CoCO2

Prototype system for a Copernicus CO_2 service

Toolbox to derive customized model forcing data and for assessing errors of simulated terrestrial CO2 fluxes from data base of biogenic CO2 flux measurements

Jacob A. Nelson, MPG-Jena

Co-ordinated by ECMWF

D5.2: Toolbox to derive customized model forcing data and for assessing errors of simulated terrestrial CO2 fluxes from data base of biogenic CO2 flux measurements

Dissemination Level: Public

Author(s): Jacob A. Nelson (MPG-Jena) and Sophia Walther (MPG-Jena)

Date: 16/12/2022
Version: 1.0
Contractual Delivery Date: 31/12/2022
Work Package/ Task: SP5/ T5.2
Document Owner: MPG-Jena
Contributors: CEA, ECMWF, ICOS ERIC, iLab, ULUND, WU, CMCC, DLR, M1-M24, MPG-Jena
Status: Final

CoCO2: Prototype system for a Copernicus CO_2 service

Coordination and Support Action (CSA)
H2020-IBA-SPACE-CHE2-2019 Copernicus evolution – Research activities in support of a European operational monitoring support capacity for fossil CO2 emissions

Project Coordinator: Dr Richard Engelen (ECMWF)
Project Start Date: 01/01/2021
Project Duration: 36 months

Published by the CoCO2 Consortium

Contact:
ECMWF, Shinfield Park, Reading, RG2 9AX,
richard.engelen@ecmwf.int

The CoCO2 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 958927.

Table of Contents

1 Executive Summary 5
2 Introduction 5
2.1 Background 5
2.2 Scope of this deliverable 5
2.2.1 Objectives of this deliverable 5
2.2.2 Work performed in this deliverable 6
2.2.3 Deviations and counter measures 6
3 Uncertainty and error assessment data base and toolbox 6
3.1 Data base overview 7
3.1.1 *{site}_flux.nc* files 8
3.1.2 *{site}_meteo.nc* files 11
3.1.3 *{site}_meteo_gf.nc* files 13
3.1.4 *{site}_rs.nc* files 14
3.1.5 File metadata 15
3.2 Toolbox 16
4 Conclusion 18
5 References 19

Figures

Figure 1 Flowchart of the data base and toolbox outline in this document. 7
Figure 2 Map and time span of all data currently in the data base. Each point on the map represents one site, colored by plant functional type. Each point on the time series represents one site-year, colored again by plant functional type, with the vertical spacing relating to latitude. 7
Figure 3: Number of site-years by data product in time. The FLUXNET2015 data product includes site years only through 2014. Most of the ICOS Drought 2018 sites are superseded by data in the Warm Winter 2020 data product. 8
Figure 4: Hourly flux data for a two week period from the Hainich Forest site in Germany. 10
Figure 5: Hourly meteorological data for a two week period from the Hainich forest site in Germany. 12
Figure 6: Monthly means of meteorological data, with lines indicating the gap filled data and points indicating measured data, for the full coverage period. Data from the Las Majadas del Tietar South site in Spain. 14
Figure 7: Screenshot of the modelevaluation.org home page, accessed on 14-12-2022. 17
Figure 8: Example of evaluation metrics calculated from modelevaluation.org, accessed on 14-12-2022. 18

Tables

Table 1: Overview of variables contained in the *{site}_flux.nc* data file. 9
Table 2: Overview of meteorological variables with a time component in the {site}_meteo.nc files 11
Table 3: Overview of ancillary variables without a time component, giving site characteristics. 12
Table 4: Overview of the remote sensing variables contained in the {site}_rs.nc file. 15
Table 5: Example metadata from the ES-LM2_flux.nc file. 15

1 Executive Summary

In order to facilitate an assessment of uncertainty and error of estimated biogenic CO2 fluxes, this document reports on a framework linking the up-to-date globally distributed eddy covariance data with a land surface model evaluation toolbox. The database of eddy covariance data includes the measured biogenic fluxes (NEE and GPP) and meteorological data, as well as meteorological data from ERA5 that were calibrated to measured meteorology to facilitate gap-filling and extension back in time. In addition, automated quality control information for each variable is available. All data have been standardized regarding units, metadata, variable names, and file format in a way that is consistent with existing model evaluation exercises. This data base is then hosted at modelevaluation.org, which acts as a toolbox to allow for automated statistical checks and analysis, which can be customized and updated as needed. Furthermore, coordination within partners has created a pipeline linking the data production and model evaluation components for more seamless and continuous updates in the future.

2 Introduction

2.1 Background

WP5 aims at better and systematically quantifying sources of uncertainty in the inversions, one of them being inaccuracies in the prior estimates of biogenic carbon exchange. Information on the uncertainty of such flux products is often limited to certain uncertainty components, or not reported at all, and therefore very inconsistently available across products. This task 5.2 addresses the need for systematic and standardized assessments of the accuracy of biogenic flux estimates. For this, it interfaces with WP7 and targets the only available actual measurements of biogenic fluxes: the ones at eddy-covariance sites. Those allow evaluations of the flux estimates based on simulations for flux towers. We built a data base of in-situ flux observations and related meteorological and remote sensing measurements, all of which are relevant not only for the simulation of terrestrial carbon exchange, but also for characterizing the error and uncertainty of the flux estimates. Within the modelevaluation.org toolbox, that has been designed for model evaluation at flux towers, the standardized data set allows for the systematic and accurate assessment of uncertainties in a consistent manner across different data sets of biogenic priors. It also allows customizing and extending evaluations to new times and sites. Information on the error and uncertainty characteristics of the prior gained from the use of such a toolbox will directly contribute to improving accuracy of modelling and inversion exercises performed on both the global scale in WP3 and on regional to national scales in WP4. Such information will also feed back to developers of biogenic model flux products, e.g. in WP2. Implementing the dataset into an existing toolbox has the additional advantages of coordinating efforts both within and outside of this project, and fostering sustained use in the scientific community, which will prolong beyond the project lifetime.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

Here we describe the data base and toolbox used to facilitate the assessment of uncertainty and error of estimated biogenic CO2 fluxes. The tool and associate data support the efforts of WP3 in establishing error covariance parametrizations for the assimilation of biogenic priors as well as benchmarking land surface model developments. The core work is in developing a data pipeline linking the in situ eddy covariance networks with model evaluation protocols which can both be used within the project as well as in future endeavours by the community.

2.2.2 Work performed in this deliverable

- Collected requirements on model forcing data (e.g. variables and resolutions) and needed functionality of the evaluation toolbox.
- Defined the needed experimental set-ups, file formats, required variables for diverse model structures, as well as defining units, metadata, and data policy requirements.
- Established links between eddy covariance networks and model evaluation toolbox (modelevaluation.org) in order to obtain the most up to date in situ data, both now and in the future.
- Enhancement of quality control protocols to automatically detect potential issues in reported in situ flux and meteorological data, reported as additional flags in the data set.
- Inclusion of additional empirical meteorological forcing data for model inputs and spin-ups which spans the full time series.
- Inclusion of remotely sensed vegetation indices from MODIS based on state-of-the-art methodologies.
- Coordination with an existing analysis toolbox (modelevaluation.org) to facilitate customized model evaluation protocols based on project needs.

2.2.3 Deviations and counter measures

The planned delivery date was November 30th 2022. The final delivery of this report was delayed until the middle of December, because of unexpected illnesses.

3 Uncertainty and error assessment data base and toolbox

The two core parts of this deliverable are described below, being a) the data base of in situ data (Section 3.1) and b) the associated toolbox (Section 3.2). Both of these parts required extensive coordination among the collaborating parties with the goal of building a data-to-model-evaluation pipeline. This pipeline is meant to be longer lasting and extendible, giving both the most up to date in situ data and the flexibility to design model evaluation schemes in a way that can be customized and that is consistent with existing efforts. The pipeline does currently not allow for automated updates of newly available or processed eddy covariance and ancillary data. However, it was heavily discussed how that can be achieved in the future.

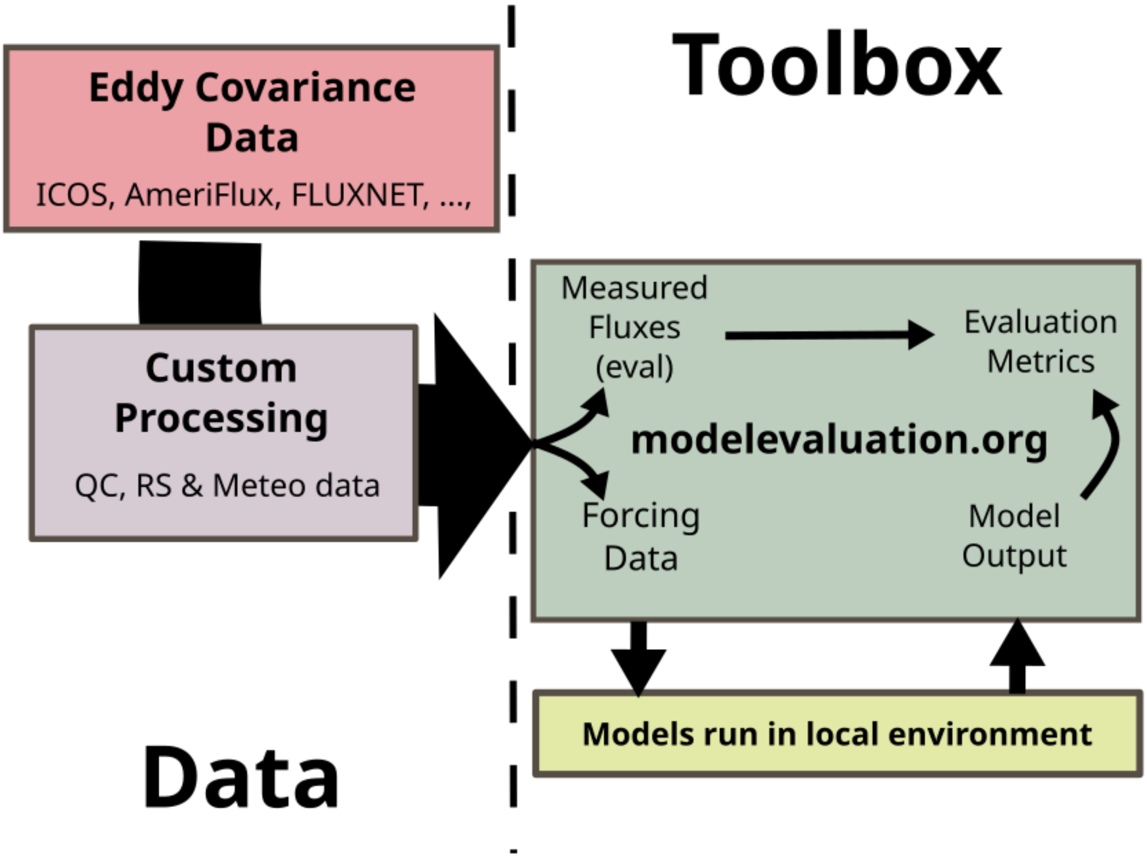


Figure 1 Flowchart of the data base and toolbox outline in this document.

3.1 Data base overview

The data base consists of data from 1,918 site-years from 257 sites aggregated from four different eddy covariance collections. Data span the MODIS era from 2002-2020 and are spatially skewed towards the northern hemisphere, particularly Europe and North America.

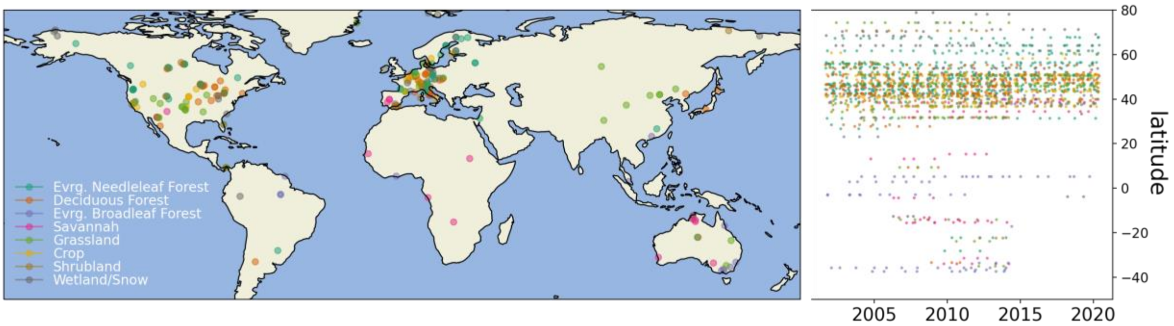


Figure 2 Map and time span of all data currently in the data base. Each point on the map represents one site, colored by plant functional type. Each point on the time series represents one site-year, colored again by plant functional type, with the vertical spacing relating to latitude.

Apart from the standard FLUXNET2015 dataset (Pastorello et al. 2017), data released from regional networks, particularly the Drought 2018 (Drought 2018 Team and ICOS Ecosystem Thematic Centre 2020) and Warm Winter 2020 (Warm Winter 2020 Team and ICOS Ecosystem Thematic Centre 2022) datasets from the Integrated Carbon Observation System (ICOS, https://www.icos-cp.eu) and the Ameriflux FLUXNET data product (as of 2022) are part of the data base. All data is covered under the CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/) license which allows the data to be shared and adapted with appropriate attribution, the details of which (including doi links) are included in the metadata of each file (see subsection 3.1.5).

All eddy covariance data has been uniformly processed using the ONEFlux data processing pipeline (Pastorello et al. 2020), giving consistent units, gap filling, and initial quality control. In cases where a site was present in more than one data product (e.g. Hainich Forest data is found in the FLUXNET2015, Drought 2018 and Warm Winter 2020 products), the version with the longest time series was used. All datasets are aggregated to a common hourly time step and files are saved in the netcdf format, with variable names and units consistent with the Assistance for Land-surface Modelling Activities (ALMA) convention (http://www.lmd.jussieu.fr/~polcher/ALMA/) where possible.

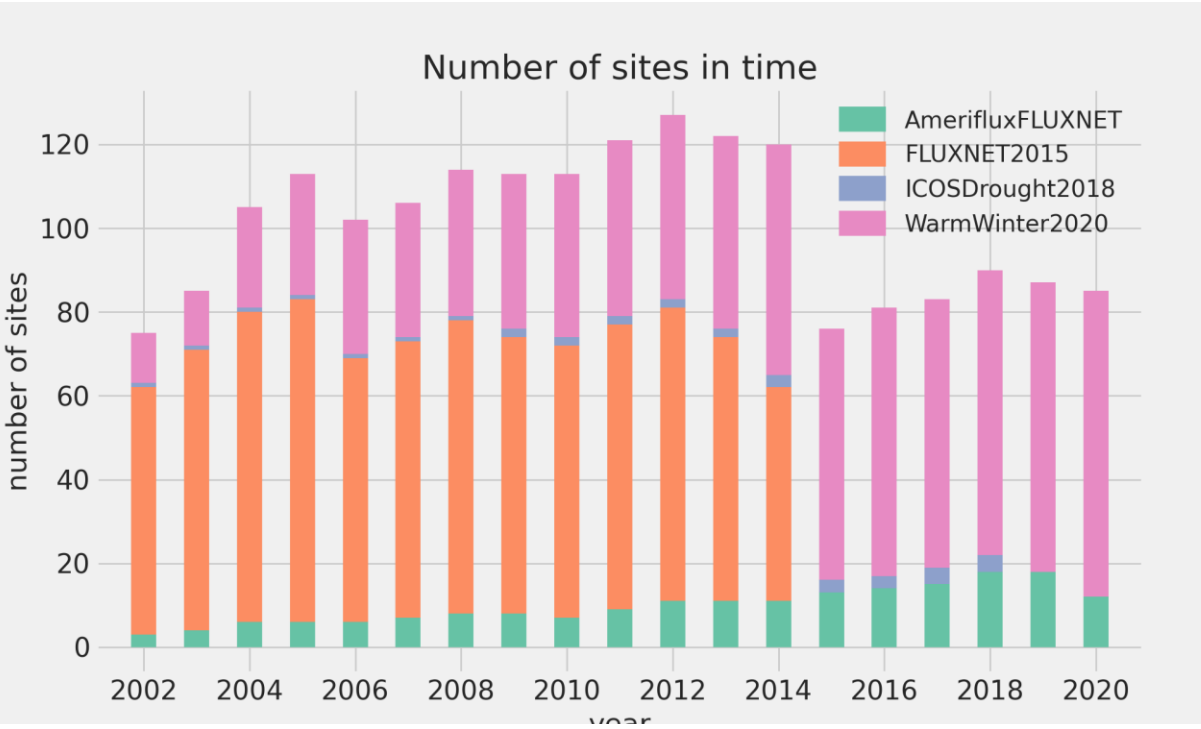


Figure 3: Number of site-years by data product in time. The FLUXNET2015 data product includes site years only through 2014. Most of the ICOS Drought 2018 sites are superseded by data in the Warm Winter 2020 data product.

In addition to the energy/carbon flux data and meteorological data reported from each site, data from remote sensing and downscaled meteorological data based on ERA5 (Hersbach et al. 2018) products are included for each site, resulting in four separate data files: flux data (*{site}_flux.nc*), measured meteorological data (*{site}_meteo.nc*), gap filled meteorological (*{site}_meteo_gf.nc*), and remote sensing (*{site}_rs.nc*). File names give the site code of each site, followed by the data file type, e.g. the flux data for Hainich Forest has the file name DE-Hai_flux.nc. Each file type and the variables contained within is further described below.

3.1.1 *{site}_flux.nc* files

The *{site}_flux.nc* data files contain not only the carbon fluxes needed for assessment of biogenic priors (NEE, GPP, and reco, see Table 1), but also the energy fluxes measured at each site. Each variable has a corresponding quality control as {variable}_qc (e.g. NEE_qc, see subsection 3.1.1.1 for more details). All turbulent fluxes (NEE, GPP, reco, Qh and Qle) have been filtered to exclude uncertain estimates from gap-filling. Furthermore, the carbon fluxes have been further filtered to exclude low turbulent conditions using the variable ustar threshold method. For each flux, after quality filtering, the resulting gaps were then gap filled using the Marginal Distribution Sampling (MDS) algorithm (Reichstein et al. 2005). GPP and reco were derived from NEE using the night-time partitioning algorithm. All details of the processing steps can be found in Pastorello et al. 2020, and the names of the corresponding ONEFLUX variables are found in Table 1.

Table 1: Overview of variables contained in the *{site}_flux.nc* data file.

variable	**unit**	**long name**	**ONEFLUX variable**
NEE	kg/m2/s	Net Ecosystem Exchange	NEE_VUT_50
GPP	kg/m2/s	Gross Primary Production	GPP_NT_VUT_50
reco	kg/m2/s	Ecosystem Respiration	RECO_NT_VUT_50
Qh	W/m2	Sensible heat flux	H_F_MDS
Qle	W/m2	Latent heat flux	LE_F_MDS
Qg	W/m2	Ground heat flux	G_F_MDS
rnet	W/m2	Net surface radiation	NETRAD

An example of the hourly flux data for a two week period can be seen in Figure 4 for the Hainich Forest site (code DE-Hai).

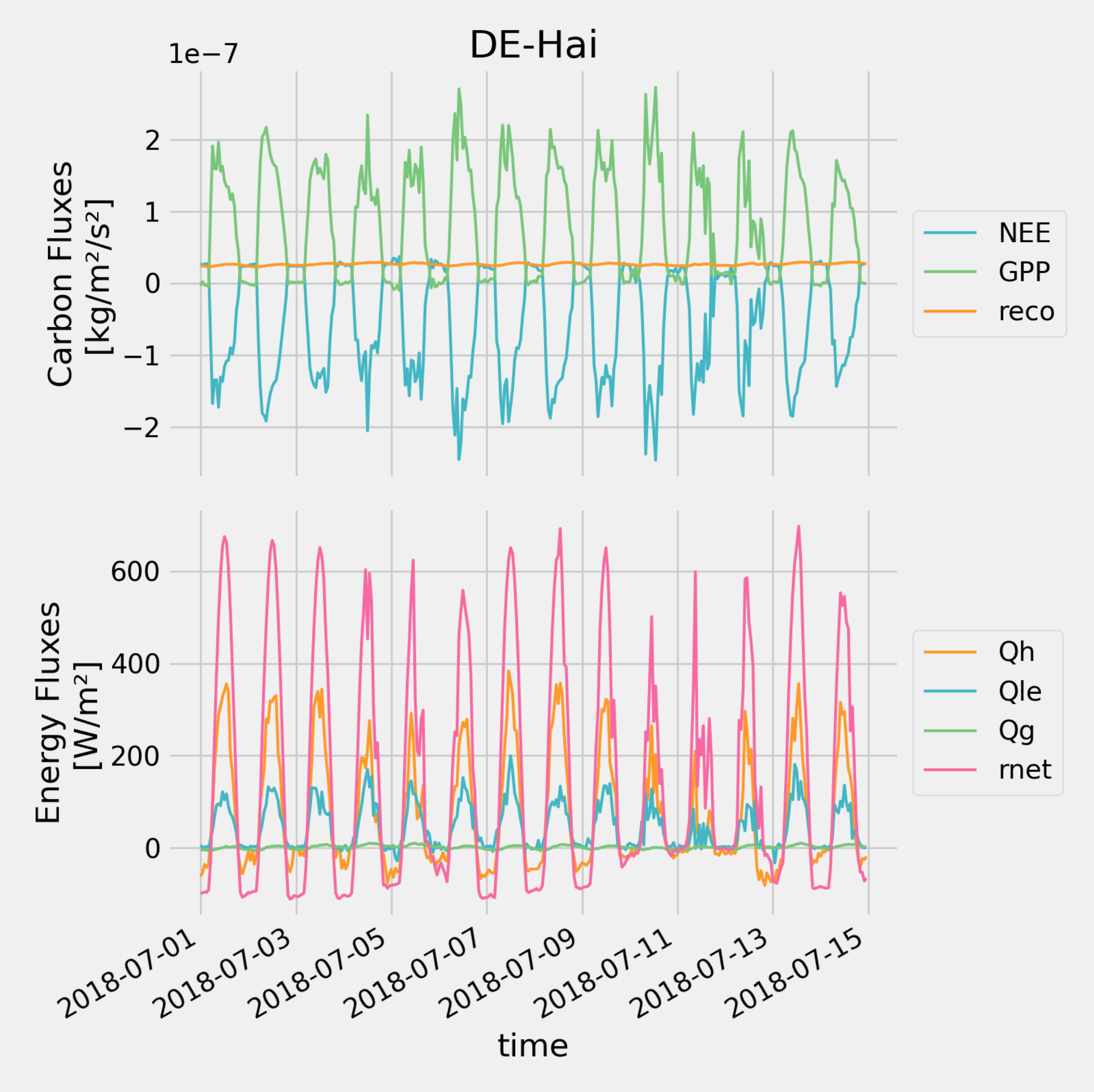


Figure 4: Hourly flux data for a two week period from the Hainich Forest site in Germany.

3.1.1.1 Additional quality control

In addition to the quality control reported from the ONEFLUX data processing, many flux and meteorological variables (i.e. NEE, GPP, reco, Tair, vpd, SWdown, Qh, Qle, rnet) have an additional automated quality flagging. The resulting flag gives four indicated quality levels based on the expected and statistically inferred relationships between variables, both within a site and across all sites (Jung et al., in preparation). Flagged data was not removed, but instead simply indicated in the quality flag, with the flags denoting:

- 0: all original data
- 1: good quality gap filled
- 2: flagged by automatic QC algorithm
- 3: bad quality

In this way, data used in each assessment can be modular depending on the needs. Any meteorological data not meeting the assigned quality control threshold can be replaced by the corresponding gap filled version in the *{site}_meteo_gf.nc* files (see section 3.1.3).

An overall uncertainty for each data point and variable is unfortunately not available. Even though, the uncertainty of some individual processing steps can be estimated (e.g. due to the u* filtering), the calculation of a solid overall uncertainty is currently not possible as it seems that the effect of several sources of uncertainty can not be quantified at present. A closer interaction among modelling, data assimilation, and eddy covariance experts can help in deriving and improving relevant uncertainty quantification in the future.

3.1.2 *{site}_meteo.nc* files

Meteorological data in the *{site}_meteo.nc* file consists of meteorological variables reported from the in situ measurements, as well as additional site characteristics such as IGBP vegetation type and soil properties.

3.1.2.1 Meteorological data

All meteorological data in this file is from the reported site measurements. In the case of RH and Qair, the data was calculated from the reported vpd, Tair and Psurf. SWdown_clearsky was calculated from the site latitude, longitude, and time based on the ONEFLUX code (Pastorello et al. 2020). Tair, vpd, and SWdown all have the additional automated quality control flags (see subsection 3.1.1.1), while all variables have standard quality control flags designating original, gap-filled, or missing data.

Table 2: Overview of meteorological variables with a time component in the {site}_meteo.nc files.

variable	unit	long name	gap filled
CO2air	mol/mol	Near surface CO2 concentration	no
Tair	K	Near surface air temperature	yes
vpd	hPa	Vapor pressure deficit	yes
RH	%	Relative humidity	yes
Qair	kg/kg	Near-surface (usually, 2 meter) specific humidity	yes
Precip	kg/m2/s	Precipitation rate	yes
Psurf	Pa	Surface Pressure	yes
SWdown	W/m2	Downward short-wave radiation	yes
LWdown	W/m2	Downward long-wave radiation	no
Wind	m/s	Near surface wind speed	yes
SWdown_clearsky	W/m2	Downward short-wave radiation assuming clear sky	yes

An example of the hourly meteorological data for a two week period can be seen in Figure 5 for Hainich Forest site (code DE-Hai).

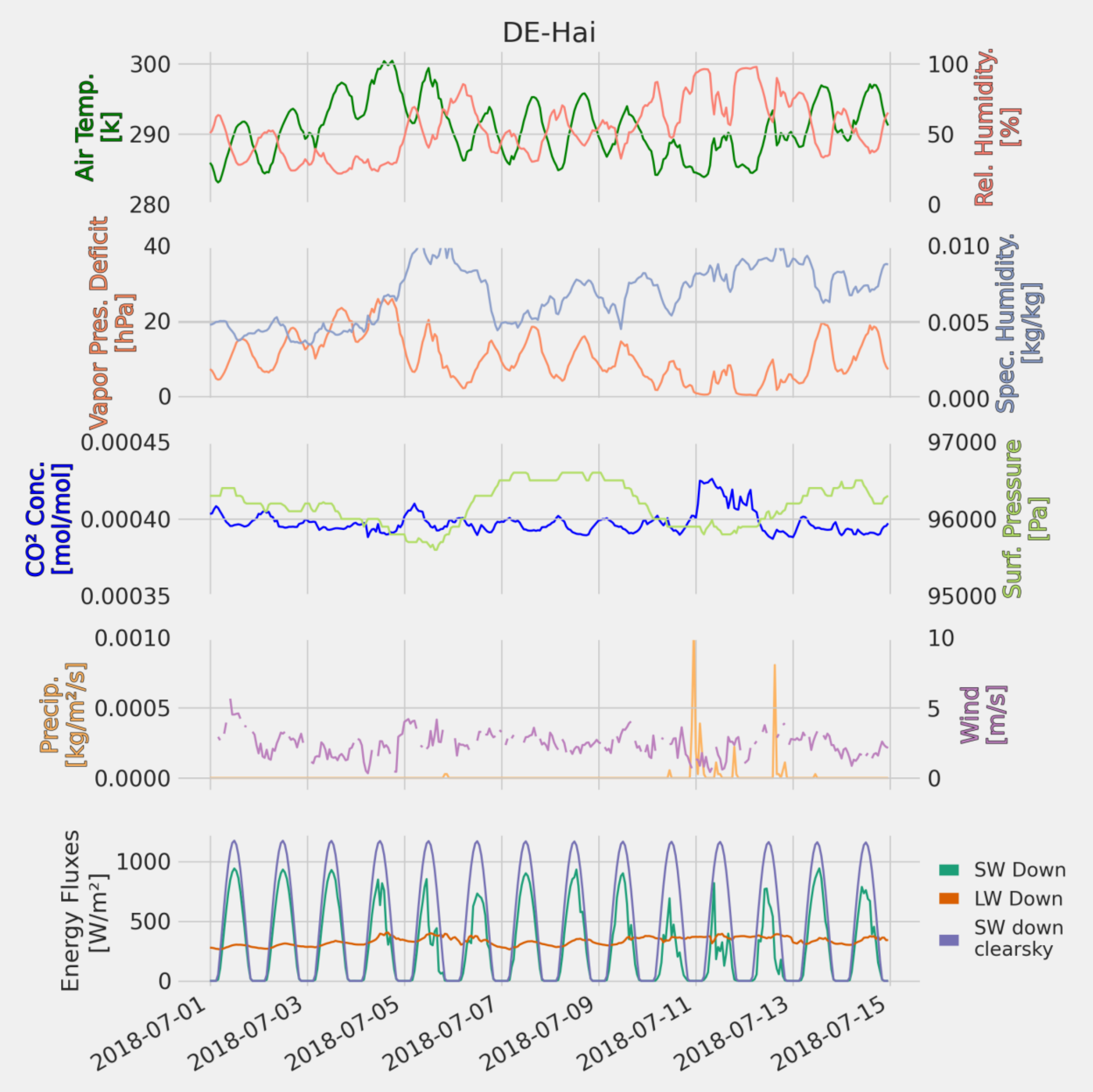


Figure 5: Hourly meteorological data for a two week period from the Hainich forest site in Germany.

3.1.2.2 Site characteristics

In addition to the reported meteorological data, each *{site}_meteo.nc* file contains site characteristic data. IGBP vegetation types are reported from the site principal investigators, and any sites with missing classifications were identified using reported literature. Soil characteristics were derived from SOILGRIDS (Hengl et al. 2017), with reported values being a depth weighted mean from all pixels (250 m gridded product) within a 1 kilometer radius of the tower.

Table 3: Overview of ancillary variables without a time component, giving site characteristics.

variable	unit	long name
IGBP_veg_short	-	IGBP vegetation type (short)

AWCh1	%	Available soil water capacity (volumetric fraction) for h1
AWCh2	%	Available soil water capacity (volumetric fraction) for h2
AWCh3	%	Available soil water capacity (volumetric fraction) for h3
AWCtS	%	Saturated water content (volumetric fraction) for tS
BLDFIE	kg m-3	Bulk density (fine earth) in kg / cubic-meter
CECSOL	cmolc kg-1	Cation exchange capacity of soil in cmolc/kg
CLYPPT	%	Clay content (0-2 micro meter) mass fraction in %
CRFVOL	%	Coarse fragments volumetric in %
ORCDRC	g kg-1	Soil organic carbon content (fine earth fraction) in g per kg
PHIHOX	index*10	Soil pH x 10 in H2O
PHIKCL	index*10	Soil pH x 10 in KCl
SLTPPT	%	Silt content (2-50 micro meter) mass fraction in %
SNDPPT	%	Sand content (50-2000 micro meter) mass fraction in %
WWP	%	Available soil water capacity (volumetric fraction) until wilting point

3.1.3 *{site}_meteo_gf.nc* files

For a subset of meteorological variables, the dataset contains additional gap-free data based on ERA5 gridded data (Hersbach et al. 2018), scaled to site level. Each variable was scaled to site level by training a Random Forest model (Scikit-learn, Pedregosa et al. 2011), to predict the site measured data, using the ERA5 data as features from the grid pixel (0.5° grid) containing the tower, as well as potential radiation. The quality of the gap-filling is generally high according to visual inspections (see e.g. Fig.5) and cross-validation statistics (not shown), while accurate gap-filling of precipitation will always be very hard.

In the case of precipitation, due to the stochastic nature of the variable, a stochastic simulator based on a predicted distribution of precipitation for each hour was used, and therefore the gap filled version of precipitation may not reflect the actual conditions, but rather aims to fit diurnal and seasonal rainfall frequency and intensity patterns. The intended use for these data is for gap filling missing or low quality data from the measured meteorological data. Also, as these data extend to the full time period, this data can be used for model spin up over longer periods, particularly for sites with short temporal coverage. The rightmost column ("gap filled") in Table 2 indicates which variables are included in both the *{site}_meteo.nc* and *{site}_meteo_gf.nc* files.

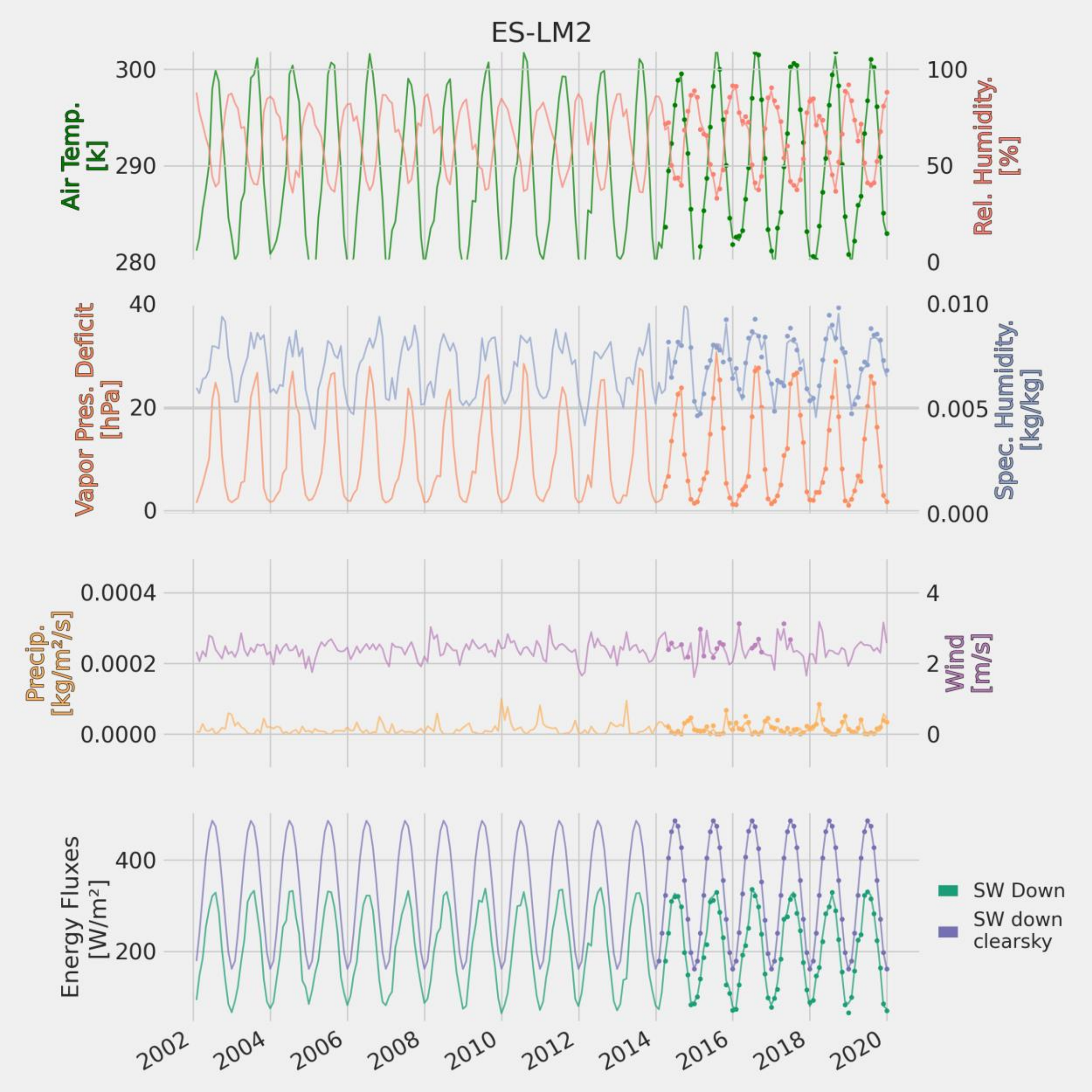


Figure 6: Monthly means of meteorological data, with lines indicating the gap filled data and points indicating measured data, for the full coverage period. Data from the Las Majadas del Tietar South site in Spain.

3.1.4 *{site}_rs.nc* files

Remote sensing data is derived from MODIS data from the FluxnetEO dataset v1 (Walther & Besnard et al. 2022). Data within a 1 km radius from a tower has been filtered for good quality data and gap filled giving continuous time series of clear sky land surface temperature (LST) and vegetation indices (EVI, NIRv, and NDWI). An evaluation of the gap-filling is available in Walther & Besnard et al. 2022. In addition, estimates of LAI and fPAR are included, which were calculated from a regression model predicting LAI from MODIS EVI and NDVI. Model parameters were estimated using the max LAI reported from a subset of sites and the yearly max EVI and NDVI from MODIS. All remote sensing data are at a daily temporal resolution.

Table 4: Overview of the remote sensing variables contained in the {site}_rs.nc file.

variable	unit	long name
LST_TERRA_Day	K	Day-time land surface temperature
LST_TERRA_Night	K	Night-time land surface temperature
EVI	-	Enhanced vegetation index
NIRv	-	Near-infrared reflectance of vegetation
NDWI_band7	-	Normalized difference water index
LAI	m2 m-2	Leaf area index
fPAR	-	Fraction of absorbed photosynthetically active radiation

3.1.5 File metadata

Additional metadata corresponding to data origin and site characteristics are included in each netcdf file. The most important of which is the licence, citation, and attribution information for each data source. While all data is covered by the CC BY 4.0 licence, attribution requirements and best practices are slightly different both for different data types (i.e. remote sensing vs. eddy covariance) and different data sources (e.g. FLUXNET2015 vs. ICOS). Care should be taken to properly attribute each source in any publication or further application of these data. Furthermore, this data base is only intended to be used in a model evaluation scheme, and should not be used as an alternate distribution mechanism for the original data, and instead any non-model evaluation use cases should be derived from the original data sources.

Table 5: Example metadata from the ES-LM2_flux.nc file.

attribute	example value
created_by	jnelson@bgc-jena.mpg.de
creation_date	2022-11-21
data_policy	CC-BY-4.0
data_policy_details	# ICOS Data Licence - Summary----ICOS data is licensed under the Creative Commons Attribution 4.0 International licence (CC BY 4.0).----The summary of (and not a substitute for) the licence can be found here :https://creativecommons.org/licenses/by/4.0/legalcode----## You are free to:----- Share — copy and redistribute the material in any medium or format— Adapt — remix, transform, and build upon the material for any purpose, even commercially----## Under the following terms:----- Attribution — You must give appropriate credit, provide a link to the licence, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.— No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the licence permits.----The licensor cannot revoke these

	freedoms as long as you follow the licence terms.----# How to cite ICOS data----For most of the ICOS data we will provide you with the citation to use through the landing page of the data object; from this page you will always be able to download the data object again. The Persistent Digital Identifiers (PID) or Digital Object Identifier (DOI) of the data object will always resolve to its landing page when you submit this to the handle system (https://hdl.handle.net/yourpidhandle or https://doi.org/yourdoihandle).----You may also use or link to pictures from the ICOS websites or use visualisations if you add the link to the original and the text "ICOS RI, licensed under CC4BY" followed by the link to the original.----# FAIR use of ICOS data----By downloading the ICOS data product you agree to the licencing conditions that apply to the data (CC4BY). Under this license derived products and redistribution are allowed, but you are required to always inform your users of the original source of the data used, refer them to the license text and the original source at ICOS for possible updates or uploads.----Use of the data requires proper reference and citation of the ICOS data, using the exact citation (including the provided DOI or PID) as provided at the moment of upload from ICOS, if applicable.----We ask you to inform the data providers, traceable through the metadata connected to the provided DOI or PID, when the data is used for publication(s), and to offer them the possibility to comment and/or offer them co-authorship or acknowledgement in the publication when this is justified by the added value of the data for your results.----The ICOS data products are provided "as is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose, title and non-infringement. In no event shall the copyright holders or anyone distributing the ICOS data products be liable for any damages or other liability, whether in contract, tort or otherwise, arising from, out of or in connection with the ICOS data products.
doi	https://doi.org/10.18160/2G60-ZHAK
pft	SAV
site_name	Majadas del Tietar South

3.2 Toolbox

These data are made available in the context of model and uncertainty evaluation via the modelevaluation.org platform, which is a state-of-the-art platform designed to uniformly assess model uncertainties. The system was built to facilitate the Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER, Best et al. 2015), and allows modelling teams to download standardized forcing data directly associated with a model evaluation experiment and upload the corresponding model runs to trigger an automated assessment. Assessment scripts can be customized to the experimental design, as well as incorporating the existing model evaluation schemes developed for the PLUMBER and the follow-up PLUMBER2 projects.

Beyond model evaluation routines already implemented in modelevaluation.org in the context of previous projects, tailored evaluation metrics for the objectives of CoCO2 and potential follow-ups should be developed and implemented by the modelling and data-assimilation teams according to a protocol. Such a model simulation and evaluation protocol may include the sensitivity of model performance metrics to the strictness of data quality filtering for fluxes but also meteorological forcing data.

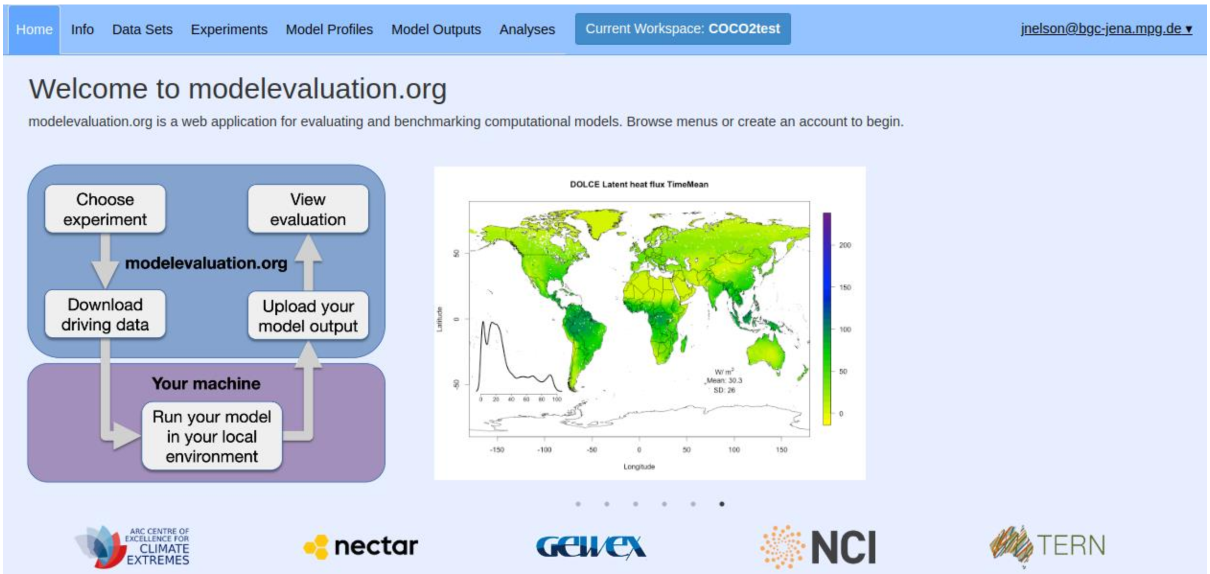
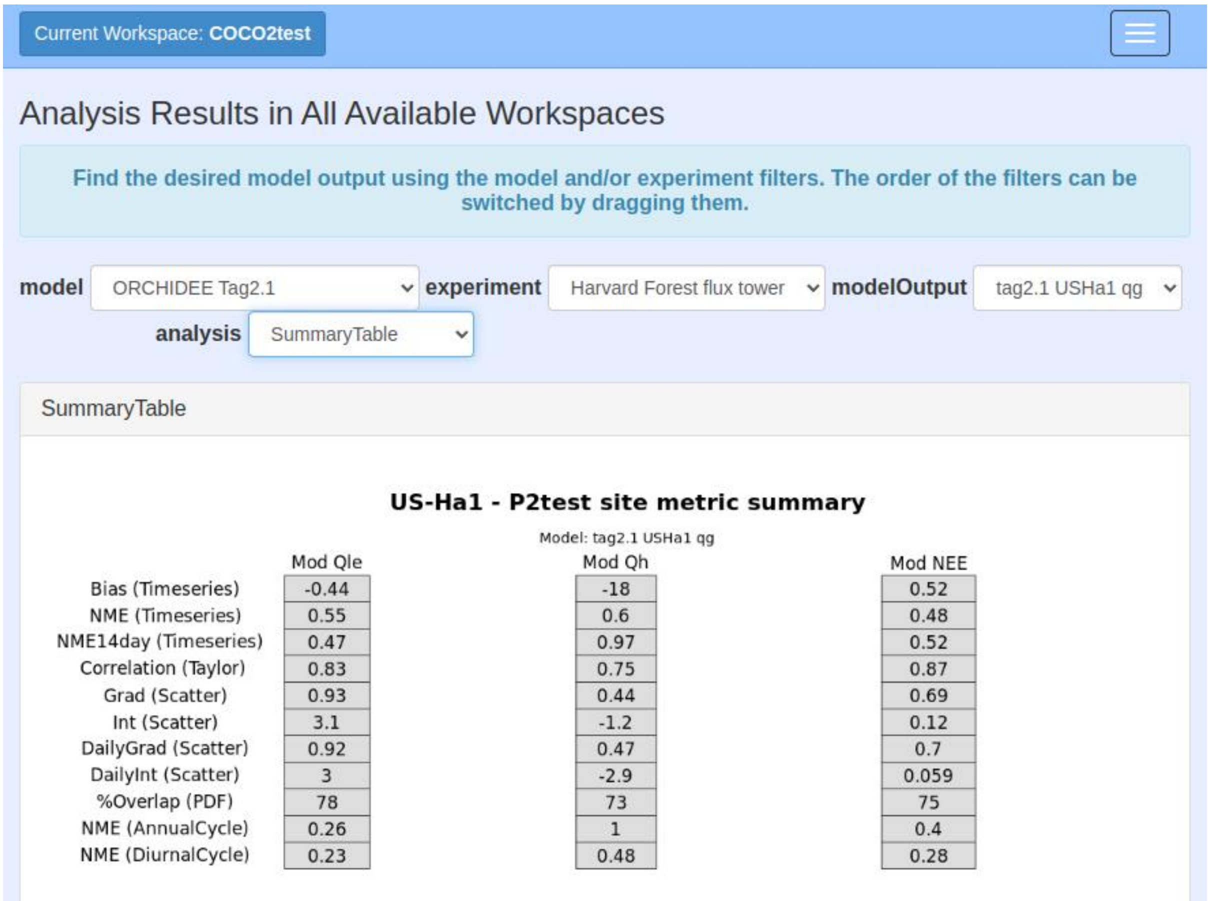


Figure 7: Screenshot of the modelevaluation.org home page, accessed on 14-12-2022.



	Mod Qle	Mod Qh	Mod NEE
Bias (Timeseries)	-0.44	-18	0.52
NME (Timeseries)	0.55	0.6	0.48
NME14day (Timeseries)	0.47	0.97	0.52
Correlation (Taylor)	0.83	0.75	0.87
Grad (Scatter)	0.93	0.44	0.69
Int (Scatter)	3.1	-1.2	0.12
DailyGrad (Scatter)	0.92	0.47	0.7
DailyInt (Scatter)	3	-2.9	0.059
%Overlap (PDF)	78	73	75
NME (AnnualCycle)	0.26	1	0.4
NME (DiurnalCycle)	0.23	0.48	0.28

Figure 8: Example of evaluation metrics calculated from modelevaluation.org, accessed on 14-12-2022.

Accessing the data and experiments requires that the modelling teams create a free account, after which they can access the appropriate workspace which holds the data and analysis scripts for the experiment. These can be updated and refined in time according to the needs of the project.

4 Conclusion

The work described here outlines a framework that allows for flexible assessment of uncertainty and error of estimated biogenic CO_2 using in situ eddy covariance data. The data base here provides the necessary measurements of CO_2 fluxes, meteorological forcing data, remotely sensed vegetation indices, and site characteristics necessary to use in modelling activities. The data was assembled in a way to both meet the requirements within the project, but also to be consistent with existing model evaluation activities to give both consistency error and uncertainty reporting, as well as to facilitate longer term undertakings both within and outside of the project. The additional components of gap filled meteorological data and automated quality control make the dataset well-tailored to modelling and error assessment. Considerable effort in coordination between partners such as CMCC, ICOS ERIC, and the modelevaluation.org team to build a longer term pipeline will allow for more continuous updates of in situ data. The resulting toolbox is also flexible enough to provide for future customized extensions based on needs.

5 References

Best, M.J. *et al.* (2015) 'The Plumbing of Land Surface Models: Benchmarking Model Performance', *Journal of Hydrometeorology*, 16(3), pp. 1425–1442. Available at: https://doi.org/10.1175/JHM-D-14-0158.1.

Drought 2018 Team and ICOS Ecosystem Thematic Centre (2020) 'Drought-2018 ecosystem eddy covariance flux product for 52 stations in FLUXNET-Archive format'. Available at: https://doi.org/10.18160/YVR0-4898.

Hengl, T. *et al.* (2017) 'SoilGrids250m: Global gridded soil information based on machine learning', *PLOS ONE*. Edited by B. Bond-Lamberty, 12(2), p. e0169748. Available at: https://doi.org/10.1371/journal.pone.0169748.

Hersbach, H.,et al. (2018): ERA5 hourly data on single levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 28-Jan-2019), https://doi.org/10.24381/cds.adbb2d47

Pastorello, G. *et al.* (2017) 'The FLUXNET2015 dataset: The longest record of global carbon, water, and energy fluxes is updated', *Eos*, 98.

Pastorello, G. *et al.* (2020) 'The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data', *Scientific Data*, 7(1), p. 225. Available at: https://doi.org/10.1038/s41597-020-0534-3.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Reichstein, M. et al. (2005) 'On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm', Global Change Biology, 11(9), pp. 1424–1439. Available at: https://doi.org/10.1111/j.1365-2486.2005.001002.x.

Walther & Besnard *et al.* (2022) 'Technical note: A view from space on global flux towers by MODIS and Landsat: the FluxnetEO data set', *Biogeosciences*, 19(11), pp. 2805–2840. Available at: https://doi.org/10.5194/bg-19-2805-2022.

Warm Winter 2020 Team and ICOS Ecosystem Thematic Centre (2022) 'Warm Winter 2020 ecosystem eddy covariance flux product for 73 stations in FLUXNET-Archive format—release 2022-1'. Available at: https://doi.org/10.18160/2G60-ZHAK.

Document History

Version	Author(s)	Date	Changes
	Name (Organisation)	dd/mm/yyyy	
V1.0	Jacob A. Nelson	16/12/2022	

Internal Review History

Internal Reviewers	Date	Comments
Marko Scholze	20/12/2022	See comments in document
Hans Chen	20/12/2022	See comments in document

Estimated Effort Contribution per Partner

Partner	Effort
Organisation	effort in person month
MPG	9
Total	**9**