Supplement of Earth Syst. Sci. Data, 15, 4519–4535, 2023
https://doi.org/10.5194/essd-15-4519-2023-supplement

Supplement of

An integrated and homogenized global surface solar radiation dataset and its reconstruction based on a convolutional neural network approach

Boyang Jiao et al.

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

Text S1 Convolutional Neural Network (CNN) deep learning model (convolutional layer, loss function)

Convolutional layer using partial convolution and mask update: The partial convolution operation and the mask update function are called the partial convolution layer (Liu et al., 2018). The partial convolution operation and the mask update function are called the partial convolution layer. The partial convolution at each position can be expressed as

$$x' = \begin{cases} W^T \left(X \odot M \right) \dfrac{sum(1)}{sum(M)} + b, & if\ sum(M) > 0 \\ 0, & otherwise \end{cases} \tag{S1}$$

$\odot$ denotes element-by-element multiplication, where 1 and M in the above equation have the same shape, and all elements in 1 are 1. Eq. (1) illustrates that our output value depends only on the valid input and that $\frac{sum(1)}{sum(M)}$ is used to adjust the amount of change in the valid value of the input.

$$m' = \begin{cases} 1, & if\ sum(M) > 0 \\ 0, & otherwise \end{cases} \tag{S2}$$

After each partial convolution operation, use equation (2) to update the mask Eq. (2) indicates that we mark that position as valid whenever the convolution can adjust its output according to at least one valid value. In other words, marking 1 where there is a value and 0 for the default part is the so-called binary mask. This approach can be implemented in any deep learning structure as part of a forward delivery. With enough partial convolutions, the input values will all eventually become valid, i.e., any masks will all become 1. Partial convolution layers can be implemented by extending the existing standard Pytorch library. The most straightforward implementation is to define a binary mask of the shape $\mathrm{C} \times \mathrm{H} \times \mathrm{W}$ that is the same size as its associated image and feature values. And then, update the mask using a fixed convolutional layer of the same size and operation as the partial convolutional layer, with the same weight (weight of 1) and no bias.

The model loss function is set for each pixel reconstruction accuracy and the transition smoothness of the repaired missing measurements to their surroundings. Let the input image be I_i, the initial binary mask be M, the predicted value be I_{out}, and the actual value be I_{gt}. Eq. (3) and Eq. (4) calculate the loss value for each pixel, where Eq. (3) calculates the default value portion of the loss value and Eq. (4) calculates the actual value portion of the loss value.

$$\mathcal{L}_{hole} = ||(1 - M) \odot \left(I_{out} - I_{gt} \right)||_1 \tag{S3}$$

$$\mathcal{L}_{valid} = ||M \odot \left(I_{out} - I_{gt}\right)||_1 \tag{S4}$$

Define the Perceptual Loss function (Eq. (5)) and the Style Loss function (Eq. (6) and (7). Where I_{comp} denotes the original data, where the valid value is the true value and K_n denotes the normalization factor.

$$\mathcal{L}_{perceptual} = \sum_{n=0}^{N-1} ||\Psi_n(I_{out}) - \Psi_n\left(I_{gt}\right)||_1 + \sum_{n=0}^{N-1} ||\Psi_n\left(I_{comp}\right) - \Psi_n\left(I_{gt}\right)||_1 \tag{S5}$$

$$\mathcal{L}_{style_{out}} = \sum_{n=0}^{N-1} ||K_n((\Psi_n(I_{out}))^T(\Psi_n(I_{out})) - (\Psi_n\left(I_{gt}\right))^T\left(\Psi_n\left(I_{gt}\right)\right))||_1 \tag{S6}$$

$$\mathcal{L}_{style_{comp}} = \sum_{n=0}^{N-1} ||K_n((\Psi_n(I_{comp}))^T(\Psi_n\left(I_{comp}\right) - (\Psi_n\left(I_{gt}\right))^T\left(\Psi_n\left(I_{gt}\right)\right))||_1 \tag{S7}$$

Finally, the Total Variation Loss function is defined in equation (8). This loss function effectively smoothes the image, reducing the total variation of the signal and removing unwanted details while retaining essential details such as edges.

$$\mathcal{L}_{tv} = \sum_{(i,j)\in P,(i,j+1)\in P} \left|\left|I_{comp}^{i,j+1} - I_{comp}^{i,j}\right|\right|_1 + \sum_{(i,j)\in P,(i+1,j)\in P} \left|\left|I_{comp}^{i+1,j} - I_{comp}^{i,j}\right|\right|_1 \tag{S8}$$

First, we set the batch size to 16 in the first 500000 iterations and fine-tuned it to 18 in the last 10000000 iterations, for a total of 1500000 iterations, to suppress the overfitting phenomenon generated during the training process, and validate the model every 10000 times and early stopping if the validation shows a decreasing trend, the final number of training times used is 1100000. Second, L2 regularization is also added to regulate the loss function. The initial hyper-parameters of the model are set as follows; learning rate of 2e-4 and learning finetune of 5e-5.

The final loss function equation (9) is constructed by combining all the loss functions necessary for image restoration, and a validation set of 100 images confirms this equation's hyperparameters.

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{valid} + 6\mathcal{L}_{hole} + 0.05\mathcal{L}_{perceptual} + 120\left(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}\right) \\ + 0.1\mathcal{L}_{tv} + \alpha\|\omega\|_2{}^2\end{aligned} \tag{S9}$$

Table S1: CMIP6 numerical models for training the neural network. CMIP6 Historical monthly experiments between 1955 and 2014 are applied to train the CMIP6-AI.

	Source ID	N°	Ensemble
1	ACCESS-ESM1-5	40	r1i1p1f1-r40i1p1f1
2	CNRM-CM6-1	30	r1i1p1f2-r30i1p1f2
3	CNRM-ESM2-1	11	r1i1p1f2-r11i1p1f2
4	EC-Earth3	22	r1i1p1f1-r4i1p1f1; r6i1p1f1; r7i1p1f1; r9i1p1f1; r10i1p1f1-r19i1p1f1; r21i1p1f1-r25i1p1f1
5	EC-Earth3-CC	10	r1i1p1f1; r4i1p1f1; r6i1p1f1-r13i1p1f1
6	MRI-ESM2-0	12	r1i1p1f1-r10i1p1f1; r1i2p1f1; r1i1000p1f1

Table S3 Trends and their 95% confidence ranges in various data sources global SSR change (units: W/m^2 per decade). * Indicate trends that are significant at the 5% level.

Type	1955-1991	1991-2018	1955-2018
$SSRI_{grid}$	-1.995 ± 0.251*	0.999 ± 0.504*	-0.494 ± 0.228*
$SSRIH_{grid}$	-1.776 ± 0.230*	0.851 ± 0.410*	-0.554 ± 0.197*
$SSRIH_{20CR}$	-1.276 ± 0.205*	0.697 ± 0.359*	-0.434 ± 0.148*
ERA5	-1.162 ± 0.319*	0.653 ± 0.350*	-0.180 ± 0.176*

Table S4 Trends and their 95% confidence ranges in continental and hemispheric $SSRIH_{20CR}$ change (Units: W/m^2 per decade). * Indicate trends that are significant at the 5% level.

Continental	Time period /Trend	Time period /Trend
North America	1955-1973	1973-2018
	-3.588 ± 1.290*	1.074 ± 0.278*
South America	1955-1990	1990-2018
	-0.408 ± 0.619	0.049 ± 0.768
Europe	1963-1978	1978-2018
	-2.180 ± 1.866*	1.081 ± 0.312*
Africa	1955-1991	1991-2018
	-1.506 ± 0.496*	0.340 ± 0.998
Asia	1955-1990	1990-2018
	-1.633 ± 0.473*	0.435 ± 0.505
North Hemisphere	1955-1991	1991-2018
	-1.457 ± 0.246*	0.887 ± 0.415*
South Hemisphere	1955-1991	1991-2018
	-0.708 ± 0.330*	-0.076 ± 0.656*

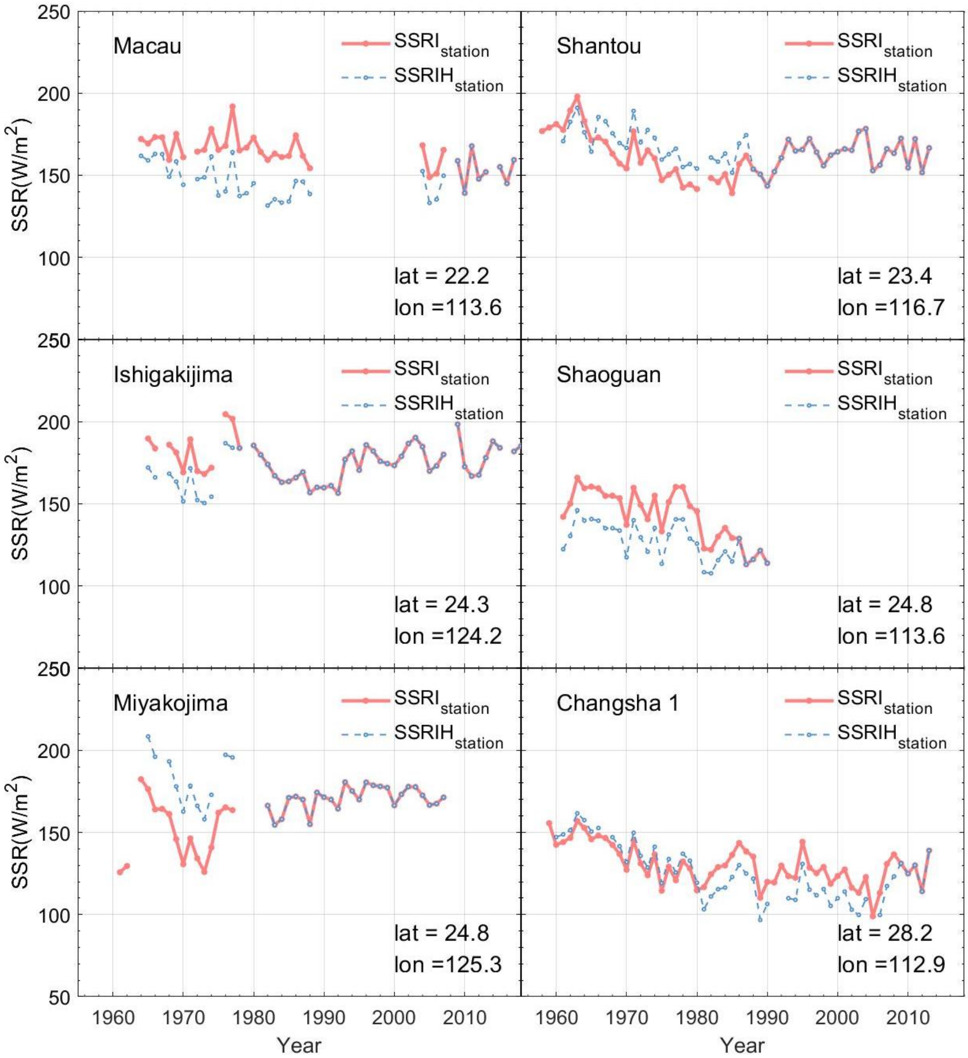


Figure S1-1 Annual variation of SSR calculated from the original station SSR series (SSRI$_{station}$, blue line), the station SSR series after homogenization (SSRIH$_{station}$, red line).

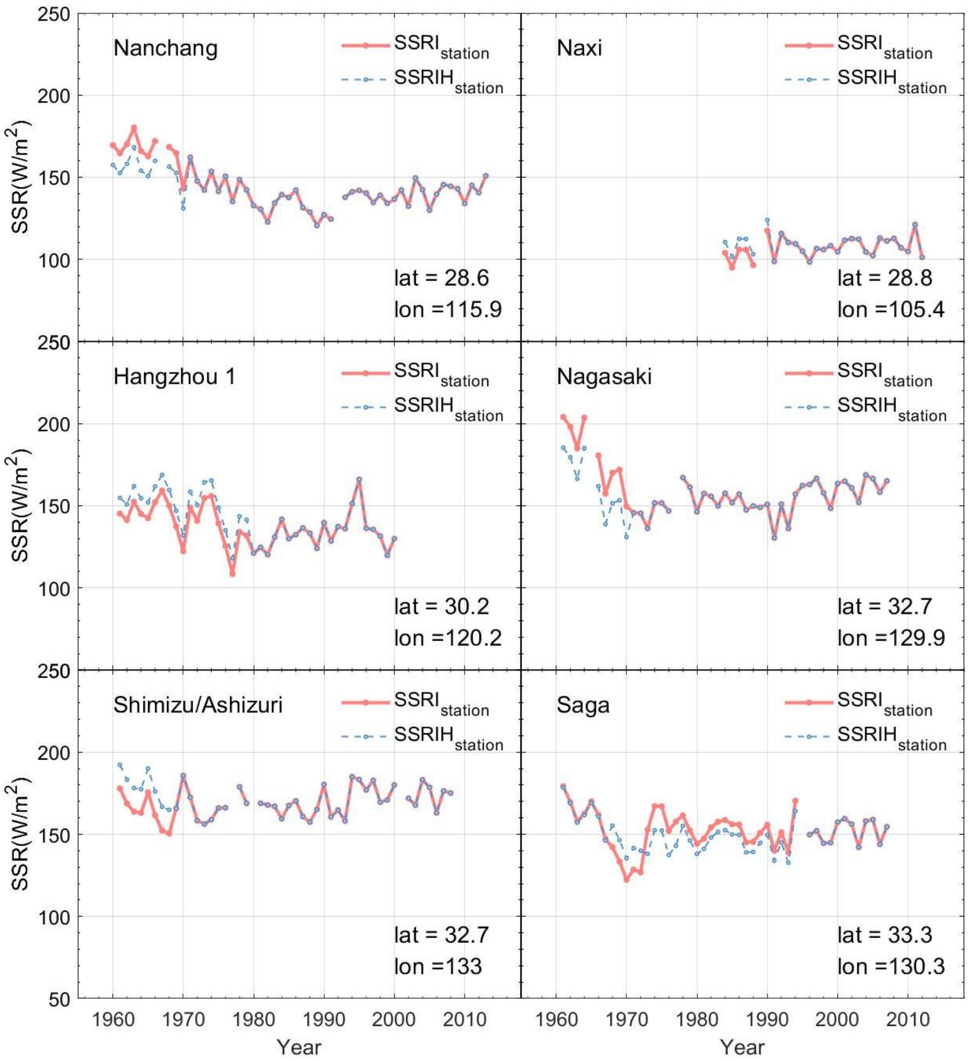


Figure S1-2 Annual variation of SSR calculated from the original station SSR series ($SSRI_{station}$, blue line), the station SSR series after homogenization ($SSRIH_{station}$, red line).

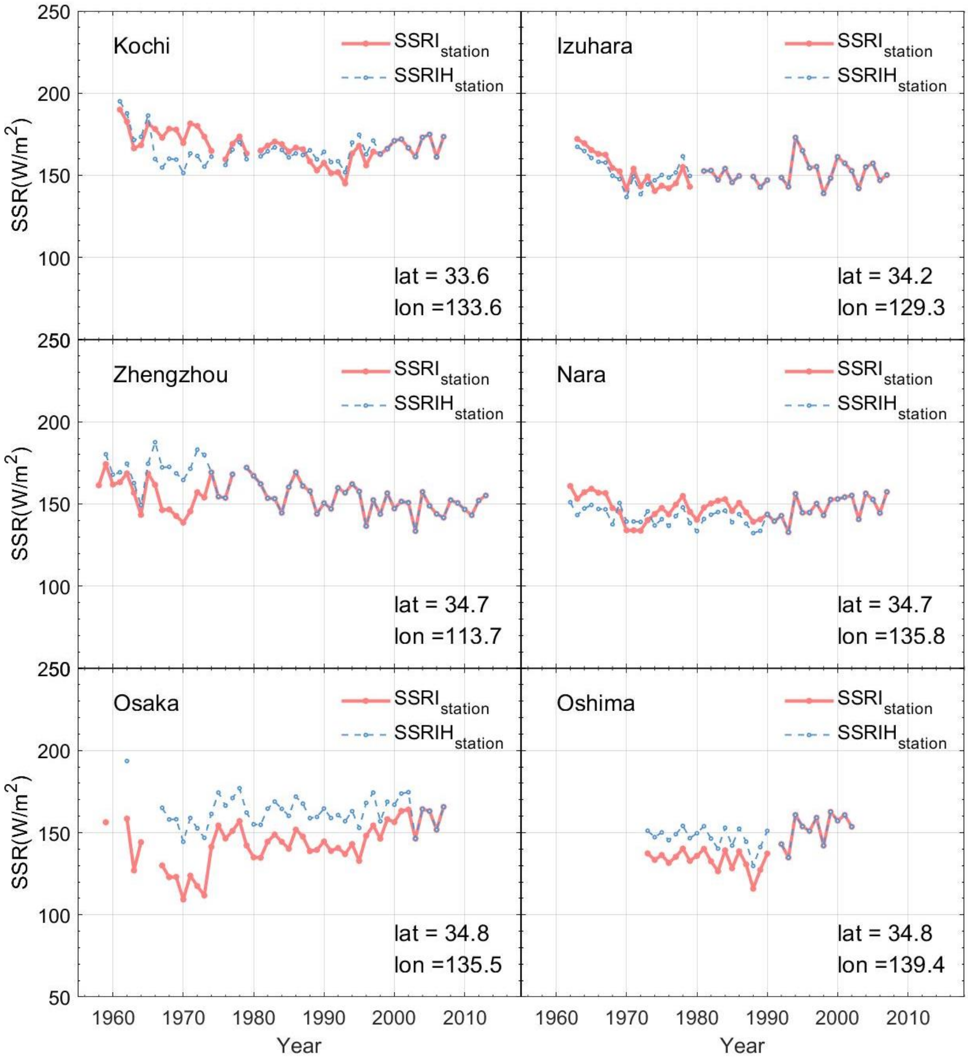
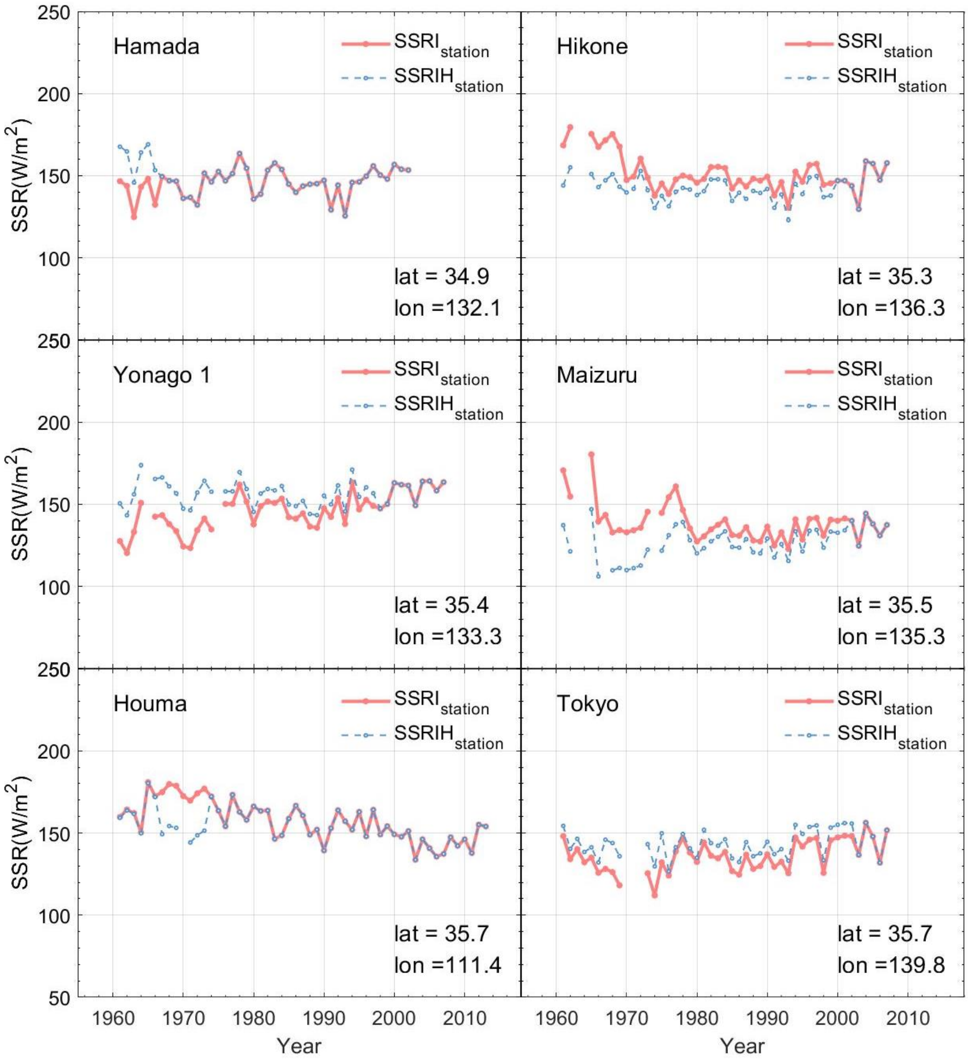
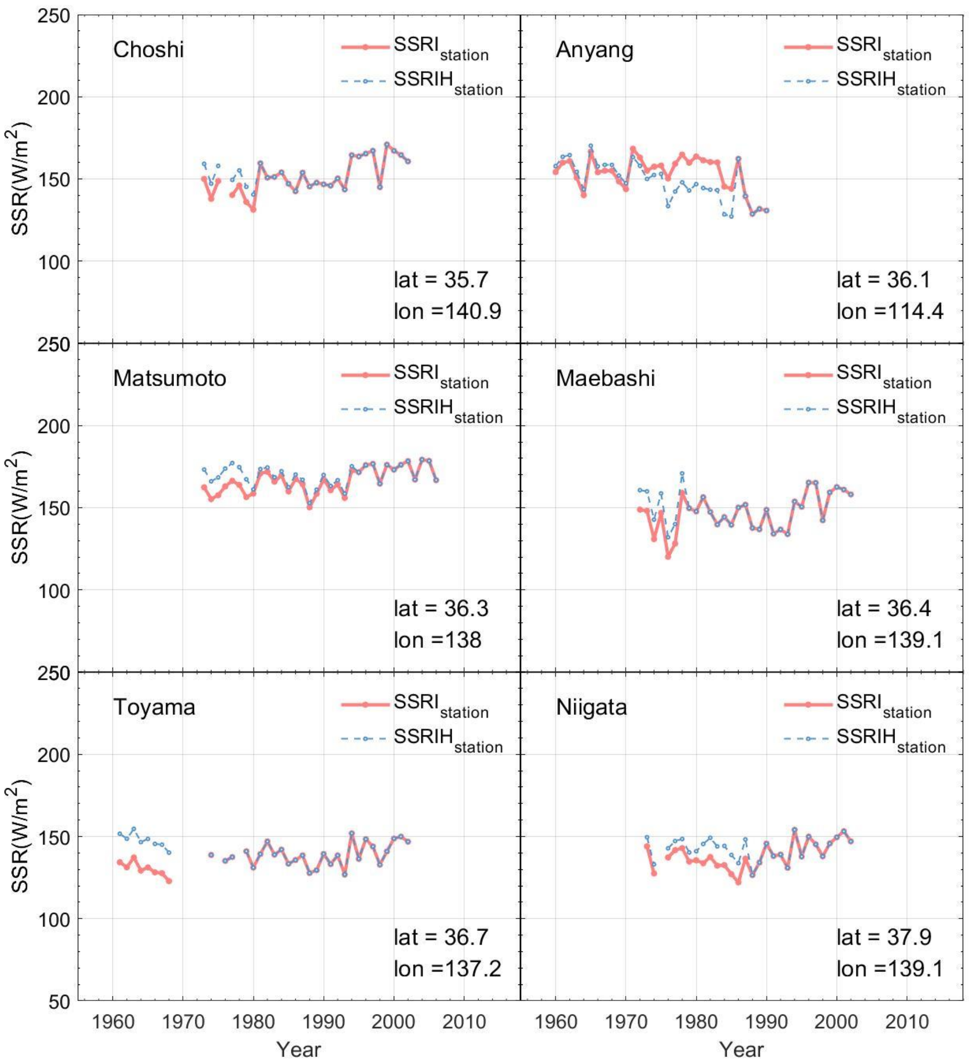


Figure S1-3 Annual variation of SSR calculated from the original station SSR series ($SSRI_{station}$, blue line), the station SSR series after homogenization ($SSRIH_{station}$, red line).



79
80 **Figure S1-4 Annual variation of SSR calculated from the original station SSR series (SSRI$_{station}$, blue line),**
81 **the station SSR series after homogenization (SSRIH$_{station}$, red line).**



82
83 **Figure S1-5 Annual variation of SSR calculated from the original station SSR series ($SSRI_{station}$, blue line),**
84 **the station SSR series after homogenization ($SSRIH_{station}$, red line).**

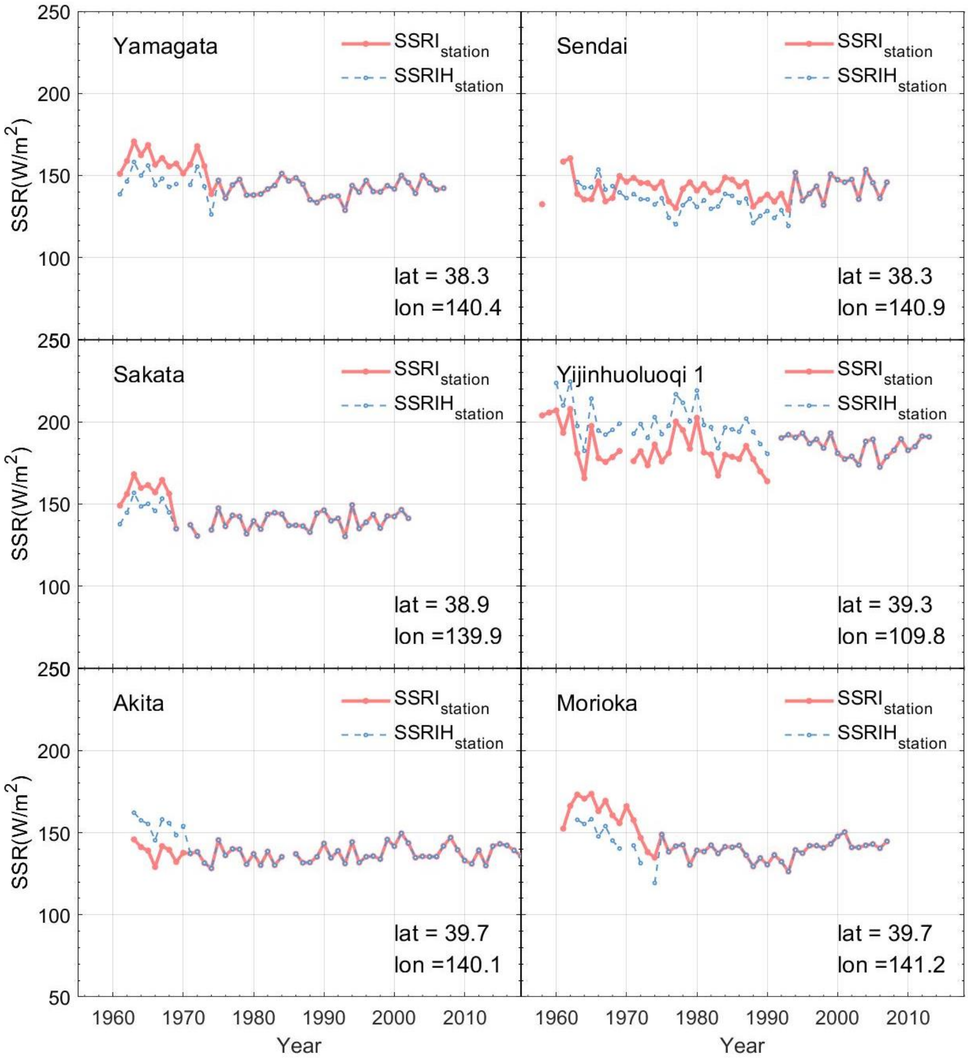
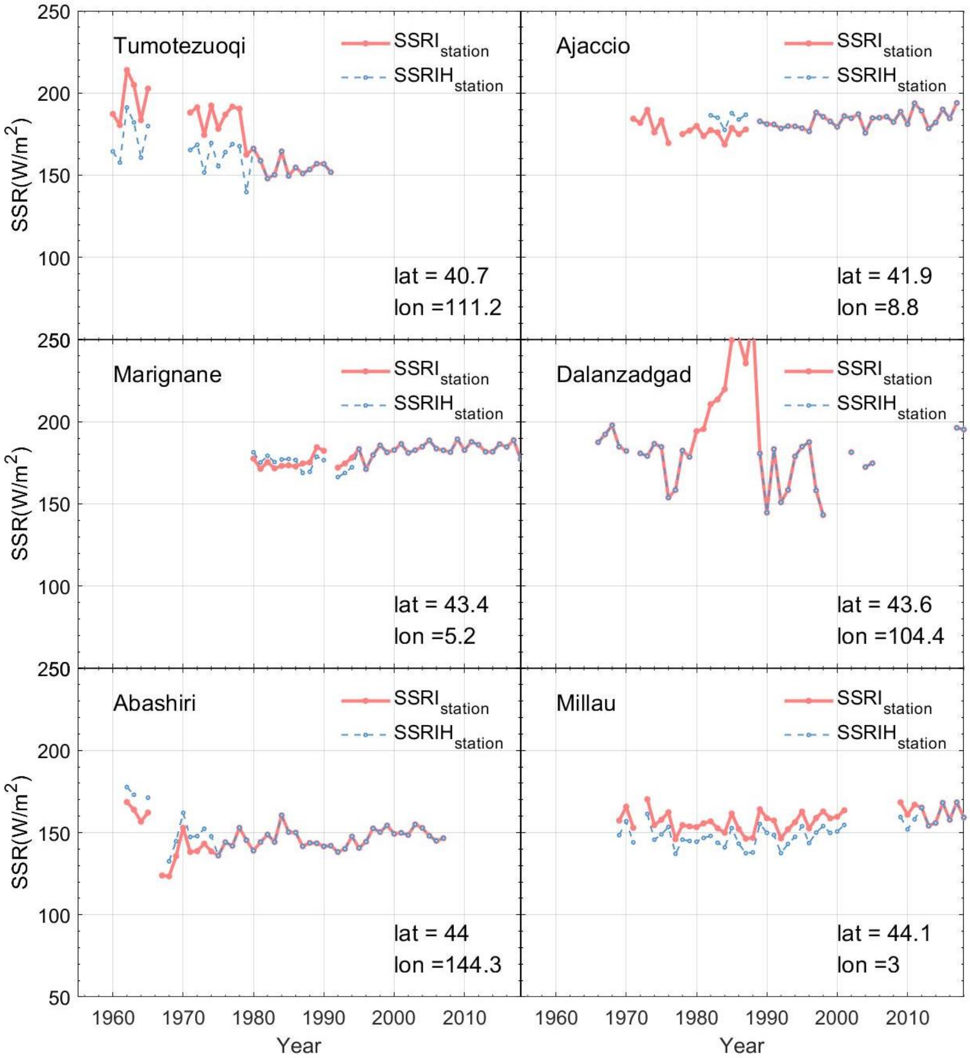


Figure S1-6 Annual variation of SSR calculated from the original station SSR series ($SSRI_{station}$, blue line), the station SSR series after homogenization ($SSRIH_{station}$, red line).



88
89 **Figure S1-7 Annual variation of SSR calculated from the original station SSR series ($SSRI_{station}$, blue line),**
90 **the station SSR series after homogenization ($SSRIH_{station}$, red line).**

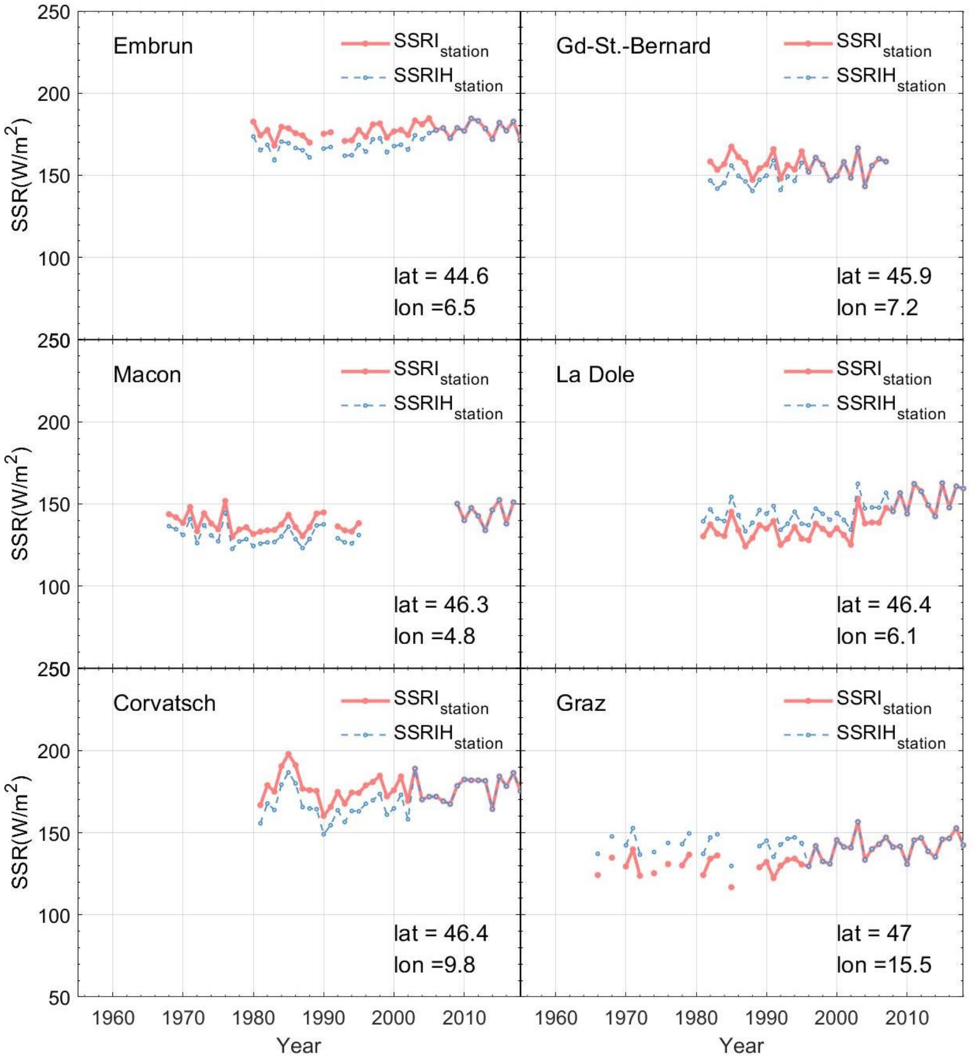


Figure S1-8 Annual variation of SSR calculated from the original station SSR series ($SSRI_{station}$, blue line), the station SSR series after homogenization ($SSRIH_{station}$, red line).

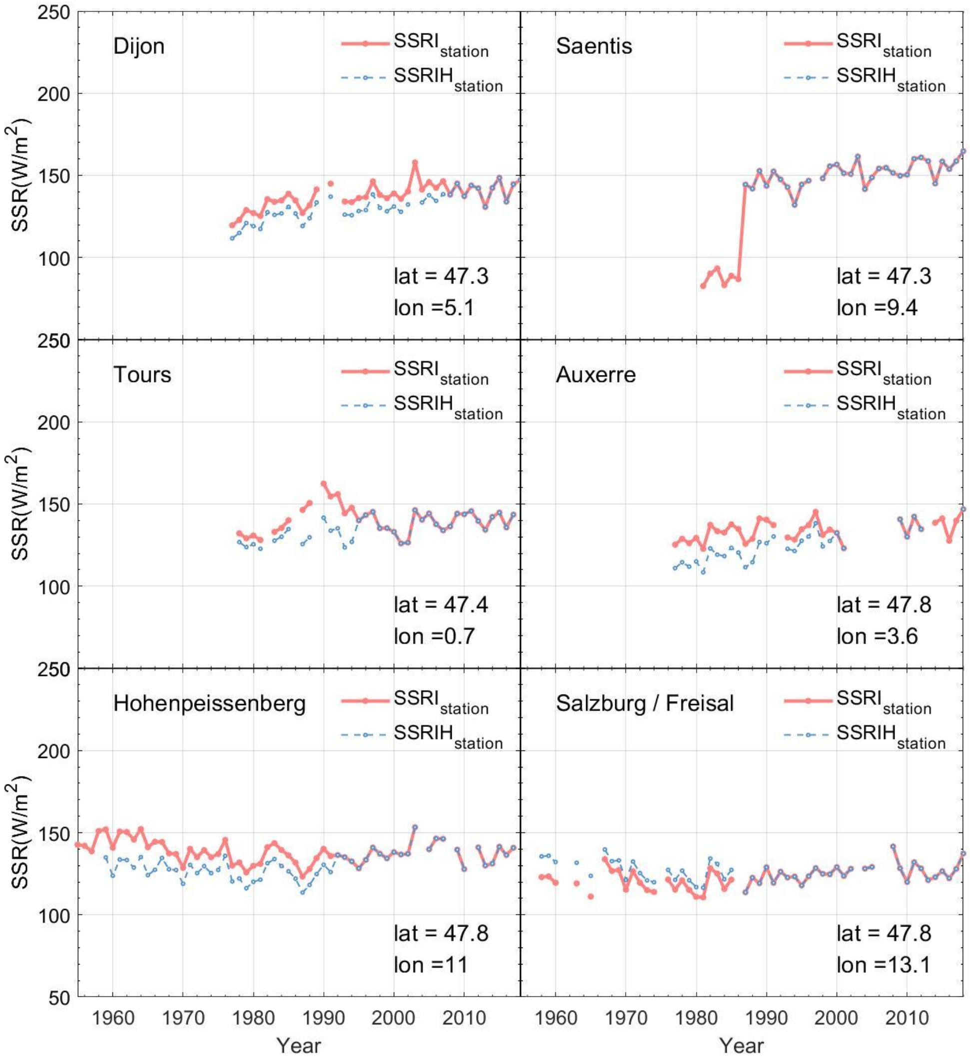


Figure S1-9 Annual variation of SSR calculated from the original station SSR series ($SSRI_{station}$, blue line), the station SSR series after homogenization ($SSRIH_{station}$, red line).

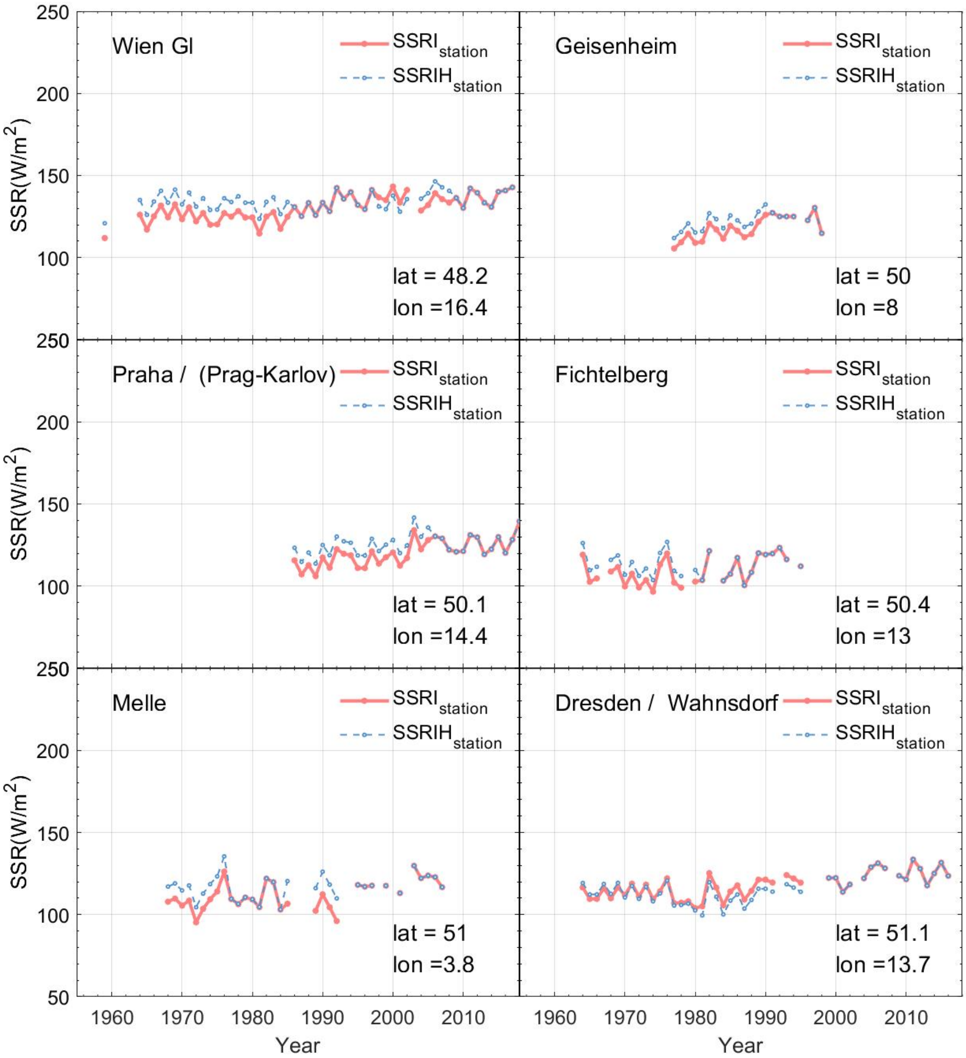
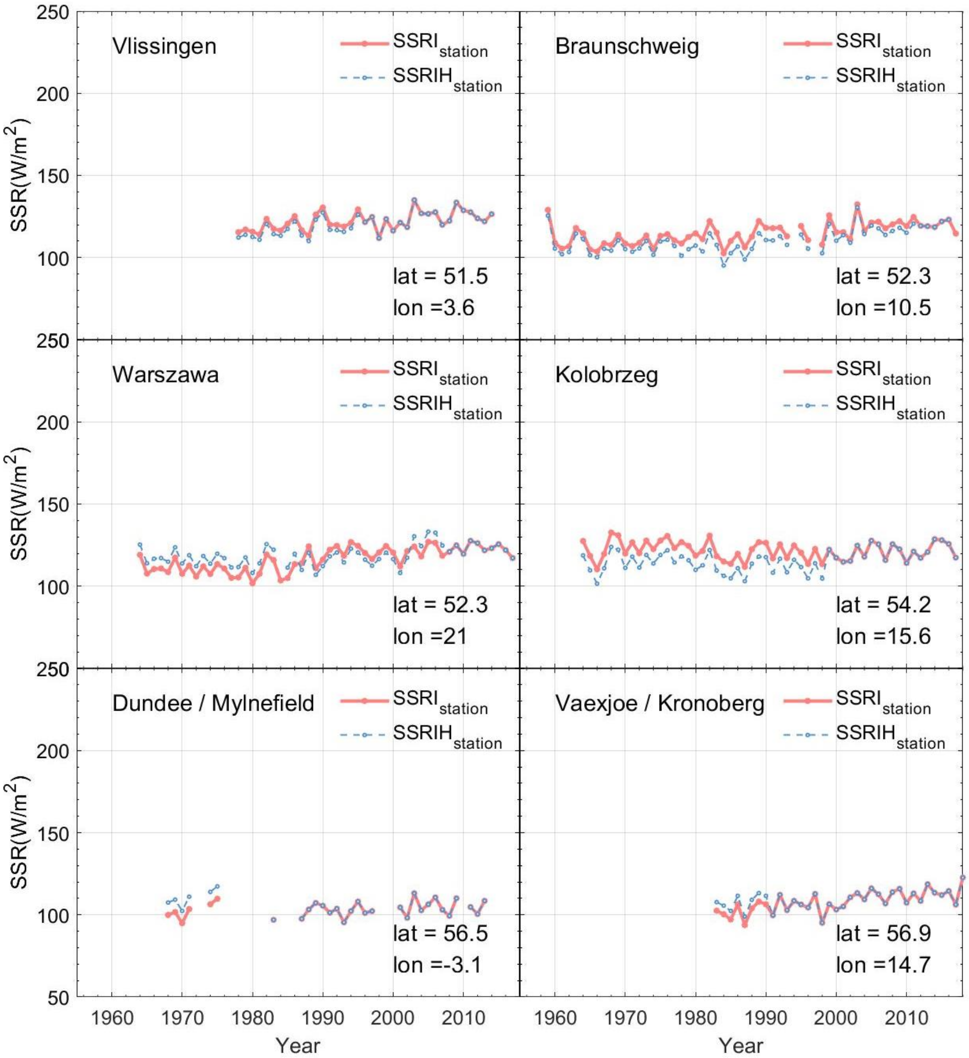


Figure S1-10 Annual variation of SSR calculated from the original station SSR series ($SSRI_{station}$, blue line), the station SSR series after homogenization ($SSRIH_{station}$, red line).



100

101 **Figure S1-11 Annual variation of SSR calculated from the original station SSR series (SSRI$_{station}$, blue line),**

102 **the station SSR series after homogenization (SSRIH$_{station}$, red line).**

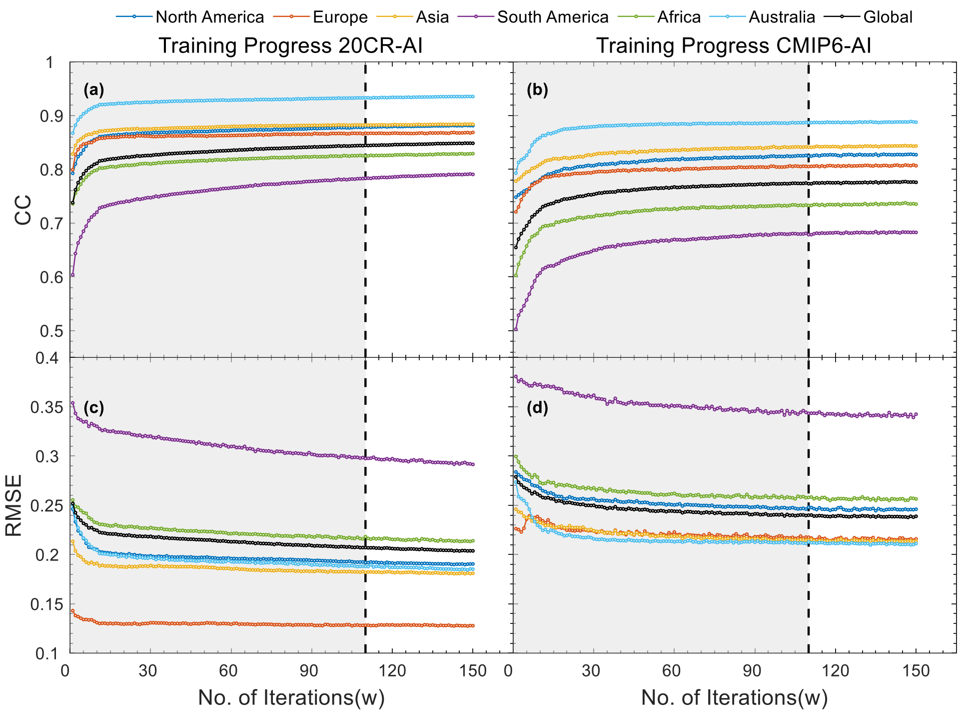
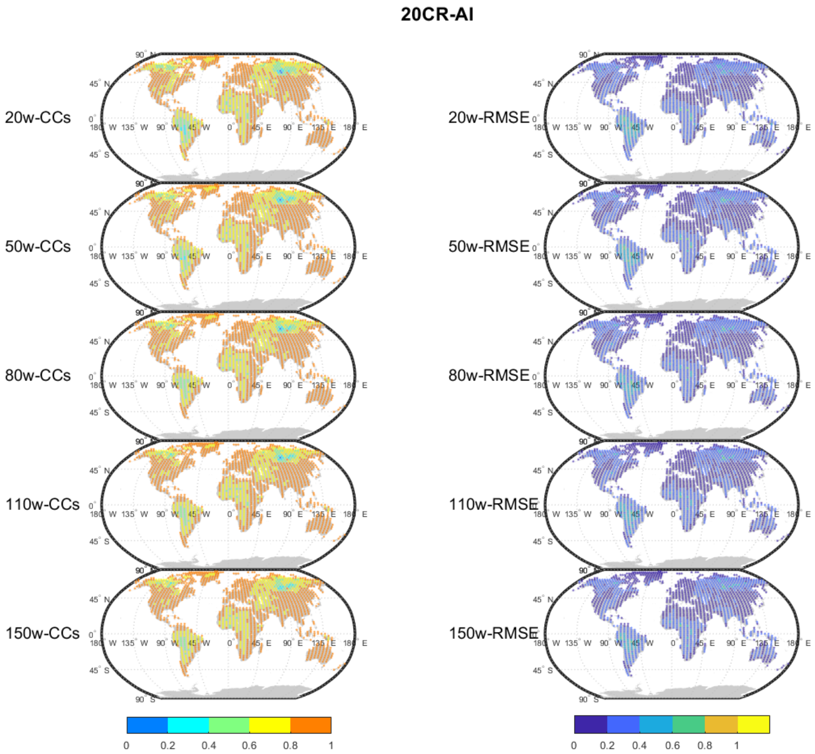


Figure S2: 20CR-AI (CMIP6-AI) reconstruction model evaluation. Figure S3 (a /b) and (c /d) show the correlation coefficient (CC) and root mean squared error (RMSE) of the 20crAI /CMIP6AI model reconstruction results with the validation set for the different number of iterations.

108



109 **Figure S3: 20CR-AI reconstruction model evaluation. The left and right panels show the spatial distribution**
110 **of the CC and the RMSE of the 20CR-AI model reconstruction results with the 20CR validation set for the**
111 **different number of iterations, respectively.**

112

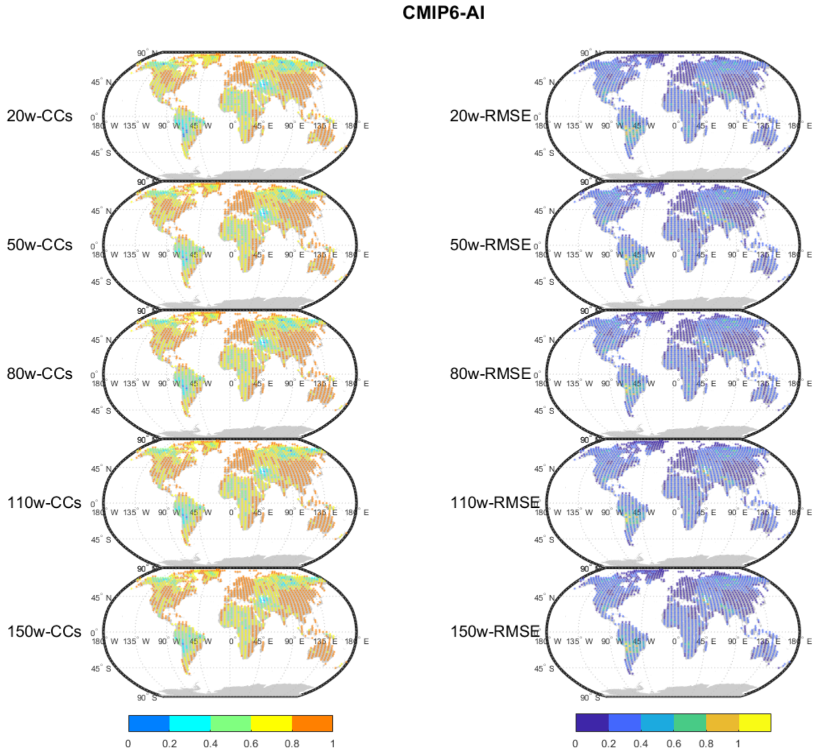


Figure S4: same as Figure S3, but for CMIP6-AI.

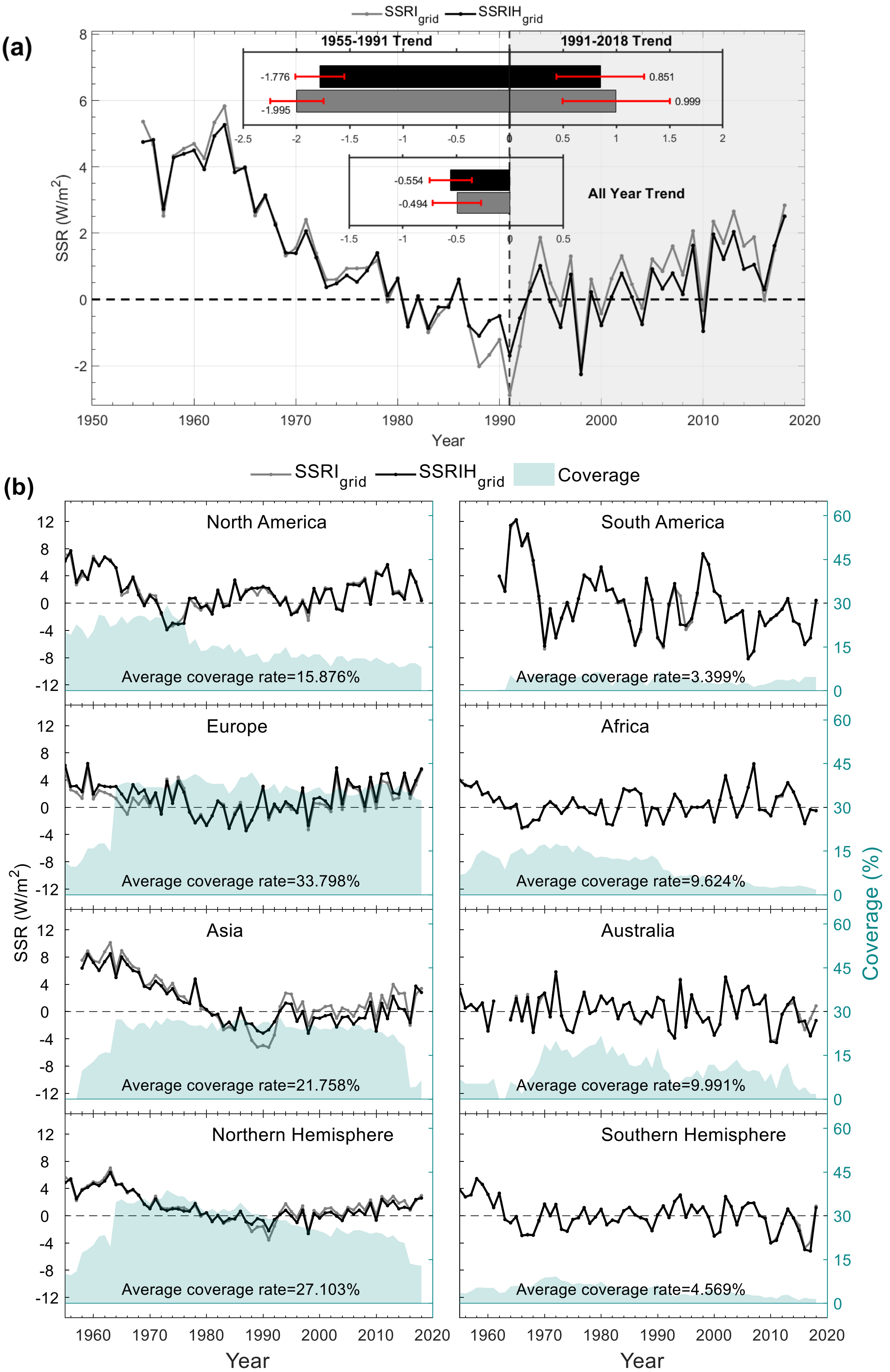


Figure S5: Time series of the annual global (a) /regional (b) SSR anomaly variations (relative to 1971-2000) before /after homogenization. The Grey /black solid line represents SSR before homogenization ($SSRI_{grid}$) /$SSRIH_{grid}$ annual anomalies. The histograms represent the decadal trends of the $SSRI_{grid}$ /$SSRIH_{grid}$ (unit: W/m^2 per decade) and their 95% uncertainty range during three periods 1955-1988, 1988-2018 and 1955-

122 **2018.**

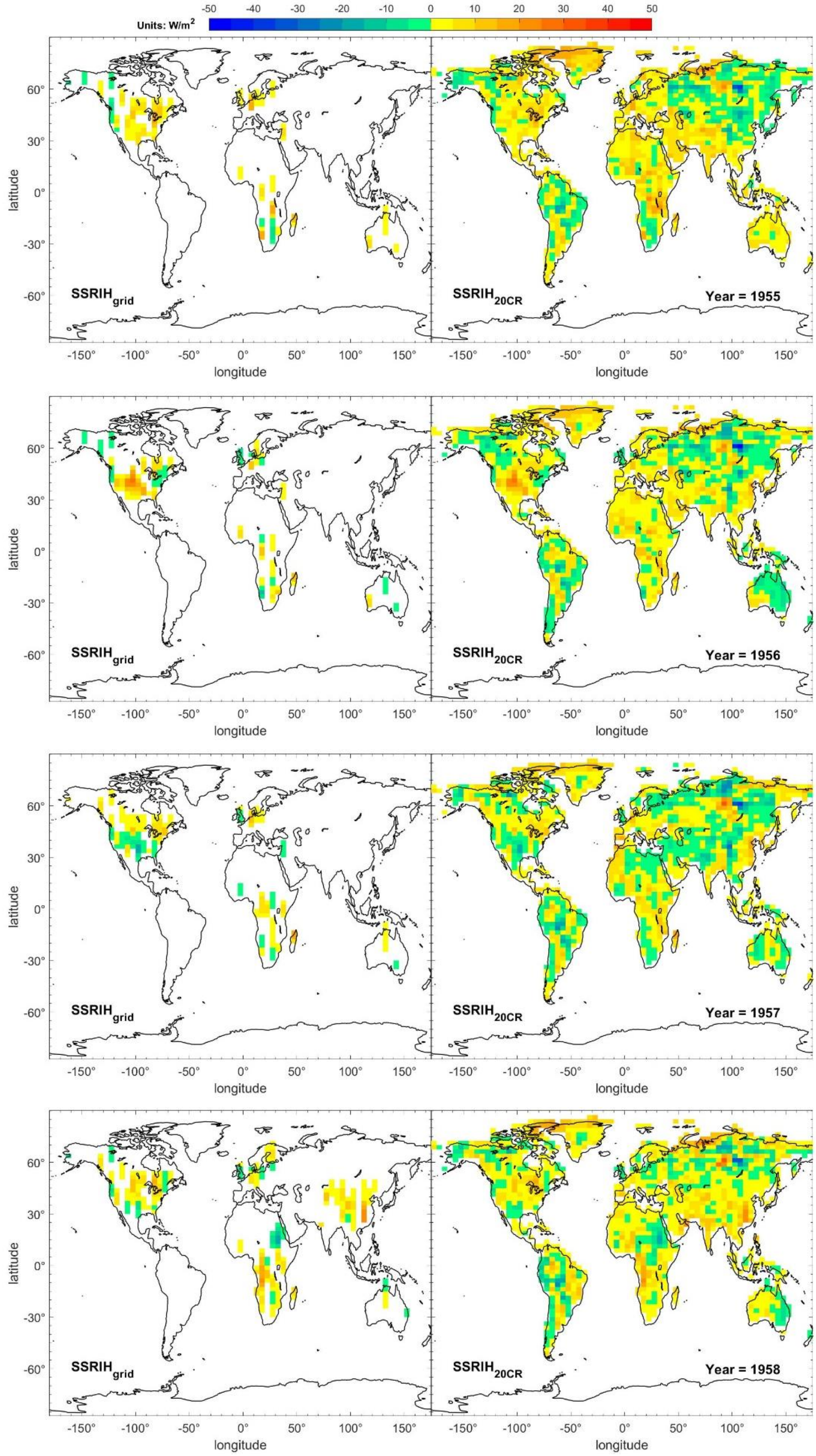


Figure S6-1: Spatial distribution of SSRIH$_{grid}$ (column 1) and the SSR of reconstruction based on the 20CR-AI model (SSRIH$_{20CR}$ (column 2)) in typical years (1955-1958).

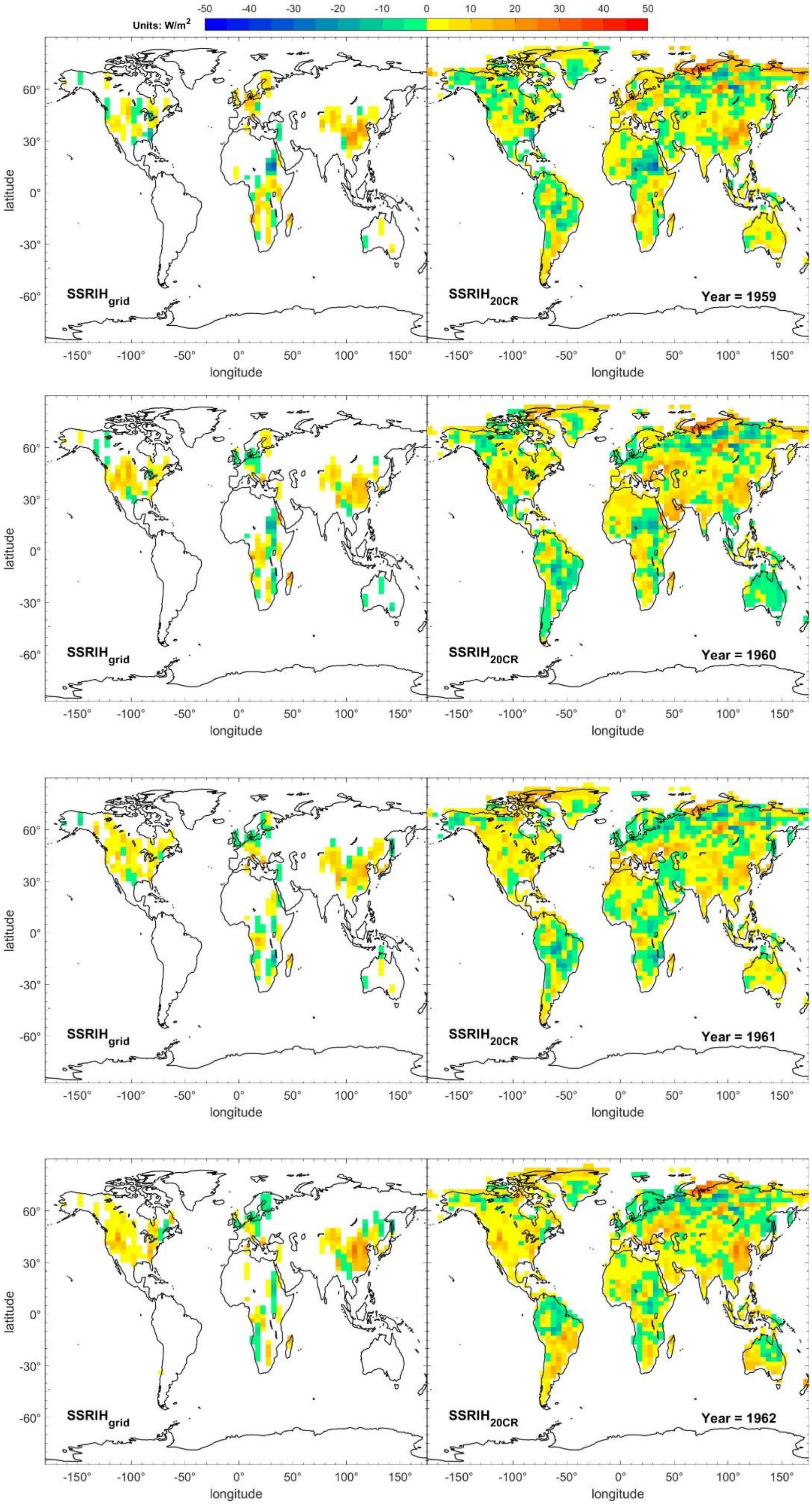


Figure S6-2: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (1959-1962).

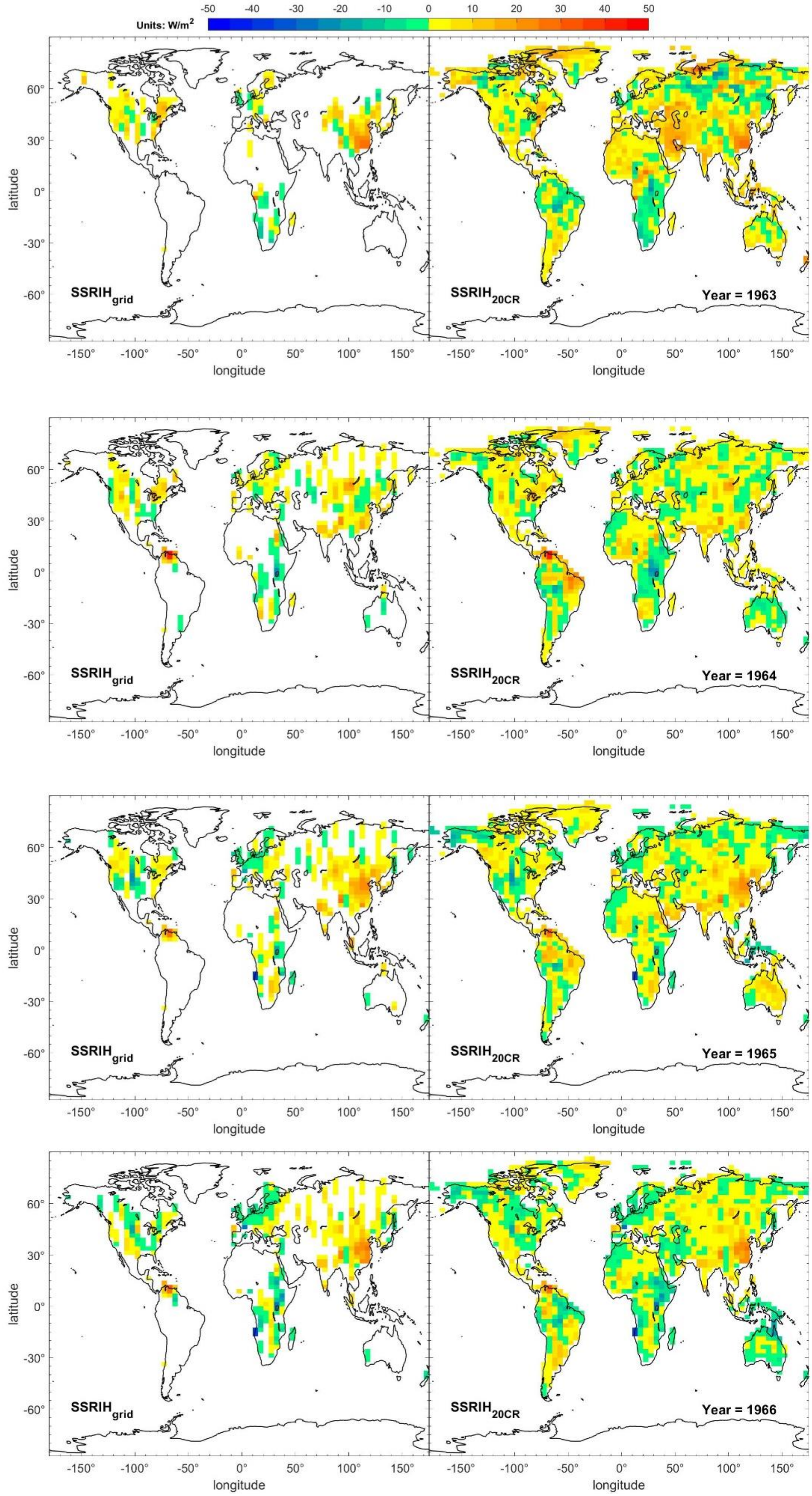


Figure S6-3: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (1963-1966).

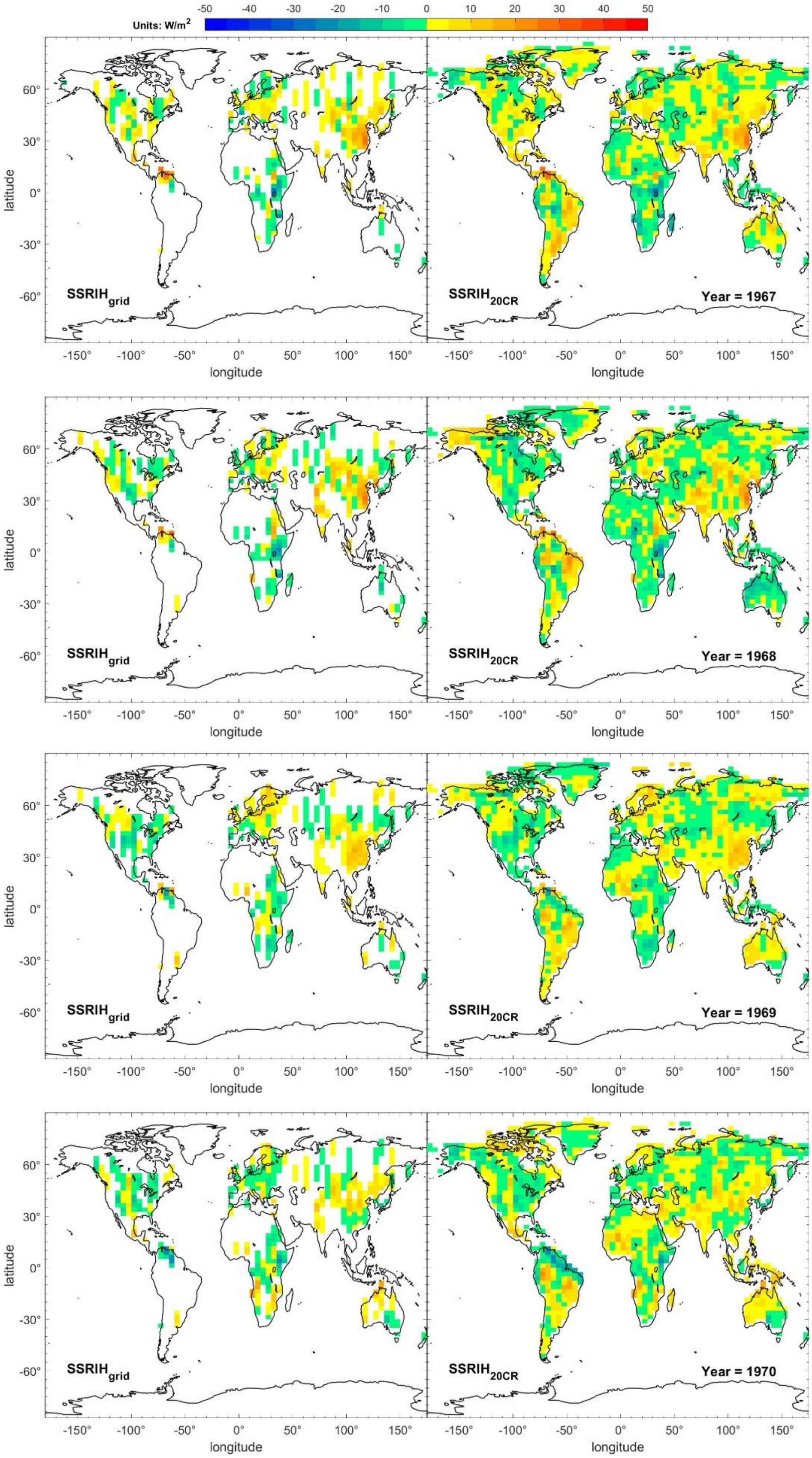


Figure S6-4: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (1967-1970).

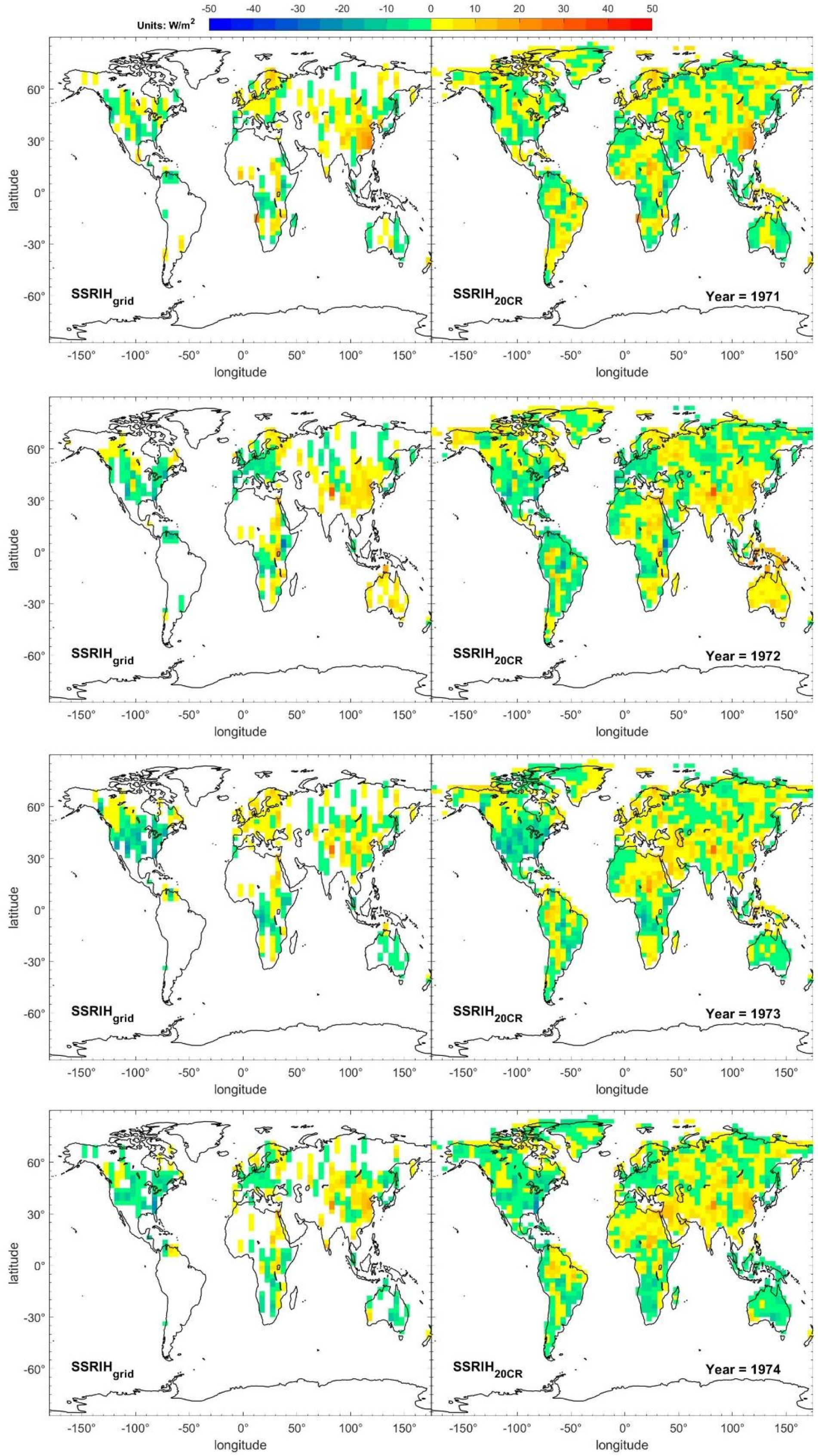


Figure S6-5: Spatial distribution of $SSRIH_{grid}$ (column 1) and $SSRIH_{20CR}$ (column 2) in typical years (1971-1974).

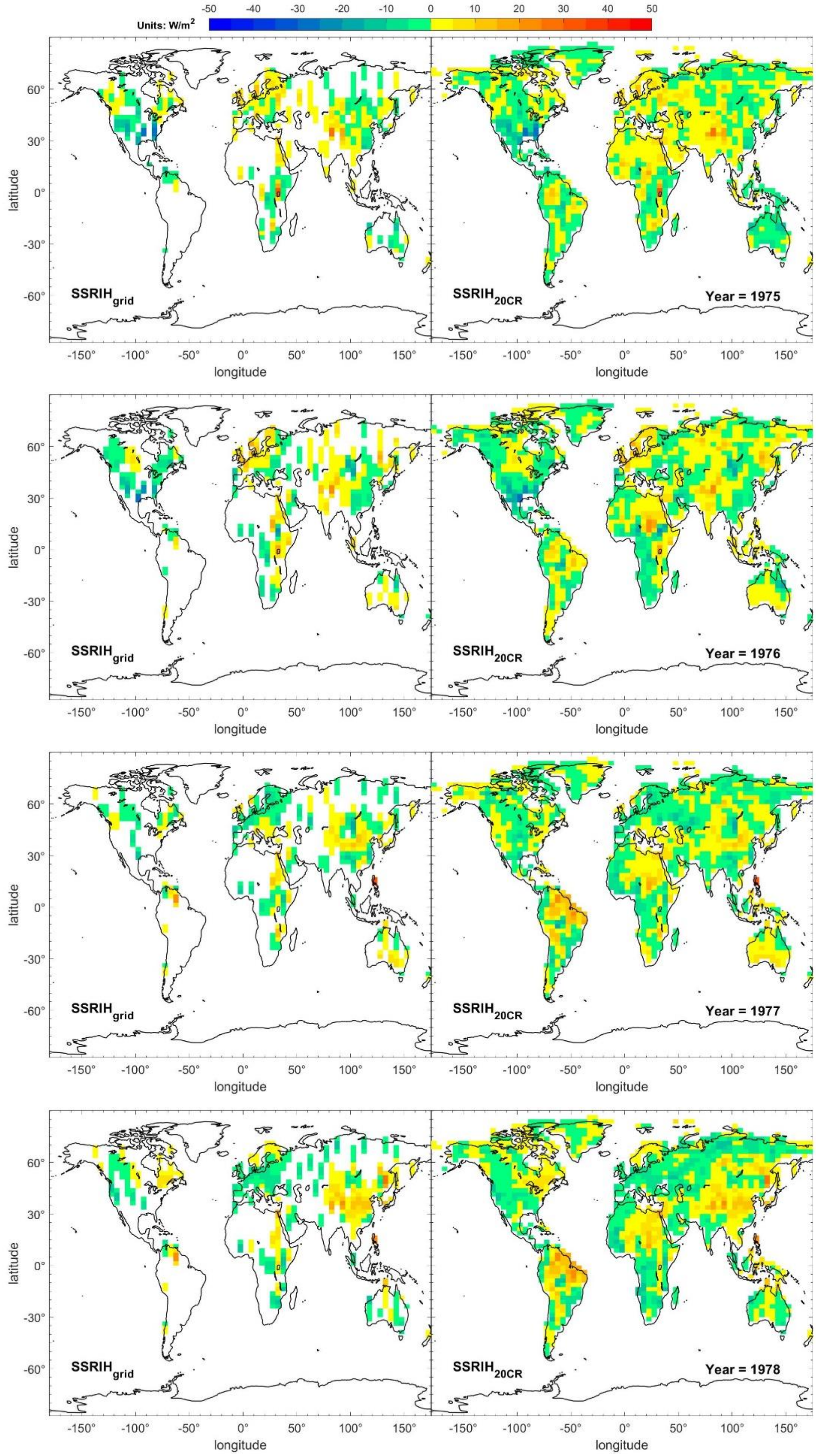


Figure S6-6: Spatial distribution of SSRIHgrid (column 1) and SSRIH20CR (column 2) in typical years (1975-1978).

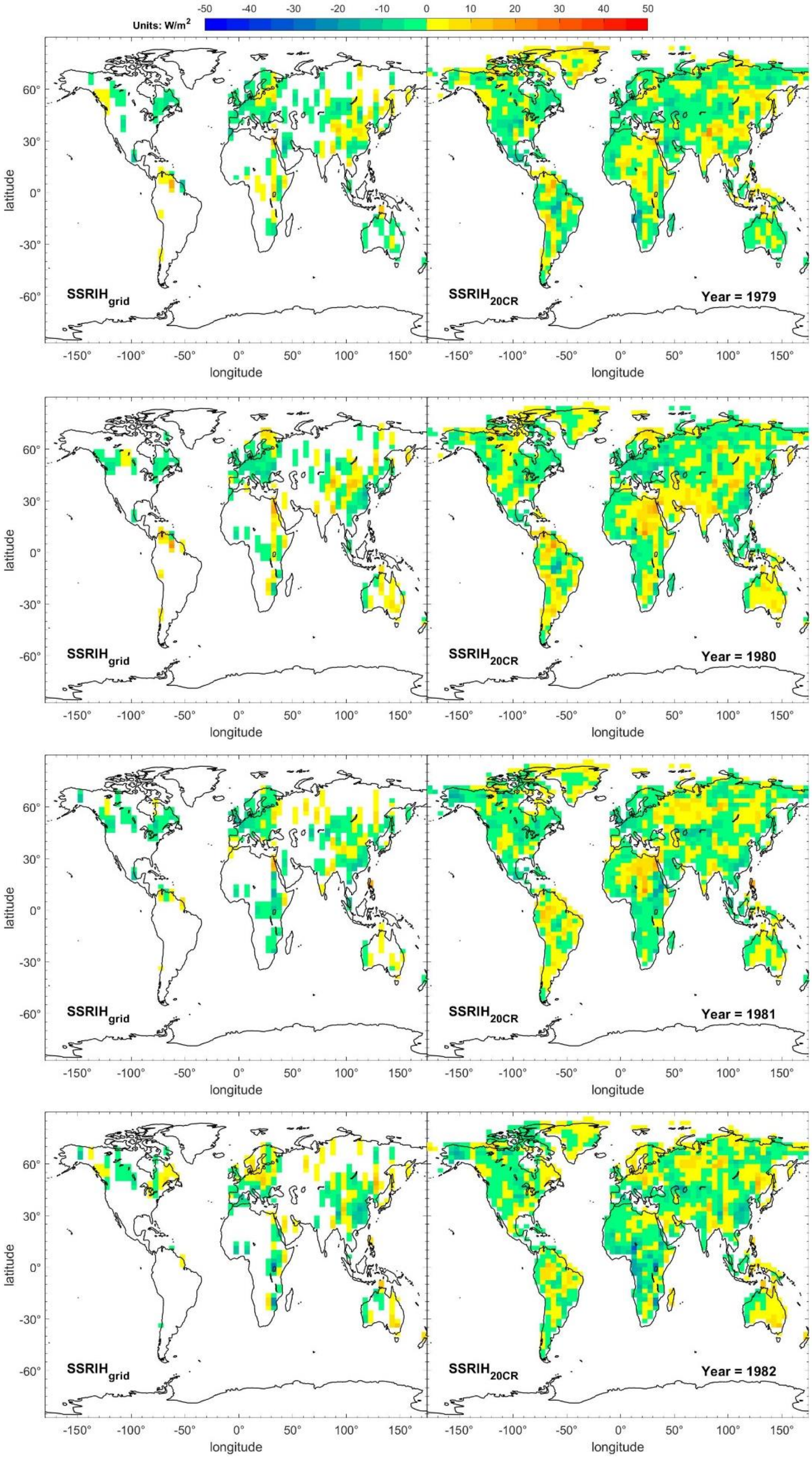


Figure S6-7: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (1979-1982).

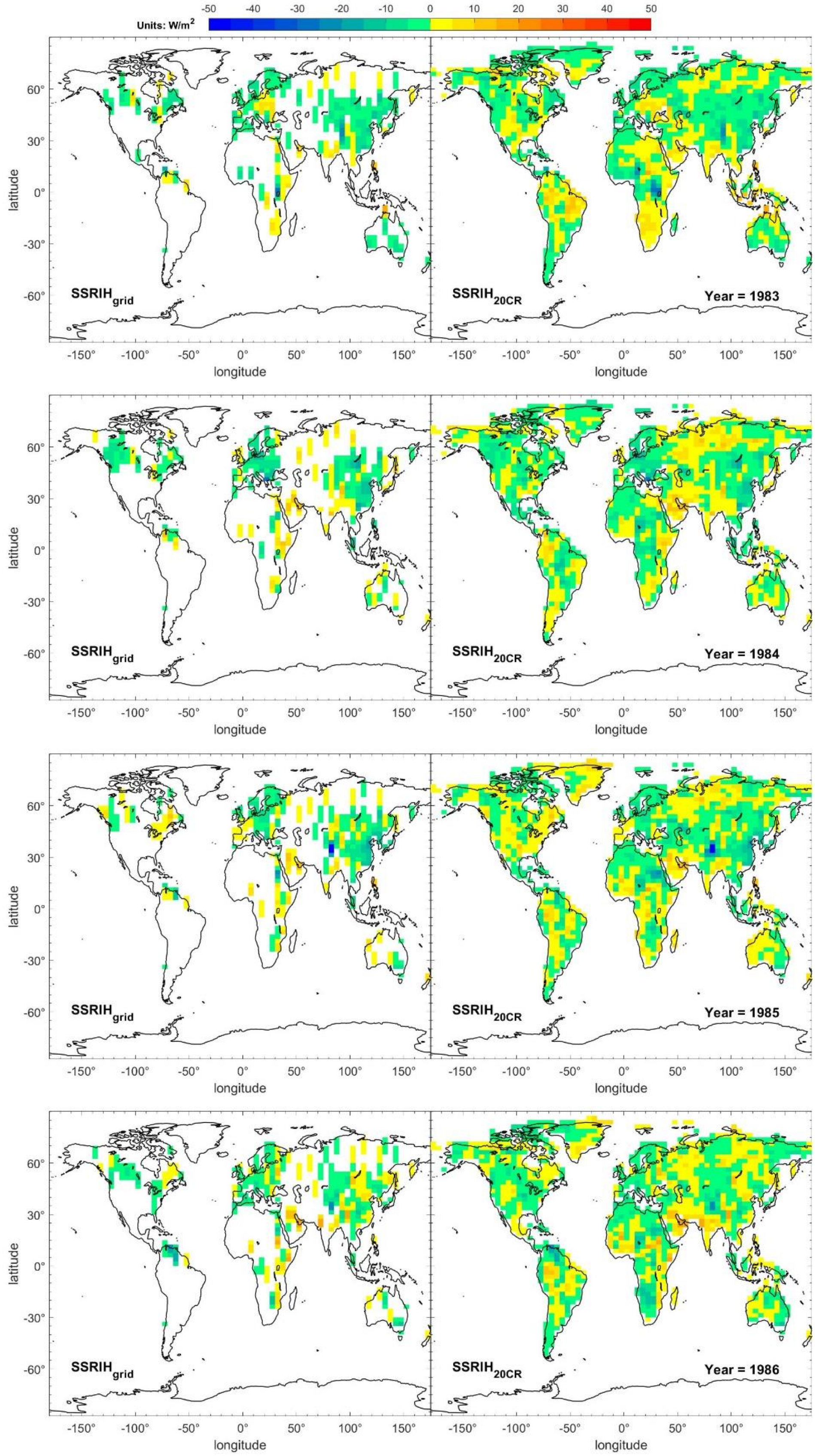


Figure S6-8: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (1983-1986).

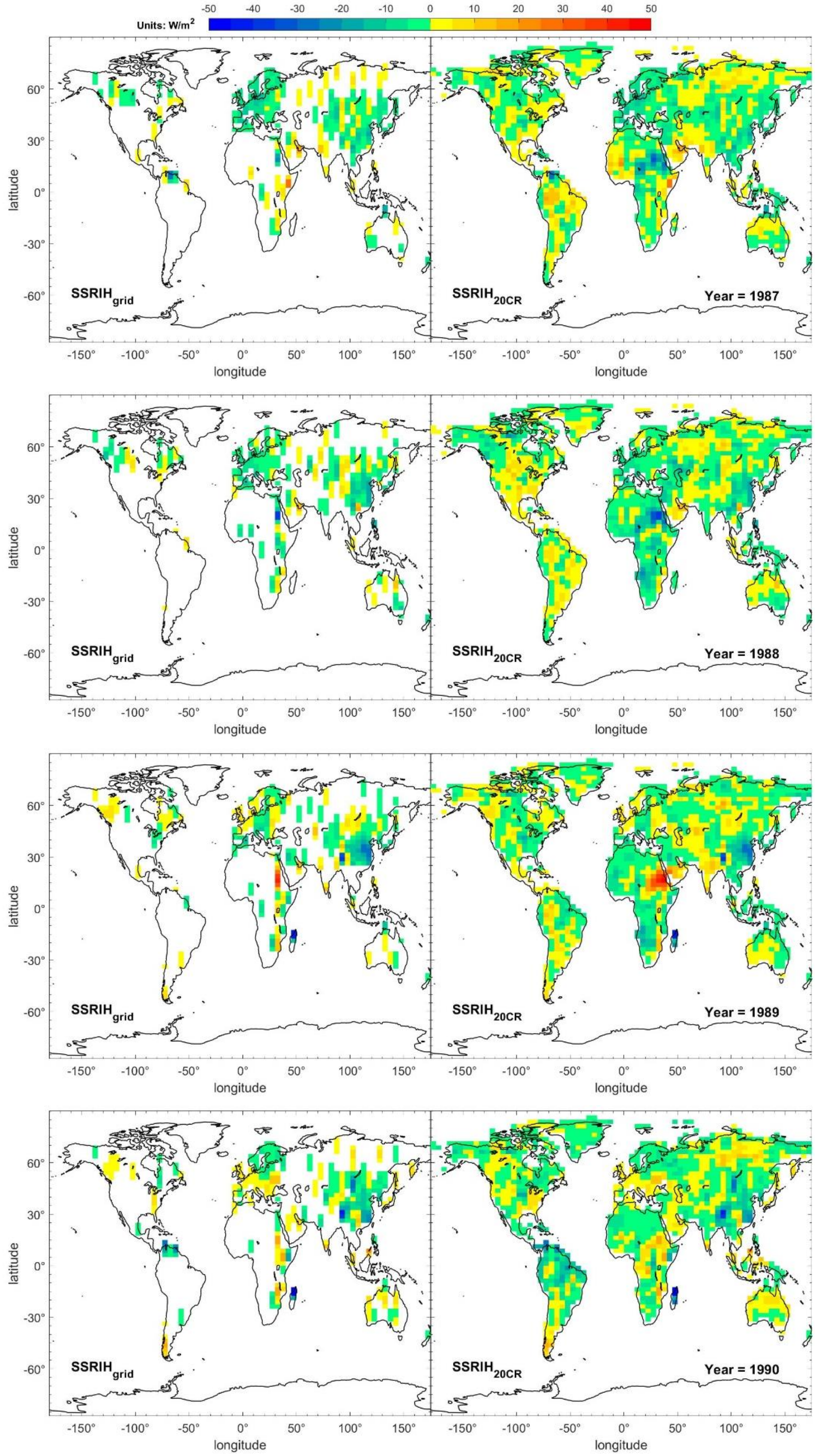


Figure S6-9: Spatial distribution of $SSRIH_{grid}$ (column 1) and $SSRIH_{20CR}$ (column 2) in typical years (1987-1990).

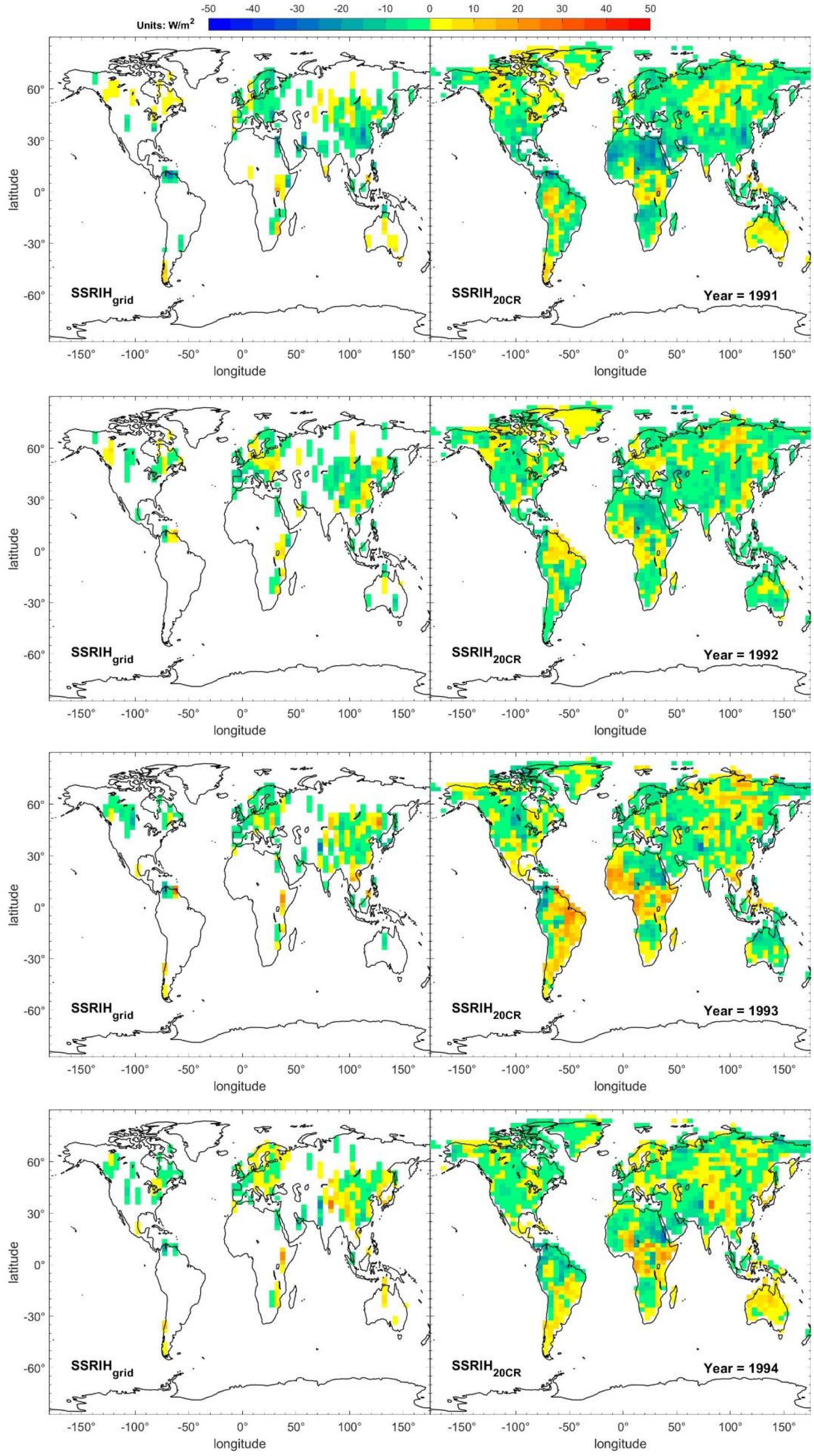


Figure S6-10: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (1991-1994).

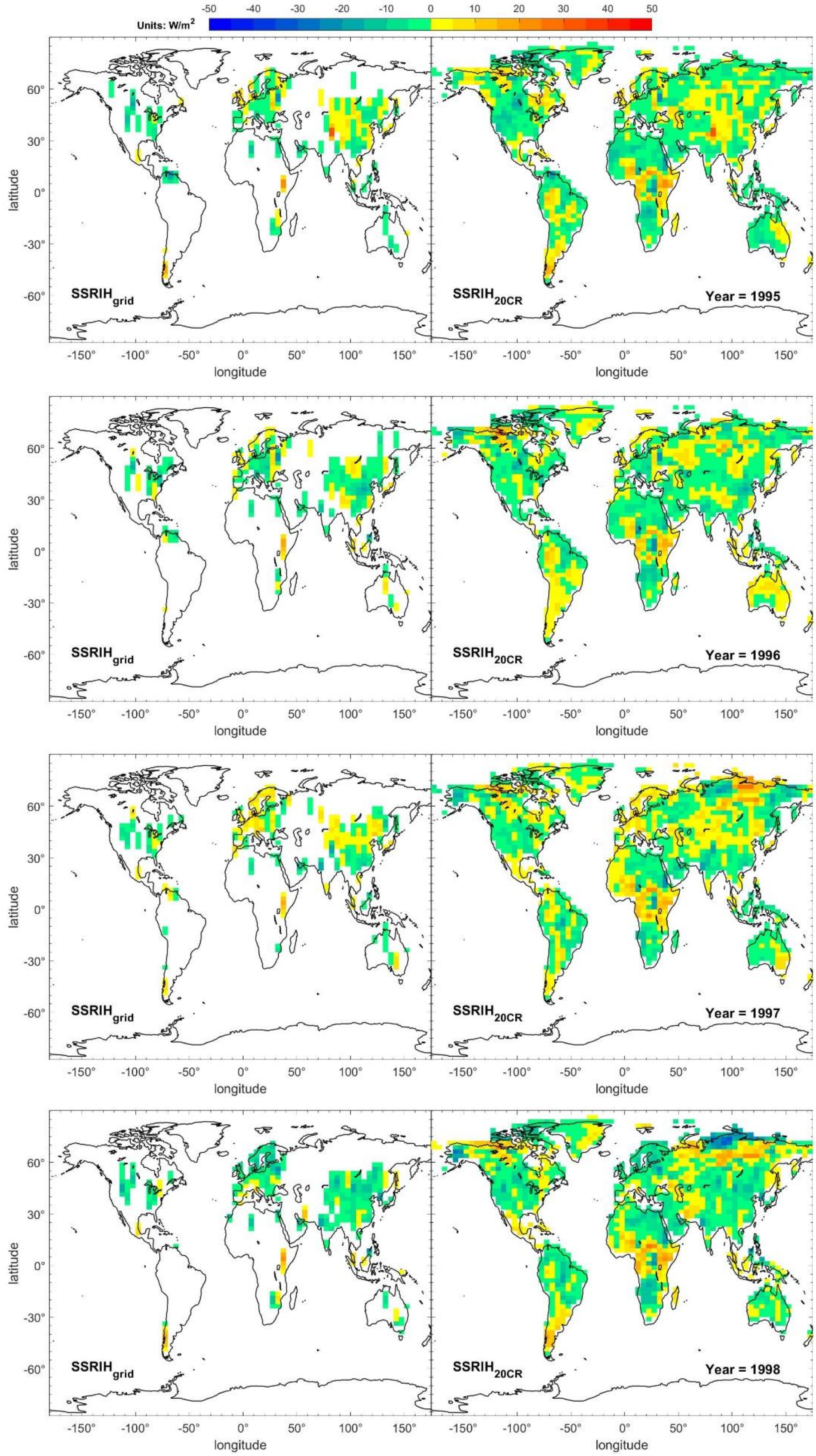


Figure S6-11: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (1995-1998).

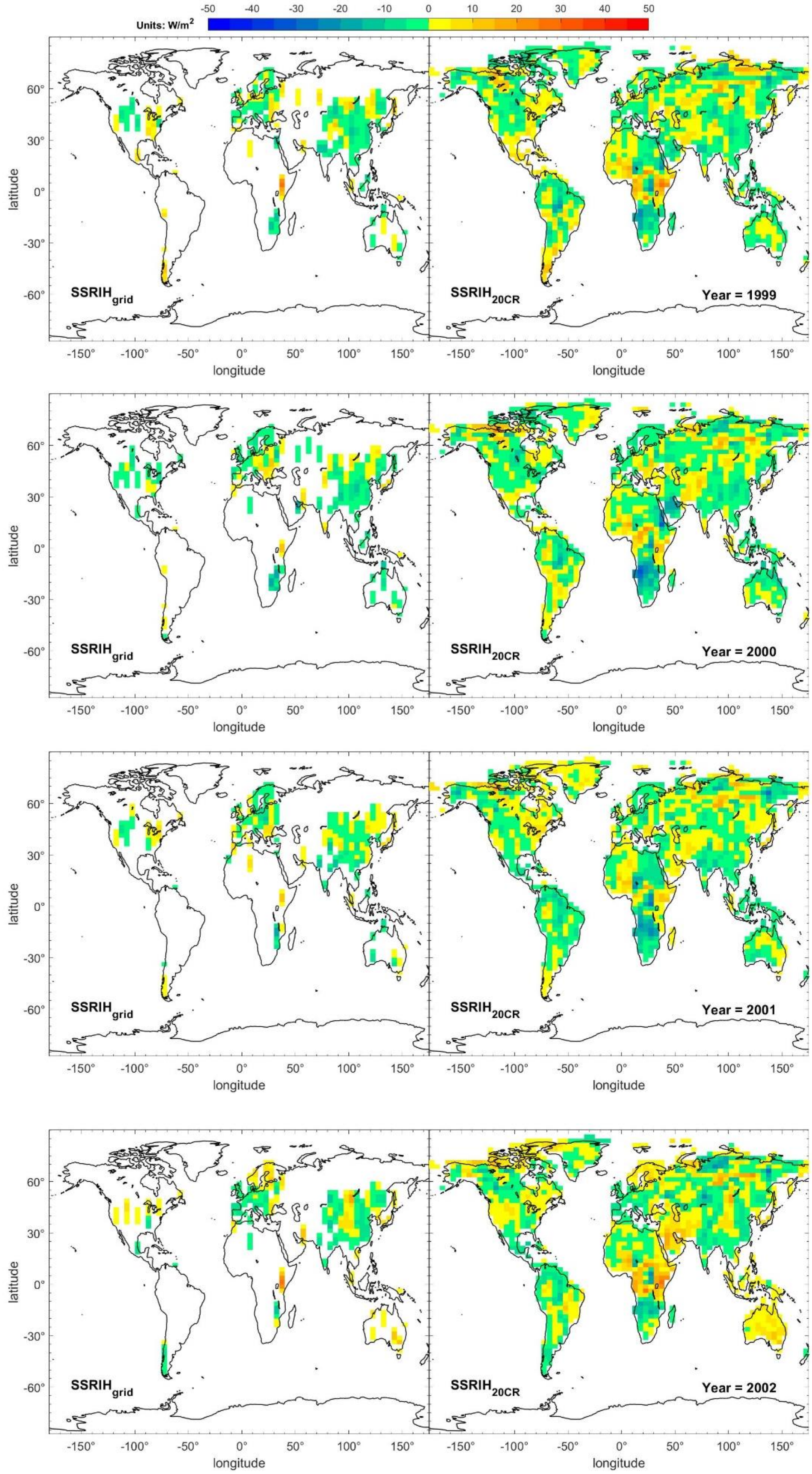


Figure S6-12: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (1999-2002).

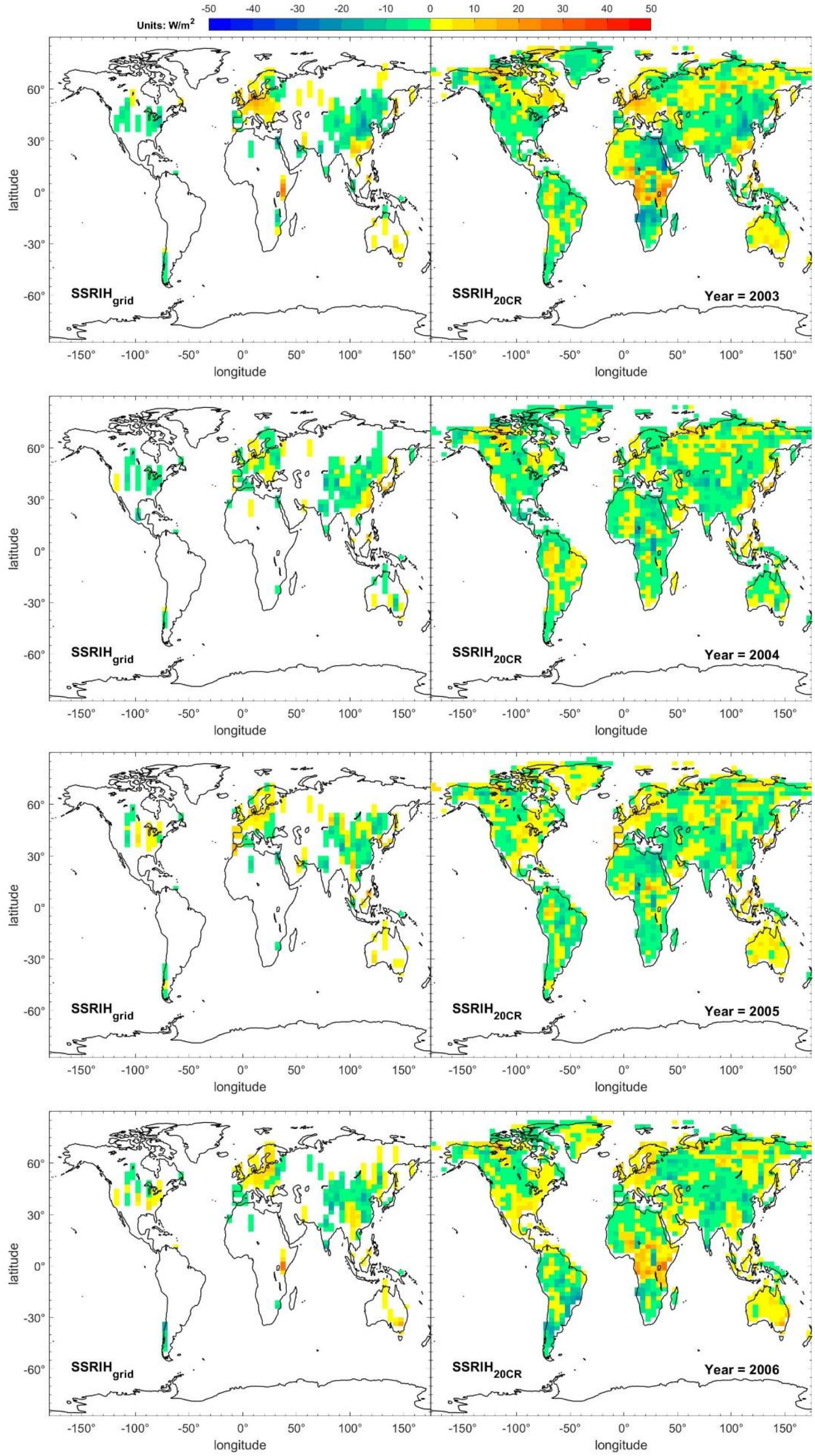


Figure S6-13: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (2003-2006).

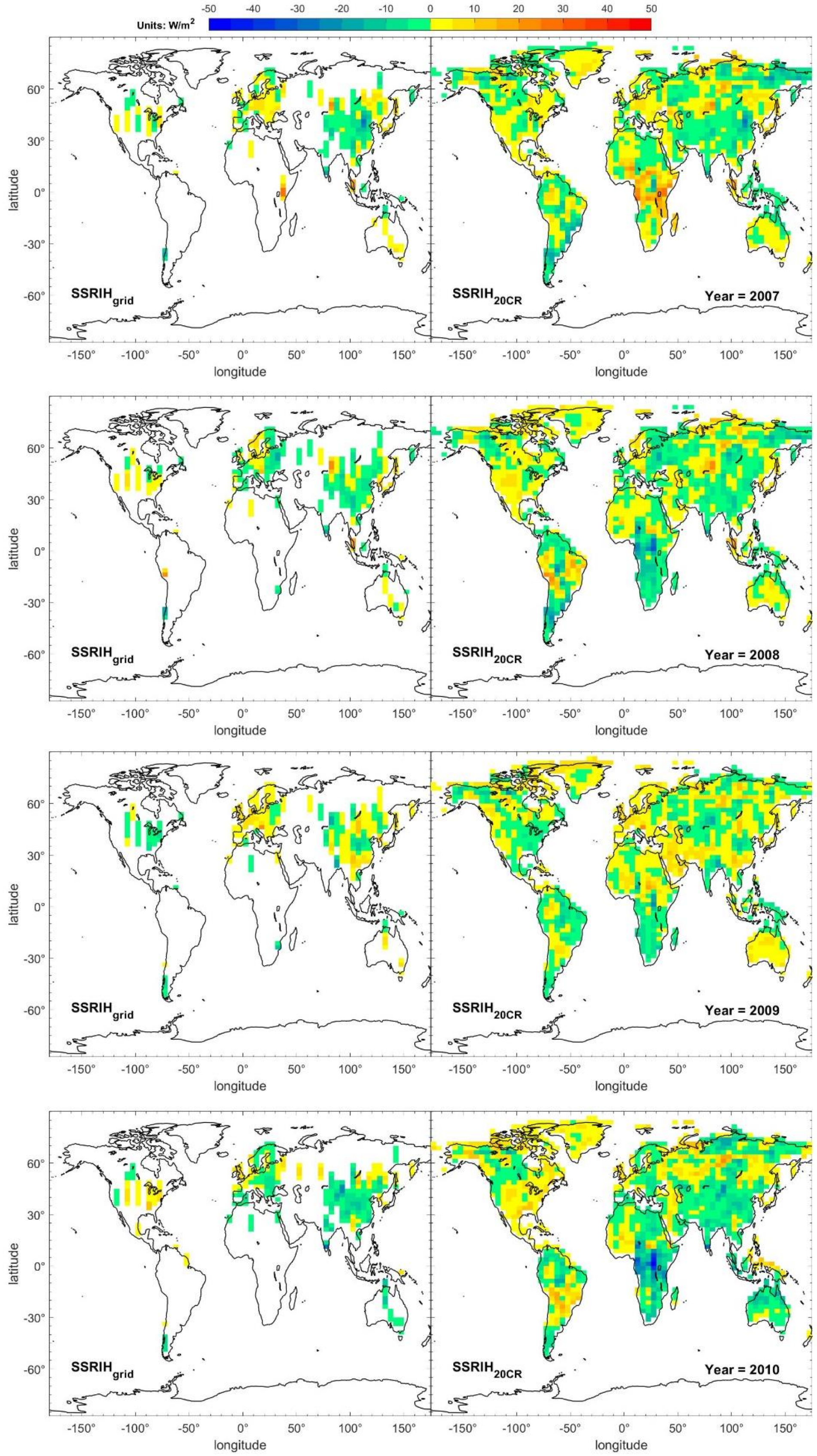


Figure S6-14: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (2007-2010).

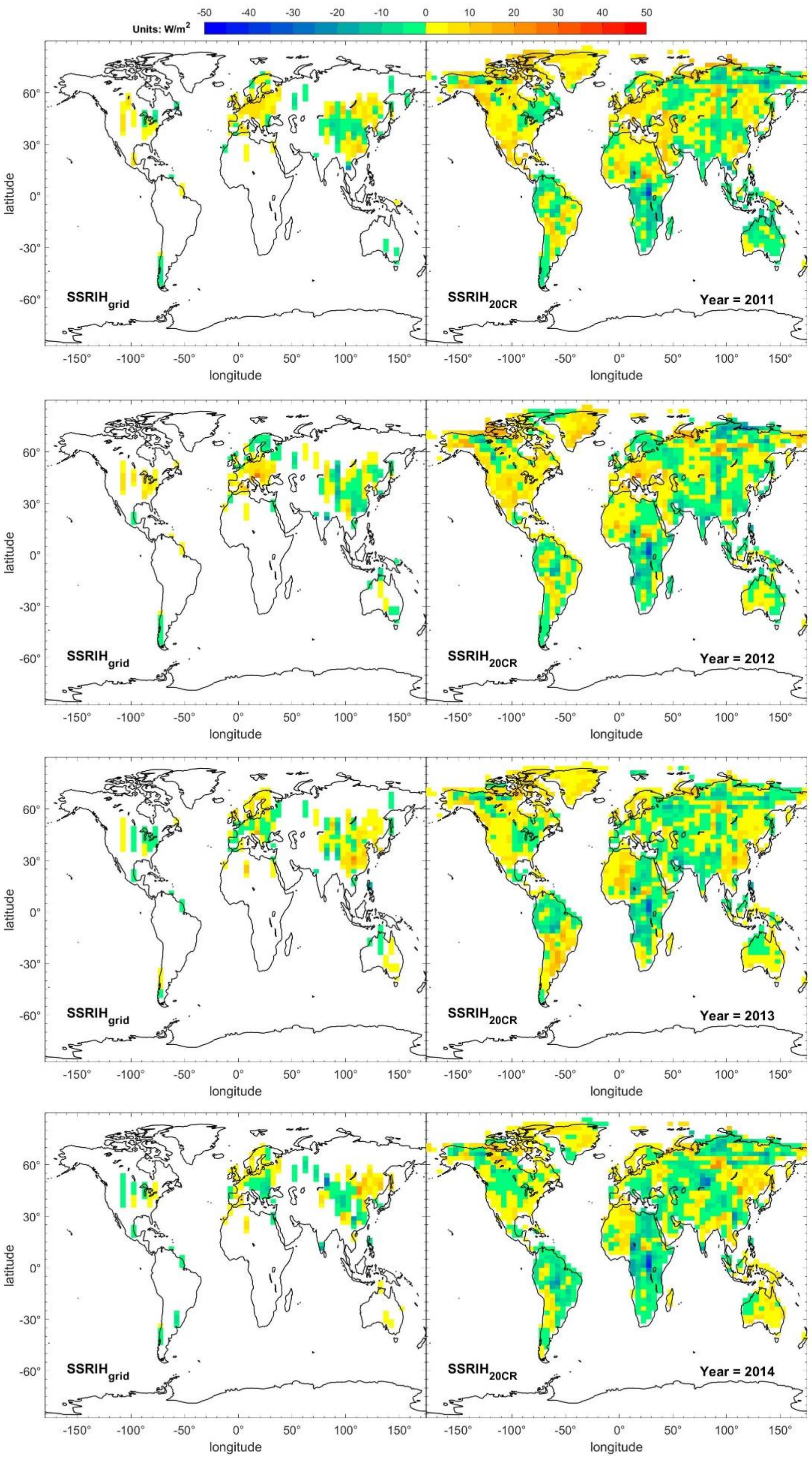


Figure S6-15: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (2011-2014).

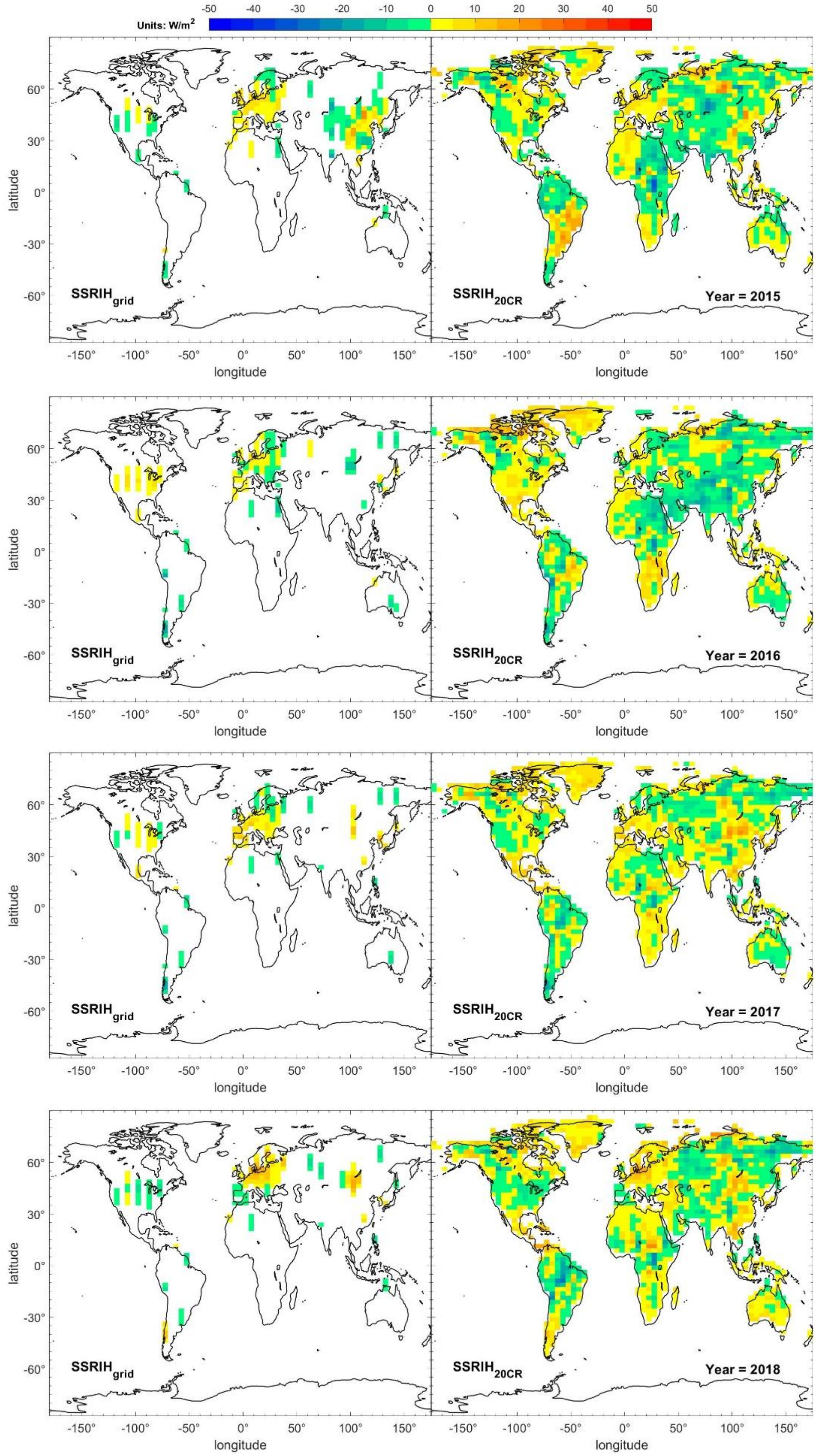


Figure S6-16: Spatial distribution of SSRIH$_{grid}$ (column 1) and SSRIH$_{20CR}$ (column 2) in typical years (2015-2018).

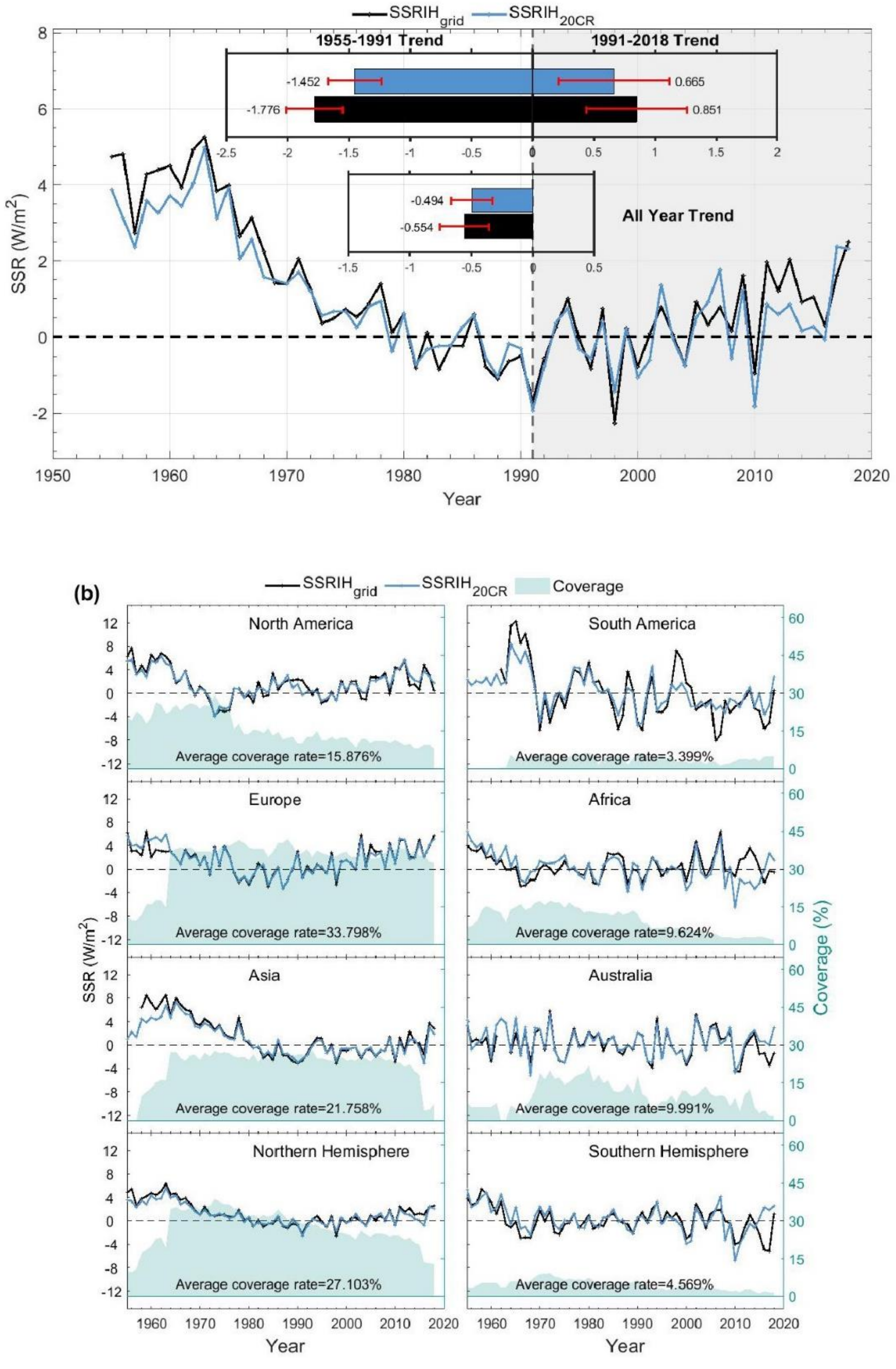


Figure S7: Global and regional (except for Antarctica) land annual SSR anomaly variations (relative to 1971-2000) before /after reconstruction. The Black solid line represents the SSRIH$_{grid}$ annual anomalies. The solid blue line represents the reduced SSRIH$_{20CR}$ annual anomalies. The histograms represent the decadal trends of the SSRIH$_{grid}$ /SSRIH$_{20CR}$ (unit: W/m2 per decade) and their 95% uncertainty range from 1955 to 1991, 1991-2018 and 1955-2018, and the SSRIH$_{20CR}$ is reduced to the grid boxes with *in situ* observations.

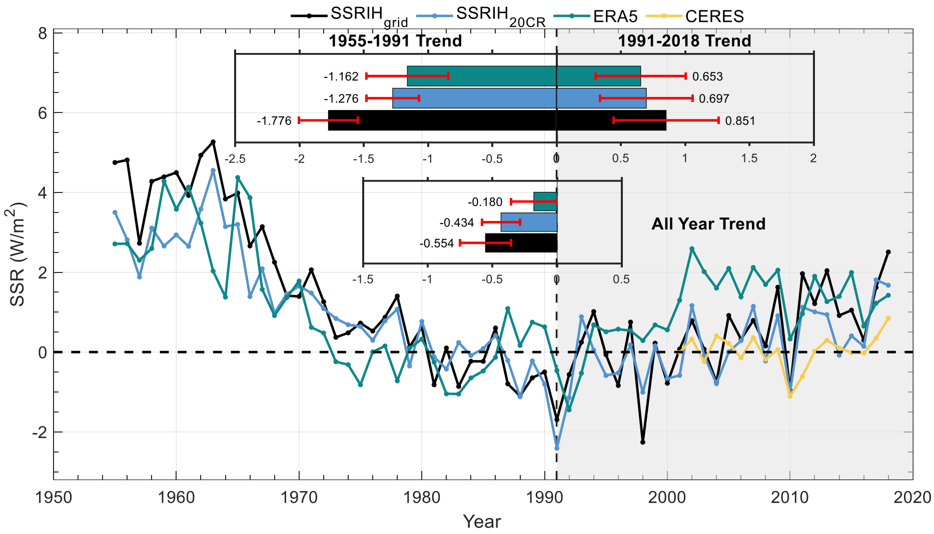


Figure S8: Global land (except for Antarctica) annual SSR anomaly variations (relative to 1971-2000) before /after reconstruction. The Black solid line represents the $SSRIH_{grid}$ annual anomalies. The solid blue line represents the $SSRIH_{20CR}$ annual anomalies. The solid green line represents the ERA5 annual anomalies. The solid yellow line represents the CERES annual anomalies. The histograms represent the decadal trends of the $SSRIH_{grid}$ /$SSRIH_{20CR}$ /ERA5 (unit: W/m^2 per decade) and their 95% uncertainty range from 1955 to 1991, 1991-2018 and 1955-2018.

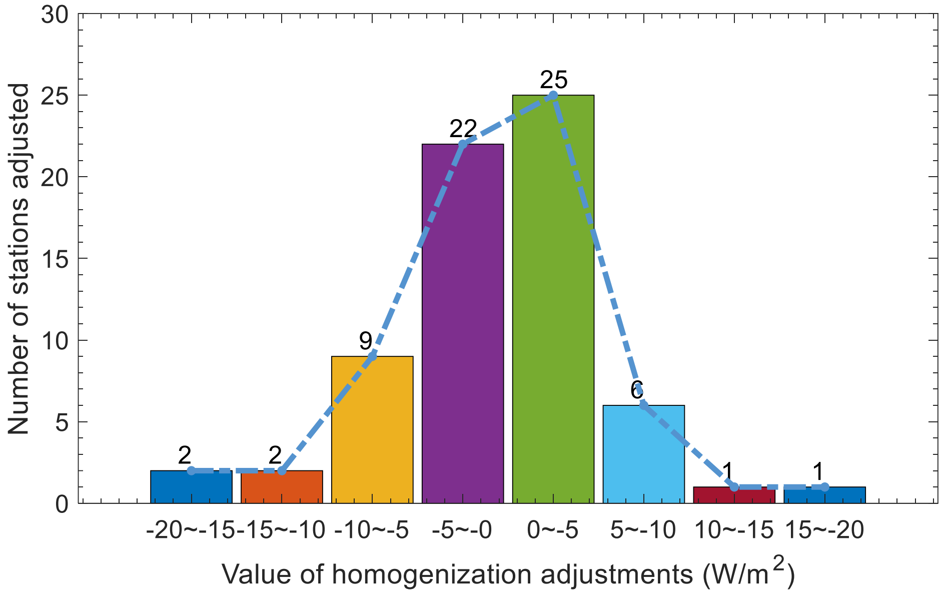


Figure S9: Distribution of annual SSR homogenization adjustments.

(The histogram is based on adjustments from all 66 stations adjusted in this paper)

Reference

Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., and Catanzaro, B.: Image Inpainting for Irregular Holes Using Partial Convolutions, Cham, 89-105, doi: org/10.1007/978-3-030-01252-6_6, 2018.