusively intended for TSO observations, no flatflelding error is taken into account in the ETC. We, therefore, are convinced that our calculations are representing the correct noise estimates appropriate for transit observations. Note that there are still uncertainties in the estimates based on the ETC, certainly at the level of transit observations. Some uncertainties on the exact values of the detector gain, telescope throughput, photon conversion efficiency, and background levels in the slitless sub-array, remain. The referee is indeed correct to say that with this we can state that our noise estimates are mostly consistent with the ETC, but given all uncertainties we can not rule out

that some systematics still remain in the data. We, therefore, removed the sentence the referee requested for section 1.6 of the supplement.