

CEDAR Project

upsCaling Ecosystem Dynamics with ARtificial intelligence

CEDAR-GPP Product Version 1 User Guide

2024/01

Yanghui Kang, Maoya Bassiouni, Max Gaber, Xinchun Lu, Trevor Keenan
University of California, Berkeley

Table of Contents

1. <i>Overview</i>	1
2. <i>Model setups</i>	1
3. <i>File structure and characteristics</i>	2
4. <i>File Naming Convention</i>	2
5. <i>Guidance on dataset selection</i>	3
6. <i>Support</i>	3

1. Overview

CEDAR-GPP is a global Gross Primary Productivity (GPP) data product, including monthly GPP estimates at 0.05° spatial resolution from 1982 to 2020. These datasets were generated by upscaling eddy covariance measurements to the global scale using machine learning based on satellite and climate data. CEDAR-GPP uniquely incorporated the effect of elevated atmospheric CO₂ on photosynthetic light use efficiency, known as the direct CO₂ fertilization effect (CFE), using both data-driven and theoretical approaches. Built on a broad range of satellite-derived land surface properties and climate variables, CEDAR-GPP provides comprehensive observation-based estimates of GPP spatiotemporal dynamics, providing important indications of global long-term photosynthesis changes. The acronym CEDAR stands for upsCaling Ecosystem Dynamics with ARtificial intelligence.

Table 1. Specifications of the CEDAR-GPP model setups.

Model Setup Name	Temporal range	Direct CO ₂ Fertilization Effects		GPP Partitioning Method
		Configuration	Method	
ST_Baseline_NT	Short-term (ST) 2001 – 2020	Baseline	Not incorporated	Night-time (NT)
ST_Baseline_DT				Daytime (DT)
ST_CFE-Hybrid_NT		CFE-Hybrid	Theoretical	NT
ST_CFE-Hybrid_DT				DT
ST_CFE-ML_NT		CFE-ML	Data-driven	NT
ST_CFE-ML_DT				DT
LT_Baseline_NT	Long-term (LT) 1982 – 2020	Baseline	Not incorporated	NT
LT_Baseline_DT				DT
LT_CFE-Hybrid_NT		CFE-Hybrid	Theoretical	NT
LT_CFE-Hybrid_DT				DT

2. Model Setups

CEDAR-GPP comprises ten model setups based on combinations of temporal spans (ST vs LT), methods of incorporating the direct CO₂ fertilization effect (Baseline, CFE-Hybrid, CFE-ML), and GPP partitioning methods (NT vs DT) (Table 1).

- **Temporal Span:** The short-term (ST) model configuration produced GPP estimates from 2001 to 2020, and the long-term (LT) configuration spanned from 1982 to 2020. Each temporal configuration uses a different set of input variables depending on their availability. Inputs for the ST configuration included MODIS, CSIF, BESS PAR, ESA-CCI soil moisture, ERA5-Land, as well as PFT and Koppen Climate zone. The long-term models used GIMMS NDVI4g and LAI4g data, ERA5-Land, PFT and Koppen climate.
- **Direct CO₂ fertilization:** CEDAR-GPP comprised three configurations regarding the direct CO₂ fertilization effect (CFE). The “Baseline” configuration did not consider these effects; The “CFE-Hybrid” configuration incorporated the effects via eco-evolutionary theory; the “CFE-ML” configuration inferred the direct effects from eddy covariance data using machine

learning, providing data-driven estimates without prescribed assumptions of the CO₂ sensitivity. Due to the limited availability of eddy covariance observations before 2001, we did not apply the CFE-ML approach to the long-term setups, as the machine learning model could not robustly extrapolate CO₂ fertilization effect beyond the training period.

- **GPP partitioning method:** Separate models were trained for GPP target variables derived from the night-time (NT) and daytime (DT) partitioning approaches.

Additionally, for each model setup in Table 1, 30 machine learning models were trained based on bootstrapping training samples to quantify uncertainties. We provided the ensemble mean and standard deviation from the model ensembles.

3. File Structure and Characteristics

Format:	NetCDF
Spatial Resolution:	0.05 degree
Temporal Resolution:	Monthly
Temporal Coverage:	Short-term (ST): 2001-2020; Long-term (LT): 1982 - 2020
Image Dimension:	Rows: 3600, Columns: 7200
Units:	gCm ⁻² day ⁻¹
Fill Value:	-9999
Scale Factor:	0.01
Data Type:	uint16
File Size:	Approximately 99 MB per file
Data variables:	“GPP_mean” and “GPP_std” representing the mean and standard deviation of monthly GPP derived from a model ensemble of 30 members

4. File Naming Convention

Files are named following the convention:

CEDAR-GPP_<version>_<model-setup>_<YYYYMM>.nc

Where <model-setup> is a composite string of

<temporal_span>_<CFE_option>_<GPP_partitioning>

- **<temporal_span>:** Defines the dataset’s temporal coverage. “ST” indicates the short-term span from 2001 to 2020, while “LT” refers to the long-term span from 1982 to 2020.
- **<CFE_option>:** indicates the approach to direct CO₂ fertilization. 'Baseline' indicates no incorporation of this effect, 'CFE-ML' represents direct CO₂ fertilization incorporated by ML, 'CFE-Hybrid' implies direct CO₂ fertilization incorporated by theoretical approaches.
- **<GPP_partitioning>:** specifies method used for eddy covariance GPP partitioning. 'NT' indicates the night-time method, while 'DT' indicates the day-time method.

Example: The file name “CEDAR-GPP_v01_ST_CFE-ML_NT_200408.nc” refers to a file from the version 1 of the project, based a model setup with a short-term (ST) temporal span, the

incorporation of the direct CO₂ fertilization effect from the CFE-ML configuration, and based on eddy covariance GPP derived from the night-time (NT) approach. The data covers April, 2004.

5. Guidance on Dataset Selection

We provide a structured approach to selecting the most appropriate CEDAR-GPP dataset for users' research and applications.

Study period considerations

The Short-Term (ST) setup is ideal for studies focusing on periods after 2000. These models are constructed using a broader range of explanatory predictors, offering higher precision and smaller random errors. The Long-Term (LT) datasets should be used for research assessing GPP dynamics over a longer time period (before 2001). It's important to note that trends from the ST and LT datasets are not directly comparable, as they were derived from different satellite remote sensing data.

CO₂ Fertilization Effect configurations:

The CFE-Hybrid and CFE-ML setups are preferable when assessing temporal GPP dynamics, especially long-term trends. The CFE-Hybrid setup includes a hypothetical trend for the direct CO₂ effect, while CFE-ML is purely data-driven and does not make any specific assumption about the sensitivity of photosynthesis to CO₂. Averaging the CFE-Hybrid and CFE-ML estimates is acceptable, with the difference between them reflecting the uncertainty surrounding the direct CO₂ effect. Note that the Baseline setup should not be used to study long-term GPP dynamics, especially those induced by elevated CO₂. Baseline setup may be useful to directly compare with other remote sensing-derived GPP datasets that do not consider the direct CO₂ effect. Differences between these setups regarding mean GPP spatial patterns, seasonal and interannual variations are minor.

GPP partitioning methods:

We recommend using the mean value derived from both the "NT" (Nighttime) and "DT" (Daytime) models. The difference between these two provides insight into the uncertainties arising from the partitioning approaches used in GPP estimation from eddy covariance measurements.

We encourage users to contact the authors for any questions or additional information needed to make an informed choice about dataset selection.

6. Support

For any questions related to this dataset, please contact:

Yanghui Kang

Email: kangyanghui@gmail.com, yanghuikang@berkeley.edu