AMTD 8, 4049-4093, 2015

Item-by-item reply to Referee #1

Referee: yellow text
Authors: plain text

GENERAL COMMENTS
The manuscript is a follow-up paper to Masiello et al (2013), who presented a Kalman Filter approach for retrieving land surface temperature and emissivity from SEVIRI data. In the current paper the authors now focus on validating their KF-based retrieval against in situ observations at two stations and perform inter-comparisons with other data sources. The manuscript in general is well organized, and provides a coherent description of the performed work, although some copy-editing by a native English speaker would be beneficial. I remark on some language issues in the specific comments but was not able to address all of them. Overall the manuscript presents a quite interesting new approach for retrieving LST and emissivity from geostationary instruments and the presented validation and inter-comparison exercises are important for showing the capabilities of the proposed approach. The general topic of the manuscript fits well within the scope of AMT and I highly recommend the manuscript to be published once the following issues have been taken into account.
We thank the referee for the fruitful and thorough review that hopefully will lead to a an improved version of the manuscript

By far the strongest aspect of the paper is the validation of the retrieved LST products against in situ measurements at the sites in Gobabeb and Evora. It shows that the KF-based algorithm presented by the authors is capable of delivering SEVIRI-based LST with an accuracy which is more or less comparable to that of the existing LSA- SAF retrieval methodology. There are, however, several weaknesses in the current manuscript that should be addressed before a publication can be considered.
We will address this supposed weaknesses in the rest of this document.

The overall weakest point of the manuscript is the inter-comparison against other data products of surface temperature and emissivity and I think the paper requires major re- visions along those lines. In some cases these sections describe what are in my opinion apple-versus-orange comparisons such as for example the comparison between the KF-based retrieval of SEVIRI SST (skin temperature) with the bulk temperature from the OI-based Reynolds AVHRR product.
We think that this comment is not balanced in view of what we have actually shown in the paper. The comparison with the AVHHR OI SST is only an additional comparison which conclude a long analysis performed with MODIS and ECMWF instantaneous products for T_s . On average the difference Bulk-Skin is expected to be less than 0.3 K (e.g., Schluessel et al JGR, doi: 10.1029/JC095iC08p13341), which is what we have exactly found in our analysis. This 0.3 K average difference cannot mask the dynamical behavior of T_s,, that is amplitude and phase of the seasonal cycle, which is what we want to compare.

Furthermore the inter-comparisons are being carried out over monthly (e.g. Figure 13) and even yearly (Figure 14) averages. When using such long averaging periods the random error in the products is reduced considerably and the resulting statistics appear much more impressive than they are in reality.
We think this comment is lacking appreciation of the fact that our analysis has shown averages complemented with standard deviations. As the referee knows, the standard deviation measures the strength of the random component, which therefore has not been masked, but rigorously assessed. Moreover, we have shown the histogram of the whole population so that the reader can effectively see the amount and strength of outliers. In the revised version of the paper we will show the histogram of the SEVIRI/MODIS analysis as well (see Fig. 1 in this reply).

I think the authors should perform a proper spatial inter-comparison between the various products using swath-level data (for the LEO instruments). The 15 minute temporal sampling of SEVIRI makes the time-matching to the overpass time of MODIS/AVHRR quite straightforward to do. This is generally how sensor intercomparisons are being carried out in validation studies and it will provide a much more realistic estimate of the algorithm performance.

In fact this what we did. SEVIRI T_s has been directly inter-compared with MODIS (MOD28 Sea Surface Temperature 5-Minute L2 Swath 1 km) data points, see, e.g., the maps in Fig, 13 and 14 in the paper. To improve this analysis, in view of the referee comment, we will add the map of the standard deviation (see Fig. 2 in this reply) and, as already, said the SEVIRI/MODIS histogram (again see Fig. 1 below). In this way, the level of data aggregation used includes: a) actual differences with no space-time average (the histogram), b) time average (the maps); c) spatial average and time average, monthly sequences.

I also had the feeling while reading through the paper that the authors in some cases described only those those statistics that appeared favourable for their algorithm while ignoring others that would show their retrieval in a less positive way. In similar sections a bit further on, the statistics then showed a slightly different picture and those statistical metrics that previously were considered important are not mentioned anymore. I urge the authors to be more consistent in showing and describing all relevant statistical metrics in the text and to openly discuss them, even if they do not necessarily support the hypothesis that the suggested retrieval is superior to existing ones.

Again we think that this comment is a bit unbalanced in view of what we have shown in the paper. The sections for Evora and Gobabeb show, in fact exactly the same kind and type of figures and statistical analysis (including the linear regression fit shown in Fig.s 4 and 8). The referee should also consider that our approach simultaneous, and dynamically, retrieve T_s and ε, when comparing it to existing ones.

There are also several statements in the current manuscript that are not properly sup ported by the results (see specific comments). I encourage the authors to go over the manuscript again and make sure that every claim is backed up by real data and analysis (and that adequate evidence is provided in the text particularly when making bold claims).

Finally, while surface temperatures are validated against in situ station data and com- pared against other data sources, the KF-based emissivity products are not validated to the same extent. I think it would be beneficial for the paper to perform a direct spatial comparison of the full-disk emissivity maps (Figures 18-20) against existing emissivity maps from other data sources, e.g. from MODIS or ASTER GED.

In the revised paper a comparison of the emissivity maps with those derived from the MODIS based UW-BFEMIS data base will be shown. The comparison refers to the year 2007. An example of the map of differences at 10.8 μm is shown in Fig. 3 at the end of this document.

SPECIFIC COMMENTS

P4050L12: "yields an accuracy of". What is this metric? RMSE? Please be specific.

Yes, rms. We have corrected in the revised version

P4050L17: "we have that emissivity". Something is superfluous here. Please revise this sentence.

We have rephrased.

P4051L7: "exemplified" -> "demonstrated" or similar

Changed to demonstrated

P4051L16: "is a breakthrough". Isn't that a bit grandiose? Especially since the KF algorithm provides similar accuracy as the LSA-SAF one? Please be modest and let the community decide on the merit of your approach.

This is a novel approach. We have rephrased.

P4051L22: "boosting" This sounds odd, please consider replacing with "complement ing" or similar
Changed to complementing

LP4051L29: Namib -> Namibia
Changed to Namibia

P4052L4: "analyses for sea surface" Please be specific about what products you are using.
Corrected to sea surface temperature,

P4052L7: "reference"? What makes it reference? Consider dropping this word.
Reference deleted

P4052L8: Better write "The results from the validation exercise are reported in Sect. 4"
Changed as suggested

P4052L9: Section 2 needs a section on the actual SEVIRI data you used and how it was pre-processed before applying the Kalman filter. You need to provide information on how you performed cloud-masking and similar processing issues.
The revised version will have a new paragraph (beginning of section 2) to deal with SEVIRI data source and cloud mask. SEVIRI data were downloaded from the EUMETSAT UMARF platform. We have added a new paragraph in section 2 of the revised version to clarify data source.

P4052L12: "Diverges"? Better write "increases" or similar.
Changed to increases.

P4052L16: "reasonable" Says who? Have you tested this? How did you arrive at the 70 degree threshold? This needs justification.
The limit of 70° is normally agreed with to be reasonable for a plane parallel atmosphere based on the classical and well-known work by Herman et al Applied Optics, 33(9), pp 1760-1770, 1994 (reference has been now added to the revised manuscript).

P4054L4: "sequence of data" -> "periods". Also, this entire sentence reads very awkward. Please revise or consider splitting up into two sentences.
In the revision, the sentence has been rephrased.

P4054L9: Please be specific. Which MODIS product did you use? MOD28?
Yes, MOD28 Sea Surface Temperature 5-Minute L2 Swath 1 km. This will be clarified in the revised paper.

P4054L9: I think this OI-based product is not suitable for comparison with your SEVIRI retrieval as it provides bulk temperature not skin temperature.
We do not agree with. Bulk and skin T_s have peak difference within ±1 and average differences of about 0.3 K (e.g., Schluessel et al JGR, doi: 10.1029/JC095iC08p13341), which is perfectly in agreement with the values of 0.30 K we have found in our analysis,

P4054L24: This section needs more information about how the interpolation in space and time was carried out.
It is a linear interpolation in time. We will clarify in the revised paper.

P4055L2: Please make a note here that when you refer to full disk you actually mean all pixels with ZVA < 70 deg
We will clarify in the revised version of the paper.

P4056L7: (T, Q, O) Please define those parameters
Temperature, humidity and ozone profiles, We will clarify in the revised version

P4056L12: Again, please provide information about how the interpolation was carried out.
It is a linear interpolation in time. We will clarify in the revised paper.

P4056L24: Replace the semicolon with "and"

Replaced

P4057L7: Please provide a short explanation about why you perform a logit-transform.
A short explanation will be added.

P4058L4: Either "posterior" or "a posteriori"
It will be changed to a-posteriori

P4060L14: "clearness" -> "clarity"
It will be changed to clarity

P4060L23: Please also provide the RMS difference also in Figures 4 and 8
Rms (root mean square error) has been now provided in Fig. 4 and 8.

P4061L3: "very few data points". Please provide the number of data points going into each month (maybe in the Figure? or at least as a rough range in the text). Also, describe what do you consider too few data points to be significant and how did you arrive at that number?
Because of clouds. It will be clarified in the text. As said in section 4.1 we selected only the retrievals which reached convergence according to the Cost Function χ^2 criterion.

P4062: "we now see a striking improvement". This is worded a bit strangely - it's not really an "improvement" as such since the plot is for an entirely different station. Also it does not really appear "striking" to me. Maybe write something like "we can observe a closer agreement of the linear fit line with the 1:1 line than for Figure 4" or similar.
We will rephrase.

P4062L10: "slope close to unity". The slope was not used as a metric for Evora, maybe because there it worse than for LSA-SAF? Please be consistent when describing statistical metrics.
The slope for EVORA has been consistently shown for Evora in Fig. 4, the same as in Fig, 8. The slope does not say anything new if not already shown with bias, SD and rms. A straight line is determined by two parameters: slope and intercept. We understand that the description added in the Gobabeb section can be misleading and we have dropped it in the new version. The deviation of the linear fit from the bisector is a sign of bias, and for Evora, the KF bias is less than that of LSA SAF.

P4062L14: I believe you mixed up KF and LSA-SAF here? I fail to see how KF is supposed to "perform slightly better" than LSA-SAF. The bias for KF is twice as high as for LSA-SAF??
Yes, it is the reverse. We will correct.

P4064L11: Why monthly averages?
We want to check for possible seasonal dependence. Also, please, consider that the average is complemented with the Standard Deviation, therefore, together they synthesize the salient feature of the given sample: bias and random component.

P4064L18: A yearly average? This reduced the random error so much that the resulting metrics are not indicative of the true performance of the algorithm.
Once again, this sounds a bit inappropriate; we gave also the Standard Deviation, which in fact is the strength of the random component. Moreover, we now have added the histogram to the revised version of the paper (Fig. 1 in this reply).

P4065L5 and this entire section: Again, I think you need to use an actual SST (skin temp) retrieval to perform such a comparison. You cannot validate a skin temperature product with an OI-based bulk temperature product.
SEVIRI vs AVHRR OI SST is a further comparison, which comes after presenting the comparison between SEVIRI and MODIS. Again, the referee opinion contrast with the well-assessed fact (e.g. Schluessel et al JGR, doi: 10.1029/JC095iC08p13341) that bulk and skin temperature differs on average within $\approx$ 0.3 K. This bias is in agreement with our findings, a fact that supports the validity of the KF scheme. We have clarified the matter in section 4.3.3 (revised version of the paper). We think that the difference between bulk and skin has been introduced and discussed at a length in section 4.3.3, and no confusion could

arise. To improve clarity of the presentation we have also re-titled this section in *Comparison with AVHRR OI SST analysis.*

P4066L4: "really run" -> "really be run"
It will be changed to *really be run*

P4066L14: But in the introduction you say that "the limit of 70 degree is reasonable". I think you will need to perform a proper analysis of this.
Please do not confuse SEVIRI instrument limitations with those of the algorithms.

P4066L20: "seas of sand". This is very poetic but probably not too suitable for a scientific article.
Sea of sand or Sand seas is the scientific way to refers to the formation of large sand dunes in the Sahara and other deserts.

P4067L6: "exemplified". Better write "demonstrated" or similar
Changed to demonstrated.

P4067L10/11: But the behaviour of the algorithm in the presence of data gaps has not been tested specifically. I think this statement is too bold without actually testing this in detail. Please consider revising.
We disagree with. We will add a paragraph to the beginning of section 4 to clarify the matter. KF is sequentially operated and it is never re-initialized; it goes through data voids and undetected cloudy radiances until it reaches the last data point in the record.

P4067L13: "better than 1.5C" This needs to be qualified. Over which surfaces? Based on which comparison?
Evora and Gobabeb, of course

P4067L14: "bias less" -> "bias of less"
It will be corrected.

P4067L23: Again, this has not been tested properly. Either add some text or figures in the results section specifically studying the impact of the KF-based approach in the presence of clouds or consider revising this statement.
See point P4067L10/11.

P4068L1: "safely applied to the SEVIRI full disk". This statement is not supported by evidence. In fact in Section 4.4. the authors themselves suggest that unrealistically low temperature values in South America could be due to the "SEVIRI point spread function at this angle". Without specifically testing the angle-dependence of the algorithm properly I do not think that this statement holds.
Again, please do not confuse SEVIRI limitations with KF capability. This is a problem of the instrument. The problem has to be checked, but has nothing to do with the algorithm.

Figure 4 Caption: R^2 is not the linear correlation coefficient but the coefficient of determination
They are the same thing.

Figures 5-7: The caption should mention that this is for Evora. Figure 6b: Move the legend so it doesn't cover the time series
No data has been covered. We have resized the legend, eliminated the box and left only the legend text for clarity.

Figure 7: I think these plots should show both the actual data (maybe in a light color) and over it the time series smoothed with the moving average filter. Also, please show the corresponding emissivity from other sources (e.g. UW and/or LSA-SAF) for reference.
Ok for the unsmoothed series. Based on a request by reviewer 3, they will be shown with a 3 hours window in order to provide a proper comparison with Gobabeb emissivity (Figure 10 in the paper). For the rest, there is no LSA-SAF emissivity retrieval to compare with. UW-

BFEMIS or similar emissivities with 15 minutes or 1day resolution simply do not exist.
Figure 8 Caption: R^2 is not the linear correlation coefficient but the coefficient of determination
Again, they are the same thing.

Figure 11a: It could be helpful to show here also the time series of the difference between KF and in situ temperature, in order to be able to discern where they differ (this is hard to see in Figure 11a).
The difference will be plotted in the revised manuscript.

Figure 12: This histogram would be a lot more useful if you limited the x axis limits to something like the range of -10 C to 10 C or so and increased the number of classes/bins.
We have limited to -10 to 10. Actually the class bins have width of ≈0.5 K, we have increased classes (see Fig. 1 in this reply).

Figure 13 Caption: Note that this is for 2013
Noted.

Figures 18-20. These emissivity maps are nice. You should perform a spatial comparison (e.g. difference maps) of these maps against other emissivity database, e.g. from MODIS or ASTER.
A comparison will be performed and shown with the UW-BFEMIS database derived from MODIS (e.g. figure 3 in this report) for the year 2007.

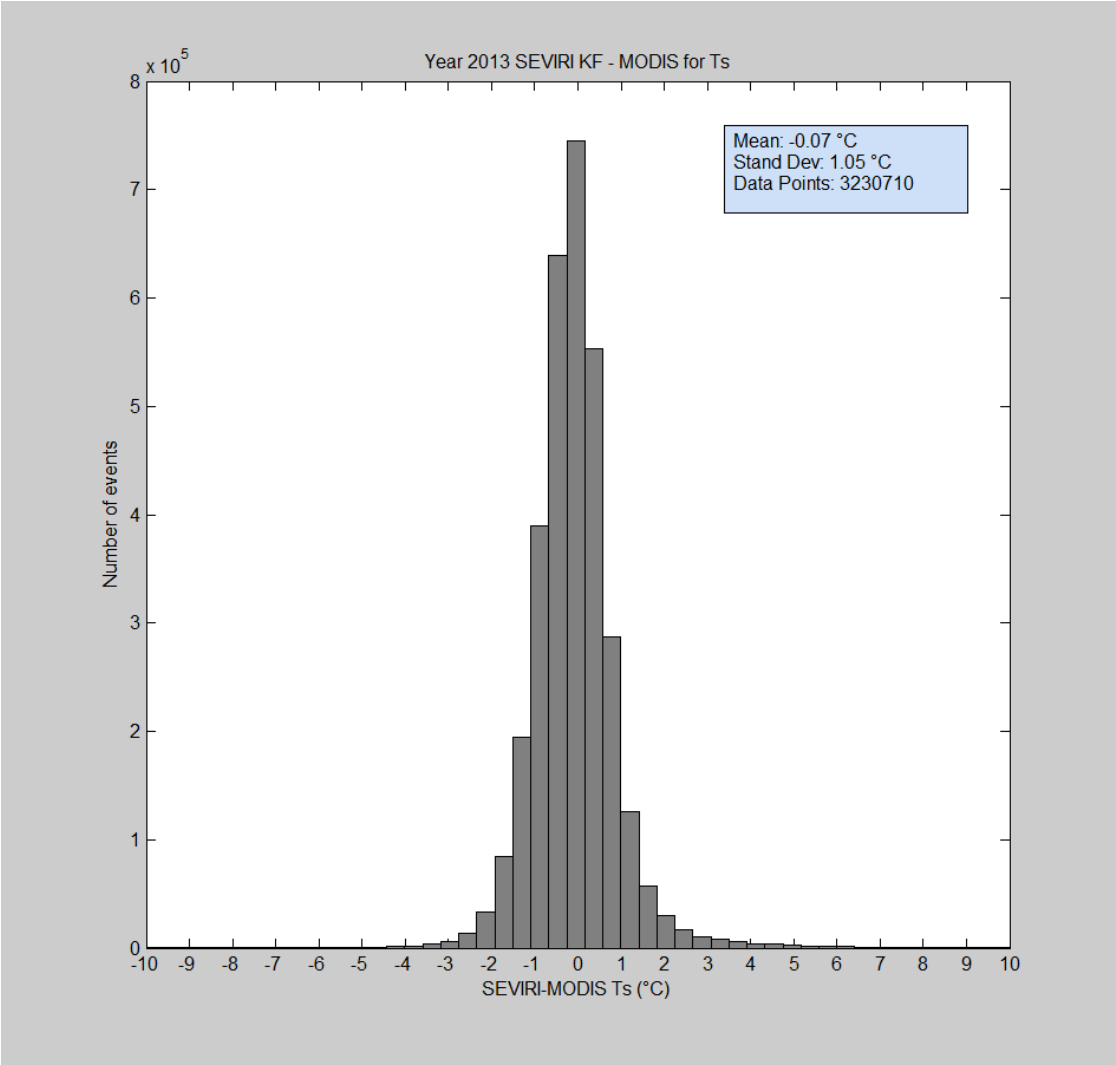

Figure 1 (will be shown also in the paper). Histogram of the difference of the skin temperature, SEVIRI-MODIS.

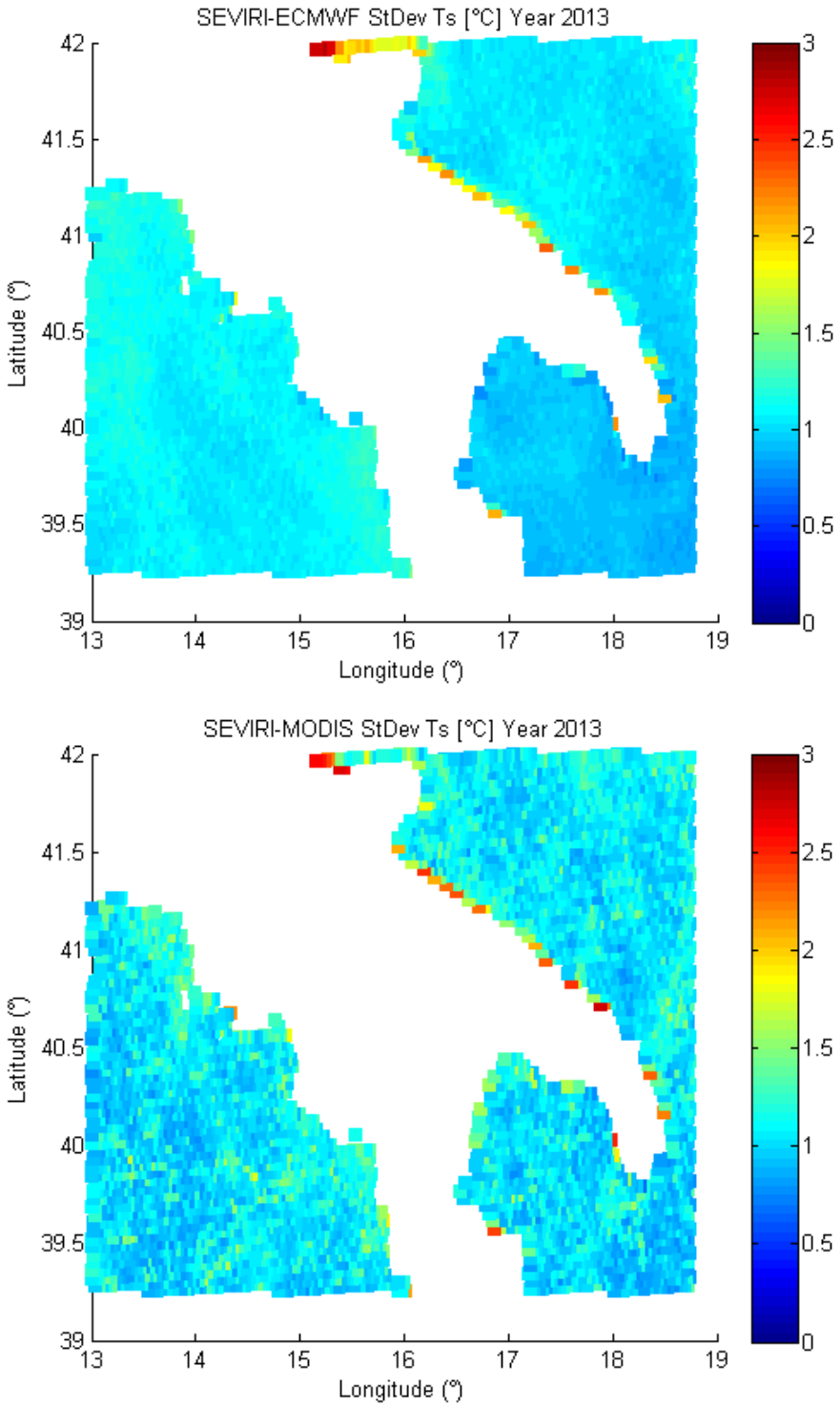

Figure 2 (will be shown in the paper). Comparison of SEVIRI retrieved skin temperature with ECMWF analysis (above) and MODIS (bottom). The maps shows the standard deviation of the skin temperature difference.

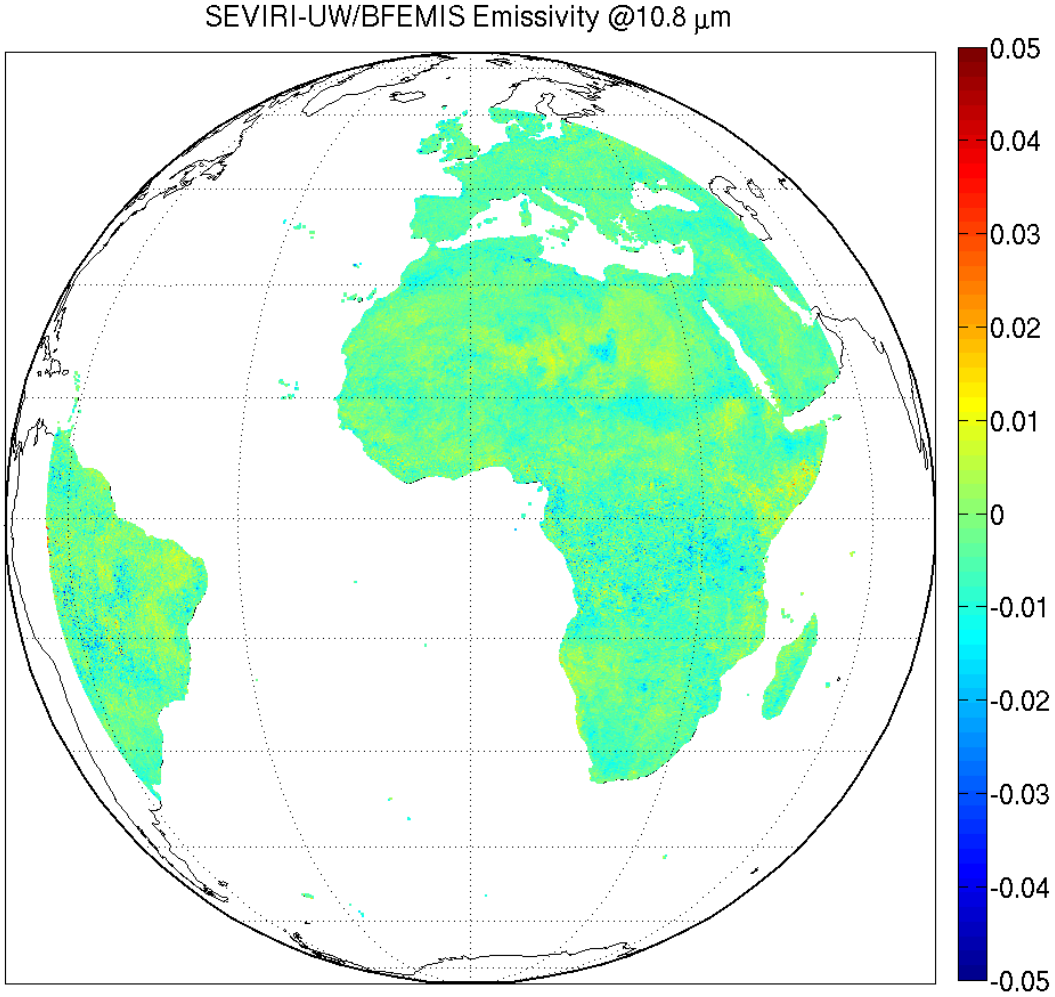

Fig. 3 (will be shown in the paper together with the maps at 8.7 µm and 12 µm). Map of the difference of the monthly average emissivity at 10.8 µm between SEVIRI KF and UW-BFEMIS for November 2007. UW-BFEMIS has been remapped to SEVIRI using a high spectral resolution algorithm which is first applied to the ten hinge points UW/BFEMIS emissivities to generate the high spectral resolution emissivity spectrum, which, in turn, is convolved to the SEVIRI spectral response. Details of this procedure can be found in Masiello 2013b, 2014 (reference list in the paper).