Peer Review File

Manuscript Title: Inhomogeneous terminators on the exoplanet WASP-39~b

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

This is an interesting paper that reports potentially important results concerning limb differences. The results are at least qualitatively in line with expectations from models. But that doesn't necessarily mean that the results are correct in the details, and I think that some additional investigation and clarification are needed. The authors have done a lot of work that I appreciate, but I think more is needed to firmly establish such an important result.

The authors may be aware that there is another similar paper on this same topic using data for the same planet, with a similar result in terms of temperature difference, but not as similar in cloud properties. Given that the total light blocked by the limb is sensitive to scale height (via temperature) and cloud content, and that it's well known that there can be a degeneracy between these variables, I think more clarification/discussion is needed concerning the trade-off between cloud height and temperature. For example, I would expect to see cloud height (or some other cloud property) and temperature plotted versus each other as a panel in corner plots for each limb, and those plots are missing from this paper. Given that an independent analysis finds a somewhat different result, I think that the cloud-temperature degeneracy needs to be further explored and clarified.

Figure 3 – the data show consistently greater differences between the terminators than the GCMs predict. The discussion in lines 347-368 is focused on what seem like minor aspects of the data-model comparison, while ignoring the "elephant in the room" that the limb differences in the data are consistently greater than predicted by the models. For example near 4.3 microns there is consistently an enormous difference between the limbs that the models do not come even close to reproducing. I think the observed differences are suspiciously large, and the authors should reconsider that they are not necessarily observing true limb differences in scale height. The differences in spectral shape shown in Figure 4 may be a clue that could be investigated in more depth. For example, are other effects such as transit time (see below) or stellar limb darkening asymmetries a potential explanation for the apparent difference in the transit spectrum at morning and evening limbs?

As the authors note, there are other potential explanations for limb differences such as the transit time. When one limb has a greater scale height, then the opacity-weighted center of the planet is shifted toward that limb, and that will mimic a different transit time, and vice-versa. The authors explore this in Figure 5, by adding a 3-sigma offset to the time of transit. But that's not the same as allowing the central transit time to vary as a fit parameter when solving for the spectra of the planetary limbs, and I think that needs to be done for a convincing demonstration.

Also, panels c and d of Figure 5 show large differences between limbs for differences in eccentricity and limb darkening. Some additional explanation is needed as to how the authors can rule out stellar limb darkening being the source of the differences they derive. For example, do the authors know that the stellar limb darkening is perfectly symmetric, or does stellar rotation cause a subtle ingress vs. egress difference in the limb darkening that could affect the transit result? (That's not the same as the transit light source effect). Does overlap between the planet's spectrum and spectral lines of the star play a role, given the different stellar vs. planet Doppler shifts at ingress and egress?

Line 430, it isn't adequate to simply refer to a paper in preparation without describing at least in outline how the data are treated. Please explain the Carter and May paper in at least general terms.

Line 597, spell evening

Referee #2 (Remarks to the Author):

The manuscript reports on a potential wavelength-dependent asymmetry in the transit of exoplanet WASP-39b. This asymmetry is interpreted as a consequence of a temperature or cloud difference on the planet's 'evening' or 'morning' terminator. Such an asymmetry is expected from models but has not previously been directly observed.

Such a detection, if real, is of significant interest. However, based on the manuscript as it is currently written, I am unconvinced that the claimed detection is statistically significant. I have made a number of comments which mostly relate to investigating the statistical significance in a more rigorous manner. As the manuscript is currently presented, I am unconvinced that the authors have demonstrated that the effect they claim to observe is a real astrophysical effect, rather than an effect caused by modelling assumptions. Unless the result holds up under further tests such as those I propose below, I am unconvinced it warrants publication in its current form.

1. Figure 1 provides a clear and well-designed illustration of the methodology, for which the authors deserve credit. Only a single wavelength band is shown. How is the specific wavelength (4.38 micron) chosen? From Figure 2, it appears this point may be chosen as one of the outlier points where the difference between evening and morning terminator is largest.

2. Even when choosing this specific wavelength, by the author's own admission (caption of Fig 1), the significance of the effect is dificult to conclude from a single wavelength. It would be useful to show transits for each individual wavelength (or binned down to e.g. R = 30 as in Fig 2). While each transit has a different, wavelength-dependent, depth, if the effect is significant on a wavelength-average basis (claimed by the authors at the ~400 ppm level), it should be possible to simply add all transits and produce a wavelength-independent fit to demonstrate the (wavelength-averaged) significance. Even without that (over)simplification, it should be possible to fit each wavelength individually, but provide an average residual plot (Fig 1c and Fig 1d, but with residuals averaged over all wavelengths, to show the alleged significance of catwoman vs batman). This would help convince the reader that the difference is indeed driven by the differing models (primarily at ingress and egress) rather than by some other effect in either data or model.

3. When computing the significance of the wavelength-averaged difference, an error bar of 88 ppm is provided. How is this uncertainty calculated? It would be useful to ascertain to which degree this difference is driven by an offset at specific wavelengths. For example, how significant is the effect if the most prominent feature (at 4-4.5 micron) is taken out? More generally, the authors could perform a bootstrap approach (drawing random (sub)samples over the wavelength range, with replacement) to establish the robustness of the observed offset.

4. In the main manuscript, multiple references are made to the work of Carter & May, which is in preparation. Although poor practice, this is not particularly problematic when only used to support a claim that NIRSpec/PRISM observations have minimal instrumental systematics. However, the work in preparation also appears to describe the Eureka! code which is key to this work, and the motivation for using only data in the 2-5 micron range. Both the latter claim and the description of

the code should be explained within this manuscript, so the details can be scrutinised. In particular, data in the 0.7-2 micron range is discarded. How would the results change if these data were included? Are the results sensitive to the exact amount of data that are discarded?

5. A number of parameters are fixed throughout the fits (e.g., period, a/Rstar, b). There are multiple problems with this. Firstly, the values adopted come from a paper in prep. that can't be scrutinised (the authors are presumably aware of this problem and seeking to address this during the refereering process, an unhelpful practice). Secondly, the values adopted presumably have uncertainties, which aren't incorporated in the analysis here. Thirdly, it is unclear whether these model parameters have the potential to influence the conclusions of the analysis. The best approach would be to model these parameters simultaneously as part of the wavelength-dependent fit, or at least incorporate their uncertainties through the use of a prior. At the very least, the authors should investigate whether changes in these parameters could affect the catwoman fit posteriors.

6. In the methods, it is claimed that catwoman NE / catwoman MM / half-ingress/egress "seem to be consistent with each other within the error bars". From Fig 4, it's not obvious that this claim is true, and it should be quantified. In particular, the NE results seem to consistently find a lower morning depth than the two other methods, and it is therefore likely to show a more significant morning-evening difference. Even though the authors argue that the NE approach is more conservative due to the freedom in limb darkening parameters, what would the result look like if one of the other methods were to be adopted? The morning-evening difference for these other methodologies should be quantified. Similarly, the wavelength-dependent mid-transit time method is discussed and shown in Figure 4, but the results are not quantitatively interpreted. What do they show?

7. In the methods, the limb darkening robustness test is valuable but the results are concerning. These results highlight that a small issue with one of the other transit parameters (here, limb darkening) can cause a spurious effect. These results highlight the need to fully propogate uncertainties on other parameters (P, a/Rstar, b) as per one of my previous comments. They also raise the question that the limb darkening parameters, even when allowed to deviate in the NE approach, may still be incorrect (e.g. if the stellar parameters were wrong). What would happen if the prior was further broadened, e.g., to 0.2 instead of 0.1 on the limb darkening parameters? What if no informative prior was adopted at all? How do the limb darkening parameters compare in the (asymmetric) catwoman fit with those that are obtained when running a (symmetric) batman fit, and how does each of these compare to the theoretically predicted values?

8. The robustness against eccentricity tests are again valuable but the results concerning. A case is made that eccentricity-related biases can't cause the results, as they would work the opposite way as the observed effect. While this may be true for a range of eccentricities, is this true for a range of omega values? What if omega were offset by 180 degrees, e.g., 190 degrees rather than 10 degrees? It's hard to know whether such values would be allowable by a general transit fit, but the transit fit referenced is (once again) in a paper-in-preparation so this is impossible to verify (but given the low \$e\$, a range of \$omega\$ is likely allowable). Regardless of what is found in Carter & May, given the clear reason for concern highlighted in Fig 5d a range of omega and eccentricity should be tested and the full results should be shown visually. If changes in \$e\$ could plausibly influence the result, this suggests there is also reason to be concerned about changes in impact parameter \$b\$ (as \$b\$

changes ingress/egress shape, where e only scales the total transit duration).

9. It's not clear why the issue of planetary rotation is dismissed. From the analysis in the methods, it appears differences at the 234 ppm level can be expected, with an uncertainty of 144 ppm. At 2sigma, that puts the effect in the range of [0,522] ppm. Given that the reported effect of morning-to-evening is at the 405 ppm level (as reported in the abstract), the conclusion would seem to be that the planetary rotation effect is hard to quantify, not that it is insignificant.

Minor comments:

1. Some affiliations in the author list have incorrect use of superscript (e.g., Kiefer, Mancini).

2. In the abstract, the sentence "the evening spectrum is [..] ppm larger than the morning spectrum" is hard to parse. It's not immediately obvious what "a larger spectrum" means. Do the authors mean more flux is received across all wavelengths, on average?

3. In the main manuscript, reference is made to several works that use models to show the evening terminator is expected to be hotter than morning terminators. Though the models are complex and contain many ingredients, it may helpful to briefly summarise the expected physical reason(s) for this (eastward winds?).

4. In the main manuscript, the authors say that "catwoman is arguably the most conservative approach". Why?

Author Rebuttals to Initial Comments:

Response to the Nature Review, Round # 1

We would like to first thank the Editor and the reviewers for the insightful comments made on the manuscript. We believe the concerns raised by the referees were all fair and sound, and we have taken the time to thoroughly address them; we detail the changes in the responses below. They are implemented and highlighted in the manuscript text as **red, bolded sentences**.

In addition, we have made the requested editorial changes to the paper; in particular:

We now respond to each of the points made by the referees in order. We start with the comments of Referee # 1, which are quoted below in grey, followed by our response in blue:

Referee # 1: The authors may be aware that there is another similar paper on this same topic using data for the same planet, with a similar result in terms of temperature difference, but not as similar in cloud properties. Given that the total light blocked by the limb is sensitive to scale height (via temperature) and cloud content, and that it's well known that there can be a degeneracy between these variables, I think more clarification/discussion is needed concerning the trade-off between cloud height and temperature. For example, I would expect to see cloud height (or some other cloud property) and temperature plotted versus each other as a panel in corner plots for each limb, and those plots are missing from this paper. Given that an independent analysis finds a somewhat different result, I think that the cloud-temperature degeneracy needs to be further explored and clarified.

Answer: We thank the referee for this comment. We have added the requested corner plots in Figure 7. In addition, we have added some discussion of these possible degeneracies in the CHIMERA atmospheric retrievals discussion in the Methods section. In summary, we find no strong correlation between the retrieved cloud properties and the morning/evening temperatures, which we attribute to the fact that the constraints on the properties are coming both from the absolute morning/evening depth differences (which define mostly cloud properties) and the feature amplitudes of H_2O and CO_2 on each spectrum (which define both).

Referee # 1: Figure 3: the data show consistently greater differences between the terminators than the GCMs predict. The discussion in lines 347-368 is focused on what seem like minor aspects of the data-model comparison, while ignoring the “elephant in the room” that the limb differences in the data are consistently greater than predicted by the models. For example near 4.3 microns there is consistently an enormous difference between the limbs that the models do not come even close to reproducing. I think the observed differences are suspiciously large, and the authors should reconsider that they are not necessarily observing true limb differences in scale height. The differences in spectral shape shown in Figure 4 may be a clue that could be investigated in more depth. For example, are other effects such as transit time (see below) or stellar limb darkening asymmetries a potential explanation for the apparent difference in the transit spectrum at morning and evening limbs?

Answer: We thank the referee for this comment. With respect to the possibilities of the limb differences being driven by other transit parameters and/or assumptions, we assess those concerns below through a variety of tests that include different limb-darkening assumptions (including limb-darkening asymmetries), transit parameter priors exploration and others. In summary, we showcase how none of the other parameters can give rise to the level of limb differences we observe.

As for the GCM models themselves, Figure 3 was updated after some corrections to the procedure that post-processes GCM output to synthetic transmission spectra. While the overarching conclusion that none of the GCMs are able to match the observed difference between the evening and morning limb spectra is still true, the updated Figure 3 now shows that some GCMs provide a qualitative match to the evening or morning limb spectra separately. Specifically, the observed morning limb spectrum best matches the results from the GCM assuming condensate clouds, while the observed evening limb spectrum best matches the results from the GCM assuming an aerosol-free atmosphere at chemical equilibrium. Notably, the latter GCM also provides the best match for the observed 4.3 micron peak in the evening limb spectrum. The implication of this observations–model comparison is that, firstly, GCMs could disagree with observations for different reasons, and secondly, while none of the participating GCMs are currently able to capture the full picture, some of them are close to capturing a part of it. Obtaining such an insight is invaluable for a correct future direction of GCM development.

Referee # 1: As the authors note, there are other potential explanations for limb differences such as the transit time. When one limb has a greater scale height, then the opacity-weighted center of the planet is shifted toward that limb, and that will mimic a different transit time, and vice-versa. The authors explore this in Figure 5, by adding a 3-sigma offset to the time of transit. But that’s not the same as allowing the central transit time to vary as a fit parameter when solving for the spectra of the planetary limbs, and I think that needs to be done for a convincing demonstration.

Answer: Based on a comment made by Referee #2 (see comment number 5 for Referee #2 below), we performed the same limb-asymmetry fits described in the manuscript, but allowing for P, a/R*, b and t0 to vary as well, where priors for these parameters were set based on the white-light lightcurve fits described in Carter & May et al. (in review). We obtain the very same results as the original limb spectra presented in our work, albeit with slightly larger uncertainties. This translates on an average depth difference between the morning and the evening that is only 5% more uncertain than when those are held fixed — and thus our result of a > 4-sigma larger morning than evening still holds even in this case, implying our conclusions and results are unchanged by this.

Referee # 1: Also, panels c and d of Figure 5 show large differences between limbs for differences in eccentricity and limb darkening. Some additional explanation is needed as to how the authors can rule out stellar limb darkening being the source of the differences they derive. For example, do the authors know that the stellar limb darkening is perfectly symmetric, or does stellar rotation cause a subtle ingress vs. egress difference in the limb darkening that could affect the transit result? (That's not the same as the transit light source effect). Does overlap between the planet's spectrum and spectral lines of the star play a role, given the different stellar vs. planet Doppler shifts at ingress and egress?

Answer: As the referee suggests, we are indeed assuming symmetric limb-darkening as the stellar rotation rate is low and typical for an old, G-type star like WASP-39 (vsini = 1.5 +/- 0.6 km/s; Faedi+2011). Nonetheless, we decided to perform a test using the stellar and transit parameters for WASP-39b, along with obliquity measurements of the system in Mancini+2018 (which are consistent with 0), which we used as input for PyTransit — a transit modelling tool which can simulate gravity darkening. We found that any possible effects of asymmetric limb-darkening are well below what we measure in our morning/evening spectra, much smaller than 100 ppm differences when compared to symmetric lightcurves. With respect to the comment on overlap between planetary and stellar spectrum: at ingress/egress, the planet's velocity is +/- 12 km/s, which would require a spectral resolution of R = 25k to resolve. Given the resolution of our presented transmission spectra is R = 100, the planet's spectral features are unresolvable from the stellar lines and so differences in the velocities between the planetary and stellar lines cannot be responsible for differences in the ingress/egress spectra.

Referee # 1: Line 430, it isn't adequate to simply refer to a paper in preparation without describing at least in outline how the data are treated. Please explain the Carter and May paper in at least general terms.

Answer: The Carter and May paper is now in review and we have updated our references accordingly. The text has also been updated to more accurately reflect the initial data products used, and the light curve extraction performed in the Carter and May publication.

Referee # 1: Line 597, spell evening

Answer: We thank the referee for this catch — we have fixed this typo in the manuscript.

We now answer to the comments made by Referee # 2:

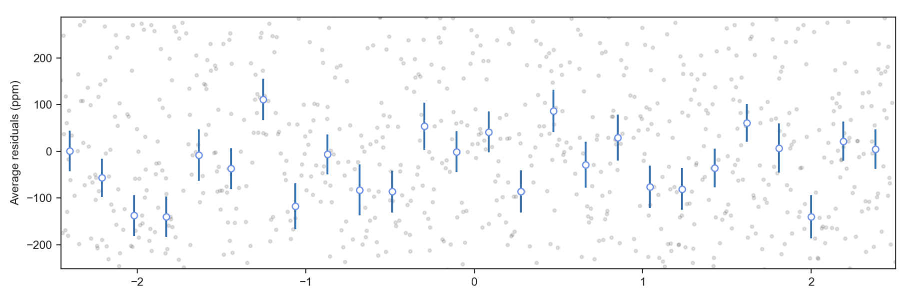
1. Figure 1 provides a clear and well-designed illustration of the methodology, for which the authors deserve credit. Only a single wavelength band is shown. How is the specific wavelength (4.38 micron) chosen? From Figure 2, it appears this point may be chosen as one of the outlier points where the difference between evening and morning terminator is largest.

Answer: This wavelength was chosen as it is right in the middle of the largest feature in the transmission spectrum: the CO2 feature. We have clarified this in the text caption of the figure.

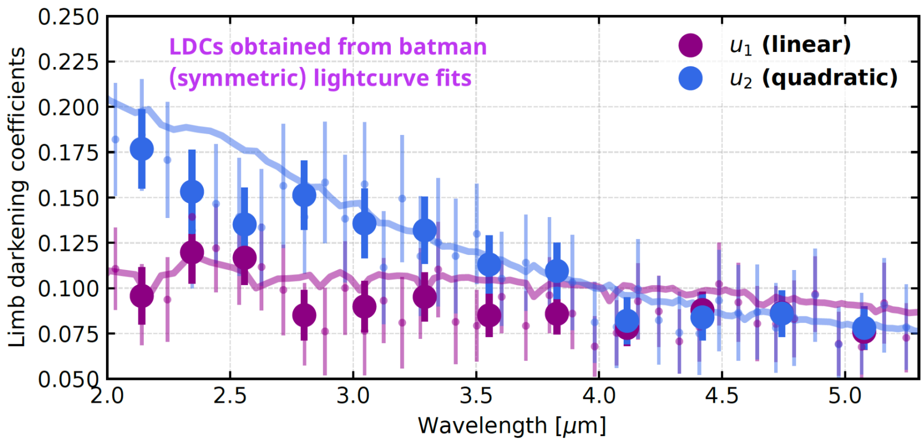
2. Even when choosing this specific wavelength, by the author's own admission (caption of Fig 1), the significance of the effect is dificult to conclude from a single wavelength. It would be useful to show transits for each individual wavelength (or binned down to e.g. R = 30 as in Fig 2). While each transit has a different, wavelength-dependent, depth, if the effect is significant on a wavelength-average basis (claimed by the authors at the ~400 ppm level), it should be possible to simply add all transits and produce a wavelength-independent fit to demonstrate the (wavelength-averaged) significance. Even without that (over)simplification, it should be possible to fit each wavelength individually, but provide an average residual plot (Fig 1c and Fig 1d, but with residuals averaged over all wavelengths, to show the alleged significance of catwoman vs batman). This would help convince the reader that the difference is indeed driven by the differing models (primarily at ingress and egress) rather than by some other effect in either data or model.

Answer: The results for each individual wavelength showcase behaviors similar to what is observed in the transmission spectrum at individual wavelengths, namely, that the morning/evening deviation is relatively smaller for wavelengths outside the CO2 feature (the next response touches on this point in more detail). As such, we believe showcasing other wavelengths doesn't add much more information than the illustration already provided in Figure 1.

As for the second set of comments, we don't agree with the referee that the results should be seen in the averaged-residual plots directly. The main reason for this is given in our experiments in Figure 5 — i.e., that a small shift in limb-darkening coefficients can remove the effect entirely in a batman (i.e., symmetric) lightcurve fit. This underscores why it is so difficult to see the effect directly in the lightcurves. Performing the last of the suggestions the referee mentioned (i.e., average the residuals in a batman, i.e., symmetric lightcurve fit), we indeed don't see any particular signal in the residuals:



And, as expected, we see limb-darkening coefficients which are shifted by about ~0.02 with respect to the ones obtained with the asymmetric fits., As per our experiments in Figure 5, this can easily account for a ~400 ppm offset between the morning and the evening (compare against Figure 6 in the manuscript):



We again note that our experiments in Figure 5 are in fact driven by the desire to convince the reader that if no effect was present, we shouldn't be able to detect anything on our asymmetric lightcurve fits. But we do. Based on the battery of tests below, we believe the effect is indeed real.

3. When computing the significance of the wavelength-averaged difference, an error bar of 88 ppm is provided. How is this uncertainty calculated? It would be useful to ascertain to which degree this difference is driven by an offset at specific wavelengths. For example, how significant is the effect if the most prominent feature (at 4-4.5 micron) is taken out? More generally, the authors could perform a bootstrap approach (drawing random (sub)samples over the wavelength range, with replacement) to establish the robustness of the observed offset.

Answer: In our analysis, the difference at each wavelength is calculated first by subtracting morning depth posterior samples from evening depth posterior samples at each individual wavelength, which in turn gives a posterior distribution for the evening-to-morning difference, X_i, at each wavelength *i* between the evening and the morning; this process accounts for the covariance between the morning and evening samples (bottom panel of Figure 1, left). Using those posterior distributions for the differences, X_i we then obtain the posterior distribution of the ***mean difference*** across all wavelengths by simply sampling from the posterior distributions of these differences and computing the mean — i.e., we obtain the derived posterior distribution

$$\overline{X} = \sum_{i=1}^{N} X_i / N,$$

where N is the number of wavelength bins. The wavelength-averaged difference we report then is the median of the posterior distribution of $\overline{X}$, and the errors are the 68% quantiles of this posterior distribution (which is 88 ppm).

Repeating the process by masking the 4-4.5 micron region, we still see an evening-to-morning depth difference of 282 +/- 91 ppm, which is > 3-sigma away from zero. This has a median slightly smaller —although consistent at ~1-sigma with— the 405 +/- 88 ppm difference we report as the wavelength-averaged difference. Performing a bootstrap by sampling 80% of the wavelength range at a time, and following the same process as above, we unsurprisingly obtain a wavelength-averaged difference of 384 +/- 90 ppm — which is consistent with our reported value. This, once again, showcases that the differences are robust to selections of the wavelength range over which they are taken.

4. In the main manuscript, multiple references are made to the work of Carter & May, which is in preparation. Although poor practice, this is not particularly problematic when only used to support a claim that NIRSpec/PRISM observations have minimal instrumental systematics. However, the work in preparation also appears to describe the Eureka! code which is key to this work, and the motivation for using only data in the 2-5 micron range. Both the latter claim and the description of the code should be explained within this manuscript, so the details can be scrutinised. In particular, data in the 0.7-2 micron range is discarded. How would the results change if these data were included? Are the results sensitive to the exact amount of data that are discarded?

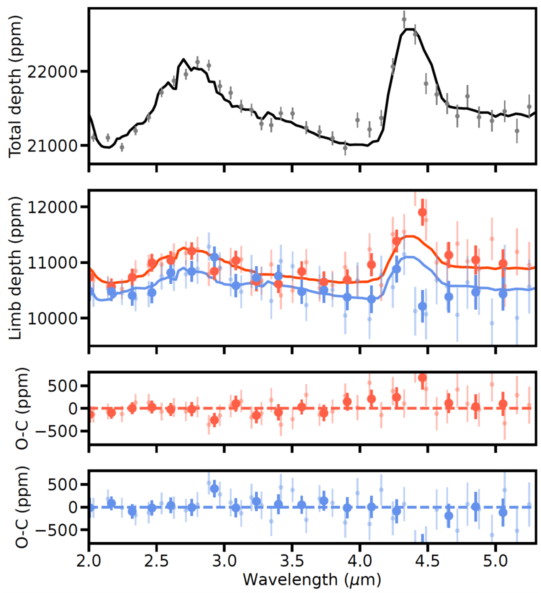
Answer: The Carter and May paper is now in review and we have updated our references accordingly. We have provided further clarity on the motivation behind discarding the data <2 micron in the text. The effect on our results through the inclusion of the short-wavelength data provides little meaning, as the data are fundamentally corrupted by the effects of detector saturation. This paper, however, doesn't describe the Eureka! code, which is thoroughly documented on the Github page we cite in our manuscript as well as in Bell et al. (2022; https://arxiv.org/abs/2207.03585), and papers in the literature such as Rustamkulov et al. (2023; https://ui.adsabs.harvard.edu/abs/2023Natur.614..659R/abstract) for NIRSpec/PRISM specifically.

5. A number of parameters are fixed throughout the fits (e.g., period, a/Rstar, b). There are multiple problems with this. Firstly, the values adopted come from a paper in prep. that can't be scrutinised (the authors are presumably aware of this problem and seeking to address this during the refereeing process, an unhelpful practice). Secondly, the values adopted presumably have uncertainties, which aren't incorporated in the analysis here. Thirdly, it is unclear whether these model parameters have the potential to influence the conclusions of the analysis. The best approach would be to model these parameters simultaneously as part of the wavelength-dependent fit, or at least incorporate their uncertainties through the use of a prior. At the very least, the authors should investigate whether changes in these parameters could affect the catwoman fit posteriors.

Answer: We thank the referee for this comment — we have run the suggested experiment, on which at each wavelength we now leave the period, a/R*, b and t0 as free parameters, with priors set by the white-light lightcurve fits obtained by Carter & May et al. (in review). In particular we:

- Set a normal prior for P with (mu, sigma) = (4.0552842, 3.6e-06) days.
- Set a normal prior for a/R* with (mu, sigma) = (11.39, 0.012)
- Set a normal prior for b with (mu, sigma) = (0.4498, 0.0022)
- Set a normal prior for t0 with (mu, sigma) = (2459771.335647, 1.4e-05) days.

As described in the manuscript, these free parameters go on top of the relative flux offset, jitter term, slope, morning/evening limbs, and limb-darkening coefficient priors for each wavelength.

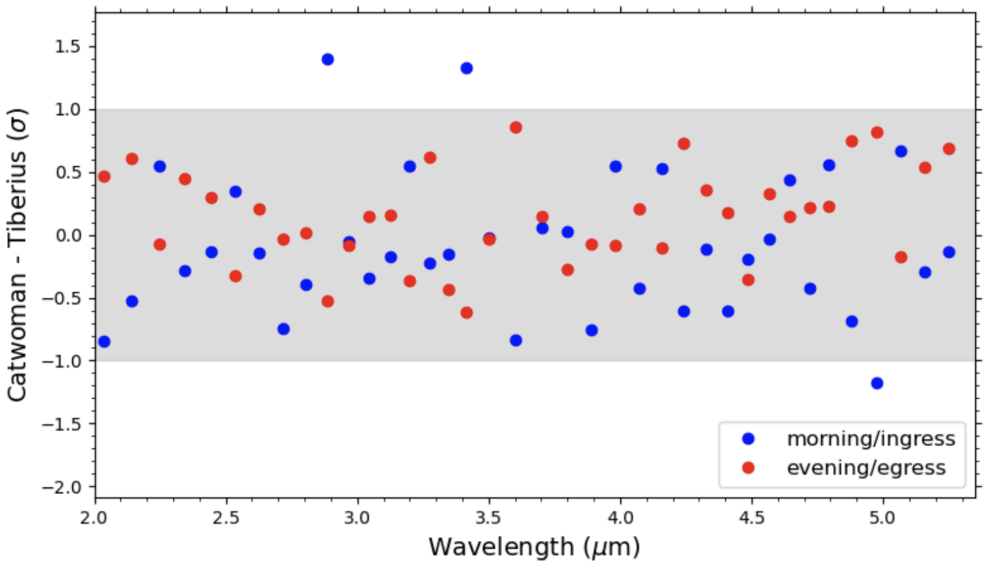


The figure to the left shows the results of this fit, in the very same format as Figure 5, as points with errorbars. The models shown here are the very same models plotted in Figure 5, for ease of comparison.

From these experiments, we conclude the results are virtually the same as leaving the above parameters fixed, which is in turn consistent with our experiments performed on, e.g., the impact of t0 given its precise knowledge/prior for WASP-39b. In this case, the average depth difference between the morning and evening is 383 +/- 94 ppm — which is slightly more uncertain than the result obtained when fixing the parameters, but still consistent with a > 4-sigma morning/evening depth difference.

6. In the methods, it is claimed that catwoman NE / catwoman MM / half-ingress/egress "seem to be consistent with each other within the error bars". From Fig 4, it's not obvious that this claim is true, and it should be quantified. In particular, the NE results seem to consistently find a lower morning depth than the two other methods, and it is therefore likely to show a more significant morning-evening difference. Even though the authors argue that the NE approach is more conservative due to the freedom in limb darkening parameters, what would the result look like if one of the other methods were to be adopted? The morning-evening difference for these other methodologies should be quantified. Similarly, the wavelength-dependent mid-transit time method is discussed and shown in Figure 4, but the results are not quantitatively interpreted. What do they show?

Answer: We have computed the differences between the morning and evening spectra in units of sigma between the catwoman NE and the half-ingress/egress (i.e., with uncertainties coming from both of those methodologies), which we show below:



As described in the text, the results are indeed consistent at 1-sigma (the catwoman MM versus half-ingress/egress shows a very similar result).

We have added some text qualitatively discussing the mid-transit time method results. In essence, they seem to show an offset of a few tens of seconds in particular around where the H_2O and CO_2 features were expected, albeit with relatively large uncertainties. These are also the regions that visually also show the largest morning-to-evening differences, in particular in the fixed limb-darkening methodologies (which the timing methodology also assumed).

7. In the methods, the limb darkening robustness test is valuable but the results are concerning. These results highlight that a small issue with one of the other transit parameters (here, limb darkening) can cause a spurious effect. These results highlight the need to fully propogate uncertainties on other parameters (P, a/Rstar, b) as per one of my previous comments. They also raise the question that the limb darkening parameters, even when allowed to deviate in the NE approach, may still be incorrect (e.g. if the stellar parameters were wrong). What would happen if the prior was further broadened, e.g., to 0.2 instead of 0.1 on the limb darkening parameters? What if no informative prior was adopted at all? How do the limb darkening parameters compare in the (asymmetric) catwoman fit with those that are obtained when running a (symmetric) batman fit, and how does each of these compare to the theoretically predicted values?

Answer: We thank the referee for this comment. We indeed performed experiments further relaxing the limb-darkening assumptions. If we assume uniform priors on q1 and q2 between 0 and 1, we obtain a wavelength-averaged depth difference of 445 +/- 107 ppm. This is still a 4-sigma detection, with a value that is consistent with the limb-darkening prior we decided as the best to use in the manuscript. As for the question of the limb-darkening coefficients on the symmetric fit, we have already answered that question in our response to comment number 2.

8. The robustness against eccentricity tests are again valuable but the results concerning. A case is made that eccentricity-related biases can't cause the results, as they would work the opposite way as the observed effect. While this may be true for a range of eccentricities, is this true for a range of omega values? What if omega were offset by 180 degrees, e.g., 190 degrees rather than 10 degrees? It's hard to know whether such values would be allowable by a general transit fit, but the transit fit referenced is (once again) in a paper-in-preparation so this is impossible to verify (but given the low e, a range of $omega$ is likely allowable). Regardless of what is found in Carter & May, given the clear reason for concern highlighted in Fig 5d a range of omega and eccentricity should be tested and the full results should be shown visually. If changes in e could plausibly influence the result, this suggests there is also reason to be concerned about changes in impact parameter b (as b changes ingress/egress shape, where e only scales the total transit duration).

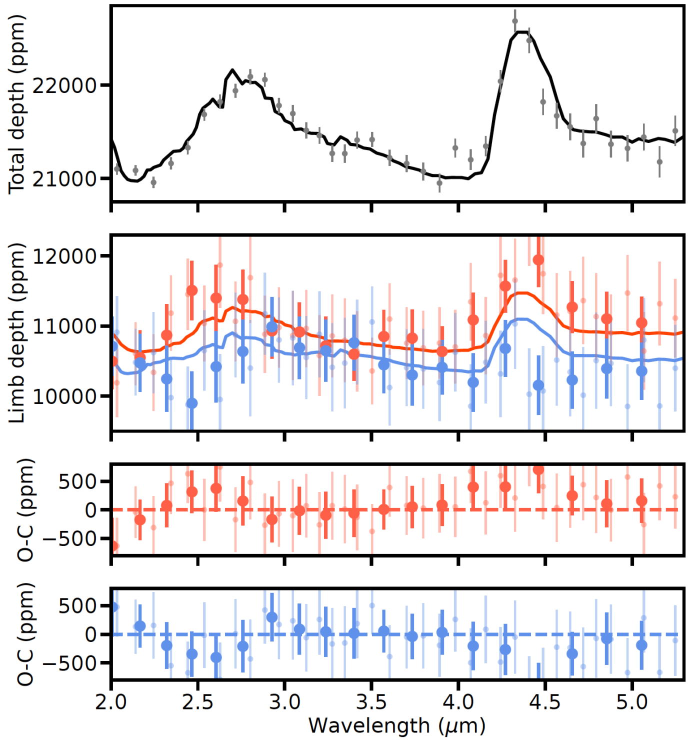
Answer: We thank the referee for this comment. In order to fully account for it with all the information available, we have decided to perform two extra analyses which we believe very thoroughly assess the referee's concern on this topic, namely:

1. We have performed a new joint fit to all the WASP-39b data accounting for all known information at the time to obtain the best constraints possible on the eccentricity and argument of periastron. This includes both the e cos(omega) constraint from WASP-39b's secondary eclipse mentioned in the manuscript, as well as a constraint in the stellar density we obtain using the prescription in Brahm et al. (2018, 2019).

2. Using those constraints, we have then fitted the spectroscopic lightcurves with limb models using all those constraints as priors instead of leaving them fixed in the analysis — including eccentricity and argument of periastron (i.e., similar to the experiment made above for a circular model, but here for an eccentric model).

The new joint fit to all the WASP-39b data is virtually the same as the one performed in Carter & May, but now include the constraints on ecos(omega) obtained by the eclipse measurements mentioned in the text, i.e., the prior on ecos(omega) = 0.0007 +/- 0.0017. This gives back a slightly better constraint on the eccentricity and omega, of ecos(omega) = 0.00087 +/- 0.0014 and esin(omega) = 0.0014 +/- 0.006. We note that even with this prior, a circular model for the joint fit is preferred based on the bayesian evidence of the fit.

We then go ahead and perform the same experiments we performed in our response to comment number 5 above, but also leaving ecos(omega) and esin(omega) as priors in the wavelength-dependent asymmetric lightcurve fits.



The figure on the left shows the results from this new lightcurve fit. As can be observed, this gives rise to more uncertain evening and morning spectra, which in turn actually leads to, again, larger evenings than mornings — on average, a difference of 582 +/- 188 ppm. This is a 3-sigma detection of evenings being larger than mornings in this case — a detection consistent with the detection made when assuming a circular model (shown above). Our experiment showcases that, while the eccentricity uncertainty indeed impacts on the quality of the spectra, the detection still stands even when accounting for this uncertainty. We have added text describing this experiment and its findings in the manuscript, on the "Transit light curve analysis" sub-section of the "Methods" section.

9. It's not clear why the issue of planetary rotation is dismissed. From the analysis in the methods, it appears differences at the 234 ppm level can be expected, with an uncertainty of 144 ppm. At 2sigma, that puts the effect in the range of [0,522] ppm. Given that the reported effect of morning-to-evening is at the 405 ppm level (as reported in the abstract), the conclusion would seem to be that the planetary rotation effect is hard to quantify, not that it is insignificant.

Answer: We agree with the referee. We have changed the text accordingly to further this point in the manuscript.

Minor comments:

1. Some affiliations in the author list have incorrect use of superscript (e.g., Kiefer, Mancini).

Answer: We thank the referee for this catch. We have tried hard to correct this, but this seems to be a problem with the template. We will deal with this at the editorial stage.

2. In the abstract, the sentence "the evening spectrum is [..] ppm larger than the morning spectrum" is hard to parse. It's not immediately obvious what "a larger spectrum" means. Do the authors mean more flux is received across all wavelengths, on average?

Answer: We have changed the wording from "spectrum" to "transit depths".

3. In the main manuscript, reference is made to several works that use models to show the evening terminator is expected to be hotter than morning terminators. Though the models are complex and contain many ingredients, it may helpful to briefly summarise the expected physical reason(s) for this (eastward winds?).

Answer: We believe the final sentences of the first paragraph in the introduction already cover these in detail through the references cited in our work.

4. In the main manuscript, the authors say that "catwoman is arguably the most conservative approach". Why?

Answer: This method is the only one that directly includes the morning and evening terminators in a single fit to the data. We added wording of this in the text.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have done a lot of work to revise and improve this paper, and I'm happy with their revisions, except in one (minor) respect. Their reply on the possible overlap between stellar and planetary lines shows that they mid-understood my initial comment (I should have been clearer).

The low resolution used in this paper is exactly the point: it's a potential liability when considering line overlap. Overlap between real stellar and planetary lines occurs at high resolution. The star has carbon monoxide lines whose wavelength will vary from ingress to egress due to stellar rotation. Carbon dioxide lines in the planet could accidently overlap stellar CO lines (or other stellar lines), depending on the exact wavelengths. Because the stellar lines could be narrow and deep, the star is darker in the stellar line cores, and the transit absorption signal could be reduced by the overlap. That effect would be diluted and not visible at low resolution. And the effect could be asymmetric because of stellar rotational Doppler shifts. I think it's necessary to check this using a high resolution stellar model spectrum (e.g., PHOENIX). If the effect is negligible, I don't think it's necessary to add much discussion to the paper, perhaps just a brief comment that potential asymmetric line overlap was checked and found to not be a problem (hopefully, it isn't).

Referee #2 (Remarks to the Author):

I thank the authors for their detailed responses to my previous points. While I appreciate the authors making a clear effort to address some of my concerns, I feel that several important issues remain outstanding. As such, I remain unconvinced that the current manuscript has demonstrated that the observed transit asymmetry is a result of the properties of the planetary atmosphere as claimed. I believe that some other scenarios, relating to the transit fit, the use of tight priors during the modelling, or the intrinsic uncertainty on stellar limb darkening, remain at least as plausible.

I have written my remaining concerns below, in three specific, related comments:

1. I previously remarked (my previous comment 2) that the authors should consider showing Fig 1 (and its residuals) for each wavelength individually, or for a range of wavelength bins. In response, the authors claim that such a plot is not useful, because "a small shift in limb-darkening coefficients can remove the effect entirely in a batman, symmetric, light curve fit". The authors also provided (in the response only) a wavelength-averaged residual plot which does not show an asymmetry.

While I appreciate the authors' transparency, I find this plot to be cause for serious concern about their main result. In the manuscript, Fig 3 shows an effect in the residuals for a specific wavelength bin (4.38 microns). While this wavelength shows the center of the CO2 feature and therefore the strongest effect, one would generally expect an asymmetry to show up at all wavelengths. If that is true, then a wavelength-averaged residual plot should show a *stronger* asymmetry than a residual plot at a single wavelength -- instead, the plot the authors provided seems to no longer show any asymmetry. I do not understand the authors' expalnation as to why the effect, if real, should not at all show up in this wavelength-averaged residual plot.

The authors attribute this lack of an effect to "limb darkening coefficients which are shifted by about 0.02 with respect to the ones obtained with the asymmetric fits", which "can easily account for a 400 ppm offset between the morning and evening". Comparing the figure the authors provided in their response, with their Fig 6 in the manuscript, a ~0.02 shift in limb darkening parameters can indeed be seen. The implied conclusion from the authors seems to be that this is therefore expected, and that a real asymmetry in the planet atmosphere can be masked by incorrect treatment of limb darkening. They don't discuss what seems to be the more obvious conclusion, which is that a real limb darkening issue can instead appear as an apparent asymmetry. It seems to be completely plausible that limb darkening parameters are incorrectly measured by such a small offset. If a batman fit, which assumes a symmetric planet model, can easily account for the observed ~400 ppm offset, by simply changing the limb darkening values by a small fraction, isn't the most likely explanation that any observation the authors have made is simply related to limb darkening, rather than to potential planet atmosphere asymmetries?

I note that the other reviewer has raised a similar concern relating to limb darkening, which I also don't find has been convincingly addressed. The authors have not shown to me why what they have observed is not an effect caused by the treatment of stellar limb darkening. After all, limb darkening treatment is inevitably highly simplified in these models, which use a parametric description of a complex stellar phenomenon.

2. I previously remarked that parameters such as period, a/Rstar, and impact parameter (b) were fixed, when they may in fact correlate with the observed asymmetry or even cause a spurious result when these parameters are fixed to incorrect values. In response, the authors have run a number of fits allowing for several of these parameters to vary, within certain normal priors. The authors then show that their results are largely unaffected when parameters are varied within these prior bounds, relative to when these parameters are fixed.

While I applaud the authors for having run this experiment, this only addresses part of my concern. The experiment the authors have run shows that the claimed uncertainty (from Carter & May) on these parameters isn't significant enough to affect the results. It doesn't, however, show whether a slight bias in any of these parameters could produce the observed effect.

As an example, the authors use a prior on impact parameter of 0.4498 +- 0.0022. That is a very tight prior. While impact parameters can sometimes be precisely measured for giant planets, there are various biases that could affect it (e.g., a small amount of stellar dilution, a slight transit distortion due to finite integration time, star spots, a low level of unaccounted for transit timing variations, etc.). A much more convincing experiment would be the use of a much broader prior, such as b =

0.45 +- 0.05.

Similarly, the authors used a normal distribution with width of ~1 second (1.4e-5 days) when varying t0. Again, while t0 can generally be precisely measured, one can easily imagine an offset or bias in its determination (especially given that the work by Carter & May upon which these parameters are based remains under review). The precision on a/Rstar is contingent on the choices made for eccentricity, which could influence this parameter (even for very low eccentricity values, as are expected for a short period planet).

It would be much more convincing to use much broader priors in these tests, and show at which level parameters such as P / a/Rstar / b / t0 would have to be uncertain or incorrect before they impact the main result.

It is a good sign that the fit, with tight priors, produces similar results than with fixed values, but it remains entirely possible a small issue with any of these parameters could cause the entire observed asymmetry. I note that the other reviewer also asked for a fit that allowed the central transit time to vary -- I don't find that varying them with tight priors is a convincing response to this concern. Even if the authors are confident each of these parameters are indeed known very precisely, it would be useful to know to which degree the claimed main result is contingent on these prior parameters.

3. Related to the previous point, I previously remarked that in addition to uncertainties (or offsets) in a/Rstar, period, or impact parameter, uncertainties or offets in eccentricity could possibly cause the observed effect. In response, the authors included priors on e and $omega$ in their fit to propagate uncertainty.

This has not addressed my main question, which I therefore repeat here: "Regardless of what is found in Carter & May, given the clear reason for
concern highlighted in Fig 5d a range of omega and eccentricity should be tested and the full results should be shown visually."

Rather than addressing this concern directly, the authors included a fit which allowed for a range of eccentricity, albeit with tight priors: ecos(omega) = 0.00087 +- 0.0014 and esin(omega) = 0.0014 +- 0.006. Given that, even within these very tight priors, the result has already lost some of its significancy (though still at ~3 sigma), it seems entirely plausible that a slightly broader range of allowable eccentricities could have a major impact on the claimed result. It would be easy to test significantly broader eccentricity priors (corresponding to, e.g., e in [0,0.05]).

If the authors can address the above concerns, then I believe their result is of high significance. If not, the authors should at least provide a significantly more careful discussion reflecting that their interpretation is only one of several possible options.

Author Rebuttals to First Revision:

Response to the Nature Review, Round # 2

We would like to once again thank the referees for their input on this second round of comments. We believe most of the concerns raised by the referees were fair and sound, and we have taken the time to thoroughly address them; we detail the changes in the responses below. They are implemented and highlighted in the manuscript text as **red, bolded sentences**. As we discuss below, we believe this and our previous response has thoroughly addressed how our result stands with the current best orbital parameter constraints in the system, and thus consider that our results are robust in light of those.

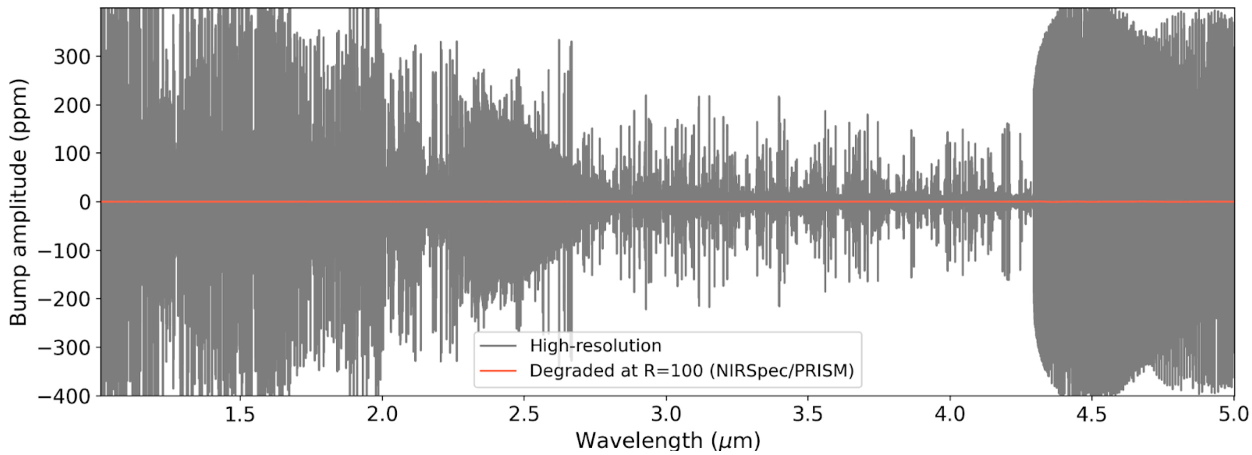
We respond to each of the points made by the referees in order. We start with the comments of Referee # 1, which are quoted below in grey, followed by our response in blue:

Referee # 1: The authors have done a lot of work to revise and improve this paper, and I'm happy with their revisions, except in one (minor) respect. Their reply on the possible overlap between stellar and planetary lines shows that they mid-understood my initial comment (I should have been clearer).

The low resolution used in this paper is exactly the point: it's a potential liability when considering line overlap. Overlap between real stellar and planetary lines occurs at high resolution. The star has carbon monoxide lines whose wavelength will vary from ingress to egress due to stellar rotation. Carbon dioxide lines in the planet could accidently overlap stellar CO lines (or other stellar lines), depending on the exact wavelengths. Because the stellar lines could be narrow and deep, the star is darker in the stellar line cores, and the transit absorption signal could be reduced by the overlap. That effect would be diluted and not visible at low resolution. And the effect could be asymmetric because of stellar rotational Doppler shifts. I think it's necessary to check this using a high resolution stellar model spectrum (e.g., PHOENIX). If the effect is negligible, I don't think it's necessary to add much discussion to the paper, perhaps just a brief comment that potential asymmetric line overlap was checked and found to not be a problem (hopefully, it isn't).

Answer: We thank the referee for the clarification on the exact questions about this line overlap point. Following the referee's suggestion, using PHOENIX models for a star like WASP-39b (5500 K, log(g) of 4.5 and zero metallicity; obtained from here), we decided to perform a calculation in the worst-case scenario: one in which we assume a homogeneous star on which half of the stellar surface is blueshifted, and the other half is redshifted due to the stellar rotational velocity. Given that (as discussed in our previous report) for WASP-39b v*sini* is 1.5 km/s with an error of 0.6 km/s, we used a worst-case scenario in which WASP-39b rotates at a value 5-sigma away from 1.5 km/s, i.e., at a velocity of 4.5 km/s.

With this, we generated a box-shaped transit light curve, on which the planet first transits the blue-shifted side of the star, and then the red-shifted side of the star, and we calculate the amplitude of the “bump” in this box-shaped lightcurve by the transit of the planet from the blue to the red-shifted side of the star, considering this lightcurve is normalized by the sum of both halves. We note this is a worst-case scenario effect because the effect is likely smaller than this as the planet would transit regions of continuously lower velocity as it approaches the center of the star during transit. Our results show that, indeed, the dilution effect mentioned by the referee exists — at high-resolution (R ~ 100,000; grey, below) this bump is of a couple hundreds parts-per-million (ppm) around the CO lines, i.e., above about 4.3 um — and smaller, of order ~100 ppm between 2.8 to 4.3 um. However, when we degrade the spectra at the resolution of NIRSpec/PRISM (about R = 100, red below), these lightcurve bumps, while still existing, fall far beyond the detectability of JWST time-series precision — below 1 ppm:



We thus conclude that, while this effect in theory exists even at low resolution spectra such as the one observed for NIRSpec/PRISM (as predicted by the referee), this effect is negligible in our case. Following the suggestion by the referee, we briefly mention this experiment in the Appendix.

We now answer the comments made by Referee # 2:

1. I previously remarked (my previous comment 2) that the authors should consider showing Fig 1 (and its residuals) for each wavelength individually, or for a range of wavelength bins. In response, the authors claim that such a plot is not useful, because "a small shift in limb-darkening coefficients can remove the effect entirely in a batman, symmetric, light curve fit". The authors also provided (in the response only) a wavelength-averaged residual plot which does not show an asymmetry.

While I appreciate the authors' transparency, I find this plot to be cause for serious concern about their main result. In the manuscript, Fig 3 shows an effect in the residuals for a specific wavelength bin (4.38 microns). While this wavelength shows the center of the CO2 feature and therefore the strongest effect, one would generally expect an asymmetry to show up at all wavelengths. If that is true, then a wavelength-averaged residual plot should show a *stronger* asymmetry than a residual plot at a single wavelength -- instead, the plot the authors provided seems to no longer show any asymmetry. I do not understand the authors' expalnation as to why the effect, if real, should not at all show up in this wavelength-averaged residual plot.

The authors attribute this lack of an effect to "limb darkening coefficients which are shifted by about 0.02 with respect to the ones obtained with the asymmetric fits", which "can easily account for a 400 ppm offset between the morning and evening". Comparing the figure the authors provided in their response, with their Fig 6 in the manuscript, a ~0.02 shift in limb darkening parameters can indeed be seen. The implied conclusion from the authors seems to be that this is therefore expected, and that a real asymmetry in the planet atmosphere can be masked by incorrect treatment of limb darkening. They don't discuss what seems to be the more obvious conclusion, which is that a real limb darkening issue can instead appear as an apparent asymmetry. It seems to be completely plausible that limb darkening parameters are incorrectly measured by such a small offset. If a batman fit, which assumes a symmetric planet model, can easily account for the observed ~400 ppm offset, by simply changing the limb darkening values by a small fraction, isn't the most likely explanation that any observation the authors have made is simply related to limb darkening, rather than to potential planet atmosphere asymmetries?

I note that the other reviewer has raised a similar concern relating to limb darkening, which I also don't find has been convincingly addressed. The authors have not shown to me why what they have observed is not an effect caused by the treatment of stellar limb darkening. After all, limb darkening treatment is inevitably highly simplified in these models, which use a parametric description of a complex stellar phenomenon.

Answer: We thank the referee for this feedback. In response, we would like to note that we have at least partially addressed this concern with our experiments shown in Figure 5. In particular, that experiment showcases that if the limb-darkening is held fixed in the fit, it can produce spurious limb-asymmetries. However, as we mention in the text, doing the same experiment with the limb-darkening coefficients being fitted does ***not*** produce spurious limb-asymmetries. That is: even in the case in which we are ignorant about the limb-darkening coefficients, our catwoman parametrization does not force limb-asymmetries to account for the offset on the limb-darkening coefficients — instead, the limb-darkening coefficients being fit as free parameters arrive at the correct (offset) values, and we retrieve the null result for limb-asymmetries we ingested. Our data experiment as showcased in our response to the referee aims to showcase the inverse argument: that if there *are* limb-asymmetries, then an (incorrect) symmetric transit model with limb-darkening coefficients as free parameters *can* fit the data. **If there weren't planetary limb-asymmetries, we wouldn't be able to see differences between morning and evening depths in a catwoman fit; but we *do* see them**.

We can summarize our experiments as follows:

If there were limb-asymmetries, the following statements have to be true:

1. One can model the lightcurve with a symmetric (batman) model and free limb-darkening coefficients. We indeed observe this in our JWST data.

2. One can model the lightcurve with an asymmetric (catwoman) model and free limb-darkening coefficients, recovering the limb-asymmetries. We indeed observe this in our JWST data.

If there were ***NOT*** limb-asymmetries, the following statements have to be true:

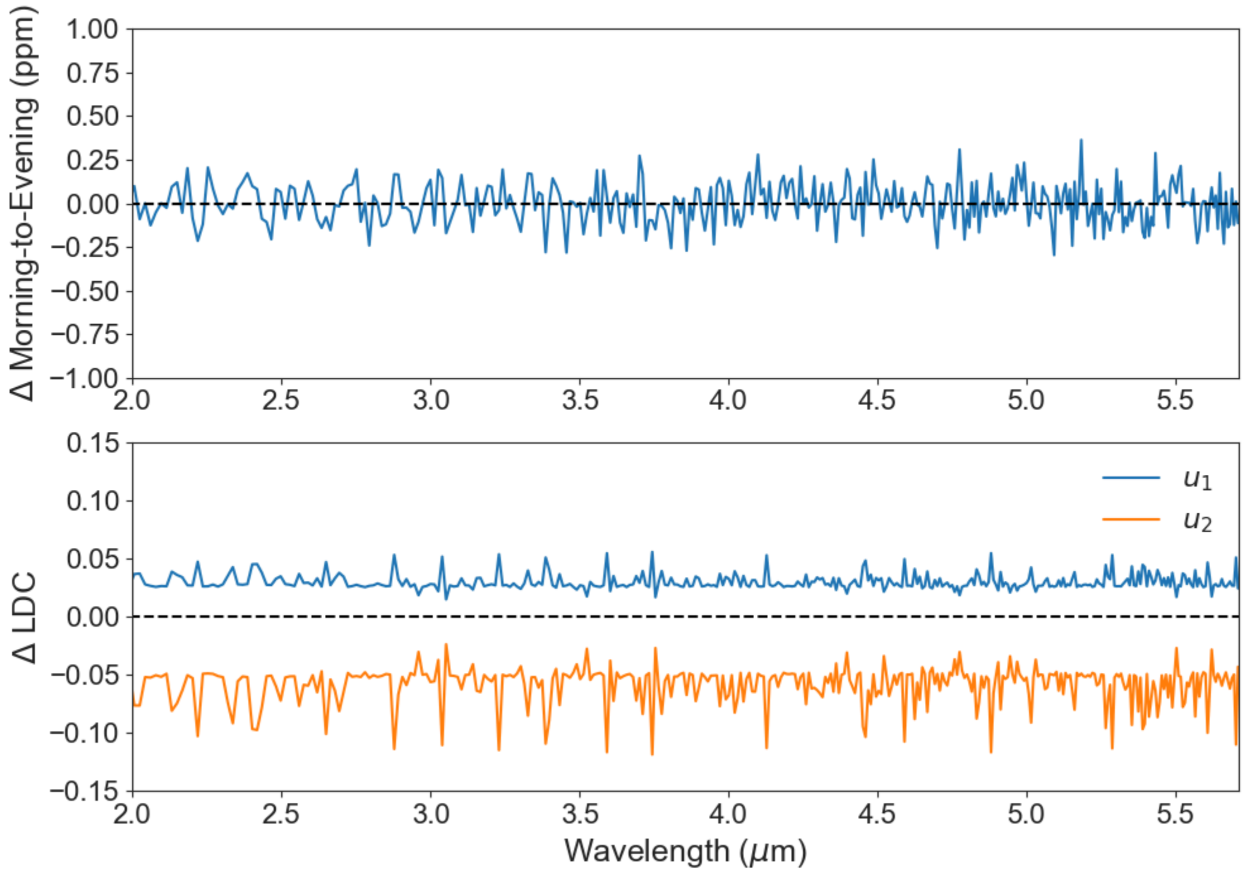
1. One can model the lightcurve with a symmetric (batman) model and free limb-darkening coefficients. We indeed observe this in our JWST data (this is the same result as 1. above).

2. One can model the lightcurve with an asymmetric (catwoman) model and free limb-darkening coefficients, and should recover no limb-asymmetries. We do ***NOT*** observe this in our JWST data.

We note, in addition, that in our injection and recovery tests using our methodology on simulated data, we do recover no limb-asymmetries when we simulate data with no limb-asymmetries. Because of this, we conclude that the limb-asymmetries model is our best explanation for the existing data, as even accounting for our ignorance on limb-darkening we are able to observe limb asymmetries.

We also note that, while as the referee correctly points out the limb-darkening effect is a complex one that is *approximated* via (two-parameter in our case) “laws”, it is also true that its wavelength-dependance (albeit perhaps not its exact form) is a well studied phenomenon. In particular, it has been showcased both in practice by other JWST studies in the literature *and* by state-of-the-art modelling that the limb-darkening effect decreases in importance (and complexity) at longer wavelengths. In fact, a very recent study on the same exoplanet host star we are studying in our work showcases that even accounting for magnetic fields the expected impact on transit lightcurves is much smaller than that what we are detecting in our catwoman lightcurve fits (Kostogryz et al., 2024; less than about 100 ppm above 2 um). Our observed planetary limb asymmetries, however, appear to be larger at larger wavelengths (in particular around the CO_2 feature), where the limb-darkening effect is the least important.

We would like to further note that limb-darkening being a symmetric effect, a mis-parametrization of the effect (e.g., using the wrong limb-darkening law) is very unlikely to give rise to limb-asymmetries. To showcase this and alleviate the referee’s concerns in this direction, we decided to run an experiment on which we studied whether a mis-parametrization of the effect (e.g., using a quadratic law when the real law generating the lightcurve was a more complex one such as, e.g., the non-linear law) can give rise to limb-asymmetries. To this end, we generated noiseless batman (i.e., symmetric) transit lightcurves modelled by a non-linear law with non-linear coefficients obtained through the limb-darkening library used in our paper (https://github.com/nespinoza/limb-darkening) for the NIRSpec/PRISM wavelength range used in our work, i.e., 2-5 um. Then, we fitted these lightcurves with a catwoman model using the same setup as that used in our work, i.e., using a quadratic law via the parametrization of Kipping+2013, and using uniform priors for the q1/q2 parameters between 0 and 1. With this experiment, we find that we indeed recover null limb asymmetries, that is, the morning-to-evening depths are consistent with zero (at better than a ppm in this noiseless example):



Adding noise similar to the one observed in our real data we get the same result. Therefore, our limb-asymmetries detection cannot be caused by a mismatch of a complex limb-darkening profile versus a simpler parametrization like the quadratic law we use if the coefficients are left as free parameters (which, as seen above and expected, are indeed different by a significant amount with the ones one would calculate from the intensity profiles — an effect that is described and studied in detail in Espinoza & Jordán, 2015 and Howarth, 2011). This highlights in turn the power of leaving the limb-darkening coefficients to be free parameters in our fits; it shields our analyses against mismatches on limb-darkening parametrizations.

2. I previously remarked that parameters such as period, a/Rstar, and impact parameter (b) were fixed, when they may in fact correlate with the observed asymmetry or even cause a spurious result when these parameters are fixed to incorrect values. In response, the authors have run a number of fits allowing for several of these parameters to vary, within certain normal priors. The authors then show that their results are largely unaffected when parameters are varied within these prior bounds, relative to when these parameters are fixed.

While I applaud the authors for having run this experiment, this only addresses part of my concern. The experiment the authors have run shows that the claimed uncertainty (from Carter & May) on these parameters isn't significant enough to affect the results. It doesn't, however, show whether a slight bias in any of these parameters could produce the observed effect.

As an example, the authors use a prior on impact parameter of 0.4498 +- 0.0022. That is a very tight prior. While impact parameters can sometimes be precisely measured for giant planets, there are various biases that could affect it (e.g., a small amount of stellar dilution, a slight transit distortion due to finite integration time, star spots, a low level of unaccounted for transit timing variations, etc.). A much more convincing experiment would be the use of a much broader prior, such as b = 0.45 +- 0.05.

Similarly, the authors used a normal distribution with width of ~1 second (1.4e-5 days) when varying t0. Again, while t0 can generally be precisely measured, one can easily imagine an offset or bias in its determination (especially given that the work by Carter & May upon which these parameters are based remains under review). The precision on a/Rstar is contingent on the choices made for eccentricity, which could influence this parameter (even for very low eccentricity values, as are expected for a short period planet).

It would be much more convincing to use much broader priors in these tests, and show at which level parameters such as P / a/Rstar / b / t0 would have to be uncertain or incorrect before they impact the main result.

It is a good sign that the fit, with tight priors, produces similar results than with fixed values, but it remains entirely possible a small issue with any of these parameters could cause the entire observed asymmetry. I note that the other reviewer also asked for a fit that allowed the central transit time to vary -- I don't find that varying them with tight priors is a convincing response to this concern. Even if the authors are confident each of these parameters are indeed known very precisely, it would be useful to know to which degree the claimed main result is contingent on these prior parameters.

3. Related to the previous point, I previously remarked that in addition to uncertainties (or offsets) in a/Rstar, period, or impact parameter, uncertainties or offets in eccentricity could possibly cause the observed effect. In response, the authors included priors on e and $omega$ in their fit to propagate uncertainty.

This has not addressed my main question, which I therefore repeat here: "Regardless of what is found in Carter & May, given the clear reason for concern highlighted in Fig 5d a range of omega and eccentricity should be tested and the full results should be shown visually."

Rather than addressing this concern directly, the authors included a fit which allowed for a range of eccentricity, albeit with tight priors: ecos(omega) = 0.00087 +- 0.0014 and esin(omega) = 0.0014 +- 0.006. Given that, even within these very tight priors, the result has already lost some of its significancy (though still at ~3 sigma), it seems entirely plausible that a slightly broader range of allowable eccentricities could have a major impact on the claimed result. It would be easy to test significantly broader eccentricity priors (corresponding to, e.g., e in [0,0.05]).

Answer (to 2 and 3 above, as they are related): We understand the referee's concerns, as we share their concerns in cases where system parameters have larger uncertainties — one of the main reasons we believe extracting limb-asymmetries is difficult in general, but can be done if orbital parameters are known with good enough precision (as is discussed in the present manuscript, and has also been discussed in the literature, see, e.g., Dobbs-Dixon et al., 2012, von Paris+2016, Espinoza & Jones, 2022). In the particular case of WASP-39b we are lucky to have the very high precision JWST ERS transit observations, taken at different epochs, using different instruments, that allow for very tight constraints on the parameters of this system — the ones analyzed and presented in Carter & May et al. (2024).

Several co-authors in this paper contributed to the analyses presented in Carter & May et al. (2024) and therefore we trust those system parameters, which have been vetted through peer-review as well by Nature Astronomy (the Carter & May et al., 2024 paper is currently undergoing final editorial editing prior to its publication — it has already been accepted). In particular, the white-light lightcurve fits in Carter & May et al. (2024) do consider and marginalize over the possibility of correlated noise in the lightcurves (which could be astrophysical, e.g., via spots, or instrumental) and consider lightcurves with different cadences (all of which are relatively high-cadence —i.e., second-level cadence— such that finite integration times are not a problem). In our previous response, we have provided access to our own fits of this dataset, including all details to perform the fits for full transparency.

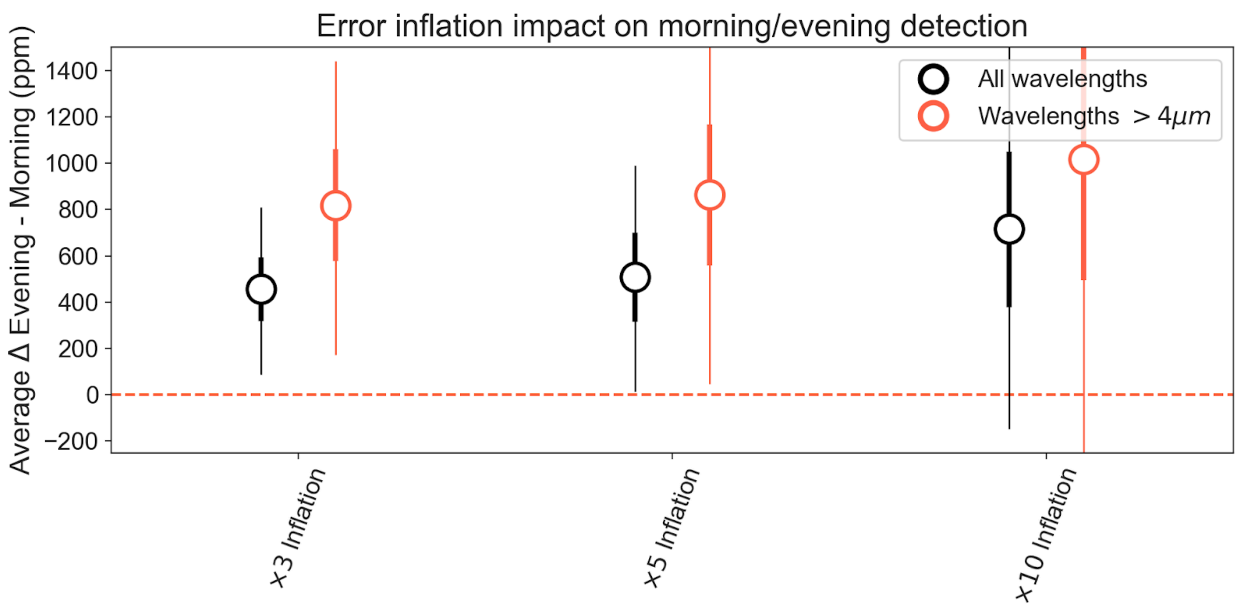
Despite all of these arguments, for transparency we still decided to perform the tests the referee suggested, performing the same exercise we did for the previous round of referee comments, but this time running wavelength-dependant fits with inflated errorbars (i.e., with broader priors, as requested by the referee) for all the system parameters mentioned in the question above in the case of a circular orbit (discussion on the eccentric orbit follows below). In particular, we ran fits inflating the errorbars on the system parameters by a factor of 3, 5 and 10 times. This corresponds to:

- For the x3 error inflation exercise, an error of 0.007 in the impact parameter, 0.036 in a/Rstar, 0.9 seconds on the period and 3.6 seconds on t0.

- For the x5 error inflation exercise, an error of 0.011 in the impact parameter, 0.06 in a/Rstar, 1.5 seconds on the period and 6 seconds on t0.

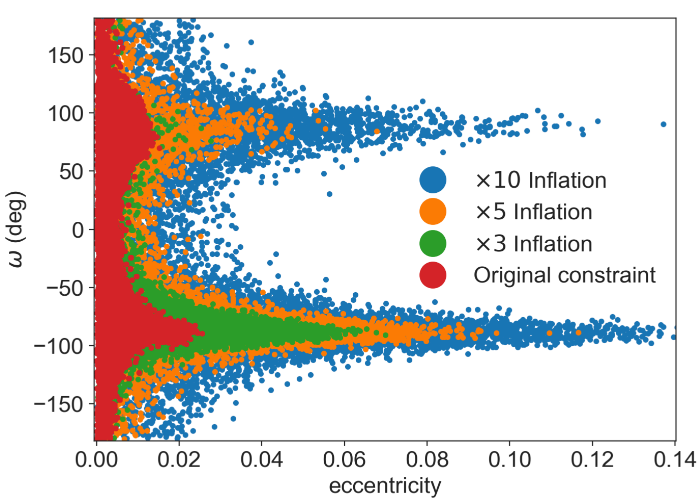
- For the x10 error inflation exercise, an error of 0.02 in the impact parameter, 0.12 in a/Rstar, 3 seconds on the period and 12 seconds on t0.

To summarize those fits, instead of showcasing the individual morning/evening spectra as we did in our first referee report, below we present the average evening-minus-morning depths of our 2-5 um spectra as a function of the level of error inflation (values above 0 imply a detection of morning/evening depth differences) for two wavelength ranges: the average difference across the entire 2-5 um range (black datapoints) and the average difference for all datapoints above 4 um (in red). The reason for the latter is to evaluate this difference where we see a clear excess on the results presented in the manuscript (i.e., around and beyond the CO_2 feature). **Thick errorbars** represent 1-sigma uncertainties, while thin errorbars represent 3-sigma uncertainties:

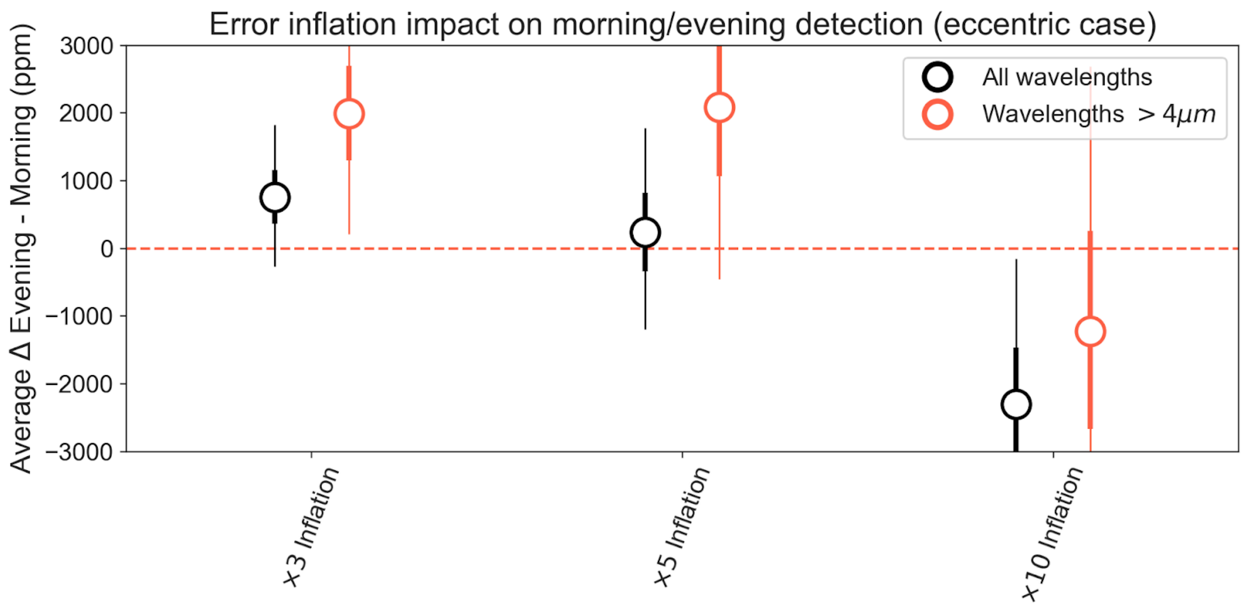


As can be seen, in this circular-case scenario (which is the preferred one given a lack of evidence in the data for an eccentric orbit), the detection stands for up to an error inflation of about x10, point at which the depth difference is consistent with zero at about 2.5-sigma.

Following the above experiment, we ran a similar one which, in addition to the error inflation on P, a/Rs, b and t0, we now also perform error inflation on our constraints on the eccentricity (e) and argument of periastron (ω) as well by the same factors. Because the priors are set on the $e\sin(\omega)$ and $e\cos(\omega)$ parameters in our fits and not directly in e and ω, we inflated the errors on $e\sin(\omega)$ and $e\cos(\omega)$ sequentially by a factor of x3, x5 and x10 and ran the same morning/evening fits once again. The mapping of each case to the eccentricity/ω plane is depicted on the left-hand side plot; this also includes our current constraint (in red, $e\sin(\omega)$ = 0.0014 $\pm$ 0.0054 and $e\cos(\omega)$ = 0.0009 $\pm$ 0.00013) for comparison/reference. Note that already when inflating errors by a factor of x3 we already incorporate eccentricities up to about 0.06, above the range suggested by the referee.



The average evening-minus-morning depths of our 2-5 um spectra as a function of the level of error inflation with the same definitions as the previous plot above are presented below:



As can be seen, in this case the detection stands for up to an error inflation of about x5 at least for wavelengths larger than 4 um, point at which this depth difference is consistent with zero at 2.4-sigma. Interestingly, the detection of larger evenings than mornings is resilient to a factor of x3 inflation in all parameters in this eccentric case for wavelengths above 4 um. Also notable is the apparent shift of the relation if the error inflation goes up to a factor of x10 — we note this is the sampler picking up only one of the modes of the solution in this case. Running the fit several times we identified in that case the errors are so large that they allow both larger evenings than mornings or larger mornings than evenings. This effect is not observed for error inflations x5 and below — in this case, only larger evenings than mornings are observed (as reported in our manuscript).

While as expected (from the circular case experiments) in the eccentric case the results are more sensitive to error inflation, it is reassuring to see that the evening-to-morning difference is particularly resilient to this inflation at wavelengths longer than 4 um, showcasing the robustness of the wavelength-dependance of the effect. Indeed, even in the x3 inflation case, the morning/evening spectrum still showcases an evident bump at those wavelengths, albeit as showcased in the results above, with a lower significance than the one observed in the spectrum using no error inflation (which was showcased in the previous referee report). We do note again that these experiments are upper limits on these significances, as we are fitting each wavelength independently — jointly fitting all wavelengths should have a much tighter constrain than the results presented above. This latter methodology is, however, very computationally expensive to perform (7 parameters per wavelength for 105 wavelengths above 2 microns implies at least a 735-parameter fit).

We conclude from our experiments that while running the experiments described above and suggested by the referee was a good exercise in understanding the detection limits on systems *like* WASP-39b, we do not believe they are representative of the current state-of-the-art *for* WASP-39b. For one, the era of JWST high-precision lightcurves completely changes what a “reasonable” errorbar is on the system parameters (e.g., see the detailed work of Eastman+2024 on how JWST lightcurves change this paradigm). Second, based on the experiments above we do agree with the referee that, as they note, our detection indeed depends on the current best estimates of the system parameters and their uncertainties (including constraints on eccentricity and argument of periastron, as showcased by our experiments), which is something that is explicit in the thorough experiments we performed in the responses and which we reported in the manuscript. One of the key messages of our work is, indeed, that performing detections like the ones we unveil is in general hard to do, but one that is *enabled* by the level of precision JWST transit lightcurves provide — i.e., obtaining precise transit parameters is essential to uncover limb-asymmetries as the ones we uncover in the present work. We have added text in our manuscript to highlight this fact in the section where the eccentricity experiments are presented and described. We don’t include the thorough experiments described above in the manuscript with the reasoning that interested readers will be able to read our exchange with the referee nonetheless were the article accepted/published in the journal.

Reviewer Reports on the Second Revision:

Referees' comments:

Referee #2 (Remarks to the Author):

I thank the authors for their responses to my previous points.

The authors have now run some of the experiments I had suggested in my initial review, where I asked them to investigate the degree to which various transit parameters could cause an apparent limb asymmetry.

Now that the authors have run the experiments (which I initially requested in review round #1), the results of these appear to be in line with my expectations, which were that the observed limb asymmetry is (in part) contingent upon the use of tight priors (here, priors following the highly precise transit parameters determined by independent work in Carter & May+ 2024). In their experiments and their response, the authors appear to now show that if some of the transit parameters were less precisely known than expected (e.g., if they were imprecise, or precise but not accurate), this could explain part or al of the observed limb asymmetry effect. For example, per the authors' tests, the significance of the effect is much smaller once an uncertainty of around 0.02 on the impact parameter is considered.

Similarly, and more concerningly, the significance of the limb asymmetry disappears entirely (or is even inversed!) once a wider range of eccentricities is considered. In this context the eccentricity-omega plot provided is misleading: while it is true that at 10x the errors in esinw and ecosw space a tail of eccentricity values above 0.06 (or higher) is allowed, the vast majority of the parameter space still covers very low eccentricities even in this case (the authors could just plot an eccentricity rather than e-omega distribution for the original, x3, x5, and x10 case to see this). Note that this issue is apparent even at just a x3 error inflation, at least when all wavelengths are considered (as they were in the main manuscript). At x5 error inflation, the limb asymmetry effect disappears, and at x10, it appears to be inverted (for reasons the authors don't seem to entirely understand).

I appreciate that the authors wish to argue that the original uncertainties, even though small, are to be trusted, and that there is no need to consider inflated error bars at all. There is some merit to this argument, given the highly precise observations from JWST and other sources. However, precision isn't necessarily accuracy. For example, impact parameters can be affected by even small levels of flux dilution, limb darkening effects, or light curve detrending. Similarly, the eccentricities used are presumably (as they come from the separate Carter & May work) the result of transit modelling itself, and so in addition to some of the same effects, are furthermore sensitive to stellar density systematics (though the fact that ecosw values are very nearly zero, presumably based on the generally reliable secondary eclipse timing, is reassuring).

Despite all of the above, on balance I believe that the authors have demonstrated that the claimed limb asymmetry effect is at least plausible. Given the general importance of detecting such an effect, I am happy to recommend even a plausible detection for publication.

However, I do believe it should be made clear in the main manuscript to which degree the detection is contingent upon both the *precise* and *accurate* transit parameters. As such, I disagree with the authors' argument that the experiments they have now run should only be part of the exchange with the referee, where it may be missed by many reasons. Given the authors have already conducted the relevant experiments, I ask that: 1) the experiments are described and included in the methods of the manuscript, including some of the figures provided in their response to me, 2) that the authors describe in the main manuscript how, *if* parameters were less precise or inaccurate, this could under certain circumstances provide an alternative explanation for the observations, and 3) that the importance of transit parameter precision and accuracy and the potential impact of these on the key results at least be acknowledged and briefly summarised in the main body text of the manuscript.

Author Rebuttals to Second Revision:

Response to the Nature Review, Round # 3

We would like to once again thank the referees for their input on this last round of referee comments. Following the referee's and editorial suggestions, which are implemented in **red, bolded sentences** in the manuscript when relevant. We have:

1. Changed the title of the paper from "Detection of inhomogeneous terminators on the exoplanet WASP-39 b" to "**Inhomogeneous terminators on the exoplanet WASP-39 b**", in order to comply with Nature title style.

2. Following the editor's suggestion of adding text "(...) along the lines [of what the referee] recommends: your interpretation is plausible, but for it to be correct the uncertainties need to be both precise and accurate. (...)" to the boldface summary paragraph, that goes along the recommendation by the editor to write: "(...) '"Here we report observations of the morning and evening transmission spectra of WASP-39b. Under the assumption that the uncertainties are both precise and accurate the evening transits are on average deeper than the morning ones, which we conclude point to an evening terminator that is 177+86-57 K warmer than the morning. ..." (Or something along those lines which conveys the real uncertainty in your conclusions.)'", we have added "**Under the assumption of precise and accurate orbital parameters on WASP-39 b,**" before the phrase "**here we report the detection of inhomogeneous terminators on the exoplanet WASP-39~b, which allows us to retrieve its morning and evening transmission spectra in the near-infrared (2-5 um) using JWST**" in the summary paragraph.

3. Following the first referee ask that "(...) the experiments [that the referee asked in our responses] are described and included in the methods of the manuscript, including some of the figures provided in their response to me", we have added the last plots showing the error inflation impact on limb asymmetry differences in the Methods section, under the "Robustness against eccentricity" sub-section, along with the following text explaining the results: " (...) **In order to further explore the impact of our orbital parameter estimation on our detection of inhomogeneous terminators on WASP-39~b, we ran experiments on which we artificially enlarged the errorbars on the orbital parameters by 3, 5 and 10-fold, for both circular (i.e., $e=0$) and eccentric (i.e., free $e\cos (\omega)$ and $e\sin (\omega)$) cases, and compared the average evening-to-morning depths for each case. We compared these differences both across the entire wavelength range and for wavelengths above 4 μm, where we see the largest deviations on evening-to-morning depths. The results from our experiment are presented in Figure \ref{fig:errorinflation}. As can be observed, for our circular case, error inflations of up to 5 times our reported errorbars still support our detection of inhomogeneous terminators, whereas for the eccentric case, error inflations of up to 3 times only support this detection at**

wavelengths above 4 μm. These experiments highlight the dependency on the detection of this effect on the accuracy and precision of orbital parameters for WASP-39~b --- which might also be the case for detecting the effect on other exoplanets. (...)".

4. Following the second and third referee ask that we describe in the main manuscript how "(...) *if* parameters were less precise or inaccurate, this could under certain circumstances provide an alternative explanation for the observations" (second ask) and that "(...) the importance of transit parameter precision and accuracy and the potential impact of these on the key results at least be acknowledged and briefly summarised in the main body text of the manuscript" (third ask), we added the following text to the main part of the manuscript: "**This detection of evening-to-morning spectral differences is, in turn, robust at the 3-sigma level even when accounting for the current best uncertainties on the orbital parameters of WASP-39 b. The result does, however, depend on the accuracy and precision of those parameters up to a factor of a few of their current state-of-the-art errorbars, which highlights the importance of accurate and precise determination of exoplanetary orbital parameters when attempting to perform morning-to-evening spectroscopy as we do in our work (see Methods)**".