Many thanks to Referee #1 (Corwin Wright) for appreciating our work and the very helpful comments that will significantly help to improve the manuscript!

Please find below our point-by-point reply to the reviewer concerns. Comments by Reviewer #1 are given in red, our reply is given in black, and changes in the manuscript are indicated in blue.

Reply to the Main Concerns by Reviewer # 1:

(Main Comment 1:) I do have a probable answer to one question raised in the manuscript. In several places (e.g lines 532-543, lines 657-664, line 730) the authors highlight that their results appear quantitatively inconsistent with a previous study (W2013a, reference below). Specifically, the wave intermittencies measured in the current study are consistently higher than those seen in W2013a. However, I believe this arises from a key methodological difference. In W2013a, we used an approach (described by WG2013, reference below) which selects for multiple waves in a given HIRDLS measurement, and which typically (see W2015, reference below) identifies four discrete waves at any one measurement location. In contrast, the current study identifies at most a single wave at each point (L173). These 'additional' waves in W2013a tend to be lower-amplitude and have smaller momentum fluxes (WG2013, W2015) than the 'main' waves measured by the method used here, and will hence tend to strongly pull intermittency values down. I think this methodological difference is likely to explain most if not all of the differences between the current study and W2013a.
I emphasise that this difference is not an mistake in experimental design in the current manuscript and I have no objections to the authors using a more cautious approach and identifying at most one wave as they do here - it is a perfectly valid choice. The WG2013 method has the advantage that it detects more smaller waves from the same data, but in particularly for HIRDLS can be negatively affected by a known fault with the instrument which will introduce a small height-varying population of nonexistent waves into the data (W2015) and hence would have to be treated cautiously for a study of this type. We had not identified this problem at the time of writing W2013a, but it shouldn't affect the results presented there too much as the effect is very small at the low altitudes that study focuses on, and even at high altitudes is only a few percent of the measured waves - i.e. I suspect that the differences between the current study and W2013a will be almost entirely methodological rather than due to this data issue.
[W2013a] Wright, Osprey and Gille, JGR 2013: 10.1002/jgrd.50869 [WG2013] Wright and Gille, GRL 2013: 10.1002/grl.50378 [W2015] Wright, Osprey and Gille, ACP 2015: 10.5194/acp-15-8459-2015

Thank you very much for this clarification! In our manuscript we already suggested that differences might occur due to a population of small-amplitude waves that lead to a different shape of the PDFs and reduce intermittency, however we did not know the exact reason for these low amplitude waves.
We will add this explanation in the revised manuscript after former l. 532 when the difference in magnitude is first mentioned and potential reasons are given.

"The likely main reason for this difference in magnitude are differences in the gravity wave analysis technique. While in our study we focus on only the strongest gravity wave at a given altitude, the method used by Wright et al. (2013) selects for multiple waves in a given HIRDLS measurement, and which typically identifies four discrete waves at any one measurement location (Wright and Gille, 2013; Wright et al., 2015). These additional waves usually have lower amplitudes and carry small momentum fluxes. This large population of relatively small absolute momentum fluxes will considerably pull down the level of intermittency, while relative variations of intermittency should be still dominated by the largest events."

In addition, we have shortened the part summarizing other potential reasons for differences in magnitude.

Further, we will mention in the summary that a direct comparison of intermittency is only possible if similar analysis methods are used.

(Main Comment 2) The methodological choice to normalise the distributions (L238-247) does make sense when the results are considered, but probably needs a little more justification. For QBO regions, where filtering varies strongly from year to year, the logic is clear and coherent, but I am less clear on the justification for doing so at extratropical latitudes (at least outside SSW periods). If this was a minor point I would be happy with the current presentation, but since this choice underpins most of the results presented I think it needs to be justified a bit more strongly.

There are several reasons for normalizing the distributions:

- Different parameters can be directly compared using the same scales
- Normalizing distributions accounts for spatial gradients within the area considered. These spatial gradients can introduce spurious intermittency. This is particularly important for PDFs that are created from quite large regions, but also for global distributions that result from gridding using a set of lon/lat bins. For example, for creating global distributions of SABER gravity wave momentum fluxes we require relatively large lon/lat-bins for averaging because of the sparsity of data. In an appendix we now show the global distributions of Gini coefficients for SABER momentum fluxes without normalizing the distributions beforehand. These distributions are clearly biased due to spatial gradients within the bins. This effect is strongest at southern hemisphere mid to high latitudes during austral winter.
- Normalizing the distributions generally reduces the width of PDFs (intermittency) in all cases, which shows that normalization is generally beneficial.

In the revised manuscript, we have given further reasoning after former l.240 why we apply normalization. This is followed by a detailed description of the normalization procedure. Further, we have added in an appendix the distributions of Gini coefficients for SABER gravity wave absolute momentum fluxes without normalization being applied. These distributions clearly show spurious enhancements of intermittency in regions of strong spatial gradients.

(Main Comment 3) In section 5.1, I was still confused after several readings as to exactly how the gradient effect (line 444 onwards) was being compensated for - if the normalisation is taking place at the level of the bin, then how does this reduce spatial biasing due to gradients within the bin? I suspect I am misunderstanding something here, and as such would appreciate this section being made clearer.

The problem is that by forming a PDF (or quantifying intermittency in another way) one assumes that all data points considered follow the same distribution with the same mean and the same standard deviation. This, however, is clearly not the case if there are horizontal gradients caused by variations of the overall global distribution within an area considered. These variations are compensated for by normalizing the single values by the temporally and spatially varying global distribution of medians.

This reasoning will be given in the revised manuscript after former l.240.

(Main Comment 4) A minor concern I do have is that the manuscript feels very long and could probably benefit from trimming, but this is not a critical problem and the paper does contain a lot of data which does justify this. If the authors do choose to trim it, I think it would be better to do so by slimming down each section rather than removing some parts entirely, and by reducing repetition between sections.

As recommended, the manuscript has been shortened in some places.

Reply to the Additional Comments by Reviewer # 1:

(1) L131: note that the resolution of HIRDLS drops sharply above 60km, averaging 2km above this level - see e.g. Figure 5.1.1 of the HIRDLS Data Quality Document HIRDLS DQD: https://docserver.gesdisc.eosdis.nasa.gov/repository/Mission/HIRDLS/3.3_Product_Documentation/3.3. **DQD_V7.pdf**

This will be mentioned in the revised paper by adding:

"(∼2 km above 60 km altitude)"

(2) L148: what kind of high-pass filtering is being applied here? Some types could introduce small waves, which may pull measured intermittencies down.

The high-pass filtering applied here is just fitting in the vertical a single sinusoidal wave of vertical wavelength between 40 and 80km. Since this vertical wavelength is relatively long, its variations in the vertical are very moderate within the 10km vertical intervals we are using for fitting wave amplitudes.
We will add the following information for clarification.

"This high-pass is performed by fitting and subtracting a sinusoidal wave of vertical wavelength of 40 km or longer, individually for each altitude profile."

(3) L201 and other places: while it's reasonable clear that by 'average' the authors mean 'mean', the mode, median and mean are all types of average - I would suggest switching from 'average' to 'mean' throughout to avoid any confusion.

Thank you very much for this pointing this out! For clarification, we have changed the text in former l. 201 to:

"... values are multi-year means of medians, and not multi-year means of arithmetic mean values."
Further, we now use 'mean' instead of 'average', where applicable.

(4) L273: additionally, presumably SABER cannot access the smallest horizontal wavelengths due to the inter-profile spacing being about double that of HIRDLS. This is mentioned in s4.2.2 later, but this is the first mention where that effect is relevant so it might help to put it here instead.

This effect was already mentioned in former lines 274–276. For better clarification we have added:

"(about twice the along-track sampling step of HIRDLS)"

(5) L339 onwards: similar high intermittencies are seen in the open Southern Ocean using AIRS data in Hindley et al 2019 (their figure 9). While this has a very different observational filter, it may be relevant to the discussion here too.
Hindley et al, ACP 2019: doi:10.5194/acp-19-15377-2019

Indeed, it is quite remarkable that the intermittency is very similar although the instruments and their observational filters are very different! We have added another paragraph after former l. 355 that mentions the AIRS results by Hindley et al. (2019).

"It is also noteworthy that gravity wave observations by the Atmospheric Infrared Sounder (AIRS) show similar characteristics at mid and high southern latitudes (Hindley et al., 2019). Although AIRS has a very different observational filter and observes only gravity waves of vertical wavelength longer than about 12 km (see, for example, Ern et al., 2017; Meyer et al., 2018), very strong intermittency is found over South America and the Antarctic Peninsula, and somewhat weaker, but still strong, intermittency over the Southern Ocean."

(6) Section 4.2.4 - I like this!

Thank you very much for appreciating our work!

(7) L470: I understand here that you mean individual profiles (i.e. single profiles against altitude), but the text as-written could be read as referring to single levels *within* a profile (i.e. a single altitude level of a given profile) - it might help to clarify this.

Thank you very much for pointing this out! We now use the expression 'for each individual altitude profile'.

"... Based on gravity wave amplitudes, calculation of potential energies from satellite observations can be performed for each individual altitude profile, ..."

(8) L484: This normalisation depends on the averaging method used and more detail about this would help - for example, are the averaged values being stored at the grid-corner or grid-centre before being interpolated to the profile location?

As averaging method we selected the median, which is quite convenient: if the single values follow a lognormal distribution, the PDFs of the normalized distribution will be centered at zero on a log scale. The median values are stored at the grid centers, and linear interpolation in longitude and latitude is applied to obtain an interpolated value at the profile locations.

In the revised manuscript, we will add a detailed description of the normalization procedure after former l.240. See also Minor Comment (10) by Reviewer # 2, and Minor Comment 2 by Reviewer # 3.

(9) L515 onwards: could the lack of shorter vertical wavelengths and/or different normalisation areas cause this higher Gini coefficient estimate?

Indeed, the SABER lack of shorter vertical wavelengths could be a candidate for differences between SABER and HIRDLS Gini coefficients. Such effects were already covered in the manuscript by stating: "... differences in the HIRDLS and SABER sensitivity functions for detecting gravity waves."
The size of the normalization areas should not be relevant, because for E_{pot} we use relatively small lon/lat bins for both instruments.

In addition to these points, in reply to Reviewer # 3, Minor Comment # 1, we also added some discussion on potential effects of the different line of sight orientations.

(10) L544: this is remarkable given the very different observational filters involved - do you think this is a real similarity and (e.g.) that the same intermittency effects are being observed uniformly across the GW spectrum despite the very different physical scales, do you think the methods are actually observing the same waves, or do you think it's just a coincidence?

It looks like a real similarity, at least for a certain part of the gravity wave spectrum. We have added this speculation after former l.556:

“Because the observational filters of limb sounders and superpressure balloons are very different, and also the gravity wave spectrum in the simulations will be different, not necessarily the same waves are being observed. Therefore, the agreement in Gini coefficient magnitudes suggests that, at least over a certain part of the gravity wave spectrum, the statistical distributions of momentum fluxes are similar.”