CARDAMOM-FluxVal Version 1.0: a FLUXNET-based Validation System for CARDAMOM Carbon and Water Flux Estimates

Yan Yang[1], A. Anthony Bloom[1], Shuang Ma[1], Paul Levine[1], Alexander Norton[1], Nicholas C. Parazoo[1], John T Reager[1], John Worden[1], Gregory R. Quetin[2], T. Luke Smallman[3,4], Mathew Williams[3,4], Liang Xu[1], Sassan Saatchi[1]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109.
[2]Department of Earth System Science, Stanford University, Stanford, CA 94305, U.S.A.
[3]School of Geosciences, University of Edinburgh, Edinburgh, EH9 3FF, United Kingdom.
[4]National Centre for Earth Observation, University of Edinburgh, Edinburgh, EH9 3FF, United Kingdom.

Correspondence to: Yan Yang (yan.yang@jpl.nasa.gov)

Abstract. Land-atmosphere carbon and water exchanges have large uncertainty in terrestrial biosphere models (TBMs). Using observations to reduce TBM structural and parametric errors and uncertainty are critical priorities for both understanding and accurately predicting carbon and water fluxes. Recent implementations of the Bayesian CARDAMOM model-data fusion framework have yielded key insights into ecosystem carbon and water cycling. CARDAMOM estimates parameters for an associated TBM of intermediate complexity (DALEC). These CARDAMOM analyses—informed by co-located C and H_2O flux observations—have exhibited considerable skill in both representing the variability of assimilated observations and predicting withheld observations. CARDAMOM and DALEC have been continuously developed to accommodate new scientific challenges and an expanding variety of observational constraints. However, so far there has been no concerted effort to globally and systematically validate CARDAMOM performance across individual model-data fusion configurations. Here we use the FLUXNET-2015 dataset—an ensemble of 200+ eddy covariance flux tower sites—to formulate a concerted benchmarking framework for CARDAMOM carbon (photosynthesis and net C exchange) and water (evapotranspiration) flux estimates (CARDAMOM-FLUXVal version 1.0). We present a concise set of skill metrics to evaluate CARDAMOM performance against both assimilated and withheld FLUXNET-2015 photosynthesis, net CO2 exchange and evapotranspiration estimates. We further demonstrate the potential for tailored CARDAMOM evaluations by categorizing performance in terms of (i) individual land cover types, (ii) monthly, annual and mean fluxes, and (iii) length of assimilation data. The CARDAMOM benchmarking system—along with CARDAMOM driver files provided—can be readily repeated to

support both the intercomparison between existing CARDAMOM model configurations and the formulation, development and testing of new CARDAMOM model structures.

1 Introduction:

Terrestrial biosphere models (TBMs) are a key tool to understanding and resolving the state of terrestrial ecosystems and their sensitivity to climate. Of particular importance are land-atmosphere CO_2 fluxes; as the land biosphere is currently a net sink absorbing nearly a third of anthropogenically emitted CO_2 (Friedlingstein et al., 2020). However, despite the importance of TBMs in understanding the role of terrestrial ecosystems in the earth system, model structural and parametric uncertainty remain major sources of error and bias impacting terrestrial carbon cycle modeling (Bonan et al., 2019; Quetin et al., 2020), presenting a major challenge to robust prediction of the magnitude of the land sink in coming decades (Booth et al., 2012; Arora et al., 2020). Improved representation and expression of the ecosystem processes of carbon, water and energy exchanges from and to the atmosphere can improve empirical modelling or data-driven predictions of the key components of the land surface and Earth system and reduce uncertainties (Jung et al., 2020, 2019; Reich, 2010; Tramontana et al., 2016). Model-data fusion (MDF) approaches merging terrestrial biosphere models with observations (Fox et al., 2009; Hill et al., 2012; Keenan et al., 2012; MacBean et al., 2016; Xiao et al., 2014) improve biogeochemical model accuracy and skill by incorporating data from field-based measurements and satellite based remote sensing observations and their associated uncertainties into model calibration. MDF hence offers a much-needed capability to reconcile uncertain model processes with the ever-increasing volume of Earth Observation datasets (Caldararu et al., 2012; Quetin et al., 2020; Richardson et al., 2011; Rowland et al., 2014; Smallman et al., 2017). Specifically, data constrained processes should improve the accuracy of estimates of global plant and soil C dynamics, their exchanges with each other and with the atmosphere, and enable quantification of their uncertainty (Bloom et al., 2016). MDF representations of terrestrial ecosystem C cycling combines the advantage of having a process-based, mathematically refined expression of the ecosystem C budget, and parameter estimation that takes external constraints with their uncertainties into consideration. Contingent on the accuracy of particular model's C cycle mechanisms, MDF can improve simulation results – relative to both assimilated datasets and withheld data from validation – due to improved parameter estimates of biogeochemical processes that may be introduced or influenced by external forcing (Bloom et al., 2020).

The CARbon DAta-MOdel fraMework (CARDAMOM) MDF system approach has been applied to a range of scales and with a wide range of in-situ and satellite datasets to (i) constrain terrestrial C cycle states and processes within a Bayesian model-data fusion framework, and (ii) diagnose these analyses to address questions or test hypotheses on the current and evolving state of the terrestrial C balance (Bloom et al., 2016; Smallman et al., 2017; Yin et al., 2019; Exbrayat et al., 2019; Quetin et al., 2020; Bloom et al., 2020, amongst others). The Data Assimilation Linked Ecosystem Carbon (DALEC; (Williams et al., 2005) model is a key component of CARDAMOM framework describing the ecosystem carbon and water cycles. The DALEC

model has multiple versions varying in structural complexity and process representation(Famiglietti et al., 2021), including alternate forms of climate sensitive phenology (Smallman et al., 2017), time dependent autotrophic respiration processes (Rowland et al., 2014), an array of hydrological representations (Bloom et al., 2016; Bloom and Williams, 2015; Exbrayat et al., 2019; Fox et al., 2009; Quetin et al., 2020; Rowland et al., 2014; Smallman and Williams, 2019; Spadavecchia et al., 2011), expanded representation of heterotrophic respiration sensitivity to climate and explicit representations of ecosystem level water-use efficiency (Bloom et al., 2020) among other model structures.

Invariably, observations play a critical role in (i) informing uncertain processes and reducing model error, (ii) providing a quantitative metric for validating model performance, and (iii) motivate subsequent model process representations. In particular, FLUXNET—an ensemble of C and H_2O flux estimates from 200+ eddy covariance flux tower sites—has been instrumental in the calibration and validation of land surface models (Williams et al., 2009). As one of the most complete and sophisticated field-based databases of land surface fluxes, FLUXNET provides gap-filled measurements of tower-based micrometeorology and eddy covariance estimates of exchanges of carbon dioxide, water vapor, and energy between the biosphere and atmosphere (Schwalm et al., 2010; Pastorello et al., 2020). With the increasing availability (in terms of both spatial coverage and record length) of eddy covariance measurements over participating FLUXNET sites, data-driven methods, or data-assimilation models have become popular and delivered progressively more accurate retrieval results with the aid of remote sensing data for large-scale studies (Anderson et al., 2007; Gonsamo et al., 2012; Velpuri et al., 2013). Gross primary productivity (GPP) and net ecosystem exchange (NEE) are two of the key fluxes in the terrestrial C cycle related to plant growth and the net C sink through vegetation, but they are difficult to measure due to the complications between processes in the biosphere (Gilmanov et al., 2003; Wang et al., 2006). Evapotranspiration (ET) is another key measure related to water, energy and carbon fluxes quantifying the combined process of transpiration, soil evaporation and canopy intercepted rainfall evaporation. The FLUXNET dataset in its entirety is particularly well suited for benchmarking and validating CARDAMOM C and H_2O flux estimates, and number of CARDAMOM-DALEC implementations across FLUXNET sites have demonstrated the scientific and technical merits of assimilating and predicting withheld observations (Bloom and Williams, 2015; Famiglietti et al., 2021; Smallman et al., 2017).

Overall, systemically challenging existing CARDAMOM model structures against observations—and using these outcomes to formulate new model structures—is a necessary process for advancing understanding and prediction of terrestrial C and H_2O fluxes. Among some of the key questions motivating CARDAMOM model-data fusion development decisions are: when trained with observations, do CARDAMOM models improve representation of principal carbon and water dynamics across terrestrial ecosystems? Which CARDAMOM model structures or model-data fusion configurations exhibit optimal predictive skill against withheld flux observations? For a given CARDAMOM model structure, is predictive skill constant, regardless of the training/prediction window, or the length of calibration period correlated with prediction error? Which model parameters or processes are key to the improvement of predictive skill? These questions have continually motivated—and will continue

to motivate—the development of CARDAMOM model structures and associated model-data fusion configurations. Consequently, systematic and easily repeatable evaluations of CARDAMOM outputs against a broad set of C and H_2O fluxes observation would amount to an indispensable strategy for supporting CARDAMOM model developments.

Here, we present "CARDAMOM-FluxVal version 1.0", a concerted FLUXNET-based validation framework to support a global evaluation of CARDAMOM model-data fusion approaches. CARDAMOM-FluxVal provides a validation test-bed for benchmarking CARDAMOM model structures against FLUXNET-2015 GPP, NEE and ET datasets. To demonstrate the operation of the validation framework, we present quantitative assessments of the performance of two example CARDAMOM model configurations—one solely trained by satellite and inventory datasets, and the other trained with an additional constraint using observations from FLUXNET sites. The methodology is described in section 2. In section 3, we present a concise set of validation metrics (against assimilated and withheld FLUXNET observations), and further evaluate performance sensitivity to the choice of constraining variables, temporal length of data assimilation, and particular land cover types. Finally, in section 4 we summarize the strengths and limitations of our CARDAMOM validation approach and outline its potential applications for (i) benchmarking and inter-comparing current and future CARDAMOM configurations, and (ii) provide recommendations and guidance to conduct scientific investigations.

2 Methods

The method section includes descriptions of the CARDAMOM implementation across FLUXNET-2015 sites (Section 2.1), satellite and inventory-based observations used for assimilation (Section 2.2), and the statistical measures used in model validation and extended evaluations (Section 2.3).

2.1 CARDAMOM implementation across FLUXNET-2015 sites

The components needed to configure CARDAMOM at each FLUXNET site namely include (a) time series of meteorological forcing variables for the DALEC model, (b) a collection of observational constraints on DALEC states and fluxes, and (c) additional attributes relating to CARDAMOM prior probability and likelihood functions (Bloom et al., 2020). At each site, we built standalone CARDAMOM "driver" files, which consist of (i) 2001-2015 ERA-interim meteorological forcings from the nearest 0.5 degree grid, based on each site's latitude and longitude value, and (ii) FLUXNET and ancillary observations, including leaf area, and biomass (see section 2.2 and **Fig. S1**). We configured the CARDAMOM model across all FLUXNET-2015 sites during the period of 2001-2015 (204 sites in total, see section 2.2). The observational timespan for each site is from a few months to 15 years, depending on the site characteristics. We chose to implement CARDAMOM for the entirety of the 2001-2015 period at each site in order to exclude the effect of varying CARDAMOM simulation lengths in the subsequent CARDAMOM evaluations. A summary of all FLUXNET-2015 sites used in CARDAMOM-FluxVal here is included in the

supplement (**Table S5**). The aforementioned datasets amount to baseline datasets for the entire CARDAMOM-FluxVal (version 1.0) system. The CARDAMOM-FluxVal driver files are available in the manuscript supplement (**Table S6**).

At each FLUXNET site, we used CARDAMOM Bayesian model-data fusion methodology (Bloom et al., 2020) to calibrate the DALEC model parameters and initial conditions, and to validate DALEC model simulations against a subset of withheld data. In particular, the observations *assimilated* into CARDAMOM were used to optimize DALEC model parameters and initial conditions in order to statistically minimize model-data mismatches. The observations *withheld* from CARDAMOM were used to validate DALEC carbon and water fluxes outside the training window, i.e., in the absence of data constraints. Depending on the scientific or technical objectives, the CARDAMOM-FluxVal analyses can be configured to exclude any subset of FLUXNET or ancillary data for validation purposes. To exemplify both the assimilation and validation aspects of CARDAMOM-FluxVal, we opted for two distinct CARDAMOM configurations (**Fig. 1**):

CARDAMOM analysis A1: the CARDAMOM DALEC model is constrained by the first 50% of FLUXNET data at each site; 50% of FLUXNET data is withheld for validation.

CARDAMOM analysis A2: the CARDAMOM DALEC model is constrained by 0% of FLUXNET data at each site; 100% of FLUXNET data is withheld for validation.

In both A1 and A2, we used the same ancillary data (satellite-based leaf area index, biomass), cost function configurations and DALEC model version. For the sake of brevity, the cost function and DALEC model version are described in the manuscript supplement. To configure the A1 scenario, we spilt the FLUXNET data from each of the site into two periods based on data acquisition time for tower sites with valid observations for the study period from 2001 to 2015.

2.2 Observations

A common set of observations are assimilated into both the A1 and A2 analyses; these consist of (1) timeseries of monthly Moderate Imaging Spectroradiometer (MODIS) leaf area index (LAI) from the MOD15A2H product (Myneni et al., 2015) for the period of 2001-2016, (2) a single estimate of the global above- and below-ground biomass (ABGB) in 2015 produced from a combination of field plots, airborne Lidar and satellite data using the machine learning approach (Yu, 2013). To find corresponding mapped values that match FLUXNET data measurements, we aggregated the mapping products (MODIS LAI and ABGB) from their original resolutions to 1-km spatial resolution and extract LAI and ABGB values at all FLUXNET locations.

For the A1, we also included the gap-filled monthly flux measurements from the FLUXNET2015 Dataset (Pastorello et al., 2020) that includes ecosystem-scale data on CO_2, water, and energy exchange between the biosphere and the atmosphere, and other meteorological and biological measurements collected at sites from the multiple regional flux networks (https://fluxnet.fluxdata.org/data/fluxnet2015-dataset/). We used all 204 CC-BY-4.0 (Tier One) sites to study the data assimilation using GPP, NEE and ET together as inputs (**Table S1**). The pre-processing of FLUXNET tower measurements

includes a quality check to filter out bad-quality monthly data, and the removal of data points where the recorded measurements show constant values throughout the observational period.

2.3 Summary metrics and extended validation

Our summary metrics consist of GPP, ET and NEE evaluated on a monthly basis, annual basis, and at site level. We selected four statistical metrics to evaluate the model accuracy, parameter correlations, and residuals (**Table S2**). The Pearson's linear correlation coefficient (R) is the ratio of covariance between the modelled simulations and observations to the product of standard deviations from model simulations and observations ($0<R<1$ represents a positive correlation between model output and observed values, while $-1<R<0$ means the model outputs have a negative correlation between model output and observed values). The Nash and Sutcliffe model efficiency (MEF) quantifies the model's predictive capacity (Nash and Sutcliffe, 1970; Tramontana et al., 2016). A $0<MEF<1$ indicates the model's predictive capacity is better than the mean of observations, with a value of 1 meaning perfect predictions; while $MEF<0$ means the mean values of the observations is better than the model predictions. BIAS is defined as the mean of the residual values for model predictions and observed data. A value of BIAS near zero indicates an unbiased estimation for model predictions. The Root Mean Square Error (RMSE) is the square root of the average over squared residuals (prediction errors), and the model predictions are more accurate when RMSE is closer to 0.

For the extended evaluation, we grouped the FLUXNET 2015 sites within 6 time-window categories: data with 1:1 assimilation/prediction time ranges spanning from < 1 year, 1 - 2 years, 2 - 3 years, 3 - 4 years, 4-5 years and >5 years (all time ranges are either assimilation or prediction lengths). The number of sites varies from 17 to 67 for different categories (**Table S3**), with the most sites (67) having the range of 1~2 years, and the least sites (17) having the range of 4-5 years. We evaluated CARDAMOM performance across 12 land-cover types that comprise the FLUXNET2015 sites included in this study (**Table S3**). In summary, ENF (Evergreen Needleleaf Forest) and GRA (Grasslands) have more than 30 tower sites, while SNO (Snow) and CSH (Closed Shrublands) have only one and two sites globally. Assuming that the CARDAMOM model has valid outputs for GPP, NEE and ET across different land-cover types, we evaluated the influence of land-cover types on the prediction accuracies.

We tested the importance of model parameters on the retrievals of GPP, NEE and ET, by calculating each parameters' correlations with the model residuals. A total of 36 model parameters (model description in **SI text**) were tested and attributed into 6 groups based on their relative contributions to different biophysical processes (**Table S4**). We tested the correlations between model parameters and retrieval residuals using the R metric for independent validation data sets.

3 Results

3.1 Summary metrics for CARDAMOM FLUXNET validation

We found good agreements between median model outputs from CARDAMOM-DALEC and site-based FLUXNET observations (GPP, NEE and ET; **Fig. 2**) for the A1 scenario. Generally, data samples used in assimilation window show better agreements between observations and simulations (i.e., higher MEF and lower RMSE) than the data in the prediction window. Monthly-based comparisons, due to the seasonal variation in each variable, have a wider data range than the range of site-level data. The MEF metrics show GPP has the best simulation results in both the assimilation and prediction windows relative to NEE and ET. Furthermore, NEE presents a better MEF in the assimilation window than ET, but worse than ET in the prediction window. The same pattern is clearer in the site-level scatter plots when we only compare the long-term average observations for each FLUXNET site. In the A1 scenario, we obtained the highest MEF at the site-level comparison during the assimilation window (e.g., NEE; **Fig. 2**), but the lowest MEF during the prediction window, indicating the assimilation procedure may be overfitting to the observations.

The model-data residual analysis show that it is possible to improve the cross-validated model outputs and reduce biases and structure errors with assimilation of FLUXNET observations (**Figs. 3-4, S2-S4**). Histograms of monthly-based residuals at the monthly timescales over all sites (**Fig. 3**) show A1 gives less-biased model residuals than the outputs of A2. In general, A1 shows positive NEE bias of 0.36 gC m^{-2} per day, and negative GPP and ET biases of 0.36 gC m^{-2} per day and 0.09 gC mm per day, respectively, while A2 shows much big biases (NEE bias: +1.03 gC m^{-2} per day, GPP bias: -1.34 gC m^{-2} per day, ET biases: -0.55 mm per day). Annual-based distributions (**Fig. S2**) of model retrieval residuals show similar patterns to monthly residuals, except that A1 show tighter distributions around zero due to the average of seasonal variations. The temporal average of site level histograms (**Fig. 4**) preserves spatial characteristics of the model retrieval residuals. Unsurprisingly, A2 has more outliers than A1 at the site-level scale. Predicted absolute values (GPP, NEE and ET) instead of residual, show a wider range of distributions (**Fig. S3**) for A1 than A2, suggesting A1 runs capture more spatial and temporal variability, with higher accuracies and lower biases. The comparisons of second-order distribution (standard deviation of distribution) provide additional evidences that A1 have closer ranges to the observed distributions (**Fig. S4**).

The constrained runs of CARDAMOM model (A1) show substantial improvements in both matching the FLUXNET observations and reducing the model output uncertainties (**Fig. 1**). In other words, the added value of data in A1—relative to A2—leads to more accurate predictions of GPP and ET, and reasonable NEE. Two well-studied long-term research sites (US-Ha and US-UMB) in the United States show that the model outputs of A1 capture the stronger seasonality of NEE than the outputs of A2 (**Fig. 1B and 1C**), show the weaker seasonality patterns. Especially during the peak of growing seasons, NEE has a strong land C sink observed from tower sites, but model outputs of A2 are systematically lower in terms of C sink magnitudes. Both A1 and A2 can capture seasonal changes of GPP and ET within the model estimated confidence intervals

(CI). However, the CI bounds reduce significantly for A1 (e.g., the 90% CI bound of ET from A2 is ~±2.5 mm/day during the peak growing seasons, and it is reduced to ~±1.5 mm/day for A1 at the selected US tower sites) due to the data assimilation process using site-level observations.

3.2 Extended assessment of CARDAMOM performance

The CARDAMOM-simulated fluxes are more sensitive to certain ecosystem parameters than others (**Fig. 5**). Results show that the modelled GPP is mostly correlated with the model parameters C_1 – canopy efficiency, A_1 – autotropic respiration and W_1 – underlying water-use efficiency (see manuscript supplement for parameter details); these 3 parameters stand out as they are positively related to GPP variation with Pearson's R greater than 0.1, while the R values for all other parameters are near zero. For the NEE output, parameter I_6 – soil organic carbon (SOM) is the most negatively correlated factor with NEE and parameter T_6 – SOM turnover rate is the most positively correlated. However, none of the R values for NEE has a magnitude > 0.1. The output of ET is also correlated with 3 parameters, W_1 (underlying water-use efficiency), W_2 (runoff coefficient), and W_5 (radiation coefficient), with W_1 being negatively correlated with ET and the other two positively correlated. All three parameters stand out to be substantially different from all other model parameters, indicating the crucial impact of these parameters on the ET output. As expected, the A1 experiment shows reduced uncertainty in a few estimated parameters when compared to the A2 experiment, indicating the additional use of observational data imposes constraints on model parameters as well **(Fig. S5)**.

Based on the major land cover types classified at the FLUXNET tower sites, we investigated the effects of land cover on the performance of CARDAMOM model retrieval. Results show that the forest types, except the evergreen broadleaf forest, generally have more accurate predictions than non-forest types (**Fig. 6**). The three major types of forests – deciduous broadleaf forest (DBF), evergreen needleleaf forest (ENF) and mixed forest (MF), all have high R (> 0.8) and MEF (> 0.6) values. The relatively small uncertainty ranges (< 0.1 for R) also indicate the stable performance of these forest types. The evergreen broadleaf forest (EBF) in the tropics, though fewer sites are available (half of DBF and one third of ENF), exhibit the difficulties in retrievals with lower performance values and higher uncertainty ranges.

For non-forest sites, the retrieval accuracy varies from site to site (**Fig. 6)** and have large uncertainties. In particular, savannas, woody savannas and closed shrublands are the three land cover types showing the least accuracies and highest uncertainty, and significantly in the NEE and ET retrievals (R ~ 0.6 and MEF being negative). Other herbaceous vegetation types, including grasslands and crops, have generally better retrievals than spatially heterogeneous land cover types such as savannas, but not as good as retrievals over extratropical forests (Smallman and Williams, 2019).

The FLUXNET data set has various lengths of observations in time (**Table S3**). Separating the results by the length of assimilations, we show that the CARDAMOM model has slightly better predictions of GPP, NEE and ET when the assimilation

period is longer (**Fig. 7**). The metric MEF for GPP and NEE increases from values below zeros to the maximum positive when the assimilation period reaches 4-5 years. The median of MEF of ET always stays positive, but also has a maximum value at the length of 4-5 years for data assimilation. Meanwhile, the R values show relatively small changes for different lengths of data assimilation, and most values are above 0.8, indicating reasonable assimilations for GPP, NEE and ET in general. There is a slightly degraded performance in R (a decrease by < 0.1) and MEF (a decrease by 0.2-0.3) for the longest assimilation period (>5 years), probably due to the increased size of FLUXNET sites, resulting in the inclusion of certain sites (e.g., tropical forests and/or woody savannas) with known bad performances compared to others. For the sites with record lengths of 2-3 years, the percentage of the non-forest PFT (grassland) is higher than other year ranges. The lack of nonforest sites could possibly be the cause of worst performance for this length of observations. With long assimilation windows, there is also a general trend of reduced uncertainty for both NEE and ET predictions. GPP has a reduction in uncertainty for longer training windows till 4-5 years, and increases for the longest assimilation period (>5 years).

4 Discussion

4.1 Assessing CARDAMOM performance

The FLUXNET-based validation approach has provided some key insights on the skill of CARDAMOM-based C and H_2O flux estimates. (1) The data assimilation using FLUXNET inputs (A1) captures missing seasonal variations in the original model with lower biases and less uncertainty, compared to the model solely constrained by satellite and inventory datasets (A2); (2) The increased lengths of data assimilation can progressively improve the model performance and reduce the predictive uncertainties in all tested flux variables; (3) Land cover types still exhibit influences on the model prediction accuracy, even though the parameters were locally adjusted in the assimilation process, consistent with earlier studies using global parametrization (Smallman and Williams, 2019); (4) Certain parameters (i.e. C_1, A_1 and W_1) show more distinct correlations with model outputs, suggesting that improved prior constraints on a subset of parameters could further improve the retrieval accuracies of the corresponding outputs; (5) The validation results also highlight that more work should be focused on the tropical vegetation, where both the humid forests and savanna regions exhibit the worst performance; the lack of regular seasonal cycles may also hamper the accurate retrievals for CARDAMOM and other models (Quetin et al., 2020).

The aforementioned insights are key for identifying seasonal and inter-annual limitations in CARDAMOM model performance, limitations (or lack thereof) in the ability of CARDAMOM model structures to predict C and H2O fluxes on a range of timescales, and limitations of CARDAMOM across specific biomes or land-cover types. The results can be further used to target future CARDAMOM model developments towards identifying weaknesses in improving predictive skill. With more spatially explicit products becoming available for assimilation into CARDAMOM—such as satellite-based constraints on GPP and NEE (Quetin et al., 2020)—this study based on FLUXNET sites can also provide a quantitative characterization of CARDAMOM model structure

4.2 Limitations of FLUXNET validation approach

One noteworthy caveat is the spatial resolution representation errors in the DALEC meteorological forcing. Specifically, the meteorological data used in this study are from the ECMWF ERA-interim dataset, projected at a 0.5° resolution. The disagreement in spatial resolution may be a confounding factor for CARDAMOM FLUXNET predictions. Implementing CARMAMOM using a finer resolution meteorological forcing will help to reduce the uncertainty caused by spatial ambiguities (see SI text S2 for replacing meteorological forcing data). Potential approaches for future versions of CARDAMOM-FluxVal include (i) using gap-filled products from FLUXNET sites to configure CARDAMOM simulations, and/or (ii) transitioning to ERA5 meteorological forcing. However, the current version has not rigorously tested the new meteorological forcing data sets. And the improvement of all drivers to a finer resolution requires modification of other ancillary data sets that are used to determine variables such as CO_2 concentration, burned area and VPD **(Table S6)**, which is an ongoing effort for the new CARMAMOM version.

We also note scarcity of tropical tower sites across the FLUXNET2015 dataset (Schimel et al., 2015) may ultimately lead to biased assessments of CARDAMOM model structures. The possible heterogeneity for nonforest tower sites also causes more uncertainty in observed variables as well as the meteorological forcing due to resolution issues. On the other hand, our PFT-level analysis could also reveal potential model structure limitations in simulating certain PFT with reasonable assumptions, which needs further attention when the caveat due to observational uncertainty is ruled out. While we advocate for the use of global summary metrics to assess model structure, we also recommend users of this validation approach recognize the variable representation of biomes and vegetation classes in the available observational datasets. In addition to extended analyses (section 3.2), we also recommend projecting validation assessments into climate space (Reichstein et al., 2003).

4.3 Applications

The summary metrics (section 3.1) provide an easily reproducible set of statistics for the validation framework for monthly and inter-annual CARDAMOM carbon and water flux estimates. While our results show the importance of observational constraints (in this study, FLUXNET data), the CARDAMOM validation system can be readily applied to test additional configurations (alternative models, cost function parameters, datasets assimilated and assimilation: prediction configurations). With a number parametric and structural variations in existing CARDAMOM framework model structures (Famiglietti et al., 2021)—as well as anticipated variations among ongoing CARDAMOM developments—we highlight the need for a concerted and easily repeatable validation system. In particular, we recommend the use of the CARDAMOM-FluxVal validation approach for three categories of CARDAMOM developments:

1. **DALEC model structures.** The growing diversity of DALEC models (Famiglietti et al., 2021) provides a unique opportunity for determining which model structures and process representations best predict assimilated or withheld carbon and water fluxes. Further investigations can also be conducted with the exclusion and/or adaptation of

ecological & dynamic constraints (Bloom and Williams, 2015; Smallman et al., 2021). Models of similar complexity to DALEC can also be used.

2. **CARDAMOM cost function.** Model-data error characterization in the CARDAMOM multi-objective optimization approach discussed in (Bloom et al., 2020) are inherently limited. The FLUXNET validation approach can be used (i) for quantitative characterization of DALEC (or alternate model) accuracy and precision based on error characterization choices, and (ii) test potential improvements in error characterizations, such as optimizable uncertainty coefficients and the error models (Norton and Uryasev, 2019; Schoups and Vrugt, 2010). These analyses can be further extended to quantify the added value of individual data streams (e.g., by sequential removal of individual observation types).
3. **CARDAMOM MDF algorithms.** CARDAMOM employs an adaptive Metropolis-Hastings Markov-chain Monte Carlo. The validation framework can be used to quantify the effectiveness of DALEC predictions using faster methods (e.g., optimal estimation, Rodgers, 2000), or previously established optimization algorithms (Fox et al., 2009). Experiments could be expanded to include dedicated studies for comparing the effectiveness of CARDAMOM analyses against non-CARDAMOM model-data fusion efforts (Bacour et al., 2019; Liu et al., 2021; MacBean et al., 2016) and machine learning methodologies (Jung et al., 2020, 2019, 2017; Tramontana et al., 2016).

We anticipate that the CARDAMOM FLUXNET validation framework will provide a much-needed quantitative benchmark to support and inform future CARDAMOM framework developments. Specifically, validation and inter-comparison experiments can span well beyond the two CARDAMOM configurations presented in this study (A1 and A2), and can be adapted to suit the individual needs for CARDAMOM developments or scientific investigations.

Acknowledgments

This work was carried out at the Jet Propulsion Laboratory, California Institute for technology, under a contract with the National Aeronautics and Space Administration. We acknowledge Dr. Alexandra Konings from Stanford University for thorough manuscript feedback and review.

Author contributions: A.B. and Y.Y. designed the research framework and performed the model validation using FLUXNET data. Y.Y and S.M. tested the integrity and validity of the code and model. L.X and S.S provided the global biomass data. All authors contributed to the writing of final manuscript.

Competing interests: The authors declare that they have no competing interests.

Codes and data availability: The CARDAMOM code used in this manuscript is available at https://github.com/CARDAMOM-framework/CARDAMOM_v2.2. CARDAMOM-FLUXVAL version 1.0 code and driver

datasets (including the CARDAMOM version used in this analysis) are tagged in the GitHub link. The code, along with the full output datasets, are permanently stored in (Yang et al., 2021). Instructions on the code implementation are provided in the manuscript supplementary information (section S2).

References

Anderson, M. C., Kustas, W. P., and Norman, J. M.: Upscaling Flux Observations from Local to Continental Scales Using Thermal Remote Sensing, Agron. J., 99, 240–254, https://doi.org/10.2134/agronj2005.0096S, 2007.

Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., Schwinger, J., Bopp, L., Boucher, O., Cadule, P., Chamberlain, M. A., Christian, J. R., Delire, C., Fisher, R. A., Hajima, T., Ilyina, T., Joetzjer, E., Kawamiya, M., Koven, C. D., Krasting, J. P., Law, R. M., Lawrence, D. M., Lenton, A., Lindsay, K., Pongratz, J., Raddatz, T., Séférian, R., Tachiiri, K., Tjiputra, J. F., Wiltshire, A., Wu, T., and Ziehn, T.: Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison to CMIP5 models, Biogeosciences, 17, 4173–4222, https://doi.org/10.5194/bg-17-4173-2020, 2020.

Bacour, C., Maignan, F., Peylin, P., MacBean, N., Bastrikov, V., Joiner, J., Köhler, P., Guanter, L., and Frankenberg, C.: Differences Between OCO-2 and GOME-2 SIF Products From a Model-Data Fusion Perspective, J. Geophys. Res. Biogeosciences, 124, 3143–3157, https://doi.org/10.1029/2018JG004938, 2019.

Bloom, A. A. and Williams, M.: Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological "common sense" in a model-data fusion framework, Biogeosciences, 12, 1299–1299, 2015.

Bloom, A. A., Exbrayat, J.-F., Velde, I. R. van der, Feng, L., and Williams, M.: The decadal state of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and residence times, Proc. Natl. Acad. Sci., 113, 1285–1290, https://doi.org/10.1073/pnas.1515160113, 2016.

Bloom, A. A., Bowman, K. W., Liu, J., Konings, A. G., Worden, J. R., Parazoo, N. C., Meyer, V., Reager, J. T., Worden, H. M., Jiang, Z., Quetin, G. R., Smallman, T. L., Exbrayat, J.-F., Yin, Y., Saatchi, S. S., Williams, M., and Schimel, D. S.: Lagged effects dominate the inter-annual variability of the 2010–2015 tropical carbon balance, Biogeosciences Discuss., 1–49, https://doi.org/10.5194/bg-2019-459, 2020.

Bonan, G. B., Lombardozzi, D. L., Wieder, W. R., Oleson, K. W., Lawrence, D. M., Hoffman, F. M., and Collier, N.: Model Structure and Climate Data Uncertainty in Historical Simulations of the Terrestrial Carbon Cycle (1850–2014), Glob. Biogeochem. Cycles, 33, 1310–1326, https://doi.org/10.1029/2019GB006175, 2019.

Booth, B. B. B., Jones, C. D., Collins, M., Totterdell, I. J., Cox, P. M., Sitch, S., Huntingford, C., Betts, R. A., Harris, G. R., and Lloyd, J.: High sensitivity of future global warming to land carbon cycle processes, Environ. Res. Lett., 7, 024002, https://doi.org/10.1088/1748-9326/7/2/024002, 2012.

Caldararu, S., Palmer, P. I., and Purves, D. W.: Inferring Amazon leaf demography from satellite observations of leaf area index, Biogeosciences, 9, 1389–1404, https://doi.org/10.5194/bg-9-1389-2012, 2012.

Exbrayat, J.-F., Bloom, A. A., Carvalhais, N., Fischer, R., Huth, A., MacBean, N., and Williams, M.: Understanding the Land Carbon Cycle with Space Data: Current Status and Prospects, Surv. Geophys., 40, 735–755, https://doi.org/10.1007/s10712-019-09506-2, 2019.

Famiglietti, C. A., Smallman, T. L., Levine, P. A., Flack-Prain, S., Quetin, G. R., Meyer, V., Parazoo, N. C., Stettz, S. G., Yang, Y., Bonal, D., Bloom, A. A., Williams, M., and Konings, A. G.: Optimal model complexity for terrestrial carbon cycle prediction, Biogeosciences, 18, 2727–2754, https://doi.org/10.5194/bg-18-2727-2021, 2021.

Fox, A., Williams, M., Richardson, A. D., Cameron, D., Gove, J. H., Quaife, T., Ricciuto, D., Reichstein, M., Tomelleri, E., Trudinger, C. M., and Van Wijk, M. T.: The REFLEX project: Comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data, Agric. For. Meteorol., 149, 1597–1615, https://doi.org/10.1016/j.agrformet.2009.05.002, 2009.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S., Aragão, L. E. O. C., Arneth, A., Arora, V., Bates, N. R., Becker, M., Benoit-Cattin, A., Bittig, H. C., Bopp, L., Bultan, S., Chandra, N., Chevallier, F., Chini, L. P., Evans, W., Florentie, L., Forster, P. M., Gasser, T., Gehlen, M., Gilfillan, D., Gkritzalis, T., Gregor, L., Gruber, N., Harris, I., Hartung, K., Haverd, V., Houghton, R. A., Ilyina, T., Jain, A. K., Joetzjer, E., Kadono, K., Kato, E., Kitidis, V., Korsbakken, J. I., Landschützer, P., Lefèvre, N., Lenton, A., Lienert, S., Liu, Z., Lombardozzi, D., Marland, G., Metzl, N., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-I., Niwa, Y., O'Brien, K., Ono, T., Palmer, P. I., Pierrot, D., Poulter, B., Resplandy, L., Robertson, E., Rödenbeck, C., Schwinger, J., Séférian, R., Skjelvan, I., Smith, A. J. P., Sutton, A. J., Tanhua, T., Tans, P. P., Tian, H., Tilbrook, B., van der Werf, G., Vuichard, N., Walker, A. P., Wanninkhof, R., Watson, A. J., Willis, D., Wiltshire, A. J., Yuan, W., Yue, X., and Zaehle, S.: Global Carbon Budget 2020, Earth Syst. Sci. Data, 12, 3269–3340, https://doi.org/10.5194/essd-12-3269-2020, 2020.

Gilmanov, T. G., Verma, S. B., Sims, P. L., Meyers, T. P., Bradford, J. A., Burba, G. G., and Suyker, A. E.: Gross primary production and light response parameters of four Southern Plains ecosystems estimated using long-term CO2-flux tower measurements, Glob. Biogeochem. Cycles, 17, https://doi.org/10.1029/2002GB002023, 2003.

Gonsamo, A., Chen, J. M., Wu, C., and Dragoni, D.: Predicting deciduous forest carbon uptake phenology by upscaling FLUXNET measurements using remote sensing data, Agric. For. Meteorol., 165, 127–135, https://doi.org/10.1016/j.agrformet.2012.06.006, 2012.

Hill, T. C., Ryan, E., and Williams, M.: The use of CO2 flux time series for parameter and carbon stock estimation in carbon cycle research, Glob. Change Biol., 18, 179–193, https://doi.org/10.1111/j.1365-2486.2011.02511.x, 2012.

Jung, M., Reichstein, M., Schwalm, C. R., Huntingford, C., Sitch, S., Ahlström, A., Arneth, A., Camps-Valls, G., Ciais, P., Friedlingstein, P., Gans, F., Ichii, K., Jain, A. K., Kato, E., Papale, D., Poulter, B., Raduly, B., Rödenbeck, C., Tramontana, G., Viovy, N., Wang, Y.-P., Weber, U., Zaehle, S., and Zeng, N.: Compensatory water effects link yearly global land CO_2 sink changes to temperature, Nature, 541, 516–520, https://doi.org/10.1038/nature20780, 2017.

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., and Reichstein, M.: The FLUXCOM ensemble of global land-atmosphere energy fluxes, Sci. Data, 6, 74, https://doi.org/10.1038/s41597-019-0076-8, 2019.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., Chevallier, F., Gans, F., Goll, D. S., Haverd, V., Koehler, P., Ichii, K., Jain, A. K., Liu, J., Lombardozzi, D., Nabel, J. E. M. S., Nelson, J. A., O'Sullivan, M., Pallandt, M., Papale, D., Peters, W., Pongratz, J., Rödenbeck, C., Sitch, S., Tramontana, G., Walker, A., Weber, U., and Reichstein, M.: Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach, Biogeosciences, 17, 1343–1365, https://doi.org/10.5194/bg-17-1343-2020, 2020.

Keenan, T. F., Davidson, E., Moffat, A. M., Munger, W., and Richardson, A. D.: Using model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling, Glob. Change Biol., 18, 2555–2569, https://doi.org/10.1111/j.1365-2486.2012.02684.x, 2012.

Liu, Y., Holtzman, N. M., and Konings, A. G.: Global ecosystem-scale plant hydraulic traits retrieved using model–data fusion, Hydrol. Earth Syst. Sci., 25, 2399–2417, https://doi.org/10.5194/hess-25-2399-2021, 2021.

MacBean, N., Peylin, P., Chevallier, F., Scholze, M., and Schürmann, G.: Consistent assimilation of multiple data streams in a carbon cycle data assimilation system, Geosci. Model Dev., 9, 3569–3588, https://doi.org/10.5194/gmd-9-3569-2016, 2016.

Myneni, R., Knyazikhin, Y., and Park, T.: MOD15A2H MODIS/terra leaf area index/FPAR 8-day L4 global 500 m SIN grid V006, NASA EOSDIS Land Process. DAAC, 2015.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Norton, M. and Uryasev, S.: Maximization of AUC and Buffered AUC in binary classification, Math. Program., 174, 575–612, https://doi.org/10.1007/s10107-018-1312-2, 2019.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Ribeca, A., van Ingen, C., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., De Cinti, B., de Grandcourt, A., De Ligne, A., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., di Tommasi, P., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., Gharun, M., Gianelle, D., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, Sci. Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

Quetin, G. R., Bloom, A. A., Bowman, K. W., and Konings, A. G.: Carbon Flux Variability From a Relatively Simple Ecosystem Model With Assimilated Data Is Consistent With Terrestrial Biosphere Model Estimates, J. Adv. Model. Earth Syst., 12, e2019MS001889, https://doi.org/10.1029/2019MS001889, 2020.

Reich, P. B.: The Carbon Dioxide Exchange, Science, 329, 774–775, https://doi.org/10.1126/science.1194353, 2010.

Reichstein, M., Rey, A., Freibauer, A., Tenhunen, J., Valentini, R., Banza, J., Casals, P., Cheng, Y., Grünzweig, J. M., Irvine, J., Joffre, R., Law, B. E., Loustau, D., Miglietta, F., Oechel, W., Ourcival, J.-M., Pereira, J. S., Peressotti, A., Ponti, F., Qi, Y., Rambal, S., Rayment, M., Romanya, J., Rossi, F., Tedeschi, V., Tirone, G., Xu, M., and Yakir, D.: Modeling temporal and large-scale spatial variability of soil respiration from soil water availability, temperature and vegetation productivity indices, Glob. Biogeochem. Cycles, 17, https://doi.org/10.1029/2003GB002035, 2003.

Richardson, A. D., Dail, D. B., and Hollinger, D. Y.: Leaf area index uncertainty estimates for model–data fusion applications, Agric. For. Meteorol., 151, 1287–1292, https://doi.org/10.1016/j.agrformet.2011.05.009, 2011.

Rodgers, C. D.: Inverse Methods For Atmospheric Sounding: Theory And Practice, World Scientific, 256 pp., 2000.

Rowland, L., Malhi, Y., Silva-Espejo, J. E., Farfán-Amézquita, F., Halladay, K., Doughty, C. E., Meir, P., and Phillips, O. L.: The sensitivity of wood production to seasonal and interannual variations in climate in a lowland Amazonian rainforest, Oecologia, 174, 295–306, https://doi.org/10.1007/s00442-013-2766-9, 2014.

Schimel, D., Pavlick, R., Fisher, J. B., Asner, G. P., Saatchi, S., Townsend, P., Miller, C., Frankenberg, C., Hibbard, K., and Cox, P.: Observing terrestrial ecosystems and the carbon cycle from space, Glob. Change Biol., 21, 1762–1776, https://doi.org/10.1111/gcb.12822, 2015.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resour. Res., 46, https://doi.org/10.1029/2009WR008933, 2010.

Schwalm, C. R., Williams, C. A., Schaefer, K., Arneth, A., Bonal, D., Buchmann, N., Chen, J., Law, B. E., Lindroth, A., Luyssaert, S., Reichstein, M., and Richardson, A. D.: Assimilation exceeds respiration sensitivity to drought: A FLUXNET synthesis, Glob. Change Biol., 16, 657–670, https://doi.org/10.1111/j.1365-2486.2009.01991.x, 2010.

Smallman, T. L. and Williams, M.: Description and validation of an intermediate complexity model for ecosystem photosynthesis and evapotranspiration: ACM-GPP-ETv1, Geosci. Model Dev., 12, 2227–2253, https://doi.org/10.5194/gmd-12-2227-2019, 2019.

Smallman, T. L., Exbrayat, J.-F., Mencuccini, M., Bloom, A. A., and Williams, M.: Assimilation of repeated woody biomass observations constrains decadal ecosystem carbon cycle uncertainty in aggrading forests, J. Geophys. Res. Biogeosciences, 122, 528–545, https://doi.org/10.1002/2016JG003520, 2017.

Smallman, T. L., Milodowski, D. T., Neto, E. S., Koren, G., Ometto, J., and Williams, M.: Parameter uncertainty dominates C cycle forecast errors over most of Brazil for the 21st Century, Earth Syst. Dyn. Discuss., 1–52, https://doi.org/10.5194/esd-2021-17, 2021.

Spadavecchia, L., Williams, M., and Law, B. E.: Uncertainty in predictions of forest carbon dynamics: separating driver error from model error, Ecol. Appl., 21, 1506–1522, https://doi.org/10.1890/09-1183.1, 2011.

Tramontana, G., Jung, M., Camps-Valls, G., Ichii, K., Ráduly, B., Reichstein, M., Schwalm, C. R., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences Discuss., https://doi.org/10.3929/ethz-b-000118483, 2016.

Velpuri, N. M., Senay, G. B., Singh, R. K., Bohms, S., and Verdin, J. P.: A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET, Remote Sens. Environ., 139, 35–49, https://doi.org/10.1016/j.rse.2013.07.013, 2013.

Wang, W., Davis, K. J., Cook, B. D., Butler, M. P., and Ricciuto, D. M.: Decomposing CO2 fluxes measured over a mixed ecosystem at a tall tower and extending to a region: A case study, J. Geophys. Res. Biogeosciences, 111, https://doi.org/10.1029/2005JG000093, 2006.

Williams, M., Schwarz, P. A., Law, B. E., Irvine, J., and Kurpius, M. R.: An improved analysis of forest carbon dynamics using data assimilation, Glob. Change Biol., 11, 89–105, https://doi.org/10.1111/j.1365-2486.2004.00891.x, 2005.

An improved analysis of forest carbon dynamics using data assimilation - Williams - 2005 - Global Change Biology - Wiley Online Library: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2486.2004.00891.x?casa_token=hdKTJT94NH4AAAAA%3AJ96UguJjaCF4e7EMO1-qMEqDaEIw6Ci4ajswkVqqFOMm6s9y8JBGAo9kU9Qw_tRVxsLhhz8J9rZmuwodCg, last access: 20 February 2020.

Williams, M., Richardson, A. D., Reichstein, M., Stoy, P. C., Peylin, P., Verbeeck, H., Carvalhais, N., Jung, M., Hollinger, D. Y., Kattge, J., Leuning, R., Luo, Y., Tomelleri, E., Trudinger, C. M., and Wang, Y.-P.: Improving land surface models with FLUXNET data, Biogeosciences, 6, 1341–1359, https://doi.org/10.5194/bg-6-1341-2009, 2009.

Xiao, J., Davis, K. J., Urban, N. M., and Keller, K.: Uncertainty in model parameters and regional carbon fluxes: A model-data fusion approach, Agric. For. Meteorol., 189–190, 175–186, https://doi.org/10.1016/j.agrformet.2014.01.022, 2014.

Yang, Y., Bloom, A. A., Ma, S., Levine, P., Norton, A., Parazoo, N. C., Reager, J. T., Worden, J., Quetin, G. R., Smallman, T. L., Williams, M., Xu, L., and Saatchi, S.: CARDAMOM-FluxVal Version 1.0, https://doi.org/10.5281/zenodo.4904195, 2021.

Yu, Y.: Global Distribution of Carbon Stock in Live Woody Vegetation, PhD Dissertation, University of California, Los Angeles, 122 pp., 2013.

US-Ha1	Var	R	MEF	Bias	RMSE
A1	GPP	0.95	0.89	-0.27	1.55
	NEE	0.90	0.80	-0.03	1.40
	ET	0.96	0.90	-0.01	0.26
A2	GPP	0.96	0.72	-0.65	1.33
	NEE	0.83	-1.96	1.1	1.78
	ET	0.96	0.87	-0.09	0.24

US-UMB	Var	R	MEF	Bias	RMSE
A1	GPP	0.99	0.97	-0.14	0.70
	NEE	0.97	0.91	0.28	0.60
	ET	0.99	0.92	0.26	0.20
A2	GPP	0.98	0.95	0.02	0.84
	NEE	0.95	0.18	1.07	0.78
	ET	0.98	0.85	-0.35	0.25

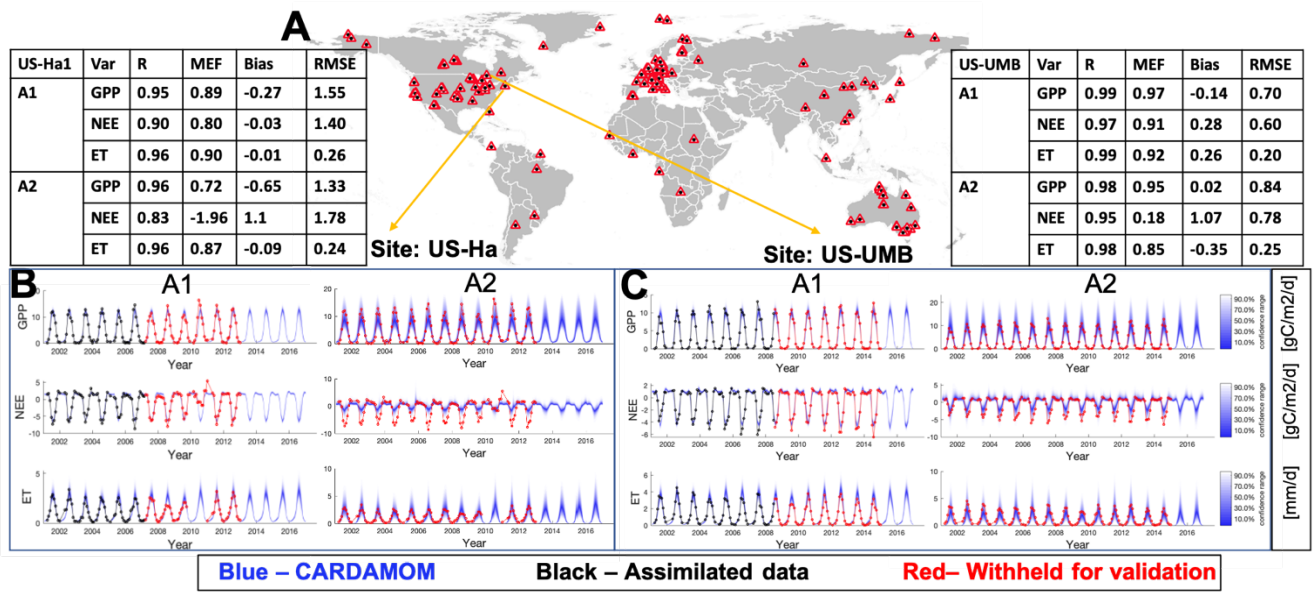


Figure 1. Performance of CARDAMOM model simulations (with 50% of Fluxnet data (Analysis 1: A1) and No Fluxnet data (Analysis 2: A2)). (A) Spatial distribution of FLUXNET tower sites (Tier-1 data). (B) The same time series for the US-Ha site; Note that blue lines in GPP, NEE and ET time series are outputs from model simulations. (C) Time series of CARDAMOM simulations for the US-UMB site; The observed time series from Flux towers are also plotted for comparison. Black lines are the first 50% of FLUXNET observations used for data assimilation, while the red lines are the rest 50% of FLUXNET observations used for validation. The validation metrics in the tables are all from the prediction window for the selected two sites.

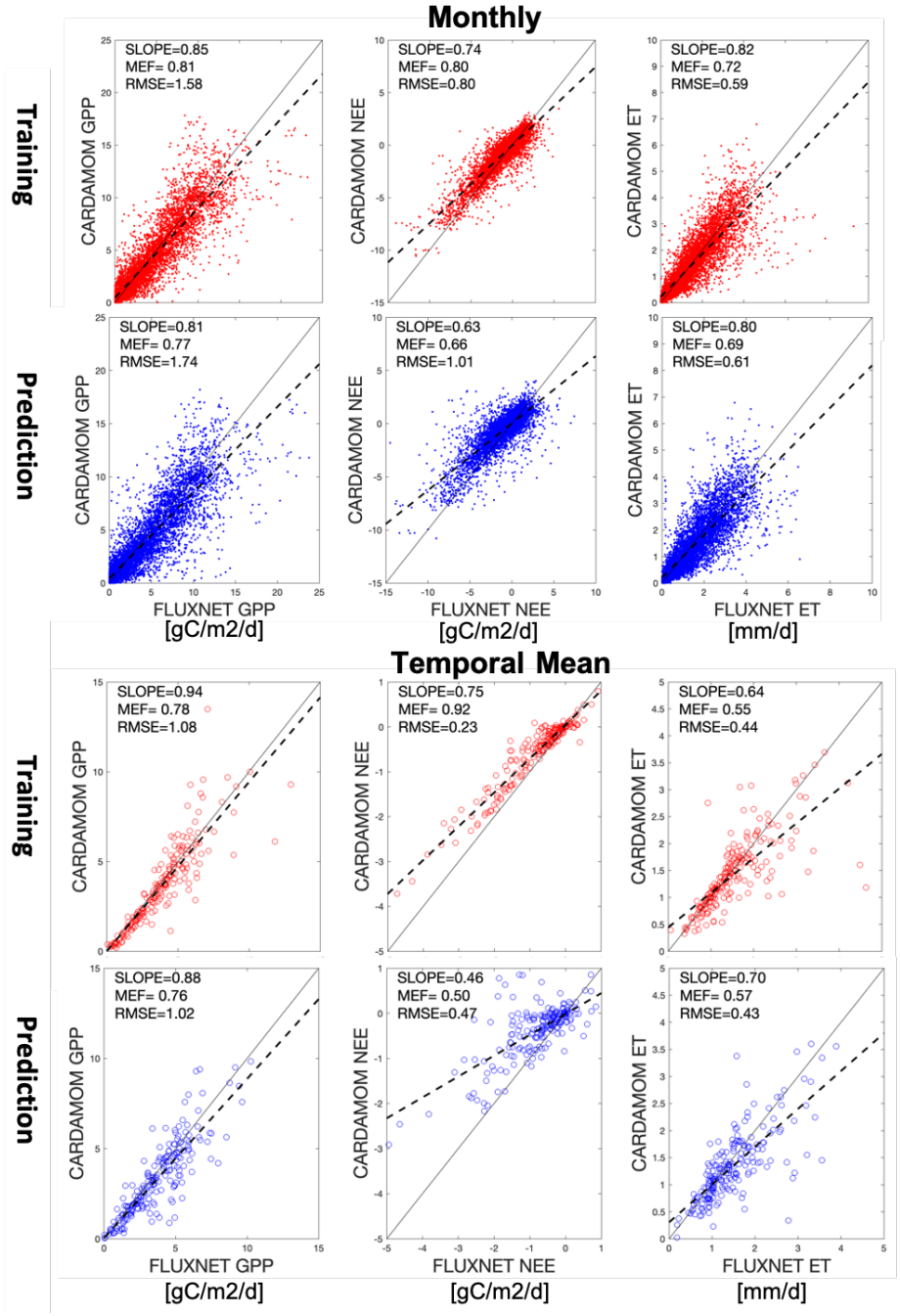


Figure 2. Scatter plots of CARDAMOM outputs (GPP, NEE and ET) versus observations from FLUXNET data (A1 scenario). Scatter plots in red are results from the assimilation window, and the scatter plots in blue are for the prediction window. We plotted both the data from monthly basis (top 2 panels) and at site level using the long-term averages (bottom 2 panels) for comparison.

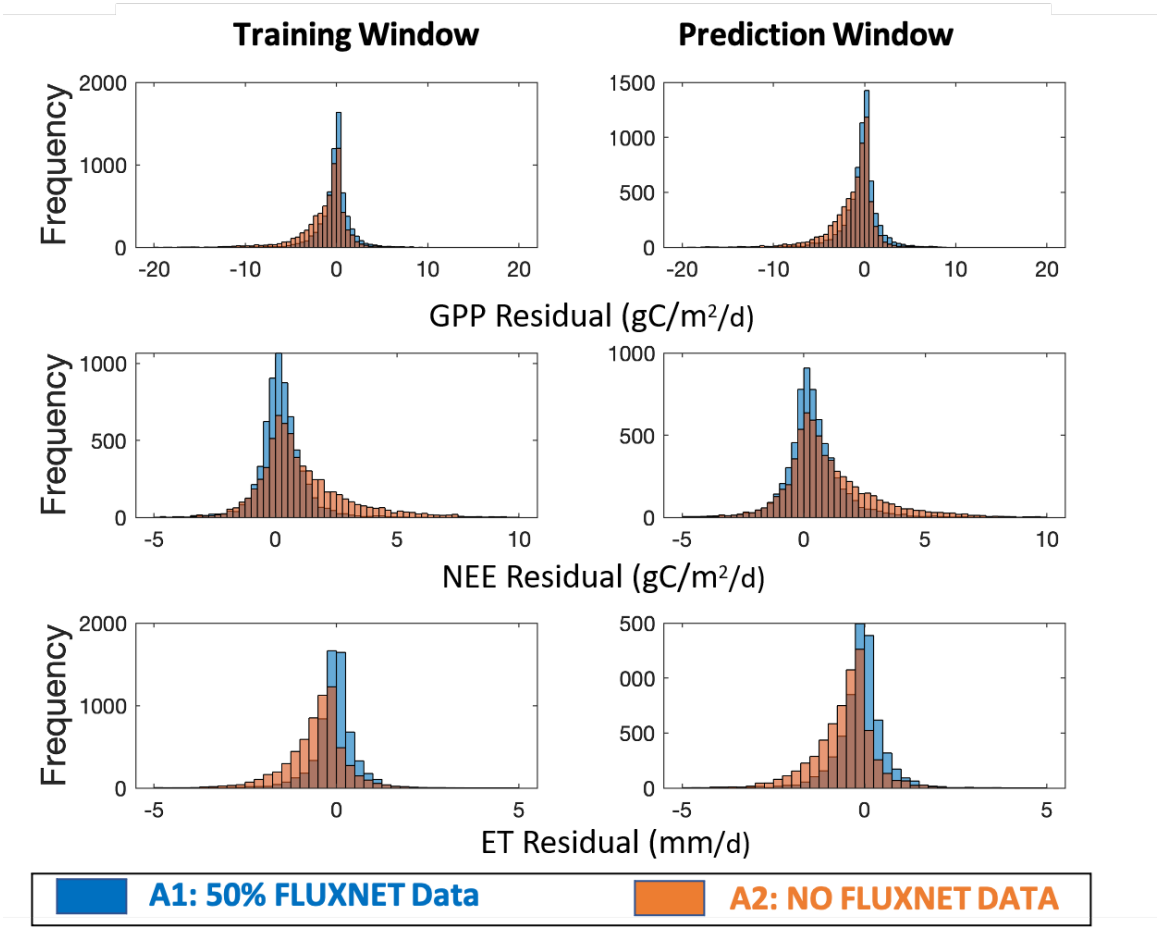
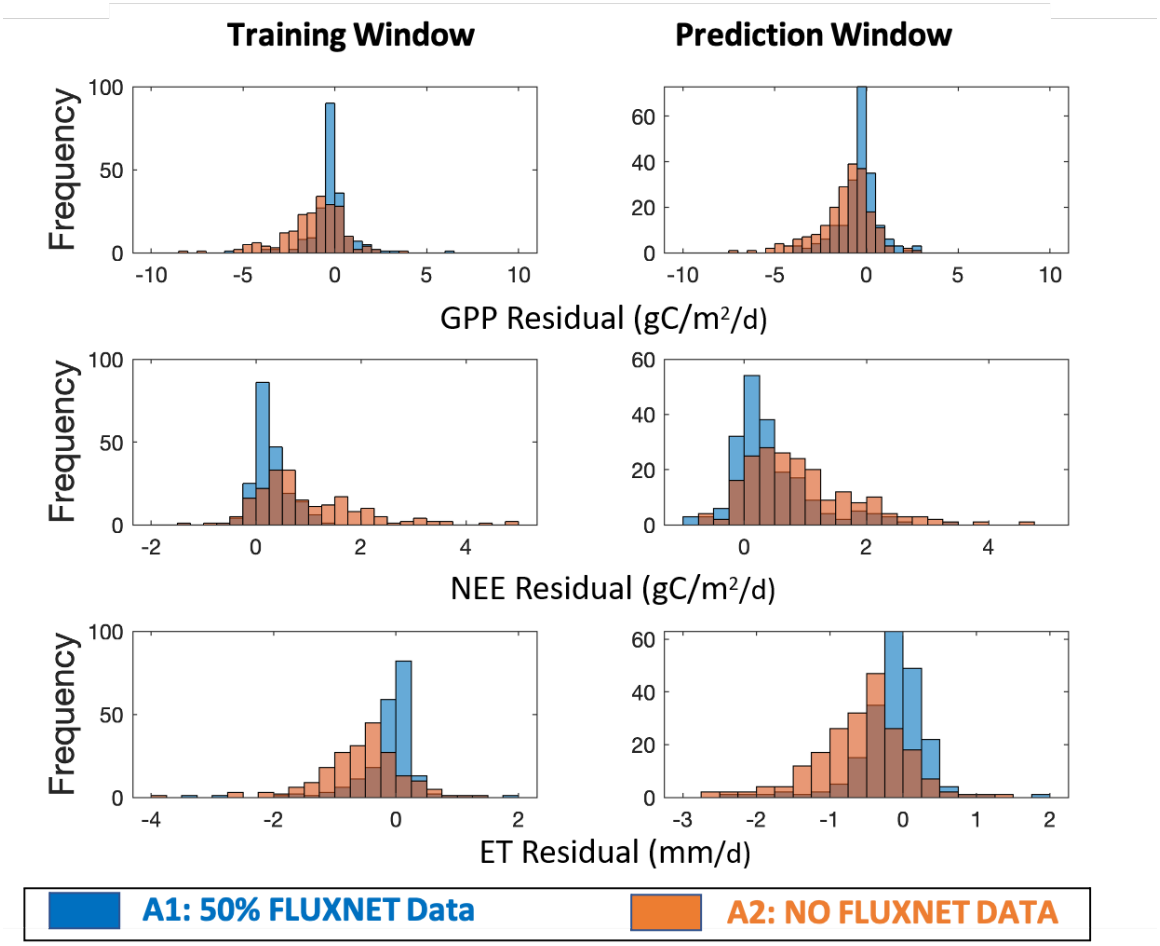


Figure 3. Histogram of monthly-based residuals over all sites for the assimilation window (left panels) and prediction window (right panels). Residuals are the differences between model outputs (GPP, NEE and ET) and observations (GPP, NEE and ET measured at tower sites). Two different CARDAMOM runs are shown as "A1" and "A2" ("A1" means model simulations using 50% FLUXNET data as constraints; "A2" means baseline model simulations with no FLUXNET data).



535 **Figure 4. Histogram of site-level residuals over all sites for the assimilation window (left panels) and prediction window (right panels). Residuals are the differences between model outputs (GPP, NEE and ET) and observations (GPP, NEE and ET measured at tower sites). Two different CARDAMOM runs are shown as "A1" and "A2" ("A!" means model simulations using 50% FLUXNET data as constraints; "A2" means baseline model simulations with no FLUXNET constraints).**

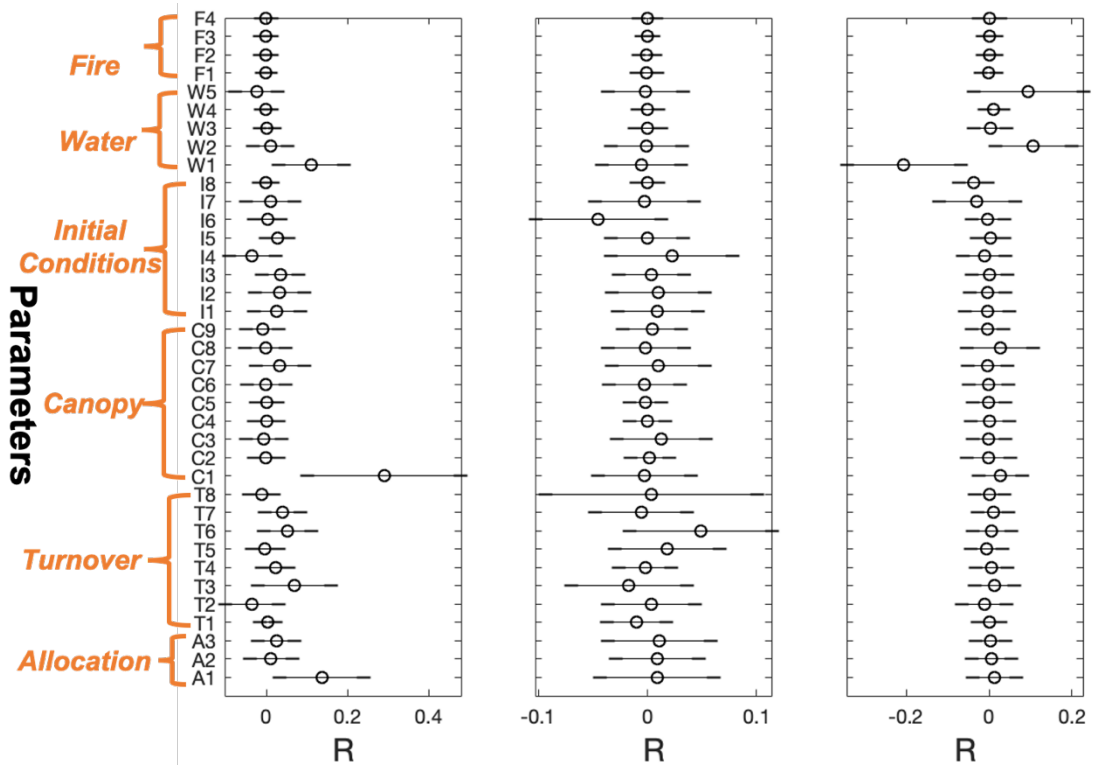


Figure 5. Correlations between the site-level model parameters and residuals of GPP (left column), NEE (central column) and ET (right column) over all sites in the prediction window. Parameters are described in Table S2.

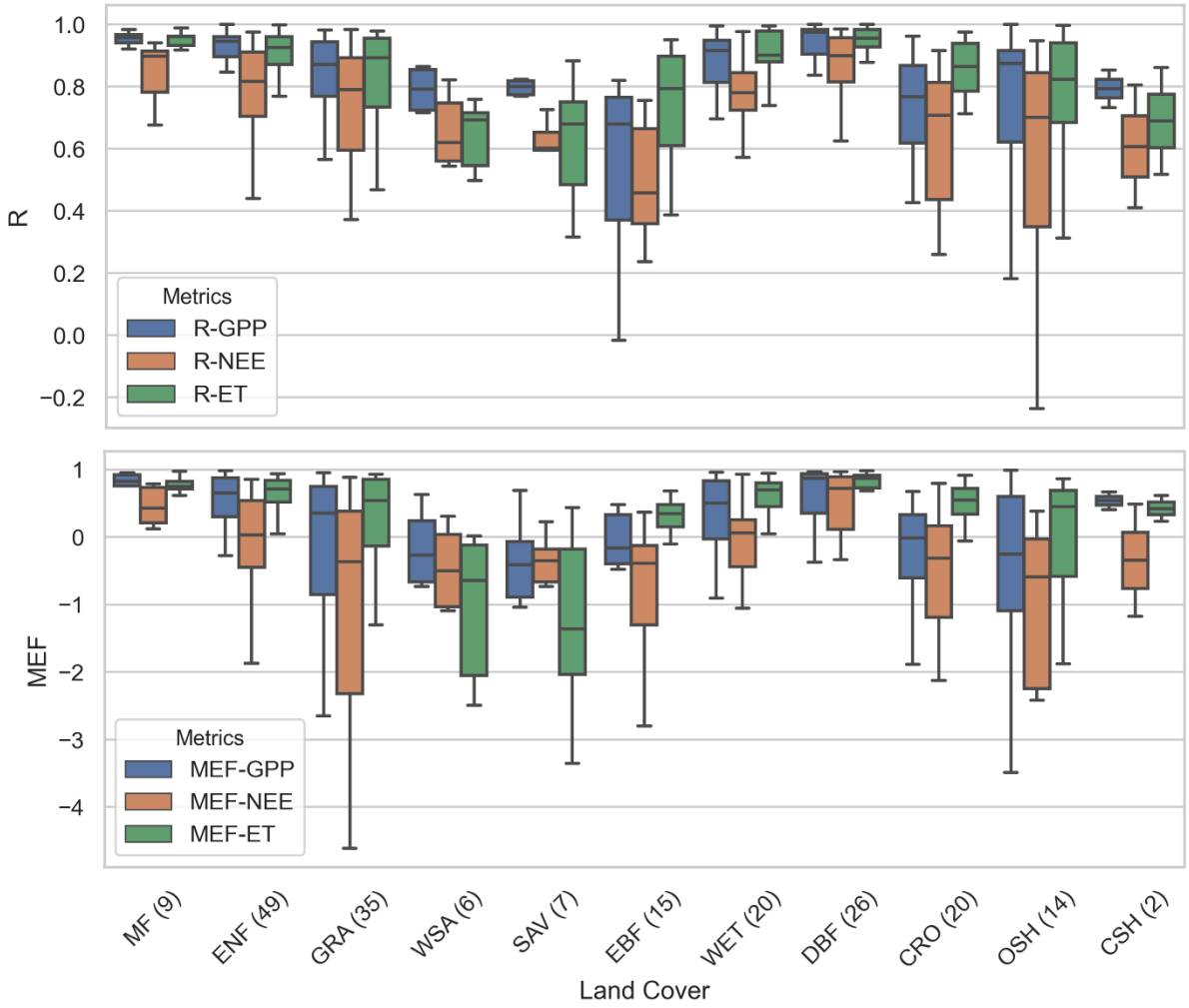


Figure 6. Box plots of correlation metrics (R and MEF) for CARDAMOM outputs (GPP, NEE or ET) versus FLUXNET tower measurements with different landcover types (A1 scenario, prediction window). The full names of land cover types can be found in Table S3. The number in parenthesis (X-axis) indicates the total available tower sites for each land cover type.

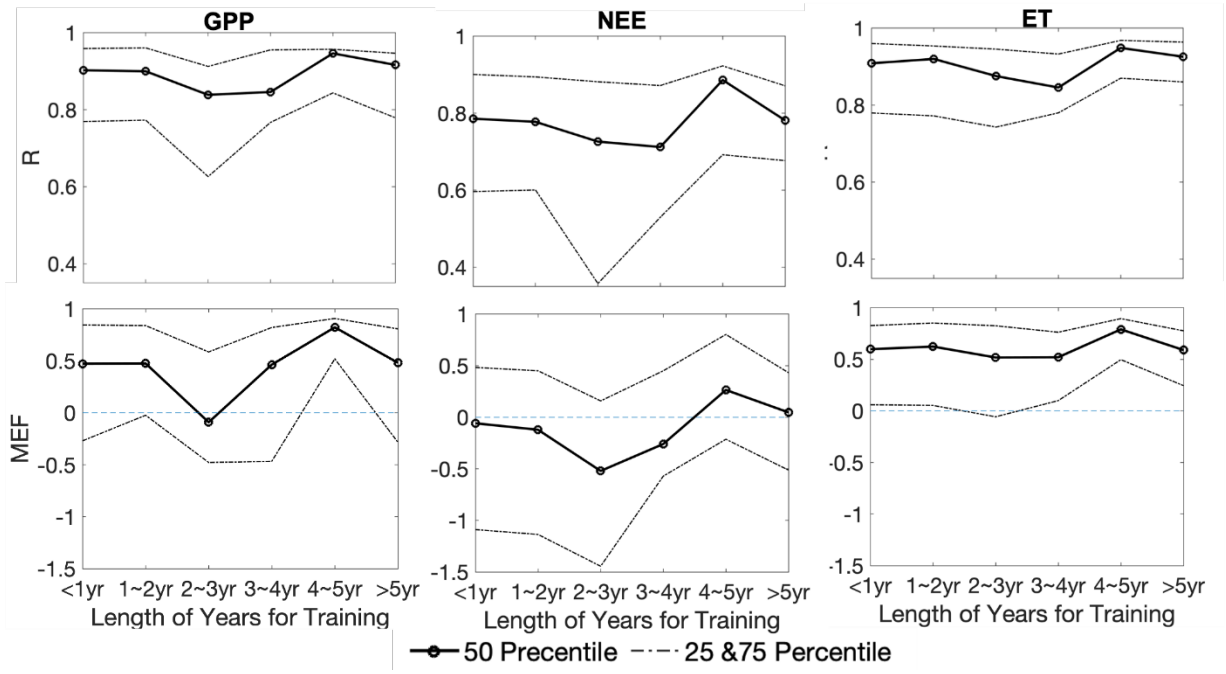


Figure 7. Correlation metrics (R and MEF) changing with different assimilation periods for CARDAMOM outputs (GPP, NEE or ET) versus FLUXNET tower measurements. The solid lines are the 50-percentile of the R and MEF, and the dash lines represent the 25- and 75-percentiles.

Table 1. Monthly-based residuals in assimilation and prediction windows (**Figure 1**)

Residuals		GPP (gC/m2/d)	NEE (gC/m2/d)	ET (mm/d)
A1	Mean (Assimilation)	-0.14	0.22	-0.06
	Std (Assimilation)	1.68	0.97	0.63
	Mean (Prediction)	-0.36	0.36	-0.09
	Std (Prediction)	1.90	1.30	0.65
A2	Mean (Assimilation)	-1.24	1.05	-0.55
	Std (Assimilation)	2.53	1.86	0.87
	Mean (Prediction)	-1.34	1.03	-0.55
	Std (Prediction)	2.49	1.92	0.82

Table 2. Annual-based residuals in assimilation and prediction windows (**Figure 2**)

Residuals		GPP (gC/m2/d)	NEE (gC/m2/d)	ET (mm/d)
A1	Mean (Assimilation)	-0.19	0.22	-0.08
	Std (Assimilation)	1.04	0.50	0.48
	Mean (Prediction)	-0.37	0.34	-0.10
	Std (Prediction)	1.16	0.73	0.45
A2	Mean (Assimilation)	-1.17	0.98	-0.56
	Std (Assimilation)	1.69	1.11	0.69
	Mean (Prediction)	0.94	0.94	-0.54
	Std (Prediction)	1.13	1.13	0.61

Table 3. Site-level residuals in assimilation and prediction windows (**Figure 3**)

Residuals		GPP (gC/m2/d)	NEE (gC/m2/d)	ET (mm/d)
A1	Mean (Assimilation)	-0.21	0.26	-0.11
	Std (Assimilation)	1.09	0.34	0.51
	Mean (Prediction)	-0.39	0.40	-0.14
	Std (Prediction)	1.05	0.71	0.47
A2	Mean (Assimilation)	-1.15	0.96	-0.59
	Std (Assimilation)	1.64	1.01	0.72
	Mean (Prediction)	-1.16	0.91	-0.57
	Std (Prediction)	1.52	1.03	0.63

565

Table 4. The BIAS, MEF, R and RMSE of GPP (unit: gC/m2/d), NEE (unit: gC/m2/d) and ET (unit: mm/d) assimilation versus the flux tower data for different land cover types.

LC	BIAS			MEF			R			RMSE		
	GPP	NEE	ET	GPP	NEE	ET	GPP	NEE	ET	GPP	NEE	ET
CRO	-0.951	0.380	-0.246	0.180	0.046	0.577	0.803	0.702	0.817	2.099	1.792	0.695
CSH	0.274	0.334	0.014	0.495	0.336	0.605	0.789	0.769	0.804	0.879	0.566	0.489
DBF	**-0.085**	**0.510**	**-0.007**	**0.830**	**0.771**	**0.800**	**0.914**	**0.914**	**0.898**	**1.875**	**1.101**	**0.527**
EBF	0.004	0.633	-0.149	0.524	-0.251	0.668	0.806	0.611	0.843	1.656	0.861	0.634
ENF	**-0.400**	**0.454**	**-0.074**	**0.719**	**0.405**	**0.647**	**0.869**	**0.862**	**0.826**	**1.636**	**0.707**	**0.544**
GRA	-0.322	0.050	0.005	0.802	0.063	0.642	0.917	0.719	0.809	1.378	0.750	0.675
MF	**-0.212**	**0.303**	**0.019**	**0.788**	**0.533**	**0.772**	**0.898**	**0.879**	**0.880**	**1.567**	**0.750**	**0.555**
OSH	-0.203	0.195	-0.101	0.525	-0.911	0.051	0.814	0.587	0.692	0.634	0.313	0.385
SAV	-0.505	0.500	-0.241	-0.181	-0.872	-0.193	0.807	0.536	0.704	0.814	0.381	0.518
WET	-0.010	0.026	-0.161	0.635	-3.034	-0.092	0.920	-0.819	0.985	0.026	0.005	0.008
WSA	-0.537	0.246	-0.205	0.649	0.322	0.373	0.872	0.753	0.777	1.326	0.841	0.702