Author's response

We thank the Reviewers and Editor for their respective feedback.

Point-by-point Answers

We thank the Reviewer for the useful and detailed comments. We followed the suggestions and modified our manuscript as described in the line-by-line answers below. Note that the reviewer's comments are in black, our answers are in red, and the changes made to the manuscript are in red italic.

The authors have addressed some of my questions, which were mostly related to some details not being clear in the text. I still think there are limitations in the in-situ validation and that more stations with similar climate should be added to the paper. I would even remove stations at lower latitudes. See below point-by-point reply.

1. I understood that the authors select the satellites in order to reduce time differences due to orbital drift and that there is no overlap between satellites. But how do you handle orbit overlap from the same satellite? At the poles there is significant overlap of the orbits resulting in multiple observations throughout the day. Are these all averaged? Do you select a specific orbit depending on the time?

For each day and pixel, for each satellite, the dataset contains values selected from two single orbit files (one for daytime and one for nighttime), i.e., no averaging is performed. The selected observation is the one closest to the zenith (NADIR condition). This ensures that the observations are made at the same viewing conditions and at nearly the same local time at a hemispheric scale. The subsection *2.1 EUMETSAT AVHRR FDR* now clarifies this point.

2.1 EUMETSAT AVHRR FDR

For each satellite, the orbit files are composited by choosing for each pixel only the observation closest to nadir. The composited files have a spatial resolution of 0.05° x 0.05° pixel size and are available for each satellite twice a day (at daytime and nighttime). This study focuses on the pan-Arctic region, therefore only data above 50° N have been processed.

3. If the calibration is limited to your area of interest, then why use stations outside such area to validate the LST?

The radiative transfer modelling is done independently for each class of TCWV and Tskin, and the corresponding split-window coefficients (SWC) are then obtained for each of these classes in an independent manner. Therefore, for each pixel, the calibration is performed for the adequate class of TCWV and Tskin, i.e., it is not based on an area of interest. The Pan-Arctic pixels are calibrated with classes corresponding to actually encountered atmospheric

conditions; the same logic applies to pixels over the USA, which are then also validated with SURFRAD stations. We modified our description of the calibration process (*3.1 Generalized Split Window Algorithm*) to clarify this point and remove the sentence that is confusing in *2.4 Auxiliary data.*

3.1 Generalized Split Window Algorithm

Table 3 summarises the construction of the simulation dataset. Finally, the calibration was performed independently for each class, and for each class, the samples were split into a training (70%) and test (30%) set, and multilinear regression was performed on the resulting BTs. Based on the test sets, look-up tables (LUT) with coefficients are created for each satellite.

4. Again, the authors agree that the SURFRAD stations have large representativeness issues, so I do not see how they can be "top-tier". Also, a lot of BSRN stations have 1 minute sampling so I don't understand why the authors don't want to use them. There are also a few high latitude stations available from the Copernicus Ground-Based Observations for Validation (GBOV) service. Regarding the accuracy of emissivity, I believe this is true for all stations and, to my knowledge, none of the stations used by the authors have in-situ measured emissivity. There are specific limitations of the retrievals close to the poles than can only be properly assessed with stations in these areas. For instance, cloud contamination tends to be much high for night-time snow covered surfaces and, as such, selecting stations with these conditions could help clarify how much cloud contamination could be in the data.

The SURFRAD stations are regarded as 'top-tier' mainly because of their long and quality-controlled time series. Furthermore, the surface around them (land cover types, seasonality, etc.) has been studied in detail. A lack of spatial representativeness is something that SURFRAD stations share with BSRN stations: this can be partially overcome by performing validations only at night-time. However, there is a lack of high-quality in-situ data obtained in large, spatially homogenous areas. GBOV mainly collects data from existing stations (SURFRAD, BSRN, e.g.). Therefore, GBOV data have the same issues, e.g., length of time series, quality, suitability (sensors) and representativeness. The emissivity at densely vegetated sites or complete snow-covered sites is close to one and can be estimated very accurately.

We have added additional station data from the Baseline Surface Radiation Network (BSRN) located in the northern high latitudes. Among the few stations available at high latitudes, we selected three stations:

ALE (Alert, Lincoln Sea, Canada)

NYA (Ny-Ålesund, Spitsbergen)

TIK (Tiksi, Siberia, Russia)

The remaining arctic stations were not considered as either not enough data points were available or the downward radiation components were missing. We show a comparison of satellite-based and in-situ LST data in Figure 5 (plot visible below). In-situ LST for the BSRN stations is computed from their radiation components as in:

Martin, M.A.; Ghent, D.; Pires, A.C.; Göttsche, F.-M.; Cermak, J.; Remedios, J.J. Comprehensive In Situ Validation of Five Satellite Land Surface Temperature Data Sets over Multiple Stations and Years. *Remote Sens.* **2019**, *11*, 479. https://doi.org/10.3390/rs11050479.

Specifically, broadband emissivity (BBE) is obtained from channel effective emissivity data provided in the ASTER GED with the linear equation described in:

J. Cheng, S. Liang, Y. Yao and X. Zhang, "Estimating the Optimal Broadband Emissivity Spectral Range for Calculating Surface Longwave Net Radiation," in *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 2, pp. 401-405, March 2013, doi: 10.1109/LGRS.2012.2206367.

The terrain corresponding to one GAC pixel around NYA is very heterogeneous (mountainous terrain), leading to poorer validation results than the two other BSRN stations. The BSRN stations are also all located close to the shoreline and snow-covered part of the year.

Due to the fact that only cloud-free data sets are used for the matchup with in-situ stations, the different cloud coverage in the Arctic modifies the number of selected matchups but not the quality of the comparison.

Section *2.4 Auxiliary data* and Table 2 (description of used stations and procedure to derive LST) has been changed accordingly. The Figure 5 has been added to the Section *4.1.2 Validation with in situ LST*

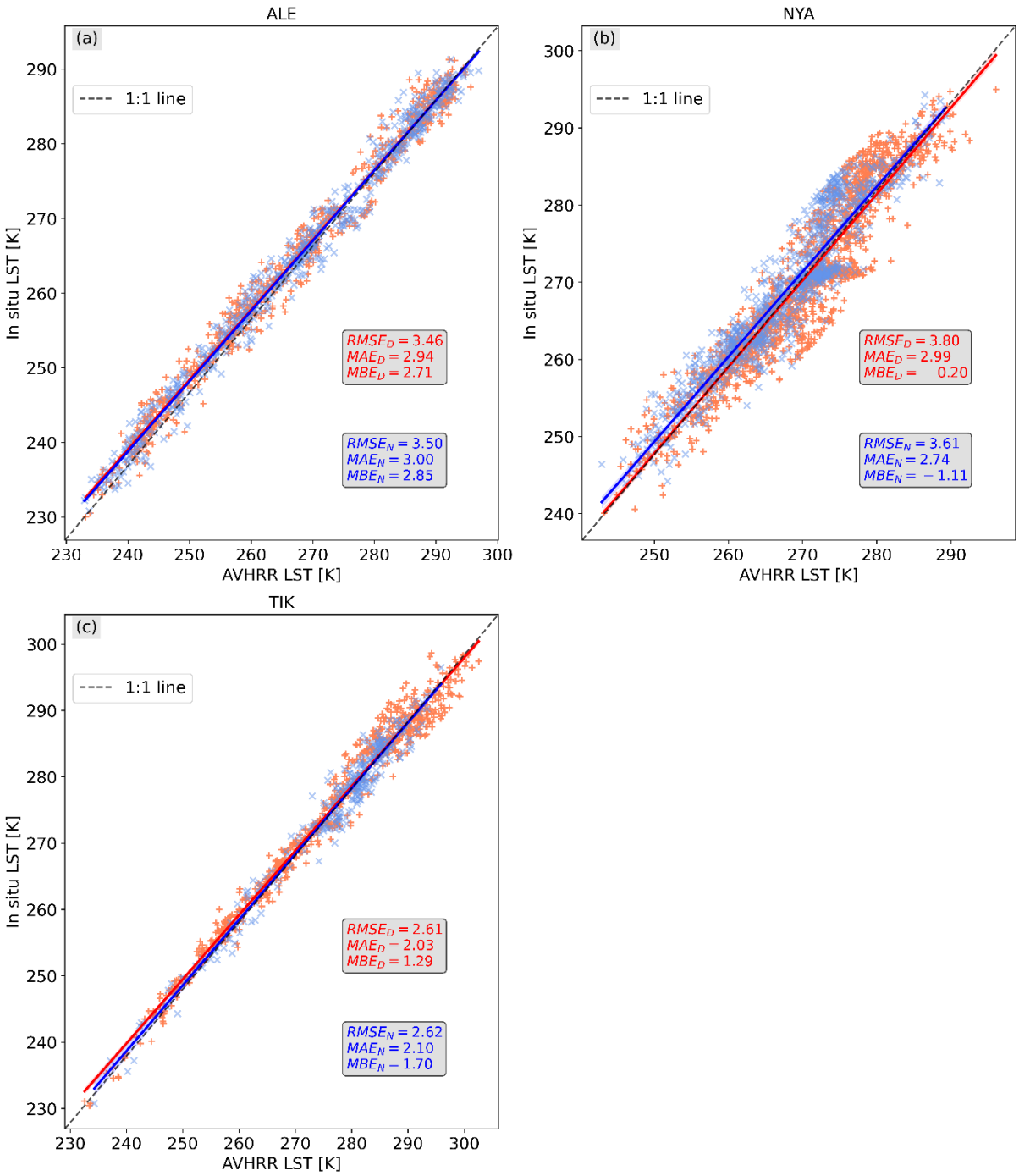

Figure 5 : AVHRR LST versus in situ LST at BSRN stations (a) Alert, Lincoln Sea (ALE), (b) Ny-Ålesund, Svalbard (NYA) and (c) Tiksi, Russia (TIK). Red represents the daytime measurements, and blue represents the nighttime measurements. Match-up periods are provided in the text.