

SCIENTIFIC DATA



OPEN

DATA DESCRIPTOR

Reference carbon cycle dataset for typical Chinese forests via colocated observations and data assimilation

Honglin He^{1,2,3,11}, Rong Ge^{1,2,4,11}, Xiaoli Ren^{1,2}, Li Zhang^{1,2,3}, Qingqing Chang^{1,2,4}, Qian Xu^{1,2,4}, Guoyi Zhou⁵, Zongqiang Xie⁶, Silong Wang⁷, Huimin Wang¹, Qibin Zhang⁶, Anzhi Wang⁷, Zexin Fan⁸, Yiping Zhang⁸, Weijun Shen⁵, Huajun Yin⁹, Luxiang Lin⁸, Mathew Williams¹⁰ & Guirui Yu^{1,2,3} ✉

Chinese forests cover most of the representative forest types in the Northern Hemisphere and function as a large carbon (C) sink in the global C cycle. The availability of long-term C dynamics observations is key to evaluating and understanding C sequestration of these forests. The Chinese Ecosystem Research Network has conducted normalized and systematic monitoring of the soil-biology-atmosphere-water cycle in Chinese forests since 2000. For the first time, a reference dataset of the decadal C cycle dynamics was produced for 10 typical Chinese forests after strict quality control, including biomass, leaf area index, litterfall, soil organic C, and the corresponding meteorological data. Based on these basic but time-discrete C-cycle elements, an assimilated dataset of key C cycle parameters and time-continuous C sequestration functions was generated via model-data fusion, including C allocation, turnover, and soil, vegetation, and ecosystem C storage. These reference data could be used as a benchmark for model development, evaluation and C cycle research under global climate change for typical forests in the Northern Hemisphere.

Background & Summary

Forests contain up to 80% of the terrestrial aboveground carbon (C) and 40% of the below-ground C and thus play a critical role in the terrestrial C cycle¹. A recent study reveals that forests now serve as a net C sink for atmospheric CO₂². However, whether the forest C sink will persist under climate change remains largely uncertain^{3,4}. Therefore, the availability of long-term and systematic observations of forest C dynamics is critical for improving the fundamental knowledge and understanding of forest C cycle processes and the robustness of forest C sink quantification and predictions.

The colocated network monitoring has developed over decades and provides a promising tool for obtaining long-term, intersite, multiple C cycle data^{5,6}. Examples of such networks include the Long-Term Ecological Research Network (LTER), UK Environmental Research Network (ECN), and the Chinese Ecosystem Research

¹Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China. ²National Ecosystem Science Data Center, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China. ³College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, 100049, China. ⁴University of Chinese Academy of Sciences, Beijing, 100049, China. ⁵South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. ⁶State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing, 100093, China. ⁷Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, 110016, China. ⁸Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, 666303, China. ⁹Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, 610041, China. ¹⁰School of GeoSciences and National Centre for Earth Observation, University of Edinburgh, Edinburgh, EH9 3FF, UK. ¹¹These authors contributed equally: Honglin He, Rong Ge. ✉e-mail: yugr@igsnr.ac.cn

Network (CERN). Among them, CERN has conducted systematic observations on the soil, atmosphere, biology, and water in accordance with unified monitoring standards since 2000^{7,8}, and it has accumulated large amounts of long-term data on Chinese forests. The eastern China monsoon region, in particular, has been revealed to be one of the most significant C sink regions worldwide due to its special monsoonal climate, high nitrogen deposition, and relatively young age structure⁹. For this important forest, CERN can provide the only comprehensive dataset covering the typical forest types in this region in the Northern Hemisphere with few human activities, such as land use and cover change, destructive logging or sampling disturbances. This dataset can serve as an important benchmark for the analysis and assessment of regional and global C dynamics under global environmental change, such as climate change, increasing CO₂ concentration and nitrogen deposition. In contrast to the numerous studies assessing aspects of the forest C cycle based on long-term and open-access data collected by ECN and LTER^{10,11}, the integration and reanalysis of CERN data are still at the early stages. Recently, some researchers began to collect CERN data to investigate C cycle states and processes in forests, e.g., biomass and soil C density^{12,13}, biodiversity¹⁴, tree mortality¹⁵, C allocation^{16,17}, and ecosystem C turnover time^{18,19}. These C cycle processes, e.g., C allocation and turnover, with various climate sensitivities collectively regulate the representation of how the forest C cycle responds to the climate^{4,20,21}. At present, due to a lack of observations, even most state-of-the-art earth system models fail to accurately represent C allocation and turnover times^{22,23}, which is largely responsible for the high uncertainty in the predictions of the forest C sink and its response to future climate change^{24–26}. Therefore, it is critical and timely to 1) integrate and produce long-term, across-site, and systematic basic C cycle datasets based on CERN observations and 2) retrieve robust key C-cycle process parameters and time-continuous ecological function dataset (i.e., C sequestration) based on these basic reference data to better evaluate the spatiotemporal C dynamics of these important forests.

Here, we generated a time series and comprehensive dataset of the atmosphere, water, biological and soil C cycle based on CERN raw observations and statistical processing with strict collection criteria and quality control. On this basis, we also conducted a model-data fusion (MDF) framework to generate another assimilated dataset, including C cycle process parameters and C sequestration function products, neither of which can be obtained solely from observations. Moreover, the MDF framework achieved temporal interpolation from the basic time-discrete C cycle data to the time-continuous C sequestration function product. Recently, similar MDF-based time-continuous product was developed as novel benchmark in the International Land Model Benchmarking (ILAMB) project on C cycle^{27,28}.

In this paper, we systematically described the estimation of basic C cycle elements and the MDF method to assimilate the C cycle parameters and sequestration functions of the Chinese Forest Carbon Cycle Dynamics (CFCCD) database (Fig. 1). This paper presents an observation-based basic dataset comprising monthly and annual atmospheric data, water data, and biological and soil C data from ten typical Chinese forests, an assimilated dataset of the C cycle parameters and time-continuous C sequestration (annual) from 2005 to 2015, and some flux data from the Chinese Terrestrial Ecosystem Flux Research Network (ChinaFLUX), which are used as auxiliary validation data in the MDF. This CFCCD database can provide as reference/benchmark for ecological modeling and C dynamics research under climate change for such typical forests in the Northern Hemisphere.

Methods

We selected 10 permanent plots with long-term observations from CERN to include typical forest types of various ages in the East China monsoon forest region, including tropical rainforests, subtropical evergreen coniferous and broad-leaved mixed forests, warm temperate deciduous broad-leaved forests and temperate coniferous and broad-leaved forests, with evident precipitation and temperature gradients from south to north (Fig. 1). The spatial representativeness of the selected 10 sites across the Chinese forest region was evaluated by calculating the Euclidean distance based on various environmental factors. The 10 sites performed well and represented more than 80% of the Chinese forest region (Fig. S1). Of these forests, the Xishuangbanna tropical seasonal rainforest (BNF), Dinghu Mountain subtropical evergreen coniferous and broad-leaved mixed forest (DHF), Ailao Mountain subtropical evergreen broad-leaved forest (ALF), and Changbai Mountain temperate deciduous coniferous and broad-leaved mixed forest (CBF) are mature natural forests; the Shennongjia subtropical evergreen deciduous broad-leaved mixed forest (SNF) and Huitong subtropical evergreen broad-leaved forest (HTF) are natural secondary forests; and the other sites, i.e., the Beijing warm temperate deciduous broad-leaved mixed forest (BJF), Maoxian warm temperate deciduous coniferous mixed forest (MXF), Qianyanzhou subtropical evergreen artificial coniferous mixed forest (QYF), and Heshan subtropical evergreen broad-leaved forest (HSF), are plantations or middle- and young-age forests. All 10 sites are well protected and subject to minimal human activities, thus reflecting the C cycle dynamics under global environmental change, e.g., climate change, increasing CO₂ and nitrogen deposition. The detailed characteristics of each plot can be found in their profiles in the CFCCD database.

There are three main steps to create the observation-based basic dataset and assimilated dataset of typical Chinese forests C cycle dynamics:

1. **Observation-based basic data acquisition.** An ensemble of daily atmospheric and water data at ten CERN sites were used as forcing datasets for MDF and future scientific analysis; biological and soil data were also collected from CERN and processed by quality control and statistical calculation as benchmark to constrain the model.
2. **Implementation of a multiple data-model fusion framework.** The Markov Chain Monte Carlo (MCMC) that integrated the Data Assimilation Linked Ecosystem Carbon (DALEC) model with multiple and dynamic observational data was used to retrieve C-cycle process parameters in a realistic disequilibrium state.

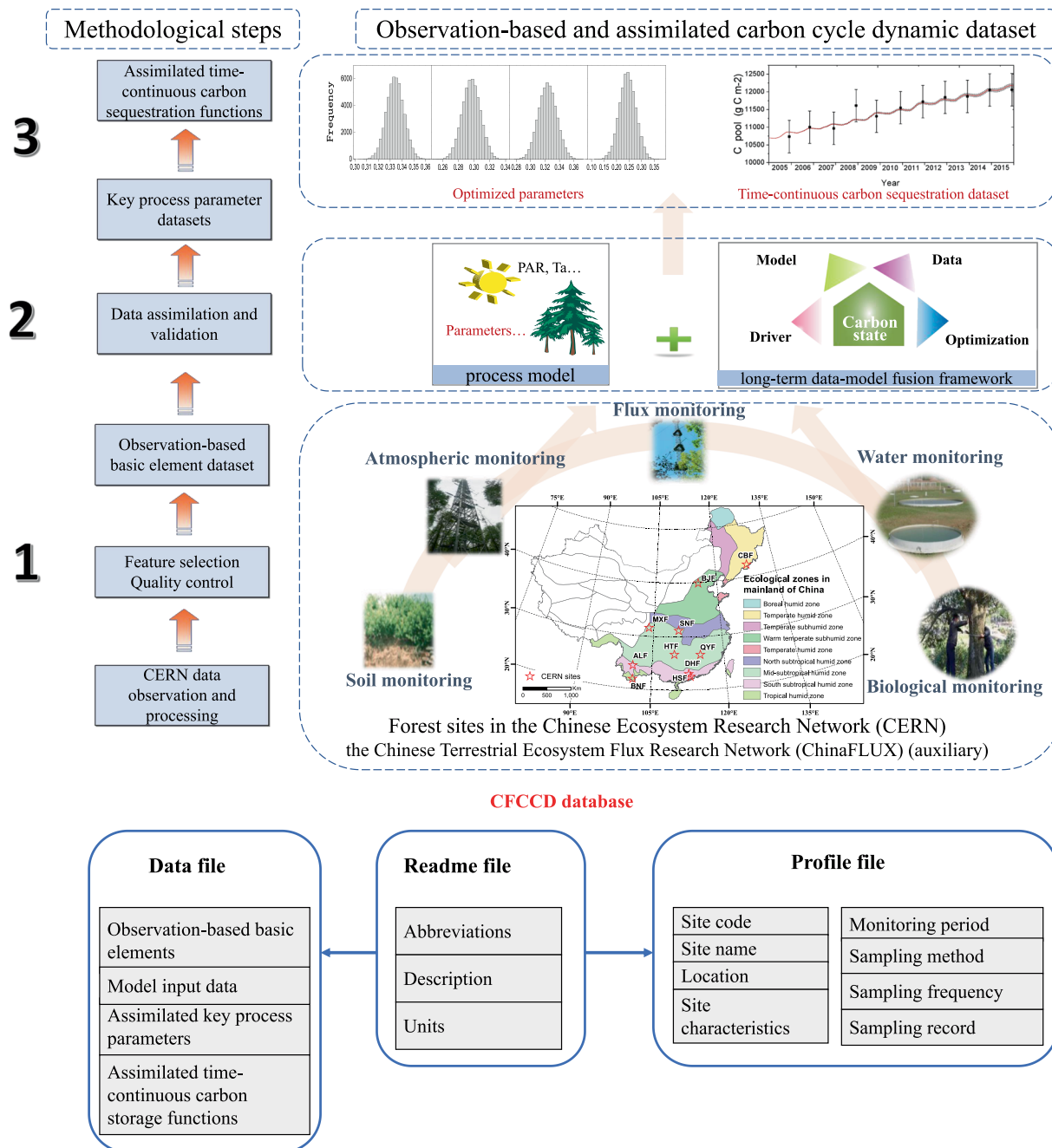


Fig. 1 Schematic overview of the methodology and data products via the colocated CERN observations and long-term data assimilation. The flow diagram shows the methodological steps (left) and the data product systems (right) for the observation-based and assimilated databases. The bottom panel shows the general view of the data structure of the Chinese Forest Carbon Cycle Dynamics (CFCCD) database.

3. Key process parameters and C function data assimilation. The key parameters of the process-based C cycle model (DALEC) were determined via the model-data fusion method; then the ecosystem C sequestration datasets were simulated by forward running the DALEC model with optimized parameters and then validated based on observational data and other previous studies.

Each step is explained in more detail below.

Observation-based basic data acquisition. *Atmospheric and water data.* In situ observations of daily air temperature (Ta), photosynthetically active radiation (PAR), relative humidity (RH), precipitation (Precip), and soil moisture (Sw) at the 10 sites from 2005 to 2015 were obtained from the CERN scientific and technological resources service system (<http://www.cern.org.cn/>). These atmospheric and water data were mostly observed by an automatic meteorological station at each site. Among them, the PAR was estimated by a LI-COR LI-190SZ

Quantum Sensor; Ta and RH were measured by a QMT110 sensor; Sw was estimated by a soil moisture neutron probe or the Time-Domain Reflectometry (TDR) soil moisture probe; the associated saturated soil water capacity (Sc) was measured by the cutting ring method to sample soil in each field campaign and the oven-drying method to measure saturated moisture content after the soil was soaked in water for 48 h; and Precip was artificially observed by CERN staff using an SM1-1 rain gauge. These monitoring data were collected in keeping with CERN's protocols of observation and quality control^{29,30}.

There were occasional missing data in time-continuous meteorological observations; therefore, the data were processed by standardized gap filling³¹. Specifically, for Ta, PAR, and RH, which were applied as model driver, we used a linear interpolation method to interpolate continuous missing data with less than three observations; otherwise, we established a regression model using the CERN observations and other observations from adjacent stations of the China Meteorology Administration (756 meteorological stations; <http://data.cma.cn/en>) to interpolate continuous missing data with more than three observations.

Biological data. Biomass. At each site, the diameters at breast height (DBHs) and tree heights were measured for each tree in a regular inventory performed at least once every five years. The allometric equations of the DBH and/or tree heights with the biomasses of different plant tissues (i.e., leaves, branches, stems and roots) were developed at each site for various species based on the felled standard trees in the destructive plot. Then, we calculated the biomasses for the ten ecosystems using these allometric equations (FA02 table downloaded from <http://www.cnern.org.cn/>), which all passed the significance test (0.01 level) and have the R² most above 0.9 when its estimation compare to observations from standard trees. For some unfelled species under protection, the allometric equations were obtained from Luo *et al.*³², which were developed based on national inventories and meta analyses from the published literature.

Litterfall. The aboveground litterfall biomass was measured monthly by ten replicates with 1 m × 1 m baskets during the growing season or once during the nongrowing season. All collected litter was dried at 70 °C for 24 h in the laboratory and then weighed. To avoid the effects of wind on the measurement of litterfall biomass within an individual month, annual litterfall biomass data were finally adopted for each site.

LAI. The leaf area index (LAI) at each site was measured optically with an LAI-2000 plant canopy analyzer (LI-COR, Lincoln, NE, USA) at least quarterly every year.

Soil data. Soil organic matter (SOM) was measured by the potassium dichromate oxidation titrimetric method. Soil bulk density (SBD) was measured by the cutting ring method in each field campaign at 10 forest sites. Soil particle size (i.e., soil mechanical composition) was measured by the laser particle analyzer. At least three samples were collected from each of the five soil layers (0–10, 10–20, 20–40, 40–60, and 60–100 cm) once every five years. SOC. The soil organic C (SOC) content was calculated from SOM, SBD, and volume percentage of gravel with particle size >2 mm at 10 forest sites as follows³³:

$$SOC = \sum_{i=1}^n 0.58 \times H_i \times B_i \times O_i \times (1 - \theta) \times 100 \quad (1)$$

where SOC is the soil organic C density (g C/m²) of all *n* layers, *H_i* is the soil thickness (cm), *B_i* is the soil bulk density (g/cm³), *O_i* is the SOM content of the *i_{th}* layer (%), and *θ* is the volume percentage (%) of gravel with particle size >2 mm. In the absence of soil bulk density or soil organic matter content measurements in some layers, the missing soil measurements corresponding to specific soil depths of these forest ecosystems were supplemented according to the empirical formulas of the relationships between SOM/soil bulk density and soil depth in different layers, which were developed based on the long-term and across-site CERN soil observations³⁴.

All these raw atmospheric, biological, and soil data mentioned above can be directly download from CERN scientific and technological resources service system (<http://www.cnern.org.cn/data/initDRsearch>) or obtained after online application via protocol sharing.

Auxiliary flux data. Net ecosystem exchange (NEE). These data were obtained from ChinaFLUX (<http://www.chinaflux.org/>), covering CBF, QYF, and BNF. The data were aggregated to the daily time step from half-hourly CO₂ flux data measured by the eddy covariance technique and processed with quality control and gap filling procedures³⁵.

Implementation of MDF method. The assimilated data were retrieved from a multiple data-model fusion method (Fig. 2). Specifically, the long-term and dynamic observations of biomass, litterfall, LAI and SOC were used as the model constraint data; Ta, PAR, and RH were used as the meteorological driving data; and the metropolis simulated annealing algorithm, a variation in the MCMC technique^{36,37}, was applied to retrieve the C cycle parameters (e.g., C allocation and C turnover times) against the observations and prior knowledge. Then, we forward-simulated the model to produce the dynamic and time-continuous changes in ecosystem C sequestration function.

Since the dynamic C cycle observations provided an effective solution to constrain the C cycle states without the steady state assumption (SSA), the novelty of our MDF framework involves estimating these C cycle dynamics in better agreement with the actual dynamic disequilibrium state³⁸. Therefore, the uncertainty in allocation and turnover parameters and in C pool states have largely been reduced based on the time-series observations under the non-SSA (NSSA)^{21,39,40}, thereby significantly enhancing the model's ability to predict the C sequestration function^{19,41,42}.

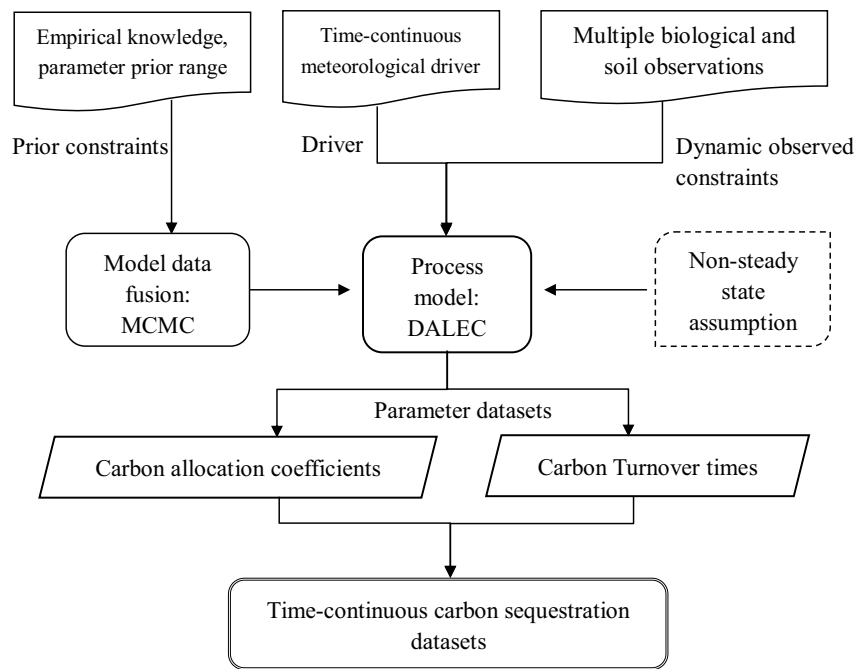


Fig. 2 Flowchart of the generation of assimilated datasets in a multiple- and long-term data assimilation framework.

Carbon cycle process model description. DALEC is a box model of C pools connected via fluxes running at a daily time step and has been applied extensively to the MDF research^{21,43}. Its main structure (i.e., C cycle, C allocation, and turnover process) is generally consistent with state-of-the-art process-based models (Fig. S2; Table S1), with five pools (i.e., foliage (Cf), fine root (Cr), woody (Cw, including branches, stems, and coarse roots), litter (Clit) and SOM (Csom)) for evergreen forests and an additional labile pool (Clab) of stored C that supports leaf flushing for deciduous forests. The C cycle was initiated with the canopy C influx: gross primary productivity (GPP), which was predicted by the Aggregated Canopy Model (ACM)⁴⁴ (Appendix S1). After GPP is consumed by autotrophic respiration (Ra), the remaining photosynthate (NPP) is allocated to plant tissue pools (Cf, Cr, or Cw). The C exiting from all C reservoirs was based on a first order differential equation with various turnover rates, with temperature and moisture dependency on the turnover from the litter and soil pools. In contrast to the original DALEC model only with temperature scalar f_{T_w} , here we added a new function f_{S_w} to express soil moisture pressure on litter and soil decomposition processes (Appendix S1). In general, the C pools and fluxes in DALEC were iteratively calculated at a daily time step and determined as a function of the key turnover and allocation parameters. A detailed model description can be found in Williams *et al.*⁴⁵ and Fox *et al.*⁴⁶.

Multiple data-model fusion at the nonsteady state. In a realistic disequilibrium state, C pools are time-variant, i.e., the C efflux is not equal to the C influx ($\frac{dC}{dt} \neq 0$); thus, the MDF was run via the dynamic and long-term CERN observations to constrain the DALEC model at the non-steady state (Eq. 2). Here, to avoid the uncertainty arising from the spin-up process under SSA, we determined the initial state of the C pools by the initial observations of C stocks or by optimization (i.e., Clab, which cannot be directly observed). Then, the turnover and allocation parameters were retrieved under the disequilibrium state with dynamic environmental forcing. This method avoids the considerable uncertainties when invoking the SSA to estimate the initial state of C pools and the C cycle parameters (e.g., allocation coefficients and turnover rates)^{39,40,47}, which could lead to obvious biases in C sequestration¹⁹.

$$\begin{cases} \Delta C_i \neq 0 \\ C_i(t+1) = C_i(t) + I_i(t) - k_i C_i(t), i = 1, 2, \dots, n \\ C_i(t=0) = C_{i0} \end{cases} \quad (2)$$

where C_i , I_i , and k_i represent the size, input and turnover rate of the i_{th} C reservoir, respectively; C_{i0} represents the initial state of the i_{th} C reservoir; t represents the specific model-running time step (daily step); and ΔC_i represents the i_{th} C pool change between t day and $t+1$ day when applicable into actual calculation. According to the Bayesian theory, the posterior distributions of the parameters are calculated by maximizing the likelihood function (Eq. 3).

$$L = \prod_{j=1}^m \prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-(x_{j,i} - \mu_{j,i}(P))^2 / 2\sigma_j^2} \quad (3)$$

where L is the integrated likelihood function; m is the number of multiple data types; n is the number of data points categorized by the j_{th} data type; $x_{j,i}$ is the measured value composed of dynamic C cycle observations; $\mu_{j,i}(P)$ represents the modeled fluxes and stocks based on parameters under the NSSA (P); and σ_j is the standard deviation of each data point classified by the j_{th} data type. Moreover, we imposed a sequence of ecological and dynamic constraints on the model parameter inter-relationships and pool dynamics (Appendix S2), which can significantly reduce uncertainty in model parameters and simulations⁴⁸. The more detailed disequilibrium method can be found in our latest study¹⁹.

Key C-cycle process parameters and C sequestration data assimilation. *Key process parameter estimation.* Here, we mainly focus on how the C input (i.e., the net primary productivity) partitioned into various plant pools (i.e., foliar, wood, and fine roots), i.e., allocation coefficients, which could be directly determined from the optimized parameters (Fig. S3) of the DALEC model after the step 2: MDF method. Another key process parameter, C turnover time, needs further simple statistical calculation based on the model simulations with optimized parameters. Turnover time is commonly estimated by the equation “ $\tau = \text{stock}/\text{flux}$ ”^{20,49}. Since the C inflow is not equal to the C efflux in the realistic dynamic disequilibrium state, the turnover time should be defined as the ratio between the mass of a C pool and its outgoing flux⁵⁰. Note that with few natural and anthropogenic disturbances in these well-protected CERN sites^{12,18}, the C efflux is approximately equivalent to the R_h from soil and litterfall (mortality) and R_a (growth) from vegetation. Hence, the turnover time for vegetation, soil, and whole ecosystem can be derived as follows:

$$\tau_{veg} = \frac{C_{live}}{I_{live} - \Delta C_{live}} = \frac{C_{live}}{\text{litterfall} + R_a} \quad (4)$$

$$\tau_{soil} = \frac{C_{dead}}{I_{dead} - \Delta C_{dead}} = \frac{C_{dead}}{R_h} \quad (5)$$

$$\tau_{eco} = \frac{C_{eco}}{I_{eco} - \Delta C_{eco}} = \frac{C_{live} + C_{dead}}{R_a + R_h} \quad (6)$$

where τ_{veg} , τ_{soil} and τ_{eco} refer to the biomass, soil and whole-ecosystem turnover times, respectively; C_{live} , C_{dead} and C_{eco} refer to the live biomass C pool size (C_f , C_r , and C_w), dead organic C pool size (C_{soil} and C_{litter}), and the whole-ecosystem C pool size, respectively; I_{live} , I_{dead} and I_{eco} refer to the C input into the live biomass C pool, dead organic C pool, and whole ecosystem C pool, respectively; ΔC_{live} , ΔC_{dead} and ΔC_{eco} refer to the changes in the live biomass C pool, dead organic C pool size, and whole-ecosystem C pool size, respectively; and R_a , R_h and litterfall refer to the autotrophic and heterotrophic respiration, and turnover from all live C pools (i.e., foliage, fine root and woody pools), respectively, as calculated from the DALEC output driven with the estimated parameters during 2005–2015. Since the C reservoirs, fluxes, and turnover times are instantaneous values, we used the averages of the fluxes and reservoirs over multiple years to reflect the average turnover time during a specific period (i.e., 2005–2015).

Time-continuous C sequestration estimation. The optimized parameter values under the NSSA along with the initial observations of the corresponding C pool sizes were used in forward modeling driven by dynamic environmental variables from 2005 to 2015 to obtain the time-continuous soil and vegetation C storage⁵¹. The difference between the ecosystem C inflow (GPP) and ecosystem respiration ($R_a + R_h$) is used to examine the ecosystem C sequestration, i.e., net ecosystem productivity (NEP). Similarly, the difference between the ecosystem C inflow (GPP) and ecosystem autotrophic respiration (R_a) is used to examine the net primary ecosystem productivity (NPP).

Data Records

The CFCCD database consists of three dataset types (Fig. 1) that were recorded in a series of Microsoft Excel files, which can be found on the Figshare repository at (<https://doi.org/10.6084/m9.figshare.12331400.v2>)⁵². Among them, the ‘profile file’ (CFCCD Profile.xlsx) includes site and observation information, such as site code; site name; site plot area; site coordinates (longitude and latitude); site characteristics; site disturbance information; sampling method, frequency and sampling period associated with each atmospheric, water, biological, and soil variable. The ‘readme file’ (CFCCD Readme.xlsx) explains the abbreviations used in the ‘data file’ and ‘profile file’ and provides the units of all variables included. The ‘data file’ provides 3 datasets: (a) the observation-based basic element dataset, i.e., monthly/yearly observation-based basic C cycle elements with quality control and statistical calculation, including six atmospheric and water datasets as meteorological drivers (i.e., Ta, Precip, PAR, RH, Sw, and Sc), three biological C dynamic datasets (i.e., biomass; litterfall; and LAI), four soil C dynamic datasets (i.e., SOC, SOM, SBD, and soil texture), and one C flux dataset at some of the sites (NEE at CBF, QYF, and BNF); (b) the model input dataset, i.e., all the time-continuous meteorological drivers at model-running time step (daily step) used in model simulation and assimilation; (c) the assimilation dataset, including the assimilated parameter dataset based on the MDF method, i.e., allocation coefficients and turnover times retrieved specifically for each site; and the assimilated annual time-continuous ecosystem C sequestration functions consisting of vegetation C stock, soil C stock, and ecosystem productivity as well as respiration (Table 1).

Specifically, the CFCCD ‘data folder’ includes records of 10 forest sites. The data time series at most of the sites cover the period from 2005 to the latest available year (2015), but those of the SNF, which were later incorporated into CERN, are from 2008 to 2015. The average C stock of the ten typical ecosystems is 21.6 kg C m⁻². From north

Observation-based basic elements	Atmospheric element	air temperature (Ta)
		photosynthetically active radiation (PAR)
		relative humidity (RH)
	Water element	soil moisture (Sw)
		soil saturated moisture capacity (Sc)
		precipitation (Precip)
	Biological element	litterfall
		leaf area index (LAI)
		biomass for different plant tissues
	Soil element	soil organic carbon (SOC) density
		soil organic matter content (SOM)
		soil bulk density (SBD)
		soil texture (soil mechanical composition)
Auxiliary carbon flux	NEE	
Model input	Time-continuous meteorological drivers at model-running time step	
Assimilation dataset	Assimilated process parameters	carbon allocation coefficients
		vegetation, soil and ecosystem turnover times
	Assimilated C storage functions	soil, vegetation, and ecosystem carbon stocks
		ecosystem productivity, respiration and carbon sink

Table 1. Element-parameter-function system for the data file in the Chinese Forest Carbon Cycle Dynamics (CFCCD) database.

Site_code	Biomass_C (g C m ⁻²)	SOC (g C m ⁻²)	Total_C (g C m ⁻²)	NEP (g C m ⁻² yr ⁻¹)
CBF	4489 ± 618	8891 ± 765	13380 ± 1371	314 ± 37
BJF	5930 ± 552	4231 ± 329	10160 ± 879	266 ± 62
MXF	3214 ± 345	14795 ± 933	18009 ± 1269	371 ± 114
SNF	10381 ± 852	13377 ± 72	23758 ± 924	351 ± 37
HTF	12917 ± 525	7806 ± 324	20722 ± 848	231 ± 56
QYF	6542 ± 1233	6721 ± 572	13263 ± 1792	467 ± 118
ALF	24155 ± 612	29383 ± 452	53538 ± 1063	302 ± 65
DHF	13501 ± 894	9527 ± 268	23028 ± 631	232 ± 27
HSF	5339 ± 516	9068 ± 625	14407 ± 1138	301 ± 81
BNF	16141 ± 758	9896 ± 179	26037 ± 936	266 ± 72

Table 2. Summary statistics (mean ± standard deviation of mean) for the assimilated dataset of biomass (Biomass_C), soil (SOC), ecosystem carbon stocks (Total_C), and net ecosystem productivity (NEP) during the 2005–2015 period at the ten forest sites.

to south, with the increase in temperature and precipitation, the vegetation and soil stocks show a significant increase, indicating that the C stocks of the forest ecosystems in warmer and humid regions are higher than those in cold and dry regions (Table 2). Among the different C pools in the ecosystem, the soil C stock is the largest, accounting for 53.2% of the total C stock of the ecosystem; as the temperature and precipitation increase, the proportion of C stocks distributed in the soil gradually decreases, while the proportion of C stocks distributed in the vegetation gradually increases. In the past 10 years, all ten forest ecosystems function as C sinks, indicating a large C sequestration capacity in eastern China monsoon forests under climate change. Based on the observation-based basic C cycle dataset, the optimized C cycle parameters, and the C sequestration function product, we can obtain a clear and transparent map showing how the C flows in different forest ecosystems (Fig. 3).

Technical Validation

Observation-based validation. In terms of the observation-based dataset, the data at all CERN sites were obtained via internationally, widely used field samplers (e.g., rain gauge used for meteorological data), quantification methods (e.g., the allometric method used for biomass data), and laboratory analysis methods (e.g., the potassium dichromate oxidation titrimetric method used for SOM data). In addition, all measurements were subject to standard uniformity procedures from sampling to storage methods, which were undertaken by trained personnel at each CERN station. The CERN also presents detailed information on the protocols for standard observation, measurement, and laboratory analysis methods for the users to evaluate for themselves^{29,30,53,54}.

Moreover, CERN has a three-level data quality control and validation system consisting of each station, sub-centers (e.g., atmosphere, biology, soil, and water subcenters) and comprehensive center. To further improve the data quality, we also established a collaborative quality control framework among data users at the co-located

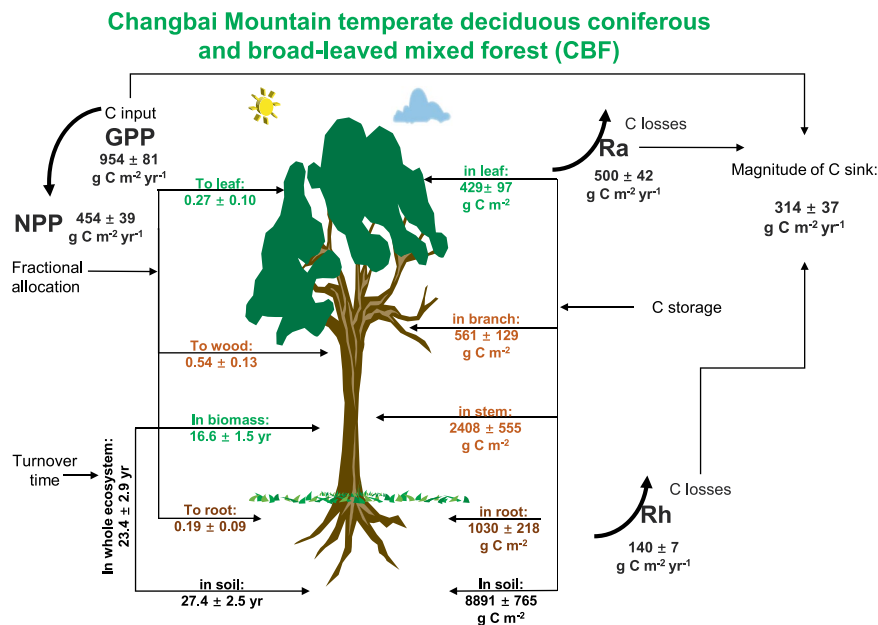


Fig. 3 General map of how carbon flows in an ecosystem based on the Chinese Forest Carbon Cycle Dynamics (CFCCD) database, taking the Changbai Mountain temperate deciduous coniferous and broad-leaved mixed forest (CBF) as an example. The statistic values refer to the mean and the standard deviation of mean for various variables.

network level and producers at the site level, focusing on the integrity, consistency, and reliability of the long-term, multisite and multielement observations during the production of the CFCCD database (Fig. S4). Specifically, we carried out data integrity analysis, consistency checks, and outlier elimination through time-series comparison, multisite comparison, multifactor comparison, and comparison with published literature, and then interpolated the missing data:

(1) Integrity check

Here, we mainly verified whether the observation frequency and sample information are complete, whether the metadata information of the data is missing, and preliminarily confirmed the degree of the missing data.

(2) Consistency check

The time consistency, spatial consistency, terminology consistency and element correlation of the observed data were systematically checked to determine, for example, whether the plant names were consistent in the inter-year community surveys, whether the names of the sample plots and the spatial sample areas were consistent in various sampling years, and whether the trend in the temporal variation in the elements followed relevant prior ecological knowledge.

(3) Detection of outliers

Statistical methods (such as the $3\text{-}\sigma$ criterion) are used to eliminate the abnormal values for the soil, biological, atmospheric and water elements. At the same time, the remaining existing observation values are compared with the results in the literature to validate and eliminate the abnormal values. Communication and confirmation are made with the staff at each station in terms of the input of the raw data, the calibration of the measuring instruments, the consistency of statistical calculation methods and the correction methods for the raw measurement. Finally, after this strict quality control, the missing data were interpolated in accordance with different methods for different types of data, as described in detail in the Methods sections 2.1.1, 2.1.2, and 2.1.3.

Assimilated dataset validation using in-situ measurements. For the assimilated dataset, we validated the performance by comparing the simulated vegetation and the soil C stocks and fluxes with the corresponding observations. The results showed good agreement, with the scatter points following the 1:1 line (Fig. 4). The Taylor diagram showed that five stock-related variables had high correlation and low bias relative to observations. Specifically, the determination coefficients (R^2) for the C stock-related variables varied between 0.91 and 0.95, and the root-mean-square errors (RMSEs) were small relative to their magnitudes. In addition to the biomass data, litterfall and SOC, we also added two datasets of C fluxes to validate the MDF performance, including the net ecosystem exchange (NEE) from ChinaFLUX, including CBF, QYF, and BNF, and soil respiration (R_s) data measured using static chamber-gas chromatography techniques at CBF, QYF, DHE, HSE, and BNF⁵⁵. In contrast, the R^2 values for C fluxes (NEE and R_s) were slightly lower (0.55–0.63), but the RMSEs were only 0.81 and $0.39 \text{ g C m}^{-2} \text{ d}^{-1}$, respectively, which fell well within the range of the C flux validation in MDF studies on eastern China forest ecosystems (e.g., Zhang *et al.*¹⁸). Moreover, we compared the directly observed pool-based

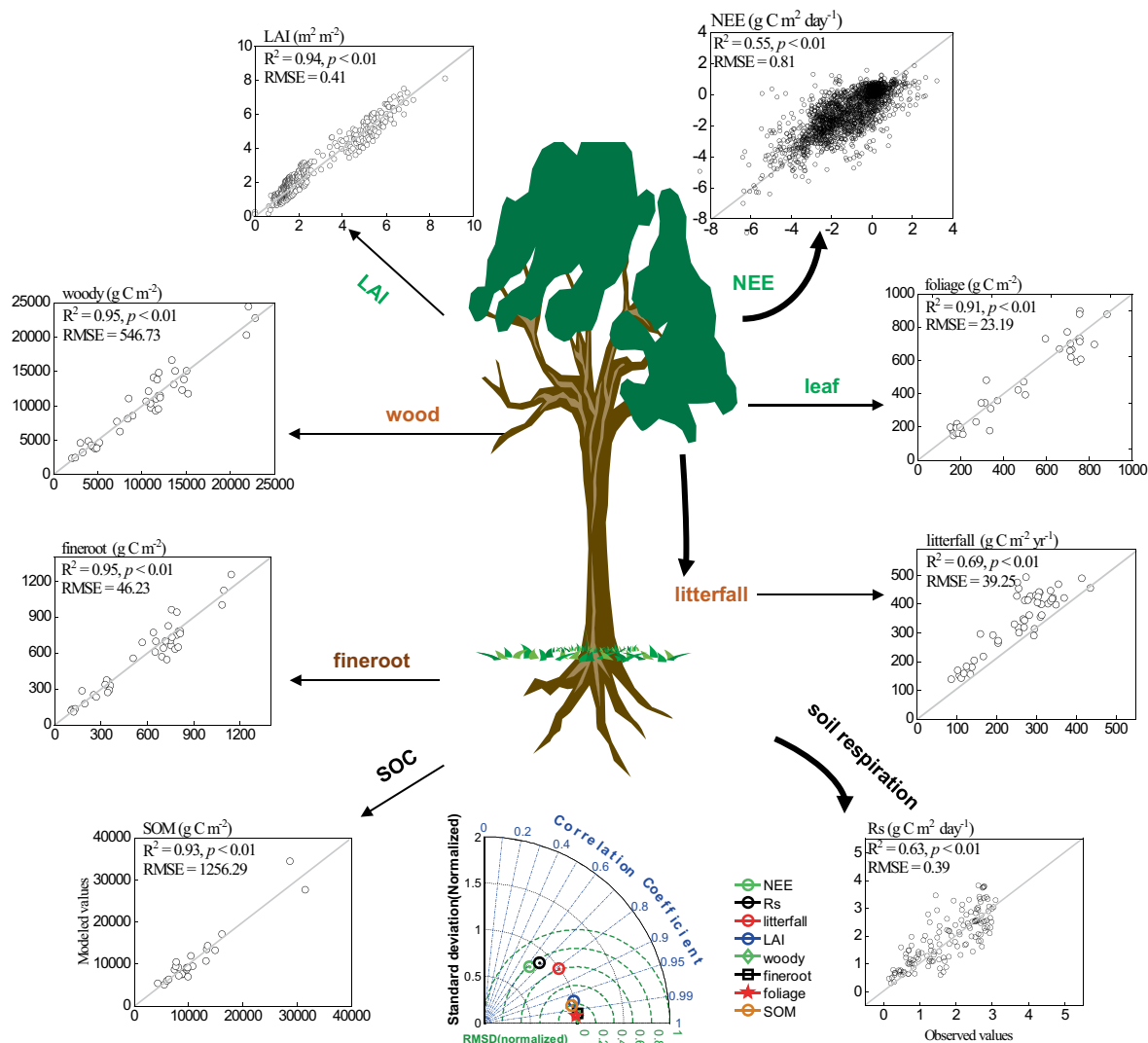


Fig. 4 Comparisons between observed (in the x-axis) and modeled values (in the y-axis) at all sites under the data assimilation framework. The Taylor diagram (bottom panel) presents statistical tests (i.e., correlation coefficient, standard deviation, and root-mean-square deviation (RMSD)) for key fluxes and state variables as a summary of quality of fit.

increment (i.e., $NEP = \Delta\text{Biomass} + \Delta\text{SOC}$) to the modelled NEP via MDF; the result also showed a high consistency between the pool-based observation and model simulations based MDF (Fig. S5; $R^2 = 0.74$, $p < 0.01$).

Assimilated dataset validation using previous studies. Moreover, the observation-based constraint datasets and assimilated datasets were reviewed by international peers; several papers associated with this database have been published/submitted^{13–15,18,19}. The optimized parameters (i.e., plant allocation, and the estimations of τ_{veg} , τ_{soil} , and τ_{eco}) under dynamic disequilibrium all showed high consistency with the existing empirical research based on field observations or experiments^{20,25,56–61} (Fig. 5). This indicated the reliability and robustness of our assimilated parameters under the realistic disequilibrium state.

Uncertainties. Model-data fusion is a powerful method for the generation of improved simulation results via the combination of models with various data streams^{62,63}. Model structure, model parameters, model assumption and observation data were the main sources of the uncertainties⁶³.

The DALEC model is widely applied in various ecosystem across global scales with good performance²¹, and the model formulation bore similarities with the state-of-the-art process-based models^{45,46}. Although we assumed that soils are a single homogeneous pool, which disregards the reality that soils consist of C that turnovers at different rates ranging from fractions of a year to centuries⁶⁴, it has been challenging among earth system models to separate soils into different pools and quantify each pool's turnover time due to lack of corresponded observed data⁶⁵. DALEC assumes a single homogeneous soil pool thus to better assimilate with available observed information from CERN. Besides, our study mainly used the mean turnover time for whole vegetation or soil pool. Therefore, the single-soil-pool structure should not have significant on the estimation of mean soil turnover

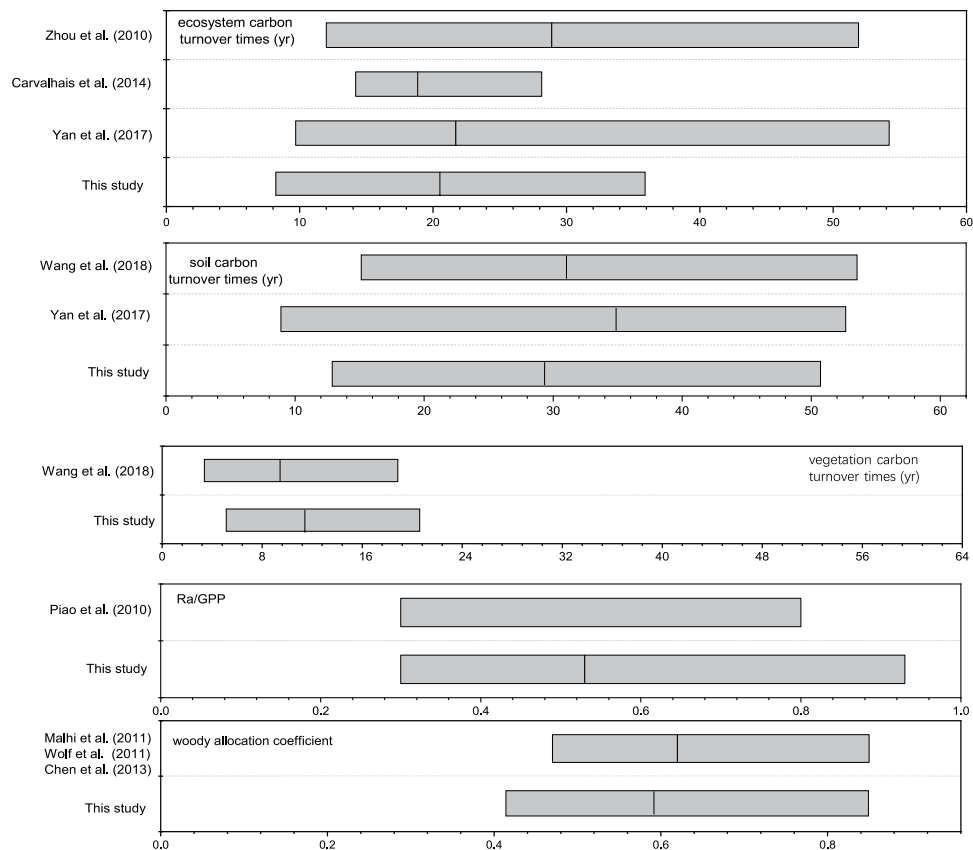


Fig. 5 Comparison of assimilated parameters with those derived from other empirical studies.

time and its further analysis. Besides, some forests are aggrading, and we conducted the MDF under the realistic dynamic disequilibrium assumption. Although the forest age was not directly considered in the model, the non-steady state estimation in this study based on long-term observational data (stock increments in aggrading forests) implicitly incorporated the age-structure-related effect on C cycle dynamics^{19,42}, thus providing a proper estimation on carbon allocation or turnover process, as well as the C sequestration function.

We further conducted a sensitivity analysis to quantify the uncertainty sensitivity to input model parameters. The response variables for the sensitivity analysis are total annual NEE, GPP, Reco, and mean annual C pools. By modifying each parameter $\pm 10\%$, we calculated the percentage change in the response variables (VR) and the sensitivity index (β , ratio of the % change in response variable to % change in a parameter)⁶⁶. Sensitivity analysis indicated that carbon fluxes and pools experienced similar sensitivity patterns (Table S2) considerably affected by parameters related to photosynthesis and C allocation. They were less affected by turnover as well as coefficient of correction. Since we collected field observations of LCMA, this measure was set as constant so that the model uncertainty over key photosynthesis parameters could be decreased, allowing emphasis on analyzing variation in allocation coefficients. Here the model allocation was well constrained by the time-series LAI and biomass of various plant tissues (Fig. S3), and thereby the allocation coefficients (especially allocation to wood and autotrophic respiration) showed high consistency with the empirical studies (Fig. 5).

The challenge of acquiring long-term and multiple observations covering different C cycle process is one of the inherent limitations in process-based model data fusion⁴⁰. Here we mainly collected the biomass and SOC observations. Since CERN sites do not conduct the DOC fluxes observation, which is a small proportion of SOC⁶⁷, the DOC fluxes were not produced in the model-data fusion analysis. This would bring uncertainty to carbon sink estimation in forests suffered soil erosion or land use change such as deforestation. Fortunately, the permanent plots at CERN sites are all protected well, but we still expect improved representations of carbon-water interaction process (e.g., DOC) into calibrated process-based models, to further help reduce the biases for the C balance of ecosystems regionally and globally.

Despite these inevitable uncertainties, the optimized key parameters and simulated C sequestration result are consistent with the site observational data and close to that of previous studies using different approaches (Figs. 4, 5 and S5). Overall, the CFCCD database provides high-quality open-access information on decadal C cycle dynamics in typical forests in China. The CFCCD database is the most comprehensive and up-to-date database covering decadal C cycle dynamics over the most representative forests in China and the Northern Hemisphere using measurement-based colocated networks; MDF-retrieved C cycle parameters, which are difficult to solely obtain from observations; and time-continuous C storage functions for long-term C cycle state evaluation under climate change. This reference dataset can be used to investigate the long-term trends in ecological C cycle dynamics, to identify the forest C sink distribution in soil and vegetation and their association with C-cycle process parameters,

and to evaluate and improve the ability of C cycle process models as benchmarks. Such knowledge will have strong implications improving our ability to evaluate and understand forest C cycle responses to global change and will be important in the implementation of C sequestration and mitigation by policy makers.

Code availability

The DALEC model and the model-data fusion code used to generate the assimilated data products can be obtained through the GitHub repository at (<https://github.com/ultradove/model-data-fusion>). Further questions can be directed towards: Rong Ge (ge7218@163.com).

Received: 4 June 2020; Accepted: 18 December 2020;

Published online: 02 February 2021

References

- Dixon, R. K. *et al.* Carbon Pools and Flux of Global Forest Ecosystems. *Science* **263**, 185–190 (1994).
- Pan, Y. *et al.* A Large and Persistent Carbon Sink in the World's Forests. *Science* **333**, 988–993 (2011).
- Goodale, C. L. *et al.* Forest carbon sinks in the northern hemisphere. *Ecol. Appl.* **12**, 891–899 (2002).
- Friedlingstein, P. *et al.* Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks. *J. Climate* **27**, 511–526 (2014).
- Hampton, S. E. *et al.* Big data and the future of ecology. *Front. Ecol. Environ.* **11**, 156–162 (2013).
- Jonathan, T. O. *et al.* Special online collection: dealing with data[J]. *Science* **331**(6018), 639–806 (2011).
- Huang, T. Q. & Dong, N. CHINESE ECOSYSTEM RESEARCH NETWORK (CERN)—Basic Information, Achievements and Perspectives. *Adv. Earth. Sci.* **20** (2005).
- Fu, B. *et al.* Chinese ecosystem research network: Progress and perspectives. *Ecol. Complex.* **7**, P225–233 (2010).
- Yu, G. *et al.* High carbon dioxide uptake by subtropical forest ecosystems in the East Asian monsoon region. *P. Natl. Acad. Sci. USA* **111**, 4910–4915 (2014).
- Sier, A. R. J. & Monteith, D. T. The UK Environmental Change Network after twenty years of integrated ecosystem assessment: Key findings and future perspectives. *Ecol. Indic.* **68**, 1–12 (2016).
- Peters, D. *et al.* *Long-Term Trends In Ecological Systems: A Basis For Understanding Responses To Global Change.* (USDA Agricultural Research Service, Washington, DC, 2012).
- Zhou, G. *et al.* Old-growth forests can accumulate carbon in soils. *Science* **314**, 1417–1417 (2006).
- Bin, W. & Xiaosheng, Y. Comparison of carbon content and carbon density of four typical zonal forest ecosystems. *J. Hunan Agri. Univ.* **36**, 464–469 (2010).
- Wu *et al.* Biogeographic patterns and influencing factors of the species diversity of tree layer community in typical forest ecosystems in China. *Acta Ecol. Sin.* **38**, 7727–7738 (2018).
- Zhou, G. *et al.* A climate change-induced threat to the ecological resilience of a subtropical monsoon evergreen broad-leaved forest in Southern China. *Global Change Biol.* **19**, 1197–1210 (2013).
- Song, X., Zeng, X. & Tian, D. Allocation of forest net primary production varies by forest age and air temperature. *Ecol. Evol.* **8**, 12163–12172 (2018).
- Luo, Y. *et al.* Ecosystem Carbon Allocation of a Temperate Mixed Forest and a Subtropical Evergreen Forest in China. *J. Res. Ecol.* **9**, 642–653 (2018).
- Zhang, L., Luo, Y., Yu, G. & Zhang, L. Estimated carbon residence times in three forest ecosystems of eastern China: Applications of probabilistic inversion. *J. Geophys. Res.* **115**, 137–147 (2010).
- Ge, R. *et al.* Underestimated ecosystem carbon turnover time and sequestration under the steady state assumption: A perspective from long-term data assimilation. *Global Change Biol.* **25**, 938–953 (2018).
- Carvalhais, N. *et al.* Global covariation of carbon turnover times with climate in terrestrial ecosystems. *Nature* **514**, 213–217 (2014).
- Bloom, A. A. & Exbrayat, J. Der Velde, I. R. V., Feng, L. & Williams, M. The decadal state of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and residence times. *P. Natl. Acad. Sci. USA* **113**, 1285–1290 (2016).
- De Kauwe, M. G. *et al.* Where does the carbon go? A model-data intercomparison of vegetation carbon allocation and turnover processes at two temperate forest free-air CO₂ enrichment sites. *New Phytol.* **203**, 883–899 (2014).
- Negrón-Juárez, R. I., Koven, C. D., Riley, W. J., Knox, R. G. & Chambers, J. Q. Observed allocations of productivity and biomass, and turnover times in tropical forests are not accurately represented in CMIP5 Earth system models. *Environ. Res. Lett.* **10**, 064017 (2015).
- Friend, A. D. *et al.* Carbon residence time dominates uncertainty in terrestrial vegetation responses to future climate and atmospheric CO₂. *P. Natl. Acad. Sci. USA* **111**, 3280–3285 (2014).
- Yan, Y., Zhou, X., Jiang, L. & Luo, Y. Effects of carbon turnover time on terrestrial ecosystem carbon storage. *Biogeosciences* **14**, 5441–5454 (2017).
- Xia, J., Yuan, W., Wang, Y. & Zhang, Q. Adaptive Carbon Allocation by Plants Enhances the Terrestrial Carbon Sink. *Sci. Rep.-UK.* **7**, 3341 (2017).
- Slevin, D., Tett, S. F. B., Exbrayat, J., Bloom, A. A. & Williams, M. Global evaluation of gross primary productivity in the JULES land surface model v3.4.1. *Geosci. Model Dev.* **10**, 2651–2670 (2016).
- López-Blanco, E. *et al.* Evaluation of terrestrial pan-Arctic carbon cycling using a data-assimilation system. *Earth Syst. Dynam.* **10**, 233–255 (2019).
- Liu, G. R. *et al.* *Protocols For Standard Atmosphere Environmental Observation And Measurement In Terrestrial Ecosystems.* (China Environmental Science Press, Beijing, 2007).
- Yuan, G., Tang, D., & Sun, X. *Protocols For Standard Water Observation And Measurement In Terrestrial Ecosystems.* (China Environmental Science Press, Beijing, 2007).
- Liu, H. *et al.* Photosynthetically active radiation dataset in Chinese Ecosystem Research Network (2005–2015). *China Scientific Data* **2**, 1–10 (2017).
- Luo, Y. J., Wang, X. K., and Lu F. *Comprehensive Database Of Biomass Regressions For China'S Tree Species.* (Beijing: China Forestry Publishing House, 2015).
- Post, W. M., Pastor, J., Zinke, P. J. & Stangenberger, A. G. Global patterns of soil nitrogen storage. *Nature* **317**, 613–616 (1985).
- Chai, H. *et al.* Vertical distribution of soil carbon, nitrogen, and phosphorus in typical Chinese terrestrial ecosystems. *Chinese Geogr. Sci.* **25**, 549–560 (2015).
- Li, C., He, H. L., Liu, M., Su, W. & Yu, G. R. The design and application of CO₂ flux data processing system at ChinaFLUX. *Geo Information Science* **10**, 557–565 (2008).
- Hurtt, G. C. & Armstrong, R. A pelagic ecosystem model calibrated with BATS data. *Deep-Sea Res. Pt. II* **43**, 653–683 (1996).
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
- Luo, Y. & Weng, E. Dynamic disequilibrium of the terrestrial carbon cycle under global change. *Trends Ecol. Evol.* **26**, 96–104 (2011).

39. Carvalhais, N. *et al.* Implications of the carbon cycle steady state assumption for biogeochemical modeling performance and inverse parameter retrieval. *Global Biogeochem. Cy.* **22**, 1081–1085 (2008).
40. Zhou, T., Shi, P., Jia, G. & Luo, Y. Nonsteady state carbon sequestration in forest ecosystems of China estimated by data assimilation. *J. Geophys. Res.-Biogeo.* **118**, 1369–1384 (2013).
41. Safta, C. *et al.* Global sensitivity analysis, probabilistic calibration, and predictive assessment for the data assimilation linked ecosystem carbon model. *Geosci. Model Dev.* **8**, 1899–1918 (2014).
42. Smallman, T. L., Exbrayat, J., Mencuccini, M., Bloom, A. A. & Williams, M. Assimilation of repeated woody biomass observations constrains decadal ecosystem carbon cycle uncertainty in aggrading forests. *J. Geophys. Res.* **122**, 528–545 (2017).
43. Richardson, A. D. *et al.* Estimating parameters of a forest ecosystem C model with measurements of stocks and fluxes as joint constraints. *Oecologia* **164**, 25–40 (2010).
44. Williams, M. *et al.* Predicting gross primary productivity in terrestrial ecosystems. *Ecol. Appl.* **7**, 882–894 (1997).
45. Williams, M., Schwarz, P. A., Law, B. E., Irvine, J. & Kurpius, M. R. An improved analysis of forest carbon dynamics using data assimilation. *Global Change Biol.* **11**, 89–105 (2005).
46. Fox, A. M. *et al.* The REFLEX project: Comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data. *Agr. Forest Meteorol.* **149**, 1597–1615 (2009).
47. Carvalhais, N. *et al.* Identification of vegetation and soil carbon pools out of equilibrium in a process model via eddy covariance and biometric constraints. *Global Change Biol.* **16**, 2813–2829 (2010).
48. Bloom, A. A. & Williams, M. Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological “common sense” in a model–data fusion framework. *Biogeosciences* **12**, 1299–1315 (2014).
49. Koven, C. D. *et al.* Controls on terrestrial carbon feedbacks by productivity versus turnover in the CMIP5 Earth System Models. *Biogeosciences* **12**, 5211–5228 (2015).
50. Schwartz, S. E. Residence times in reservoirs under non-steady-state conditions: application to atmospheric SO₂ and aerosol sulfate”. *Tellus* **31**, 530–547, <https://doi.org/10.3402/tellusa.v31i6.10471> (2016).
51. Zhou, T. & Luo, Y. Spatial patterns of ecosystem carbon residence time and NPP-driven carbon uptake in the conterminous United States. *Global Biogeochem. Cy.* **22**, 3411–3434 (2008).
52. He, H., Ge, R., Yu, G. Reference carbon cycle dataset for typical Chinese forests via colocated observations and data assimilation. *figshare* <https://doi.org/10.6084/m9.figshare.12331400.v2> (2020).
53. Sun, B., Shi, J. & Yang, L. *Protocols For Standard Soil Observation And Measurement In Terrestrial Ecosystems.* (China Environmental Science Press, Beijing, 2007).
54. Wu, D. X., Wei, W., & Zhang, S. *Protocols For Standard Biological Observation And Measurement In Terrestrial Ecosystems.* (China Environmental Science Press, Beijing, 2007).
55. Zheng, Z. M. *et al.* Spatio-temporal variability of soil respiration of forest ecosystems in China: influencing factors and evaluation model. *Environ. Manage.* **46**, 633–642, <https://doi.org/10.1007/s00267-010-9509-z> (2010).
56. Zhou, T., Shi, P., Jia, G., Li, X. & Luo, Y. Spatial patterns of ecosystem carbon residence time in Chinese forests. *Science China Earth Sciences* **53**, 1229–1240 (2010).
57. Wang, J. *et al.* Soil and vegetation carbon turnover times from tropical to boreal forests. *Funct. Ecol.* **32**, 71–82 (2018).
58. Piao, S. *et al.* Forest annual carbon cost: a global-scale analysis of autotrophic respiration. *Ecology* **91**, 652–661 (2010).
59. Chen, G., Yang, Y. & Robinson, D. Allocation of gross primary production in forest ecosystems: allometric constraints and environmental responses. *New Phytol.* **200**, 1176–1186 (2013).
60. Malhi, Y., Doughty, C. & Galbraith, D. The allocation of ecosystem net primary productivity in tropical forests. *Philos. T. R. Soc. B* **366**, 3225–45 (2011).
61. Wolf, A., Field, C. B. & Berry, J. A. Allometric growth and allocation in forests: a perspective from FLUXNET. *Ecol. Appl.* **21**, 1546–1556 (2011).
62. Wang-Erlandsson, L., van der Ent, R. J., Gordon, L. J. & Savenije, H. H. G. Contrasting roles of interception and transpiration in the hydrological cycle – Part 1: Temporal characteristics over land. *Earth Syst. Dynam.* **5**, 441–469 (2014).
63. Keenan, T. F., Carbone, M. S., Reichstein, M. & Richardson, A. D. The model-data fusion pitfall: assuming certainty in an uncertain world. *Oecologia* **167**, 587–597 (2011).
64. Davidson, E. & Janssens, I. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **440**, 165–173 (2006).
65. Luo, Y. *et al.* Toward more realistic projections of soil carbon dynamics by earth system models. *Global Biogeochem. Cy.* **30**, 40–56 (2016).
66. Larocque, G., Bhatti, J., Boutin, R. & Chertov, O. Uncertainty analysis in carbon cycle models of forest ecosystems: Research needs and development of a theoretical framework to estimate error propagation. *Ecol. Model.* **219**, 400–412 (2008).
67. Marinari, S., Liburdi, K., Fliessbach, A. & Kalbitz, K. Effects of organic management on water-extractable organic matter and mineralization in european arable soils. *Soil Till. Res.* **106**, 211–217 (2010).

Acknowledgements

We thank the CERN staff for their dedication to observation and data processing. This study was supported by the National Key Research and Development Program of China (grant no. 2016YFC0500204) and the Strategic Priority Research Program of the Chinese Academy of Sciences (grant no. XDA19020301).

Author contributions

H.H., R.G. and G.Y. designed the database; G.Y., Z.X., S.W., H.W., G.Z., Q.Z., A.W., Z.F., Y.Z., W.S., H.Y. and L.L. collected the data from CERN; H.H., R.G., X.R., L.Z., Q.C. and Q.X. compiled and generated the observation-based and assimilated dataset; M.W. developed the DALEC model and EDC concept. H.H., R.G., and X.R. established the data assimilation framework; and H.H. and R.G. wrote the paper with contributions from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-021-00826-w>.

Correspondence and requests for materials should be addressed to G.Y.

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

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2021