We thank the Referee for constructive criticism and comments which significantly improved our manuscript. In the following, we provide point-by-point replies to all issues raised. Note that the reviewer's comments are in black text, our answers in red and the changes in the manuscript are indicated in red italic.

1. It is not completely clear what is the added value of having this specific dataset derived for the artic region if the authors are simply averaging all observations within a day. There are already datasets available based on AVHRR that provide daily composites (e.g. GLASS, LSA-SAF). In my opinion, it would have been more beneficial to explore the multiple passages of the different AVHRR to try to reconstruct the diurnal cycle. That would have made the dataset more unique and more useful. Having averages of whatever observations exist in a day can create high instabilities in day-to-day variability, depending on what sensors are available and cloud coverage.

We agree with the reviewer and apologize for not making this sufficiently clear in the manuscript. The presented pan-Arctic AVHRR LST dataset does not simply average all provided LST observations within one day, which are typically two for an individual satellite. In order to generate time series from the LST observations of the different NOAA satellites, these are selected based on their overpass time (i.e., not averaged); the details are presented in Table 4. In contrast, for the *EUMETSAT* Polar System (EPS) series (MetOp-1, -2, -3), which has highly stable overpass times, the observations are averaged (see Table 4). Only the stability analysis and trend analysis are performed on monthly mean values, i.e., for each individual satellite a time series of monthly mean composites is created. Subsection *#3.4 Time series generation* now explains and clarifies the above points and the data description in subsection *#2.1 EUMETSAT AVHRR FDR* has been expanded.

The GLASS product (Zhou et al.2019; Ma et al. 2020) is based on the Long-Term Dataset Records (LTDR) (Pedelty et al., 2007) and is built on the SeeBor V5.0 (Borbas et al. 2005) data and TIGR2000 V1.2. Compared to these two profile databases, the novel calibration database from Ermida et al. (2022) is based on the recent ERA5 reanalysis and therefore exhibits high temporal and spatial coverage as well as improved good vertical resolution (137 levels). In addition, the profiles were selected with a dissimilarity criterion, ensuring that less common atmospheric conditions are also included. Furthermore, the GLASS product has considerable data gaps above 45° latitude, which can be attributed to cloud masking (see Fig. 11 from Ma et al., 2020), and to its emissivity computation relying on visible channels, which are unavailable during the polar night. Our pan-Arctic AVHRR LST dataset utilizes a probabilistic cloud mask provided in the CLARA-A3 dataset (Karlsson et al., 2023). In addition, in our workflow, emissivity retrieval for the snow- and ice-covered areas is based on snow water equivalent (SWE) data retrieved from passive microwave radiometer (PMR), which are also available during polar night (Solberg et al. 2021).

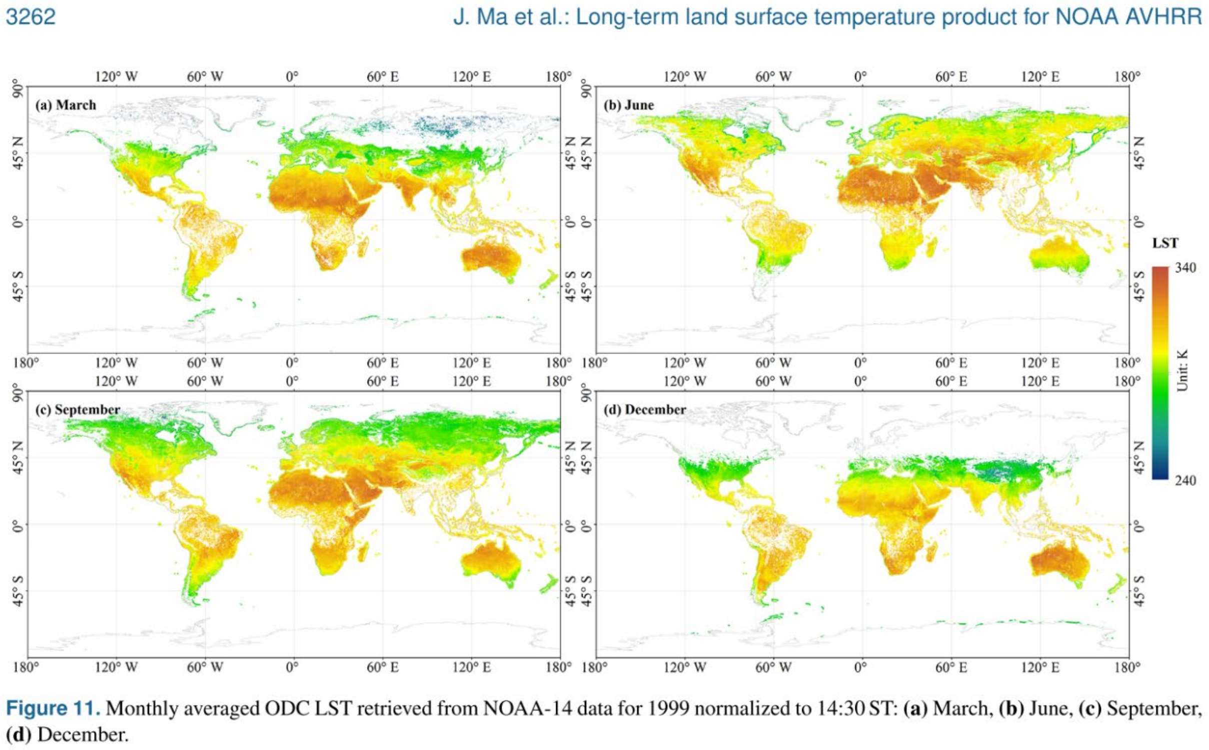
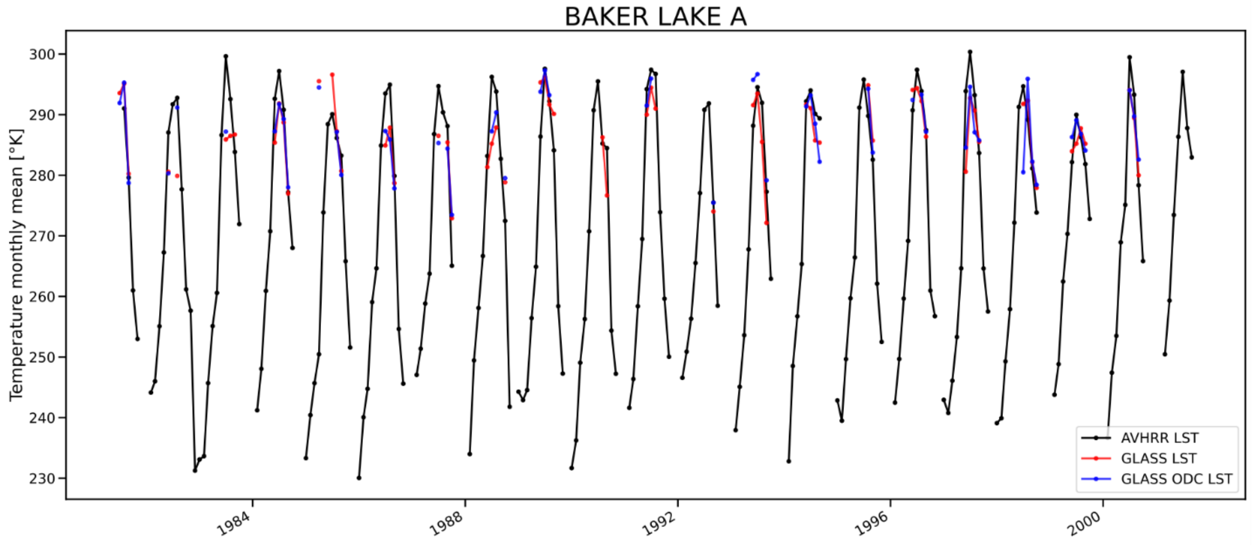
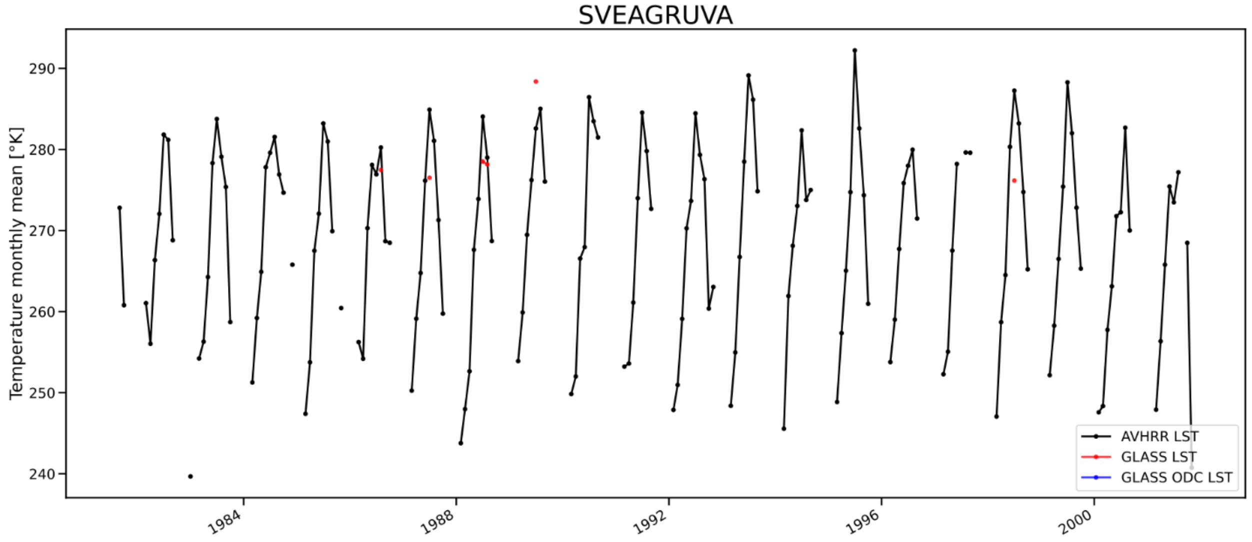

3262 J. Ma et al.: Long-term land surface temperature product for NOAA AVHRR

Figure 11. Monthly averaged ODC LST retrieved from NOAA-14 data for 1999 normalized to 14:30 ST: **(a)** March, **(b)** June, **(c)** September, **(d)** December.

In order to illustrate the better availability of our pan-Arctic AVHRR LST dataset compared to the GLASS product, in section *#4.1 LST validation results* a subsection has been added that showcases the differences between the two products. Both plots below show the differences between our pan-Arctic AVHRR LST dataset (black), the GLASS product (red) and the ODC corrected GLASS product (blue). In the high northern latitudes, the GLASS product is only available during summer months. This is particularly visible for the SVEAGRUVA site (SVALBARD), where very few GLASS observations are available. Also at BAKER LAKE A our product presents considerably more and slightly higher values, which can be explained by the different cloud masking and emissivity computation.





Concerning the LSA SAF LST products mentioned by the Referee: the EDLST dataset from LSA-SAF is based on AVHRR-MetOp, provides uncertainty estimates and has been intensively validated. However, the EDLST time series starts with the first MetOp satellite in 2015 (https://lsa-saf.eumetsat.int/en/data/products/land-surface-temperature-and-emissivity/), while our dataset covers a 40-year period of AVHRR instruments. The other LSA-SAF LST products are based on MSG/SEVIRI data, i.e., they do not cover the high latitudes.

The introduction of the manuscript has been modified to emphasize these differences and the benefits of our Pan-Arctic AVHRR LST dataset compared to already existing datasets.

2.1 EUMETSAT AVHRR FDR

The FDR contains AVHRR reflectance and brightness temperatures for each available orbit and channel. The daily AVHRR data from one satellite provides nearly complete coverage of the globe. The dataset provides for each satellite twice-daily composites (one daytime overpass and one nighttime overpass). AVHRR GAC measurements have been processed using the PyGAC software –a Python software package to read and transform AVHRR data in GAC format- (https://pygac.readthedocs.io/en/latest/#), including the conversion from counts to reflectance or brightness temperature and cross-calibration of the visible channels of the AVHRR sensor.

3.4 LST AVHRR time series generation

Depending on the heterogeneity of the land cover, between four and nine AVHRR LST GAC pixels are extracted around each station. Pixels that have a cloud probability higher than 0.1 are removed, and the average of the remaining pixels is computed. Daytime data from NOAA-7, 9, 11, 14, 16, 18 and 19 (satellites with ascending (northbound) equator crossing times), as well as the entire MetOp series (satellites with descending (southbound) equator crossing times), are considered for constructing the time series. The considered period for each satellite is chosen to minimise orbital drift and avoid the outage periods (EUMETSAT, 2023d). The retained periods are listed in Table 4.

Once the relevant periods are extracted, outlier detection is performed based on a 10-day rolling window analysis and detected outliers are removed. Daily temperature variability is very high (Mildrexler et al., 2011), and AVHRR-derived LST time series are subject to noise, therefore, monthly means are computed from the concatenated day time series for further analysis.

#4.1.3 Comparison with the GLASS dataset

The pan-Arctic AVHRR LST dataset is compared against the well-established GLASS product (Zhou et al. 2019, Ma et al. 2020), that provides twice daily LST observation for the whole globe for the 1980-2000 period. Figures 5 and 6 present a comparison of monthly means at two stations located in the Arctic (BAKER LAKE A and SVEAGRUVA). The classical GLASS LST, the orbital drift corrected (ODC) GLASS LST and the pan-Arctic AVHRR LST are compared. In the high northern latitudes, the GLASS product is only available during summer months. This is particularly visible for the SVEAGRUVA site (SVALBARD), where very few GLASS observations are available. Also at BAKER LAKE A our product presents considerably more and slightly higher values, which can be explained by the different cloud masking and emissivity computation.

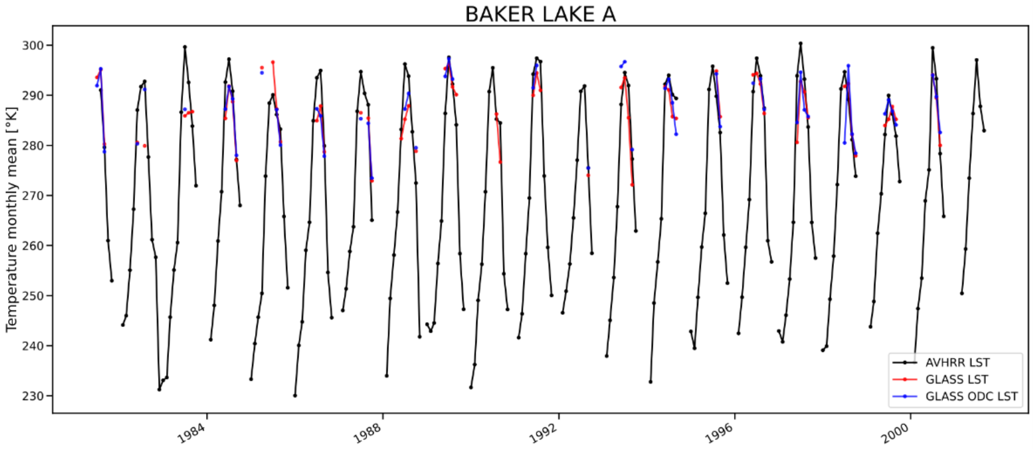


Figure 5. Monthly means LST product comparisons at BAKER LAKE A.

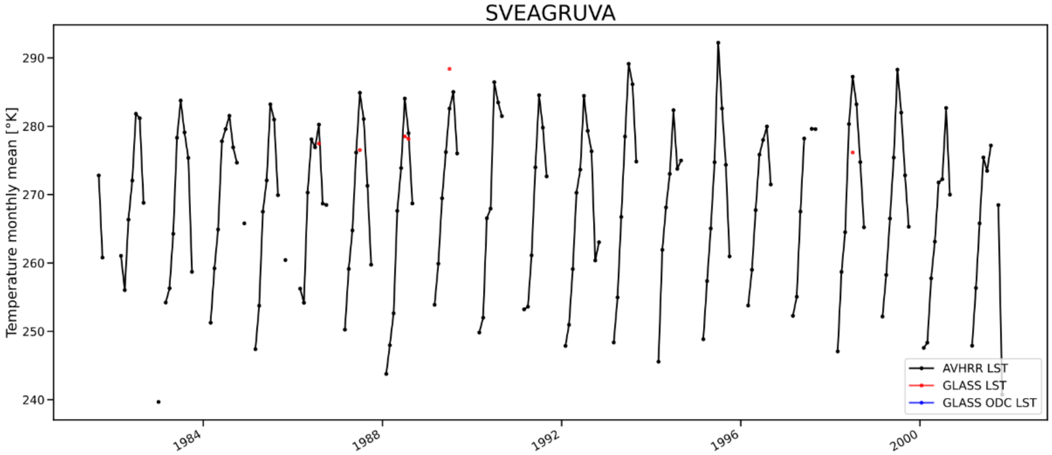


Figure 6. Monthly means LST product comparisons at SVEGRUVA.

2. For the same reason, I'm not convinced the dataset is appropriate for trend, and specially not for anomaly analysis. If the time of observation that goes into the average keeps changing, then there is just too much instability in the series.

Again, we agree and apologize for not describing our approach to AVHRR LST time series generation clearly enough; please refer to our in-depth answer to point #1.

Subsection *#3.4 Time series generation* has been modified and expanded to clarify the differences in generating LST time series for the NOAA and MetOp satellites. For each individual satellite the selected time period (Table 4) has been chosen to minimize the effect of orbit drift. Furthermore, winter data (December and January) are analyzed separately from the summer data to investigate the influence of the orbital drift on the trend analysis.

3. Also, in terms of algorithm calibration, here there was a unique opportunity to explore an algorithm more suited for the specific conditions of the Artic. That maybe would allow using a higher range of view angles, resulting in an even larger sampling of observations through the day. The same in terms of the calibration database, why not tailor the database to the more specific conditions of the Artic? Using a generic algorithm and database that are valid over the whole globe is something that is already available in other products.

The Generalized Split Window (GSW) algorithm we have employed (Wan & Dozier, 1996) is well-established and used for operational LST products (e.g., LST products from LSA-SAF). This algorithm is optimal for sensors with two TIR channels centered at 11 and 12 µm, which is the case of AVHRR. The GSW algorithm was compared against other retrieval algorithms: for LST retrievals from Sentinel-3/SLSTR by Yang et al. (2020), where it presented the highest accuracy overall, in line with similar studies performed for other sensors. The GSW can be tailored and adapted for every region with the appropriate split-window coefficients.

Our area of interest starts at 50° latitude and encompasses the whole pan-Arctic region. The climate zones in this area differ strongly from each other, e.g., the Siberian tundra from the high mountains in Alaska or the great plains in southern Canada. The clear-sky database created by Ermida et al. (2022) is built on ERA5 data resampled with a dissimilarity criterion and includes satellite observation to determine realistic surface conditions, as opposed to the SeeBor database (Borbas et al. 2005), built from ERA-40 data. Currently most LST products (including the GLASS product) rely on the SeeBor database. The recent ERA5 exhibits significant improvement in the lower layers of the atmosphere, which improves the simulation of satellite observations performed in wavelengths more sensitive to the surface. The GSW is trained independently for each class of total column water vapor and surface temperatures. Only profiles suitable for our area of interest have been chosen in the training and testing phase.

The above points have been clarified and a more detailed description of the calibration database as well as the criteria for selecting atmospheric profiles for an LST retrieval algorithm optimized for the pan-Artic region has been included in the manuscript.

It is true that satellites have a higher coverage nearer the poles. This allows to choose scenes with viewing angles closer to nadir, which have the advantage of providing smaller footprints and higher quality data, e.g., in terms of cloud contamination and surface anisotropy. This is independent of the chosen LST algorithm. The split-window coefficients (SWC) were computed for angles up to 70°, but in the final product, all pixels with a satellite viewing angle higher than 40° were masked out to keep only the best quality data.

2.4 Auxiliary data

Atmospheric profiles from the Clear-Sky Database developed at LSA-SAF (Ermida and Trigo, 2022) are used for the RT modelling (RTM). This database contains atmospheric profiles such as temperature, specific humidity and ozone on 137 model levels (full vertical resolution), sampled from ERA5 for the 2009-2019 period. The sampling technique follows the method from Chevallier et al. (2000). Surface variables like T2M, surface pressure, Tskin and emissivity are obtained from the combination of ERA5 and satellite data to ensure the best possible representation of the surface conditions. Column variables, such as TCWV and total cloud cover (TCC) are also present in the database. The atmospheric profiles are classified on TCWV varying from 0 to 60 mm and TS ranging from 190 to 340 K. The profiles belonging to our area of interest are selected.

3.1 Generalised Split Window algorithm

Based on the test sets, look-up tables (LUT) with coefficients are created for each satellite. The LUTs are organized into classes of TCWV and Tskin, allowing to allocate the right SWC to the encountered atmospheric conditions. Mean absolute error (MAE), the coefficient of determination (R2) and root mean square error (RMSE) are computed for all coefficients to keep track of the general performance of the RTM

4. With respect to the LST validation, the authors only used KIT and SURFRAD stations. None of the stations is within the study area and therefore are not representative of the presented LST dataset. This is very clear when looking at figures 4 and 10. These stations' LSTs lowest values are around 260K, while most of the Artic is well bellow this value. There is a very with range of surface flux stations within the considered area (AmeriFlux, Fluxnet, BSRN) or even in Antarctica, which has much more similar conditions. The authors should have tried to use more stations that encompass the specifics of the polar climate. It's true that these stations tend to be more heterogeneous, but the SURFRAD stations are also very heterogeneous.

We agree that for a broader validation that is more representative of the low temperatures, high-latitude sites would be highly desirable. However, high quality in situ data from dedicated LST validation sites are rare and most of the existing stations (SURFRAD, BSRN, ...), as mentioned by the reviewer, have spatial representativeness issues.

In our study we decided to only use top tier in situ LST validation data. Therefore, we only consider stations that have been investigated within the ESA GlobTemperature and the LST CCI projects in terms of their suitability for validating satellite LST and undergone quality controls. Following recommended validation protocols, in situ measurements need to have a high temporal frequency (sampling rate ranging from 1 to 3 min, according to Guillevic (2018)) to avoid additional uncertainty due to temporal mismatch / interpolation. BSRN and FluxNet only provide data averaged over a 30 min or one hour period. Furthermore, accurate emissivity information needs to be available to convert measurements of brightness temperature into in situ LST observations.

Data from the Atmospheric Radiation Measurement Climate Research Facility *US* Department of Energy (ARM) site at the North Slope of Alaska (NSA) are available from 2007 to 2012, have undergone quality control procedures and previously been used in the ESA GlobTemperature project, i.e., they meet the above stated criteria. Therefore, we integrated the in situ LST data from the NSA site into our validation and updated the corresponding figure (Fig. 4) and table (Table 2) accordingly. The surface is very heterogeneous at the NSA site, the station being close to lagoons (North Salt Lagoon and Imikpuk Lake), and very close to the coast. This explains why the performances are much worse during summertime than during winter when the entire area is snow and ice covered.

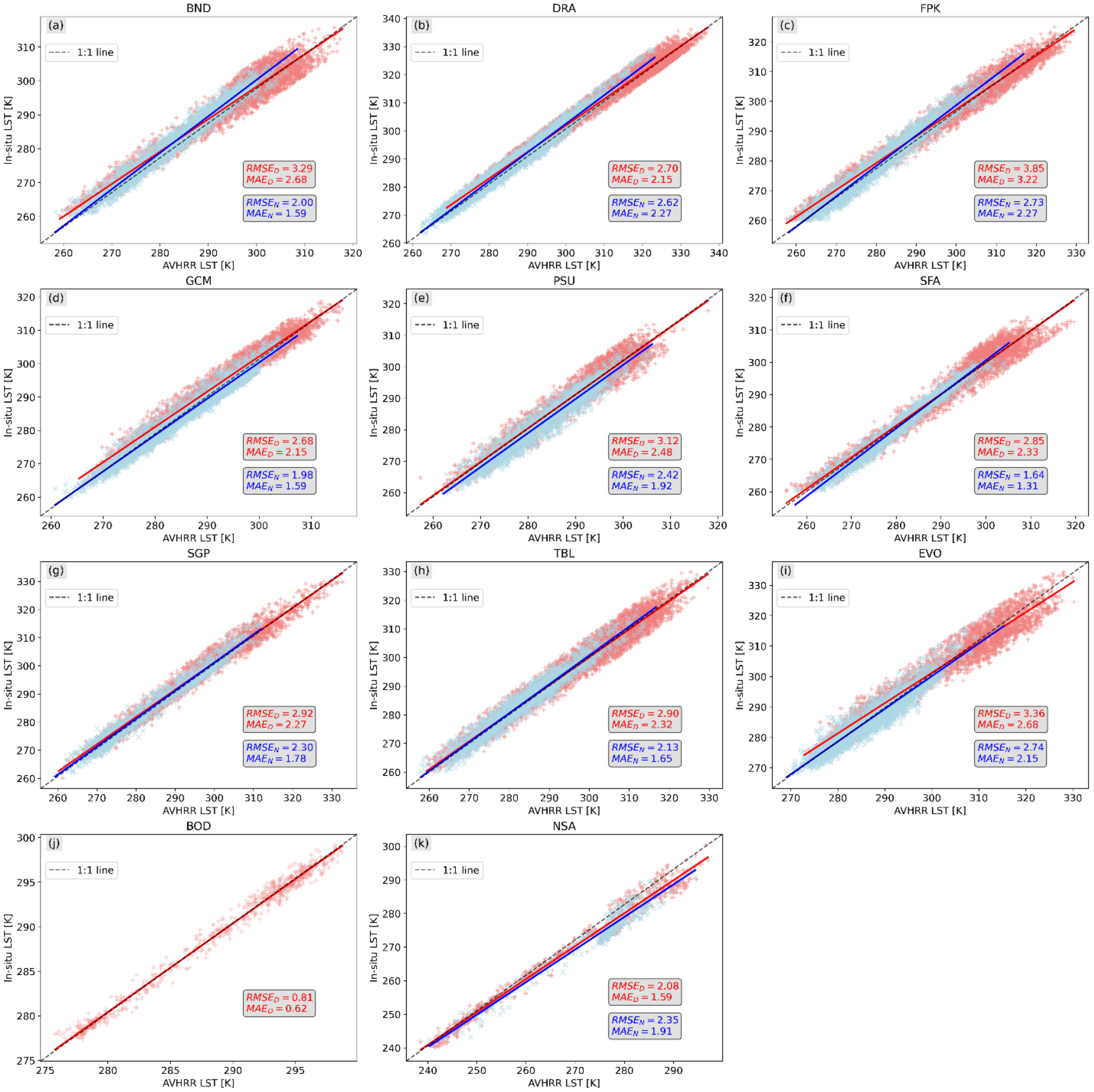


Figure 4. AVHRR LST versus in situ LST at (a) Bondville (BND), (b) Desert Rock (DRA), (c) Fort Peck (FPK), (d) Goodwin Creek (GCM),(e) Penn. State Univ (PSU), (f) Sioux Falls (SFA), (g) Southern Great Plains (SGP), (h) Table Mountain (TBL), (i) Evora (EVO), (j) Lake Constance (BOD), (k) North Slope of Alaska (NSA). Red represents daytime measurements and blue represents nighttime measurements. Match-up periods are provided in the text.

5. There is a long discussion on whether the problems in stability seen when comparing Tair with T2M and LST being related to day/night problems. It’s not clear to me why the authors did not separate daytime from nighttime observations. This would make

comparing with Tair_max and Tair_min more easy to interpret. For T2M, it's not clear from the text but it seems the authors are averaging all hours of the day? The ERA5-land provides a seamless diurnal cycle with hourly frequency, why not compute the daily max and min to obtain variables comparable to Tair?

Thank you for pointing this out. We agree with the Referee: our description of the use of LST daytime data only was not clear. We now describe the time series generation more clearly. The goal was to prove the overall stability of our product based on 17 different satellites, using the ERA5-Land product. In that respect, we based our analysis on monthly mean T2M data from ERA5-Land. The highlighted topic by the Referee will be considered in our next analysis.

6. Why do you use ERA5 in some cases and MERRA-2 in other? ERA5 has better spatial and temporal resolution.

Two datasets were chosen for this study: MERRA-2 and ERA5-Land.

MERRA-2 assimilates space-based observations of aerosols and accounts for ice sheets, ensuring accurate data for regions like Greenland and Antarctica. The GLASS product (Ma, 2020) has been generated with MERRA-2 products: using MERRA-2 for our study simplifies comparisons between the GLASS and our product. Our study uses skin temperature and total column water vapor from MERRA-2 to determine the atmospheric conditions at each pixel and to select the SW coefficients from the look up table.

In contrast, air temperature data from ERA5-Land are used for the correlation and stability analysis. Using separate sources of reanalysis data helps keeping the stability analysis independent from the SW coefficient assignment process. Furthermore, ERA5-Land has been fully validated in the Arctic region in a previous study on trend analysis (Rantanen, 2023, ARCLIM atlas). Finally, in the stability analysis we compared point data (weather station data), satellite data and a reanalysis product: in order to optimize the spatial representativeness in the point-to-pixel comparisons, the reanalysis product with the highest spatial resolution has been chosen.

The above choices are now explained in a corresponding paragraph added to subsection *#2.4 Auxiliary data*.

2.4 Auxiliary data

Skin temperature (Tskin) and Total Column Water Vapor (TCWV) from the MERRA-2 reanalysis dataset (M2T1NXSLV, variables are labelled TS and TQV). The data come at hourly temporal resolution with a spatial resolution of 0.5° x 0.625°. Nearest neighbour resampling was performed to match the AVHRR spatial resolution and scanline time, i.e. as in the work *of Ma et al. (2020). MERRA-2 is preferred over other reanalysis products with finer spatial resolution to allow comparison with the GLASS product (Ma et al. 2020) and to keep the LST retrieval independent from ERA5-Land, which will be used for the stability analysis.*

7. Is/will this dataset be made available publicly? What is the format? What is the projection? More technical details about the dataset are needed.

This product is part of a PhD project, and part of ongoing research, but we consider releasing it after completion of the project.

The PyGAC AVHRR FDR from EUMETSAT (2023) is available in the Network Common Data Form (NetCDF) format and so is our Pan-Arctic LST product. We kept the same data structure as the original FDR. The dataset covers the pan-Arctic region (− 180°, 90°, 180°, 50°) at a spatial resolution of 0.05 × 0.05° pixel size. Technical details regarding format and projection are added to the manuscript as well as details on the format of the EUMETSAT dataset.

2.1 EUMETSAT AVHRR FDR

The IR calibration procedure is satellite-specific, with no cross-calibration between satellites for IR channels (EUMETSAT, 2023d). The PyGAC AVHRR FDR from EUMETSAT (2023) is available in the Network Common Data Form (NetCDF) format and covers the entire globe (− 180°, 90°, 180°, -90°) at a spatial resolution of 0.05 × 0.05° pixel size. This study focuses on the pan-Arctic region, therefore only data above 50° N have been processed.

References

Borbas, E.E.; Seemann, S.W.; Huang, H.L.; Li, J.; Menzel, W.P. Global profile training database for satellite regression retrievals with estimates of skin temperature and emissivity. In Proceedings of the International TOVS Study Conference-XIV, Beijing, China, 25–31 May 2005.

EUMETSAT (2023): AVHRR Fundamental Data Record - Release 1 - Multimission, European Organisation for the Exploitation of Meteorological Satellites, DOI: 10.15770/EUM_SEC_CLM_0060. http://doi.org/10.15770/EUM_SEC_CLM_0060

Guillevic, P., Göttsche, F., Nickeson, J., Hulley, G., Ghent, D., Yu, Y., Trigo, I., Hook, S., Sobrino, J.A., Remedios, J., Román, M. & Camacho, F. (2018). Land Surface Temperature Product Validation Best Practice Protocol. Version 1.1. In P. Guillevic, F. Göttsche, J. Nickeson & M. Román (Eds.), Good Practices for Satellite-Derived Land Product Validation (p. 58): Land Product Validation Subgroup (WGCV/CEOS), doi:10.5067/doc/ceoswgcv/lpv/lst.001

Karlsson, K.-G., Stengel, M., Meirink, J. F., Riihelä, A., Trentmann, J., Akkermans, T., Stein, D., Devasthale, A., Eliasson, S., Johansson, E., Håkansson, N., Solodovnik, I., Benas, N., Clerbaux, N., Selbach, N., Schröder, M., and Hollmann, R.: CLARA-A3: The third edition of the AVHRR-based CM SAF climate data record on clouds, radiation and surface albedo covering the period 1979 to 2023, Earth Syst. Sci. Data, 15, 4901–4926, https://doi.org/10.5194/essd-15-4901-2023, 2023.

Ma, J., Zhou, J., Göttsche, F.-M., Liang, S., Wang, S., and Li, M.: A global long-term (1981–2000) land surface temperature product for NOAA AVHRR, Earth Syst. Sci. Data, 12, 3247–3268, https://doi.org/10.5194/essd-12-3247-2020, 2020

Pedelty J., Devadiga S., Masuoka E., Brown M., Pinzon J., Tucker C., *et al.*
Generating a long-term land data record from the AVHRR and MODIS Instruments. Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007, Barcelona (2007), pp. 1021-1025

Solberg, R., G. Schwaizer, T. Nagler, S. Wunderle, K. Naegeli, K. Luojus, M. Takala, J. Pulliainen, J. Lemmetyinen, and M.
Moisander (2021) ESA CCI+ Snow ECV: Product User Guide, version 3.1, December 2021.

Wan, Z. and Dozier, J.: A generalized split-window algorithm for retrieving land-surface temperature from space, IEEE T. Geosci. Remote, 34, 892–905, https://doi.org/10.1109/36.508406, 1996.

Yang, J., Zhou, J., Göttsche, F. M., Long, Z., Ma, J., and Luo, R.: Investigation and validation of algorithms for estimating land surface temperature from Sentinel-3 SLSTR data, Int. J. Appl. Earth Obs., 91, 102 136, https://doi.org/10.1016/j.jag.2020.102136, 2020.

Zhou, J., S. Liang, J. Cheng, Y. Wang, and J. Ma, 2019: The GLASS land surface temperature product. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **12**, 493–507, https://doi.org/10.1109/JSTARS.2018.2870130.