scientific data

OPEN

DATA DESCRIPTOR

A long-term reconstructed TROPOMI solar-induced fluorescence dataset using machine learning algorithms

Xingan Chen[1], Yuefei Huang[1,2,3], Chong Nie[4,5], Shuo Zhang[1]✉, Guangqian Wang[1], Shiliu Chen[1] & Zhichao Chen[1]

Photosynthesis is a key process linking carbon and water cycles, and satellite-retrieved solar-induced chlorophyll fluorescence (SIF) can be a valuable proxy for photosynthesis. The TROPOspheric Monitoring Instrument (TROPOMI) on the Copernicus Sentinel-5P mission enables significant improvements in providing high spatial and temporal resolution SIF observations, but the short temporal coverage of the data records has limited its applications in long-term studies. This study uses machine learning to reconstruct TROPOMI SIF (RTSIF) over the 2001–2020 period in clear-sky conditions with high spatio-temporal resolutions (0.05° 8-day). Our machine learning model achieves high accuracies on the training and testing datasets ($R^2 = 0.907$, regression slope = 1.001). The RTSIF dataset is validated against TROPOMI SIF and tower-based SIF, and compared with other satellite-derived SIF (GOME-2 SIF and OCO-2 SIF). Comparing RTSIF with Gross Primary Production (GPP) illustrates the potential of RTSIF for estimating gross carbon fluxes. We anticipate that this new dataset will be valuable in assessing long-term terrestrial photosynthesis and constraining the global carbon budget and associated water fluxes.

Background & Summary

Accurate quantification of gross primary production (GPP) through photosynthesis is essential for studies of ecosystem function, carbon cycle, human welfare, and net-zero carbon emission[1–4]. Various methods have been developed to estimate GPP at the global scale, which can be divided into three main categories: enzyme kinetic (process-based) models[5–7], light use efficiency (LUE) models[8–12], and data-driven approaches[13–17]. While a wide range of global GPP estimates is available, the significant discrepancies in GPP estimates generated by different methods remain one of the most uncertain aspects in quantifying the global carbon cycle[18–22]. Over the past decade, advances in global remote sensing of solar-induced chlorophyll fluorescence (SIF) have made it possible to inform on vegetation photosynthetic activity at a global scale[23–30], providing new opportunities for accurate GPP estimates.

SIF is a small fraction of re-emitted light accompanying the absorption of photosynthetically active radiation (PAR) by excited chlorophyll-a molecules in the spectral range from 650 to 800 nm[31]. The first approved global mission designed explicitly for SIF measurement of terrestrial vegetation, the FLuorescence EXplorer (FLEX), was selected as the eighth Earth Explorer mission of the European Space Agency and will be launched in 2025[32]. The global SIF datasets currently used are estimated from atmospheric sensors because they have the required spectral resolution and signal-to-noise ratio (details of the sensors are given in Table 1). However, the existing SIF records have long been limited by their low spatial resolution and sparseness in data acquisition. For instance, the Global Ozone Monitoring Experiment-2 (GOME-2)[24] and the SCanning Imaging Absorption SpectroMeter for Atmospheric CHartographY (SCIAMACHY)[26] provide spatially continuous coverage of SIF

[1]State Key Laboratory of Hydroscience and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing, 100084, China. [2]The Key Laboratory of Ecological Protection and High Quality Development in the Upper Yellow River, Qinghai Province, China. [3]State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining, Qinghai, 810016, China. [4]Chinese Research Academy of Environmental Sciences, Beijing, 100012, China. [5]National Joint Research Center for Yangtze River Conservation, Beijing, 100012, China. ✉e-mail: zhangs2019@tsinghua.edu.cn

Sensor	GOSAT	GOME-2	SCIAMACHY	OCO-2	TanSat	TROPOMI	FLEX
Launch time	2009/6	2007/1	2002/3	2014/7	2016/12	2017/10	2025
Overpass time	13:30	9:30	9:30	13:15	13:00	13:30	10:00
Spatial coverage	sparse	continuous	continuous	sparse	sparse	continuous	continuous
Footprint size	10 km	40 × 80 km	30 × 240 km	1.5 × 2.25 km	2 × 2 km	3.5 × 5.5 km	300 m
Temporal resolution	3 days	1.5 days	6 days	16 days	16 days	1 day	27days

Table 1. Space-borne instruments currently in orbit enabling SIF estimation.

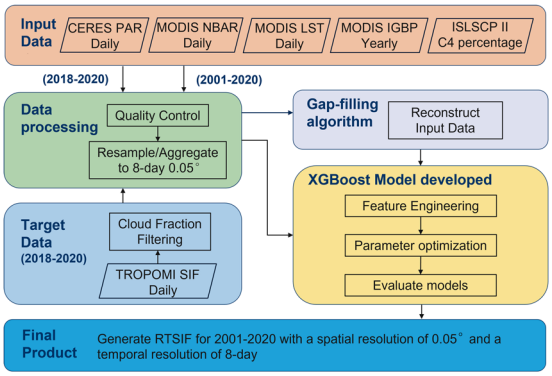

Fig. 1 The workflow to generate RTSIF.

but with large footprint size (hence low spatial resolution, Table 1). Conversely, the Greenhouse Gases Observing Satellite (GOSAT)[23] and the Orbiting Carbon Observatory-2 (OCO-2)[25] offer much smaller footprint size, but with sparse and thus spatially discontinuous measurements.

The above dilemma is partially addressed by the TROPOspheric Monitoring Instrument (TROPOMI) on the Copernicus Sentinel-5P mission thanks to the significantly increased spatiotemporal resolution and data coverage[27]. TROPOMI has almost global coverage (except for small gaps between footprints) and high spatial resolution (3.5 km × 5.5 km at nadir)[33]. Compared with the earlier missions, TROPOMI has a daily revisit time to provide a significant increase in the number of clear-sky measurements. However, Sentinel-5P was launched in October 2017, and the TROPOMI SIF data are only available since April 2018, limiting its use for long-term applications.

This study uses machine learning algorithms to reconstruct TROPOMI SIF (RTSIF) for a longer period to alleviate the issue above. RTSIF is generated based on the Caltech TROPOMI SIF data[27], the nadir bidirectional reflectance distribution adjusted reflectance (NBAR)[34], land surface temperature (LST)[35], and land cover data[36] from the Moderate Resolution Imaging Spectroradiometer (MODIS), the PAR data[37] from the Earth's Radiant Energy System (CERES), and the vegetation type data[38] from the International Satellite Land Surface Climatology Project, Initiative II (ISLSCP II). This dataset extends the time coverage of the TROPOMI SIF data and provides a long-term, high-resolution, and global SIF record. RTSIF is in good agreement with TROPOMI SIF and has been evaluated against the GOME-2 and OCO-2 SIF. We further demonstrate the consistency between RTSIF and tower measured SIF and GPP. The proposed dataset provides a new dataset for SIF evaluation and could benefit related ecosystem, carbon cycle, and net-zero carbon emission studies.

Methods

Framework overview. Figure 1 illustrates the overall framework used to generate RTSIF. Based on the LUE concept, SIF can be expressed as follows according to Zhang *et al.*[39] and Zhang *et al.*[40]:

$$SIF = PAR \times fPAR_{chl} \times FE \quad (1)$$

where $fPAR_{chl}$ is the fraction of PAR absorbed by chlorophyll ($APAR_{chl}$) and FE is the fluorescence efficiency. Since SIF originates from the solar energy absorbed by chlorophyll-a molecules[41], it is highly correlated with $APAR_{chl}$, the product of $fPAR_{chl}$ and PAR[42–44]. Previous studies have shown that $fPAR_{chl}$ can be estimated from surface reflectance using radiative transfer models[45], and thus PAR and surface reflectance have been widely used to reconstruct SIF[46–51]. Previous studies have also shown that the high correlation between SIF and $APAR_{chl}$ is limited to unstressed conditions[52], while drought and other environmental stresses can affect FE. LST can be used as a proxy of thermal stress in predictive models of SIF[53–56]. In this study, we further consider that including biome type may improve the prediction accuracy of the SIF model given the plant structural and physiological differences in different biomes and different photosynthetic pathways in C3 and C4 plants. We finally selected surface reflectance, PAR, LST, land cover, and C3/C4 fraction as input variables for the RTSIF modeling.

Data source	Dataset	Derived variables	Spatial resolution	Temporal resolution	Available at
TROPOMI	TROPOMI SIF	SIF	ungridded	Daily	ftp://fluo.gps.caltech.edu/data/tropomi/
MODIS	MOD11C1	LST	0.05° × 0.05°	Daily	https://doi.org/10.5067/MODIS/MOD11C1.006[85]
	MCD12C1	Land cover	0.05° × 0.05°	Yearly	https://doi.org/10.5067/MODIS/MCD12C1.006[86]
	MCD43C4	NBAR	0.05° × 0.05°	Daily	https://doi.org/10.5067/MODIS/MCD43C4.006[87]
CERES	SYNI	PAR	1° × 1°	Daily	https://doi.org/10.5067/Terra+Aqua/CERES/SYN1degDay_L3.004A[88]
ISLSCP II	C4 vegetation percentage map	C4 percentage	1° × 1°	invariant	https://doi.org/10.3334/ORNLDAAC/932[82]

Table 2. Datasets used in developing the machine learning model for RTSIF and their characteristics.

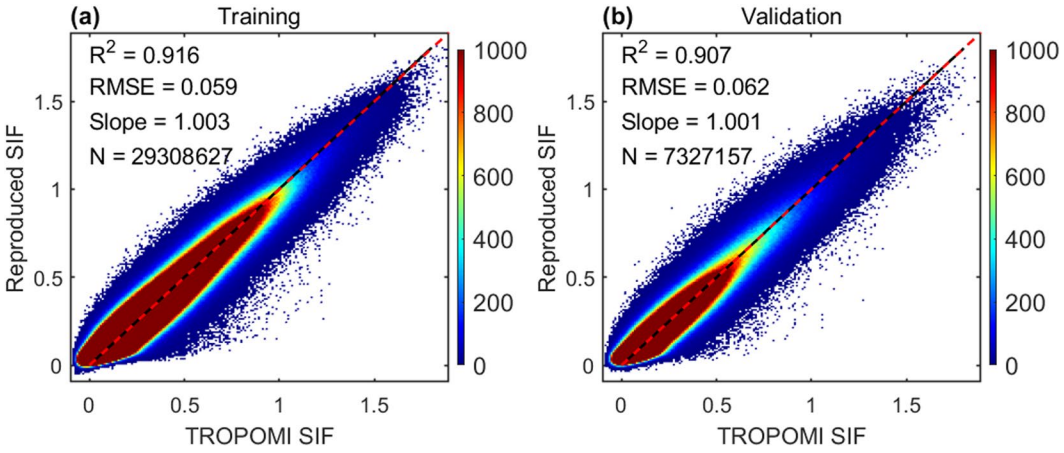

Fig. 2 Performance of the XGBoost model in reproducing TROPOMI SIF over the training and testing data. The shading color represents the density of the scatterplot. Black lines represent the regression slope, and the red dotted lines represent the 1:1 line. The regression is forced to pass the origin. All values are in the unit of $mWm^{-2}nm^{-1}sr^{-1}$.

Data Sets. We used multiple datasets as input to generate RTSIF. All the datasets used are summarized in Table 2 and described in detail as follows.

The Caltech TROPOMI SIF data between March 2018 and December 2020 were used for model training and evaluation. We followed the filtering scheme in the original reference[27] to retain daily average clear-sky SIF data with cloud fractions less than 0.1, and excluded the data with a sun zenith angle (SZA) greater than 60° and a view zenith angle (VZA) greater than 70°. The ungridded data through the filtering scheme were aggregated to 0.05° grids at an 8-day resolution, the grid size of which was close to the footprint size of the TROPOMI SIF data. Averaging the multiple observations reduces the uncertainty in the original SIF retrievals by $\sqrt{n}$ (n is the effective number of observations in the grid cell)[25]. For each 0.05° grid, only the SIF footprint covering the center of the grid was recorded as valid retrievals, and the SIF values were calculated only when more than four valid retrievals were included. We used the SIF values at 740 nm from the 743–758 nm retrieval window, which is optimal for high retrieval precision and low sensitivity to clouds[33].

Ancillary input data including the MODIS land products, the CERES products, and the ISLSCP II products were used to generate RTSIF. The MODIS products included LST (MOD11C1[35]), land cover (MCD12C1[36]), and seven bands for nadir bidirectional reflectance distribution adjusted reflectance (NBAR; MCD43C4[34]). To reduce the uncertainty in the SIF modeling, only high-quality MOD11C1 (QA < 2) and MCD43C4 (QA < 2) data were used and aggregated to an 8-day average. Gap-filling and smoothing algorithms were used to reconstruct the 8-day MOD11C1 and MCD43C4 data[57] and replace the poor observations caused by bad atmospheric conditions. We used an updated land cover map (MCD12C1) for each year. PAR data (SYNI PAR[37]) from the CERES products were used, aggregated to 8-day, and interpolated to 0.05° using bilinear interpolation. The ISLSCP II C4 vegetation map was used for natural C4 vegetation distribution[38], assuming that all the vegetation types within each 1° grid cell shared the same C3/C4 ratio.

Data-Driven approach. Extreme Gradient Boosting (XGBoost) is an enhanced version of the machine learning algorithm named Gradient Boosted Decision Tree (GBDT)[58]. It constructs enhanced trees that can handle complex nonlinear relationships[59,60]. As a boosting algorithm, XGBoost consists of multiple decision trees, each of which is trained with the residual error of the predicted result from the previous decision tree, and finally iterates the results of all the decision trees before producing the final result. Compared with other traditional GBDT algorithms that only use first-order derivatives, XGBoost performs a second-order Taylor expansion on the loss function between computed results and actual observations to accelerate the convergence of the model

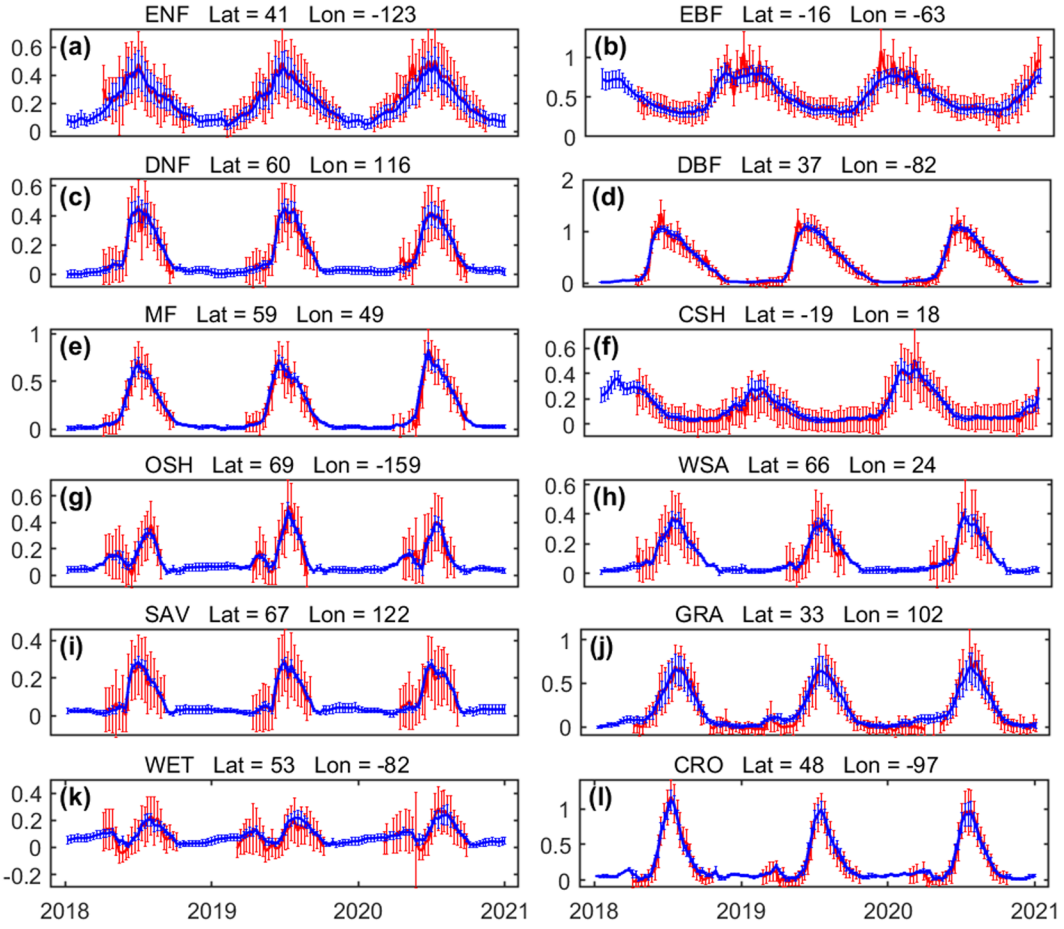

Fig. 3 Time series of RTSIF and TROPOMI SIF for selected 1° grid cells. All the samples from the training data and the testing data were used. The red line represents TROPOMI SIF, and the blue line represents RTSIF. The error bars represent the standard deviation of the TROPOMI SIF footprint and RTSIF used to generate 1° grid. The MODIS MOD12C1 land cover dataset was used to select these example grid cells. All the values are in the unit of $mWm^{-2}nm^{-1}sr^{-1}$.

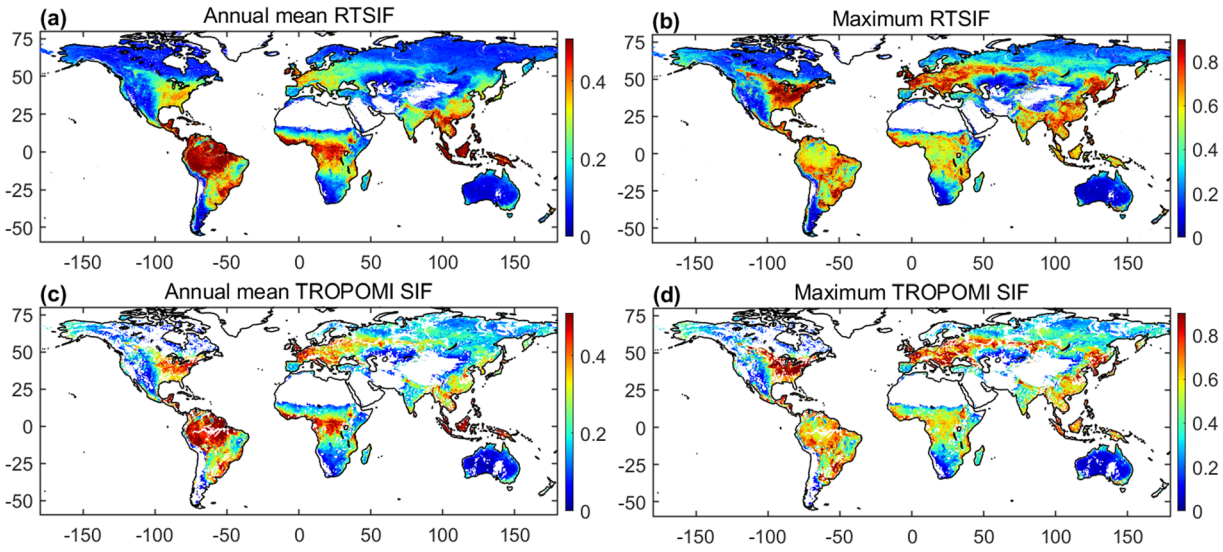

Fig. 4 Spatial pattern of average and maximum (90^{th} percentile) daily values for RTSIF (**a** and **b**) and TROPOMI SIF (**c** and **d**) in 2019. All the values are in units of $mWm^{-2}nm^{-1}sr^{-1}$.

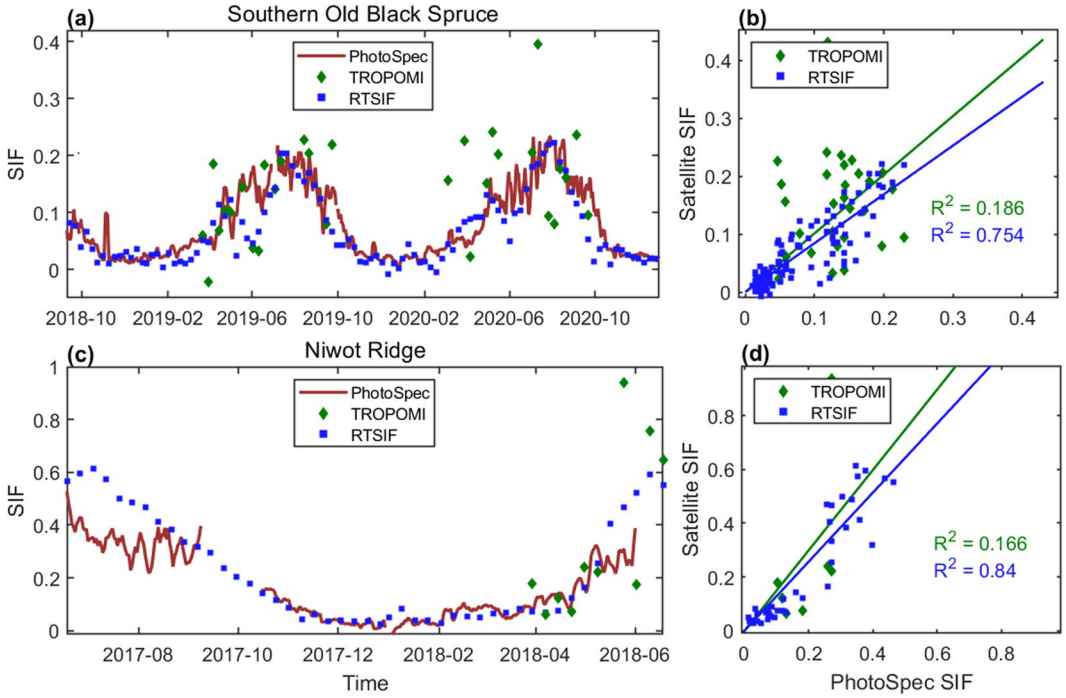

Fig. 5 Comparison between RTSIF, TROPOMI SIF, and tower-based SIF measurements. The red line is plotted using the daily average SIF collected by PhotoSpec, presented as 5-day moving averages. The green dots represent the average of the TROPOMI SIF footprint aggregated to 8-day. The blue dot represents the RTSIF value of the pixel where the site is located. The comparison between TROPOMI SIF and RTSIF with the tower-based SIF was based on 8-day averages (**b** and **d**). All the values are in the unit of $mWm^{-2}nm^{-1}sr^{-1}$.

during training and provide higher efficiency in finding the optimal solution. In addition, XGBoost has a regularization term to control the complexity of the model, which can effectively avoid overfitting. The TROPOMI SIF and the input variables constitute a dataset containing a large number of data samples (about 36 million). The current machine learning algorithms have difficulties in processing large datasets using existing packages[61], while XGBoost employs software and hardware optimization techniques to make it possible to process tens of millions of training data. In this study, XGBoost was implemented using the Python library XGBoost (https://github.com/dmlc/xgboost). Before training, each variable was standardized by its mean and deviation. We split the data into the training group (80%) and the testing group (20%). Many hyperparameters in XGBoost affect the model performance, and a grid search was performed for the hyperparameters with 10-fold cross-validation to find the best combination of the parameters based on the Root Mean Square Error (RMSE) metric[62]. The optimized hyperparameters are compiled in Supplementary Table S1.

Data Records

Our long-term global SIF dataset, RTSIF, is available at https://doi.org/10.6084/m9.figshare.19336346.v2[63]. The data record contains global RTSIF data from January 2001 to December 2020 at a 0.05°/8-day resolution. There are 46 GeoTiff files per year, one for each 8-day period. The unit is $mWm^{-2}nm^{-1}sr^{-1}$. The file name RTSIF_<YYYY>-<MM>-<DD>.tif provides information on the year, month, and start date of the 8-day period. Considering that deserts and glaciers have no vegetation, those pixels are flagged.

Technical Validation

Model validation. We tested the performance of the XGBoost model with the optimal hyperparameters. The model reproduces the TROPOMI SIF with a determination coefficient R^2 of 0.916, a RMSE of 0.059 $mWm^{-2}nm^{-1}sr^{-1}$ during training, and an R^2 of 0.907, and an RMSE of 0.062 $mWm^{-2}nm^{-1}sr^{-1}$ during testing (Fig. 2), suggesting that our optimized XGBoost model is not overfitting. The slope of the fit between the reproduced and observed SIF values is close to 1, indicating that there is no systematic discrepancy. We also investigated the performance of the model for each land cover type defined in the MCD12C1 dataset. For most land cover types, the reproduced and observed TROPOMI SIF values have R^2 values over 0.8 (Table. S2).

We compared RTSIF and TROPOMI SIF for 1° selected grid cells representative of the 12 vegetated biomes (locations shown in Fig. S1b). RTSIF can accurately capture seasonal and interannual variations in TROPOMI SIF for most biome types. The standard deviation in the RTSIF data is typically smaller than that in the originally retrieved TROPOMI SIF, indicating reduced noise in the RTSIF dataset. RTSIF also fills the gaps where no TROPOMI SIF data are available (Fig 3).

To further illustrate the spatial variation of RTSIF, we show the global mean and maximum values of RTSIF in 2019 (Fig. 4). The average daily SIF has the highest values in the tropics, intermediate values in southern

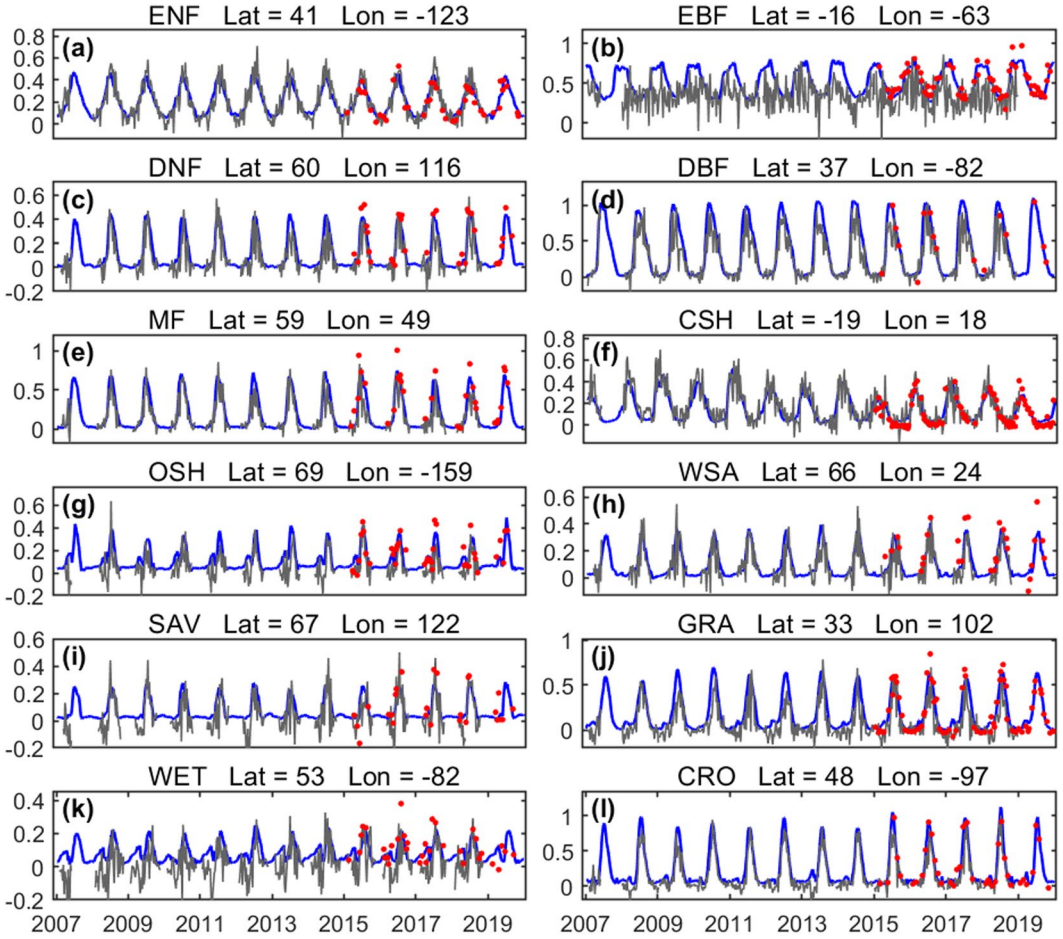

Fig. 6 Time series of RTSIF, OCO-2 SIF, and GOME-2 SIF for selected regions. The blue line represents the RTSIF and the gray line represents the GOME-2 SIF. The red dots represent the OCO-2 SIF measurements which are not continuous. All the values are in the unit of $mWm^{-2}nm^{-1}sr^{-1}$.

China, central Europe, and the eastern United States, and the lowest values in barren regions. The maximum daily SIF is found mainly in the North American corn belt, South Asia, central Europe, and tropical rainforests, consistent with the high productivity in these regions[64]. The annual average SIF and the maximum daily SIF show similar spatial patterns as those in TROPOMI SIF.

Comparison of RTSIF with tower-based SIF. Recently several studies have reported SIF measurements from ground towers[65–69], providing a valuable opportunity to verify the temporal variation observed in RTSIF. We compared the tower-based SIF observations at the Southern Old Black Spruce[65] (53.98°N, 105.12°W) and the Niwot Ridge sites[69] (40.03°N, 105.55°W) with RTSIF. The ground tower SIF data were collected using a scanning spectrometer (PhotoSpec) for far-red (745–758 nm) SIF and retrieved by the singular value decomposition (SVD) method scaled to 750 nm. For comparison, we scaled the ground SIF to 740 nm using a wavelength scaling factor of 1.17 and aggregated the hourly data to the daily timescale[51]. Our results show good agreement between RTSIF and tower-based SIF (Fig. 5), with an R^2 of 0.754 at the Southern Old Black Spruce site and an R^2 of 0.84 at the Niwot Ridge site. Although mismatches were found between RTSIF and SIF measurements at the Niwot Ridge Site, which is possibly due to inconsistency between tower footprint and RTSIF pixel size and landscape heterogeneity. RTSIF captures the seasonal changes of the tower-based SIF at both sites well reprocudes, successfully locating the timing of spring onset and autumn senescence.

Comparison of RTSIF with other SIF products. We further compared the RTSIF dataset with the retrievals of OCO-2 SIF and GOME-2 SIF[24,25] (Fig. 6). OCO-2 SIF was retrieved at 757 nm, and a wavelength scale factor of 1.56 was required to convert the wavelength from OCO-2 (757 nm) to 740 nm[27]. We used OCO-2 (2015–2020) and GOME-2 (2007–2019) SIF data and aggregated all the clear-sky and good-quality measurements to 1° with an 8-day temporal resolution by using the same cloud filtering threshold (less than 0.1). All the data show similar seasonal variations in the most selected areas of typical biomes except over broad-leaf evergreen forests. The disagreement is mostly due to the low signal-to-noise ratio of GOME-2, which led the GOME-2 SIF cannot capture seasonal changes (blue lines in Fig. 6b)[50]. In addition, the large footprint of GOME-2 SIF makes it more sensitive to cloud contamination in subpixels leading to underestimated SIF values[70]. Notably, GOME-2 SIF showed large fluctuations (even negative values) during the non-growing season at some sites caused by

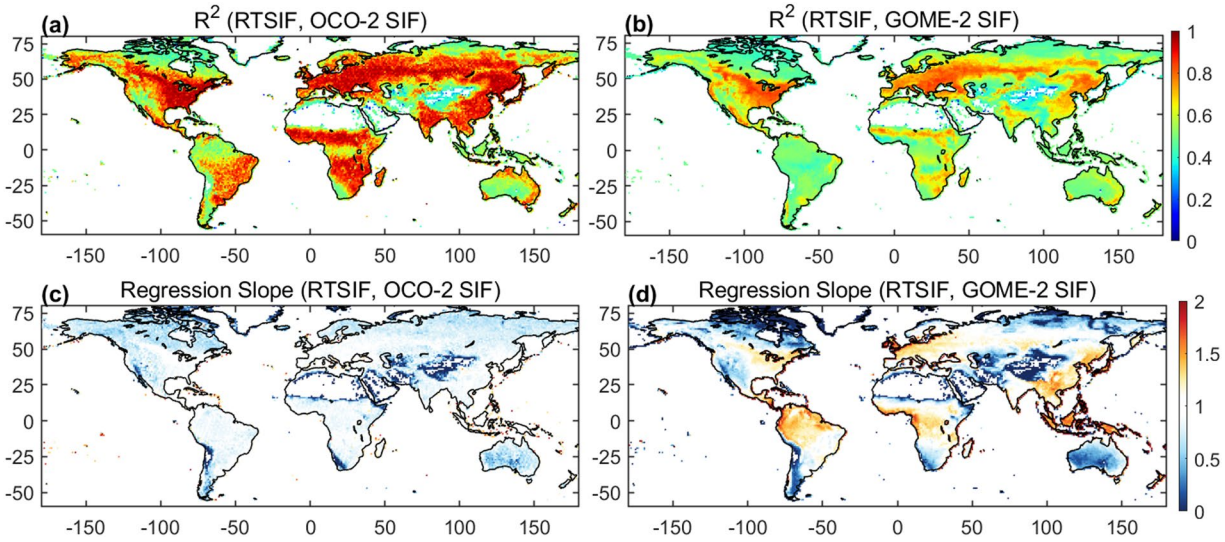

Fig. 7 Comparison of the RTSIF, OCO-2 SIF, and GOME-2 SIF datasets. R^2 and regression slope for RTSIF versus OCO-2 SIF (**a** and **c**) and GOME-2 SIF (**b** and **d**). The regression is forced to pass the origin. The white area represents the barren region. The data between 2015–2020 (OCO-2 SIF) and 2007–2019 (GOME-2 SIF) were used for comparison.

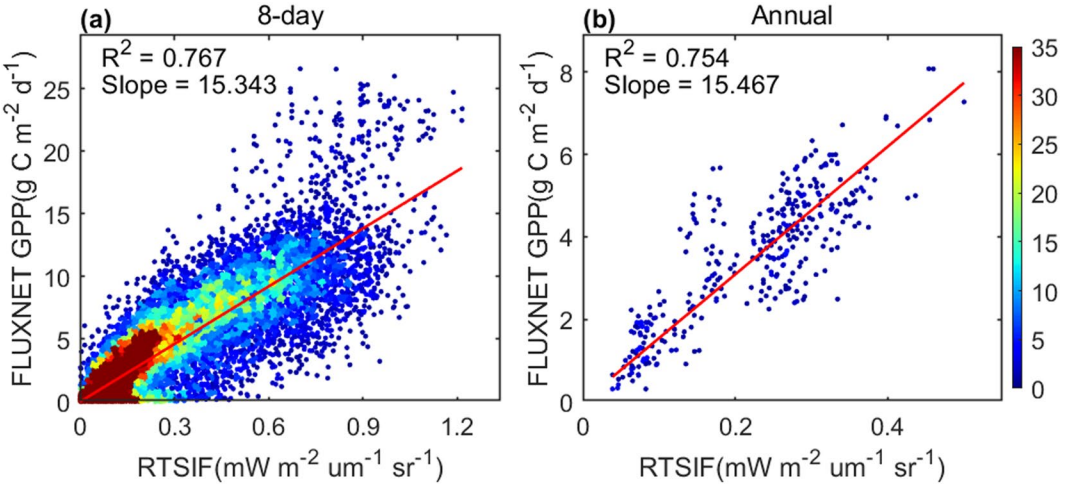

Fig. 8 Relationship between RTSIF and FLUXNET GPP at a 8-day timescale (**a**) and the annual scale (**b**). The shading color represents the density of the scatterplot. The regression is forced to pass the origin.

snow contamination (Fig. 6k,i)[71,72]. RTSIF agrees well with OCO-2 SIF as the training TROPOMI SIF with high signal-to-noise ratios and spatial resolutions has demonstrated agreement with OCO-2 SIF[27] and fills the gap where OCO-2 SIF is discontinuous both spatially and temporally.

At the global scale, RTSIF shows good agreement with OCO-2 SIF and GOME-2 SIF in most regions with an $R^2 > 0.7$ (Fig. 7a,b). The R^2 between RTSIF and OCO-2 SIF is higher than that between RTSIF and GOME-2 SIF due to the reasons mentioned in the previous paragraph. The regression slopes of RTSIF with OCO-2 SIF and GOME-2 SIF are close to 1. However, in regions with persistent cloud cover (e.g., tropical rainforests and Western Europe), the regression slope of RTSIF with GOME-2 SIF is larger than 1 (Fig. 7d), suggesting that GOME-2 SIF is underestimated due to cloud cover in these regions. Although we filter the GOME-2 SIF data with a cloud fraction of 0.1, the large footprint size in GOME-2 SIF (~40 km) makes it impossible to remove all the subpixel cloud contamination[51]. Because our model is trained with clear-sky data (although these areas usually have high cloud coverage, there are still a large amount of clear-sky data), RTSIF is less affected by cloud cover. In addition, there is no significant increase in noise in the TROPOMI SIF due to the South Atlantic Anomaly (SSA)[73], and RTSIF should reproduce SIF values for parts of South America with higher accuracy than OCO-2 and GOME-2 SIF. Overall it can be concluded that RTSIF provides consistent and spatially continuous SIF estimates compared to the other two products.

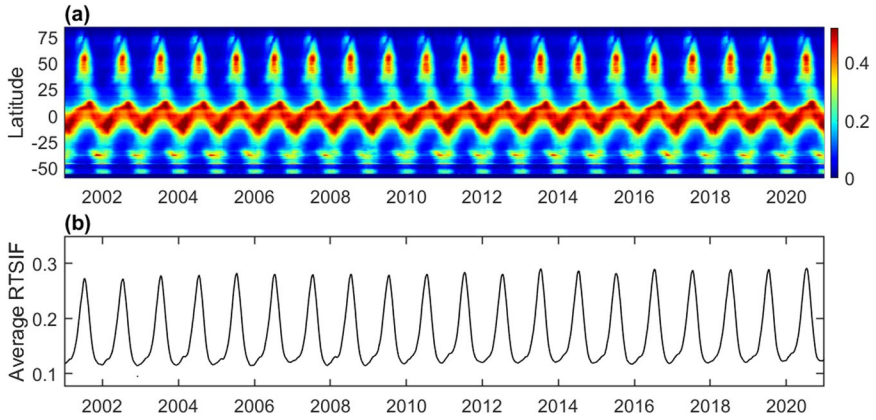

Fig. 9 Seasonal and interannual variation of daily SIF. (**a**) Latitudinal averages of SIF for each 8-day period. (**b**) The global average of SIF for each 8-day period. All the values are in units of $mWm^{-2}nm^{-1}sr^{-1}$.

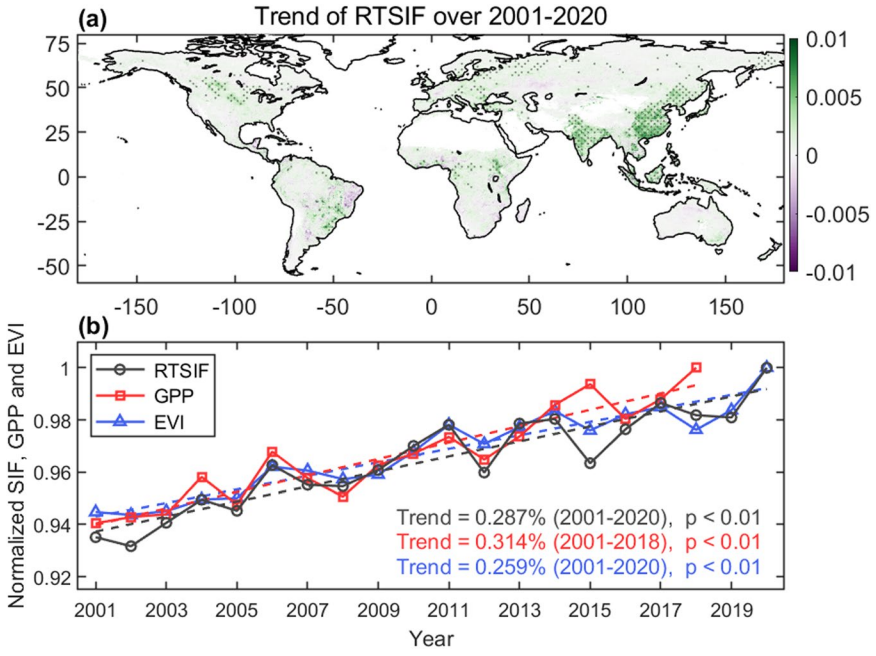

Fig. 10 (**a**) Spatial distribution of the trends of annual average RTSIF during 2001–2020. Sen's slope estimator is used to calculate the trend. Dots represent the locations where the trend is significant ($p < 0.05$) through a Mann–Kendall test. All the values are in the unit of $mWm^{-2}nm^{-1}sr^{-1}yr^{-1}$. (**b**) Inter-annual variations and trends of normalized global average RTSIF, EVI, and VPM GPP from 2001 to 2020.

Comparison of SIF with Tower GPP estimates. To further evaluate the RTSIF product, we explored the relationship between RTSIF and GPP using GPP estimates from the FLUXNET 2015 Tier 1 dataset[74]. The daily GPP estimates were calculated using the average of GPP estimates from the nighttime (GPP_NT_VUT_REF) and daytime (GPP_DT_VUT_REF) partitioning methods[75,76]. Only the GPP estimates with more than four consecutive days of high quality (QA = 1) measurements were used when aggregated to an 8-day resolution. Considering the inconsistency between the flux tower footprint and the RTSIF pixel size, we only selected sites where the biome type in the RTSIF grid is homogeneous and the same as that at the flux tower site. We finally collected 76 sites from 171 flux sites with more than two years of GPP data. The detailed descriptions of these flux tower sites, including site code, location, and biome type are provided in Supplementary Table S3. There is a linear relationship between RTSIF and GPP in both 8-day and annual timescale (Fig. 8), indicating that RTSIF is tightly related to GPP.

To investigate whether the SIF-GPP relationship is universal for different biomes, we compared the relationship between biome-specific RTSIF and GPP (Table S4 and Fig. S2). RTSIF was in good agreement with GPP for almost all biomes at the 8-day timescale, indicating strong SIF-GPP correlations for different biomes. The agreement between RTSIF and GPP was good at the annual scale in mixed forests, woody savannas, savannas, and grasslands. GPP and RTSIF showed an overall regression slope of 15.343 ($g\ C\ m^{-2}\ day^{-1}/mWm^{-2}\ nm^{-1}\ sr^{-1}$)

in the 8-day timescale and 15.467 (g C m^{-2} day^{-1}/mWm^{-2} nm^{-1} sr^{-1}) in the annual timescale, with different biomes showing significant differences. Specifically, a larger slope was found in evergreen needleleaf forests due to their distinct canopy structure, resulting in stronger reabsorption of SIF.

Temporal patterns of the long-term RTSIF. We further investigated the seasonal variation of RTSIF. Fig. 9a demonstrates the seasonal variation of RTSIF in different latitudes. The northern and southern hemispheres show clear seasonal variations with repeated high values in summer. On the other hand, the tropical regions show persistently high SIF values across seasons. Globally averaged SIF shows clear seasonality (Fig. 9b).

Between 2001 and 2020, the annual average of SIF increased in China and India, and decreased in parts of the tropical rainforest (southern Amazonia and eastern Brazil), consistent with findings in previous studies[77–80] (Fig. 10a). The global average annual RTSIF over the last 20 years has a significant positive trend (0.3% yr^{-1}, $p < 0.01$), consistent with those observed in other reconstructed SIF products[47,50] (Fig. S3). The interannual variability and positive trend of RTSIF are similar to those observed for MODIS EVI (enhanced vegetation index)[81] and VPM GPP[57], but RTSIF shows larger interannual variabilities (Fig. 10b).

Code availability

The code for generating the RTSIF is available at https://github.com/chen-xingan/Reconstruct-TROPOMI-SIF.git.

Received: 19 October 2021; Accepted: 4 July 2022;
Published online: 20 July 2022

References

1. Canadell, J. G. *et al*. Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks. *Proc. Natl. Acad. Sci.* **104**, 18866–18870 (2007).
2. Beer, C. *et al*. Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate. *Science* **329**, 834–838 (2010).
3. Park, T. *et al*. Changes in timing of seasonal peak photosynthetic activity in northern ecosystems. *Global. Change. Biol.* **25**, 2382–2395 (2019).
4. Wang, T. *et al*. Emerging negative impact of warming on summer carbon uptake in northern ecosystems. *Nat. Commun.* **9**, 5391 (2018).
5. Farquhar, G. D., Von Caemmerer, S. & Berry, J. A. A biochemical model of photosynthetic CO_2 assimilation in leaves of C3 species. *Planta* **149**, 78–90 (1980).
6. Chen, J. M. *et al*. Effects of foliage clumping on the estimation of global terrestrial gross primary productivity. *Global. Biogeochem. Cy* **26**, GB1019 (2012).
7. De Pury, D. G. G. & Farquhar, G. D. Simple scaling of photosynthesis from leaves to canopies without the errors of big-leaf models. *Plant Cell Environ.* **20**, 537–557 (1997).
8. Zhang, Y. *et al*. Development of a coupled carbon and water model for estimating global gross primary productivity and evapotranspiration based on eddy flux and remote sensing data. *Agr. Forest. Meteorol.* **223**, 116–131 (2016).
9. Monteith, J. L. Climate and the efficiency of crop production in Britain. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **281**, 277–294 (1977).
10. Running, S. W. *et al*. A continuous satellite-derived measure of global terrestrial primary production. *Bioscience.* **54**, 547–560 (2004).
11. Yuan, W. *et al*. Global estimates of evapotranspiration and gross primary production based on MODIS and global meteorology data. *Remote Sens. Environ.* **114**, 1416–1431 (2010).
12. Ruimy, A., Dedieu, G. & Saugier, B. TURC: A diagnostic model of continental gross primary productivity and net primary productivity. *Global. Biogeochem. Cy* **10**, 269–285 (1996).
13. Jung, M. *et al*. The FLUXCOM ensemble of global land-atmosphere energy fluxes. *Sci. Data* **6**, 190076 (2019).
14. Bodesheim, P., Jung, M., Gans, F., Mahecha, M. D. & Reichstein, M. Upscaled diurnal cycles of land–atmosphere fluxes: a new global half-hourly data product. *Earth Syst. Sci. Data* **10**, 1327–1365 (2018).
15. Joiner, J. *et al*. Estimation of Terrestrial Global Gross Primary Production (GPP) with Satellite Data-Driven Models and Eddy Covariance Flux Data. *Remote Sens.* **10**, 1346 (2018).
16. Xiao, J. *et al*. Data-driven diagnostics of terrestrial carbon dynamics over North America. *Agr. Forest. Meteorol.* **197**, 142–157 (2014).
17. Ichii, K. *et al*. New data-driven estimation of terrestrial CO_2 fluxes in Asia using a standardized database of eddy covariance measurements, remote sensing data, and support vector regression. *J. Geophys. Res. Biogeosci.* **122**, 767–795 (2017).
18. Cai, W. *et al*. Improved estimations of gross primary production using satellite-derived photosynthetically active radiation. *J. Geophys. Res. Biogeosci.* **119**, 110–123 (2014).
19. Ma, J., Yan, X., Dong, W. & Chou, J. Gross primary production of global forest ecosystems has been overestimated. *Sci. Rep.* **5**, 10820 (2015).
20. Cai, W. *et al*. Large Differences in Terrestrial Vegetation Production Derived from Satellite-Based Light Use Efficiency Models. *Remote Sens.* **6**, 8945–8965 (2014).
21. Jung, M. *et al*. Uncertainties of modeling gross primary productivity over Europe: A systematic study on the effects of using different drivers and terrestrial biosphere models. *Global. Biogeochem. Cy* **21**, GB4021 (2007).
22. Yuan, W. *et al*. Global comparison of light use efficiency models for simulating terrestrial vegetation gross primary production based on the LaThuile database. *Agr. Forest. Meteorol.* **192-193**, 108–120 (2014).
23. Frankenberg, C. *et al*. New global observations of the terrestrial carbon cycle from GOSAT: Patterns of plant fluorescence with gross primary productivity. *Geophys. Res. Lett.* **38**, L17706 (2011).
24. Joiner, J. *et al*. Global monitoring of terrestrial chlorophyll fluorescence from moderate-spectral-resolution near-infrared satellite measurements: methodology, simulations, and application to GOME-2. *Atmos Meas Tech* **6**, 2803–2823 (2013).
25. Frankenberg, C. *et al*. Prospects for chlorophyll fluorescence remote sensing from the Orbiting Carbon Observatory-2. *Remote Sens. Environ.* **147**, 1–12 (2014).
26. Joiner, J. *et al*. Filling-in of near-infrared solar lines by terrestrial fluorescence and other geophysical effects: simulations and space-based observations from SCIAMACHY and GOSAT. *Atmos Meas Tech* **5**, 809–829 (2012).
27. Köhler, P. *et al*. Global Retrievals of Solar-Induced Chlorophyll Fluorescence With TROPOMI: First Results and Intersensor Comparison to OCO-2. *Geophys. Res. Lett.* **45**, 10,456–410,463 (2018).
28. Joiner, J. *et al*. First observations of global and seasonal terrestrial chlorophyll fluorescence from space. *Biogeosciences* **8**, 637–651 (2011).
29. Guanter, L. *et al*. Retrieval and global assessment of terrestrial chlorophyll fluorescence from GOSAT space measurements. *Remote Sens. Environ.* **121**, 236–251 (2012).
30. Du, S. *et al*. Retrieval of global terrestrial solar-induced chlorophyll fluorescence from TanSat satellite. *Sci. Bull.* **63**, 1502–1512 (2018).
31. Baker, N. R. Chlorophyll fluorescence: a probe of photosynthesis *in vivo*. *Annu. Rev. Plant Biol.* **59**, 89–113 (2008).
32. Drusch, M. *et al*. The FLuorescence EXplorer Mission Concept—ESA's Earth Explorer 8. *Ieee. T. Geosci. Remote* **55**, 1273–1284 (2017).

33. Guanter, L. *et al*. The TROPOSIF global sun-induced fluorescence dataset from the Sentinel-5P TROPOMI mission. *Earth Syst. Sci. Data*, **13**, 5423–5440 (2021).
34. Roesch, A. Use of Moderate-Resolution Imaging Spectroradiometer bidirectional reflectance distribution function products to enhance simulated surface albedos. *J. Geophys. Res.* **109** (2004).
35. Wan, Z. New refinements and validation of the collection-6 MODIS land-surface temperature/emissivity product. *Remote Sens. Environ.* **140**, 36–45 (2014).
36. Sulla-Menashe, D., Gray, J. M., Abercrombie, S. P. & Friedl, M. A. Hierarchical mapping of annual global land cover 2001 to present: The MODIS Collection 6 Land Cover product. *Remote Sens. Environ.* **222**, 183–194 (2019).
37. Su, W., Charlock, T. P., Rose, F. G. & Rutan, D. Photosynthetically active radiation from Clouds and the Earth's Radiant Energy System (CERES) products. *J. Geophys. Res.* **112** (2007).
38. Still, C. J., Berry, J. A., Collatz, G. J. & Defries, R. S. Global distribution of C3and C4vegetation: Carbon cycle implications. *Global. Biogeochem. Cy* **17**, 6-1-6-14 (2003).
39. Zhang, Y. *et al*. Spatio-temporal convergence of maximum daily light-use efficiency based on radiation absorption by canopy chlorophyll. *Geophys. Res. Lett.* **45**, 3508–3519 (2018).
40. Zhang, Z. *et al*. The potential of satellite FPAR product for GPP estimation: An indirect evaluation using solar-induced chlorophyll fluorescence. *Remote Sens. Environ.* **240**, 111686 (2020).
41. Baker, N. R. Chlorophyll Fluorescence: A Probe of Photosynthesis *In Vivo*. *Annu. Rev. Plant. Biol.* **59**, 89–113 (2008).
42. Du, S., Liu, L., Liu, X. & Hu, J. Response of canopy solar-induced chlorophyll fluorescence to the absorbed photosynthetically active radiation absorbed by chlorophyll. *Remote Sens.* **9**, 911 (2017).
43. Rossini, M. *et al*. Analysis of Red and Far-Red Sun-Induced Chlorophyll Fluorescence and Their Ratio in Different Canopies Based on Observed and Modeled Data. *Remote Sens.* **8**, 412 (2016).
44. Verrelst, J. *et al*. Global sensitivity analysis of the SCOPE model: What drives simulated canopy-leaving sun-induced fluorescence? *Remote Sens. Environ.* **166**, 8–21 (2015).
45. Zhang, Q. *et al*. Estimating light absorption by chlorophyll, leaf and canopy in a deciduous broadleaf forest using MODIS data and a radiative transfer model. *Remote Sens. Environ.* **99**, 357–371 (2005).
46. Zhang, Y., Joiner, J., Alemohammad, S. H., Zhou, S. & Gentine, P. A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks. *Biogeosciences* **15**, 5779–5800 (2018).
47. Li, X. & Xiao, J. A Global, 0.05-Degree Product of Solar-Induced Chlorophyll Fluorescence Derived from OCO-2, MODIS, and Reanalysis Data. *Remote Sens.* **11**, 517 (2019).
48. Yu, L., Wen, J., Chang, C. Y., Frankenberg, C. & Sun, Y. High-Resolution Global Contiguous SIF of OCO-2. *Geophys. Res. Lett.* **46**, 1449–1458 (2019).
49. Ma, Y., Liu, L., Chen, R., Du, S. & Liu, X. Generation of a Global Spatially Continuous TanSat Solar-Induced Chlorophyll Fluorescence Product by Considering the Impact of the Solar Radiation Intensity. *Remote Sens.* **12**, 2167 (2020).
50. Gentine, P. & Alemohammad, S. H. Reconstructed Solar-Induced Fluorescence: A Machine Learning Vegetation Product Based on MODIS Surface Reflectance to Reproduce GOME-2 Solar-Induced Fluorescence. *Geophys. Res. Lett.* **45**, 3136–3146 (2018).
51. Wen, J. *et al*. A framework for harmonizing multiple satellite instruments to generate a long-term global high spatial-resolution solar-induced chlorophyll fluorescence (SIF). *Remote Sens. Environ.* **239**, 111644 (2020).
52. Yang, X. *et al*. Solar-induced chlorophyll fluorescence that correlates with canopy photosynthesis on diurnal and seasonal scales in a temperate deciduous forest. *Geophys. Res. Lett.* **42**, 2977–2987 (2015).
53. Hain, C. R., Crow, W. T., Mecikalski, J. R., Anderson, M. C. & Holmes, T. An intercomparison of available soil moisture estimates from thermal infrared and passive microwave remote sensing and land surface modeling. *J. Geophys. Res.* **116**, D15107 (2011).
54. Anderson, M. C., Norman, J. M., Mecikalski, J. R., Otkin, J. A. & Kustas, W. P. A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 2. Surface moisture climatology. *J. Geophys. Res.* **112**, D11112 (2007).
55. Scherrer, D., Bader, M. K.-F. & Körner, C. Drought-sensitivity ranking of deciduous tree species based on thermal imaging of forest canopies. *Agr. Forest. Meteorol.* **151**, 1632–1640 (2011).
56. Duveiller, G. *et al*. A spatially downscaled sun-induced fluorescence global product for enhanced monitoring of vegetation productivity. *Earth Syst. Sci. Data* **12**, 1101–1116 (2020).
57. Zhang, Y. *et al*. A global moderate resolution dataset of gross primary production of vegetation for 2000–2016. *Sci. Data* **4**, 170165 (2017).
58. Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785-794 (Association for Computing Machinery).
59. Hengl, T. *et al*. SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE* **12**, e0169748 (2017).
60. Li, Y., Li, M., Li, C. & Liu, Z. Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. *Sci. Rep.* **10**, 9952 (2020).
61. Tan, W., Wei, C., Lu, Y. & Xue, D. Reconstruction of All-Weather Daytime and Nighttime MODIS Aqua-Terra Land Surface Temperature Products Using an XGBoost Approach. *Remote Sens.* **13**, 4723 (2021).
62. Adnan, M., Alarood, A. A. S., Uddin, M. I. & Ur Rehman, I. Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. *PeerJ Comput. Sci.* **8**, e803 (2022).
63. Chen, X. A long-term reconstructed TROPOMI solar-induced fluorescence dataset using machine learning algorithms. *figshare* https://doi.org/10.6084/m9.figshare.19336346.v2 (2022).
64. Guanter, L. *et al*. Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence. *Proc. Natl. Acad. Sci.* **111**, E1327–E1333 (2014).
65. Pierrat, Z. *et al*. Diurnal and seasonal dynamics of solar-induced chlorophyll fluorescence, vegetation indices, and gross primary productivity in the boreal forest. *J. Geophys. Res. Biogeosci*., e2021JG006588 (2022).
66. Magney, T. S. *et al*. Mechanistic evidence for tracking the seasonality of photosynthesis with solar-induced fluorescence. *Proc. Natl. Acad. Sci.* **116**, 11640–11645 (2019).
67. Grossmann, K. *et al*. PhotoSpec: A new instrument to measure spatially distributed red and far-red Solar-Induced Chlorophyll Fluorescence. *Remote Sens. Environ.* **216**, 311–327 (2018).
68. Li, Z. *et al*. Solar-induced chlorophyll fluorescence and its link to canopy photosynthesis in maize from continuous ground measurements. *Remote Sens. Environ.* **236**, 111420 (2020).
69. Magney, T. S. *et al*. Mechanistic evidence for tracking the seasonality of photosynthesis with solar-induced fluorescence. *Proc. Natl. Acad. Sci*. 201900278 (2019).
70. Wei, X., Wang, X., Wei, W. & Wan, W. Use of Sun-Induced Chlorophyll Fluorescence Obtained by OCO-2 and GOME-2 for GPP Estimates of the Heihe River Basin, China. *Remote Sens.* **10**, 2039 (2018).
71. Walther, S. *et al*. Satellite chlorophyll fluorescence measurements reveal large-scale decoupling of photosynthesis and greenness dynamics in boreal evergreen forests. *Global. Change. Biol.* **22**, 2979–2996 (2016).
72. Köhler, P., Guanter, L. & Joiner, J. A linear method for the retrieval of sun-induced chlorophyll fluorescence from GOME-2 and SCIAMACHY data. *Atmos. Meas. Tech.* **8**, 2589–2608 (2015).
73. Parazoo, N. C. *et al*. Towards a Harmonized Long-Term Spaceborne Record of Far-Red Solar-Induced Fluorescence. *J. Geophys. Res. Biogeosci.* **124**, 2518–2539 (2019).
74. Pastorello, G. *et al*. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci. Data* **7** (2020).

75. Reichstein, M. *et al*. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global. Change. Biol.* **11**, 1424–1439 (2005).
76. Lasslop, G. *et al*. Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. *Global. Change. Biol.* **16**, 187–208 (2010).
77. Chen, C. *et al*. China and India lead in greening of the world through land-use management. *Nat. Sustain.* **2**, 122–129 (2019).
78. Tong, X. *et al*. Increased vegetation growth and carbon stock in China karst via ecological engineering. *Nat. Sustain.* **1**, 44–50 (2018).
79. Miettinen, J., Shi, C. & Liew, S. C. Deforestation rates in insular Southeast Asia between 2000 and 2010. *Global. Change. Biol.* **17**, 2261–2270 (2011).
80. De, S. V. *et al*. Land use patterns and related carbon losses following deforestation in South America. *Environ. Res. Lett.* **10**, 124004 (2015).
81. Huete, A. *et al*. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **83**, 195–213 (2002).
82. Still, C. J., Berry, J. A., Collatz, G. J. & Defries, R. S. ISLSCP II C4 Vegetation Percentage, *ORNL Distributed Active Archive Center*, https://doi.org/10.3334/ORNLDAAC/932 (2009).
83. Pierrat, Z. & Stutz, J. Tower-based solar-induced fluorescence and vegetation index data for Southern Old Black Spruce forest, *Zenodo*, https://doi.org/10.5281/ZENODO.5884643 (2022).
84. Magney, T. *et al*. Canopy and needle scale fluorescence data from Niwot Ridge, Colorado 2017-2018, *CaltechDATA*, https://doi.org/10.22002/D1.1231 (2019).
85. Wan, Z., Hook, S. & Hulley, G. MOD11C1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 0.05Deg CMG V006, *NASA EOSDIS Land Processes DAAC*, https://doi.org/10.5067/MODIS/MOD11C1.006 (2015).
86. Friedl, M. & Sulla-Menashe, D. MCD12C1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 0.05Deg CMG V006, *NASA EOSDIS Land Processes DAAC*, https://doi.org/10.5067/MODIS/MCD12C1.006 (2015).
87. Schaaf, C. & Wang, Z. MCD43C4 MODIS/Terra+Aqua BRDF/Albedo Nadir BRDF-Adjusted Ref Daily L3 Global 0.05Deg CMG V006, *NASA EOSDIS Land Processes DAAC*, https://doi.org/10.5067/MODIS/MCD43C4.006 (2015).
88. Doelling, D. CERES Level 3 SYN1DEG-DAYTerra+Aqua HDF4 file - Edition 4A, *NASA Langley Atmospheric Science Data Center DAAC*, https://doi.org/10.5067/TERRA+AQUA/CERES/SYN1DEGDAY_L3.004A (2017).

Acknowledgements

This study is financially supported by the National Natural Science Foundation of China (No. 91847301and No.51809007), the Central Funds Guiding the Local Science and Technology Development of Qinghai Province (2021ZY024), Major Basic Research Development Program of the Science and Technology Agent, Qinghai Province (2019-SF-146) and the State Key Laboratory of Hydroscience and Engineering-Tsinghua (No. 2019-KY-01). We thank the California Institute of Technology for providing access to the TROPOMI SIF. Data used in this study: CERES PAR product is from https://ceres.larc.nasa.gov/data/data-product-dois/; MODIS products are from https://ladsweb.modaps.eosdis.nasa.gov/; ISLSCP II C4 vegetation map product is from https://doi.org/10.3334/ORNLDAAC/932[82]; FLUXNET products are from https://fluxnet.fluxdata.org/; TROPOMI SIF and COC-2 SIF are from ftp://fluo.gps.caltech.edu/data/; GMOE-2 SIF is from https://avdc.gsfc.nasa.gov/pub/data/satellite/MetOp/GOME_F/; Tower-based SIF measurements form https://doi.org/10.5281/zenodo.5884643[83] and https://doi.org/10.22002/D1.1231[84] VPM GPP product is from http://data.tpdc.ac.cn/en/data/582663f5-3be7-4f26-bc45-b56a3c4fc3b7/; We would like to thank these organizations for providing the easily accessible data. XGBoost was implemented using the Python library XGBoost (https://github.com/dmlc/xgboost). The authors thank Figshare for archiving the dataset. Finally, we thank the editor and two anonymous reviewers for their suggestions on the article.

Author contributions

Xingan Chen and Shuo Zhang designed research and wrote the paper. Yuefei Huang and Guangqian Wang designed the research. Chong Nie, Shiliu Chen, and Zhichao Chen processed the data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01520-1.

Correspondence and requests for materials should be addressed to S.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.