We'd like to thank the reviewers for their helpful comments. Those addressing the issue of negative weights in our average and how those relate to the error covariance matrix were particularly helpful to our understanding. We have updated our text to reflect this updated viewpoint.

Responses to comments from Reviewer #1

> The authors study the correlations between errors in
> X_{CO2} satellite retrievals, based on reference lidar
> measurements, and discuss various ways to account for
> them in atmospheric inversions. The paper looks a bit
> like the clean minutes of a brainstorming meeting:
> every sentence is well written but the logical flow is
> curvy and difficult to follow. The authors have not
> done enough to make their thoughts accessible and to
> take the text beyond elaborate speculation, perhaps
> simply because their thinking is not yet ripe for
> publication. Maybe it doesn't matter, as the paper
> will be cited anyway given the role of this activity
> for the OCO-2 team, but for the few who will bother
> to read it, it may be a daunting task, perhaps in the end wasted.

We hope our replies to the detailed comments below will address this reviewer's issues.

> I am listing a number of comments here to help clarify the presentation.

> Footnote 1, p. 25: the disclaimer here is a bit hidden, but
> it is actually essential. Basically, if the "good reasons" listed
> here are correct, all results of the paper can be ignored. This
> observation could be fatal for the patient reader who painfully
> reaches this page… In the end, nothing is given to convince the
> reader that the MFLL-OCO-2 differences do indeed represent OCO-2
> errors, that the two scales of correlation lengths found (10 and
> 20 km) should be used at all in OCO-2 error models. It's embarrassing.

No need for embarrassment here. We have attempted to calculate a correlation length scale using the limited data that is available for that purpose. When more data become available to test our conclusions, that should certainly be done -- if the reviewer can suggest some additional data that we could use to refine these estimates, we would be happy to work with it. In the meantime, the derivations presented here using the newly-calculated correlation length scale remain valid regardless of what precise value is used for that quantity. We do hope that the reviewer does not mean to suggest that no one can publish using new data -- little progress would be made in such a world.

> There is good and interesting math here, but the authors belittle
> it by arbitrarily rejecting certain math results: why should the
> negative weights not be physical (l. 20) or considered undesirable
> (l.331)? They simply follow from the authors' correlation model:
> if the authors are not satisfied with this consequence, they should
> change the model rather than fooling the math.

Based on this comment as well as one of Reviewer #2's below, we have looked into this issue more and we agree that there is nothing inherent in the structure of the error covariance matrix that precludes the individual weights in our weighted average from assuming negative values – for the error correlation models that we use, these negative values are to be expected. The main constraint provided by a positive definite covariance matrix is that the *sum* of all the weights must not be negative or zero – if that criterion is met, a weighted average may always be calculated (since this implies that the inverse of the covariance matrix is also positive definite, ensuring that the denominator in (6), $\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}$, is always positive).

In terms of computing our weighted average, the implication of this is that there is nothing in our correlation models to prevent the mean value from falling outside the range of the values going into the average. If we wish to enforce that more stringent criterion, then we may require that each individual weight be non-negative, as we have done in the paper.

While there apparently is no fundamental requirement that the individual weights in a weighted average are non-negative, one often sees this given as a basic requirement. Wikipedia, while not an unimpeachable academic reference, reflects broader practice when they give on their "weighted arithmetic mean" page the mathematical definition of such a mean as follows:
"Formally, the weighted mean of a non-empty finite multiset of data $\{ x_1, x_2, \ldots , x_n \}$, with corresponding *non-negative* weights $\{ w_1, w_2, \ldots , w_n\}$ is
$\langle x \rangle = \text{sum}\{w_i x_i\} / \text{sum}\{w_i\}$ ...
Therefore, data elements with a high weight contribute more to the weighted mean than do elements with a low weight. *The weights cannot be negative.* Some may be zero, but not all of them (since division by zero is not allowed)." [emphasis ours]

The practical effect of any individual weight being allowed to be negative is that the mean value computed may fall outside the range of the data going into the average – a result that is contrary to the very idea of an average. We are certainly justified in imposing the requirement that all individual weights be taken to be non-negative (and at least one positive) as an additional constraint to ensure that we obtain an average that falls inside the data range for the work we are presenting here. Such a choice is hardly "arbitrary". It simply constitutes an additional constraint that we choose to add, in addition to the assumed correlation structure, to obtain results that don't swing

outside the range of the input values. We reject the reviewer's assertion that such an approach is incompatible with the use of this correlation model and that we are "fooling the math". There are good reasons to do it.

Requiring each individual weight to remain non-negative, rather than the sum of the weights, can be thought of as a more local constraint than the global constraint imposed by positive definiteness. In the text, we show that it may be used as a filter to throw out those scenes that cause problems locally. This added constraint may be useful in guiding us to design a form for the error correlations that has the local boundedness constraint that we seek: the constant correlation model violates this local constraint generally, while the exponentially-decaying correlation model only requires about 5% of the data to be thrown out to satisfy it – perhaps if we keep looking we can find some other simple error correlation model that will satisfy both the local and global constraints on the weights.

We have added new discussion in the text explaining that the correlation models we assume should be expected to drive individual weights in our average into the negative range, there being no requirement against that in the math. And we have added additional wording to the text explaining that we have forced the individual weights in the average to be non-negative in order to prevent the averages from falling outside the range of the values input to the average, a result that we feel is undesirable in our averages.

> l.35-37: the sentence seems general, but does not apply to GOSAT in practice.

The sentence in question is this: "The resolution is limited both for computational reasons (the models must be run many dozens of times across the measurements to obtain the inverse estimate) and because the spatial coverage of the satellite measurements is currently not dense enough to resolve spatial scales much finer than this when solving at typical time scales (the gap in longitude between subsequent passes of a typical low-Earth- orbiting (LEO) satellite is ~25°, resulting in gaps of between 3° and 4° across a week, gaps which are generally never filled in further due to the repeat cycle of the satellite's orbit). "

The reviewer is correct that the longitude-separation calculation at the end of the sentence does not apply to GOSAT data taken over land, which may include data for 3, 5, or more data points taken in the cross-scan direction (or effectively 3, 5, or more parallel tracks of data per orbit groundtrack). For GOSAT, the resulting gap size would have to be divided by 3, 5, or more, resulting in potentially a finer scale being resolvable over land.

To be clearer, we have added "taking a single swath of data along its orbit path" after "(LEO) satellite".

> l.43: why would it make little sense to assimilate the measurements
> individually? From the text, it is obvious that it is so much easier
> than trying averages. So, this can make a lot of sense. Personally,
> I would still prefer the averaging but for reasons that are not
> discussed here (numerical stability, but this is only a feeling).

The sentence in question: "The individual OCO-2 retrievals are generally averaged together along-track across some distance closer to the model grid box size before being assimilated in the inversion: this is because the modeled measurements to which the true measurements will be compared in the inversion are available only at the grid box resolution, so it makes little sense to assimilate each measurement individually when assimilating a coarse-resolution summary value will do just as well."

The answer is that it reduces the computational load related to the assimilation of the data by as much as a factor of 240 (the maximum number of individual scenes per 10-second span) if one averages the data beforehand. We don't understand why the reviewer feels that it would be easier to process each point individually in the assimilation, rather than averaging beforehand, unless he/she intends to ignore error correlations altogether in the process (and even then, the computational load would still be up to 240 times greater). If the same error correlations are to be considered in both cases, the effort required to do so is the same, whether done inside the code or outside, beforehand. But the computational savings obtained by pre-computing the averages remains, either way.

> l.48: interesting comment… The dependence of the correlations on
> the scene questions the representativeness of the ACT measurements
> used here. This key element is only briefly touched on in the warning in line 181.

The sentence spanning line 48 is this one: "CO_2 mixing ratios in the upper part of the atmospheric column (at all levels but the immediate surface layer) feel the influence of multiple flux locations at the surface due to atmospheric mixing, causing errors in adjacent measurements to be highly correlated there. "

It is not clear to us how the comment relates to this sentence. It is also not clear to us what argument the reviewer is trying to make here. If the comment implies that one cannot attempt to use a correlation model with a blanket correlation length applying across all data points, when local correlation lengths are higher or lower than that due to local conditions, then we must disagree. It is still better to try to use a more-accurate correlation model, even if the non-negativity constraint must be added as well, than not to attempt to account for error correlations at all.

> l.123: OCO-2 was already defined in l. 39. Same comment for ASCENDS a bit later.

We thank the reviewer for catching these -- we have removed the extraneous re-definitions.

> l.168: the motivation behind removing the linear trend is obscure.
> For the constant value removal, I do not see how this affects the
> calculation of the autocovariances.

Pertains to this sentence: "We then detrend this weighted difference timeseries across each flight leg – subtracting either a single constant value or a linear trend Y(x), where x is the along-track distance."

The reviewer is correct that a constant offset should not effect the spectrum returned by a harmonic analysis, insofar as the harmonic analysis usually solves for the average of the timeseries as part of the analysis such that it does not affect the harmonic terms (that is, the harmonic terms are solved for to describe the *variability* of the time series, which is described as deviations from the mean). This depends on whether one's particular routine does in fact solve for the mean and remove it or not, however. Our routine does not, so we are forced to pre-compute it and pre-subtract it before presenting the variability timeseries to the routine. Since the main factor that we address here is whether removing the linear trend or not makes a difference, we will refer to the case where only the constant offset (but not the linear trend) has been removed as the "linear trend not removed" case.

Linear trends are often also removed from a timeseries of data before performing a correlation analysis, in order to avoid including the harmonic terms needed to describe the trend in with the terms needed to describe the stationary variability. Not removing the trend in a harmonic analysis is equivalent to assuming that the trend repeats itself in a sawtooth-like pattern when the data span is repeated end-to-end. The spectrum needed to describe the trend is red: not removing the trend before doing a harmonic analysis will result in a spectrum in which the lower frequencies elements are exaggerated, making it more difficult to assess the longer frequency terms that pertain only to the stationary variability (i.e. the non-trend part).

To be conservative, we have chosen to use the spectrum computed with the trend not removed, giving the longer correlation length (20 km vs. 15 km) and resulting in less information being retained as the data are averaged across a given span length. We would also be comfortable with the more aggressive assumption that the trend should be removed, however (giving the shorter ~15 km length scale). We have chosen to show results for both cases to help illustrate the uncertainty in the calculation of the length scale.

Also, to be clearer, we have also replaced "Constant value removed" with "Linear trend not removed" in Figure 1.

> l.208 and 219: Kalman filters are used to control fluid state variables,
> but not boundary conditions such as surface fluxes. They are outside the
> scope of the discussion, unless the authors refer to simplifications like

> the Kalman smoother, but in this case the assimilation window covers
> periods much larger than the observation error correlations lengths which
> are discussed here.

On line 208: "Most estimation methods used in global atmospheric trace gas inversion work (Bayesian synthesis inversions, Kalman filters, variational data assimilation) combine measurement information in different timespans as:"

On line 219: "This assumption of uncorrelated errors between different timespans is built into the derivations of these inverse methods explicitly, for example in the Kalman filter, in which the dynamical errors related to propagating the measurement information from time to time are assumed to be uncorrelated with the measurement errors themselves."

The reviewer is incorrect in asserting that boundary conditions such as surface fluxes cannot be included as control variables in Kalman filters. For the CO_2 flux estimation problem that we are dealing with here, one can envision a state vector that includes both the current 3-D CO_2 field and surface fluxes in its state vector to be estimated -- new measurements would modify both the CO_2 field and the surface fluxes, the two of which would be expected to be highly correlated. This would not be as effective as including multiple past fluxes in that state, as well (a situation that would lead to it becoming effectively a fixed-lag Kalman smoother), but that does not mean that it would be an illegitimate model. So Kalman filters are certainly not outside the scope of discussion. And the Kalman smoother is not a simplification of the Kalman filter -- the reverse is actually true (the filter is a simplification of the more general case of the smoother, for only one lag step in the state). It is not clear what point is being made related to the assimilation window length.

> Section 3.1.3: Some main elements (the tridiagonal influence matrix, the statistically-
> optimal inflation factor) have been shown years ago by Chevallier et al
> (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2007GL030463)
> in a short paper. It may have been too brief, but was much more accessible, I think.

Yes, we have been remiss in not mentioning this reference, given the good work that Chevallier has done in characterizing errors in our flux inversion problem and given the similar approach that he took in that reference to what we have used here (especially the use of the tri-diagonal covariance inverse, but also the use of an inflation factor, even if the precise form for computing it was not provided in that reference). This was an accidental oversight. We hope to rectify that here by adding the following sentence after equation (17): "(Note that Chevallier et al. (2007) handled exponentially-decaying correlated errors as well and used a similar tridiagonal matrix for the inverse of the covariance.)"

Response to Comments from Reviewer #2:

> General comments.
> Authors propose an innovative approach to the construction of the
> error covariance matrixes for X_{CO2} observations by OCO-2 satellite.
> The method relies on comparison of OCO-2 retrievals to the collocated
> X_{CO2} observations by an airborne lidar made along ground tracks of
> several tenth of km in length, and thus avoids use of model simulations
> for estimating the amplitude and spatial statistics of the OCO-2
> retrieval errors, which constitutes a major innovation presented in
> this work. Authors derive spatial error correlation lengths in the
> range of 10 to 20 km, which is a new result, based solely on
> observations. Authors consider 2 variants of the spatial error
> correlation models that are used to evaluate the errors and errors
> correlations for target 10 second averages derived from intermediate
> 2-second mean data, and propose ways to overcome related technical
> difficulties. The paper is well written and can be accepted after minor revisions.

> Detailed comments

> Although the problems and remaining deficiencies are extensively
> discussed in the Summary and conclusions, it is worth revisiting
> a couple of topics below:

> The paper would benefit from deeper look at the problem with negative
> covariance elements, addressed in Sect. 3.2. Similar problem with
> spatial correlations would appear with covariance matrixes for
> surface fluxes errors in the same inverse problem for which the data
> uncertainty matrix is constructed. But the problem of negative elements
> for surface fluxes is not widely known, thus there is a possibility
> that the problem here is caused by an unrealistic design of the
> covariance matrix, but not by statistical properties of the data in hand.

Please see the response above to Reviewer #1's comments on lines 20 and 331.
We argue there that we are justified in demanding non-negative weights for the average as an additional constraint we impose upon the problem to keep the average from falling outside the range of the input values

Reviewer #2 argues above that a similar negative weight problem ought to exist when handling correlated errors in the fluxes (as, for example, when averaging fluxes to a coarser time resolution, while properly accounting for the off-diagonal terms in the error covariance matrix), but this has never been described, suggesting that there should not be a problem with negative weights in specifying correlated data errors, either, and that some error has been made in the assumptions used in this paper, or in

the construction of the covariance matrix. While that is not the case, this analogy has in fact been very helpful to us in adjusting our thinking, and we thank the reviewer for bringing it up.

Any covariance matrix with off-diagonal terms (including the ones we construct using our correlation models) can potentially have the elements of vector $\mathbf{R}^{-1}\mathbf{1}$ (the individual weights In our average, as we define then) assume negative values. This would be true both of the measurement error covariance matrices that we deal with in this paper, as well as the flux error covariance matrices that the reviewer brings up in this comment. The reason for this is that there is no fundamental requirement in the form of the covariance matrices that requires each individual weight to stay non-negative. Instead, the key criterion is that the *sum* of the weights of the terms going into the average must be greater than zero. In that case, the denominator $\mathbf{w}^T\mathbf{1} = \mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}$ in (5) and (6) always remains greater than zero and the weighted average is always well-defined. The sum of the weights, $\mathbf{w}^T\mathbf{1}$, stays positive whenever the covariance matrix, $\mathbf{R}$, is positive definite (since then $\mathbf{R}^{-1}$ is also positive definite, and therefore $\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1} > 0$, by definition of positive definiteness). Steps are always taken in estimation methods to ensure that covariance matrices remain positive definite, in the face of round-off error or noise driving them in that direction: divergence occurs in the Kalman filter, for example, when this condition is violated, and dynamic noise is added when propagating the covariance matrix forward in time to prevent this problem. So there is does not need to be a problem in the design of our covariance matrix for this to happen – it is a natural feature of covariance matrices with off-diagonal terms. And it (the existence of negative weights) ought to occur when dealing with flux error covariance matrices, as well, even if this is not generally well known.

There is no problem, as long as the user is satisfied with the net impact of the data (the mean data value in the case where the data is averaged before assimilating, or the net effect of the data in the inversion if each individual scene is assimilated separately) potentially falling outside the range of that given by the individual data when these correlation models are used. However, we do see this as a problem – we do not want our average to fall outside the range of our input data -- and seek to mitigate it, while hopefully retaining the other advantages of the correlation models. To do that, we must impose an additional, more stringent, constraint in demanding that each individual weight be non-negative, rather than just the sum of the weights. This criterion (that each individual weight be non-negative) keeps the average value from straying outside the range of the data being averaged – this is certainly reasonable, and, since it is an effect that is caused by assuming a non-zero error correlation model itself in the first place, it is solved by adding an additional constraint outside of the correlation model. We have modified our text in Section 3.2 to discuss the points raised above. We continue to demand the additional constraint that each individual weight be non-negative as before, but the additional text puts that decision into better perspective.

> Unfortunately authors do not provide information on the amplitude of

> the correlating component – what is a fraction of the (OCO-2 – lidar)
> difference that is correlated at 10 km or less scale. If that is only
> a fraction of the 1.5-2 ppm of error as found by comparison with TCCON,
> then the origin and spatial scales of remaining is unknown, thus it can
> be treated as random and uncorrelated. Bell et al (2020) notice that
> the correlations at local scale are pretty low, they write: "We conclude
> from these low correlations that for an average scene with no strong
> variability in the X_{CO2} field, OCO-2 and the MFLL do not typically "see"
> the same small-scale features". In practice, the adopted level of
> correlation coefficients as shown on Eq. 15 do not appear justified by
> the comparison with MFLL and may possibly come from a separate source.

Yes, the coefficients shown in (15) come from a separate source – an independent analysis make by Dr. Susan Kulawik – we have added a note in the text that she did not use the MFLL data in her analysis, but rather based it in part on TCCON and *in situ* aircraft observations. Actually, we feel that the +0.3 correlation coefficient calculated by Kulawik for use over land agrees quite well with the correlations that we computed using the MFLL data, shown in Figure 1, when applied to 10-second averages: the integral of +0.3 from 0 to 67.5 km gives a value of about 200, compared to an integral under the blue line on the left panel of Figure 1 of something very close to that across the same range on the x axis. This agreement, based on very different data sources, gives us some confidence in the results of our MFLL-OCO2 analysis.

The reviewer is correct that we should have reported the magnitude of the variability corresponding to the correlations we report. This can be computed directly from the spectrum shown in Figure 1 by multiplying the blue curve by the normalization factor that we have divided by to get the correlation spectrum. For the case in which we remove the trend, that factor is 0.84 sigma when we analyze x_i / sigma_i, and 0.59 ppm when we analyze x_i itself without dividing by the associated uncertainty. For the case in which we do not remove the trend, the values are 1.04 sigma and 1.003 ppm. This means that the RMS difference between the MFLL and OCO-2 values is about 1.0 ppm for the non-detrended case, so that the correlation coefficient of +0.3 associated with separation distances of from 20 to 40 km is describing an RMS variability of about 1.04 * sqrt(0.3) = 0.57 ppm. Or in terms of the fraction of the variability described at scales of 10 km or less that the reviewer asks about, anywhere from 100% of the variability at zero separation distance to about sqrt(0.37)=60% of the variability (in terms of concentration difference instead of its square) at a separation distance of 8 km (the first dot to the right of zero in Figure 1). This is a substantial portion of the full variability, so we don't think it would be fair to discount it as being unimportant. It is also not fair to discount it compared to the RMS difference of 1.5 to 2 ppm (OCO-2 compared to TCCON) since that number includes biases in both TCCON and OCO-2 that would drop out in any analysis of variability between the two. The true RMS error in the OCO-2 data taken over land is down in the 1.2-1.5 ppm range at the moment, so the 1 ppm RMS difference between MFLL and OCO-2 is capturing about half of that – the

other half could plausibly be attributed to longer correlation scales than those that can be assessed here. We have added text discussing the magnitude of the variability associated with the MFLL – OCO-2 differences we have analyzed here.

> Minor comments, technical corrections:

> L9 'Errors in the CO_2 retrieval method have long been thought to be
> correlated at these fine scales' – It would be more accurate/safe to
> say that data are correlated, rather than the errors.

By data, we both mean retrievals here. We actually do mean to point to correlations in the errors rather than in the data themselves, since the errors are what are being quantified in the covariance matrices we discuss. Since the data are obviously strongly correlated due to their large background (~400 ppm) and large variability on seasonal and synoptic timescales, it might appear safer to only talk about them, but that is really besides the point. We know that the accuracy of the retrievals is strongly tied to the accuracy of the assumed surface properties, atmospheric scatterers, and atmospheric gases needed as part of the retrieval, and the *a priori* values assumed for these in the retrievals all have errors, so it is not surprising that we should talk about correlations between errors in the retrieved values themselves, rather than just in their values.

> L44 Alternatively, one can call this 'summary value' an 'average value'

Ok, we made that change.

> L48 Authors imply the errors here correspond to model-observation
> difference, and contributed by model errors related to smoothing due
> to coarse model resolution. Its better to define somewhere above this
> point what is implied by 'errors'.

The errors referred to here are measurement errors, more specifically errors between the retrieved and true X_{CO2} values. We think the sentence on lines 45-46 already does a good job specifying the errors we are discussing: "Whether the individual OCO-2 retrievals or some coarser-resolution average measurement are assimilated into the inverse model, correlations between the errors in the individual CO_2 measurements must be considered." The following sentences delve into what factors cause errors in the X_{CO2} retrievals, and some of these factors do involve modeled variables used in the retrievals. We do not get into the exact cause of those errors, because that is not our focus here.

> L86 Adding some MIP paper reference should be useful here (eg Crowell et al. 2019).

Yes it would – we have added a reference to Crowell et al. (2019) at the end of this long sentence.

> L167 Any rationale for detrending Y rather than X itself?

Analyzing the timeseries $Y_i = X_i$ / sigma$_i$ instead of X_i itself is better, because it places the proper weight on deviations that are large in a statistical sense (in sigma space) rather than in an absolute sense. If just X_i was analyzed, it would be dominated by large deviations in places where the measurements are less certain, and the parts of the time series containing the most reliable data would be de-emphasized. That said, we also performed the analysis directly on the X_i time series and the spectrum did not change much. A note on this has been added to the text.

> L203 As the MFLL data are first aggregated to 7-9 km blocks, it
> appears that one needs to clarify here on how the analysis would
> become useful for finer scales.

As noted here, because of this initial aggregation into 7-9 km blocks, scales finer than this cannot be addressed by this analysis. To be able to say something at finer scales, this blocking would have to be done across a shorter length scale -- this could be done down to the 2.3 km length of an individual OCO-2 cross-scan, at the expense of increased noise in the MFLL measurements. Whether the best fit to the spectrum would remain in the 15-20 km range determined here or would decrease would remain to be seen. A note to this effect has been added to the text.

> L221 Although temporally uncorrelated errors are convenient for
> Kalman filters, it does seem to be an excessive requirement, need to
> add a reference to appropriate text, if exists.

This sentence is being referred to: "This assumption of uncorrelated errors between different timespans is built into the derivations of these inverse methods explicitly, for example in the Kalman filter, in which the dynamical errors related to propagating the measurement information from time to time are assumed to be uncorrelated with the measurement errors themselves."

Kalman filters were designed originally for use inside real-time control loops, e.g., for a missile, rocket, or statically-unstable fighter aircraft. All the information needed to define the estimate at a given time is derived from new measurements at that time and the previous estimate propagated forward in time with a dynamical model. The errors between the different values in the new data vector may be correlated, and the dynamical errors may be correlated, but explicit correlations between the state estimates at different times can only be included by adding the state estimates at previous times into the state vector (a step that effectively turns the Kalman filter into a fixed lag Kalman smoother). We had added a reference to Applied Optimal Estimation, Gelb ed., The M.I.T. Press, 1974, 374 pp. to support this.

> L627 Mistype: correct 'Zendolo' to 'Zenodo'

Corrected, thanks for catching that.

> L660 For web document, need to give url.

Thanks for catching this. We have added the link:
https://cce.nasa.gov/ascends_2015/ASCENDS_FinalDraft_4_27_15.pdf