# An integrated and homogenized global surface solar radiation dataset and its reconstruction based on a convolutional neural network approach<s>an artificial intelligence approach</s>

Boyang Jiao[1,#], Yucheng Su[2], Qingxiang Li*[1,#], Veronica Manara[3],Martin Wild[4]

[1]School of Atmospheric Sciences, Sun Yat-sen University, and Key Laboratory of Tropical Atmosphere−Ocean System, Ministry of Education, Zhuhai 519082, China

[2]Meteorological Bureau of Zhuhai, Zhuhai 519082, China

[3]Department of Environmental Science and Policy, Università degli Studi di Milano, via Celoria 10, 20133, Milano, Italy

[4]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

[#]Southern Laboratory of Ocean Science and Engineering (Guangdong Zhuhai), Zhuhai 519082, China

*Correspondence to:* Qingxiang Li (liqingx5@mail.sysu.edu.cn)

**Abstract**

Surface solar radiation (SSR) is an essential factor in the flow of surface energy, enabling accurate capturing of long-term climate change and understanding the energy balance of Earth's atmosphere system. However, the long-term trend estimation of SSR is subjected to significant uncertainties due to the temporal inhomogeneity and the uneven spatial distribution of the *in-situ* observations. This paper develops an observational integrated and homogenized global-terrestrial (except for Antarctica)) stational SSR dataset (SSRIH$_{station}$) by integrating all available SSR observations, including the existing homogenized SSR results. The series ~~are~~ is then interpolated in order to obtain a 5°×5° resolution gridded dataset (SSRIH$_{grid}$). On this basis, we further reconstruct a long-term (1955-2018) global land (except for Antarctica) SSR anomalies dataset with a 5°×2.5° resolution (SSRIH$_{20CR}$) by training improved partial convolutional neural network deep learning methods based on the reanalysis 20CRv3. Based on this, we analysed the global land (except for Antarctica) /regional scale SSR trends and spatiotemporal variations: the reconstruction results reflect the distribution of SSR anomalies and have high reliability in filling and reconstructing the missing values. At the global land (except for Antarctica) scale, the decreasing trend ~~ofthe~~ of the SSRIH$_{20CR}$ (-1.276±0.205 W/m$^2$ per decade) is slightly smaller than the trend of the SSRIH$_{grid}$ (-1.776±0.230 W/m$^2$ per decade) from 1955 to 1991. The trend of SSRIH$_{20CR}$ (0.697±0.359 W/m$^2$ per decade) from 1991 to 2018 is also marginally lower than that of the SSRIH$_{grid}$ (0.851±0.410 W/m$^2$ per decade). At the regional scale, the difference between the SSRIH$_{20CR}$ and SSRIH$_{grid}$ is more significant in years and areas with insufficient coverage. Asia, Africa, Europe and North America cause the global dimming of the SSRIH$_{20CR}$, while Europe and North America drive the global brightening of the SSRIH$_{20CR}$. Spatial sampling inadequacies have largely contributed to a bias in the long-term variation of global /regional SSR. This paper's homogenized gridded dataset and the Artificial Intelligence reconstruction gridded dataset (Jiao and Li, 2023) are all available at https://doi.org/10.6084/m9.figshare.21625079.v1.

## 1 Introduction

Energy flows at the Earth's surface play an essential role in climate change and human activity and link to physical processes such as global warming, glacier retreating, hydrological cycle, and carbon budget (Hoskins and Valdes, 1990; Peixoto et al., 1992; Trenberth and Fasullo, 2013; Wild, 2012). As a critical factor characterizing surface energy flows, Surface Solar Radiation (SSR) largely determines the climatic conditions and ecological environment in which we live. Therefore, a more accurate and comprehensive analysis of the SSR fluxes will help better understand the Earth's atmospheric system. *In-situ* observations provide the most accurate baseline data for measuring SSR. They allowed for the first time the detection of decadal changes in SSR known as "dimming and brightening" (Wild et al., 2005), especially considering that they cover a longer period concerning another type of data like for example satellite data (Pfeifroth et al., 2018). ~~even~~ Even ~~if~~ observational data often have uneven distribution and missing data with respect to the satellite data, especially in areas with complex orography (Manara et al., 2020).

The sources of *in-situ* SSR observations are mainly collected from the Global Energy Balance Archive (GEBA) (Wild et al., 2017) and the World Radiation Data Centre (WRDC) (Tsvetkov et al., 1995). Furthermore, other SSR station series are obtained from the high quality Baseline Surface Radiation Network (BSRN) (Driemel et al., 2018) and the data centres of individual national hydrometeorological services. However, two issues still need to be addressed: 1) the inhomogeneity of station data resulting from station relocations and instrumentation changes severely impacts the climate change assessment. For the regions with a relatively high density of stations, like Europe (Manara et al., 2019; Manara et al., 2016; Sanchez-Lorenzo et al., 2013a; Sanchez-Lorenzo et al., 2015; Sanchez-Lorenzo et al., 2013b), Japan (Ma et al., 2022) and China (Ju et al., 2006; Wang, 2014; Wang et al., 2015; Wang and Wild, 2016; Yang et al., 2018b; You et al., 2013), much previous work has redefined the degree and timing of "dimming and brightening" by addressing the inhomogeneity of the SSR data series. For example, in Spain, the average annual homogenized SSR series has a significant increasing trend (+ 3.9 W/m$^2$ per decade) during the 1985–2010 period (Sanchez-Lorenzo et al., 2013a). The period of dimming observed in Italy's homogenized SSR series is not apparent in the 1960s and early 1970s when the raw series (inhomogenized) are taken into account (Manara et al., 2016). The direct measurements of SSR show a level trend from 1961 to 2014 over Japan, while their homogenization series display a decreasing trend

71  (0.8-1.6 W/m$^2$ per decade) (Ma et al., 2022). In China, homogenization largely eliminated the dramatic

72  non-climatic rise of the early 1990s and also reduced the increasing trend from 1990 to 2016 (Yang et

73  al., 2018b). However, most of the research was still limited to regional scales. 2) The issue of limited

74  spatial sampling of long observational stations and their uneven distribution especially over areas with

75  complex orography. Considerable efforts have been devoted to filling in /interpolating the missing values

76  in climate datasets ("spatial analysis") (Collins, 1996; Erxleben et al., 2002; Scudiero et al., 2016). The

77  traditional spatial interpolation methods commonly used include Inverse Distance Weighted (Fisher et

78  al., 1993; Shepard, 1968), Kriging (Krige, 1951), Thin-Plate Splines (Bookstein, 1989) et cetera. Since

79  the 1980s, physical parametric interpolation (Feng and Wang, 2021; Tang et al., 2019) and Bayesian

80  fusion schemes (Aguiar et al., 2015) based on multi-source observational data were widely used, when

81  the emergence of highly accurate and relatively precise satellite data. However, the resulting fusion

82  datasets cover a too short period to investigate their decadal and multi-decadal variations and to study

83  the underlying causes. The spatial, temporal, and spectral coverage of a single satellite is limited, and

84  multiple satellite data are therefore often used in tandem with each other; however, such a

85  discontinuity in time and space can introduce inhomogeneity into a dataset (Evan et al., 2007; Feng

86  and Wang, 2021; Shao et al., 2022). Reanalysis products are an important complement containing

87  long-term SSR data, therefore have been widely used in climate studies (Huang et al., 2018; Jiao et

88  al., 2022; Urraca et al., 2018; Zhou et al., 2018a; Zhou et al., 2017) due to the dynamically consistent

89  and spatiotemporally complete atmospheric fields with high resolution and open access to data.

90  However, existing studies have shown that reanalysis products generally overestimate multi-year

91  mean SSR values compared to observations over land (He et al., 2021). With the continuous

92  development of climate system simulations, model data from the Coupled Model International

93  Program (CMIP) have become an important resource for conducting climate change research (Gates

94  et al., 1999; Zhou et al., 2019). Previous studies have shown that the models used in CMIP6

95  overestimate the global mean SSR (He et al., 2023; Jiao et al., 2022; Wild, 2020). The rise of deep

96  learning and big data techniques has brought about an explosion of artificial intelligence (AI). Machine

97  learning is increasingly being used in spatial interpolation, such as the spatial reconstruction of surface

98  temperature datasets (Huang et al., 2022; Kadow et al., 2020; Cao et al., 2022), the spatial and temporal

99  reconstruction of turbulence resolution (Fukami et al., 2021), etc. Furthermore, it shows high accuracy

100  and low uncertainty in reproducing and predicting SSR (Leirvik and Yuan, 2021; Tang et al., 2016; Yang

101 et al., 2018a; Yuan et al., 2021). However, long-term homogenized SSR datasets with global terrestrial

102 coverage have yet to be developed, resulting in significant uncertainties in assessing global SSR variation

103 (Jiao et al., 2022).

104     Therefore, developing a more homogeneous and comprehensive global long-term SSR climatic dataset

105 that provides a better benchmark for observational constraints on the global surface energy balance

106 /budget remains a valuable and challenging task. This paper first homogenizes and grids the most

107 extensive collection of available global SSR station observations. Then, the missing grid boxes /years

108 are spatially interpolated using a convolutional neural network (CNN) approach to obtain a globally

109 covered land surface SSR anomalies dataset. Finally, the reconstructed datasets are initially analysed and

110 evaluated. Thus, the paper is divided into seven main sections. The data resources are introduced in

111 Section 2. Section 3 presents the data homogenization, and the CNN model reconstruction methods. The

112 data homogenization and verification are shown in Section 4. Section 5 gives the AI reconstruction results.

113 Section 6 is the availability of the datasets. Conclusions are provided at the end of the paper.

114 **2 Data**

115 Nine SSR datasets are collected to derive the global SSR variable. In particular, six datasets contain data

116 from observational stations (Section 2.1): two global ground-based measurement datasets (GEBA,

117 WRDC) and four homogenized products at regional and country levels (Europe, China, Japan and Italy).

118 Three of the adopted datasets are reanalysis data (Section 2.2.1): Fifth generation European Centre for

119 Medium-Range Weather Forecasts (ECMWF) reanalysis (ERA5), 20th Century Reanalysis version 3

120 (20CRv3) reanalysis data and the Coupled Model Intercomparison Project Phase 6 (CMIP6) historical

121 simulation output (125). Specifically, the ERA5 data are used to fill the data over oceans and Antarctica

122 (Section 3.2.1), 20CRv3 data and CMIP6 simulations are used for the AI model training (Section 5.1)

123 and reconstruction. All have been listed in Table 1.

124 **2.1 In situ observational Data**

125 **2.1.1 Global datasets**

126 There are two main sources of raw SSR data (see Table 1): the ETH Zurich GEBA with monthly data

127 from 2,445 globally distributed stations, starting from 1922 until 2020, and the WRDC dataset with

monthly globally distributed data from 1136 stations since 1964. The first one is available for download at https://geba.ethz.ch (Last access: 2022.7. 2). The second one published the first SSR radiation balance data in 1965 and then its publication has been issued four times a year since 1993 and is available for download at http://wrdc.mgo.rssi.ru/ (Last access: July 2021).

**2.1.2 National (regional) homogenized station datasets**

1) Chinese homogenized SSR dataset

The China Meteorological Radiation Fundamental Elements Monthly Value Data Set has been downloaded at http://www.nmic.cn. The homogenized SSR dataset in China is released by the National Meteorological Information Centre (NMIC), China Meteorological Administration (CMA) (Yang, 2016). The data are available for the period between Jan 1950 to Dec 2014, and the follow-up data are extended with raw observations from NMIC. They used the sunshine duration (SSD) data from nearby stations to construct an arguably better reference to identify inhomogeneities in the SSR data. Then, a combined metadata and the maximum penalty t-test (PMT) method was used to detect the change points. Finally, they were adjusted by a quantile matching (QM) algorithm (Wang and Feng, 2013). The final homogenized SSR station dataset was converted to gridded data using the first difference method (FDM (Peterson et al., 1998)) and is available for download at http://www.nmic.cn. Last Access: September 2022.

2) Japanese homogenized SSR dataset

Ma et al. (Ma et al., 2022) released a Japanese SSR homogenized dataset in 2022 spanning the period between 1870 and 2015. First, they homogenized SSD based on PMF (penalized maximal F test) and QM algorithms. They then used the homogenized SSD from the previous step as a reference series, combined with metadata and PMT, to detect change points. Finally, they adjusted the change points by the QM algorithm. For more details on data descriptions, the adopted methodology and downloading data refer to https://data.tpdc.ac.cn/en/data/45d73756-3f5a-4d27-82a4-952e268c20e8/, Last Access: March 2022.

3) European homogenized SSR data

A homogenized dataset of European SSR stations was developed by Sanchez-Lorenzo et al. (Sanchez-Lorenzo et al., 2015) and is currently available as a full public download at https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2015JD023321. They selected the 56 longest

157 Central European SSR series available in GEBA dataset with data for the period comprised between

158 1922 and 2012. They adjusted them to ensure temporal homogeneity homogenizing the data with the

159 Standard Normal Homogeneity Test (Alexandersson, 1986) and the Craddock test (Craddock, 1979).

160 4) Italian homogenized SSR dataset

161 The Italian homogenized SSR datasets are those published by (Manara et al., 2019; Manara et al.,

162 2016). As candidate stations to use as reference series, they selected the ten series located in the same

163 area of the series to be tested and that series correlate well with the test one. In particular, they tested the

164 change points with the Craddock test (Manara. et al., 2017) and when a break is identified by more than

165 one reference series the preceding portion of the series is corrected, leaving the most recent portion

166 unchanged. In this way, the SSR stations were homogenized, and then the missing values were

167 interpolated.

168 **2.2 Other datasets**

169 **2.2.1 Reanalysis**

170 ERA5 can be used to fill in SSR data from the oceans and Antarctica and carry out the global

171 reconstruction, taking into account its high spatial resolution and reliable performance of SSR (Jiao et

172 al., 2022; Liang et al., 2022). After the reconstruction, we removed the data for the ocean reanalysis and

173 maintain the data only in the land area (except for Antarctica). In addition, two SSR data products

174 (20CRv3, CMIP6) are used to train AI models. These are:

175 1) ERA5 (space-filling data): ERA5 is the fifth generation of the European Centre for Medium-Range

176 Forecasting reanalysis product, which currently publishes data from 1950 to the present (Hersbach et al.,

177 2020). In addition, ERA5 has an hourly output and an uncertainty estimate from the ensemble. The data

178 is based on the Integrated Forecasting Model Cy41r2 run in 2016, which contains a 4D-Var assimilation

179 scheme. In ERA5, SSR is obtained from a Rapid Radiation Transfer Model (RRTM) (Mlawer et al.,

180 1997). The present study utilizes monthly SSR data for the period 1955-2018 from ERA5 with a

181 resolution of $0.25°\times0.25°$ (last accessed in July 2022). It can be downloaded at

182 https://cds.climate.copernicus.eu

183 2) 20CRv3 (data for AI model training): The 20CR Project is an effort led by NOAA's Physical

184 Sciences Laboratory and CIRES at the University of Colorado, supported by the Department of Energy,

185 to produce reanalysis datasets spanning the entire 20th century and much of the 19th century (Slivinski

et al., 2019). 20CR provides a comprehensive global atmospheric circulation data set from 1850 to 2015. Its chief motivation is to provide an observational validation dataset, with quantified uncertainties, for assessing climate model simulations of the 20th century. 20CR uses an ensemble filter data assimilation method which directly estimates the most likely state of the global atmosphere every three hours and estimates the uncertainty in that analysis. The most recent version of this reanalysis, 20CRv3, provides 8-times daily estimates of global tropospheric variability across 75 km grids, spanning 1836 to 2015 (with an experimental extension from 1806 to 1835). The present study uses monthly SSR data of 20CRv3 (NOAA /CIRES /DOE 20CR, 80 members) from 1955-2015. We selected all 80 members of the 20CR as input (1 for evaluation and to test reconstruction, the other 79 for training the CNN model). The SSR of 20CRv3 has a spatial resolution of 0.7°×0.7° (Last accessed: May 2022). The download is available at https://portal.nersc.gov/archive/home/projects/incite11/.

**2.2.2 CMIP6 models output**

3) CMIP6 models output (data for AI model training): the Coupled Model Intercomparison Project, driven by the World Climate Research Program, is now in its 6th phase. Specifically, CMIP6 is considered as the current state of the art way of producing future climate simulations, including predicting future SSR based on different climate scenarios (Zhou et al., 2018b). It provides an important resourceexcellent resource for studying current and future climate change (Eyring et al., 2016). The historical simulations of CMIP6 are designed to reproduce observed climate and climate change, constrained by radiative forcing. Its historical simulation spans between 1850 and 2014. In this study, we selected 125 members out of a total of 507 members from several CMIP6 large ensemble models (with more than 10 realizations/runs) with high correlation coefficients with observations as input to train and validate the CNN model (1 for evaluation and to test reconstruction, the other 124 for training the CNN model)we selected 125 members from historical simulations of a large sample (a total of 507 members). These 125 members match better with the *in situ* observations than the other (507-125) members. We selected the monthly downward shortwave radiation from 1955 to 2014 (see Table S1 in the Supplemental Material (SM)). Last access July 2022. Download at: https://esgf-node.llnl.gov/search/cmip6.

213    **3 Methods**

214    **3.1 Data Quality Control (QC) and homogenization**

215    The SSR data homogenization method is only applied to the two inhomogenized *in-situ* observations

216    datasets (GEBA and WRDC). The Quality Control (QC) and homogenization flowchart (Figure 1) is

217    divided into three steps: 1. QC; 2. Homogenization; 3. Integration and consolidation.

218    **3.1.1 QC**

219    The QC of SSR data includes the following steps:

220    1) Simple integration: integration of the GEBA (2445) and WRDC (1136) datasets removing stations

221    with no data and leaving 2681 stations.

222    2) Removing duplicate stations: a. Stations with similar latitude and longitude. We consider two

223    stations with totally identical latitude and longitude to be the same station; b. Stations less than 10km

224    apart. We averaged the duplicate stations in this a and b case; c. Special duplicate stations: Stitching

225    together data of the duplicate stations based on metadata from CMA.

226    3) Remove stations or years /months for which a climatic analysis cannot be established: we remove

227    stations with records of less than ten years and values more than three~~five~~ times (3σ criterion (Olanow

228    and Koller, 1998) the standard deviation of the SSR anomalies.

229    4) Candidate stations (487) with a record length greater than 15 years in the period 1971-2000 are

230    selected. We added stations (715) with more than 10 years of SSR records to increase the number of

231    available stations for a better homogenization of the candidate stations (Figure 2).

232    **3.1.2 Station series homogenization**

233    This paper uses the RHtestV4 software package to test and adjust the SSR station data for homogeneity

234    (http://etccdi.pacificclimate.org/software.shtml) (Wang and Feng, 2013). The package is based on the

235    empirical penalty functions PMF (Wang, 2008b) and PMT (Wang, 2008a; Wang et al., 2007) for the

236    homogenization test. It takes into account the lag-1 autocorrelation of the time series. It embeds a multiple

237    linear regression algorithm to significantly reduce the problem of an unbalanced distribution of pseudo-

238    identification rates and test efficacy. Also, RHtestV4 uses the QM algorithm (Vincent et al., 2012; Wang

239    et al., 2010) and Mean-Adjustments to adjust the identified change points.

240    The specific steps are as follows:

241     1) Building the reference series

242       a. We processed the data from all stations series (715) into the annual first differences (FD) series

243     $e_i$(Eq. (1)) (Peterson et al., 1998).

244       b. We calculated the correlation of the annual FD series between the series from the potential reference

245     pool and the candidate stations.

246       c. We calculated the distance between the potential reference pool stations and candidate stations.

247       d. We selected potential stations according to the correlation coefficient (CC >= 0.6) between the series

248     from potential reference pool and candidate stations. And the potential stations also satisfy the limits in

249     distances (<= 500km) between the potential pool stations and candidate stations.

250       e. We obtain the reference FD series ($Re$)based on the $m$ potential reference series (Pe$_i$) and the CCs

251     ($c_i$) between the potential reference series (Pe$_i$) and candidate stations series (Eq. (2)).

252       f. The synthesized reference FD series ($Re$) (Eq. (2)), plus the average of all potential reference series

253     ($\bar{R}$), yields the final annual reference series ($R$) (Eq. (3)).

$$e_i = x_i - x_{i+1}$$

$$\text{i=1, 2, …, n-1} \qquad (1)$$

$$R_e = \frac{\sum_{i=1}^{m} Pe_i * c_i^2}{\sum_{i=1}^{m} c_i^2} \qquad (2)$$

$$R = R_e + \overline{R} \qquad (3)$$

254     $e_i$   Annual FD series,

255     $x_i$   Raw observational station SSR in the year i,

256     $Re$   Final reference series,

257     $Pe_i$   Potential reference series,

258     $c_i$   CC between the potential reference series and the candidate stations series.

259     2) Testing and adjusting the candidate series

260       The homogenization test algorithm used in this paper is the PMT. This method is a reference series-

261     dependent test for a normalized candidate series. It assumes that the linear trend of the time series is zero

262     and uses the degree of mean deviation at different points in the series to find change points. Furthermore,

263     it eliminates the effect of different sample lengths on the test results. At the same time, the method

264     introduces an empirical penalty factor, which effectively improves detection. We used the PMT to test

265     the homogeneity of the candidate series based on the reference series established in 1). We then adjusted

266     the statistically significant(p>0.05) changepoints obtained using the mean adjustment method (p>0.05).

267     We homogenize the monthly series for 66 stations (see Figure S1 in the SM).

268     **3.1.3 Integration and consolidation**

269     As can be seen from Figure 1, the candidate stations (487) are relatively sparse. To better adapt deep

270     learning methods for the dataset reconstruction later, we adjusted, added and integrated station series

271     based on the results of homogenized data from other scholars: 1) We added stations with more than 10a

272     overall (1955-2018) records but no more than 15a during the 1971-2000 period, and removed those

273     stations that were clearly inhomogeneous (25) and some years of the station (3); 2) We subsequently

274     integrate monthly SSR series for 116 stations based on the results of homogenization by other scholars

275     (China (56), Japan (8), Europe (2) and Italy (50)). After the above steps, we end up with a homogenized

276     dataset containing 944 stations (Figure 3). The details of the processing and classification are shown in

277     Table S2 (see in the SM).

278     **3.2 CNN model reconstruction methods**

279     The CNN deep learning model network architecture uses a U-shaped structure similar to the U-net

280     (Ronneberger et al., 2015). The advantage of using this model is: 1) both high and low-frequency

281     information of the picture can be retained, and when reconstructing the SSR data, not only the grid point

282     information close to the missing measurement point will be considered, but also information from more

283     distant locations (which may be remotely correlated with that missing measurement point); 2) This makes

284     the model convergence faster and more economical in terms of computational resources. The upper part

285     of the U-shaped structure, which has no down samples or a low number of down samples, represents the

286     high-frequency information of the graph. These sections contain much of the detail in the graph and the

287     relationships between similar grid points are conveyed by this section. The lower half of the U-shaped

288     structure is down-sampled more often and represents the lower frequency information of the graph. The

289     global radiation of a wide range of undulations is transmitted by it, and then the information at the various

290     levels of the U-shaped structure is connected and transmitted through the skip connection, allowing the

291     whole network to remember all the information of the picture very well. The model uses nearest

292     neighbour upsampling in the decoding phase, the skip links will concatenate two feature maps and two

293     masks as the feature and mask inputs for the next part of the convolution layer. The input to the last part

294  of the convolution layer will contain the original input image concatenated with the holes and the original

295  mask, allowing the model to replicate the ~~non-hole~~gap-free pixels. The complex and variable nature of

296  the sea-land boundary then has a significant impact on the reconstruction, when we reconstruct the global

297  land SSR data. Therefore, we use partial convolution at the image boundaries with a suitable image

298  padding, ensuring that the padding content at the image boundaries is not affected by values outside the

299  image. The deep learning models' convolutional layers and loss functions have been described ~~in SM~~in

300  the SM.

301  We further reconstruct a long-term (1955-2018) global SSR anomalies dataset (SSRIH$_{20CR}$) by using

302  improved partial CNN deep learning methods based on a "perfect" dataset. CNN consists of three parts.

303  A convolutional layer to reduce the number of weights by extracting local features, a pooling layer to

304  reduce peacekeeping and prevent overfitting, and a fully connected layer to output the desired result. In

305  this paper, a modified CNN network is used to model the reconstruction of the SSR data, with the

306  convolutional layer replaced by a partial convolution method and mask update. This method is the latest

307  in image restoration effects and can restore irregular holes, an advantage over other image restoration

308  methods that can only restore rectangular holes. Therefore, this paper uses the modified CNN model

309  (Kadow et al., 2020) to recover the missing part of the global terrestrial SSR (except Antarctica). The

310  specific reconstruction steps and processes are as in Figure 4.

311  **3.2.1 Data pre-processing**

312  The homogenized station data is converted to grid box anomalies using the Climate Anomalies Method

313  (CAM) (Jones et al., 2001). CAM is a commonly used method for converting station anomaly data to

314  gridded data. We divide all global areas into a $5° \times 5°$ grid, after which we calculate the SSR anomalies

315  (relative to 1923-2020) within the grid box by averaging the anomalies of all stations (at least 1 station

316  in it). If there are more than one site exists in the same grid box, the record length of this grid box is the

317  total length of all sites in that grid box. Finally, we removed the values that were more than ~~five~~ three

318  times the standard deviation of the SSR anomaly time series after gridding. SSRs are all processed as

319  daily average anomalies, i.e., monthly anomalies divided by 30 (each month is approximated as 30 days).

320  We multiplied all the values by 30 again when the reconstruction is complete. The global land (except

321  for Antarctica) distribution and coverage of SSRs after gridding are shown in Figure 5 a, b.

322    As seen in Figure 5a, the SSR is spatially sparsely distributed across South America and Africa. As

323    shown in Figure 5b, SSR coverage increased yearly from 1950 until the mid-1970s, when it slowly

324    decreased. In 2013, the coverage rate decreased sharply due to untimely data submission. Considering

325    the SSR coverage above, we only kept the years (1955-2018) with data coverage of more than 8% of

326    global land (except for Antarctica) areas.

327    Comparisons show that the ERA5 has high spatial resolution and relatively reliable performance in

328    the temporal variations and long-term trends (Liang et al., 2022; Jiao et al., 2022). To obtain a higher

329    data coverage and ensure that the AI model runs well, we used the ERA5 to fill the SSR of homogenized

330    global gridded SSR in the Antarctic and ocean areas. However, if we use the SSR of ERA5 to directly

331    fill the SSR of homogenized global gridded SSR (SSRIH$_{grid}$) in the Antarctic and on the ocean areas,

332    then the relatively weaker ocean SSR variations (variabilities, decadal changes, trends, etc.) from ERA5

333    will inevitably introduce certain systematic biases in land SSR reconstruction due to the SSRs have the

334    lower coverage on the land. Therefore, we designed an algorithm to avoid excessive diffusion of SSR

335    system bias in terrestrial areas: we first calculated the ratios $\gamma_i (i=1, 2, 3, ...., n)$ between the SSR from

336    ERA5 and from SSRIH$_{grid}$ on the land in all $n$ years. For a single grid box, the $\gamma_i$ have small changes

337    and are regarded as a constant $\gamma_{median}$ (Eq (4)), and the $\gamma_{median}$ vary by latitude and longitude both on

338    the marine and the land areas. We then extrapolated the $\gamma_{median}$ for all the grid boxes along the land

339    and sea boundaries. If there is no observation there, then the adjacent ocean ERA5 SSR is used to take

340    its place after it is adjusted according to the differences between the SSR variations (represented by the

341    linear trends) for the different underlying surfaces (Eq (5).

$$\gamma_{median} = Median(\frac{OBS_{i\_land}}{ERA5_{i\_land}}), \tag{4}$$

$$OBS_{i\_O\&L}(land) = ERA5_{i\_O\&L}(Ocean) * \gamma_{median} * \frac{T_O}{T_L}, \tag{5}$$

$$i = 1,2,3......,n$$

342    $\gamma_{median}$: The median value of the ratios of OBS and ERA5 land SSR series,

343    $OBS_{i\_land}$: Land SSR for the year i from SSRIH$_{grid}$ in a single grid,

344    $ERA5_{i\_land}$: Land SSR for the year i from ERA5in a single grid,

345    $OBS_{i\_O\&L}(land)$: LandSSRalong the sea-land boundary(land) for the year $i$ from SSRIH$_{grid}$,

346    $ERA5_{i\_O\&L}(Ocean)$: Ocean SSR along the sea-land boundaryfor the year $i$ from ERA5,

347    $T_O$: Trend of ERA5 SSR on ocean areasin all $n$ years,

348     $T_l$: Trend of ERA5 SSR on areas in all $n$ years.

**3.2.2 AI Model reconstruction**

350     We use a server (configured with processor Intel (R) Core (TM) i7-8700 CPU @ 3.20GHz 3.19 GHz,

351     RAM 32G, 64-bit OS, GPU model 516.94, NVIDIA GeForce 1080T version, Python 3.9.12 64-bit,

352     CUDA 10.1) for AI models training. The specific training steps are as follows:

353     1) A total of 768 missing value masks (monthly masks between 1955 and 2018) were prepared for

354     training and validation using '1' for existing and '0' for missing values;

355     2) The 20CRv3 /CMIP6 training set (monthly values between 1955 and 2015 /2014) and missing value

356     masks are fed into the 20CR-AI /CMIP6-AI model for training;

357     3) We perform 1,500,000 training sessions with an interval of 10,000 sessions for the training output

358     model.

359     Afterwards, the two AI models are validated against the root mean squared error (RMSE) /CCs of the

360     reconstructed SSRs (SSR$_{20CR}$/SSR$_{CMIP6}$). The validation set SSRs, and the optimal number of training

361     cycles is 1,100,000 (see Figure S2, Figure S3 and Figure S4 in the SM). The initial hyper-parameters of

362     the model are set as follows; learning rate of 2e-4 and learning finetune of 5e-5. First, we set the batch

363     size to 16 in the first 500000 iterations and fine-tuned it to 18 in the last 10000000 iterations, for a total

364     of 1500000 iterations, to suppress the overfitting phenomenon generated during the training process, and

365     validate the model every 10000 times and early stopping if the validation shows a decreasing trend, the

366     final number of training times used is 1100000. Second, L2 regularization is also added to regulate the

367     loss function (see Eq. (9) in the SM).

368     The training result models generated by the different AI models are obtained separately for the

369     different training sets. The model is first used to reconstruct a reanalysis validation set with the same

370     missing value mask as the original observation dataset. This is followed by a validation of the

371     reconstruction against the original reanalysis dataset (calculation of CC and RMSE) to understand the

372     discrepancies in the model reconstruction.

**4 Data homogenization and verification**

374     We homogenized the original monthly stations /gridded SSR time series (SSRIH$_{station}$ /SSRIH$_{grid}$) using

375     the method in section 3.1.2. We selected six continental regions, excluding Antarctica and the Arctic,

376    from the eight continents of the world defined by Xu et al. (Xu et al., 2018) (Asia, Africa, South America,

377    Europe, North America, Australia, Antarctica and the Arctic). The decreasing trend of the $SSRIH_{grid}$ is

378    consistent with the original gridded SSR series ($SSRI_{grid}$) during 1955-1991 while the increasing trend

379    during 1991-2018 is weaker. At the regional scale, the $SSRIH_{grid}$ has a generally similar variation to the

380    $SSRI_{grid}$, and the $SSRIH_{grid}$ ~~is~~ usually more representative of climate change than $SSRI_{grid}$ at individual

381    stations.

382    Figure S5 (see in the SM) illustrates the long-term variations of global (Figure S5 (a) in the SM) and

383    continental land SSR (Figure S5 (b) in the SM) from the $SSRI_{grid}$ and $SSRIH_{grid}$ (except for Antarctica)

384    during 1955-2018. The most prominent change revolves around the adjustment around 1992: the SSR

385    anomalies were systematically adjusted upward from 1987 to 1992, while the SSR anomalies were

386    systematically adjusted downward from 1993 onwards. Thus, there is a significant decreasing trend for

387    both global land $SSRI_{grid}$ (-1.995±0.251 W/m$^2$ per decade) and global land $SSRIH_{grid}$ (-1.776±0.230

388    W/m$^2$ per decade) (except for Antarctica) from 1955 to 1991. While the increasing trend of the global

389    land $SSRIH_{grid}$ from 1991 to 2018 is 0.851±0.410 W/m$^2$ per decade, slightly smaller than the increasing

390    trend of the $SSRI_{grid}$ (0.999±0.504 W/m$^2$ per decade). It is worth noting that 1992 happened to be the

391    second year of the eruption of Mount Pinatubo, and the homogenized SSR data integrated in this paper

392    may be affected by this event. But overall, the homogenization also has limited effects on the global SSR

393    variations from Figure S5 (see in the SM), which is consistent with the influence of data homogenization

394    on a wide range of surface air temperatures (Brohan et al., 2006; Xu et al., 2013).

395    At the regional scale, the differences between the $SSRIH_{grid}$ and $SSRI_{grid}$ are more pronounced in Asia

396    and Europe (see Figure S5(b)in the SM). Asia's homogenized SSR show that the regional average SSR

397    has been declining significantly over the period 1958-90; this dimming trend mostly diminished over the

398    period 1991-2005 and was replaced by a brightening trend in the recent decade. The $SSRIH_{grid}$ in Asia is

399    higher than the $SSRI_{grid}$ from 1985 to 1990 and lower than the $SSRI_{grid}$ from 2012 to 2015. The $SSRIH_{grid}$

400    shows a more moderate short-term increase in Europe from 1960 to 1980. Note also that the Australian

401    raw data prior to 1988 were artificially detrended because at the time the Australia Weather Service was

402    afraid that the instruments would drift. Therefore, they detrended them and unfortunately did not store

403    the raw data, and the SSR evolution in Australia is artificial with no trend (Wild et al., 2005). In addition,

404    the $SSRI_{station}$ and $SSRIH_{station}$ comparisons for all 66 stations are shown in Figure S1 (see in the SM).

405 **5 AI reconstruction and comparison**

406 **5.1 Training of the AI model**

407     We produce two (20CRv3 /CMIP6) separate training and validation sets: we select the 1th member

408 data of the reanalysis data and the model data, respectively, as the validation set, and the remaining 79

409 (124) ensemble members as the training sets, where each ensemble member included 732 (720) months

410 of SSR data. Each validation set included 732 (720) samples, while the training sets contained 57828

411 (89280) ensemble members. All the above data, including the *in-situ* observations, are then resampled to

412 monthly anomalies of $5° \times 2.5°$.

413 We reconstruct the SSR of 20CRv3 /CMIP6 with missing values based on 20CRv3 /CMIP6 datasets

414 using the method in section 3.2 and obtain two reconstructions, $SSR_{20CR}$ and $SSR_{CMIP6}$, respectively. The

415 SSR of 20CRv3/CMIP6 with missing values uses the $SSRIH_{grid}$ mask between 1955 and 2015 /2014. We

416 compare the global land (except for Antarctica) /regional annual anomalies variation of $SSR_{20CR}$

417 /$SSR_{CMIP6}$. The results show that $SSR_{20CR}$ is significantly more consistent with the validation set than

418 $SSR_{CMIP6}$.

419     Figure 6(a) shows that the RMSE/CC of the $SSR_{20CR}$ (0.~~25~~247 W/m$^2$ /0.97~~0~~ W/m$^2$) are smaller /larger

420 than those of $SSR_{CMIP6}$ (0.~~52~~518 W/m$^2$ /0.93~~7~~ W/m$^2$) with the original 20CR /CMIP6 dataset. The

421 20CR-AI model has a better reconstruction ability for SSR at the global land (except for Antarctica) scale.

422 The RMSEs of the $SSR_{20CR}$ ($SSR_{CMIP6}$) are 1.46~~0~~ (2.41~~3~~) W/m$^2$, 1.~~11~~109 (1.83~~829~~) W/m$^2$, 2.~~22~~219

423 (2.~~60~~596) W/m$^2$ and 1.~~29~~286 (2.24~~235~~) W/m$^2$ in North America, Europe, Asia, and Northern

424 Hemisphere, whereas these values are 1.~~12~~116 (1.77~~766~~) W/m$^2$, 0.62~~2~~ (1.60~~2~~) W/m$^2$, 1.~~88~~877 (1.84~~839~~)

425 W/m$^2$ and 0.77~~2~~ (1.~~68~~679) W/m$^2$ in South America, Africa, Australia, and Southern Hemisphere

426 concerning the original 20CR /CMIP6 dataset, respectively. In other words, the RMSEs of the $SSR_{20CR}$

427 are smaller than those of $SSR_{CMIP6}$ for the original 20CR /CMIP6 dataset except for Australia. In

428 addition, the CCs of the $SSR_{20CR}$ ($SSR_{CMIP6}$) are 0.~~96~~958 (0.83~~0~~) W/m$^2$, 0.~~96~~958 (0.99~~987~~) W/m$^2$, 0.~~89~~

429 886 (0.~~67~~669) W/m$^2$, 0.93~~0~~ (0.~~97~~965) W/m$^2$, 0.~~94~~938 (0.93~~0~~) W/m$^2$, 0.94~~3~~ (0.~~92~~916) W/m$^2$, 0.~~94~~936

430 (0.~~88~~875) W/m$^2$ and 0.90~~3~~ (0.82~~2~~) W/m$^2$ in North America, Europe, Asia, Northern Hemisphere, South

431 America, Africa, Australia, and Southern Hemisphere with respect to the original 20CR /CMIP6

432 dataset, respectively. That is, the CCs of the $SSR_{20CR}$ are larger than those of $SSR_{CMIP6}$ to the original

433 20CR /CMIP6 dataset except for Europe.

434  Based on the above comparison, the higher uncertainty for CMIP6 model output possibly biases the

435  CMIP6-AI method. Thus, the accuracy of the $SSR_{20CR}$ is higher than that of the $SSR_{CMIP6}$ at both global

436  land (except for Antarctica) and regional scales. Therefore, we choose the reconstruction results of the

437  20CR-AI model as the final AI reconstruction dataset, and subsequent analysis in the following sections

438  is only based on this dataset.

439  **5.2 Comparison of the spatial and temporal variation characteristics**

440  We investigate the long-term trends and spatial and temporal variation of the $SSRIH_{20CR}$, compare the

441  differences between the $SSRIH_{20CR}$ and $SSRIH_{grid}$, and suggest: the area and magnitude of the high and

442  low centres of the $SSRIH_{20CR}$ are the same as those of the $SSRIH_{grid}$; the results of the global land (except

443  for Antarctica) reconstruction are consistent with "dimming and brightening"; the global dimming is

444  primarily dominated by decreasing trends in Asia, Europe Africa and North America, whereas Europe

445  and North America are contributors to the increasing trends.

446  Figure 7 shows the spatial distribution of the $SSRIH_{grid}$ and $SSRIH_{20CR}$ for the three months (July 1960,

447  July 1980, and July 2000). Figure S6 (see in the SM) displays the spatial distribution of annual $SSRIH_{grid}$

448  and $SSRIH_{20CR}$ from 1955 to 2018. Figure 7 also shows the area and the magnitude of the high and low

449  centres in the $SSRIH_{20CR}$ are the same as in the $SSRIH_{grid}$. The $SSRIH_{20CR}$ is mainly positive anomalies

450  in Africa and the Eurasian continent in July 1960, especially in India and the Middle East. Afterwards,

451  India showed a continuous and steady decline in SSR. This confirms the well-known phenomenon of

452  global dimming over India (Wild et al., 2009; Soni et al., 2016; Soni et al., 2012; Padma Kumari et al.,

453  2007; Kambezidis et al., 2012). In Australia, the $SSRIH_{20CR}$ is dominated by negative anomalies in July

454  1980 and positive anomalies in July 1960 and July 2000. In Greenland, the $SSRIH_{20CR}$ shows a large

455  positive anomaly during three months. In northern Russia, there is a high value in July 2000. The

456  reconstruction can better reflect the anomaly distribution of observation information, and the grid boxes

457  with the missing values are infilled and reconstructed, which has high reliability.

458  Figure 8 illustrates global land (except for Antarctica) annual anomalies variation and long-term trend

459  of the $SSRIH_{20CR}$ for the period of 1955-2018, 1955-1991 and 1991-2018. Table S3 in the SM

460  demonstrates the trends ~~assessment~~ of global SSR change evaluation ~~in~~ for various data sources ~~global~~

461  ~~SSR change from~~on different scales. Also, we compare the differences between the $SSRIH_{20CR}$ and

462  $SSRIH_{grid}$. The minimum value of the $SSRIH_{20CR}$ occurred in 1991 (-2.411 W/m$^2$). The decreasing trend

463     of the $SSRIH_{20CR}$ from 1955 to 1991 (-1.276±0.205 W/m$^2$ per decade) is slightly lower than that of the

464     $SSRIH_{grid}$ (-1.776±0.230 W/m$^2$ per decade). After that, the $SSRIH_{20CR}$ turns to an increasing trend of

465     0.697 ± 0.359 W/m$^2$ per decade from 1991 to 2018. This suggests that the difference between $SSRIH_{20CR}$

466     and $SSRIH_{grid}$ may be caused by the results observed in limited data coverage (such as in Africa and

467     North America) (Figure 9). After homogenization and reconstruction, the trend (-1.276 W/m$^2$ per decade)

468     from 1955 to 1991 corresponds to an overall reduction of -4.6 W/m$^2$ over the dimming period, while that

469     (0.697 W/m$^2$per decade) from 1991 to 2018 correspond to an overall increase of 2 W/m$^2$ over the

470     brightening period. This is in amazing agreement with the -4 W/m$^2$ for the dimming period and the 2

471     W/m$^2$ for the brightening period based on an overall surface energy budget assessment ((Wild, 2012) see

472     their Figure 1). Also, similar conclusions (incomplete coverage of observational data lead to an

473     underestimation of global warming trends) have been confirmed in global warming research (Gulev et

474     al., 2021; Li et al., 2021).

475     Figure 9 demonstrates the long-term annual anomaly variations of the $SSRIH_{20CR}$ in different regions

476     and its results compared to the $SSRIH_{grid}$. Table S4 in the SM shows the ~~trends~~ evaluation in continental

477     and hemispheric $SSRIH_{20CR}$ change trends ~~from~~on different scales. The $SSRIH_{20CR}$ shows a similar

478     annual anomaly variation to the global land (except for Antarctica) average trend in North America and

479     Asia, reaches a minimum in the late 1970s or early 1990s, and follows a moderate reversal. In Europe,

480     the $SSRIH_{20CR}$ shows a decrease (-2.180 ± 1.866 W/m$^2$ per decade) between 1963 and 1978 before

481     turning to brightening (1.081 ± 0.312 W/m$^2$ per decade). In South America and Australia (Southern

482     Hemisphere), the $SSRIH_{20CR}$ shows no significant variation. In Africa, the $SSRIH_{20CR}$ has a dimming

483     trend (-1.506 ± 0.496 W/m$^2$ per decade) from the 1950s to the 1990s, after which it remains levelled off

484     (0.340 ± 0.998 W/m$^2$ per decade). The $SSRIH_{20CR}$ shows a decreasing trend (-1.457 ± 0.246 W/m$^2$ per

485     decade) until the 1990s in the Northern Hemisphere and a brightening (0.887 ± 0.415 W/m$^2$ per decade)

486     afterwards. The annual average anomaly variations in regions and globally show that Asia, Africa,

487     Europe and North America are the four contributors to the global dimming, while Europe and North

488     America are two major contributors to the "brightening". This is in general agreement with the results

489     obtained by previous machine learning (Yuan et al., 2021). In addition, the discrepancy between the

490     $SSRIH_{20CR}$ and $SSRIH_{grid}$ is more significant in low-coverage areas (right) than in high-coverage regions

491     (left). It is particularly pronounced before 1980 and in South America. This suggests that the limited

492     surface observations are not representative of the continental variation in SSR.

493 To sum up, the AI reconstruction of this paper helps to decrease the uncertainties in SSR variations in

494 both spatial scales. Further, it shows that there may be a sampling error in the variations of the global

495 land (except for Antarctica) and regional SSR before reconstruction, leading to a systematic deviation in

496 the long-term trend of global land (except for Antarctica) or regional SSR.

497 **6 Data availability**

498 Both the $SSRIH_{grid}$ (the homogenized monthly gridded SSR data over 1923-2020) and the $SSRIH_{20CR}$

499 (the monthly 20CR-AI model reconstructed SSR data for 1955-2018) are currently publicly available on

500 the figshare website under DOI at https://doi.org/10.6084/m9.figshare.21625079.v1 (Jiao and Li, 2023).

501 These datasets are also available at http://www.gwpu.net for free.

502 **7 Conclusion**

503 In this study, we integrate global station observations based on the raw observational SSRs from GEBA

504 and WRDC, combined with existing homogenized SSR datasets from other scholars. Also, we

505 homogenize the globally distributed station data using the RHtestV4 software package. An improved

506 CNN deep learning algorithm is subsequently used to reconstruct the SSR anomalies. Thus, a

507 reconstructed SSR anomaly dataset, $SSRIH_{20CR}$, is obtained based on training sets (20CRv3), for the

508 years 1955-2018, with a resolution of $5°×2.5°$. The main results are as follows:

509  1) The first integrated and homogenized global SSR monthly dataset is developed, which contains 944

510 stations in total and covers the longest periods (from the 1920s to recent years). A $5°×5°$ grid boxes

511 version of the monthly SSR anomalies dataset is derived.

512  2) This paper develops $5°×2.5°$ full-coverage monthly land (except for Antarctica) SSR anomalies

513 reconstructed datasets based on the above observations, using the 20CRv3 to train the AI model.

514 Comparative validations /evaluations show that the $SSRIH_{20CR}$ provides a reliable benchmark for global

515 SSR variations.

516  3) On average, the global annual SSR variations based on the $SSRIH_{grid}$ are not significantly different,

517 except that the increasing (brightening) trend after 1991 is a little smaller for the latter. The short-term

518 brightening SSR in Europe from the 1970s- to the 1980s disappear at the regional scale. At the same time,

519 the brightening SSR after the 1990s in Asia slowed or postponed.

**Author contributions**

Boyang Jiao: Software, Data curation, Writing- Original draft preparation, Visualization, Investigation.

Yucheng Su: Software, Data curation.

Qingxiang Li: Methodology, Supervision, Conceptualization, Validation, Writing - Review and Editing.

Veronica Manara: Providing the homogenized Italian dataset, Writing - Review and Editing.

Martin Wild: Writing - Review and Editing.

**Competing interests**

At least one of the (co-) authors is a member of the editorial board of Earth System Science Data.

**Disclaimer**

Publisher's note: Copernicus Publications remains neutral about jurisdictional claims in published maps and institutional affiliations.

**Financial support**

This study is supported by the Natural Science Foundation of China (Grant: 41975105) and the National Key R&D Program of China (Grant: 2018YFC1507705; 2017YFC1502301). The Global Energy Balance Archive (GEBA) is co-funded by the Federal Office of Meteorology and Climatology MeteoSwiss within the framework of GCOS Switzerland. Global dimming and brightening research at ETH Zurich are supported by the Swiss National Science Foundation (Grant No. 200020 188601). Veronica Manara was supported by the "Ministero dell'Università e della Ricerca" of Italy [grant FSE – REACT EU, DM 10/08/2021 n. 1062].

539 **Reference**

540 Aguiar, L. M., Pereira, B., David, M., Díaz, F., and Lauret, P.: Use of satellite data to improve solar

541  radiation forecasting with Bayesian Artificial Neural Networks, Solar Energy, 122, 1309-1324,

542  doi:10.1016/j.solener.2015.10.041, 2015.

543 Alexandersson, H.: A homogeneity test applied to precipitation data, Journal of Climatology, 6, 661-675,

544  doi:10.1002/joc.3370060607, 1986.

545 Bookstein, F. L.: Principal warps: Thin-plate splines and the decomposition of deformations, IEEE

546  Transactions on pattern analysis and machine intelligence, 11, 567-585, doi:10.1109/34.24792, 1989.

547 Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., and Jones, P. D.: Uncertainty estimates in regional and

548  global observed temperature changes: A new data set from 1850, Journal of Geophysical Research:

549  Atmospheres, 111, doi:10.1029/2005JD006548, 2006.

550 Cao, Y., Jiao, B., Lan, X., Tan, J., Yang, Y., Sun, W., Li, Z., Luo, J., and Li, Q.: Reconstruction og China

551  global Merged Surface Temperature (CMST) based on an artifical intelligence approach

552  (submitted), Enviormental Scinence & Technology, 2022.

553 Collins, F. C.: A comparison of spatial interpolation techniques in temperature estimation, The 3rd

554  International Conference/Workshop on Integrating GIS and Environmental Modeling, Santa

555  Barbara, Santa Fe, NM; Santa Barbara, CA, 21-26 January1996 of Conference.

556 Craddock, J. M.: Methods of comparing annual rainfall records for climatic purposes, Weather, 34, 332-

557  346, doi:10.1002/j.1477-8696.1979.tb03465.x, 1979.

558 Driemel, A., Augustine, J., Behrens, K., Colle, S., Cox, C., Cuevas-Agulló, E., Denn, F. M., Duprat, T.,

559  Fukuda, M., and Grobe, H.: Baseline Surface Radiation Network (BSRN): structure and data

560  description (1992–2017), Earth System Science Data, 10, 1491-1501, doi:10.5194/essd-10-1491-

561  2018, 2018.

562 Erxleben, J., Elder, K., and Davis, R.: Comparison of spatial interpolation methods for estimating snow

563  distribution in the Colorado Rocky Mountains, Hydrological Processes, 16, 3627-3649,

564  doi:10.1002/hyp.1239, 2002.

565 Evan, A. T., Heidinger, A. K., and Vimont, D. J.: Arguments against a physical long-term trend in global

566  ISCCP cloud amounts, Geophysical Research Letters, 34, 10.1029/2006gl028083, 2007.

567 Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview

568    of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and

569    organization, Geoscientific Model Development, 9, 1937-1958, doi:10.5194/gmd-9-1937-2016,

570    2016.

571  Feng, F. and Wang, K.: Merging high-resolution satellite surface radiation data with meteorological

572    sunshine duration observations over China from 1983 to 2017, Remote Sensing, 13, 602,

573    doi:10.3390/rs13040602, 2021.

574  Fisher, N. I., Lewis, T., and Embleton, B. J.: Statistical analysis of spherical data, Cambridge university

575    press, doi:10.1017/CBO9780511623059, 1993.

576  Fukami, K., Fukagata, K., and Taira, K.: Machine-learning-based spatio-temporal super resolution

577    reconstruction of turbulent flows, Journal of Fluid Mechanics, 909, doi:10.1017/jfm.2020.948, 2021.

578  Gates, W. L., Boyle, J. S., Covey, C., Dease, C. G., Doutriaux, C. M., Drach, R. S., Fiorino, M., Gleckler,

579    P. J., Hnilo, J. J., Marlais, S. M., Phillips, T. J., Potter, G. L., Santer, B. D., Sperber, K. R., Taylor,

580    K. E., and Williams, D. N.: An Overview of the Results of the Atmospheric Model Intercomparison

581    Project (AMIP I), Bulletin of the American Meteorological Society, 80, 29-55, 10.1175/1520-

582    0477(1999)080<0029:Aootro>2.0.Co;2, 1999.

583  Gulev, S. K., Thorne, P. W., J. Ahn, F. J. D., Domingues, C. M., Gerland, S., Gong, D., Kaufman, D. S.,

584    Nnamchi, H. C., Quaas, J., Rivera, J. A., Sathyendranath, S., Smith, S. L., Trewin, B., Shuckmann,

585    K. v., and Vose, R. S.: In: Climate Change 2021: The Physical Science Basis., Climate Change 2021:

586    The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the

587    Intergovernmental Panel on Climate Change, in, edited by: [Masson-Delmotte, V., Zhai, P., Pirani,

588    A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M.,

589    Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and

590    (eds.)], B. Z., Cambridge University Press. 2021., 287–422. Cambridge University Press, 2021.

591  He, J., Hong, L., Shao, C., and Tang, W.: Global evaluation of simulated surface shortwave radiation in

592    CMIP6 models, Atmospheric Research, 292, 10.1016/j.atmosres.2023.106896, 2023.

593  He, Y., Wang, K., and Feng, F.: Improvement of ERA5 over ERA-Interim in simulating surface incident

594    solar radiation throughout China, Journal of Climate, 34, 3853-3867, 2021.

595  Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey,

596    C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold,

597    P., Biavati, G., Bidlot, J. R., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M.,

598       Dragani, R., Flemming, J., Forbes, R. G., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B.,

599           Hogan, R. J., Holm, E. V., Janisková, M., Keeley, S. P. E., Laloyaux, P., Lopez, P., Lupu, C., Radnoti,

600           G., Rosnay, P. d., Rozum, I., Vamborg, F., Villaume, S., and Thepaut, J.-N.: The ERA5 global

601           reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999 - 2049,

602           doi:10.1002/qj.3803, 2020.

603       Hoskins, B. J. and Valdes, P. J.: On the existence of storm-tracks, Journal of Atmospheric Sciences, 47,

604           1854-1864, doi:10.1175/1520-0469(1990)047<1854:OTEOST>2.0.CO;2, 1990.

605       Huang, B., Yin, X., Menne, M. J., Vose, R., and Zhang, H.-M.: Improvements to the Land Surface Air

606           Temperature Reconstruction in NOAAGlobalTemp: An Artificial Neural Network Approach,

607           Artificial Intelligence for the Earth Systems, 1-35, doi:10.1175/AIES-D-22-0032.1, 2022.

608       Huang, J., Rikus, L. J., Qin, Y., and Katzfey, J.: Assessing model performance of daily solar irradiance

609           forecasts over Australia, Solar Energy, 176, 615-626, 10.1016/j.solener.2018.10.080, 2018.

610       Jiao, B. and Li, Q.: Global Integrated and Homogenized Solar surface Radiation Datasets,

611           doi:10.6084/m9.figshare.21625079.v1,   2023.

612       Jiao, B., Li, Q., Sun, W., and Martin, W.: Uncertainties in the global and continental surface solar

613           radiation variations: inter-comparison of in-situ observations, reanalyses, and model simulations,

614           Climate Dynamics, 1-18, doi:10.1007/s00382-022-06222-3, 2022.

615       Jones, P., Osborn, T., Briffa, K., Folland, C., Horton, E., Alexander, L., Parker, D., and Rayner, N.:

616           Adjusting for sampling density in grid box land and ocean surface temperature time series, Journal

617           of Geophysical Research: Atmospheres, 106, 3371-3380, doi:10.1029/2000JD900564, 2001.

618       Ju, X., Tu, Q., and Li, Q.: Homogeneity test and reduction of monthly total solar radiation over China, J

619           Nanjing Inst Meteorol, 29, 336-341, 2006.

620       Kadow, C., Hall, D. M., and Ulbrich, U.: Artificial intelligence reconstructs missing climate information,

621           Nature Geoscience, 13, 408-413, doi:10.1038/s41561-020-0582-5, 2020.

622       Kambezidis, H. D., Kaskaoutis, D. G., Kharol, S. K., Moorthy, K. K., Satheesh, S. K., Kalapureddy, M.

623           C. R., Badarinath, K. V. S., Sharma, A. R., and Wild, M.: Multi-decadal variation of the net

624           downward shortwave radiation over south Asia: The solar dimming effect, Atmospheric

625           Environment, 50, 360-372, 2012.

626       Krige, D. G.: A statistical approach to some basic mine valuation problems on the Witwatersrand, Journal

627           of the Southern African Institute of Mining and Metallurgy, 52, 119-139,

doi:10.10520/AJA0038223X_4792, 1951.

Leirvik, T. and Yuan, M.: A machine learning technique for spatial interpolation of solar radiation observations, Earth and Space Science, 8, e2020EA001527, doi:10.1029/2020EA001527, 2021.

Li, Q., Sun, W., Yun, X., Huang, B., Dong, W., Wang, X. L., Zhai, P., and Jones, P.: An updated evaluation of the global mean land surface air temperature and surface temperature trends based on CLSAT and CMST, Climate Dynamics, 56, 635-650, doi: 10.1007/s00382-020-05502-0, 2021.

Liang, H., Jiang, B., Liang, S., Peng, J., Li, S., Han, J., Yin, X., Cheng, J., Jia, K., and Liu, Q.: A global long-term ocean surface daily/0.05° net radiation product from 1983–2020, Scientific Data, 9, 1-17, doi:10.1038/s41597-022-01419-x, 2022.

Ma, Q., Wang, K., He, Y., Su, L., Wu, Q., Liu, H., and Zhang, Y.: Homogenized century-long surface incident solar radiation over Japan, Earth System Science Data, 14, 463-477, doi:10.5194/essd-14-463-2022, 2022.

Manara, V., Bassi, M., Brunetti, M., Cagnazzi, B., and Maugeri, M.: 1990–2016 surface solar radiation variability and trend over the Piedmont region (northwest Italy), Theoretical and Applied Climatology, 136, 849-862, doi:10.1007/s00704-018-2521-6, 2019.

Manara, V., Brunetti, M., Celozzi, A., Maugeri, M., Sanchez-Lorenzo, A., and Wild, M.: Detection of dimming/brightening in Italy from homogenized all-sky and clear-sky surface solar radiation records and underlying causes (1959–2013), Atmospheric Chemistry and Physics, 16, 11145-11161, doi:10.5194/acp-16-11145-2016, 2016.

Manara, V., Stocco, E., Brunetti, M., Diolaiuti, G. A., Fugazza, D., Pfeifroth, U., Senese, A., Trentmann, J., and Maugeri, M.: Comparison of Surface Solar Irradiance from Ground Observations and Satellite Data (1990–2016) over a Complex Orography Region (Piedmont—Northwest Italy), Remote Sensing, 12, 3882, 2020.

Manara., V., Michele., B., ., M. M., ., S.-L. A., and Martin, W.: Homogenization of a surface solar radiation dataset over Italy, AIP Conference Proceedings, 22 February 2017, doi: org/10.1063/1.4975544, 2017.

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., and Clough, S. A.: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave, Journal of Geophysical Research: Atmospheres, 102, 16663-16682, https://doi.org/10.1029/97JD00237, 1997.

Olanow, C. W. and Koller, W. C.: An algorithm (decision tree) for the management of Parkinson's disease:

Treatment guidelines, Neurology, 50, S1-s88, 1998.

Padma Kumari, B., Londhe, A. L., Daniel, S., and Jadhav, D. B.: Observational evidence of solar dimming: Offsetting surface warming over India, Geophysical Research Letters, 34, https://doi.org/10.1029/2007GL031133, 2007.

Peixoto, J. P., Oort, A. H., and Lorenz, E. N.: Physics of climate, Springer1992.

Peterson, T. C., Karl, T. R., Jamason, P. F., Knight, R., and Easterling, D. R.: First difference method: Maximizing station density for the calculation of long-term global temperature change, Journal of Geophysical Research: Atmospheres, 103, 25967-25974, doi:10.1029/98JD01168, 1998.

Pfeifroth, U., Sanchez-Lorenzo, A., Manara, V., Trentmann, J., and Hollmann, R.: Trends and Variability of Surface Solar Radiation in Europe Based On Surface- and Satellite-Based Data Records, Journal of Geophysical Research: Atmospheres, 123, 1735-1754, doi: 10.1002/2017JD027418, 2018.

Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention, 234-241, doi:10.48550/arXiv.1505.04597,

Sanchez-Lorenzo, A., Calbó, J., and Wild, M.: Global and diffuse solar radiation in Spain: Building a homogeneous dataset and assessing their trends, Global and Planetary Change, 100, 343-352, doi:10.1016/j.gloplacha.2012.11.010, 2013a.

Sanchez-Lorenzo, A., Wild, M., and Trentmann, J.: Validation and stability assessment of the monthly mean CM SAF surface solar radiation dataset over Europe against a homogenized surface dataset (1983–2005), Remote sensing of environment, 134, 355-366, doi:10.1016/j.rse.2013.03.012, 2013b.

Sanchez-Lorenzo, A., Wild, M., Brunetti, M., Guijarro, J. A., Hakuba, M. Z., Calbó, J., Mystakidis, S., and Bartok, B.: Reassessment and update of long-term trends in downward surface shortwave radiation over Europe (1939–2012), Journal of Geophysical Research: Atmospheres, 120, 9555-9569, doi:10.1002/2015JD023321, 2015.

Scudiero, E., Corwin, D. L., Morari, F., Anderson, R. G., and Skaggs, T. H.: Spatial interpolation quality assessment for soil sensor transect datasets, Computers and Electronics in Agriculture, 123, 74-79, doi:10.1016/j.compag.2016.02.016, 2016.

Shao, C., Yang, K., Tang, W., He, Y., Jiang, Y., Lu, H., Fu, H., and Zheng, J.: Convolutional neural network-based homogenization for constructing a long-term global surface solar radiation dataset, Renewable and Sustainable Energy Reviews, 169, 10.1016/j.rser.2022.112952, 2022.

688 Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data, Proceedings of the

689     1968 23rd ACM national conference, 517-524,    doi:10.1145/800186.810616,

690 Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., Allan, R.,

691     Yin, X., Vose, R., and Titchner, H.: Towards a more reliable historical reanalysis: Improvements for

692     version 3 of the Twentieth Century Reanalysis system, Quarterly Journal of the Royal

693     Meteorological Society, 145, 2876-2908, doi:10.1002/qj.3598, 2019.

694 Soni, V. K., Pandithurai, G., and Pai, D. S.: Evaluation of long-term changes of solar radiation in India,

695     International Journal of Climatology, 32, 540-551, https://doi.org/10.1002/joc.2294, 2012.

696 Soni, V. K., Pandithurai, G., and Pai, D. S.: Is there a transition of solar radiation from dimming to

697     brightening over India, Atmospheric Research, 169, 209-224, 2016.

698 Tang, W., Yang, K., Qin, J., Li, X., and Niu, X.: A 16-year dataset (2000–2015) of high-resolution (3 h,

699     10 km) global surface solar radiation, Earth System Science Data, 11, 1905-1915, doi:10.5194/essd-

700     11-1905-2019, 2019.

701 Tang, W., Qin, J., Yang, K., Liu, S., Lu, N., and Niu, X.: Retrieving high-resolution surface solar radiation

702     with cloud parameters derived by combining MODIS and MTSAT data, Atmospheric Chemistry

703     and Physics, 16, 2543-2557, doi:10.5194/acp-16-2543-2016, 2016.

704 Trenberth, K. E. and Fasullo, J. T.: Regional energy and water cycles: Transports from ocean to land,

705     Journal of Climate, 26, 7837-7851, doi:10.1175/JCLI-D-13-00008.1, 2013.

706 Tsvetkov, A., Wilcox, S., Renne, D., and Pulscak, M.: International solar resource data at the World

707     Radiation Data Center, American Solar Energy Society, Boulder, CO (United States), 1995.

708 Urraca, R., Huld, T., Martinez-de-Pison, F. J., and Sanz-Garcia, A.: Sources of uncertainty in annual

709     global horizontal irradiance data, Solar Energy, 170, 873-884, 10.1016/j.solener.2018.06.005, 2018.

710 Vincent, L. A., Wang, X. L., Milewska, E. J., Wan, H., Yang, F., and Swail, V.: A second generation of

711     homogenized Canadian monthly surface air temperature for climate trend analysis, Journal of

712     Geophysical Research: Atmospheres, 117, doi:10.1029/2012JD017859, 2012.

713 Wang, K.: Measurement biases explain discrepancies between the observed and simulated decadal

714     variability of surface incident solar radiation, Scientific reports, 4, 1-7, doi:0.1038/srep06144 2014.

715 Wang, K., Ma, Q., Li, Z., and Wang, J.: Decadal variability of surface incident solar radiation over China:

716     Observations, satellite retrievals, and reanalyses, Journal of Geophysical Research: Atmospheres,

717     120, 6500-6514, doi:10.1002/2015JD023420, 2015.

Wang, X. L.: Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal t or F test, Journal of applied meteorology and climatology, 47, 2423-2444, doi:10.1175/2008JAMC1741.1, 2008a.

Wang, X. L.: Penalized maximal F test for detecting undocumented mean shift without trend change, Journal of Atmospheric and Oceanic Technology, 25, 368-384, doi:10.1175/2007JTECHA982.1, 2008b.

Wang, X. L. and Feng, Y.: RHtestsV4 user manual, Climate Research Division, Atmospheric Science and Technology Directorate, Science and Technology Branch, Environment Canada, 28, 2013.

Wang, X. L., Wen, Q. H., and Wu, Y.: Penalized maximal t test for detecting undocumented mean change in climate data series, Journal of Applied Meteorology and Climatology, 46, 916-931, doi:10.1175/JAM2504.1, 2007.

Wang, X. L., Chen, H., Wu, Y., Feng, Y., and Pu, Q.: New techniques for the detection and adjustment of shifts in daily precipitation data series, Journal of Applied Meteorology and Climatology, 49, 2416-2436, doi:10.1175/2010JAMC2376.1, 2010.

Wang, Y. and Wild, M.: A new look at solar dimming and brightening in China, Geophysical Research Letters, 43, 11,777-711,785, doi:10.1002/2016GL071009, 2016.

Wild, M.: Enlightening global dimming and brightening, Bulletin of the American Meteorological Society, 93, 27-37, doi:10.1175/BAMS-D-11-00074.1, 2012.

Wild, M.: The global energy balance as represented in CMIP6 climate models, Clim Dyn, 55, 553-577, 10.1007/s00382-020-05282-7, 2020.

Wild, M., Trüssel, B., Ohmura, A., Long, C. N., König-Langlo, G., Dutton, E. G., and Tsvetkov, A.: Global dimming and brightening: An update beyond 2000, Journal of Geophysical Research: Atmospheres, 114, https://doi.org/10.1029/2008JD011382, 2009.

Wild, M., Ohmura, A., Schär, C., Müller, G., Folini, D., Schwarz, M., Hakuba, M. Z., and Sanchez-Lorenzo, A.: The Global Energy Balance Archive (GEBA) version 2017: A database for worldwide measured surface energy fluxes, Earth System Science Data, 9, 601-613, doi:10.5194/essd-9-601-2017, 2017.

Wild, M., Gilgen, H., Roesch, A., Ohmura, A., Long, C. N., Dutton, E. G., Forgan, B., Kallis, A., Russak, V., and Tsvetkov, A.: From dimming to brightening: Decadal changes in solar radiation at Earth's surface, Science, 308, 847-850, doi:10.1126/science.1103215, 2005.

Xu, W., Li, Q., Wang, X. L., Yang, S., Cao, L., and Feng, Y.: Homogenization of Chinese daily surface air temperatures and analysis of trends in the extreme temperature indices, Journal of Geophysical Research: Atmospheres, 118, 9708-9720, doi:10.1002/jgrd.50791, 2013.

Xu, W., Li, Q., Jones, P., Wang, X. L., Trewin, B., Yang, S., Zhu, C., Zhai, P., Wang, J., and Vincent, L.: A new integrated and homogenized global monthly land surface air temperature dataset for the period since 1900, Climate Dynamics, 50, 2513-2536, doi:10.1007/s00382-017-3755-1, 2018.

Yang, L., Zhang, X., Liang, S., Yao, Y., Jia, K., and Jia, A.: Estimating surface downward shortwave radiation over China based on the gradient boosting decision tree method, Remote Sensing, 10, 185, doi:10.3390/rs10020185, 2018a.

Yang, S.: Chinese Monthly homogenized  surface solar radiation datasets (V 1.0), 2016.

Yang, S., Wang, X. L., and Wild, M.: Homogenization and trend analysis of the 1958–2016 in situ surface solar radiation records in China, Journal of Climate, 31, 4529-4541, doi:10.1175/JCLI-D-17-0891.1, 2018b.

You, Q., Sanchez-Lorenzo, A., Wild, M., Folini, D., Fraedrich, K., Ren, G., and Kang, S.: Decadal variation of surface solar radiation in the Tibetan Plateau from observations, reanalysis and model simulations, Climate dynamics, 40, 2073-2086, doi:10.1007/s00382-012-1383-3, 2013.

Yuan, M., Leirvik, T., and Wild, M.: Global trends in downward surface solar radiation from spatial interpolated ground observations during 1961–2019, Journal of Climate, 34, 9501-9521, doi:10.1175/JCLI-D-21-0165.1, 2021.

Zhou, C., He, Y., and Wang, K.: On the suitability of current atmospheric reanalyses for regional warming studies over China, Atmospheric Chemistry and Physics, 18, 8113-8136, 10.5194/acp-18-8113-2018, 2018a.

Zhou, C., Wang, K., and Ma, Q.: Evaluation of Eight Current Reanalyses in Simulating Land Surface Temperature from 1979 to 2003 in China, Journal of Climate, 30, 7379-7398, https://doi.org/10.1175/JCLI-D-16-0903.1, 2017.

Zhou, W., Gong, L., Wu, Q., Xing, C., Wei, B., Chen, T., Zhou, Y., Yin, S., Jiang, B., Xie, H., Zhou, L., and Zheng, S.: Correction to: PHF8 upregulation contributes to autophagic degradation of E-cadherin, epithelial-mesenchymal transition and metastasis in hepatocellular carcinoma, Journal of Experimental & Clinical Cancer Research, 37, 10.1186/s13046-018-0944-7, 2018b.

Zhou, W., Gong, L., Wu, Q., Xing, C., Wei, B., Chen, T., Zhou, Y., Yin, S., Jiang, B., Xie, H., Zhou, L.,

778    and Zheng, S.: Correction to: PHF8 upregulation contributes to autophagic degradation of E-

779    cadherin, epithelial-mesenchymal transition and metastasis in hepatocellular carcinoma, J Exp Clin

780    Cancer Res, 38, 445, 10.1186/s13046-019-1452-0, 2019.

781

**Captions of tables and Figures**

Table 1: List of information on the various types of data used in this paper.

Figure 1: Flowchart of quality control (QC) (first step), homogenization (second step) and integration (third step).

Figure 2: Spatial distribution of candidate stations ("*") and added stations ("+"). The different colour bars represent the length of the station record in months (Units: Month).

Figure 3: Spatial distribution of stations after homogenization (Units: Month), different colours represent the length of station records in months

Figure 4: Flowchart of AI reconstruction.

Figure 5: (a) Spatial distribution of 5°x5°grid boxes (SSRIH$_{grid}$) obtained interpolating the homogenized global land (except for Antarctica) SSR series. The different colours represent the length (the sum of all records) of the station record, Units: Year. (b) Grid box coverage for the homogenized global land (except for Antarctica) SSR (SSRIH$_{grid}$) except for Antarctica.

Figure 6: Reconstruction capabilities of the AI model.

Figure 7: Spatial distribution of the SSRIH$_{grid}$ (a1-3) and SSRIH$_{20CR}$ (b1-3) in typical months. 1-3 is July 1960, July 1980, and July 2000, respectively.

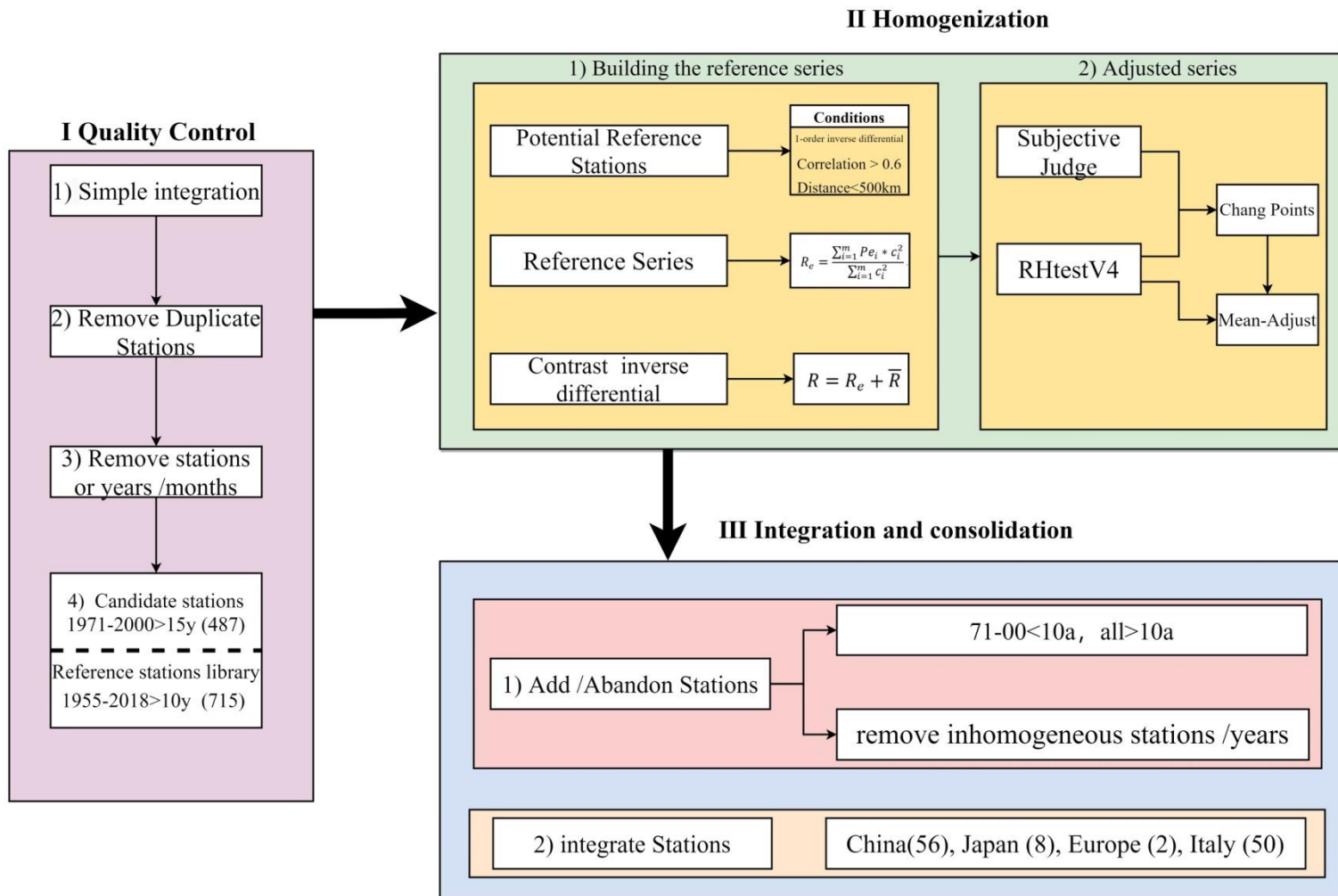
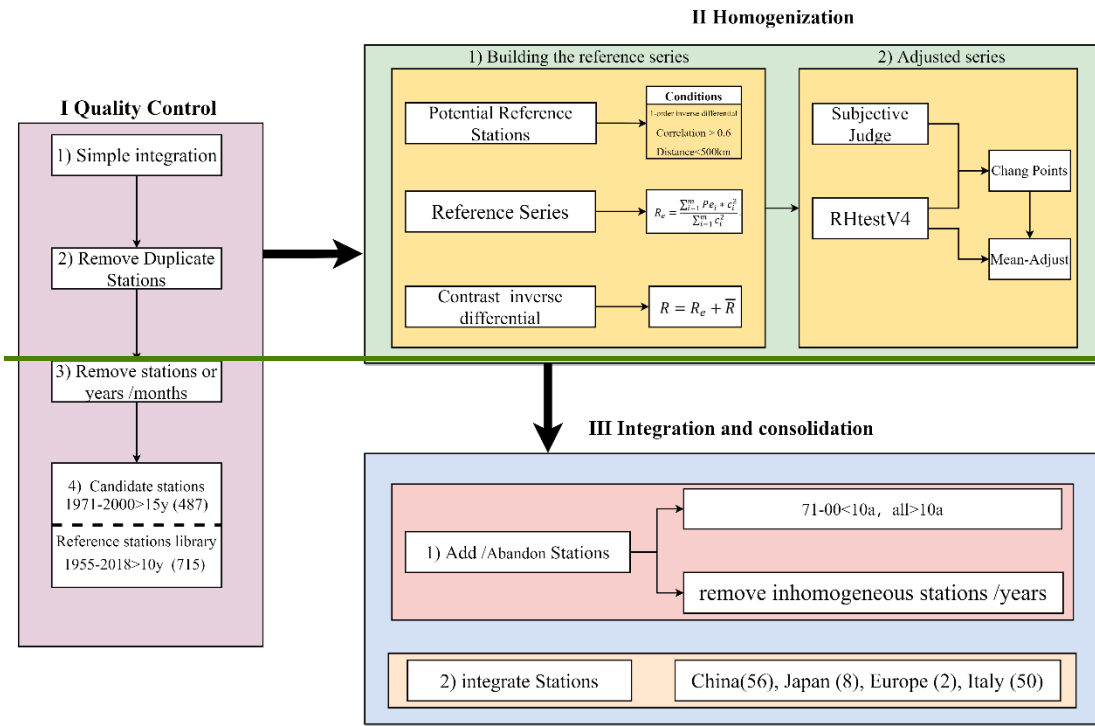
Figure 8: Global land (except for Antarctica) time series of the annual anomaly variations SSR (relative to 1971-2000) before/after reconstruction.

Figure 9: Same as Figure 8, but for regional annual anomaly variations.

812 **Table 1: List of information on the various types of data used in this paper**

|  | Abbreviation | Resolution | Time | Reference |
|---|---|---|---|---|
| *In-situ*-Raw | GEBA (Station) | Monthly | 1922-2020 | (Wild et al., 2017) |
|  | WRDC (Station) | Monthly | 1964-2017 | (Tsvetkov et al., 1995) |
| *In-situ*-Homo | China (Station) | Monthly | 1950-2016 | (Yang et al., 2018b) |
|  | Japan (Station) | Monthly | 1870-2015 | (Ma et al., 2022) |
|  | Europe (Station) | Monthly | 1922-2012 | (Sanchez-Lorenzo et al., 2015) |
|  | Italy (Station) | Monthly | 1959-2016 | (Manara et al., 2016; Manara et al., 2019) |
| Reanalysis / Model | ERA5 (Grid) | Monthly/ 0.25°×0.25° | 1950-2020 | (Hersbach et al., 2020) |
|  | 20CRv3 (Grid) | Monthly/ 0.7°×0.7° | 1940-2015 | (Slivinski et al., 2019) |
|  | CMIP6 (Grid) | Monthly/- | 1940-2014 | (Eyring et al., 2016) |

813

814

815

816    **Figure 1: Flowchart of quality control (QC) (first step), homogenization (second step) and integration (third step).**

817

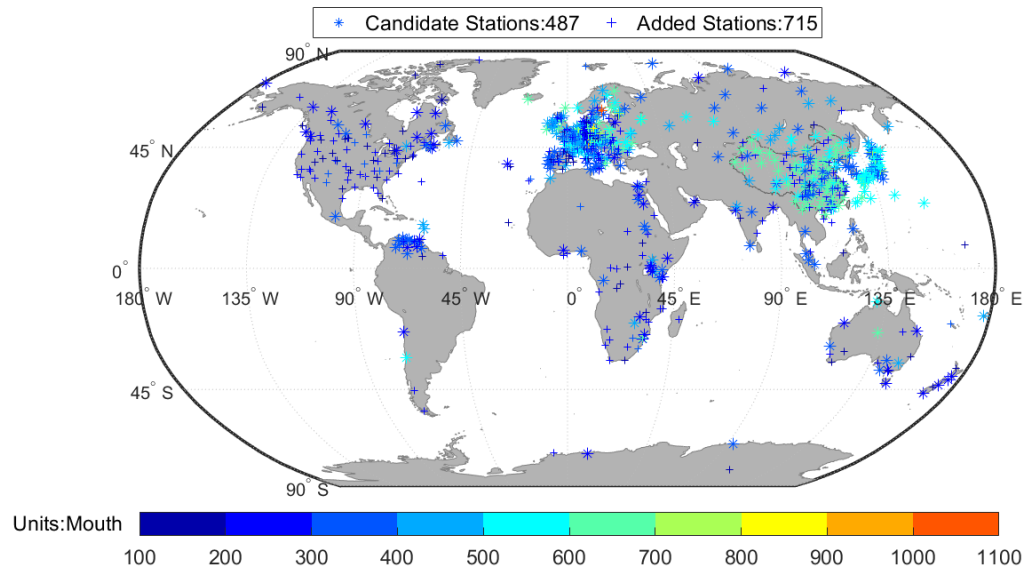
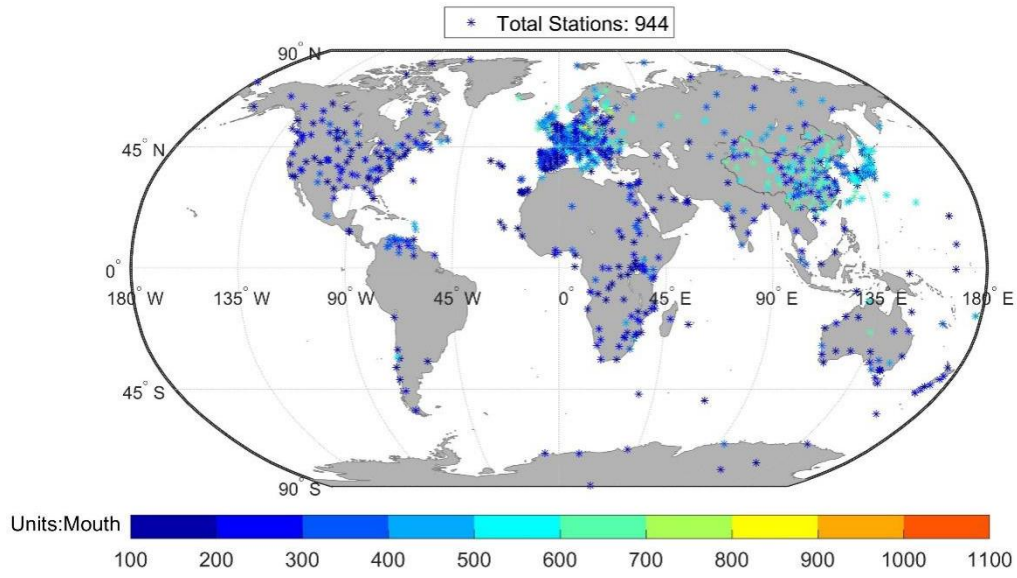
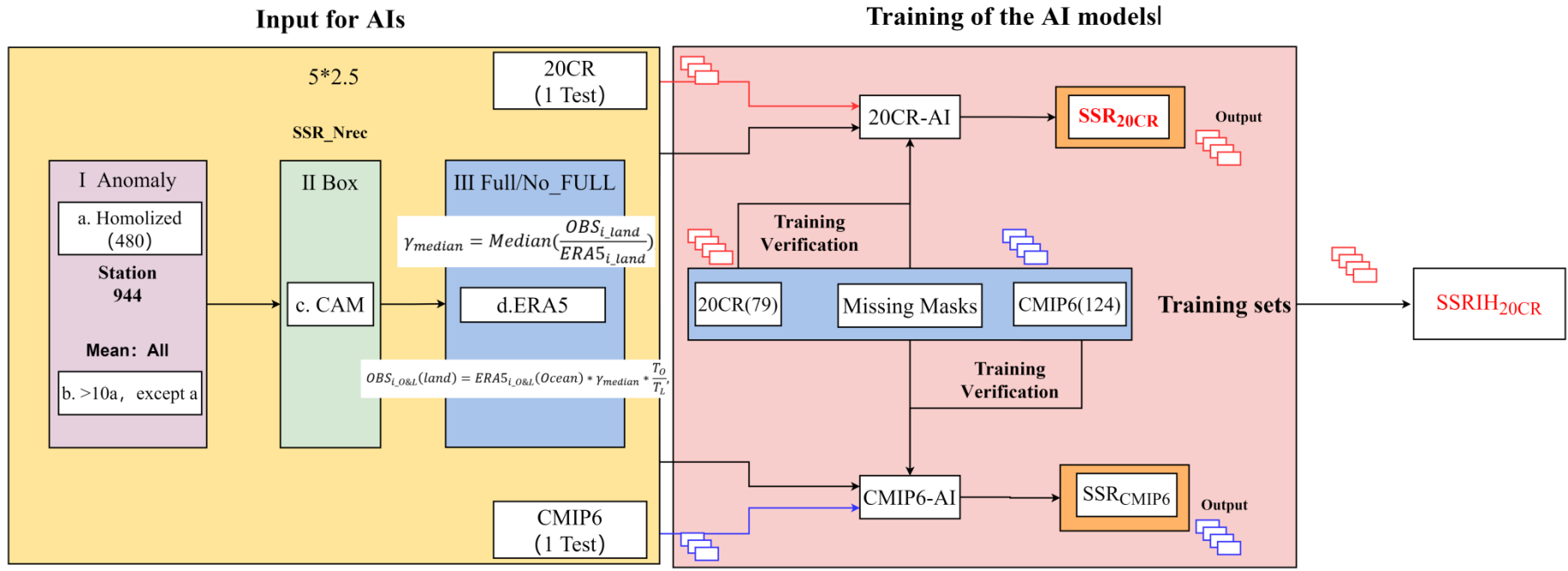
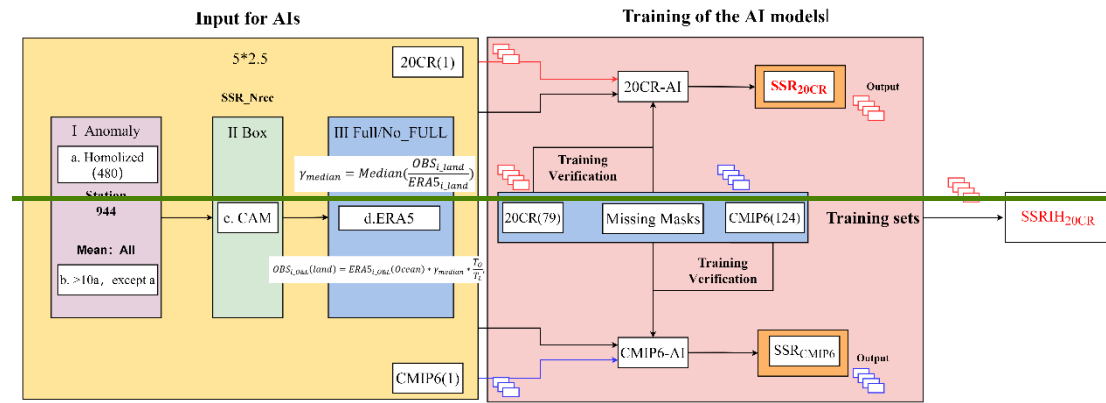
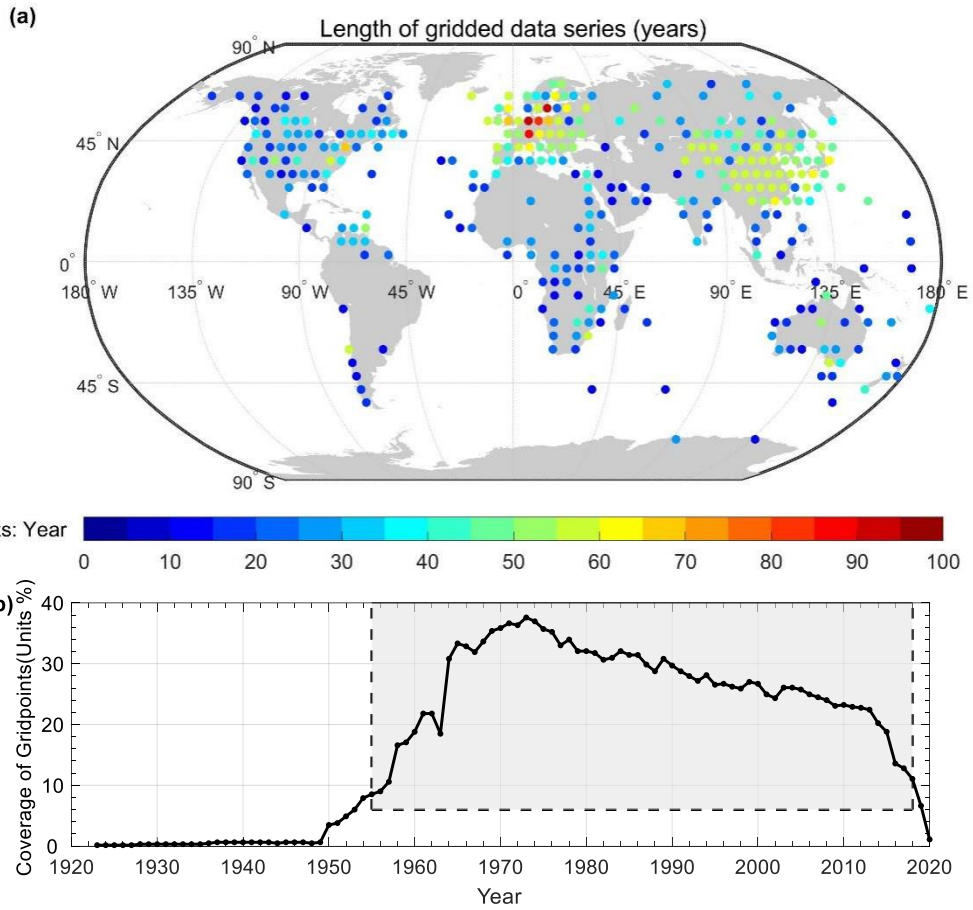



Figure 2: Spatial distribution of candidate stations ("*") and added stations ("+"). The different colour bars
represent the length of the station record in months (Units: Month).

820
821 **Figure 3: Spatial distribution of stations after homogenization (Units: Month), different colours represent the**

822 **length of station records in months.**

824

825 **Figure 4: Flowchart of AI reconstruction.**

826

827
828 **Figure 5: (a) Spatial distribution of 5°×5°grid boxes (SSRIH$_{grid}$) obtained interpolating the homogenized**
829 **global land (except for Antarctica) SSR series. The different colours represent the length (the sum of all**
830 **records) of the station record, Units: Year. (b) Grid box coverage for the homogenized global land (except**
831 **for Antarctica) SSR (SSRIH$_{grid}$) except for Antarctica.**

832

833

834

835

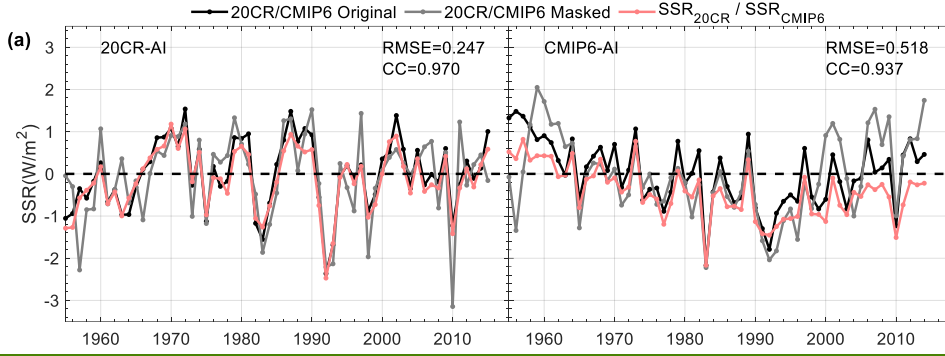
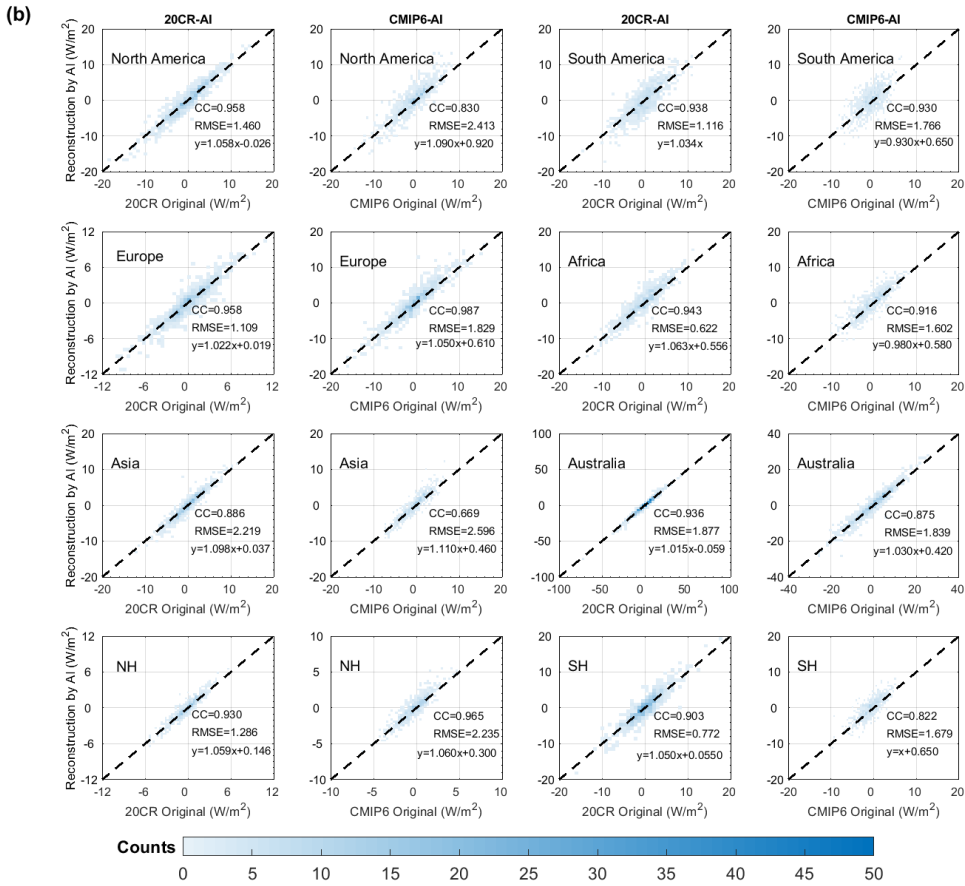
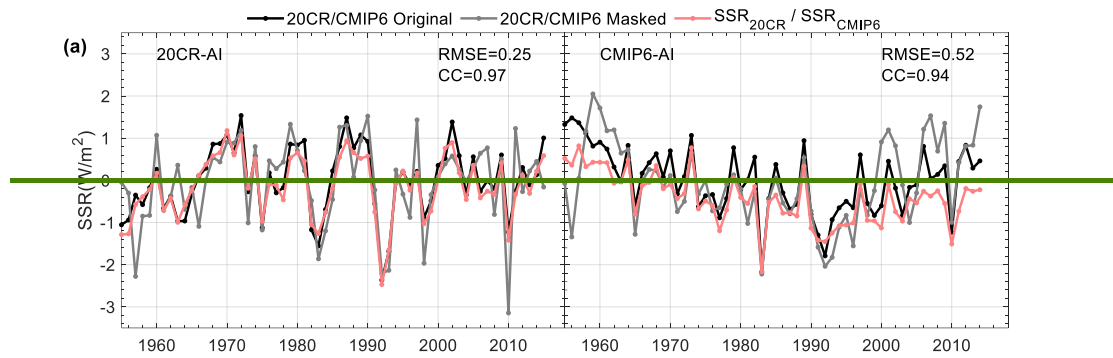
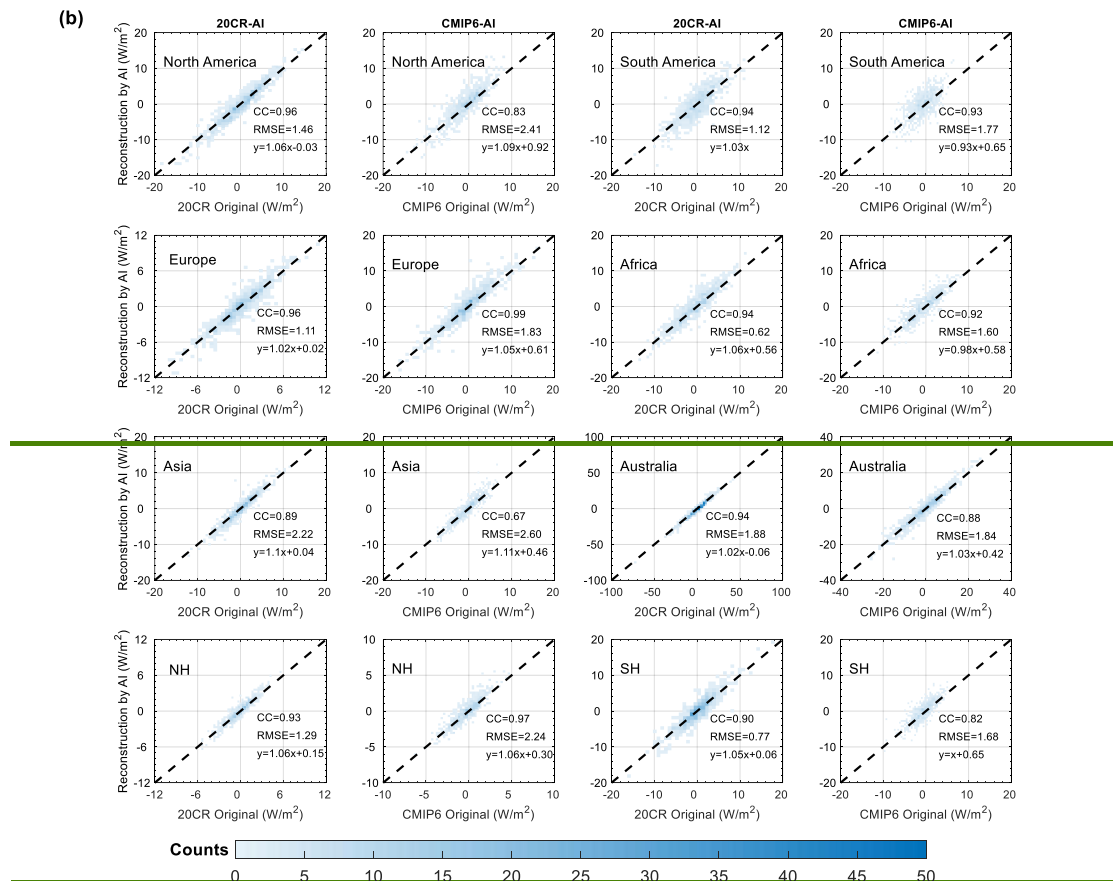
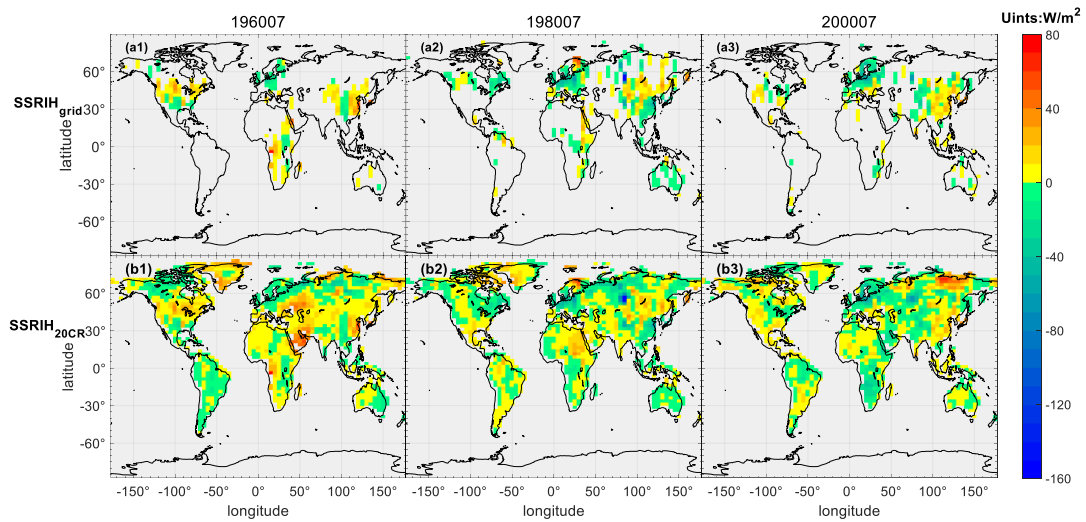
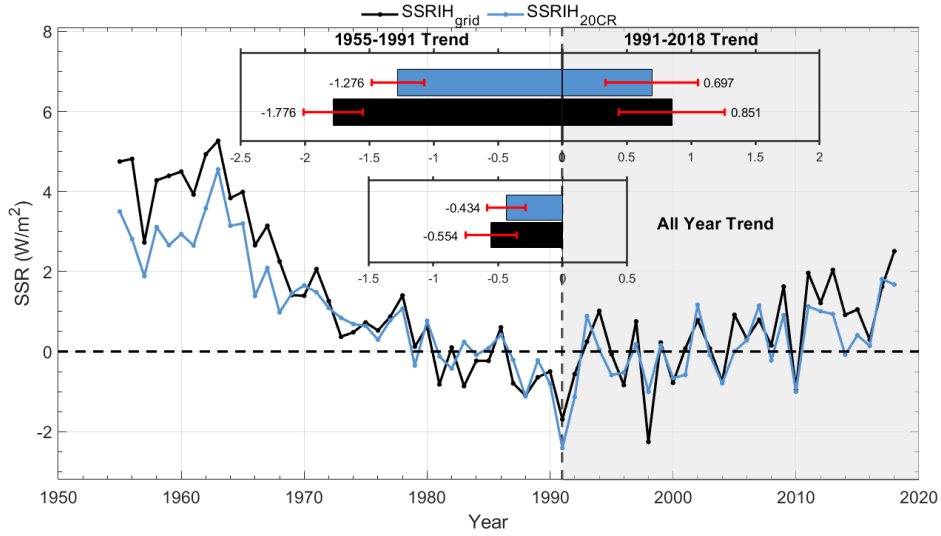



Figure 6: Reconstruction capabilities of the AI model. (a) Global land (except for Antarctica) means time-series analysis and AI model reconstruction evaluation. The red line is the SSR of the reconstruction based on the 20CR-AI /CMIP6-AI model ($SSR_{20CR}$ /$SSR_{CMIP6}$); The grey line is the masked datasets with missing values of the $SSRIH_{grid}$. The solid black line is the 20CR and CMIP6 validation set (the SSR from the 1th member of 20CRv3 /CMIP6). (b) Comparisons of the $SSR_{20CR}$ (columns 1, 3) /$SSR_{CMIP6}$ (columns 2, 4) with the SSR from the 20CR and CMIP6 validation set. Colour bars represent counts with the same values for both. Figures also show the $SSR_{20CR}$ ($SSR_{CMIP6}$) correlation coefficient (CC), root mean squared error (RMSE) and fitting equation compared to the original dataset in different regions.

844

845 **Figure 7: Spatial distribution of the SSRIH_grid (a1-3) and SSRIH_20CR (b1-3) in typical months. 1-3 is July**

846 **1960, July 1980, and July 2000, respectively.**

847

848 **Figure 8: Global land (except for Antarctica) annual SSR anomaly variations (relative to 1971-2000)**

849 **before/after reconstruction. The Black solid line represents the SSRIH$_{grid}$ annual anomalies. The solid blue**

850 **line represents the SSRIH$_{20CR}$ annual anomalies. The histograms represent the decadal trends of the**

851 **SSRIH$_{grid}$ /SSRIH$_{20CR}$ (unit: W/m$^2$ per decade) and their 95% uncertainty range from 1955 to 1991, 1991-**

852 **2018 and 1955-2018.**

853

854

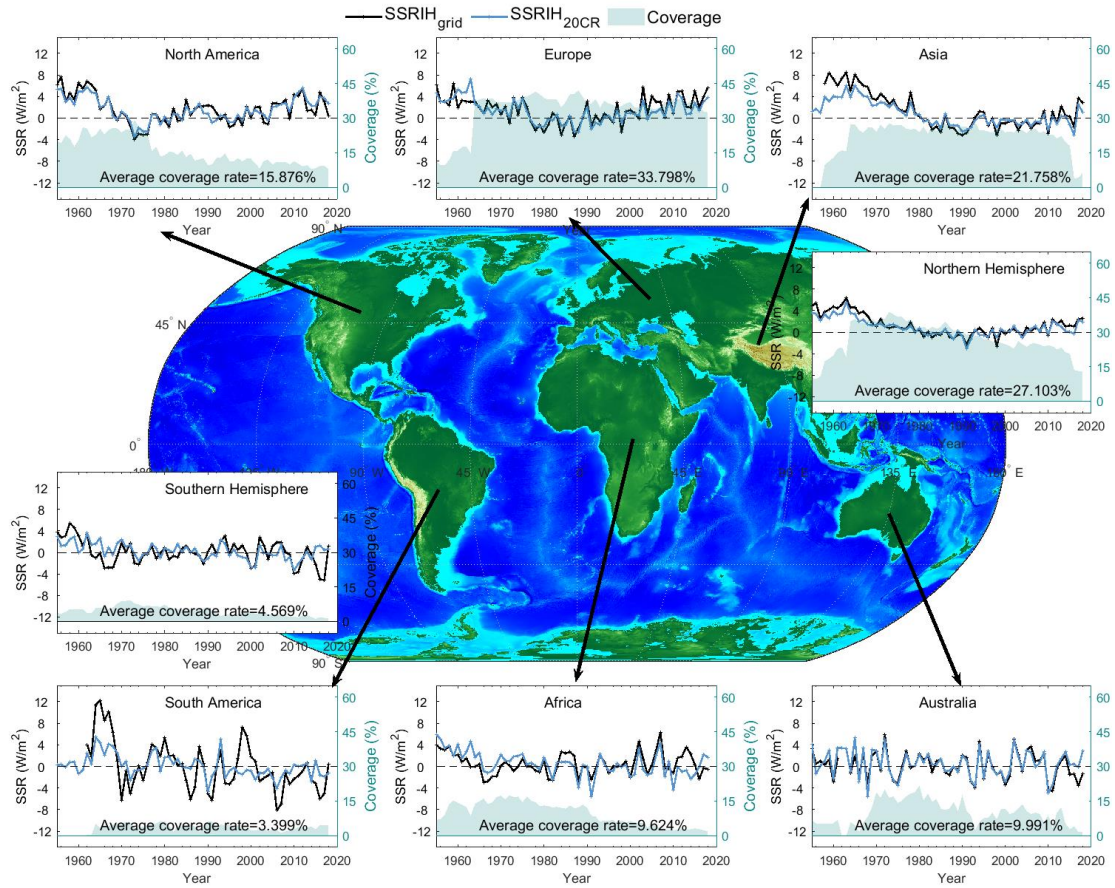
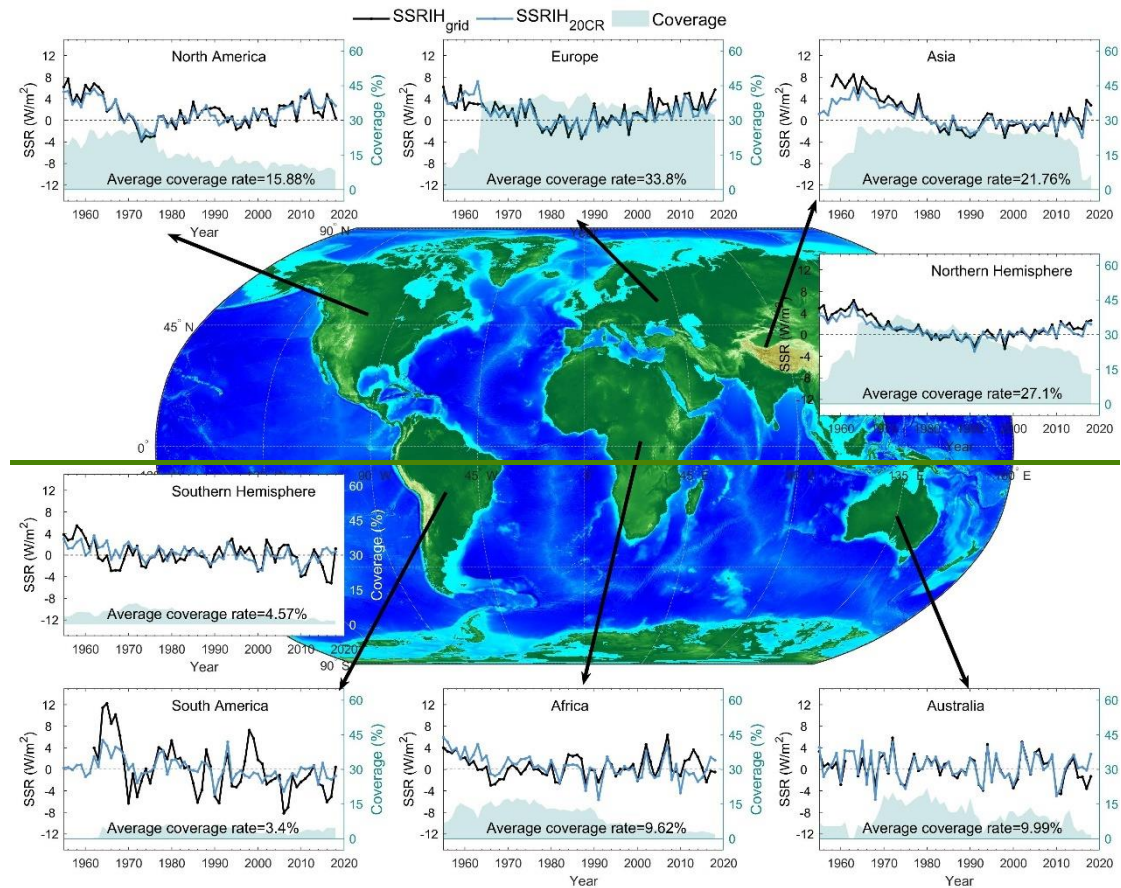



Figure 9: Same as Figure 8, but for regional annual anomaly variations. The green colour filling diagram
represents the variation in grid box coverage (before reconstruction).