We sincerely thank the reviewers for their very helpful feedback on this paper. We address all their comments and suggestions below.

Reviewer comments are in black italicised text.

Our responses are in blue, regular text.

Where practical/necessary, we provide a screenshot of the track-changes document to show the changes that we have made (in outlined boxes). In these, text that is removed is ~~struck through~~ and coloured red, while new text is underlined and coloured blue.

Ian Ashpole (on behalf of both authors).

REVIEWER # 1

Received and published: 18 May 2022

23 "These L3L L3W differences are clearly linked to retrieval sensitivity differences"

While reading the abstract it was unclear if the authors were taking into consideration the fact that CO emissions over water are negligible and whether that affects the difference between L3L and L3W.

This is a fair point, but it is also important to consider that due to the relatively long-lived, well-mixed nature of atmospheric CO, the portion of coastal L3 grid boxes situated over water are unlikely to represent "pristine" conditions. Moreover, there are a large portion of coastal L3 grid boxes where emissions from the land-based component are also negligible, due to the lack of large anthropogenic CO sources or natural emission hotspots (see distribution of coastal L3 grid boxes in Figure 2). Finally, we do demonstrate cases where retrieved surface-level CO concentrations in L3L are less than in L3W for a given L3 grid box, and we would not expect to see this if the differences are simply related to differences in emissions. It therefore seems unlikely that the L3L-L3W CO concentration and trend differences presented are strongly impacted by land-water emission differences within these grid boxes – especially against the weight of evidence that links these to well-understood retrieval sensitivity contrasts. However, we do understand that it is important to mention that there are plausible physical factors that could contribute to L3L-L3W differences in some circumstances (i.e. that not all differences are solely a retrieval artefact), and we have included a paragraph in Section 3.2.1 (where L3L-L3W retrieved VMR differences are shown and discussed) to address this:

> ~~priori VMRs used in the retrievals.~~ It should be noted that there are additional physical factors that could plausibly play a role in generating the L3L – L3W retrieved VMR difference that is observed, in addition to retrieval sensitivity. Given that most CO sources are land-based, a decrease in VMRs from land to water might be expected, especially in the LT. However, this assumption only seems reasonable where large CO sources are proximal to the coastline, as it is unrealistic to expect gradients as large as we observe in background CO (which coastal grid boxes far from large CO sources are more likely to represent) across the relatively small distance covered by a L3 grid box. Given the relatively long-lived, well-mixed nature of atmospheric CO, VMRs retrieved at a given location are a function of both local emissions *and* transport, and the portion of coastal L3 grid boxes situated over water therefore do not represent pristine conditions in

screenshot continues on next page

comparison to the adjacent land-based portion of the grid boxes. This is verified by comparing a priori VMRs (also shown in Figure 4), which suggest the land-water difference in CO concentrations should be negligible (mean L3L – L3W a priori VMR difference = 0.69 ppbv, compared to a mean retrieved VMR difference of 10.29 ppbv). The above reasoning can also be applied to the question of whether wind direction is responsible for creating the observed L3L – L3W difference in retrieved VMRs: It could be hypothesised that a prevailing

onshore wind may lead to CO concentrations being higher over land than water, yet the negligible L3L – L3W a priori VMR difference, the fact that atmospheric CO is well-mixed, and the clear land-water sensitivity gradient that has been demonstrated suggest that wind direction does not play a big role in creating the land-water difference observed in retrieved VMRs. To further rule out the role of wind direction, the L3L – L3W retrieved VMR comparison has been analysed alongside wind direction for several case study grid boxes, and there appears to be no notable shift in wind direction whether L3L or L3W is greater for a given grid box. Results for this analysis are given in the Supp. Mat. (SM5). The weight of evidence therefore points towards L3L – L3W retrieved VMR differences being a function of reduced retrieval sensitivity over water compared to land.

421 "As expected from the previous analysis, the land-water sensitivity contrast is greater when mean VMRs are significantly different than when not."

It's as if they assume that all land-water contrast is a processing artifact. See line 446

This comment is addressed by our response to the reviewer's comments on L23 (above) and L446 (below).

446 "An underlying assumption is that the temporal trend in "true" VMRs should not vary much across a 1º x 1º L3 grid box."

This was revisited in line 399. It seems as if the authors neglect to account for wind direction. If the prevailing wind is blowing from the ocean inland (e.g. the coast of California) then the CO concentration could be much higher over the land than the water in the same gridbox. Whereas if the wind was blowing *out to sea, one would expect far less difference between L3L and L3W. Yet it appears that the authors are not accounting for this.*

The reviewer is correct that we do not account for the effect of wind direction on L3L-L3W differences. While we agree in principle with their statements about the expected effect that wind direction could have on L-W differences in VMRs in some cases, we did not address this for the following reasons:

1. Ashpole and Wiacek (2020 AMT) explicitly considered this for the coastal city of Halifax, Canada (population > 400,000) and found that shifts in wind direction (based on ERA Interim reanalysis data) could not account for differences in retrieved CO over land and water for the L3 grid box containing Halifax.
2. Coastal marine CO is unlikely to be "background" levels due to the relatively long-lived, well-mixed nature of atmospheric CO. Additionally, not all coastal grid boxes are separated from downstream sources by thousands of km's of ocean (as in the California example – see distribution of coastal grid boxes in Figure 2).
3. The wind direction is unlikely to be onshore (synoptic or from a well-developed sea breeze) for all coastal grid boxes where L3L > L3W.
4. Hedelius et al. (2021, JGR), demonstrate that most large cities do not stand out against the surrounding region in trend analyses based on MOPITT data (a maximum of 21 of 500 cities compared saw faster trends than the regional average).

However, to evaluate any potential impact of wind direction on our results, we have compared wind direction (taken from ERA Interim data, u and v wind vectors at 10-m level) for 6 of the case-study L3 grid boxes containing large cities considered in Sect 3.4 for days when L3L > L3W and L3L < L3W. We found no marked differences in wind direction on these days for any of the grid boxes considered, giving confidence that wind direction does not have a large impact on our results. Results are presented for the case study of the grid box containing Dubai below (Figure R1), and we include results for all 6 cities in the supplementary material (SM5). As with the comment above, we do acknowledge that it is important to mention that there are plausible physical factors that could contribute to L3L-L3W differences in some circumstances (i.e. that not all differences are solely a retrieval artefact), and we have included a paragraph in Section 3.2.1 (where L3L-L3W retrieved VMR differences are shown and discussed) to address this (see paragraph inserted above in response to comment for L23).

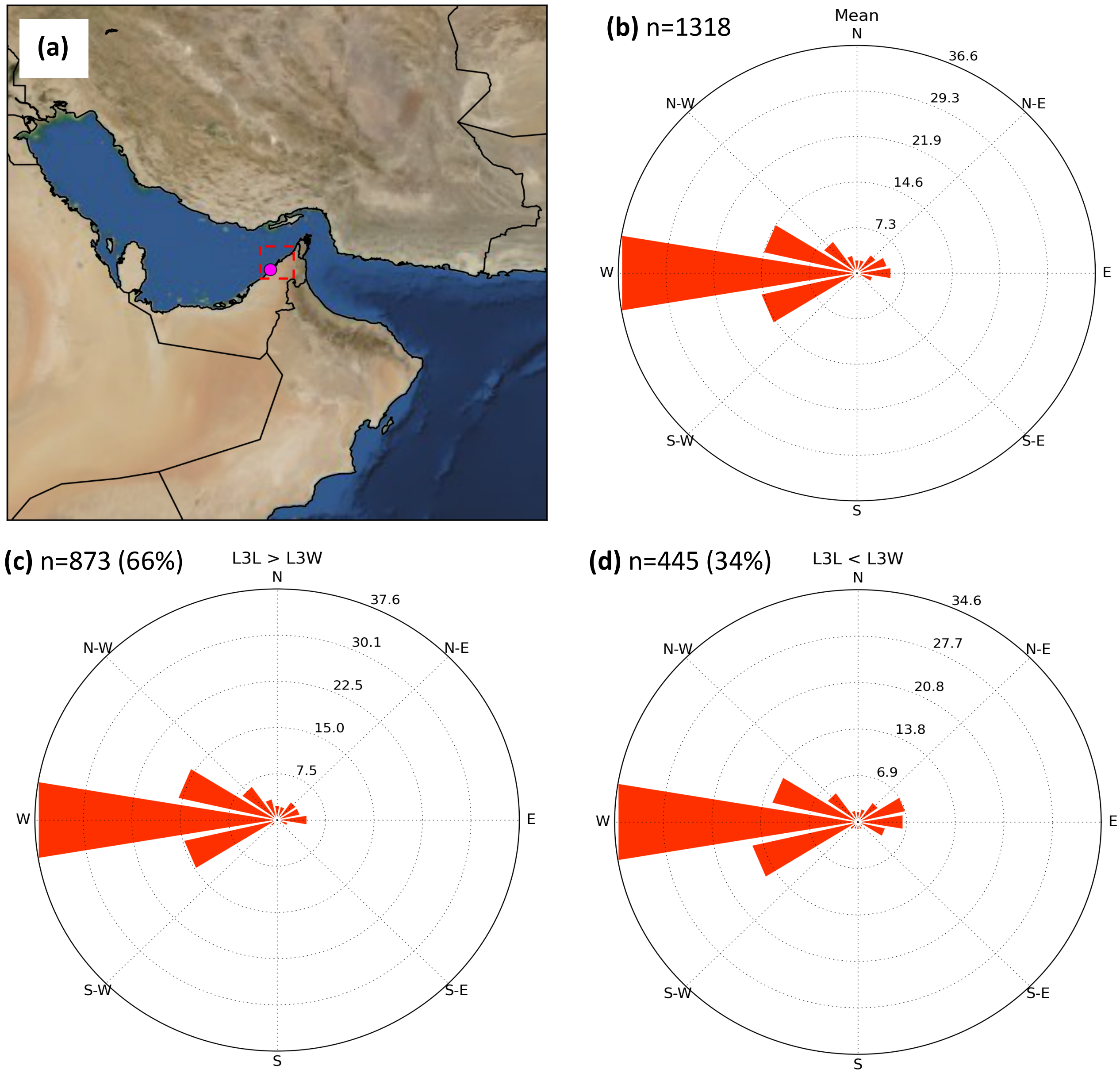

Figure R1: Analysis of surface wind direction for days when there is a surface level VMR retrieval over both land and water in the L3 grid box containing Dubai ("L3L" and "L3W", resp.) Wind direction data are calculated from u and v vector components at 10-m above the surface, taken from daily mean ERA Interim data. Wind rose barbs in panels b-d depict the direction in which wind is blowing *from* – e.g. predominantly *from the west* in all cases shown. Note that only the sub-period 01-01-2002 to 31-08-2017 is considered in this analysis (as opposed to the full study period of 2001-08-25 to 2019-02-28 in the submitted manuscript) owing to local availability of ERA Interim data. **(a)** NASA Blue Marble image of the region surrounding Dubai. The boundaries of the L3 grid box containing the city are shown by red-dashed lines, with the city location indicated by the pink marker. **(b)** Wind rose showing wind direction taken from the grid box containing Dubai in ERA Interim data. Wind direction data are taken for all days with L3L and L3W data for this grid box for the study period, and the wind rose displays the mean wind direction for all these days. Number of days represented is given by n value in panel label. **(c)** As b, but only for days when retrieved surface-level VMR is greater in L3L than L3W ("L3L > L3W"). **(d)** As b, but for days when L3L < L3W.

110 "It cannot be overlooked that working with L3 data thus requires fewer computing resources and less technical proficiency"

Agreed. Furthermore gridded products are accessible by many more tools that users are familiar with such as Panoply.

Agreed. We have added mention of the availability of simple-to-use tools for working with gridded products such as MOPITT L3:

> size (~25 MB vs ~450 MB respectively, for a single daily, global file). It cannot be overlooked that working
>
> 4
>
> with L3 data thus requires fewer computing resources and less technical proficiency, with a range of simple-to-use tools available for working with gridded products. L3 products thus make the MOPITT data more easily accessible, especially to less-expert users, who may lack the expertise required to scrutinize the data

177 "Validation results are comparable to V8. It is expected that the main conclusions of this paper to hold for V9, since the land-water sensitivity contrast remains and L3 processing method appears to be unchanged."

Actually, V9 discards far fewer L2 pixels due to cloudiness than V8 which may affect the results. I suggest the authors repeat some experiments with V9 to confirm their assumption. Yes, they are correct that the L2 → L3 processing method is unchanged.

Thank you for confirming that the L2 → L3 processing method is unchanged.

Regarding the potential for results to be different in V9 than V8: we have re-done certain sections of the analysis using V9 data to verify that the main conclusions presented in our submitted manuscript based on V8 data hold for V9 too. Results of this analysis are outlined below. We hope it is appreciated that re-doing the entirety of the analysis is not practical (or reasonable) given that the original work was undertaken before V9 was released. We have included these results in the supplementary material (SM1), and reference these where V9 is outlined:

> overview of MOPITT V9 is given by Deeter et al (2021). ~~It is expected that the main conclusions of this paper to hold for V9, since the land-water sensitivity contrast remains and L3 processing method appears to be unchanged.~~ A subset of the analysis presented in this paper has been duplicated using V9 data, and this confirms that the main conclusions drawn based on V8 data also hold for V9 (this analysis is outlined in the Supp. Mat. (SM1)). This is to be expected, given that the land-water sensitivity contrast remains in V9 and the L3 processing method is unchanged. ~~An overview of MOPITT V9 is given by Deeter et al (2021).~~

Results of analysis with V9:

1. The global land-water sensitivity contrast shown in Figure 3 (V8) is also present if Figure 3 is re-plotted using V9. This confirms that the land-water sensitivity contrast remains. Both Figures are shown in Figure R2, below.
2. The land-water sensitivity differences within coastal L3 grid boxes (demonstrated by comparing L3L and L3W for these grid boxes), shown by the scatterplots in Figure 4 (V8), is also present if the analysis is reproduces using V9 and Figure 4 replotted with these data. Both Figures are shown in Figure R3, below.
3. We repeated the comparison of a) mean surface level retrieved VMR, and b) temporal trends therein, for selected grid boxes containing large cities analysed in Sect. 3.4 of the submitted manuscript, using V9 data. Results for both V8 and V9 are shown in Table R1, below, and similar differences to V8 exist in V9. Although this analysis is restricted to L3L and L3W only, given that the L2 → L3 processing method is unchanged there is every reason to expect that similar differences would emerge for L3O V9 subsets too.

(Note that due to time and data storage limitations, the V9 analysis was restricted to the data years 2010-2015 inclusive for results 1 and 2 above. Given the clarity of the results however, we are confident that this would remain if the whole study period considered for V8 had been reanalysed using V9. L3L and L3W time series for the full period studied using V8 were able to be obtained for V9 for the analysis leading to result 3 above).

Table R1. Mean retrieved surface level VMR, and its temporal trend, in L3L and L3W, for L3 grid boxes containing selected cities from Sect. 3.4 of the submitted manuscript, using V8 and V9.

		V8			V9			V9 conclusion same as V8?
		L3L	L3W	Δ (L–W)	L3L	L3W	Δ (L–W)	
Bangkok	Mean	314.4	261.3	53.1	286.4	268.3	18.1	Y
	Trend	-3.03	-2.00	-1.03	-2.93	-2.21	-0.72	Y
Dubai	Mean	180.0	163.3	16.7	186.6	168.8	17.8	Y
	Trend	-2.90	-0.90	-2.00	-2.97	-1.06	-1.91	Y
Hong Kong	Mean	336.1	260.1	76.0	307.2	270.5	36.7	Y
	Trend	-8.06	-3.55	-4.51	-7.50	-4.37	-3.13	Y
Miami	Mean	160.7	143.5	17.2	158.7	149.8	8.9	Y
	Trend	-1.52	-0.75	-0.77	-1.44	-1.08	-0.36	Y
Sydney	Mean	94.0	86.8	7.2	89.2	88.6	0.6	Y
	Trend	-0.74	-0.24	-0.50	-0.58	-0.42	-0.16	Y
Toronto	Mean	238.4	254.5	-16.1	240.7	255.9	-15.2	Y
	Trend	-1.09	-1.99	0.9	-1.71	-1.87	0.16	Y

(a – V8)

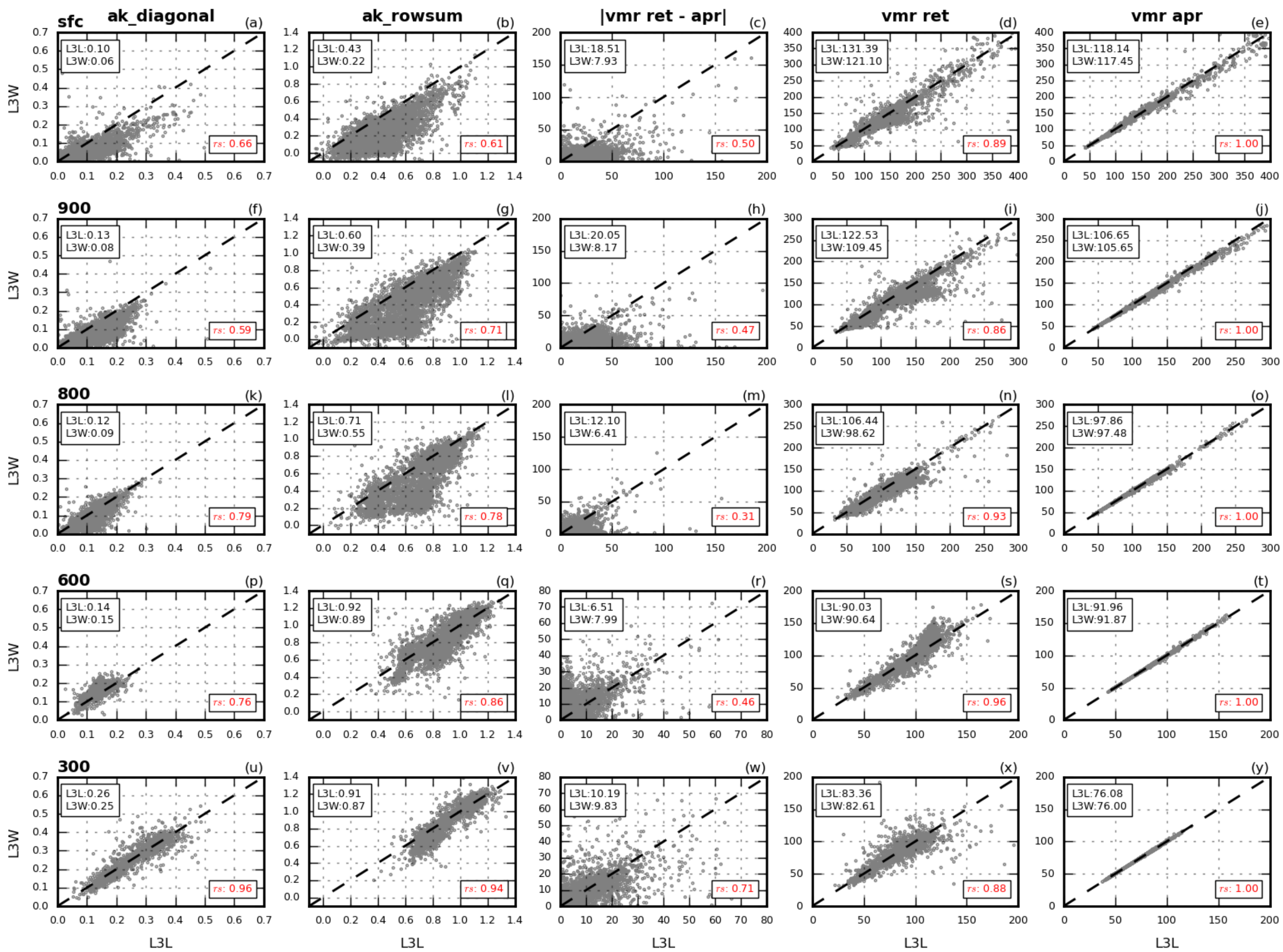

Figure R2. (a) Mean sensitivity metrics and VMRs (retrieved and a priori) from coastal L3 grid boxes, from MOPITT V8 data. Values compared in the scatterplots are mean values from matched L3L and L3W retrievals within these grid boxes. "Matched" means that only days when both L3L and L3W are present, and the L3O surface index is mixed, are used to create the mean values analysed. Shown are AK diagonal values (left column), AK rowsums (second column), absolute VMR retrieved minus a priori values (third column), retrieved (fourth column) and a priori (fifth column) VMRs, for the following levels of the retrieved profile: surface (top row), 900 hPa (second row), 800 hPa (third row), 600 hPa (fourth row), and 300 hPa (bottom row). Values in boxes in the top-left corner of each panel correspond to mean values across all L3L and L3W grid boxes. NOTE: This is a reproduction of Figure 4 from the submitted manuscript.

Figure R2 (b) shown on next page

(b – V9)

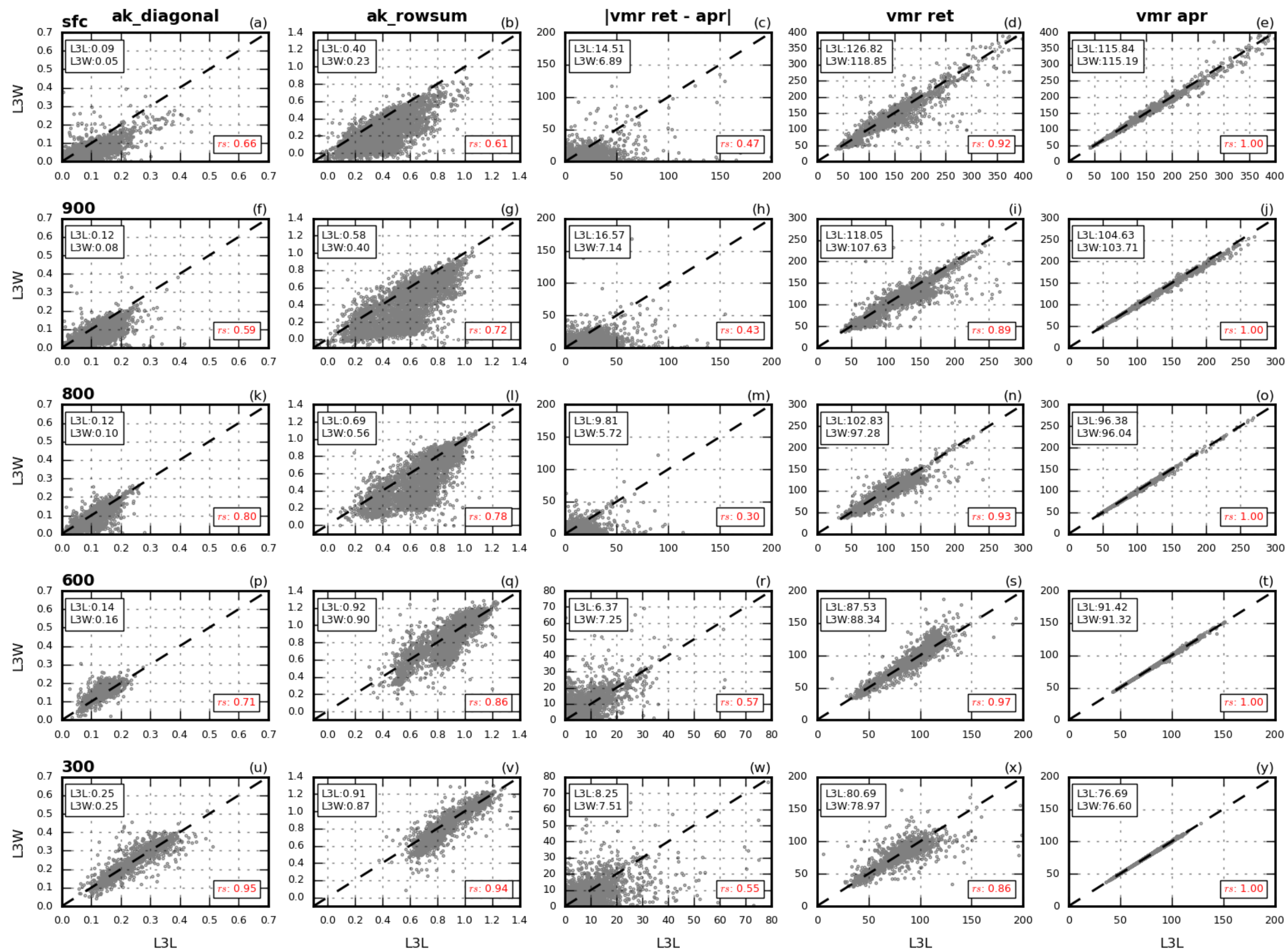

Figure R2. (b) As Figure R2 (a) except using MOPITT Version 9 data and for the sub-period 2010-2015 (inclusive), as explained in response text above.

(a – V8)

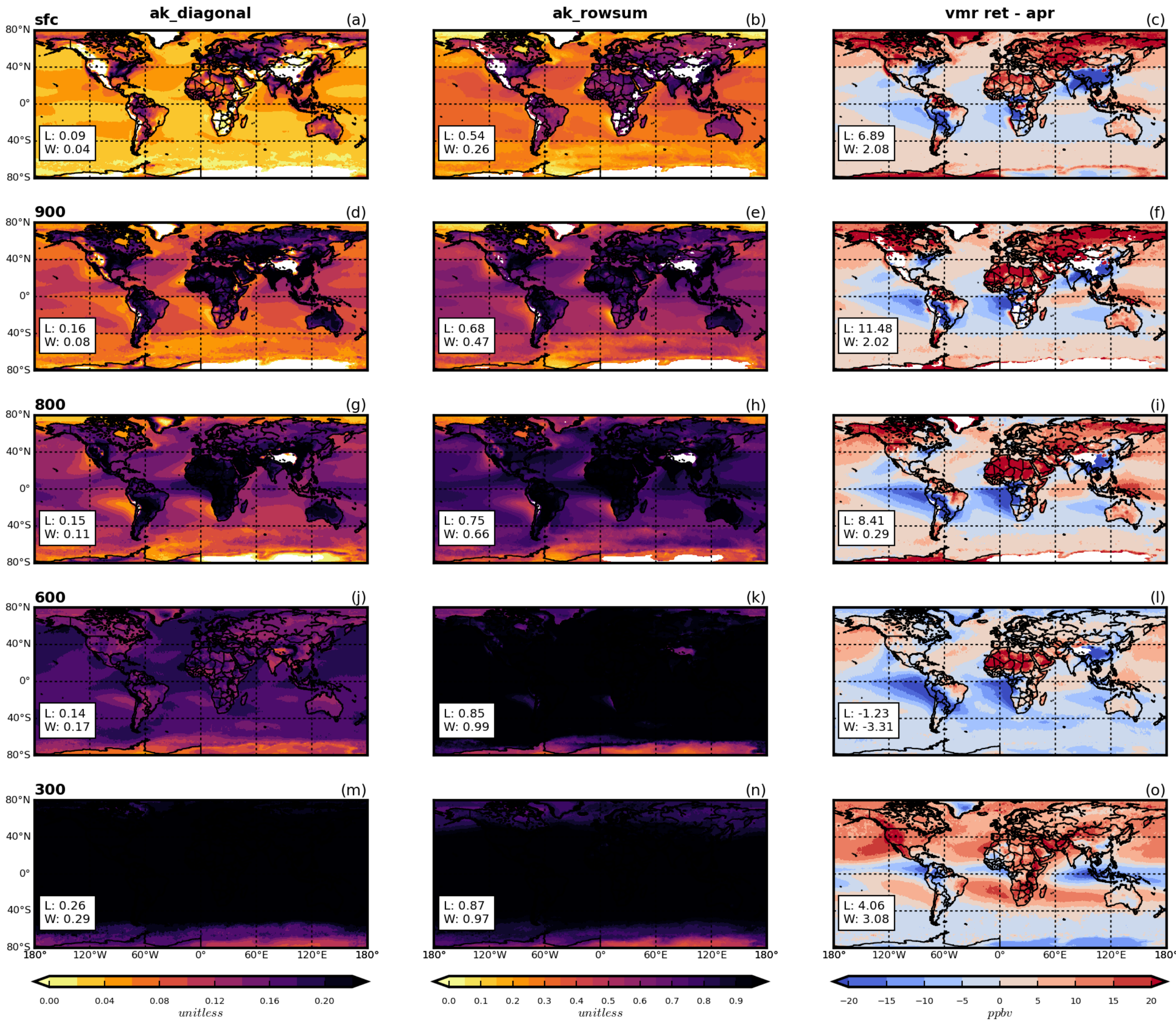

Figure R3. (a) Mean sensitivity metrics from MOPITT L3 data (Version 8), averaged across the entire study period (September 2001 – February 2019, inclusive). Shown are AK diagonal values (left column), AK rowsums (center column) and VMR retrieved minus a priori values (right column) for the following levels of the retrieved profile: surface (top row), 900 hPa (second row), 800 hPa (third row), 600 hPa (fourth row), and 300 hPa (bottom row). Values in white boxes correspond to mean values across all land ("L") and water ("W") L3 grid boxes. NOTE: This is a reproduction of Figure 3 from the submitted manuscript.

Figure R3 (b) shown on next page

(b – V9)

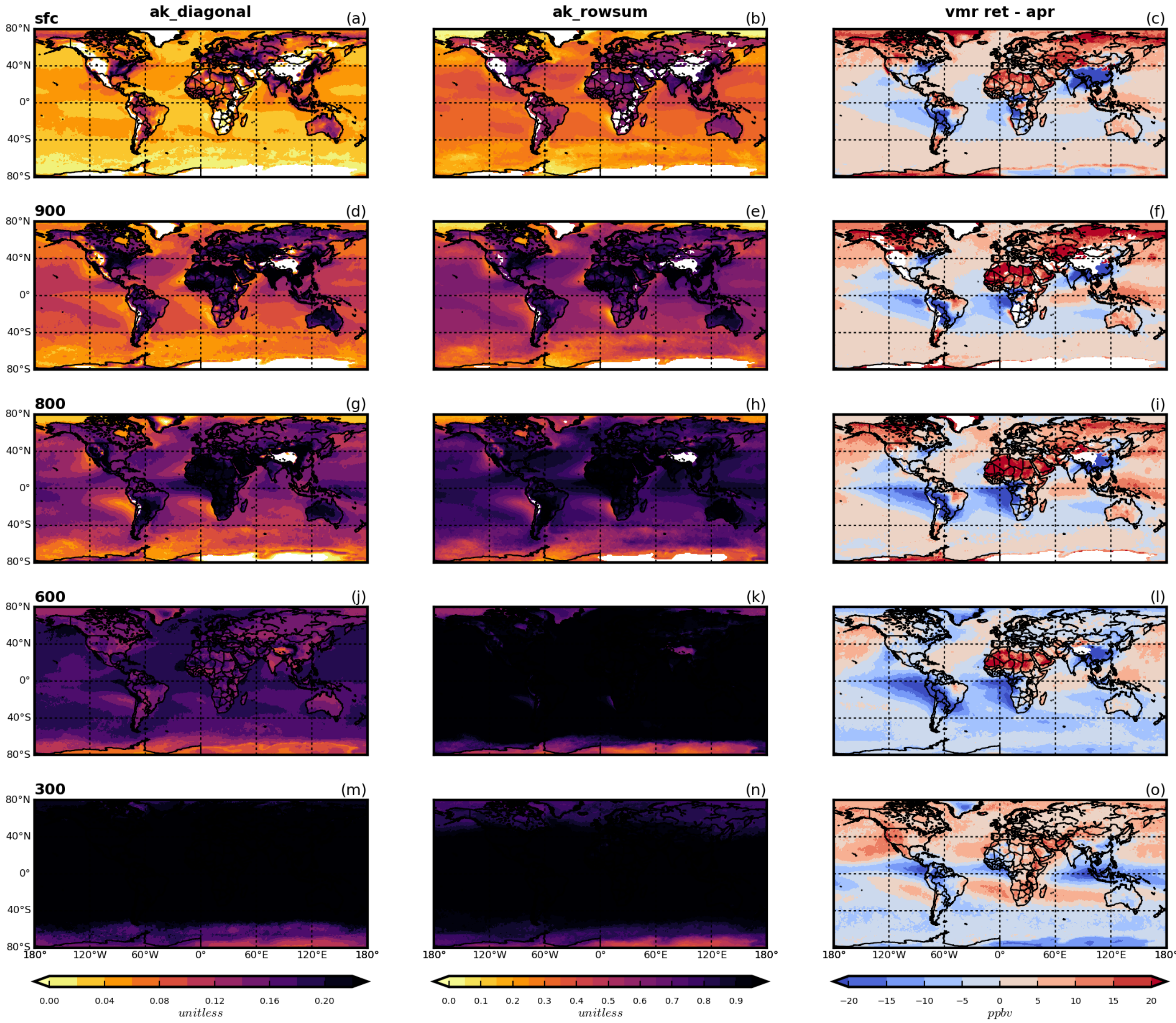

Figure R3. (b) As Figure R3 (a) except using MOPITT Version 9 data and for the sub-period 2010-2015 (inclusive), as explained in response text above.

193 "which at the time of writing, is the most recent data quality summary"

More recent data quality statements are available now. See https://asdc.larc.nasa.gov/documents/mopitt/mopitt_quality_statements.html

We thank the reviewer for pointing this out. However, our reference needs to remain to the data quality statement for Version 6 data, because none of the more recent data quality summaries explain how the Level 3 surface classification is derived (which is what our reference to the data quality statement is for). We have clarified this in the paper:

> 1 (this information is taken from the MOPITT Version 6 L3 data quality summary[1], which at the time of writing, is the most recent data quality summary to detail exactly how L3 data are created, despite more recent data quality summaries being available). Note that the L2 VMR profiles that are averaged to produce

483-485 "However, the results presented do imply a general tendency for trend underestimation in retrievals over water within coastal grid boxes compared to retrievals over land in the same grid boxes obtained at the same times, which appears to be linked to differences in retrieval sensitivity."

This feels like the most important point of the paper. Perhaps more effort to demonstrate and quantify would be helpful.

This point is explicitly discussed in Section 3.2.1, with trend differences between L3L and L3W demonstrated and quantified in Figure 5c-d and Table 2.

To help with the presentation and communication of our results, we have undertaken a thorough edit of the manuscript, where the focus has been on clarifying the points being made and, where possible, moving methodological details (which are flagged as a source of confusion by the reviewer elsewhere in this review) either to the end of paragraphs as caveats to the points, or to figure captions. The most comprehensive edits are in the Results and Discussion Section (Section 3).

We include a screenshot below of the subsection containing the above quote, to demonstrate how the point being made is now more clearly framed in the revised manuscript following the edit.

~~All together~~In summary, these results show a general tendency for trend underestimation in surface level retrievals over water compared to retrievals over land in the same coastal grid boxes obtained at the same times, which appears to be linked to differences in retrieval sensitivity.~~show that within coastal L3 grid boxes, differences in retrieval sensitivity over land and water are related to differences in temporal trends identifiable in corresponding surface level retrievals.~~ The relationships found in these analyses are not perfect because trend differences are sensitive to several other factors, in addition to differences in retrieval sensitivity. For example, a greater trend difference would be evident if the rate of change in "true" CO concentrations is faster than if it is slow/negligible, for a given sensitivity difference. Similarly, there should be zero trend difference if "true" CO concentration levels are stable over time, irrespective of the magnitude of difference in retrieval sensitivity. The accuracy of the a priori is a further complicating factor. ~~However, the results presented do imply a general tendency for trend underestimation in retrievals over water within coastal grid boxes compared to retrievals over land in the same grid boxes obtained at the same times, which appears to be linked to differences in retrieval sensitivity.~~ An underlying assumption is also that the temporal trend in "true" VMRs should not vary much across a 1° x 1° L3 grid box. Hedelius et al. (2021) lends credence to this assumption with the finding that CO trends are similar within regions spanning a few thousand kilometres (L3 grid boxes are ~ 100 km^2), and that trends within urban areas are generally indistinguishable from the trend of the broader region encompassing the urban area, despite an expectation that urban trends should exceed the regional background due to a concentration of CO emission sources here.

The discussion in the paragraph starting on line 417 is very important however it would have even more impact if it it included the consideration of one more bit of information. The skill of MOPITT retrievals of CO (VMR) is not random. It is dependent on conditions such as thermal contrast between the surface and the air, which is what the authors are describing when they see discontinuities between L3L and L3W. However another factor is that MOPITT sees CO better when there's a lot of it. The uncertainty (as measured by DFS) decreases when the CO signal is large. So if there's less CO over the ocean due to fewer sources, that will also affect the results of this analysis.

This comment has mostly been addressed above, in response to the reviewer's comments on L23, L421 and L426. To re-iterate our response: it is true that CO emissions are negligible over ocean compared to land, however the potential impact that this will have on the results of our analysis appears weak in the face of:

1. The portion of coastal L3 grid boxes situated over water being unlikely to represent "pristine" background conditions owing to the well-mixed and relatively long-lived nature of atmospheric CO. i.e. the CO loading for a given grid box is not simply a result of emission from within that grid box;

2. There being a large portion of coastal L3 grid boxes where emissions from the land-based component are also negligible, due to the lack of large anthropogenic CO sources or natural emission hotspots;
3. There being cases where retrieved surface-level CO concentrations in L3L are less than in L3W for a given L3 grid box, and we would not expect to see this if the differences are simply related to differences in emissions.
4. CO Trends detected over cities in MOPITT data not tending to stand out from the regional average despite the presumption of greater emissions there (Hedelius et al., 2021).

It therefore seems unlikely that the L3L-L3W CO concentration and trend differences presented for coastal grid boxes are strongly impacted by land-water emission differences within these grid boxes – especially against the weight of evidence that links these to well-understood retrieval sensitivity contrasts, as the review themselves outline. However, we have included a more thorough discussion of this point in the text (Section 3.2.1 – see response to comments on L23, L421 and L426 above).

Additionally, we do mention in the text (introduction section) that retrieval sensitivity is also linked to the amount of CO present:

> these a priori CO profiles are based on a monthly climatology from a chemical transport model. The degree to which a given MOPITT retrieval reflects information obtained from the observed radiances – known as "information content" – is highly spatially and temporally variable, depending on scene-specific factors such as surface temperature, thermal contrast in the lower troposphere, and the actual ("true") CO loading itself, as well as on instrumental noise (e.g. Deeter et al., 2015). The lower the retrieval information content, the closer the retrieved CO loading will be to the a priori; a model value.

451 – The word "gradient" appears but I'm having trouble understanding its definition here. Is it just the difference between temporal trends of L3L and L3W? Or is it spatial?

Apologies for the confusion – we were referring to "trend", which is the word that should have been used in the first place for clarity and consistency with the rest of our wording. Note that this sentence has been removed in the revised version of the manuscript, following a thorough edit of the text for clarity.

Table 1 – It took me several attempts to understand what the "d" column was.

Apologies for the confusion – we have changed the column heading and explained its meaning in the caption, for greater clarity.

Table 1. Mean values for selected variables from L3L and L3W for coastal L3 grid boxes, matched retrievals only. "Matched" means that only days when both L3L and L3W are present and the L3O surface index are mixed are used to create the mean values analysed. Mean values are calculated and presented separately according to the results of a 2-tailed student's t-test (unequal variance) performed on mean retrieved VMR values in L3L and L3W (n = 3971). Mean L3L – L3W differences are also shown for each subset ('L-W')

	P < 0.1 (n=2379, 60 %)			P > 0.1 (n=1592, 40 %)		
	land	water	d	land	water	d
Mean vmr_ret	129.97	117.41	12.55	133.52	126.60	6.90
Mean vmr apr	113.78	113.18	0.61	124.65	123.83	0.83
Mean ret-apr	16.18	4.24	11.94	8.87	2.77	6.09
Mean ak rowsum	0.43	0.18	0.24	0.44	0.27	0.16

	P < 0.1 ("SIGDIFF") (n=2379, 60 %)			P > 0.1 ("NOT_SIGDIFF") (n=1592, 40 %)		
	L3L	L3W	L-W	L3L	L3W	L-W
Mean vmr_ret	129.97	117.41	12.55	133.52	126.60	6.90
Mean vmr apr	113.78	113.18	0.61	124.65	123.83	0.83
Mean ret-apr	16.18	4.24	11.94	8.87	2.77	6.09
Mean ak rowsum	0.43	0.18	0.24	0.44	0.27	0.16

483-485 "However, the results presented do imply a general tendency for trend underestimation in retrievals over water within coastal grid boxes compared to retrievals over land in the same grid boxes obtained at the same times, which appears to be linked to differences in retrieval sensitivity."

This seems like a valid conclusion based on the analysis performed. There are a lot of details about the methodology to arrive at this conclusion that confused me more than served as support for clear statements such as this.

We apologise that some of our writing caused confusion to the reviewer. As outlined in response to an earlier comment, we have undertaken a thorough edit of the text, where the focus has been on clarifying the points being made and, where possible, moving extra methodological details either to the end of paragraphs as caveats to the points, or to figure captions.

613 "In these instances, L3O would therefore seem to be misclassified."

This is a valuable insight.

820-823 "there is enough evidence to support the suggestion from Ashpole and Wiacek (2020) that an additional L3 "land-only" product, created only from averaging bounded L2 retrievals performed over land – the L3L dataset that has been analysed in this paper – would be beneficial to the research community."

This recommendation will be brought to the attention of the MOPITT science team. It will hopefully be incorporated in the archival processing version.

General Comments:

These researchers took a very close look at how the MOPITT L3 product is created and have identified a flaw in the way the MOPITT team processes pixels into coastal grid cells by mixing retrievals of uneven quality. This distorts the values reported for a non-insignificant number of gridcells. Their conclusions appear valid and robust. However, I had a difficult time following the arguments and methodolgy of the paper (1). I didn't understand why they were focused on surface level retrievals instead of higher in the atmosphere where MOPITT is considerably more sensitive (2). I was curious if they would have come to the same conclusion if they looked at MOP03M (monthly mean) products which have far less random noise and greater coverage than the daily L3O products (3). In several places, the authors were making a clear distinction between two situations and I had trouble understanding the meaning of this distinctions. For example: "For other datasets, whether the marker is filled or not, and whether the lines are solid or dash/dot, depends on the outcome of an independent, 2-tailed t-test assuming unequal variance (aka "Welch's test") against L3L: filled markers and solid lines indicate the mean is significantly different to L3L ($p < 0.1$); open markers and dash/dot lines indicate there is no significant difference to L3L." This distinction was too difficult for me to understand its significance (4).

I believe the researchers can transform this paper into a valuable analysis by having a more clarified statement of their conclusions and focusing the readers' attention on the evidence that supports that point.

We thank the reviewer for their time and thoughts on this paper, helpful suggestions for improvement, and their positive comments. We have addressed their specific comments above, but note four more points from this summarising paragraph which we address in turn below: (the numbers 1-4 below correspond to the numbers we have added in parentheses to the general comment above, to highlight the separate points being made and make our response clearer).

(1) To make the arguments and methodology clearer and easier to follow, we have undertaken a thorough edit of the text, where the focus has been on clarifying the points being made and, where possible, moving methodological details either to the end of paragraphs as caveats to the points, or to figure captions. We have also clarified our methods further in Section 2 (e.g. outlining the different time series being analysed in different parts of the results section).

(2) As stated at the end of Section 3.1.2 (L383-385), this paper focuses on the surface-level of the retrieved profile, since the LT is where discrepancies are greatest, and the cause of this sensitivity contrast is well established (as outlined in the introduction). The surface-level is also of most

interest for identifying potential air quality impacts for humans (e.g. Buchholz et al., 2022). We have added this second justification to the text:

> This analysis clearly shows how L2 retrievals that are averaged together to create the L3O data over coastal grid boxes have differing degrees of sensitivity, especially in the LT. This is explicitly cautioned against in the MOPITT data user's guide (MOPITT Algorithm Development Team, 2018). The remainder of this paper focuses on the surface-level of the retrieved profile, since the LT is where discrepancies are
>
>
>
> greatest, and the cause of this sensitivity disparity is well established: differing thermal contrast conditions near to the surface over land and water; and a lack of NIR radiances being used in the retrieval over water. Furthermore, the surface-level is of most interest for identifying potential air quality impacts for humans (e.g. Buchholz et al., 2022).

(3) MOP03M products are unsuitable for this analysis as there is not a corresponding L2-monthly product to co-analyse. However, there is every reason to expect that the conclusions would be the same, since the MOP03M products are created from the L2 retrievals that we analyse in this paper. We have added a note clarifying this in the text, at the end of Section 2.4 where the products we analyse are outlined:

> Note that the analysis presented in this paper is restricted to daily products. Monthly L3 files are available, however the absence of a monthly L2 product precludes the analysis from being conducted on those data. Based on the results of the analysis of daily data, however, there is reason to also advise caution if working with coastal grid boxes in the monthly L3 product. This is because the data for those grid boxes will still be created from daily L2 retrievals over land and water, with the same implications that are discussed in this paper.
>
> **2.5. Timeseries preparation, sStatistical methods ~~used for this study~~, and additional data sources**

(4) Again, we apologise that some of our writing caused confusion to the reviewer. As outlined previously, we have undertaken a thorough edit of the text to address this, and hope that the meaning of distinctions such as that outlined is now clearer. Regarding the situation that is outlined, the text now makes clear that L3L is the dataset that others are being compared to.

REVIEWER # 2

Received and published: 21 May 2022

The manuscript describes a study of MOPITT V8 TIR-NIR surface CO retrievals over 33 coastal cities. Daily L3 data (data gridded to $1^o x 1^o$, 111x111 km^2 per pixel) and daily L2 data (22x22km^2 per pixel at nadir) are analyzed. This study's main findings are that statistics of coastal cities obtained from L3 and L2 products differ, that "mixed" L3 pixels (L3 pixels averaging both water and land L2 pixels) are not suitable to study coastal cities, and that a L3 land only product for coastal pixels is needed. In order to demonstrate these points, several comparisons and statistical analyses between land and water L3 TIR-NIR pixels (original and re-created from L2 data) are performed. The manuscript is well written.

Two major issues are described below.

1. Use of TIR-NIR data in land/water comparisons

As described in Deeter et al., 2013, among others, TIR-NIR retrievals over land and over water are fundamentally different, since NIR radiances cannot be used in the latter. The authors acknowledge the fact that retrievals over water are limited to the TIR band due to the lack of NIR signal, but don't acknowledge the implications, which are key. Using the TIR-NIR product for this study is not appropriate, since there are two effects causing land/water differences in the averaging kernels: thermal contrast effects and the lack of NIR radiances in retrievals over water. The two effects cannot be separated.

We agree with the reviewer that NIR lacking over water is a potential problem for comparing retrievals made over land and water in TIR-NIR joint products. However, far from making the TIR-NIR joint product inappropriate for this study, this actually strengthens the case for why L/W should not be averaged together in coastal L3 grid boxes in these products, e.g. if the lack of NIR radiances in the retrievals over water makes them so fundamentally different to retrievals over land, then they should not be averaged together. We feel that this emphasises the need for a study like this, which emphasises the consequences of mixing these retrievals, to be published.

For greater clarity, we have added a statement emphasising that a lack of NIR radiances hampers the retrieval over water in the Introduction section; and we reiterate at the end of Section 3.1 that NIR radiances are not used in the retrievals over water and therefore also contribute to the L-W sensitivity disparity, in addition to near-surface thermal contrast differences:

- Screenshot showing addition to Introduction section:

Retrievals that take place over water are known to have a lower information content than retrievals that take place over land. ~~This is~~ Primarily, this is due to weak thermal contrast near to the surface hampering the instrument's ability to sense CO absorption in the lowermost layers of the troposphere (Deeter et al., 2007; Worden et al., 2010), and this is confounded by a lack of NIR reflectance over water, which limits these retrievals to TIR wavelengths only. It is therefore recommended that MOPITT data users exclude these

- Screenshot showing addition to Section 3.1:

This analysis clearly shows how L2 retrievals that are averaged together to create the L3O data over coastal grid boxes have differing degrees of sensitivity, especially in the LT. This is explicitly cautioned against in the MOPITT data user's guide (MOPITT Algorithm Development Team, 2018). The remainder of this paper focuses on the surface-level of the retrieved profile, since the LT is where discrepancies are

19

greatest, and the cause of this sensitivity disparity is well established: differing thermal contrast conditions near to the surface over land and water; and a lack of NIR radiances being used in the retrieval over water. Furthermore, the surface-level is of most interest for identifying potential air quality impacts for humans (e.g. Buchholz et al., 2022).

2. Use of L3 data to study coastal cities

(1) *L3 products (either TIR-NIR or TIR) are not suited for the analysis of the coastal cities listed, given the horizontal extent of the targets. A cursory search (please see Table 1 attached) shows that 30 of the 33 cities in Fig. 9 correspond to a very small fraction of a single L3 pixel footprint. Only 3 of the 33 cities are close to covering or barely cover one L3 pixel footprint. Basing such analysis on L2 data could be an*

adequate choice, at least for some of these cities. About half of the 33 cities would not even fill the footprint of a single L2 retrieval. Only 10 of the 33 cities would fill 4 or more L2 retrieval footprints.

(2) According to the manuscript, "L3 data are better suited to long timeseries analysis than L2 data owing to their smaller size". That statement is wrong. Some tools are easier and more convenient to use than others, but that does not mean that they are better suited for a given task. Analyzing long time series with L3 data may be easier, more convenient. However, easy and convenient generally comes at a cost, in this case the quality of the analysis. The manuscript continues "working with L3 data [...] requires fewer computing resources and less technical proficiency [...] L3 products thus make the MOPITT data more easily accessible, especially to less-expert users, who may lack the expertise required to scrutinize the data for potential a priori bias." Again, a tool may be easy/convenient to use but unfit for certain tasks.

(3) Time series are at the center of this work and are the justification provided for using L3 data in the first place. The manuscript, however, does not include a single time series. It's hard to imagine that meaningful information/trends can be identified in L3 time series covering a ~6400 days range (from 25 Aug 2001 to 28 Feb 2019) but having only a few hundreds of even a few tens of days with a L3 value at all (and that L3 value coming in all cases from a single L3 pixel). This is the case for most of the cities analyzed (Fig. 9).

(4) Are those few hundreds of even few tens of L3 data points representative of the 1°x1° areas they stand for? L3 pixels (land, water, or mixed) may be produced by averaging as little as 2 L2 pixels. As an example: more than 25% of the total number of daytime pixels in a randomly selected L3 file resulted from averaging either 2 or 3 L2 measurements. These L3 pixels may not be representative of the 1°x1° area they stand for and, thus, should be filtered out so as not to corrupt the statistical results. It is unclear if such filtering was applied.

There are several different points made above, and we address each of these separately below (the point numbers below correspond to the numbers that we have inserted in parentheses above):

1. We have clarified in various points throughout the text (notably the Abstract, end of the Introduction, Methods, throughout Section 3.4 where "cities" are analysed, and in the Summary and Conclusions) that our focus is on *the grid boxes containing the cities*, as opposed to the cities themselves. These grid boxes are chosen for illustrative purposes, to demonstrate the impact of choosing to analyse L3 vs L2 data, given that such grid boxes are likely of interest for users of the data interested in e.g. air pollution and human health (whether using L3 or L2). Note that the focus on grid boxes containing cities only makes up one subsection of four in our analysis (Section 3.4):

we focus on *all* coastal grid boxes in Sections 3.1 – 3.3. To address the argument about L3 being inappropriate as the grid boxes are larger than city extents: CO is long-lived and therefore well-mixed horizontally around cities. Hedelius et al. (2021) studies this and shows that trends for cities do not stand out against the background on a scale of a couple of degrees. As outlined above, we have made additional clarifications throughout the text that we are studying the L3 grid boxes that contain these cities, not the cities themselves. If our analysis was restricted to L2 data that fall within city limits, as seems to be suggested, this would result in even fewer data days to base our analysis on, which is something that the reviewer raises as an issue later.

- Screenshot showing clarification RE grid boxes containing coastal cities in Abstract:

cities, we ask whether results of analyses are significantly different if using L3O compared to L3L. I~~i~~t is shown that mean VMRs in $L3O_L$ and L3L differ significantly for 11 of the 27 ~~cities~~ grid boxes that can be compared (there are no $L3O_L$ data for 6 of the grid boxes studied~~cities~~). The L3L – $L3O_{LM}$ mean VMR difference exceeds 10 (22) ppbv for 11 (3) of the 33 ~~cities~~grid boxes, significant in 13 cases. 9 of the 18 ~~cities~~ grid boxes where WLS analysis can be performed in $L3O_L$ feature a trend that is significantly different to L3L. The trends in $L3O_{LM}$ and L3L differ significantly for 5 of the 33 ~~cities~~grid boxes. It is concluded that a

- Screenshot showing clarification RE grid boxes containing coastal cities in Summary and Conclusions section:

• Focusing on the L3 grid boxes containing the 33 largest coastal cities in the world, mean VMRs in $L3O_L$ and L3L differ significantly for 11 of the 27 ~~cities~~ grid boxes that can be compared (40 %; there are no $L3O_L$ data for the remaining 6 cities). The L3L – $L3O_{LM}$ mean VMR difference across all 33 ~~cities~~ grid boxes is relatively small (3.7 ppbv), but this does hide some much larger discrepancies, with the difference exceeding 10 ppbv for 11 of the 33 ~~cities~~ grid boxes and 20 ppbv for 3 of them. The difference is significant for 13 of 33 ~~cities~~ grid boxes (39 %). Of the 18 ~~cities~~ grid boxes where WLS analysis can be performed in $L3O_L$, there are 9 cases where the trend is

54

significantly different to that in L3L. The trends in $L3O_{LM}$ and L3L differ significantly for 5 of the 33 ~~cities~~grid boxes

2. We accept the reviewer's point that our statement is wrong with respect to L3 data being "better suited" to long time series analysis. We have rephrased this to "L3 data are more convenient for long time series analysis than L2 data owing to their smaller file size" (see screenshot below). Regarding whether or not they are fit for the task: we contest that in most cases *except* for coastal grid boxes, L3 data *are* fit for the task of long time series analysis. This is because, as we outline, they are simply gridded area averages of bounded L2 data. Whether users prefer to analyse L2 or L3 data is then a trade-off between required spatial resolution, temporal coverage, ease of analysis, etc. In addition, we refute the reviewer's judgement that "easy and convenient generally comes at a cost"; one of the key objectives of data provision (in our opinion) should be to minimise barriers to using those data in order to maximise the scientific and societal use that they (and the investment made in acquiring them) can have. If the "quality of the analysis" is one of the costs then there is a problem with the way that the data are being delivered, and this is exactly what our paper seeks to solve.

difference that this would make on a global scale. This is necessary to understand for two reasons: firstly, L3 data are ~~better~~ more convenient for~~suited to~~ long timeseries analysis than L2 data owing to their smaller file size (~25 MB vs ~450 MB respectively, for a single daily, global file). It cannot be overlooked that working

3. The issue raised about patchy temporal coverage is unfortunately a broader problem when it comes to analysis of data acquired by satellite instruments, including MOPITT. Beyond selecting a broader averaging area (e.g. a 3 x 3 grid of L3 grid boxes), there is nothing that can be done to improve the temporal coverage in the V8 dataset. The situation is worse for L2 data (unless they averaged across a broad area, which then defeats the purpose of using L2 data). Evaluating whether the presented trends are "meaningful" requires additional verification datasets and is beyond the scope of this study, since the focus of our trend analysis is instead to demonstrate the difference in trends that can be detected using L2 vs L3 data. This was already stated in Sect. 2.5, where our trend detection method is outlined. We have now added that additional data would be needed to verify the trends if the trend values themselves were our focus, and we have reiterated this caveat at the end of Sect. 3.4.3 where trends are discussed for the case study grid boxes containing large coastal cities. Screenshots from both these sections are shown below.

- Screenshot from Section 2.5 showing clarification that the aim of the trend analysis is for dataset comparison only, and that additional data would be needed for verification:

> preferred over OLS because it is less sensitive to outliers. For simplicity, no other trend detection methods – e.g. the Thiel-Sen slope estimator – are applied to corroborate the trends that are detected with WLS, nor do we analyse additional datasets to verify them. Such extra steps would be necessary if the actual trend values were the focus of this study; however, the aim of this trend analysis is instead to identify whether the same method can yield different results depending on which of L3O, L3L or L3W is analysed.

- Screenshot from Section 3.4.3 showing clarification that the trend analysis is illustratory only, and that additional data would be needed for trend verification:

> As outlined in Sect. 2.5, it is important to note that the trends presented in this section are for illustratory purposes only, with the intention of demonstrating that different results can be obtained depending on whether L3O or L3L (and, by extension, L2) data are analysed. More focused analysis is needed to verify these trends, which is beyond the scope of this paper.

4. We do not filter our time series based on the number of L2 retrievals that are averaged together to create the L3 data which we are analysing. We do this because a) doing so would further limit the temporal sampling of the products analysed; and b) no such filtering is performed in the creation of the L3O product, available for public download, which is the focus of this study. To the reviewer's point about the representativeness of a L3 retrieval that is only created from a small number of L2 retrievals within the L3 grid box; this misses the point of our study. We are comparing the L2 retrievals performed over land and water that are averaged together to create L3O products for coastal grid boxes, and arguing that the retrievals over water shouldn't be used as they can be very different to those over land, with retrieval sensitivity differences an explanatory factor. Whether the retrievals being compared are representative of the grid box as a whole is not the focus of our study. Assuming a coastal L3 grid box has data for a given day in the as-downloaded L3O product, we compare the bounded L2 retrievals to identify whether there is evidence of a systematic difference between those made over land and water – and ultimately to demonstrate that the retrievals over water shouldn't be used when creating the L3O product.