RC1:

The manuscript has improved, but I still have some concerns.

In response to my comment #3, the authors suggest that FLUXNET measurements do have errors. Therefore, I would suggest adding information about the observational instrumentation, such as the sensors and their accuracy. I also suggested adding a description of the sources of error in the field data.

We have added a section describing the measurement technique and to-be-expected uncertainties, see Line 164-171:

While the in situ measurements are considered as ground-truth, it is necessary to mention that they have their own sources of uncertainties. Incoming shortwave and longwave radiation are measured by pyranometers and pyrgeometers. Accuracy targets for the BSRN network measurements (from 2004) are for example 2% or 5 W m^{-2} for incoming shortwave radiation and 2% or 3 W m^{-2} for incoming longwave radiation. Target uncertainties for outgoing shortwave and longwave radiation are 3% and 2% (or 3 W m^{-2}) respectively (McArthur, 2004). For the measurement of the outgoing radiation components the pyranometer/pyrgeometer is installed facing downwards. The target uncertainties are in line with the achievable accuracy of the pyranometer/pyrgeometer instruments although they might not be met under some conditions, e.g. incorrect installation at an angle or snow cover. The instruments should be calibrated every 2 years (Walter-Shea et al., 2019).

In addition, the authors stated that ‘using as many stations as possible benefits the validation, also in areas where the spatial heterogeneity is large’. However, the use of in situ data from regions with large spatial heterogeneity may lead to inaccurate or erroneous validation results.

We thank the reviewer for this comment and understand the concern. The high variability expected in heterogeneous landscapes in the end determines the need for more *in situ* measurements within the satellite footprint to be able to reduce representativeness errors and make the validation of 1 km data more meaningful. Ideally, we would have many validation sites in every heterogeneous pixel to capture the entire footprint. As unfortunately this is currently not possible, we argue that using as many sites as possible, at least to some extent, alleviates the issue as in theory the representativeness errors will average out the more *in situ* data across the domain is used. While this approach is not perfect, restricting the validation to homogeneous pixels would further reduce the limited amount of *in situ* data and make the validation less meaningful.

We rephrased L434–435 as:

Unfortunately, we cannot solve this issue but argue that using as many stations as possible benefits the validation, particularly within pixels where the spatial heterogeneity is very large.

RC2:

The manuscript has been further improved but is still not satisfactory. There are several issues in the revised manuscript. Authors can consider to further improve this manuscript if they can clearly proof the quality and novelty of their datasets. Issues I can find as follows.

1. “1 km” in the title is not proper since the first word is usually not number. Move “1 km” to the middle of the title or use “One-kilometer” may be better.

We thank the reviewer for this comment and agree. We have changed the title to *“High-resolution (1-km) all-sky net radiation over Europe enabled by the merging of land surface temperature retrievals from geostationary and polar-orbiting satellites.”*

2. If I remember correctly, the links on the published datasets should also be shown in the abstract.

Thank you, we have added the links to the abstract (L17-18).

3. Lines 349-369. As the LST data is not the main output of this study, it is not necessary to compare it with other existing LST datasets. In my previous comment, “I suggest authors highlight the novelty of their dataset by comparing with existing datasets”. It is meaningful to compare SNR datasets instead of LST datasets. Besides, the current comparisons of LST datasets are unsatisfactory as the advantages and drawbacks of these LST datasets are still not clear to us.

We thank the reviewer for this comment and agree. Important other studies were listed but an evaluation in terms of potential advantages or drawbacks was not included. We have added this now for the LST dataset and have also expanded the comparison for SNR datasets (L359–424):

Examples of other gap-free LST datsets which have recently been developed are given by Shiff et al. (2021), Xu and Cheng (2021), Jia et al. (2022) or Wu et al. (2023). The approach taken by Shiff et al. (2021) was to merge clear-sky 1 km MODIS LST with 0.2 degree modelled air temperature provided by the National Center for Environmental Prediction (NCEP) from the Coupled Forecast System Model version 2 (CFSv2) system. This was done by extracting the underlying seasonal behaviour from both input datasets by Temporal Fourier Analysis and subsequently adding the CSFv2 anomalies to the MODIS climatology on days where no clear-sky MODIS LST observation is available. Xu and Cheng (2021) demonstrated a multi-step approach based on infrared Advanced Microwave Scanning Radiometer 2 (AMSR2) brightness temperatures, MODIS LST as well as MODIS based ancillary datasets and elevation data. First, land surface temperature is retrieved from all the above datasets at 0.1 degree spatial resolution. This LST dataset is then downscaled from 0.1 degree resolution to 0.01 degree resolution by using the elevation data and MODIS NDVI. Clear-sky MODIS LST data and the retrieved all-sky LST data are then bias corrected allowing for the temporal gap-filling of the clear-sky LST retrievals. Finally, the 0.1 degree LST retrievals based on AMSR2 are assimilated into the merged 0.01 degree LST dataset by applying a multiresolution Kalman filtering approach. Jia et al. (2022) have produced all-sky diurnal, hourly LST estimates at 2 km spatial resolution based on the surface energy balance. The three step approach is based on constructing a spatiotemporal dynamic model of LST from ERA-5 in which clear-sky LST from the Advanced Baseline Imager (ABI) are assimilated. As a final step the gap-free LST record is updated by superimposing diurnal cloud effects using satellite radiation products. Wu et al. (2023) have tested an approach to produce very high-resolution, 100m gap-free LST, from a single Landsat-8 acquisition by training a Random Forest algorithm with the Landsat derived LST and ancillary variables, e.g. land cover, population density and elevation. The LST merging methodology presented in this paper shares some of the elements of the above mentioned studies, i.e. primarily the bias-correction of the coarse-scale LSAF LST observations towards Sentinel 3 (see section 3.3), as well as a Kalman Filtering approach. An in-depth validation and quantitative inter-comparison of the above mentioned products was not the aim of this study presented here. We argue however, that on a theoretical basis, the here proposed methodology has some advantages. Most of the above mentioned approaches rely on input data with a coarser spatial resolution. Shiff et al. (2021) and Jia et al. (2022) for instance use air temperature data at a 0.2 degree resolution or ERA-5 with approximately 31 km spatial resolution. Both these datasets are also model output, albeit from data assimilation systems taking a multitude of observations into account. The coarsest spatial resolution of the input datasets used in the here presented methodology are the LSAF geostationary retrievals with a pixel size of ca. 5-7 km, depending on latitude. While especially the LSAF all-sky retrievals also are based on modelling, and require ancillary information, they are optimised for the retrieval of the single target variable at high accuracy. They are also available hourly, whereas e.g. ERA-5 is provided 3-hourly or Landsat-8, used by Wu et al. (2023), is only available every few days, depending on cloud cover. Furthermore, our approach does not rely on using ancillary variables which are not directly linked to the physical processes to statistically downscale the input products, as is for example done in Wu et al. (2023) by using population density. One of the drawbacks of the here presented methodology is the lack of a dynamic temporal model which is able to propagate assimilation updates, provided by the 1 km LST retrievals from Sentinel-3, over time, which has been achieved by Jia et al. (2022). Here we thus apply the same update from when a Sentinel-3 observation is available to the subsequent time steps until the next observation is available. An additional drawback is that e.g. ERA-5 and NCEP are globally available datasets and the use of MODIS LST retrievals allows for the production of long time series. In contrast, LSAF is limited to North Africa and Europe and Sentinel-3 was only launched in 2016. The approach can however be transferred to other regions by substituting LSAF with other geostationary retrievals and using MODIS instead or in addition to Sentinel-3 to allow for an extension of the time series.

In terms of the calculation of the daily all-sky surface net radiation dataset, we argue that the approach taken is the most straightforward as it is based on the underlying physical principles of the individual radiation components. This is in contrast to studies presenting methods to produce net radiation at a similar temporal and spatial resolution which exploit statistical relationships between some well observed components, e.g. incoming radiation components from satellite, and ancillary information, e.g. land cover or NDVI, or modelled variables. Xu et al. (2022) for example train a convolutional neural network using net radiation from a selection of in-situ measurements, MERRA-2 reanalysis and AVHRR top of atmosphere (TOA) data. Jiang et al. (2023) presented two algorithms based on a Random Forest to downscale the GLASS net radiation product, either by exploiting the relationship between net radiation and shortwave radiation as well as ancillary information, including from from ground measurements, or by linking net radiation to TOA observations from the Landsat satellites and ancillary information.

The GLASS algorithm itself is based on the Multivariate Adaptive Regression Splines (MARS) model, trained with remotely sensed incoming radiation, NDVI and albedo as well as mostly MERRA-2 meteorological variables (Jiang et al., 2016). While such downscaling methodologies can work very well, and we need to note that no quantitative comparison is here performed, they rely on training a model which establishes a statistical relationship between the different input variables. These data driven approaches are very sensitive to the training data and e.g. the spatial or temporal domain for which such a model is established. Hence, a globally trained model might not capture locally specific conditions or provide accurate output for time periods not considered for the training. With in-situ training data often the limiting factor, established statistical relationships might also be only valid for these specific sites and avoiding model over-fitting can be very challenging. It can thus be beneficial if in-situ measurements are solely used for the validation of a methodology rather than the development itself. Another methodology to produce hourly *surface solar radiation at 5 km spatial resolution was developed by Tang et al., 2016. In the two step approach hourly cloud parameters are estimated with a neural network by combining cloud products from MODIS with high temporal-resolution top-of-the-atmosphere (TOA) radiance data from the geostationary Multifunctional Transport Satellite (MTSAT). Subsequently the cloud information and other auxiliary information is combined in a radiative transfer model to retrieve the surface net radiation. Conceptually, although estimating surface radiation primarily based on cloud properties, is is similar to the here presented approach in exploiting the advantages of geostationary and polar-orbiting satellite measurements and being more physically based. An overview of some further approaches to produce surface net radiation products are also given by Tang et al., 2016.*

We have also compared the SNR dataset to the state-of-the-art ERA5-Land reanalysis product, see Figure 11 and 12 and Lines 330-350:

Figure 11 shows the SNR validation for the different CCI land cover types for a LSAF only based SNR as well as the downscaled product. The Figure also includes performance metrics for the ERA5-Land product (Muñoz-Sabater et al., 2021) which were included to give some context. R is generally high for all products (ca. 0.95) for all sites with the exception of sites with land cover affected by water. There ERA5-Land outperforms the LSAF and downscaled SNR product in terms of R, likely due to a sub-optimal treatment of these areas in the processing of the input products. In terms of MSE ERA5-Land again outperforms the other products for water affected land cover. However, for the other land cover classes the LSAF SNR and downscaled products perform better with the downscaled dataset showing the lowest values. In terms of bias, ERA5-Land performs best with the downscaled data performing between ERA5 and the LSAF only SNR. Figure 11. Validation of SNR for different CCI land cover types in terms of R, RMSE, RMSPE and bias.
For the SNR products we also carry out a seasonal analysis. The results of this are shown in Figure D1 and Figure D2 in boxplot form (see annex). Table E1 and Table E2 list all performance metrics for the entire study period as well as seasonally. For the entire 2018–2019 period, R is very similar for both datasets with R=0.93 for the downscaled product and R=0.92 for ERA5-Land. In comparison to ERA5-Land, the downscaled product has a RMSE of 22.53 vs 25.7 W 2. The average bias is lower for ERA5-Land, with -1.56 vs -6.83 W 2.
The downscaled product shows a better performance for the summer period AMJ and JAS (R=0.91 and 0.93 vs 0.83 and 0.86) and the same is true in terms of RMSE (27.58 and 22.18 W 2 vs 34.79, 29.37 W 2). The seasonal bias is lower for the downscaled product. Figure 12 shows as an example the SNR for the downscaled product and ERA5-Land for the 30th of June over an area of western Europe. The increase in spatial resolution and therefore landscape details is clearly visible. The downscaled dataset both shows higher and lower values than ERA5-Land as it is able to resolve finer land surface features due to the high-resolution merged LST and Albedo inputs.

4. Lines 371-390. The comparisons of methods is not satisfactory as the advantages of the proposed method is not persuasive enough.

We have highlighted some points which we think are an important advantage of the proposed methodology. See reviewer point 3 above where we describe the amendments to the discussion section, both for the LST and SNR dataset. There we also refer to the quantitative comparison of the SNR product to ERA5-Land.

5. Figure 11. Cannot find RMSE and RMSPE in this figure, but they occur in the title. The legend of the third subfigure may be not correct. Similar issues may exist in other figures.

Thank you, this has been corrected and the manuscript has been proof-read.