Response to referee #2

Title: A global long-term (1981–2019) daily land surface radiation budget product from AVHRR satellite data using a residual convolutional neural network

MS_No: ESSD-2021-250

Thanks very much for taking your time to review this manuscript. We really appreciate all your valuable comments and constructive suggestions! The specific responses to your all comments are listed below one by one.

Major comments:

Comment 1:

One of the real advantages I see with this dataset is the long record—since the dataset starts in 1981 and has an accuracy equal to or exceeding other satellite-based estimates, this extends observation-based estimates of the surface radiation budget significantly. That could be of significant value for long-term climate studies. ***The authors could highlight this advantage more strongly in the abstract and conclusions.***

Response 1:

Thanks for your kind suggestion. We have highlighted the advantage of the long-term record of AVHRR R_n dataset, and the related content has been included in the revised manuscript.

spatial pattern and temporal evolution trends of R_n observations. The long-term record (1981-2019) of the AVHRR R_n product also shows its value in climate change studies. This dataset is freely available at https://doi.org/10.5281/zenodo.5546316 for

global climate change. Besides, compared to current satellite-derived R_n products, e.g., CERES-SYN and GLASS (2000-present), a more long record (1981-2019) of the AVHRR R_n dataset shows its value in climate change studies. However,

Comment 2:

Evaluation and training are done against multiple networks, but some of these networks are interconnected, for example, some ARM and all SURFRAD sites are included in the BSRN. As I look at the list of sites in Table S1, it appears that some of these stations are included multiple times. For example, BSRN_DRA is the same station as SF_DRA because the SURFRAD Desert Rock stations is submitted to the BSRN global network. ***This is particularly a problem if any of the independent validation stations are also included in the training dataset. Please look into this duplication.***

Response 2:

Thanks for your careful examination. After thoroughly reviewing the list of training (460) and validation (77) sites based on sites' geographic coordinates (i.e, latitude, longitude, elevation), we found several duplicate sites in the training sites group, including one ARM (ARM_E13) site and six SURFRAD (SF_TBL, SF_DRA, SF_PSU, SF_SXF, SF_FPK, SF_GCM) sites have corresponding duplicated sites in the BSRN network in the training group. Besides, several sites from the AsiaFlux, including FxMt_GCK, FxMt_MSE, FxMt_QHB, FxMt_TMK, FxMt_TSE, FxMt_QHB, may be identical to the corresponding sites of the Global FluxNet. However, the same site from different observation networks has different time periods of record, e.g., BSRN_DRA (1998-2017) and SF_DRA (1999-2019). After these duplicated sites were removed from the training dataset, the training statistics were almost the same probably because the duplicated samples are relatively small compared to the total sample population. Therefore, we deleted theses duplicated sites from Table S1 and Table 2, and the corresponding Figure 5(a) unchanged. The corresponding content was also revised.

In addition, we found out that there are three same sites in both the training and independent validation groups, as shown in Table 1, although they are nominally administrated by different observation networks. The three training sites of Lath_CN-Ha2, Lath_KR-Hnm, and Lath_ID-Pag are of the Global FluxNET. For corresponding three validation sites, the CF_HB site belongs to the ChinaFlux network; the FxMt_HFK and the FxMt_PDF sites are of the AsiaFlux network. The ChinaFlux and AsiaFlux networks are sub-network of the FluxNET project. Therefore, we believe that the respective three sites in the training group and the validation group are the same.

Table 1 Summary of duplicate site in both training and validation sites groups.

Training site name	Latitude (°)	Longitude (°)	Elevation (m)	Validation site name	Latitude (°)	Longitude (°)	Elevation (m)
Lath_CN-Ha2	37.6086	101.3269	3203	CF_HB	37.6099	101.3224	3205
Lath_KR-Hnm	34.55	126.57	7	FxMt_HFK	34.55	126.57	13.74
Lath_ID-Pag	2.345	114.036	30	FxMt_PDF	2.345	114.0364	30

Besides, we also found that several validation sites have extremely similar geographic coordinates to the training sites (Table 2). These sites are from the same observation network at local scale. These sites do not belong to the same site at both training and validation sites groups, e.g., the Lath_US-Tw1 in the training group and Lath_US-Tw1 (-2, -3) in the validation group. To deal with the issue, we have adopted the method that the mean values from these sites' measurements within 5-km extent were used to match the grid data, as mentioned in section 3 (Line 224).

Table 2 Summary of sites with the similar geographic coordinates in training and validation groups.

Training site name	Latitude (°)	Longitude (°)	Elevation (m)	Validation site name	Latitude (°)	Longitude (°)	Elevation (m)
FGI_MET0002	67.361866	26.637728	179	FGI_VUO0002	67.361883	26.643233	180
HAWS17	38.8451	100.36972	1559.63	HAWS16	38.84931	100.36411	1564.31
Lath_CA-SCB	61.3089	-121.2984	280	Lath_CA-SCC	61.3079	-121.2992	285
Lath_US-Tw1	38.1074	-121.6469	-9	Lath_US-Tw2	38.1047	-121.6433	-5
Lath_US-Tw1	38.1074	-121.6469	-9	Lath_US-Tw3	38.1159	-121.6467	-9
Lath_US-Tw1	38.1074	-121.6469	-9	Lath_US-Tw4	38.10298	-121.6414	-5
IMAU-S10	67.0005	-47.0167	1850	PM-KAN_U	67.0003	-47.0253	1840

To keep the independence of validation dataset from training samples, we removed duplicate three sites of the CF_HB, the FxMt_HFK, and the FxMt_PDF from the validation group. Note that we only use measurements of the sites with ETC coefficient of more than 0.9 to weaken upscaling errors of ground-based measurements. The ETC coefficients of the CF_HB and the FxMt_PDF are 0.7492 and 0.0337, respectively. Measurements from the two sites were previously not used in the validation activity. Therefore, we only need to delete the FxMt_HFK site with an ETC coefficient of 0.9225 from the validation group to evaluate the performance of the RCNN model again. Figure 1 shows the evaluated result based on the independent validation sites without/with the FxMt_HFK site. The uncertainty of R_n retrievals at validation sites changes slightly with RMSE values from 26.66 Wm^{-2} to 26.77Wm^{-2}.

Therefore, previous independent evaluation of R_n retrievals at validation sites is reliable although duplicate three sites are used in training and validation dataset simultaneously. We have revised Figure 5(b) and the corresponding content in the revised manuscript.

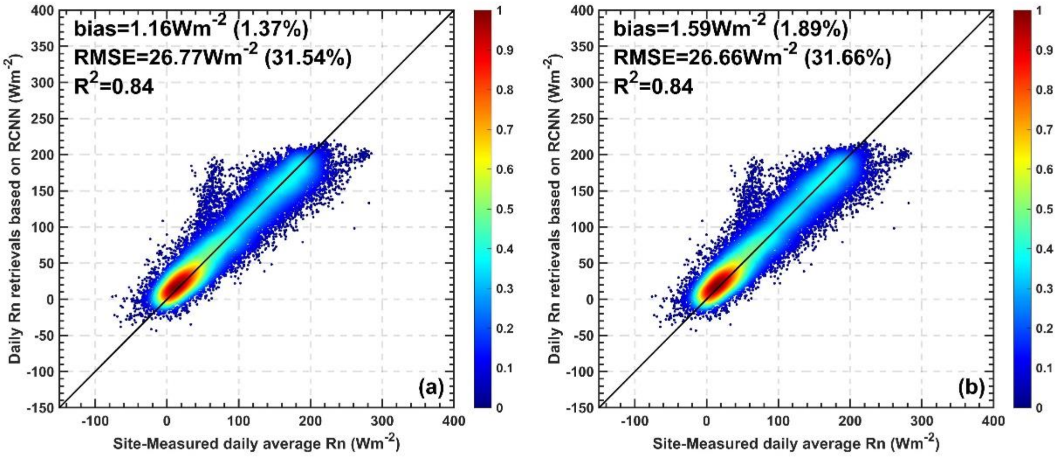

Figure 1: Evaluated results of RCNN model using independent validation dataset (a) without FxMt_HFK and (b) with FxMt_HFK site measurements, respectively.

Comment 3:

I am curious whether the results shown in Figure 7 reflect the fact that some of these networks are included in training the AVHRR dataset. It isn't clear to me from the description whether training stations were also used in this analysis, or whether this only includes independent testing stations and stations that didn't meet the reliability requirements. But even if these validation stations are independent from training data, the network of measurements around the ARM Southern Great Plains sites, for example, may be more similar to each other than a site that is located in a much different climate regime (e.g. independent sites ARM_E06 and ARM_E41 sites). That could lead to overfitting. ***It would be helpful to understand how independent this validation dataset is.***

Response 3:

The collected sites come from several local observation networks and international networks. Generally, sites of the local networks are located at the small region (e.g., ARM, HiWATER), while sites of the international networks are distributed over the globe (e.g., BSRN, FluxNet). To fully utilize these networks, we follow the idea of determination of training and validation sites in the GLASS R_n estimation algorithm (Jiang et al., 2016). For an observation network with multiple sites, we randomly selected several sites to serve as independent sites and remaining sites are used as training sites. Regarding to a local network with less sites, all sites are used as training sites to ensure the representativeness of the training dataset in characterizing spatiotemporal variation of surface R_n. Based on the strategy, the training and validation sites are finally determined. Therefore, the training and validation dataset both have great representation that reflect different surfaces (land cover types, elevations) and atmospheric properties (climate zone), which is important for evaluating model's robustness. Figure 2 shows the proportional distribution of training and validation sites under different conditions.

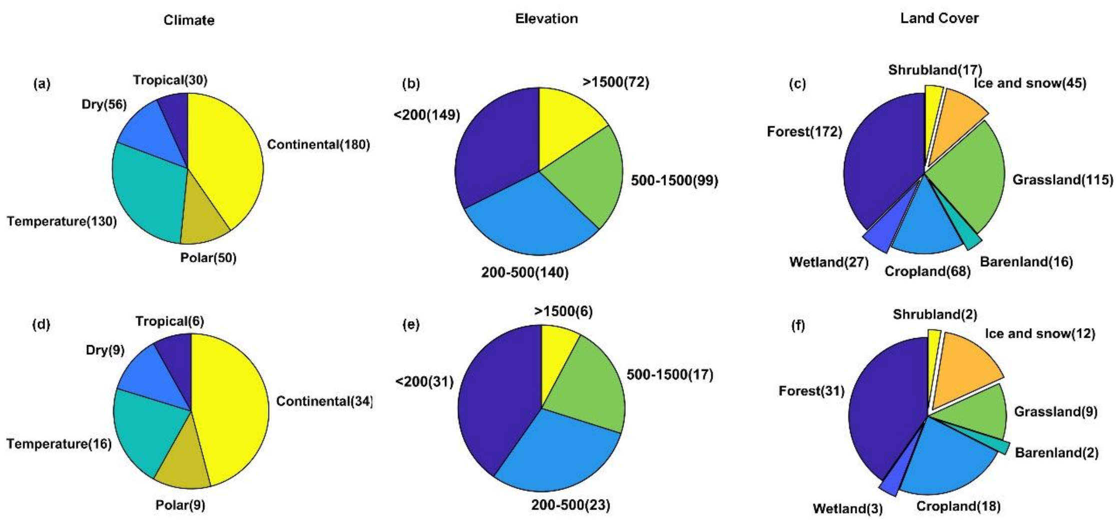

Figure 2: Proportional distributions of (a-c) training sites and (d-f) validation sites under different climates, elevation ranges, and land covers, respectively. The value in the brackets is total number of sites under specific condition.

The result in Figure 7 is obtained only using the independent validation dataset with ETC coefficients > 0.9 (reliable). We can see that AVHRR and GLASS R_n retrievals have comparable accuracies over most observation networks, and the overall validation result also illustrates that the difference of uncertainty in these two R_n datasets is small (< 1.63%). Specifically, the RMSE differences between AVHRR and GLASS R_n are -2.03 (ARM), -1.31 (BSRN), -1.34 (CEOP), -0.32 (CEOP-Int), 1.41 (EOL), -0.84 (AsiaFlux), -1.32 (FluxNet), 0.22 (PROMICE), and -0.99 (SURFRAD) Wm^{-2}, respectively. The performance of RCNN model over ARM network is better than other networks as ARM is a local network with extremely similar conditions for training and validation, which may reveal a false performance of RCNN model. However, some results from BSRN, FluxNet, EOL networks can reflect more information about RCNN model robustness at a larger spatiotemporal extent.

For the small regional network, measurements only reflect the spatiotemporal variation of R_n at a local extent. It is unsuitable to select many sites from local networks as validation sites to evaluate RCNN's independent performance when we want to retrieve surface R_n at global scale. Therefore, more sites from the international networks should be used as the validation sites. Fortunately, the number of site from the local networks is small in the validation group. Most validation sites were used come from the networks at continental or global scales. Specifically, the number of sites from the continental and global networks is more than 89%, including BSRN (2), CEOP (5), EOL (5), AsiaFlux (10), FluxNet (39), and PROMICE (7). Conversely, the number of sites from local networks is small with a proportion < 10%, including ARM (2), HiWATER (1), GAME.ANN (1). Besides, based on the response 2, there is no duplicate site in training and validation site groups, except the CF_HB, the FxMt_HFK, and the FxMt_PDF. Therefore, the independence of validation dataset is adequate to evaluate the overall performance of RCNN model at validation sites.

Comment 4:

Does Figure 14 show local time? Please label for clarity.

Response 4:

Thanks for your nice suggestion. Figure 14 shows the local time of NOAA-series satellites crossing. We have added the information in the caption of figure 14 and the corresponding phrase.

> Figure 14 shows the effect of the daily mean MERRA2 R_n on the final AVHRR R_n retrievals at different AVHRR overpass times in local time. The improved effect is slightly more significant during the afternoon than in the morning when more over-

> **Figure 14: Effect of daily mean MERRA2 R_n on AVHRR R_n retrievals at different satellite crossing times in local time. The bars indicate RMSE and lines indicate absolute biases. The shading shows the variation range of absolute bias.**

Minor comments:

Comment 1:

Line 50: “RT-based physical methods show a great generalization” I am not sure what this phrase means, please revise for clarity.

Response 1:

The phrase refers that different from empirical methods, the application of RT-based physical methods is not subjected to the limitation of training samples at a regional scale; in other words, the RT-based physical models are more applicable to a larger spatiotemporal extent. The phrase has been revised in the revised manuscript.

> empirical statistical methods. ~~RT-based physical methods show a great generalization~~ The RT-based physical methods are more applicable to a larger spatiotemporal extent because they consider the physical processes of solar radiation from the top

Comment 2:

Line 310: should it be: “data was *then* removed”?

Response 2:

Thanks for your careful examination. We have corrected the mistake in the revised manuscript.

> and divided into ten groups. One group of these data was ~~them~~ then removed as a hold-out or validation dataset and the remaining nine groups of data were treated as the training datasets. The training datasets were used to fit the RCNN model,

Comment 3:

Line 346: “for in surface radiation estimations.” Wording doesn’t seem quite right here.

Response 3:

Thanks for your careful work. We have revised the phrase in the revised manuscript.

identified as unreliability due to the presence of large water bodies within the satellite footprint. Thus, the processing of identifying reliable sites highlights the need to pay more attention to such areas for ~~in~~ surface radiation estimations.

Comment 4:

Lines 360-361: This sentence is awkwardly written and should be revised. Changing consistently to consistent, and site to sites would improve readability.

Response 4:

Thanks for your kind suggestion. We have revised the phrase according to your comments.

implementation of ETC for the selection of reliable ~~site~~ sites ensures more ~~consistently~~ consistent spatial representativeness of ground-based measurements and AVHRR data, which improves the accuracy of R_n retrievals. Indeed, the CV-derived average

Comment 5:

Line 483: should be very instead of “vary”

Response 5:

Thanks for your careful examination. We have corrected the mistake.

for the RCNN training. In addition, the AVHRR R_n retrievals show steady and ~~vary~~ very low (close to zero) biases under different conditions, while the biases of the GLASS R_n retrievals show a high degree of variation. This illustrates that the

Comment 6:

Line 545: should “produced” be replaced?

Response 6:

Thanks for your careful examination. We have revised the phrase.

alternative update times of the NOAA-series satellites. For example, NOAA-11 was successfully ~~produced~~ succeeded by NOAA-14 from 1994 to 1995. Important gaps and noise were found in the images from March to September and empty data

Comment 7:

Line 563: GLASS is misspelled GALSS

Response 7:

Thanks for your careful examination. We have corrected the misspelling.

Figure 12: Long-term temporal variation of (a) monthly average R_n and (b) monthly R_n anomalies for the AVHRR, CERES, ~~GALSS~~ GLASS and MERRA2 datasets, respectively. The shading represents the variation range (stand deviation) of the global monthly

Comment 8:

Line 572: I think that 7.08 must be 0.78. Please check.

Response 8:

Thanks for your careful examination. After looking back at the evaluated result, the R-value is 0.78, not 7.08. We have corrected the mistake.

the average R increases from 0.61 to ~~7.08~~ 0.708, and RMSE decreases from 50.12 to 46.17, respectively, for the MLR model. As the valid spatial extent increases, essential and complete spatial features are exposed and incorporated into the MLR model,

Comment 9:

Line 690: I think “satellite replacement works” should be satellite replacement work if you are referring to times when there is no satellite data because it the satellites are being worked on.

Response 9:

Thanks for your kind suggestion. We have revised the phrase according to your comments.

filling method or multi-source data-fusion algorithm is required to fill the data gaps over land, especially during periods of satellite replacement work. Third, coupled with spatially adjacent information, real-time temporal information, or historical information should be incorporated to further improve the accuracy of the R_n retrievals.

Comment 10:

Line 697: should be “covered surfaces”.

Response 10:

Thanks for your careful work. We have corrected the mistake.

covered ~~surface~~ surfaces. To address this problem, more physical knowledge is needed to fully utilize data-driven modeling to estimate surface R_n under different atmospheric and surface conditions. In particular, more attention should be paid to

Reference:

Jiang, B., Liang, S., Ma, H., Zhang, X., Xiao, Z., Zhao, X., Jia, K., Yao, Y., and Jia, A.: GLASS daytime all-wave net radiation product: Algorithm development and preliminary validation, Remote Sensing, 8, 222, 2016.