Review of: Retrieval of top-of-atmosphere fluxes from combined EarthCARE LiDAR, imager and broadband radiometer observations: the BMA-FLX product

The paper presents a description of the BM-FLX SW and LW radiance to flux conversions including the determination of the required cloud information from the MSI cloud properties. Using EarthCARE test frames it goes on to evaluate the error component in the fluxes due to the conversion process for both the ideal case of perfect cloud information and for the case of cloud information derived from other EarthCARE products. The work presented provides important information to users regarding the basis of the radiance to flux conversion and the evaluation undertaken is a valuable and necessary exercise providing an initial look at the flux performance compared to modelled fluxes.

The paper is generally well structured but in terms of clarity I think that some additions and modifications are needed in the description of the method in section 2. Some context to the evaluation would also help elucidate the scope and limitations of the comparisons presented. The goal of being able to estimate flux to 10 Wm^{-2} at the 10km by 10km for both the SW and LW is a very challenging requirement (particularly in the SW). Without knowing the details of the specification it is not immediately clear to me if this is k=1, 2 or 3 requirement on the error or an upper limit so I don't know exactly how to relate it to your results. If possible it would be good to clarify this, although I realise that these things are not always as clearly specified as they might be and may require a bit of interpretation on your part, in which case I would suggest you explain your interpretation to provide some indication of the bar you are trying to meet. I note that you mention the requirement in relation to achieving radiative closure to within 10 Wm-2 (which I appreciate is the origin in the requirement within the EarthCARE mission) but this is possibly even harder to relate to your results and to be honest I think not clearly related even at mission specification level when deriving the BBR accuracy requirement, so it may be easier to stick to discussing the flux product requirements, however this is up to you. Whatever the specifics I think it would be useful for users to be able to understand how these results indicate they should employ the data for their chosen purpose and how they might identify fluxes that meet a certain quality criteria or a means by which they may be able to meet a specific criteria. To this end it would be useful to know if quality indicators in the fluxes themselves can provide some information on accuracy of the fluxes allowing users to select an appropriate subset. Alternatively or additionally it would be useful to know if averaging over a larger domain provide generally better results (this is implicit in the bias and seems likely looking at the results and I think knowing at what scale you may meet a specific criteria would be useful). Also some discussion of how representative the errors are likely to be (given you are comparing to model fluxes), and discussion of the way forward regarding the effect of the scene identification errors would be helpful. For example, are the scene errors because these products don't meet their specifications and will be improved? Is it because the model radiances aren't as realistic as the retrieval and thus the errors are not representative of the real world or are these errors that you have to live with and can you adjust / tune the scene id to cope with them?. I don't think that addressing these things constitutes a major change to the paper in any sense.

I detail these issues below along with noting some possible issues with figures 3, 5 and 6 which may be incorrectly displayed.

Main points:

Section 2.

Lines 59 to 84. At this point in the paper you are discussing the definition of flux generally (both SW and LW), however the formulation you present is specific to the SW so is a bit confusing. I think this section should be general to both SW and LW or needs to become part of the SW section below and then be repeated for the LW. I would suggest you can generalize by stating that you are integrating the radiance field over viewing zenith and azimuth (this is true for the both LW and SW). The radiance depends on the scene the viewing geometry and additionally for the SW the solar geometry. Note in this case you would not use relative azimuth but viewing azimuth as the integration term. You would in this case also need to avoid anything specific to the shortwave method. Alternatively, you will need to formulate a separate discussion and equation for LW and SW but I don't think this is needed.

Lines 67 you 69. This is phrased a bit strangely and I am really not sure about the points these two sentences are making. Converting radiance to flux using a model of the radiation anisotropy is the method you are employing. As opposed to calculating flux from properties retrieved from radiances for example. The model used to represent non uniform variation of the radiation field (i.e. its anisotropy), is an angular distribution models (ADMs) in a general sense. You state 'The ADMs have accurately represented this variation (Su et al., 2016) and so are used as the basis for flux retrieval'. My understanding is that you are not actually directly using **the** particular ADMs described in Su et al. although they may be relevant to the accuracy of the fluxes you are using this does not automatically imply any accuracy of the ADMs you derive from those fluxes if you follow a different classification system for example. If you just mean by this sentence that ADMs can be accurate, this may be true but that some angular distribution model has been used successfully before doesn't have any bearing on the accuracy of any other ADM. So I'm a bit lost and think maybe a few things are not quite said as intended, particularly bearing in mind that at this point you are talking about both the SW and LW flux derivation. I think you want to make the following points but I'm not sure if you really want to make them all here:

- You are developing your own ADMs to represent scene anisotropy and enable fluxes to be derived from flux.
- For the SW you are using CERES fluxes to develop your model, these employ their own distinct ADMs which are of established accuracy (although this is only relevant for the SW I think so that will need to be clarified)

Line 70 to 72. I think this needs to be rephrased. An ADM doesn't estimate the flux, it represents [an estimate of] the relative angular distribution of the radiation field and can be used for the derivation of flux from a single radiance measurement. Similarly an anisotropic factor at a given set of angles does not define the ADM but can be derived from it to enable the conversion of a radiance at a given angle to flux.

Section 2.1.3 I think it is a difficult task to try and convey your selection of inputs here via a narrative style I would suggest that a table may be an easier way to convey much of this information listing all the possible inputs the source of the information and then the scene classes where they were included. Similarly, some indication in this or a separate table of the best/chosen inputs, at present I don't think this section makes it clear what the inputs end up being. I think this would clarify a lot of what is trying to be explained and could replace much of the text in this section.

Line 121. 'a predefined list of parameters that influence scene anisotropy is very vague. I assume this relates to the climatology and meteorological data you speak of above. I think a table listing all the variables considered and some discussion on those eventually considered significant would solve the ambiguity.

Line 122 'BBR-received scattering direction' I'm not sure if you mean each BBR view or if you mean the forward and backwards scattering directions generally (in which case this is part of the scene) or if you mean every bin of the ADM. Can you please clarify.

Line 125 you speak of two methods which 'largely agree' how are they combined and how is any disagreement treated, do you just AND the two sets?

Line 131 where does LAI come from is this a dynamic MODIS value integrated at footprint level or a generic value based on the static classification of the surface. Where do you get this information in the evaluation section it doesn't seem to be discussed.

Line 132 how do you 'consider' the hot spot effect as input, what is used as the input?

Line 133 to 134 you say AOD and wind speed is 'selected as parameters for network training' what do you mean by selected, is this selection the result of testing (i.e. selected by some criteria in which case please explain) or just considered to be important ('chosen')

Line 139 how are MSI-like MODIS radiances created? Is this just the nearest wavelengths or are some spectral adjustments done, are they something you create or a defined product, can you please include further detail or a reference.

Line 140 it's a very plausible assumption but would benefit from an extra line of justification maybe stating which narrow bands you use and pointing out they are the ones used to retrieve cloud properties and thus implicitly contain that information albeit extracted via some other model.

Line 143 could you clarify if the process determines if these are the best or if you choose these.

Line 146 to line 147 how is the 20% of the dataset used for validation chosen and is it truly independent? E.g. every 5th ceres footprint would not be independent every 5th orbit *might* be, but I would suggest that 1 year out of 5 would most likely be the best test. Is the CERES validation dataset the same as the cross check validation or different? How much data is this and how is it chosen. What is the result of this validation what expected best case errors are you expected from the retrieved flux on the basis of this validation?

Line 155 You say the retrieval algorithm uses two surface types and combines them, do you mean you make two separate retrievals for each class individually which are then combined or do you mean there is some sort of combined class possible in the retrieval?

Line 162 what DEM is used, and do you also do this for elevations below 0 (assuming zero is the average Geoide or is it something else? Although I'm a bit confused because surely the coregistration at the surface performed by BBR considers some sort of DEM already.

Line 164 to 166. This sentence is unclear at this point, it becomes clearer after figure 1 is discussed maybe reference this later discussion or consider rearranging such that the 2.1.5 comes before 2.14? Also I would replace cloud properties with cloud fraction as I think this is the only 'property' you are using for scene? If you are also using other properties for input it might be worth clarifying somewhere what is used for that for the oblique view although probably this shouldn't go in the scene section, although I'm not sure it would make sense to use nadir properties in that case unless that somehow corresponds to what was used in the training.

Line 166 to 175. I am confused by the discussion on cloud top height in this section on scene identification as whilst it has been described as an input it has not been discussed as part of the scene classification. Is this part of the scene classification (in which case it should have been discussed in scene definition section), does it just relate to matching scenes (in which case can you clarify this) or is this really a discussion on inputs in which case it should go elsewhere or the scope of this section should be extended.

Line 183 'providing there is no significant elevation' do the BBR products coregister at the 'surface' or at some average sea level, surely the former which would include elevation of surface?

Line 195 to 196 'CTH…is a reliable estimator for co-registering…in the thermal' was this shown in the AATSR study or some other way how is this know. One might have thought that thin cloud may pose a similar issue in the thermal as in the visible, or that wet atmospheres might may the co-registration point above a low cloud, or in clear sky make the surface a poor point of co-registration.

Line 202 how are the errors in the ADM treated when minimizing the height of co-registration, what happens in cases that are poorly behaved and a minimum isn't found?

Lines 223 to 225. I find this sentence difficult to understand. I think maybe don't use the word 'coordinates' unless you really mean 2d or 3d coordinates and or specify what coordinates you are talking about. So if you just taking about the central pixel in the nadir view, say this, if you are talking about the 3d location of the nadir reference level then specify this, if you mean the location of the ctp in the nadir then again specify.

Line 228 equation 5. Do you calculate this for each combination, i.e. nadir vs fore, nadir vs aft and aft vs fore?

Line 229 this equation does not necessarily limit the range of 0 to 200%, if you limit it artificially maybe this should be specified in the equation by an alternative formulation,i.e. as stated if < 200% and 200% if greater than that.

Line 244 do you have any evidence that using a set of regressions solve the problem and if so why/how. Maybe you need to just state the cause of the issue was a deviation from the fit used for this scene type (rather than an issue for a plane parallel assumption for example which would not be solved but a different regression). Also is this set just the splitting into 5 degress in VZA and 20 Wm-2sr-1 bins or is it something more based on cloud information could you explain somewhere in this paragraph, either state on line 244 that it is radiance and angle separated or explain it is also cloud property dependent and include this addition on line 246.

Line 265 how are uncertainty estimates associated with the fluxes derived this was derived, are they related to the MSI bt uncertainties propagate through the regression used to derive the anisotropy? Do they take into account the scatter about the original regression?

Line 267 can you state what the reference level is in clear sky, is it the surface or somewhere above and does it vary with atmospheric moisture?

Line 280 how it is evaluated from the CERES data? I assume you mean that the ceres data indicate that equal weight should be given to the forward and nadir views in combination. It is not clear what you are doing here but I suggestion (see the following point) that it might be worth a bit of explanation.

Line 281 I think this requires a bit more comment. You are saying the studies with CERES implies that the LW radiance distribution doesn't behave in a plane parallel way. However you are using plane parallel assumptions to base your radiance to flux conversions on so this would seem potentially concerning. If your comparison with CERES is based on applying your theoretical plane parallel ADMS to forward and nadir and comparing the induvial vs combined result with the CERES empirical ADM conversion then this would support this equal weighting to more closely match the empirical CERES result. But I think this needs explaining and possibly the actual improvement and discrepancies found stated here.

Section 3.

Lines 325 to 329. I think it would be helpful to have a previous section clarifying the two step comparison you will execute in 3.2 and 3.3 and the purpose and scope of the two comparisons and strengths weakness and limitations of these assessments. I think the two step approach you use is reasonable (and in fact wonder why you don't go further and substitute a 'model truth' for the X-MET data. But I think it would be useful to also discuss the limitation and issues of the model truth in the context of your method as whilst it assesses much of the performance it will obviously be limited to how the model fluxes relate to the ADMs which is not necerssarily going to be the same as the real world. You discuss some of this such as surface mis-match when you consider differences but I think it might be helpful to have a broader stage setting consideration of how the model fluxes might not match the SW adms you use, or how assumptions in the longwave such as cloud modelling or surface emissivity might differ between the model the information used to derive your ADMs.

Line 335 'being 0.1 the one with better performance' → '0.1 being found to have the best performance'. I don't understand what is being done here, what determines the best performance? Also why is this not needed to be repeated for the retrieved parameters, this would account for any offset in the retrieval. Will this be redone operationally to determine an operational threshold and if so how will performance be assessed in that case?

Figure 3 – Upper panels please label the colour bar and check that it is correct. The caption states it is cloud optical depth (at 680nm ?) but the scales seem quite extraordinary. I would suggest if you really have COT up to 800 you might consider a logarithm scales as a more useful way to provide information in these figures.

Lines 342 to 353 I think that results in table 1 need to be discussed in terms of the differences between the simulations and the FLX processor assumptions. For example the SW flux retrieval is based on observed radiance distributions, albeit somewhat indirectly (via CERES fluxes based on observationally derived ADMs), in contrast to plane parallel (?) model calculations, can this be an explanation of the variation in the sign of the bias between views in the SW do you think? What might explain the consistent offset in the LW which actually seems worst in the nadir view, is this a consistent difference between the limb darkening effect between the model simulation and your simulations or a surface emissivity effect or something you might expect to persist with real data?

Line 358 Are you saying you use only the GLCC classes and don't use anything equivalent to the X-MET to id fresh snow, later (line 372) you seem to imply you use X-MET would it not be helpful to put what you use in the ADM here if this is different, maybe you do and you just need to clarify you use GLCC updated where appropriate with X-MET here?

Line 367. I think before you are too hard on your performance under broken cloud conditions you should consider how well these are likely to be simulated. The ‘complex interaction’ seems to imply a 3d effect but are these considered by the model truth. Or do you mean to say this might be a reference level issue?

Line 370 to 374 It is not immediately obvious which has snow and which doesn’t in your assumption vs the model truth, but I think you are saying you don’t assume snow because it isn’t in the X-MET data but the model truth includes snow. Have you confirmed that if you use snow here that the problem goes away. Is using the X-MET data sensible in this case rather than using the model truth surface?

Figure 5, Although the mean lines look they are probably correct the Std an RMSE plotted on the graphs don’t seem to be correctly plotted or at least do not correspond to the values shown plotted around the mean. Furthermore, I think it would be useful to see how the errors relate to any quality indicators on the flux you have (e.g the flux uncertainties from the ADM (equation 6) you have for the SW and or the discrepancy between the fluxes derived from different views or some other measure from the retrieval if it is good or not). Maybe this could be indicated by colour coding the different points that fell above a certain retrieval goodness threshold or by including an additional plot of flux difference against retrieval uncertainty.

Also in the longwave there seems to me to be an indication of noise in the retrieved flux that might be significantly improved if you increased the averaging region (smooth the results seen here). Do you think this is the case. Would it be worth looking at what improvement this wrought to the percentage of points within the 10Wm-2 line, and would be useful information in the context of closure style comparison studies.

Lines 384 to 387. You discuss the differences between the cases but is there indication of why they behave so differently?

Line 387 Is the lack of cloud information from M-CLD just for this test of going to be true operationally and what is the implication if the latter?

Lines 394 to 401 What are the implications of this, is it expected that M-CLD will improve or can/should some adjustment be made to the result before employing them in this manner or are these results a truer representation of the FLX accuracy and indicate its reliance on M-CLD c.f. that used in the classification? Is the 0.1 optical depth threshold found to perform best earlier still valid or does it need to be adjusted?

Line 393 You state the results are significantly worse, they seem to have a lot of scatter to them and I wonder to what extent they might improve if you increased the averaging domain (possibly because the cloud information is more accurate on these scales) or is this a result a biases in the cloud information. Can you relate this to known errors in the cloud information, do they meet their requirements are they expected to improve in orbit or be improved.

Figure 6. As was the case for figure 5 although the mean lines look to correspond to the stated values the lines indicating the RMSE and Std don’t seem to be correct. The same points about showing the errors in the context of retrieval uncertainty main in relation to figure 5 apply here, it would be good know if you can tell you are likely to have poorly retrieved fluxes.

Section 4

Can you clarify how your results relate to the goal requirements, I realise this may require some interpretation on your part. I don't think that the std can be related to the 10Wm-2 requirement unless you also consider the effect of the bias. Can points meeting the requirement be identified via a flux quality indicator or by excluding some subset of scenes or by averaging over a larger domain if so what size. Can you add some discussion of what happens next, if the scene id is expected to improve in orbit, if it needs to be altered or if you can tune your flux retrieval.

Minor corrections:

Line 3 design → designed

Line 4 remove 'an algorithm' (it is a processor specifically created I assume consisting of several algorithms)

Line 6 to 7 'measurements' It is not clear here if you are talking about the radiances or the fluxes or all the products. It might be clearer to change measurements to radiances if that is the intent, radiance and fluxes of simply all products depending on what you mean.

Lines 7 and 8 'of the atmosphere in cloud condition (reference level)' doesn't make sense. Do you mean '..or in cloudy conditions at a reference level which corresponds to the radiatively most significant vertical layer of the atmosphere'

Lines 12 to 14. 'The radiance to flux conversion algorithms have been successfully validated.....' Make clear this is done here in this paper and state the result obtained in the paper, something along the lines of. 'Validation of the radiance to flux conversion through end-to-end verification using L1 and L2 synthetic data for three EarthCARE orbits. The results .....'

Line 40 (and multiple other occurrences) change 10x10km2 → 10km x 10 km throughout. Could also be 10 by 10 km or square with side of 10km or 100km2 region (it is 100km2 but you are not presenting the equation so I think 10x10km2 is a bit strange. Wehr et al 2023 used 10km x 10km so this might be safest.

Line 44 'are challenging' → is challenging (challenging relates to estimation which is singular)

Line 50 'is created per scene and constructed from'. I'm not sure if you are trying to say that the algorithm is determined from scene stratified observations from CERES and MODIS, or it you are saying that it is a scene dependent algorithm or both, could you please rephrase to clarify.

Line 57 'the mission goal of 10 Wm^{-2}' please explain what the goal is.

Lin 62 'integrating the radiance field' need to add somewhere in the sentence that the integration is **over viewing and azimuth angles.**

Line 76 could you state the time period during which the 3 views are obtained to clarify the 'almost simultaneously'. Also I would suggest replacing 'providing a detailed view of scene anisotropy' to 'providing information on scene anisotropy' as whilst three views are helpful, they are not necessarily 'detailed' information.

Line 95, 'CERES instrument' → 'CERES data' (assuming you are using a data product) and state the version of the CERES product used and the years used.

Line 104 ‘scene definition concludes the number of ANN training sets’ I don’t think that concludes is the correct word here, do you mean determines?

Line 106 ‘relies on’ → ‘consists of’

Line 106 to 107 You say there are 6 static surface types but list 7 in the brackets, so either the number stated or the listing is incorrect

Line 107 ‘type’ → ‘types’

Line 108 can you somewhere in this sentence confirm the total number of scene types (I assume this is 72 but it would be good to state

Line 110 Are the NSIDC and NESDIS snow maps combined by you or are they combined by the SSF product? If the latter then change the sentence to say you use the CERES SSF snow infor which is a combination of these, if the former then explain how you combine them.

Line 137 ‘considered’ → ‘included’

Line 148 ‘estimates and original’…’estimates compared to the original…’ Do you have this estimate?

Line 152 you mention fore and after BBR observations, but presumably it also classifies the nadir?

Line 189 surely it is ‘fore and nadir’ from AATSR not ‘nadir and aft’, the along track view is in the forward direction and obtained before the nadir view.

Line 190 can you include a reference or further detail of the NB to BB used at least the channels employed.

Line 208 equation 4, is there a reason that there is no vza dependence indicated in R and L?

Line 214 ‘ideally reflected from the same atmospheric domain’ what does this mean ‘ideally’ and ‘domain’ specifically.

Line 219 what ‘internal consistency check’ can you please explain what is done, is this just referring to the minimization in 4 or something additional, if additional how is this different from 4?

Line 222 ‘higher cloud those observed’… not sure if you mean ‘higher cloud than those observed’ or something else.

Line 235 add to last case ‘and δ = 0 for the others’

Line 242 Add reference for the GERB issue with semi-transparent cloud issues: Dewitte et al 2008 https://doi.org/10.1016/j.asr.2007.07.042 and Clearbaux et al 2009 https://doi.org/10.1016/j.rse.2008.08.016.

Line 247 to 250. I think it worth noting here again the absence of a water vapour channel and the that the channel difference fills that role.

Line 252 ‘allows to obtain’ → ‘enable sufficiently accurate anisotropy models to be derived from…’ Also do we have any idea what the theoretical error is or any reference of what is possible from theory any reference for this assertion even if it is the LW errors achieved in GERB c.f. CERES for clear sky?

Line 253, might be worth stating that azimuthal dependence is negliglbe/neglected before stating R only is a function of VZA.

Line 256 can you put Velazquez et al 2010 on zelando and get a doi to make it accessible?

Line 270 'done at' → 'obtained at a'

Line 271 'would' → 'should' or 'will'

Line 273 'and statistically reducing the' → 'reducing the'

Line 281 'in average' → 'on average'

Line 286 Can you state what cloud properties are used, just fraction and height or something more?

Line 303 'chained...to' → 'interfaces...with'

Line 349 'As expected' is it expected in the model case specifically because it is using the plane parallel assumption can you specify that.

Line 354 Can you clarify here and in the caption to figure 5 and figure 6 if the BMA-FLX results shown are the combined view result.

Line 358 'employed both' → 'both employed'

Line 381 'the longwave results remain consistent...' → 'the longwave results are identical to' (?)

Line 420 'greater dependency' → 'dependency' (given the lw has no dependency.

Line 486 – The doi associated with this paper is for the unpublished version I think it needs to be updated to https://doi.org/10.5194/amt-16-5327-2023 unless you specifically want to reference the unpublished submitted version and not the final publication.