Response to Reviewers

Anonymous Referee 1

The manuscript provides an assessment of land surface temperature (LST) simulated with different configurations of the UK Met Office Unified model. The exercise is made for a small area in the US (Arizona) taking advantage of simulations and data gathered for a particular experiment (SALSTICE), which was focused on model LST bias with respect to IASI retrievals. Model LST simulations are compared with IASI and MODIS products, as well as with in situ estimates. Model net radiation, turbulent heat fluxes at the surface and ground flux are also compared with ground observations.

The manuscript is very well written and the subject is of interest, given the limitation in the assimilation of radiances sensitive to lower troposphere over land due to the large model skin temperature biases. However, it is difficult to draw solid conclusions when different model configurations (in terms of dynamics, resolution, approach to bias correction, surface parameters) are not run for a common period. I suggest the article to be accepted subject to revisions in line with my comments below.

We thank Reviewer 1 for carefully reading our paper and providing recommendations and comments on how to improve the manuscript. We believe that the advice in this review is very useful, and contributes to a substantial improvement of the article.

1) On local estimates of LST (section 2.2.2): I fully agree with the need to account for the uncertainty in local emissivity for the LST ground estimates. From the description provided in this section, it seems you do not correct the surface leaving radiance for the downward radiation that is reflected by the surface. This may be the same order of emissivity uncertainty for the 8-14 micro-m band. Please check and modify the data and model versus in situ comparisons in the manuscript as needed.
Thank you for this suggestion, we now apply a further correction to the IRT in-situ data which accounts for the 8-14 µm downwelling longwave radiation according to Eq. (1).

$$BT_{surf,8-14\,\mu m} = \frac{1}{\varepsilon}\left(LW^{\uparrow}_{surf,8-14\,\mu m} - (1-\varepsilon)LW^{\downarrow}_{surf,8-14\,\mu m}\right) \quad (1)$$

where $BT_{surf,8-14\,\mu m}$ is the surface blackbody radiance, ε is the emissivity in the range of 0.97 $\pm$ 0.02, $LW^{\uparrow}_{surf,8-14\,\mu m}$ is the upwelling radiance at the surface in the IRT field of view, $LW^{\downarrow}_{surf,8-14\,\mu m}$ is the downwelling radiance at the surface which is reflected into the IRT field of view.

The 8-14 µm downwelling longwave ($LW^{\downarrow}_{surf,8-14}$) is modelled using the Havemann-Taylor Fast Radiative Transfer Code (HT-FRTC) (Havemann, 2006) for each of the ground sites, Lucky Hills and Kendall Grassland, which have an IRT installed. Hourly downwelling longwave radiation is calculated based on the ECMWF ERA-Interim (Dee et al., 2011) which is available every 6 hours (00, 06, 12 and 18). For the other times the ECMWF ERA-Interim atmospheric profiles have been interpolated in time. The downwelling calculation uses the 8-14 µm spectral emissivity for sandy soil from Arizona from UCSB (University of California, Santa Barbara) Emissivity Library (UCSB Library)

(https://icess.eri.ucsb.edu/modis/EMIS/html/em.html).

The IRT measurements were found to be on average (of the six years) -0.51 K colder when accounting for the reflected downwelling average for the 6 years; the smallest impact was found for the 2014 measurements (-0.43 K) and the largest impact was found in 2015 (-0.59 K).

We have revised the data presented in Figure 2&3 and modified the IRT data referred to throughout manuscript.

Two additional references have been added;

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.,

Park, B., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q.J.R. Meteorol. Soc., 137: 553-597. doi:10.1002/qj.828, 2011.

Havemann, S.: The development of a fast radiative transfer model based on an empirical orthogonal functions (EOF) technique. Proc. SPIE 6405: 64050M, doi: 10.1117/12.693995, 2006.

2) End of section 2.3 (page 6): The angular dependence of LST estimates should not be linked to atmospheric effects, as these should have been corrected during the retrieval process. Although insufficient correction for the optical path may still persist, the effects described in the text are more frequently a consequence of spatial heterogeneity (i.e., different viewing perspectives may actually yield different scenes, even if matching in time and space) and therefore are essentially dependent on the viewing & illumination geometry. That is why angular-dependent biases are mostly inexistent for night-time observations.
Thank you for highlighting the confusion in the text, we have revised the manuscript and added additional references as follows. "The angular dependence described arises due to different viewing and illumination geometry of the surface; studies have shown that factors including slope orientation relative to sun, properties of the soil and vegetation such as surface heterogeneity and the structure of the vegetation canopy, all contribute to the directional anisotropy (Duffour et al., 2016; Ermida et al., 2014; Rasmussen et al. 2010)."

Duffour, C., Lagouarde, J.P., Olioso, A., Demarty, J., Roujean, J.L.: Driving factors of the directional variability of thermal infrared signal in temperate regions., Remote Sens. Environ., 177, 2016.

Ermida, S. L., Trigo, I. F., Dacamara, C. C., Göttsche, F. M., Olesen, F. S., Hulley, G.: Validation of remotely
sensed surface temperature over an oak woodland landscape - The problem of viewing and illumination geometries., Remote Sens. Environ., 148, 16–27, 2014.

Rasmussen, M. O., Pinheiro, A. C., Proud, S. R., and Sandholt, I.: Modeling Angular Dependences in Land Surface Temperatures From the SEVIRI Instrument Onboard the Geostationary Meteosat Second Generation Satellites., IEEE Trans. Geosci. Remote Sens. 48, 3123-3133, 2010.

3) Lines 3-8 (page 8): Indicate how the change in the emissivity attributed to each tile (bare ground, grasses) change the emissivity map over the study area. I'd say that overall you have a slight decrease for GA/L6.1 and US2.2(A to D) and an increase in US2.2E due to the drastic reduction of bare ground fraction.
To summarise the emissivity changes, an emissivity map of the study region for each configuration is presented in Supplement Fig. S1. The emissivity changes relative to GA/L3.1 (Fig. S1a) and US2.2_ConfigA (Fig. S1d) result in regional decreases for GA/L6.1 (Fig. S1b, Fig. S1c) and US2.2ConfigA-D (Fig. S1e) associated with regions of larger bare soil fractions. US2.2ConfigE (Fig. S1f), in contrast, shows an increase in emissivity for the study domain related to a reduction in the bare soil cover fraction. Section 3.3 provides a more thorough discussion of the surface heterogeneity and land cover in each model configuration.

4) Lines 10-15 (page 8): Please include a short justification for the use of different ZoH/ZoM ratios in the global and limited area model versions.
The z_{OH}/z_{OM} ratio was revised between GA/L3.1 and GA/L6.1 in order improve both land surface temperature and near surface air temperatures in desert regions. The revised z_{OH}/z_{OM} ratio was adopted in the US2.2 (and other LAMs) from 2013, whilst GA/L6.1 was adopted for operational use in July 2014. We have included this text in the manuscript.

5) On the overall analysis of model simulations: As referred above in my general comment, the results of different model configurations correspond to the same period of the year (May), but for model runs performed for different years. You must ensure that when comparing these results, they are not affected by inter-annual variability. In order words, please show that the conditions observed in each May of the 2013-2018 period do not deviate greatly from the average. In the case they do, please check how that may have affected your results. This is relevant, since a dryer or rainier than usual year may lead to a significant change in vegetation cover (and therefore in surface parameters such as surface albedo and emissivity, and even ZoM), soil moisture availability (and likely in the partition between latent and sensible heat fluxes), which will certainly impact your model performance.

Thank you for this suggestion to investigate if the different case periods are affected by inter-annual variability. We have approached this by examining the in situ soil moisture measurements and using the soil moisture anomaly product from Climate Prediction Center for the six year evaluation period.

We have included in the manuscript "Variability of surface temperatures could arise due to variability in cloud cover or soil moisture. In this study we consider only clear sky situations; both the model and observational datasets have been screened to remove cloud contamination, which suggests that soil moisture variability between the analysis years could be a factor for investigation. Point scale measurements of volumetric soil moisture at the eddy-covariance sites are made at depths of 5 cm and 15 cm. A six year multi-year mean soil moisture for each site and at each soil depth has been calculated, and used to calculate a soil moisture anomaly. At both sites, the volumetric soil moisture in May is less than 0.05 kg m^{-2} (0.10 kg m^{-2}) at 5 cm (15 cm) for all years in the evaluation. The in situ volumetric soil moisture measurements suggest that the moisture levels were almost always exhausted for each May analysis period and therefore it is unlikely there was sufficient soil moisture to impact on surface temperature variability.

In support of the eddy-covariance measurements, monthly 0.5°×0.5° soil moisture and soil moisture anomaly product from Climate Prediction Center (Fan et al. 2004) were used to assess the larger scale trends in soil moisture in southeastern Arizona. The soil moisture anomaly product indicates that May 2013 and 2014 were anomalously dry (-20 to -40 mm) for an extensive region of the western US, May 2015 had a neutral soil moisture anomaly, May 2016 and 2017 had localised dry regions confined within Arizona, and May 2018 was anomalously dry (-80 mm) for an extensive region of the western US."

Reference added: Fan, Y., and van den Dool, H.: Climate Prediction Center global monthly soil moisture data set at 0.5° resolution for 1948 to present, J. Geophys. Res., 109, D10102, doi:10.1029/2003JD004345, 2004.

6) Last line of page 9 – line 7 of page 10: I'm not sure I follow what is meant here, especially with what respects the degradation in the representation of the grassland fractions. In contrast to the latter, when use a higher resolution land cover, you get better representation for bare soils: is this so? Why? Please clarify (or just rephrase).
Thank you for highlighting the confusion in the text, a similar comment was also raised by Reviewer 2. We have revised this paragraph to compare to modelled bare soil cover with the observation fractions from Scott et al. 2015, and only reference the changes in surface fractions to the four sites and not to the land classification. The revised paragraph is as follows: "The higher resolution ancillaries in the US2.2 improve the surface fractions for the two shrubland sites; the US2.2 increases the bare soil fractional cover which acts to increase the sparsity of the vegetation cover, and improves the model representation of the surface heterogeneity. At the Lucky Hills shrubland site, for example, the bare soil fraction is increased from 0.26 (GA/L3.1) to 0.48 (US2.2_ConfigA-D) and at Santa Rita Mesquite a similar increase from 0.22 (GA/L3.1) to 0.37 (US2.2_ConfigA-D) is reflected. This brings the modelled bare soil cover fractions closer to the observed fractions of 63 % for Lucky Hills Shrubland and 50 % for Santa Rita Mesquite (Scott et al., 2015). However, at the two grassland sites, Kendall Grassland and Santa Rita Grassland, there was a reduction in bare soil fractional cover between GA/L3.1 and US2.2_ConfigA. The lower cover fraction at the grassland sites is maintained in all GA/L6.1_17km configurations. At the Kendall Grassland site, for example, the bare soil fraction is decreased from 0.26 (GA/L3.1) to 0.20 (US2.2_ConfigA-D) and at Santa Rita Grassland a similar decrease from 0.16 (GA/L3.1) to 0.10 (US2.2_ConfigA-D) is reflected. This is in contrast with the observed fractions of 60 % for Kendall Grassland and 45 % for Santa Rita Grassland (Scott et al., 2015)."

7) When discussing the statistics between the various model configurations and MODIS LST products (collections 5 and 6), it would be useful to have an idea how both compare with the in situ estimates (please make sure these are properly estimated, as commented above). You may consider adding a table with a summary of all these, including an average of the in situ (or MODIS) LST per site, which would somehow answer my question above on stable the conditions are among the studied years. This may also help you check if there are years/sites for which MODIS (Aqua or Terra) presents higher biases, and therefore help you analysing your model comparisons with MODIS LST.
Thank you for this suggestion to make a more direct comparison between the model configuration, MODIS LST and the in situ IRT measurement. We have included the in-situ measurements in Figure 3 to enable this comparison, rather than adding a table of the data.

In the text we have included more reference to how the IRT measurements compare with the model and MODIS LST. In section 3.2, paragraph 3 we include: "The IRT measurements support this trend; at Lucky

Hills the bias in reduced from -9.0±3.7 K (GA/L3.1) to -3.3+2.3 K (US2.2_ConfigA), whilst the IRT measurements at Kendall Grasslands only show a 2.2 K improvement in the US2.2_ConfigA compared with GA/L3.1." In section 3.2, paragraph 6 we include: "The IRT measurements located at Lucky Hills support the development of the warm bias (0.6±5.4 K in 2013; 1.4±2.6 K in 2015)".

8) On the assessment of model biases and terrain slope: The impact of slope, especially the x-component) surely differs for Terra (morning overpass) and Aqua (afternoon overpass), if this is essentially related to the LST contrast between slopes facing/hiding from the sun. Maybe this effect is more noticeable in the afternoon, and in that case the "Aqua signature" prevails. In any case, the illumination geometry is obviously very relevant for this, and therefore results should be assessed for the two platforms separately.
Thank you for this suggestion. We have separated the Terra and Aqua platforms for calculating the daytime coefficients of correlation for the x-component and y-component of the orographic slope, and have adjusted Figure 6 accordingly and the caption for Figure 6 "N.B In panel b (c) the collection 6 (dotted) Terra and (dot-dashed) Aqua retrievals are separated for presenting the correlations with the x-component (y-component) of the orographic slope."
Within the manuscript we have revised the text as follows; "The coefficients of correlation between the LST bias and x-component (y-component) of the orographic slope have been calculated for the six year analysis period and are presented in Figure 6b (c). The solar illumination geometry of orography changes as a function of time of day, whilst the remotely sensed LST is a directional variable with each satellite platform (Terra and Aqua) maintaining the same angle with respect to the sun. Each platform measures a similar illumination geometry on each overpass, and therefore the coefficients of correlations are calculated separately for the Terra and Aqua retrievals in Figure 6b and 6c. The night-time coefficients of correlation have a value of +/-0.2 which indicates there is a relationship between the two variables, but it is weak and likely insignificant. For the x-component prior to 2018, the daytime coefficient of correlation was positively correlated with a value of 0.41±0.05 (0.28±0.05) for Aqua (Terra) retrievals; and identifies that regions of cold model LST bias are found on easterly slopes and regions of warm model LST bias are found on westerly slopes. We find a stronger correlation between the x-component of the orographic slope and the LST bias for Aqua compared with Terra, whilst the difference between the two platforms was minimal for the y-component of the orographic slope."

9) Lines 15-16 of page 14 "Our findings suggest that the daytime model LST bias could be minimised by increasing the bare soil cover fraction in the study regions". I don't think you can say this, as you are suggesting that you should change the fraction of bare soil, instead e.g., correcting, e.g., model parameters where the fraction of bare soil is low.
Thank you for this comment, by this we mean that our work has identified that the surface cover ancillary datasets do not adequately represent sparse canopies as the bare soil fractions are too low, and suggest that new developments of ancillary datasets should take this into account. We have revised this sentence as follows; "Our findings suggest that the development of surface cover ancillary datasets for sparse canopies is necessary."

10) The comparisons between model and observed net radiation and surface energy fluxes are only discussed for a single site/year. Although the issue of different models run for different periods could make the discussion difficult, it would be interesting to know how the comparison between simulations and observations evolved as the model land surface temperature changed.
Thank you for this suggestion and we agree that investigating the surface energy balance beyond the 2013 analysis we have presented in this manuscript is important. However, as you indicated it is difficult to draw conclusions for the impact of different model parameters when we are examining different time periods. Rather than interpreting the changes to the surface energy balance with the operational coupled configurations presented in this study, we are doing a follow on study which uses offline/standalone JULES driven with observations from the AmeriFlux network and for a greater number of sites, which will enable us to examine the response of the surface energy balance in greater detail. However this follow on work is beyond the scope of this manuscript, and will form a separate publication.

11) Editorial:

- Abstract: "The diurnal cycle of LST in Global Atmosphere/Land 6.1 (GA/L6.1) showed a significant improvement relative to GA/L3.1": Please be more specific (meaning quantitative) here.
Thank you for this comment. We have revised the sentence to "The diurnal cycle of LST in Global Atmosphere/Land 6.1 (GA/L6.1) showed a significant improvement relative to GA/L3.1 with the cold LST biases reduced to -1.4±2.7 K and -3.6±3.0 K for Terra and Aqua overpasses, respectively."

- lines 5-6 (page 6): Suggest replacing “to give site-specific LST for each site.” by “to give site-specific LST.”.
Changed.

- Figures 4 and 5: suggest the authors include a short title for each panel (e.g., LST bias – US2.2A), to facilitate their interpretation.
Thank you for this suggestion. We have added a title to each panel in Figure 2, 4, 5 and 7.

- line 23 (page 12): “IASI”
Changed.

- Figure 7: Please ensure each individual scatter-plot has the same range in the y- and x-axes, since we are comparing the same variable (model versus observations). For the same reason, please resize the diagrams so that they are closer to a square, i.e., so that the length in the y-axis corresponding to, say, 10 Wm^{-2} roughly matches the same length for 10 Wm^{-2} in the x-axis.
In Figure 7 we have changed the range of the latent heat flux plot to have the same axis range, and replotted the each subplot so they are square.

- Line 12 page 18: Please rephrase sentence.
We have rephrased the sentence as follows; “With recent advances in supercomputing power, the ability to perform high resolution ensemble forecasting, for example within a research LAM such as the US2.2, is becoming viable. This will provide an opportunity to evaluate the impact of forecast uncertainty on the land surface processes, rather than only for the deterministic forecast as has been carried out in this study.”