Reply to reviewer 2.

We thank the reviewer for their time and helpful suggestions, particularly as regards to the suggested rearrangement of section 2 into methods and results, which we agree improves the clarity of the paper.

In respect of their overarching comment:

First, they should clarify how their dataset improves and complements the existing TOA satellite-based products. A comparison of GERB Obs4MIP against state-of-the-art products (e.g., CERES) would be highly recommended. Second, they should provide a clearer description of the methodology, splitting between methodology and results, so manuscript readers and dataset users can find more easily all the methodological steps used to produce the dataset.

We give more detailed replies in the response to the detailed comments below, but to summarize here: the revised paper adds a discussion of the relationship between the GERB data and other observational top of the atmosphere flux products and has provided a clearer split between methodology and results as suggested. We have also added references to previous comparisons between GERB and CERES products and explain below why a further comparison between the GERB Obs4MIP and temporally interpolated CERES products does not provide any additional benefit.

We provide detailed replies to the specific comments below which we have numbered for ease of reference.

Major comments

1)-Section 1 – Introduction. I would suggest including a summary of the existing satellite-based products for RSW and OLR (e.g., CERES, NASA GEWES SRB, CLARA). It is true that most of the existing products are based on polar-orbiting instruments that focus on daily and monthly fluxes. However, products such as CERES SYN and NASA GEWEX SRB are also able to represent the diurnal TOA cycle (3-hourly resolution). The authors should highlight how their GERB Obs4MIP can complement and improve these products.

Answer: As suggested more detail has been added to section 1 to mention other satellite measurements of the TOA RSW and OLR. Specific mention has been made of CERES, ERBE and ScaRAB datasets. However, we do not discuss the NASA GEWEX SRB as this is a surface not TOA product. We also do not consider the CLARA dataset as this is not a satellite based *observation* of the RSW and OLR, but a calculation of these quantities based on satellite retrieved quantities about the state of the atmosphere, surface and cloud properties. In this sense it is more of a reanalysis product, and we don't think it is relevant to include in a discussion about observations of the OLR and RSW unless one were to extend the discussion to include reanalyses as well.

We note that the CERES instrument is polar orbiting and thus the diurnally resolved product relies on narrowband observations to supplement the baseline observations. This means that the CERES diurnally resolved products are therefore somewhat different in nature to the

GERB data presented here, as the diurnally resolved component is not based on broadband observations. A discussion of this, including references to the use of the GERB products in the development and evaluation of the CERES temporal interpolation that underlies the CERES higher temporal resolution products, has been added.

2) -Line 73 - temporal averaging: Why are the temporal averages not weighted as done for the spatial averages? The 15-minute instantaneous GERB observations are not perfectly aligned with the hourly UTC intervals. Moreover, the exact timestamp of the retrieval changes between pixels following the GERB scanning cycle. First, this temporal mismatch should be described in the manuscript. Second, due to this mismatch, authors should consider 5 GERB observations (instead of 4) and weight the observations at the edge of the hourly interval accordingly.

Answer: The original GERB observations are indeed not perfectly aligned with the hour and differ between columns. However, the GERB HR product used here has already undertaken the interpolation of the original GERB observations to provide a product which is a snapshot at the 15 minute SEVIRI observation time: this interpolation does indeed use GERB observation beyond the original hour. As this is part of the production of the GERB HR product and described in detail in the reference given when we introduce the use of the GERB HR, we don't feel it is appropriate to discuss in detail in this paper. However, we have added an additional sentence at the start of section 2.1 to emphasise that the HR product is a temporal interpolation of the original GERB observations:

> The GERB HR fluxes are a temporally interpolated resolution enhanced version of original GERB observations, derived using spatial information on the scene variation within the GERB footprint from the SEVIRI imager.

The GERB HR product replicates the SEVIRI time difference according to row, which is the result of the SEVIRI scan mechanism. The products are named according to the time of the start of the scan and run from this time for the top (northernmost) row, to 12 minutes later for the bottom (southernmost) row. Using the four HR products starting at 00, 15, 30 and 45 past the hour we ensure that all the observations from all rows have been observed during the 60 minutes of the given hour. As stated in the methods, in deriving the hourly averages for the shortwave we average albedo, essentially adjusting for the variation in incoming solar to the centre of the hour for all pixels. We have updated the paper to note that the albedo averaging and sza adjustment to the central value, mitigates not just for missing timeslots but also for the slight time offset that occurs with row. No adjustment for the variation of albedo with solar zenith or for the scene variation with time is attempted. This level of complexity is unlikely to add further benefit for this small time offset considering the likely uncertainty that would be added by any attempt to adjust these. The choice to present the RSW flux for the hour as the average albedo multiplied by the incoming solar at the centre of the 1 degree region and hour was decided during product development at the request of modellers as being most appropriate for model comparison.

3) -Line 75 - temporal averaging: explain better the albedo conversion. I assume that incoming solar radiation is calculated twice. First, at the GERB retrieval timestamp (to transform GERB instantaneous RSW observations into instantaneous albedo), and then, at the center of the UTC hourly interval (to transform hourly albedo averages into hourly RSW averages). I would suggest including the corresponding equations to clarify this part of the methodology.

Answer: This is correct but the only difference between the incoming solar calculated is the change in solar zenith angle. We have reworked this explanation to clarify that the process is really an adjustment of the solar zenith, as other terms cancel, showing the ratio used for the adjustment:

> As the total solar irradiance and the Earth-Sun distance do not change during the conversion to albedo, and back to flux, this becomes purely an adjustment in solar zenith angle to the centre of the grid box and hour bin. The process is equivalent to multiplying each flux by the ratio $\cos(\theta_{local}) / \cos(\theta_{centre})$, where θ_{local} is solar zenith angle at the HR pixel time and position and θ_{centre} is the solar zenith angle at the 1 degree latitude/longitude centre at half past the hour.

4) -Line 75 – spatial averages: Why was the albedo conversion not applied to the spatial averages? The same bias mentioned in line 77 for the temporal average could be introduced in the temporal averages if RSW instantaneous observations are systematically missing at some parts of the 1x1 degree pixel, due to the change of RSW with latitude.

Answer: The conversion to albedo is applied before *any* averaging (both spatial and temporal), and the conversion made back to flux at the 1 degree hourly scale. We have amended the sentence to explicitly state that the conversion is done before both spatial and temporal averaging.

5)-(missing) methods section. The current version of the manuscript presents a sequential structure that mixes methodology paragraphs with results paragraphs. For instance, the averaging and gap-filling processes are currently described in two different sections (section 2.1 and section 2.4), while sections 2.2 and 2.3 contain methods and results regarding the impact of missing data. I consider that the readability of the manuscript could improve by splitting methods and results. The methods section could be further split into "dataset production" (a kind of ATBD) and "dataset evaluation" (e.g., the impact of missing data before and after gap-filling). A specific section on the "dataset production" containing all the methodological steps (including an extended version of Fig 1 diagram, which currently only focuses on the averaging process) would be highly valuable for potential users.

Answer: We thank the reviewer for this excellent observation and agree their suggested change would be much more helpful to users than the current layout which describes a first attempt at an unfilled dataset and then justifies the need to improve it. We have rearranged these sections in accordance with the suggestion. We now describe the production of an unfilled and filled dataset (both are released Obs4MIPs GERB products) and then present the evaluation of each.

6)-I would also suggest adding a specific section or sub-section listing all the attributes of the final product (e.g., spatial and temporal resolution, spatial and temporal coverage, data format, data layers available in the final product, etc.).

We have added this information into the data availability section and also added a summary in the conclusions.

7)-Dataset evaluation: The manuscript would significantly improve if the authors included a validation of their dataset against an external TOA satellite-based product. The obvious choice would be using CERES products. This will not only allow benchmarking the new

dataset against state-of-the-art products but also having an independent reference to quantify the improvement obtained with some of the methodological steps proposed by the authors (gap-filling, GERB/GERB-like ratios)

Answer: The original GERB data has already been compared with the CERES products for observationally matched points (i.e. GERB points matched to direct CERES observations). This is a fundamental comparison and evaluation of the GERB observational dataset against the CERES observational dataset (Clerbaux et al 2009, doi:10.106/j.rse.2008.08.016, Parfitt et a 2016, doi: 10.1016/j.rse.2016.09.005, 2016). Furthermore, the fidelity of the CERES temporal interpolation used to produce their SYN product has also been evaluated using observations from GERB (Doelling, et al. 2013 doi: 10.1175/JTECH-D-12-00136.1 & 2016 doi: 10.1175/JTECH-D-15-0147.1). All these references have now been included in the paper.

These, much more direct, comparisons exploit the strengths of both datasets: CERES for baseline accuracy, GERB for a correct (observed) representation of the diurnal variation in outgoing fluxes. We argue that comparing the two hourly monthly mean products would essentially mimic these studies but with the complication of combining accuracy and sampling differences that have already been assessed. Hence we argue against performing this exercise. The evaluation presented in this paper deals specifically with the additional problem of creating an average where there is incomplete data which is the only additional consideration in evaluating the fidelity of the GERB Obs4MIPs monthly hourly averages.

Minor comments

8) -Line 84: "Hence, twilight and night-time RSW HR fluxes are not included in the averaging to the daily hourly scale if the central solar zenith angle is less than 85 but are used to replace grid-box values when the central solar zenith angle is equal to or exceeds 85" Could you clarify this sentence? Regarding the first part of the sentence, does it mean that 9km pixels with SZA > 85 are not used in the spatial average if the SZA at the center of the 1x1 pixel is less than 85 degrees?

Answer: Correct, if the SZA at the centre of the hour for that 1 degree region is less than 85 degrees the flux value of any GERB HR pixels with SZA > 85 are not included in the spatial or temporal averaging. This is because these pixels are not observations but just a fixed non location dependent twilight model flux used to avoid bias when diurnally averaging and would not be appropriate to include except for a twilight location. By excluding these twilight model values completely when the central point has a SZA < 85° and using only these values when the SZA ≥ 85° we preserve the diurnal bias correction and avoid them corrupting observed values. We have stated more clearly that they are excluded from the spatial and temporal averaging and noted that they are not observed quantities.

9)-Line 194: "empirical narrowband to broadband conversion" Please, include either the equation with the coefficients or a reference to a document describing this conversion.

Answer: References have been added as requested.

10)-Figures 2 & 5: Is there any reason to use these unevenly spaced categories (0, 1-5, 6-22, >22) for the number of missing days? If so, explain it. Otherwise, I would suggest using an evenly-spaced color palette or a continuous color palette

Answer: The colour coding is related to the data included in the filled and unfilled products discussed later, with up to 5 missing days being allowed in the released unfilled products and up to 22 filled days being allowed in the released filled products. No missing days are highlighted separately to show the cases not subject to error due to missing or filled data. We have added information regarding this to the figure caption, although detailed discussion is left until later in the paper.

11)-Figures 3, 4, 6, 7, 8, 9: The panel number is not seen very well. Please, take it out of the panel and increase the font size (and/or use bold text).

Answer: Panel numbers have been moved to the outside of the panels and font size increased as suggested.

12)-Figure 3: I would suggest using a diverging color palette centered around 0. Otherwise, it is difficult to interpret the differences. I would also suggest describing the four realizations in the figure caption.

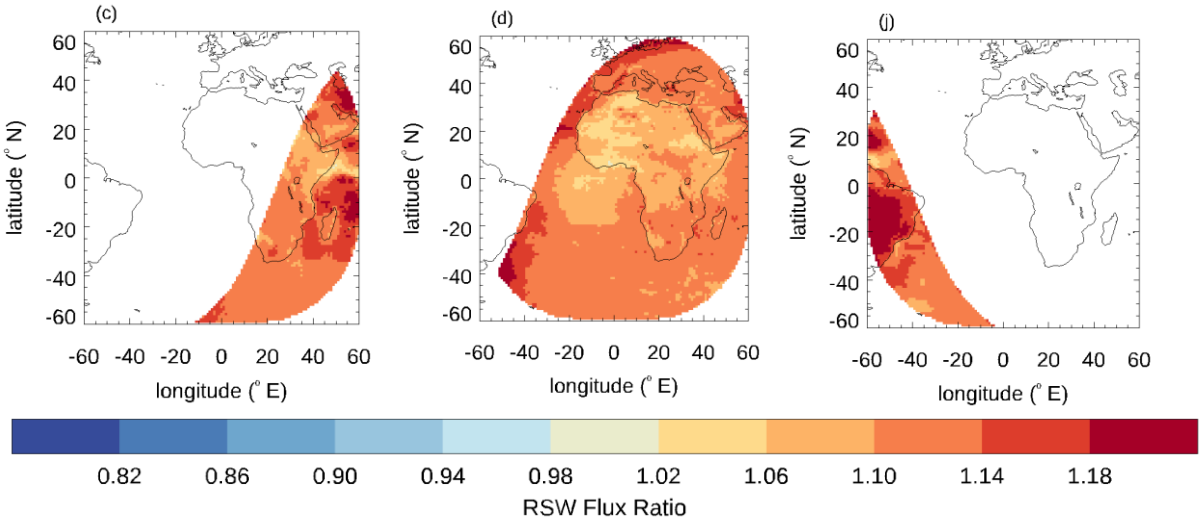
Answer: We apologise, an incorrect key was provided on this figure, a diverging colour scale was in the plot but not shown in the key. This has been corrected and a diverging colour scale is used and shown in the colour bar.

13)-Figure 4 & 8. Add a horizontal line (in the background) to better interpret the bias plots.

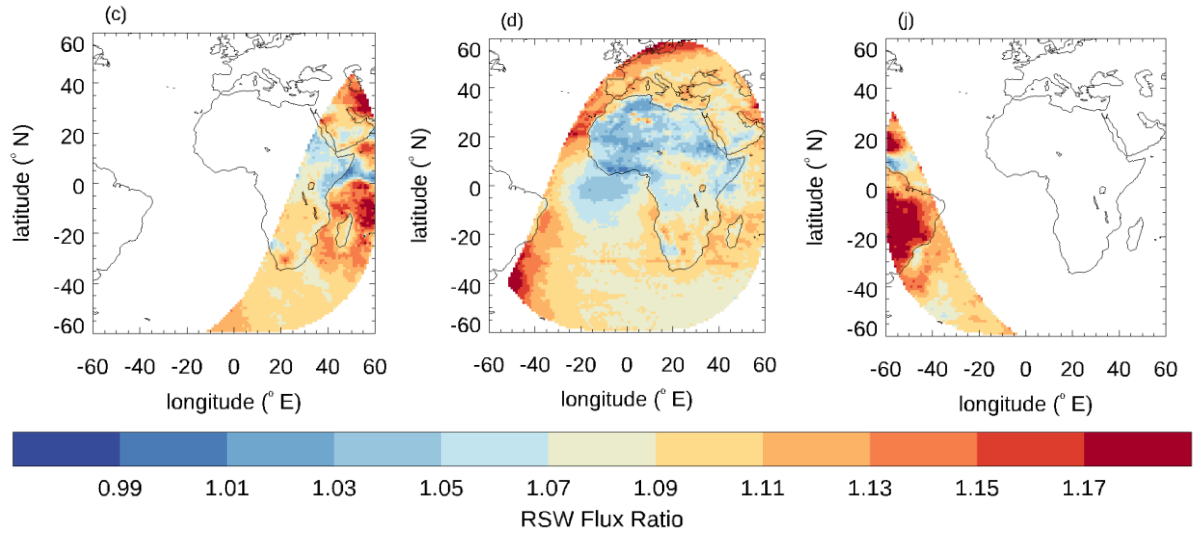
Answer: A horizontal zero line has been added to the error distribution mean panels as suggested

14)-Figure 6 & 7. As for Figure 3, use a diverging color palette centered around 1

Answer: We chose a linear colour scale for these plots because it is the spatial and temporal variation in the ratio rather than its difference from 1 that is most significant. A fixed offset would be a simple adjustment but the variability of the difference in time and space requires more complex treatment. A diverging colour scale centred around 1, although sensible in terms of 1 being an important value, would not be useful given the distribution of the ratios. For the RSW less than 1% of the points have a ratio less than 1, and for the OLR less than 2.5% of the points have a ratio greater than 1. Thus, using a colour scale centred around 1 for these plots would limit the plotted colours to around half of those available and would barely make use of the diverging nature of the scale. We show some examples below for the RSW which illustrate this issue:



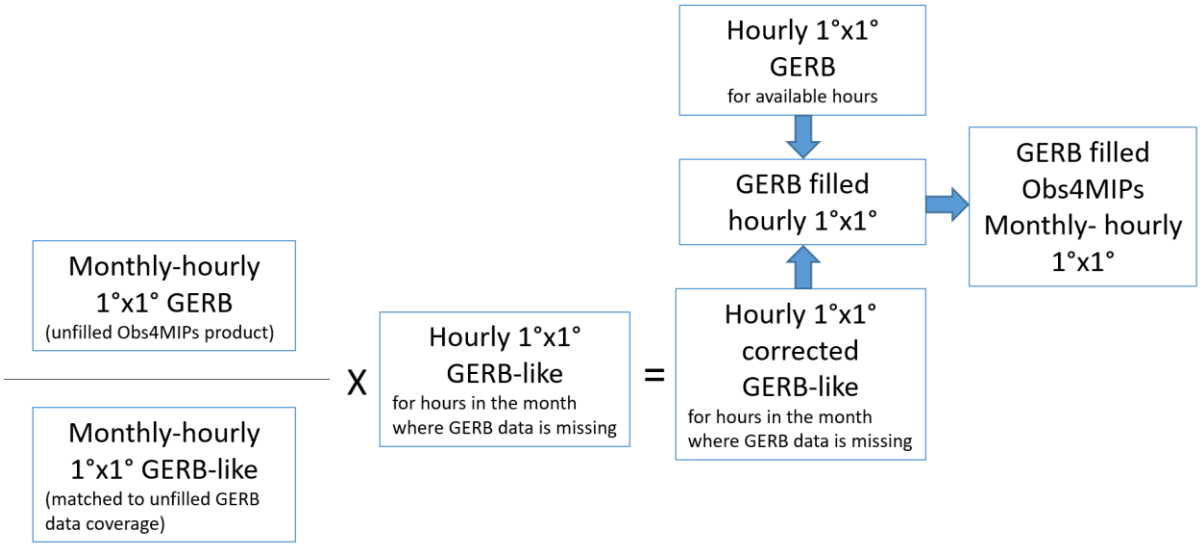
As an alternative, a diverging colour scale centred around the mean bias (1.076 for the SW and 0.99 for the OLR) could be used. However, there is nothing particularly special about the mean, either physically, or in relation to scene or angle. Thus, we feel this choice gives unwanted emphasis to the contrast between points above and below the mean even when they are quite close in value, at the expense of emphasizing the overall range of values which is more relevant to consider. An example using a diverging colour scale centred around the mean for RSW is shown below to illustrate this.



For these reasons we feel a diverging colour scale centred around 1 or the mean would be counterproductive to what these plots are trying to explore. We have added some text to the discussion of these plots to emphasise that it is variability in the ratio rather than its deviation from 1 that we are most interested in considering when determining the nature of the correction required.

15)-Section 2.5: Could you clarify if GERB-like fluxes are used (a) only to replace fully missing 1x1 degree averages, or (b) also to replace missing 9km GERB HR observations before the spatial averaging?

The corrected GERB-like fluxes are only used to replace fully missing 1x1 degree averages. This has now been more clearly explained and the following schematic added to clarify the filling process:



.

16)-Discuss the challenges to extend this methodology to GERB instruments onboard other MSG satellites, and if you have any plans to undertake this project in the near future.

A discussion of this has been added to the conclusions as suggested.