Reply to reviewer 1

Our thanks to the reviewer for their helpful suggestions. In response to their overarching comment :

“Overall, my opinion is that a more focused description of the data, highlighting strengths and weaknesses and presenting quantitative assessment of the uncertainty would improve the impact of this dataset. Although interesting, I am not sure the illustration of the dataset's use in evaluating a climate model simulation is central to a dataset description paper and would benefit from a more detailed exploration elsewhere. However, if the authors prefer this more lengthy discussion, I do not see any scientific reason to object and I consider that this work is suitable for publication with only minor modifications listed below.”

We think that the changes made to what was section 2, into a data production (new section 2) and data evaluation (new section 3) sections, more clearly emphasises the assessment of the dataset. We prefer to keep the inclusion of the brief comparison between model and data to highlight the utility of the dataset and demonstrate the information it can provide to study diurnally varying processes, such as the examples used. We agree that further papers could certainly explore this further but would not wish to remove the discussion included in this paper.

We provide answers to the specific questions below.

1) Title - I would not describe GERB Obs4MIP as a tool? How about something like:

"The GERB: a dataset for evaluating diurnal and monthly variation in top of atmosphere radiative fluxes in climate models."

Answer: Title has been changed as suggested.

2*) L8 - since GERB data has been used in model evaluation for 20 years it is not new. What is new is the Obs4MIP aspect ("newly reprocessed" may be more appropriate).*

Answer: We have changed ‘new’ to ‘newly available’. We choose not to use ‘newly reprocessed’ to avoid any implication that the underlying processing of the GERB data has changed. The GERB Obs4MIPs data presented are newly derived higher level products based on the long existing processing of the GERB data.

3) L11 "approximately" since it is not a square 60S-60N, 60W-60W domain

Answer: Added as suggested.

4) L19 ":" --> "," (or we could compromise with a ";")!

Answer: Altered to a comma as suggested

5) L20 improved fidelity, relative to GC3.1

Answer: Amended as suggested

6) L27 Since outgoing longwave is used later it may be worth stating here e.g. "emitted thermal infrared (outgoing longwave) and solar reflected..."

Answer: Amended as suggested

7) L36 also Allan et al. (2011), Examination of long-wave radiative bias in general circulation models over North Africa during May–July. Q.J.R. Meteorol. Soc., 137: 1179-1192. https://doi.org/10.1002/qj.717

Answer: Added as suggested.

8) L55 add a line pointing to the fact that the HR product is a resolution enhanced version of the GERB broadband radiative fluxes using SEVIRI narrow band measurements.

Answer: As suggested the following sentence has been added:

> 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.

9) L63 "are" --> "were"

Answer: Altered as suggested

10) L72 why is 70 degrees zenith the cut-off and not slightly more or less?

Answer: The limit follows the recommendation of the GERB quality summary beyond which errors increase and mapping to other grids becomes problematic. The sentence has been amended to indicate that the vza limit relates to the recommended valid range of the GERB products when averaging the data.

11) L74 "proceed without prejudice" is not really clear what is meant (sounds like legal speak). Is one observation representative of a 100x100km area? Presumably there are mostly lots of values per 1x1 degree box.

Answer: When there is no missing data a 1x1 degree box will, depending on its location, contain between 6 and 169 HR points for each of the four 15-minute time steps that comprise an hour. However, there are occasions when one or more of the 15-minute slots is missing, there are also much rarer occurrences when some of the HR pixels within the region are missing. In both these cases averages are made over the available points as long as there is at least one within the 1x1 degree box and hourly bin. Whilst in theory this means that one HR point could represent a whole hour of the day for the entire 1x1 degree region this is highly unlikely and if it occurs at all would likely be restricted to cases close to the edge of the valid region where there are fewer HR points to start with. We have added a sentence to explain this.

12) L76 incoming solar flux from what dataset or algorithm?

Answer: The process described reduces to an adjustment of the solar zenith angle. In practice we use the incoming solar flux used in the production of the GERB-like products. This assumes a solar constant of 1366 modulated by the calculated Earth-Sun distance for that day of the year and multiplied by the cosine of the solar zenith angle according to location. However, the assumed values of incoming solar and earth sun distance are irrelevant to the resulting product, as they do not change between the conversion from flux to albedo and back from averaged albedo to flux. Thus, the process is equivalent to adjustment to the solar zenith of the midpoint of the grid point and hour, with the overall level of flux maintaining that associated with the GERB data, and thus traceable to the observations themselves. The following sentences have been added to make this point clear:

> 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.

13) Figure 3 - the colour bar seems to bear no resemblance to the maps. Also this figure could be designed in a way to maximise the size of the panels (or reduce the size of the plot and all the dead space)

Answer: Our apologies an incorrect scale was included. This has been amended and the panels have been rearranged to make better use of the available space as suggested.

14) L187 "exiting" --> "existing"

Answer: Amended as suggested

15) L200 this paragraph could be reduced to the first line plus "(compare Figures 5 and 3)".

Answer: We agree the paragraph should be reduced but suggest as a compromise it be reduced to the following 3 sentences (note figure numbers have changed due to rearranging of these sections):

> The SEVIRI based GERB-like fluxes suffer from significantly less missing data than the original GERB record (compare Fig. 3 and Fig. 2). Except for some extended outages in the first few years which are a result of satellite level anomalies, nearly all the data missing in the GERB record are present in the GERB-like. Thus, the latter record may be useful for filling much of the missing GERB data.

16) L205 these lines are difficult to understand. Can they be written more clearly? There is a lot of detail in this section that may be unnecessary for users of the dataset so another option is to more briefly note issues and refer to prior papers if the reader is interested.

Answer: We agree and have replaced this paragraph with the sentence:

> The manner in which the GERB-like fluxes are used in the GERB processing places no requirements on their absolute accuracy and very limited requirements on their relative accuracy.

And merged this with the following paragraph which has also been slightly altered to accommodate the change.

17) Figures 6/7 - could these be designed to fit better on the page and maximise the size of the panels? Does the line at 25oS relate to a dead or damaged pixel?

Answer: The panels have been rearranged to maximise their size as suggested.

The line at around 25 S relates to GERB pixels with a longer than ideal time-constant and that are more sensitive to instrument conditions as a result. For the months shown this can cause an elevation of around $1 Wm^{-2}$ (~0.5%) to the GERB monthly hourly average flux for that region. Discussion of this seemed beyond the scope of the paper as the intent is to investigate the ability to replicate the GERB data. Although noticeable in the ratio because it is a persistent effect, the response difference is quite small and still falls within the overall accuracy statements of the GERB products. Thus, it is not of great relevance to the Obs4MIPs products themselves.

18) L251 - of course it is feasible to fill missing GERB data with GERB like - there is no need to hypothesise, just quantify the expected error and assess whether this is tolerable for the designed usage of the dataset. A key line to emphasise is that the strategy is to "fill missing GERB fluxes with the corresponding GERB-like fluxes, adjusted by the GERB/GERB-like ratio calculated at the monthly hourly mean temporal and 1º spatial scale." and now the associated errors are quantified.

Answer: Indeed, I suspect that we intended to hypothesize that it would be feasible to produce a more accurate product by using the GERBlike to fill the missing GERB data. In response to this comment and comments 20 and 21 the paragraph has been rewritten to make the point of the investigation shown in figure 8 clear. (see response to 21 for more information).

19) L259 1 degree "latitude/longitude" (throughout)

Answer: Amended as suggested throughout.

20) Figure 8 - make text bigger. It's not obvious what the use of this is since the decision to scale GERB-like with the ratio seems obvious (and the mean bias of about zero is by design)

Answer: The text size has been increased (for consistency similar changes have also been made to similar figures). With regards to the point of the plot please see the reply to the following comment in reply to point 21.

21) L270 - the improvement in agreement between GERB and GERB-like after essentially removing the bias is obvious by design. This sentence can be removed. Section 2.5 seems to be the main analysis following the rational for the approach, which could be more concisely presented as the method

Answer: We apologise that the intent of this figure wasn't clear and agree that it is more appropriately part of the method. This figure is included to show how a monthly correction impacts the match between the GERB and GERB-like daily hourly data, as the latter is the scale at which the corrected GERB-like will be used as a proxy for missing GERB data. Without this check we have not shown that a monthly correction is useful for our case. It is not a given that the monthly correction would improve the daily data. For example, if the overall bias changes from day to day, the monthly correction would not reduce the range in the daily error distribution mean, shown in the upper panels. Although it would be expected as you note to shift their values to fall around zero. Similarly, if the bias for different scenes occurring at a given location was very different, the monthly average scene bias would not provide a good location specific correction and the standard deviation of the daily error distributions would not be reduced after the monthly correction.

We appreciate that this point was not clear in the original description and the paragraph has been reworked. Also, the whole section is reordered to make clearer the distinction between method and results with this figure forming part of the filling method as you suggested.

22) Figure 9 - it's not clear to me what the benefit of showing all the unfilled results as well as all the filled results. If the idea is to compare Figure 9 and Figure 4 then an unfilled vs filled line could be presented on the same figure. Are all times lumped together? In this case massive errors due to missing parts of the diurnal cycle will be introduced won't it (which would obviously not be considered in practice)?

Answer: We have reworked this whole section to make clear that we are discussing the error in the unfilled products due to missing data (original Figure 4) and the filled products (original figure 9) due to filling this missing data. In both cases these error statistics summarize the errors at the 1 degree scale in the **monthly hourly** average. There are no times of day missing, rather this is the effect of some days missing at a given hour and, for the filled case, result in figure 9 being replaced with the adjusted GERB like.

To clarify we have added this detail into the sentence:

> "Comparing the two figures shows that filling the missing GERB fluxes with their scaled GERB-like equivalents reduces both the mean error and the spread in the error by more than a factor of 10 in all cases "

So it now reads:

> Comparing the two figures shows that filling the missing days of GERB fluxes with their scaled GERB-like equivalents before calculating the monthly hourly average, reduces both the mean and standard deviation of the error in the monthly hourly average at the 1 degree scale by more than a factor of 10 in all cases.

In response to reviewer 2 we have also reordered and reworked this whole section so that it now clearly separates the methodology for producing unfilled and filled products from the evaluation of the products. This clarifies that the two figures are making distinct points relevant to the points in the paper they are included. It is also relevant to compare the figures to demonstrate the value of filling to produce a higher fidelity average. However, as the scales used in figure 4 and 9 are very different (due to the factor of 10 reduction in error), including the results again in figure 9 for comparison would make the results of interest here

(the error after filling) difficult to see well. (Note figure numbers in the updated paper are now figures 9 (previously 4) and 10 (previously 9))

23) L306 AMIP needs to be explained

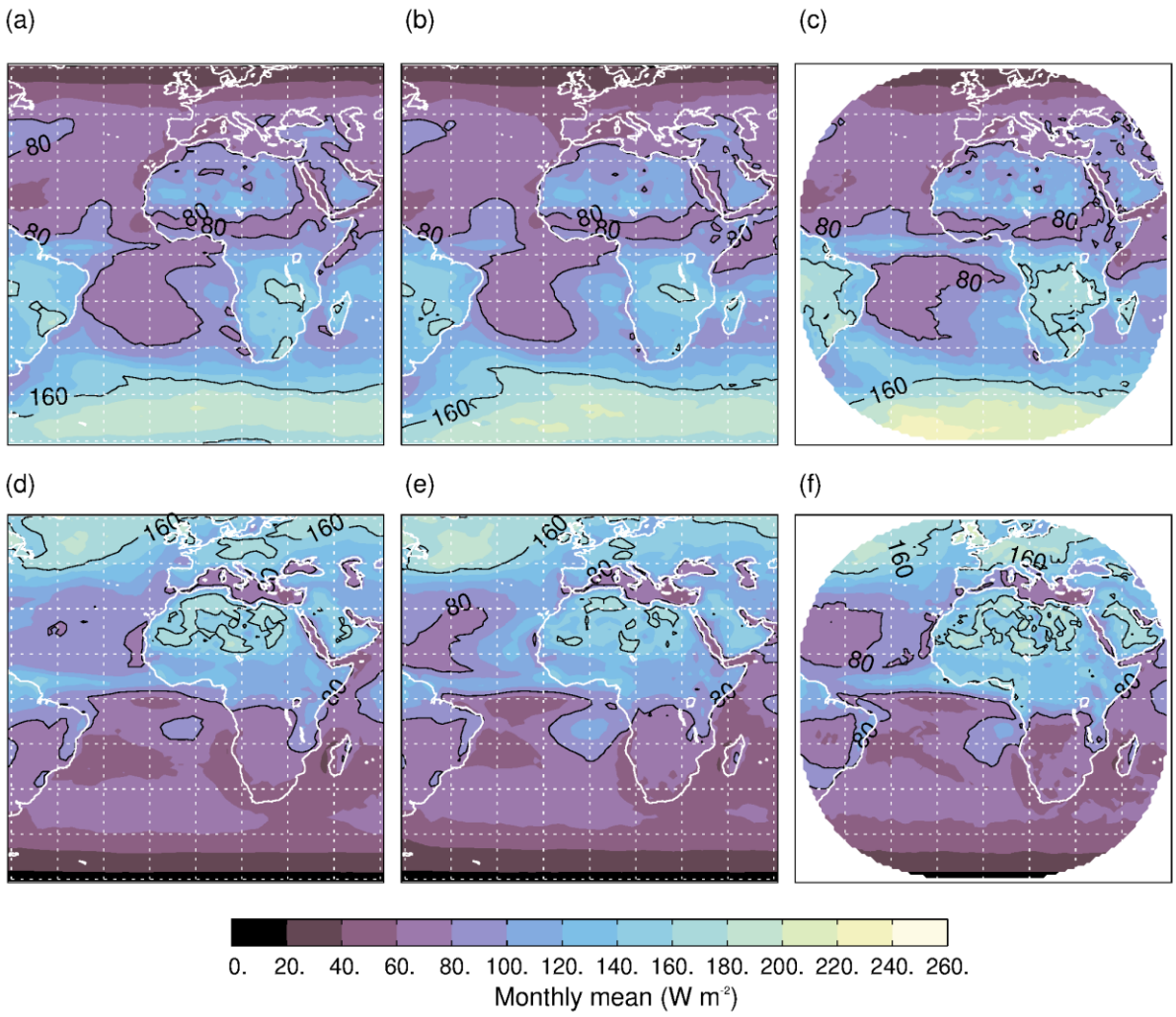
Answer: Now defined with an added reference

24) Section 3.1 - although important and well described, this is a bit of a gear change from describing the dataset. Is all this information necessary?

Answer: We have now simplified and summarised the detail on the model differences emphasising the expected changes relevant to our analysis. We have also added a sentence at the end of the introduction to make clear the intent and scope of this section.

25) (a) Figure 10/11 are very qualitative and it is rather difficult to link colours on the bar to values in the plots. A plot of differences would be more informative.

Answer: These figures are intended to show a spatially resolved picture that is familiar for orientation to indicate the nature of the information in the GERB product and to highlight regions of interest for more quantitative analysis. They are not expected to be the focus of detailed quantitative analysis. We have added a couple of sentences to make this intent clear at the point they are introduced. In addition to help with linking the colour scale to the contours we have extended the range of colours used for the figures and added two labelled contours on each plot to help orientate readers. Viz:



Due to differences between the model and GERB obs4MIPs product scales (1.875° longitude by 1.25° latitude for the model and 1 by 1 degree for the GERB) a subtraction would necessitate a rescaling and be counter to the intent here which is to illustrate the GERB product directly. We also feel that the differences would be harder to interpret, due to the effects of shifts in regimes, and less useful for orientation than the more familiar monthly average. Illustrative quantitative comparisons which are averaged over larger regions to minimise effects of the different product scales and allow for slight shifts in regime location follow later for the selected regions of study.

25) (b) Why are coincident model years not used (CMIP6 amip simulations usually end in 2014)?

Answer: Whilst coincident years were available for GC3.1 runs, only years between 2000 to 2009 were available for the GC5.0 runs. We considered the priority to match model years and use sufficient years to average internal variability as much as possible, as the focus was on comparing general model behaviour with the general behaviour seen in the data. We did consider comparison only for the overlapping years for the three datasets (i.e. 2007, 2008 and 2009), as noted in the paper the average region results for these years only, differ by no more than 3 Wm-2 from the full 2000 to 2009 average. Although not discussed in the paper we have compared for GC3.1 the decade used with 2007-2012 average which matches the GERB

observation period. The differences in this case are even smaller and do not change the conclusions drawn.

25 (c) I am not sure the Figure 12 results are very relevant to the GERB data description; a reference could suffice.

Answer: The intent of this figure is to provide detail on the general/average cloud property changes between the two models, to provide context to the flux comparison against the GERB data. It also serves to demonstrate the additional utility of diurnally resolved flux comparisons compared to diurnally averaged or single time of day cloud property comparisons.

25) (d) Figure 13/14 seem much more relevant - a legend to denote model simulation is needed and perhaps a thick mean (or median) model value would be useful. Showing albedo may reveal the diurnal cycle of stratocumulus better than RSW (which is more dependent on the insolation). Figure 14 - if the idea is to compare the model versions, it would be more useful to have the two mean lines (and perhaps shading for range) in the same plot.

We have updated these figures, adding the model simulation lines to the legend and including a line for the model mean in each panel as suggested. However, rather than show a model spread we have retained the individual year lines as we feel that the individual annual shapes and variability are more informative.

The GERB Obs4MIPs are monthly hourly mean flux products. Conversion of the monthly hourly mean flux to an albedo would require a deconvolution to account for the daily variation of the of the outgoing flux and incoming solar, this is beyond the scope of the monthly diurnal average product presented. We also note that the GERB RSW are directly linked to the observed radiance which is converted to flux via directional models and not linked to any assumption regarding the incoming solar. Thus, there is an argument that the GERB flux is the more direct observed quantity and for that reason the better quantity to compare against the model.

26) L352 - it could be mentioned that these issues are not fully solved in higher resolution simulations e.g. Watters et al. (2021) J. Clim doi: 10.1175/JCLI-D-20-0966.1.

Answer: A sentence noting this has been added including the suggested reference.

27) Table 1 - how much is positional/regime error and how much is cloud property error? How have biases changed since earlier analysis (e.g. for the NWP version comparison in Allan et al. (2007) QJRMetS clouds were instead too bright due to too much water)?

Answer: We don't have a direct comparison against the NWP model configuration used in Allan at al. (2007), but the “too few and too bright” bias in low-level clouds has been a persistent problem for many generations of models and is still present in GC3.1 and in other CMIP6 models (Konsta et al., GRL, 2022; https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021GL097593). It should also be borne in mind that Allan et al 2007 predates the recommendation to apply aging corrections to the GERB RSW. For the GERB 2 July 2006 data used in that study this correction would result

in around a 2% uplift to the GERB fluxes (assuming that the 0.976 correction recommended for the GERB 2 data was available and applied in the Allan comparison).

For our comparisons, altering the region or month under consideration changes the details of the numbers, but not the overall pattern seen. For convection for example, using the African land region from Allan et al. 2007, for either June or July, reduces the SW and increases the OLR compared to the June deep convective region used here. However, the basic result that the GERB data is brighter (by ~20Wm-2) and colder (by ~20-30 Wm-2) than the models remains. With CG5.0 showing a small movement towards the GERB observations.

Regarding the effect of regime position differences, as the region comparisons average monthly averages over a large geographical area and consider multiple years, they comprise both cloud and non-cloud regimes to some extent in all cases and are relatively insensitive to small regimes shifts. The spatially resolved averages for the regions chosen show that the regimes are contained within the chosen geographical areas for both models and observations. However, for the models, as the convection is less extensive, the regime will comprise a smaller portion of the region. You might consider this a positional difference, as a comparison that was limited to identified convection would differ. Although the fact that the convective region were smaller / less intense in the model would remain, which can be considered a cloud property problem, i.e. a lack of cloud development.

The cloud fraction and optical depth plots (original Figure 12) shows that average cloud properties have indeed changed but the effect on the average fluxes (table 1) is small in comparison to the difference between model and data. As the GERB product does not provide associated cloud property information, cloud property effects are not assessed here, but could be the subject of a future study.

The primary intent of the presented study is to compare the diurnal cycle of flux and we feel it achieves this goal. The regions chosen cover the intended regimes for both models and observations. Thus, the flux differences for the regions reflect the differences associated with the regime, but is a function of both regime extent and cloud properties.

28) L442 the model seems to overestimate OLR for both months - does cirrus outflow or water vapour contribute?

Though this would be interesting to determine in a future comparison it is beyond the scope of the paper and would likely require higher resolution data and additional information to investigate. We do not discuss it in the paper but in general water vapour comparisons between the model and reanalysis look OK. We cannot say anything definitive about cirrus outflow. As we note in the paper, our comparison between observations and model shows that for the model convection is too weak or missing. There are indications from the monthly mean that the region of deep convection-like low OLR is smaller in the models. The diurnal cycle also shows that the OLR in the models is too warm before the onset of the convection and the convective development is stunted compared to the observations.

29) L475 - it is still not clear what this factor of 10 reduction in uncertainty means. Is this just for simplistic averaging, where parts of the diurnal cycle are missing and which would not be undertaken in a serious analysis? Or is it referring to missing days?

Answer: The errors being discussed are at the scale of the product (1 degree longitude/latitude monthly hourly average) no further averaging over the diurnal cycle is being undertaken. The reworking of the original section 2 now makes clear there is an unfilled and filled product, this should make it clearer that we are comparing the errors for a given amount of missing days between the filled and unfilled products. The error in the unfilled GERB Obs4MIPs products due to missing data arises purely from having missing days of data for a given hour when we make the monthly average for that hour and is a result of uncaptured day to day variation. The error in the filled GERB Obs4MIPs products due to filling the missing days is due to the residual difference between the corrected GERB-like and the GERB data it represents. We have also added some words here to clearly define the errors we talking about:

> For a given number of missing days the residual uncertainty in the monthly hourly average at the 1 by 1 degree longitude/latitude scale due to filling is more than a factor of 10 smaller than the error in the unfilled products due to the missing data.

30) L478 Some clear statements about product uncertainty and it's recommended use could strengthen the conclusions.

Answer: Additions have been made as suggested.

31) L479 some illustrative comparison with model simulations are useful for introducing the dataset though I think the work presented here is more deserving of a separate, more detailed investigation.

Answer: We agree that an expansion of the comparison should be the subject of a more detailed investigation but feel that the comparisons presented in the paper are appropriate to the introduction intended. We feel that both the brief spatial comparison and regime specific diurnal investigation provide context and show the unique strength of the product. The presented studies also clearly illustrate how non-diurnally resolved cloud property comparisons are complementary but are, on their own, insufficient to understand improvements in cloud response on diurnal timescales.

32) Acknowledgements - the work of the GERB team could be mentioned in the acknowledgments and more specifically the contribution to initial discussions of the strategy for dealing with sunglint led by Jo Futyan and the terminator led by me could be stated.

Answer: These were not included because they pertained to the GERB data where they have long been implemented and are not specific to the Obs4MIPs products presented here. However, we agree that these features of the GERB product are fundamental to the ability to produce an average and thus along with the availability of the GERBlike data are of fundamental importance to the obs4MIPs product. We have thus added acknowledgements as suggested as well as to the RMIB team for their help with the GERBlike dataset