DR. JINSHI JIAN (Orcid ID : 0000-0002-5272-5367)
DR. BEN BOND-LAMBERTY (Orcid ID : 0000-0001-9525-4633)
DR. DALEI HAO (Orcid ID : 0000-0001-7154-6332)
DR. JIANQIU ZHENG (Orcid ID : 0000-0002-1609-9004)
MS. KALYN DORHEIM (Orcid ID : 0000-0001-8093-8397)
MS. STEPHANIE PENNINGTON (Orcid ID : 0000-0003-2685-1092)
DR. WILLIAM R WIEDER (Orcid ID : 0000-0001-7116-1985)

Article type : Primary Research Article

Leveraging observed soil heterotrophic respiration fluxes as a novel constraint on global-scale models

Jinshi Jian[1,2*], Ben Bond-Lamberty[2], Dalei Hao[3], Benjamin N. Sulman[4], Kaizad F. Patel[5], Jianqiu Zheng[5], Kalyn Dorheim[2], Stephanie C. Pennington[2], Melannie D. Hartman[6,7], Dan Warner[8], and William R Wieder[7,9]

1. State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Northwest A&F University, Yangling, 712100, China
2. Pacific Northwest National Laboratory, Joint Global Change Research Institute at the University of Maryland–College Park, College Park, MD 20740, USA
3. Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, WA, USA
4. Climate Change Science Institute and Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, US
5. Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

6. Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO, USA

7. Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, CO 80307, USA

8. Delaware Geological Survey, University of Delaware, Newark, DE, USA

9. Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO 80309, USA

* Corresponding author: Jinshi Jian

Email: jinshi@vt.edu

ORCID: 0000000252725367

Keywords: Soil organic carbon models, heterotrophic respiration, MIMICS, CASACNP, CORPSE, benchmarking

Running title: Benchmarking biogeochemical models

Abstract

Microbially-explicit models may improve understanding and projections of carbon dynamics in response to future climate change, but their fidelity in simulating global-scale soil heterotrophic respiration (R_H), a stringent test for soil biogeochemical models, has never been evaluated. We used statistical global R_H products, as well as 7,821 daily site-scale R_H measurements, to evaluate the spatio-temporal performance of one first-order decay model (CASA-CNP) and two microbially-explicit biogeochemical models (CORPSE and MIMICS) that were forced by two different input datasets. CORPSE and MIMICS did not provide any measurable performance improvement; instead, the models were highly sensitive to the input data used to drive them. Spatial variability in R_H fluxes was generally well simulated except in the northern middle latitudes (~50°N) and arid regions; models captured the seasonal variability of R_H well, but showed more divergence in tropic and arctic regions. Our results demonstrate that the next generation of biogeochemical models shows promise, but also needs to be improved for realistic spatio-temporal variability of R_H. Finally, we emphasize the importance of net primary production, soil moisture, and soil temperature inputs, and that jointly evaluating soil models for their spatial (global scale) and temporal (site scale) performance provides crucial benchmarks for improving biogeochemical models.

1. Introduction

The response of soil heterotrophic respiration (R_H) to environmental change will largely determine whether soils are a carbon sink or source in the future (Bond-Lamberty *et al.*, 2018). Despite the critical importance of microbes in driving this globally important carbon flux, current understanding of soil microbial communities and their potential responses to climate change is highly uncertain (Crowther *et al.*, 2019). Most biogeochemical models use a first-order decay process to describe soil carbon decomposition (Todd-Brown *et al.*, 2012), meaning that biological factors such as the size and composition of the decomposer microbial community, interactions of organic matter with soil minerals and aggregates, adaptations of microbial physiology, and priming are ignored (Schmidt *et al.*, 2011). This risks omitting crucial climatic feedback as the Earth system transitions to a novel and uncertain future (Wieder *et al.,* 2015a).

New models seek to represent such potential biotic feedback to environmental change, but evaluation of their carbon cycle representation and performance is nascent (Wieder *et al.*, 2015b). In recent years, many efforts have been made to include microbial biogeochemical mechanisms into a new generation of microbially-explicit models (Sulman *et al.*, 2014, 2018; Wieder *et al.*, 2014, 2015a, 2019). In theory, these models offer considerable advantages over first-order decay approaches in projecting future climate and carbon cycle feedback (e.g., they are capable of simulating population-driven dynamics independent of abiotic drivers). However, uncertainties in process representation and parameterization have led to divergent outcomes from different microbial-explicit models, leaving open questions of which if any model formulations improve predictive accuracy (Sulman *et al.*, 2018).

Evaluating global-scale soil biogeochemical models is challenging, given the lack of appropriate datasets that can serve as model benchmarks (Koven *et al.*, 2017; Collier *et al.*, 2018; Shi *et al.*, 2020). Wieder *et al.* (2018, 2019) developed a biogeochemical testbed to compare the performance of a first-order model, Carnegie-Ames-Stanford Approach (CASA-CNP) (Potter *et al.*, 1993; Randerson & Thompson, 1996; Wang *et al.*, 2010) vs. two microbially-explicit models, including MIcrobial-MIneral Carbon Stabilization (MIMICS) (Wieder *et al.*, 2014, 2015a), and Carbon, Organisms, Rhizosphere, and Protection in the Soil Environment (CORPSE) (Sulman *et al.*, 2014, 2017). Such a testbed provides a consistent environment for evaluating different models, similar to e.g. International Land Model Benchmarking (ILAMB) (Collier *et al.*, 2018). This work showed that different models lead to divergent SOC predictions with a distinct signature of heterotrophic respiration fluxes (Basile *et al.,* 2020; Wieder *et al.*, 2019).

Model performance needs to be evaluated by actual data. Wieder *et al.* (2013) and Wieder *et al.* (2015a) compared carbon storage simulation from testbed models vs. Harmonized world soil database (HWSD) carbon storage to evaluate the model performance. Basile *et al.* (2019) showed that spatial and temporal

variations in atmospheric CO_2 could be used to benchmark biogeochemical models. Predicting carbon fluxes constitutes an equally and perhaps more stringent test for models, compared with projecting SOC stocks or atmospheric CO_2 (Todd-Brown *et al.*, 2012, 2013). Heterotrophic respiration (R_H) is a direct consequence of microbial activities, and therefore accurately simulating R_H is a key metric of model performance, but R_H simulations from microbially-explicit models have never been compared with observational benchmarks at global scale. Emerging global soil respiration (R_S) databases provide an opportunity to evaluate models across different conditions based on *in situ* observations. Bond-Lamberty et al. (2010) and Jian et al. (2020) compiled published annual R_S (which also includes estimates of annual R_H) into a global R_S database (SRDB). Based on the SRDB, Jian *et al.* (2018) further compiled the daily and monthly R_S (which also includes R_H measurements) into a global daily R_S database (DGRsD), through which 7,821 daily R_H field measurements are available for model performance evaluation. Meanwhile, new global R_H datasets have been developed (Hashimoto *et al.*, 2015; Warner *et al.*, 2019; Tang *et al.*, 2020). Such field measurements and data-driven statistical R_H data products (Hashimoto *et al.*, 2015; Warner *et al.*, 2019; Tang *et al.*, 2020) offer promising opportunities to validate both existing first-order and new microbially-explicit models' performance at site, regional, and global scales.

The objectives of this study are to: 1) evaluate differences between CASA-CNP, CORPSE, and MIMICS models' predictions and observational benchmarks at multiple spatial scales; 2) investigate whether microbially-explicit models (CORPSE and MIMICS) outperform the first-order model (CASA-CNP); and 3) explore the main reasons causing R_H mismatch between models and benchmarks. We analyzed R_H simulations from CASA-CNP, CORPSE, and MIMICS with three global R_H data products derived from statistical models and 7,821 daily site-scale R_H observations from DGRsD as benchmarks (Figure 1 and Figure 2).

2. Methods

Previous studies have identified the importance of climate forcing in generating carbon cycle uncertainty (Todd-Brown *et al.*, 2013). In this study, we further examine the role of climate forcing in generating R_H uncertainty. The biogeochemical testbed is driven by external forcings that include gross primary productivity, soil and air temperature, and soil moisture (Wieder *et al.*, 2018). We generated these inputs from simulations with the Community Land Model, versions 4.5 (Oleson *et al.*, 2010) and 5.0 (Lawrence *et al.*, 2019). The simulations were run in satellite phenology mode with the default climate reanalysis for each model version: CLM4.5 uses atmospheric forcing data from National Centers for Environmental Prediction and Climatic Research Unit (CLM4.5-CRUNCEP), and CLM5.0 uses the Global Soil Wetness Project Phase 3 (CLM5.0-GSWP3). Although this approach complicates attributing differences in input

data to differences in model versions vs. climate forcings, previous work suggests that these combinations provide the most realistic soil temperatures, especially at high latitudes (Bonan et al. 2019; Lawrence et al. 2019). This study used the forcing data from CLM as boundary conditions to drive the CASA vegetation model, which generated net primary production (NPP) that was partitioned into different plant tissues and litterfall that drive the accumulation and decomposition of soil C stock simulated by CASA-CNP, MIMICS and CORPSE. We then compared the R_H fluxes simulated by the models to examine the R_H simulation uncertainty related to different forcing data (Figure 2).

Heterotrophic respiration (R_H) data from a biogeochemical testbed, statistical global R_H datasets, as well as daily timescale field R_H measurements were used in this study. For the biogeochemical testbed, the CASA-CNP, CORPSE, and MIMICS models were driven by the inputs from the Community Land Model (CLM, versions 4.5 and 5.0). These two versions of CLM were parameterized with different forcing datasets: CLM4.5 uses atmospheric reanalysis from Climatic Research Unit of National Centers for Environmental Prediction (CRUNCEP) for the period 1901-2010 (Wieder *et al.*, 2018); while CLM5.0 uses climate reanalysis from the Global Soil Wetness Project Phase 3 (GSWP3) for the period 1901-2014 (Dirmeyer *et al.*, 2006; Yoshimura & Kanamitsu, 2013). Daily modeled GPP, air and soil temperature, and soil moisture were used to drive the CASA vegetation model, which generated NPP and partitioning into different plant tissues and litterfall. Thus, the soil models experienced identical environmental drivers, when forced by CLM4.5-CRUNCEP and CLM5.0-GSWP3 inputs, respectively. This experimental design affords opportunities to evaluate uncertainties in external forcings (CLM4.5-CRUNCEP vs. CLM5.0-GSWP3 forced simulations). It also allows us to isolate the effects of model structural uncertainty among CASA-CNP, MIMICS, and CORPSE soil model formulations. This ensures that simulated R_H differences from those three models are caused by the microbial processes (first-order decay vs. microbially-explicit) and parametric differences between the models.

An exhaustive description of differences in the strengths and weaknesses of different versions of CLM is outside the scope of this paper. Lawrence et al. (2019) documented significant improvements in CLM5.0, relative to previous versions of the model. Notable improvements include reduced biases in gross primary productivity (especially across mid- and high-latitude ecosystems), improved representation of permafrost extent, and better agreement with observed terrestrial water storage anomalies. Similarly, Lawrence et al. (2019) reported better agreement with model benchmarks for air temperature, precipitation, and solar radiation for CLM simulations forced with GSWP3, compared to CRUNCEP forced runs. Thus, we assumed the improvements from these data to the testbed models will carry forward to our R_H results.

The details regarding the microbial processes in CASA-CNP, CORPSE, and MIMICS can be found at (Potter *et al.*, 1993; Randerson & Thompson, 1996; Wang *et al.*, 2010; Wieder *et al.*, 2013, 2014, 2015a, 2018; Sulman *et al.*, 2014, 2017), but we briefly described here: the CASA-CNP used first-order, linear decay rates modified by soil temperature and soil moisture to simulate microbial R_H that are proportional to soil organic matter pools (Potter *et al.*, 1993; Randerson & Thompson, 1996; Wang *et al.*, 2010); MIMICS was designed to evaluate the interactions of microbial physiology and soil properties (specifically soil texture) in moderating patterns of soil carbon persistence across large eco-climatological gradients (Wieder *et al.*, 2014, 2015a). MIMICS uses a temperature-sensitive reverse Michaelis-Menten kinetics to explicitly represent microbial activity that is moderated by substrate availability, soil temperature, and the availability of liquid soil water (Wieder *et al.*, 2014, 2015b). MIMICS simulates the activity of two microbial biomass communities that are characterized by having either rapid growth rate and low growth efficiencies or slow growth rates and higher growth efficiencies. The turnover of these microbial biomass pools are subject to density dependent microbial mortality rates as well as environmental conditions. CORPSE was developed to examine the priming (increased carbon inputs due to atmospheric CO_2 fertilization may accelerate old carbon decomposition) and protection (increased carbon inputs are protected through interactions with mineral particles) responses of SOC to rising atmospheric CO_2 (Sulman *et al.*, 2014). CORPSE also explicitly represents microbial activity and uses modified Michaelis-Menten kinetics that are sensitive to substrate availability, soil temperature and liquid water availability. CORPSE only represents a single microbial biomass pool and uses a fixed microbial mortality rate (Sulman *et al.*, 2014, 2017; Georgiou *et al.*, 2017). In these simulations MIMICS and CORPSE use the same function to modify soil C turnover based on liquid water availability (Sulman *et al.*, 2014).

Global R_H data products were used to evaluate the performance of microbially-explicit models. There are three global field-measurements-driven statistical R_H data products available, and all these R_H data products were developed based on the same global R_S database (SRDB, version 3) (Bond-Lamberty & Thomson, 2010) but used different modelling approaches. The first global R_H data product (http://cse.ffpri.affrc.go.jp/shojih/data/index.html) was developed using a Metropolis–Hastings algorithm to parameterize the relationship between R_S and climate factors (air temperature and precipitation) at the site-level. Specifically, this algorithm was used to up-scale R_S from site to globe between 1965 and 2012 at 0.5 ° × 0.5 ° spatial resolution (Hashimoto *et al.*, 2015). R_H was then calculated based on the relationship between R_H and R_S [$\ln(R_H) = 1.22 + 0.73 \ln(R_S)$] from a global meta-analysis (Bond-Lamberty *et al.*, 2004). As confidence interval (CI) of R_H was not directly reported, we calculated it based on the CI of global soil respiration (R_S) and its CI based on Hashimoto *et al.*(2015), i.e., CI(R_H) =

$4 \times \frac{51}{91}$. The second R_H product (https://doi.org/10.6084/m9.figshare.8882567) was developed using a random forest machine learning approach. Based on the third version of SRDB together with a R_H data search, 504 annual R_H observations were used to train the relationship between R_H and 9 environmental factors (including mean annual temperature, mean annual precipitation, diurnal temperature range, nitrogen deposition, Palmer Drought Severity Index, shortwave radiation, soil carbon content, soil nitrogen content, and soil water content) based on the random forest modelling approach. Then global annual R_H and related CI between 1980 and 2016 were predicted at 0.5 ° × 0.5 ° spatial resolution (Tang *et al.*, 2020). The third global R_H data product (https://daac.ornl.gov/CMS/guides/CMS_Global_Soil_Respiration.html) was developed using a quantile regression forest modelling approach. Specifically, field data was used to train the relationship between R_S and four environmental factors (mean annual temperature, mean annual precipitation, enhanced vegetation index, and mean winter precipitation) (Warner *et al.*, 2019), and mean annual R_H was then predicted at 1 km spatial resolution based on the R_H and R_S relationship [$\ln(R_H) = 1.22 + 0.73 \ln(R_S)$]. Warner *et al.* (2019) did not report CI, so we first calculated CI of R_S pixels (generated as CI of the mean of all random forest "trees" predictions at each pixel, was 4.7 Pg), and then calculated CI (R_H) = $4.7 \times \frac{49.8}{87.9}$.

Daily R_H measurements at the site scale were collected to evaluate the seasonal performance of microbially-explicit models at finer time and spatial scales. We went through the studies in SRDB which reported annual R_H measurements (red crosses in Figure 1), and checked whether detailed daily R_H measurements were reported; we then compiled these daily R_H measurements into DGRsD. In total 7,821 daily R_H observations from 254 studies were obtained (Figure 1, blue circles). We used latitude, longitude, year, and day of year to link these DGRsD data with the models' R_H outputs and thus obtain modeled R_H for all the observational sites and sampling times. Note that the spatial resolution of CASA-CNP, CORPSE, and MIMICS is 2.0° latitude × 2.5° longitude, meaning an inevitable spatial mismatch as we thus compare site-specific observations with grid cell-scale model outputs (Shao *et al.*, 2013). In addition, the CASA-CNP, CORPSE, and MIMICS runs ended in 2010 and 2014 (for CLM4.5-CRUNCEP and CLM5.0-GSWP3 forced simulations, respectively) but observational data for some sites were as late as 2017. In these cases, we used the 2000-2010 (when driven by CLM4.5-CRUNCEP) and 2000-2014 (when driven by CLM5.0-GSWP3) model average of a 3-day window around the observational date; for example, an R_H value measured on June-20th 2015 would be linked with the modeled R_H between June-17th and June-23th, averaged over the 2000-2010 period (CLM4.5-CRUNCEP forcing) or 2000-2014 period (CLM5.0-GSWP3 forcing). We did this because our focus was not on evaluating model performance at the daily scale, but rather the correctness of its overall seasonality. We then calculated

mean R_H (measured R_H as well as predicted R_H by CASACNP, MIMICS, and CORPSE) and their confidence interval (CI, within a specific group, CI = t_{score} × standard error) by day of year within six climate regions (Tropic, Arid, Temperate, Mediterranean, Boreal, and Arctic) to analyze models also reasonably capture the seasonal pattern of R_H in different climates.

Global scale R_H fluxes between 1980 and 2010 for CLM4.5-CRUNCEP, between 1980 and 2014 for CLM5.0-GSWP3 from the biogeochemical models (CASA-CNP, CORPSE, and MIMICS models) were compared with that from the global R_H statistical products as well as daily R_H measurements collected from 254 studies (Figure 1). Specifically, we investigated whether biogeochemical models can predict the magnitude and trend of global annual R_H under global climate change (Figure 2). Global annual R_H were summed up based on R_H rate and the area of each cell. We calculated global annual R_H between 1980 and 2016 when possible. We then compared R_H latitudinal patterns from the biogeochemical models vs. that from the global R_H data to investigate whether biogeochemical models well capture the R_H spatial variability. The spatial resolution of biogeochemical models and the global R_H data varied from 0.5 to 2.5 °, therefore, we averaged R_H rate by every 5 ° along the latitude gradient. Mean R_H rates along each latitude were then compared. We also compared NPP outputs from the CASA-CNP model vs. MODIS NPP (https://code.earthengine.google.com/) (Zhao *et al.*, 2005) to evaluate whether the difference between modeled R_H and benchmark R_H is related to the NPP inputs bias.

Site scale daily R_H vs. the prediction from the biogeochemical models was compared to investigate whether microbially-explicit models well predict daily and site scale R_H. Based on the latitude, longitude, and time information, we retrieved the R_H predictions from the three biogeochemical models. Linear regression was used to analyze the linear relationship between measured R_H and model predicted R_H. The raw data do not follow a normal distribution and thus it is difficult to compare the difference between measured R_H and modeled R_H. We thus used a bootstrap resampling approach to sample the mean of R_H from both measured and modeled R_H 10,000 times, and a non-parametric Wilcoxon test was then used to test whether modeled R_H were different from the measured R_H.

We further used the Wilcoxon test to compare the modeled R_H vs. measured R_H. From DGRsD, there are 254 studies with more than 5 R_H observations, for each study, we applied a non-parametric Wilcoxon test to investigate whether modeled R_H significantly differs from measured R_H. Based on the Wilcoxon-test p value and mean error (ME, i.e., mean of modeled R_H - measured R_H in each study, equation 1), all sites could be separated into three groups: overestimated ($p \leq 0.05$ and $ME > 0$), underestimated ($p \leq 0.05$ and $ME < 0$), and well estimated ($p \geq 0.05$).

$$ME = \frac{\Sigma_{i=1}^{n}(\hat{y}_i - y_i)}{n} \quad [1]$$

where $\hat{y}_i$ represents the i_{th} predicted R_H value and y_i represents the i_{th} measured R_H value (Yang *et al.*, 2014).

To investigate whether the R_H differences between models and benchmarks are related with model NPP inputs, we collected the MODIS NPP data from 2000-2019 using the platform of Google Earth Engine (https://code.earthengine.google.com/). Specifically, we used the MOD17A2H V6 NPP product, which has a 8-day temporal resolution and 500m spatial resolution (Zhao *et al.*, 2005). We further calculated the annual mean value and latitude gradient of MODIS NPP. We collected the NEE data covering 2001-2015 from the FLUXCOM project, which aims to upscale biosphere-atmosphere fluxes from local FLUXNET sites to continental and global scales (Tramontana *et al.*, 2016; Jung *et al.*, 2019). Here we used the monthly 0.5° latitude × 0.5° longitude FLUXCOM data with the setup of remote sensing to derive the annual mean values of NEE globally.

3. Results

Global annual R_H from CASA-CNP, CORPSE, and MIMICS driven by CLM4.5-CRUNCEP (51.0 Pg C yr^{-1}, all annual R_H from three models averaged, Figure 3a) forcing is in the range of that from statistical models (47.2 - 58.9 Pg C yr^{-1}, Figure 3a) (Hashimoto *et al.*, 2015; Warner *et al.*, 2019; Tang *et al.*, 2020). Latitudinally, R_H from CASA-CNP, CORPSE, and MIMICS match well with R_H from the statistical models, but exhibited a mismatch in the northern mid-latitudes and arid regions (Figure 3b). However, global annual R_H from CASA-CNP, CORPSE, and MIMICS driven by CLM5.0-GSWP3 (43 Pg C yr^{-1}, all annual R_H from three models averaged, Figure 3a) forcing is lower than that from statistical models (47.2 - 58.9 Pg C yr^{-1}). Differences in the global sum of R_H fluxes (Figure 3a) reflect higher net primary productivity with the CLM4.5 CRU-NCEP forced simulations, relative to the CLM5.0-GSWP3 forced simulations (Figure 3c). The spatial distribution of the fluxes, however, suggest that CRU-NCEP forced simulations have larger C fluxes in mid- to high latitudes (centered around 50 degrees N), which do not agree with upscaled R_H observations (Figure 3b and Figure 4).

In the site-scale comparison, when driven by CLM4.5-CRUNCEP, we found that for most of the sites, model-simulated R_H is higher than the field measured R_H (Figure 4). A non-parametric Wilcoxon test shows that within 254 studies we investigated, CASA-CNP, CORPSE, and MIMICS driven by CLM4.5-CRUNCEP forcing overestimated R_H for 169 sites (66%), underestimated R_H for 65 sites (26%), and only simulated R_H well for 8% of sites (Figure 4b). When driven by CLM5.0-GSWP3 forcing, the model

277 performance improved, as Root Mean Square Error (RMSE) decreased when compared with the results
278 driven by CRUNCEP (Figure S4). In addition, CASA-CNP, CORPSE, and MIMICS overestimated R_H
279 for 128 sites (50%), underestimated R_H for 96 sites (38%), and simulated R_H well for the remaining 30
280 (12%) sites (Figure 4e). The spatial pattern of R_H differences between CASA-CNP (averaged between
281 2000 and 2010) and Warner *et al.* (2019) showed that CASA-CNP overestimated R_H in northern middle
282 latitudes (Figure 4a). This problem was largely resolved when using GSWP3 to drive the model runs
283 (Figure 4d), however, model predicted R_H were still slightly higher than the measured R_H (Figure 5). The
284 spatial patterns of R_H from Hashimoto *et al.* (2015) and Tang *et al.* (2020) were almost identical to
285 Warner *et al.* (2019) (Figure S1). A similar conclusion could be obtained by comparing the distribution of
286 R_H simulations of CASA-CNP, CORPSE, and MIMICS with daily time scale R_H measurements from
287 individual sites (Figure 5).

288 When simulated R_H were compared with measured R_H by month and climate region, we found large
289 disagreements in most months and climate regions (Figure S2). A linear regression analysis (n=7,821)
290 showed that R_H simulations from MIMICS driven by CLM5.0-GSWP3 were weakly correlated with the
291 field observations ($R^2 = 0.11$); comparable percentages of variance explained for CASA-CNP and
292 CORPSE were 10.0% and 4.0%, respectively (Figure S3). The low correlation between field R_H
293 measurements and modeled R_H may result from the coarse spatial resolution of the testbed results (~ 2.0°
294 latitude × 2.5° longitude), and they are compared to site-level (typically 0.1-1.0 km^2) measurements.

295 The process models reasonably capture the seasonal pattern of R_H (Figure S5), but show differences when
296 separated into six climate regions, all models performance well in temperate ($R^2 > 0.40$) and boreal ($R^2 >$
297 0.50), but not for other climate regions ($R^2 < 0.20$, Figure 6). The seasonal pattern of modeled R_H was
298 also improved when driven by CLM5.0-GSWP3, especially in tropic and boreal regions (Figure 6). In
299 general, MIMICS captures the R_H seasonal variability slightly better compared with CORPSE, but similar
300 as CASA-CNP (Figure 6). However, MIMICS simulated R_H seems to have a later peak in the fall,
301 inconsistent with observations (Figure 6), a finding consistent with Basile et al. (2020). In the tropics, the
302 process models were unable to capture the large temporal amplitude of R_H (Figure 6). Similarly, in the
303 arctic, measured R_H showed a clear peak in the growing season, but models fail to capture this pattern.

304 **4. Discussion**

305 Biases in simulated R_H fluxes are strongly influenced by potential biases in plant productivity simulated
306 in the testbed models. The spatial distribution of R_H differences (biases, calculated as CASA-CNP
307 modeled R_H - observed R_H) showed a similar spatial pattern as the differences between CASA-CNP
308 modeled NPP and Moderate Resolution Imaging Spectroradiometer (MODIS) NPP (NPP differences,

Figure 4c and 4f). Specifically, large positive biases in simulated R_H fluxes from the CLM4.5-CRUNCEP forced simulations occurred in mid latitudes (~ 50° N) where we also found positive biases in simulated NPP, compared with MODIS estimates. These productivity biases were reduced in the CLM5.0-GSWP3 forced simulations, leading to an improvement in simulated R_H as well.

However, NPP inputs alone cannot explain the process models' overestimation of R_H. For example, underestimated R_H were mostly observed at central Australia and Tibet Plateau (Figure 4a, 4b, and Figure S1a), but CASA-CNP modeled NPP showed no differences compared with the MODIS NPP there (Figure 3c, Figure 4c and 4f). One possibility is that the autotrophic and heterotrophic carbon fluxes simulated by the land models will largely balance out, but observationally derived data products of NPP and R_H are not necessarily internally consistent with each other. Indeed, the Net Ecosystem Exchange (NEE) that would be derived from the difference between observationally based NPP and R_H estimates would produce large carbon sinks across the tropics and middle latitudes according to the FLUXCOM NEE products (Tramontana *et al.*, 2016; Jung *et al.*, 2019), but this carbon sink is poorly captured by these simulations (Figure 3d).

When using GSWP3 as model forcing, the site scale daily R_H comparison showed that simulated R_H from CASA-CNP, CORPSE, and MIMICS are much closer to the measured R_H, while annual R_H from model results from CLM5.0-GSWP3 forcing (43 Pg C yr^{-1}, Figure 3a) are lower than the statistical benchmarks (47.2 - 58.9 Pg C yr^{-1}) (Hashimoto *et al.*, 2015; Warner *et al.*, 2019; Tang *et al.*, 2020), What might cause these discrepancies? First, we recognize that gridded statistical estimates of R_H fluxes are themselves uncertain. For example, lower global R_H has been reported by calculating R_H from satellite-driven estimates (global $R_H = 43.6 \pm 19.3$ Pg C yr^{-1}; mean ± SD) (Konings *et al.*, 2019) and by a comprehensive global bottom-up carbon budget accounting (global $R_H = 39 \pm 6$ Pg C yr^{-1}) (Ciais *et al.*, 2020) compared with the estimates from statistical models shown in Figure 3 (Hashimoto *et al.*, 2015; Warner *et al.*, 2019; Tang *et al.*, 2020). Second, Jian *et al.*, (2018a) suggests that R_S sites' uneven distribution from the global R_S database causes about 6 Pg C overestimate of global annual R_S. Finally, Jian *et al.*, (2018b) also posited that the temporal variability of soil respiration plays an important role on global soil respiration modeling and estimates, with global soil respiration prediction based on monthly soil respiration data about 10 Pg C smaller than that based on annual data. This suggests that the mismatch between modeled R_H (driven by GSWP3 forcing) and benchmarks may be related to R_S sites spatial uneven distribution and temporal variability in R_H fluxes.

Another potential source of bias in the models, compared to observationally extrapolated R_H fluxes, is that the modeled ratio of R_H to NPP in the biogeochemical models is too high. The value of R_H to NPP ratio

reported by the IPCC assessment report is 0.9 (Stocker *et al.*, 2013), and in testbed models this ratio is almost 1 (i.e., NEE values are very close to 0, Figure 3d). R_H to NPP ratio estimates from IPCC and testbed models, however, may be too high because the models generally do not consider dissolved and particulate organic carbon losses to rivers and erosion (Cole *et al.*, 2007; Tan *et al.*, 2020), crop harvest and grazing (Guenther *et al.*, 2012; Ciais *et al.*, 2020), or carbon emission due to fire (Werf *et al.*, 2017). As a result, Tan et al. (Tan *et al.*, 2020) suggested that too much carbon is transferred to soils in the models compared to reality, which results in a higher R_H to NPP ratios. This suggests that the associated turnover time (and thus CO_2 emissions) of soil C pools may be more uncertain than currently thought (Carvalhais *et al.*, 2014).

Alternatively, higher R_H estimates from biogeochemical models could be due to model parameterizations with too-low carbon use efficiency (relative to transfers among soil C pools) (Geyer *et al.*, 2016). Microbially-explicit models advance the representation of SOC dynamics and turnover under global climate change, but parameterizing them remains an outstanding challenge (Wieder *et al.*, 2015b, 2018; Bradford *et al.*, 2016). The field observations compiled in this study can be used to constrain the model parameters for the next generation model improvement. Similarly, Zhang et al. (2020) used data from 72 sites in Europe and 134 sites in China to calibrate the parameter for SOC deprotection rate, improving the performance of MIMICS. The key parameters related to SOC decomposition in the default CORPSE and MIMICS (such as the microbial temperature sensitivity and microbial mortality rate) were parameterized using laboratory (German *et al.*, 2012) and field data (Wieder *et al.*, 2013, 2014, 2015b), but increasing field data collected from different environments across the globe should improve these models' performance. This also emphasizes the need to collect relevant environmental covariates, especially soil temperature and soil moisture data, as well as site level data on plant productivity and soil characteristics from field R_S studies.

Such model benchmarking, evaluation, and diagnosis exercises are most powerful–provide the most scientific benefit–when performed using high spatio-temporal resolution global R_H data products (Collier *et al.*, 2018). Currently, there are three global observation-driven statistical R_H data products available, but each was developed from the same underlying global R_S database (SRDB, version 3) (Bond-Lamberty & Thomson, 2010), reducing their independence and thus benefit for benchmarking. In addition, currently global R_H estimates are available only on annual timescales. Nonetheless, incorporating these observations into benchmarking packages such as ILAMB (Collier *et al.*, 2018) will provide useful tests to evaluate the representation of soil biogeochemistry in land models. Higher frequency daily and monthly R_H fluxes are likely necessary to understand and evaluate the different sensitivities in soil biogeochemical models, sub-annual biases and transient processes that produce hot spots and hot

moments (Bernhardt *et al.*, 2017), and the distribution and occurrence of extreme values across temporal and spatial scales.

Conclusion

Microbially-based soil models likely hold the key to predicting 21st-century soil carbon climate feedback accurately, but their development and evaluation remains a challenge. This study is the first to evaluate such models' heterotrophic respiration fluxes against observed R_H at both local and global scales. The three biogeochemical models we evaluated (the first-order land model CASA-CNP, and the microbial CORPSE and MIMICS models) reasonably simulate annual R_H and its spatio-temporal variability compared to three data-driven statistical global R_H data products and site-scale daily measurements. The forcing dataset CLM5.0-GSWP3 provided significantly improved results compared to CLM4.5-CRUNCEP. The spatial variability of R_H from the biogeochemical models is highly affected by the model NPP and litterfall inputs, and all models exhibited temporal biases at the site scale. We conclude that (i) it is important to improve NPP and litterfall (i.e., the carbon inputs to soil heterotrophs) in the next generation of biogeochemical models; (ii) joint evaluations of models at multiple spatial and temporal scales provides a stringent test of their performance; and (iii) microbial models' performance, at least in the group examined here, is already at least as good as traditional first-order alternatives, high temporal- and spatial-resolution datasets will be key to evaluating and improving these models in the future.

Acknowledgments: J.J was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDA20040202). B.B.-L., B.N.S., K.F.P, J. Z., K. D., S.C.P, and J.J were supported by the US Department of Energy, Office of Science, Biological and Environmental Research as part of the Terrestrial Ecosystem Sciences Program. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. W.R.W. was supported by the U.S. Department of Energy under award number BSS DE-SC0016364, US Department of Agriculture NIFA Award number 2015-67003-23485, and NASA NNX17AK19G. D.H. was supported by the U.S. National Oceanic and Atmospheric Administration (NOAA) under award number NOAA-OAR-CPO-2019-2005530.

Data and materials availability: All code used in this analysis are available at https://github.com/PNNL-TES/landmodels. Data outputs from CASA-CNP, CORPSE, and MIMICS are too big to share at GitHub, please contact the corresponding author if users are interested in using the data to reproduce all the results.

Author Contributions: W.R.W., B.B.-L., and J.J. conceived this study and designed the primary analysis. W.R.W., M.D.H. and B.N.S. prepared heterotrophic respiration outputs from CASA-CNP, CORPSE, and MIMICS driven by CLM4.5-CRUNCEP as well as CLM5.0-GSWP3. D.H., K.D., and S.P processed and analyzed the spatial distribution of heterotrophic respiration data. J.J. and W.R.W. wrote the manuscript in close collaboration with all authors. All authors provided feedback and insights in all phases.

Competing Interest Statement: None.

Figure caption

Figure 1. Geographic distribution of heterotrophic respiration (R_H) sites from the daily global soil respiration (R_S) database (DGRsD, daily R_H collected from 254 individual articles, show as blue circles) (Jian *et al.*, 2018) and a global R_S database (SRDB, red crosses) (Bond-Lamberty & Thomson, 2010); Three global R_H datasets (Hashimoto *et al.*, 2015; Warner *et al.*, 2019; Tang *et al.*, 2020) were developed based on annual R_S or R_H measurements. Point sizes represent the number of observations in each location (unique latitude and longitude).

Figure 2. Diagram shows the workflow of this study. Data driven statistical global R_H products (Hashimoto *et al.*, 2015; Warner *et al.*, 2019; Tang *et al.*, 2020) and filed R_H measurements collected from articles were used as the benchmarks for the biogeochemical models (CASA-CNP, CORPSE, and MIMICS) in this study. We compared the spatio-temporal variability of R_H to evaluate the performance of biogeochemical models. Net Primary Production (NPP) and Net Primary Exchange (NEE) from Moderate Resolution Imaging Spectroradiometer (MODIS) (Zhao *et al.*, 2005) and FLUXCOM (Tramontana *et al.*, 2016; Jung *et al.*, 2019) are used to investigate whether the R_H differences between models and benchmarks are related with model NPP inputs.

Figure 3. (a) Comparison of global annual mean heterotrophic respiration (R_H) predicted by data driven statistical models (benchmark) (Hashimoto *et al.*, 2015; Warner *et al.*, 2019; Tang *et al.*, 2020) and biogeochemical models (CASA, CORPSE, and MIMICS forced by CLM4.5-CRUNCEP and CLM5.0-GSWP3). The shaded area showed the confidence interval (CI) of R_H. **(b)** Zonal mean R_H (mean rate: g C m^{-2} yr^{-1}) along latitude, R_H predicted by Warner *et al.*, (2019) is used as a benchmark, R_H simulated by CASA-CNP using CLM4.5-CRUNCEP and CLM5.0-GSWP3 forcings are compared with the benchmark (within a specific zonal band, e.g., 85° - 90°, CI of (Warner et al., 2019) was calculated according to CI = t_{score} × standard error, CI of NPP and NEE benchmarks were calculated similarly); **(c)** Zonal mean Net Primary Production (NPP) fluxes with latitude (mean rate between 1980 and 2010: g C m^{-2} yr^{-1}), NPP (with CI) from MODIS is showed as the benchmark. **(d)** Zonal mean Net Ecosystem Exchange (NEE) with latitude (mean rate between 1980 and 2010: g C m^{-2} yr^{-1}, NEE = R_H - NPP for simulations), NEE (with CI) from FLUXCOM (Tramontana *et al.*, 2016; Jung *et al.*, 2019) is showed as the benchmark. Note that CI of MODIS NPP and FLUXCOM NEE are also shown in panel c and d, but they are too small to see clearly.

Figure 4. (a) Global spatial distribution of soil heterotrophic respiration (R_H) differences between CASA-CNP (driven by CLM4.5-CRUNCEP forcing) and data-driven statistical model result from ref (Warner *et al.*, 2019); (b) R_H differences between CASA-CNP (driven by CLM4.5-CRUNCEP forcing) and daily measured R_H from 254 studies. (c) NPP differences between CASA-CNP model and MODIS data. (d, e, and f) similar as a, b, and c, but CASA-CNP was driven by CLM5.0-GSWP3 forcing.

Figure 5. Distributions of bootstrap resampled measured daily R_H and daily R_H predicted from CASA-CNP, CORPSE, and MIMICS models driven by CLM4.5-CRUNCEP forcing and CLM5.0-GSWP3 forcing.

Figure 6. Heterotrophic respiration (R_H) seasonal pattern (averaged by day of year) across tropic, arid, temperate, mediterranean, boreal, and arctic. Panels from left to right are the comparison between measured R_H (gray) and CASA modeled R_H driven with CRUNCEP forcing, CASA modeled R_H driven with GSWP3 forcing, CORPSE modeled R_H driven with GSWP3 forcing, and MIMICS modeled R_H driven with GSWP3 forcing. The temporal trend of R_H simulated by CORPSE and MIMICS driven by CRUNCEP forcing were almost identical to R_H simulated driven by GSWP3, therefore the results were not shown. Note that only measurements from the northern hemisphere were used. R^2 and RMSE are from the linear regression between measured R_H and model predicted R_H.

References

Basile, S.J., Lin, X., Wieder, W.R., Hartman, M.D. & Keppel-Aleks, G. (2020) Leveraging the signature of heterotrophic respiration on atmospheric CO 2 for model benchmarking. *Biogeosciences*, 17(5): 1293-1308. https://doi.org/10.5194/bg-17-1293-2020.

Bernhardt E. S., Blaszczak J. R., Ficken C. D., Fork M. L., Kaiser K. E. & Seybold E. C. (2017). Control Points in Ecosystems: Moving Beyond the Hot Spot Hot Moment Concept. *Ecosystems* , *20*(4), 665–682. https://doi.org/10.1007/s10021-016-0103-y.

Bonan, G. B., Lombardozzi, D. L., Wieder, W. R., Oleson, K. W., Lawrence, D. M., Hoffman, F. M., & Collier, N. (2019). Model Structure and Climate Data Uncertainty in Historical Simulations of the Terrestrial Carbon Cycle (1850–2014). *Global Biogeochemical Cycles,* 33(10), 1310-1326. doi: 10.1029/2019gb006175.

Bond-Lamberty B., Wang C. & Gower S. T. (2004). A global relationship between the heterotrophic and autotrophic components of soil respiration? *Global change biology*, *10*(10), 1756–1766. https://doi.org/10.1111/j.1365-2486.2004.00816.x.

Bond-Lamberty B. & Thomson A. (2010). A global database of soil respiration data. *Biogeosciences* , *7*(6), 1915–1926. https://doi.org/10.5194/bg-7-1915-2010.

Bond-Lamberty B., Bailey V. L., Chen M., Gough C. M. & Vargas R. (2018). Globally rising soil heterotrophic respiration over recent decades. *Nature*, *560*(7716), 80–83. https://doi.org/10.1038/s41586-018-0358-x.

Bradford M. A., Wieder W. R., Bonan G. B., Fierer N., Raymond P. A. & Crowther T. W. (2016). Managing uncertainty in soil carbon feedbacks to climate change. *Nature climate change*, *6*(8), 751–758. https://doi.org/10.1038/nclimate3071.

Carvalhais N., Forkel M., Khomik M., Bellarby J., Jung M., Migliavacca M., … Reichstein M. (2014). Global covariation of carbon turnover times with climate in terrestrial ecosystems. *Nature*, *514*(7521), 213–217. https://doi.org/10.1038/nature13731.

Ciais P., Yao Y., Gasser T., Baccini A., Wang Y., Lauerwald R., … Zhu D. (2020). Empirical estimates of regional carbon budgets imply reduced global soil heterotrophic respiration. *National Science Review*. https://doi.org/10.1093/nsr/nwaa145.

Cole J. J., Prairie Y. T., Caraco N. F., McDowell W. H., Tranvik L. J., Striegl R. G., … Melack J. (2007). Plumbing the Global Carbon Cycle: Integrating Inland Waters into the Terrestrial Carbon Budget. *Ecosystems* , *10*(1), 172–185. https://doi.org/10.1007/s10021-006-9013-8.

Collier N., Hoffman F. M., Lawrence D. M., Keppel-Aleks G., Koven C. D., Riley W. J., … Randerson J. T. (2018). The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754. https://doi.org/10.1029/2018MS001354.

Crowther T. W., van den Hoogen J., Wan J., Mayes M. A., Keiser A. D., Mo L., … Maynard D. S. (2019). The global soil community and its influence on biogeochemistry. *Science*, *365*(6455). https://doi.org/10.1126/science.aav0550.

Dirmeyer P. A., Gao X., Zhao M., Guo Z., Oki T. & Hanasaki N. (2006). GSWP-2: Multimodel Analysis and Implications for Our Perception of the Land Surface. *Bulletin of the American Meteorological Society*, *87*(10), 1381–1398. https://doi.org/10.1175/BAMS-87-10-1381.

Georgiou K., Abramoff R. Z., Harte J., Riley W. J. & Torn M. S. (2017). Microbial community-level regulation explains soil carbon responses to long-term litter manipulations. *Nature communications*, *8*(1), 1223. https://doi.org/10.1038/s41467-017-01116-z.

German D. P., Marcelo K. R. B., Stone M. M. & Allison S. D. (2012). The Michaelis-M enten kinetics of soil extracellular enzymes in response to temperature: a cross-latitudinal study. *Global change biology*, *18*(4), 1468–1479.

Geyer K. M., Kyker-Snowman E., Grandy A. S. & Frey S. D. (2016). Microbial carbon use efficiency: accounting for population, community, and ecosystem-scale controls over the fate of metabolized organic matter. *Biogeochemistry*, *127*(2), 173–188. https://doi.org/10.1007/s10533-016-0191-y.

Guenther A. B., Jiang X., Heald C. L., Sakulyanontvittaya T., Duhl T., Emmons L. K. & Wang X. (2012). The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions. *Geoscientific Model Development*, *5*(6), 1471–1492.

Hashimoto S., Carvalhais N., Ito A., Migliavacca M., Nishina K. & Reichstein M. (2015). Global spatiotemporal distribution of soil respiration modeled using a global database. *Biogeosciences* , *12*, 4121–4132.

Jian J., Steele M. K., Day S. D. & Thomas R. Q. (2018a). Future global soil respiration rates will swell despite regional decreases in temperature sensitivity caused by rising temperature. Earth's Future, 6 (11), 1539-1554.

Jian J., Steele M. K., Thomas R. Q., Day S. D. & Hodges S. C. (2018b). Constraining estimates of global soil respiration by quantifying sources of variability. *Global change biology*.

https://doi.org/10.1111/gcb.14301.

Jian J., Vargas R., Anderson-Teixeira K., Stell E., Herrmann V., Horn M., … Bond-Lamberty B. (2020). A restructured and updated global soil respiration database (SRDB-V5). *Earth System Science Data Discussions*, 1–19.

Jung M., Schwalm C., Migliavacca M., Walther S., Camps-Valls G., Koirala S., … Walker A. (2019). Scaling carbon fluxes from eddy covariance sites to globe: Synthesis and evaluation of the FLUXCOM approach. *Biogeosciences discussions* , 1–40. https://doi.org/10.5194/bg-2019-368.

Konings A. G., Bloom A. A., Liu J., Parazoo N. C., Schimel D. S. & Bowman K. W. (2019). Global satellite-driven estimates of heterotrophic respiration. *Biogeosciences* , *16*(11), 2269–2284. https://doi.org/10.5194/bg-16-2269-2019.

Koven C. D., Hugelius G., Lawrence D. M. & Wieder W. R. (2017). Higher climatological temperature sensitivity of soil carbon in cold than warm climates. *Nature climate change*, *7*(11), 817–822. https://doi.org/10.1038/nclimate3421.

Lawrence D. M., Fisher R. A., Koven C. D., Oleson K. W., Swenson S. C., Bonan G., … Zeng X. (2019). The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4245–4287. https://doi.org/10.1029/2018MS001583.

Oleson K. W., Lawrence D. M., Bonan G. B. & Flanner M. G. (2010). Technical description of version 4.5 of the Community Land Model (CLM), NCAR Tech. *Notes (NCAR/TN-478+ STR*.

Potter C. S., Randerson J. T., Field C. B., Matson P. A., Vitousek P. M., Mooney H. A. & Klooster S. A. (1993). Terrestrial ecosystem production: A process model based on global satellite and surface data. *Global biogeochemical cycles*, *7*(4), 811–841. https://doi.org/10.1029/93GB02725.

Randerson J. T. & Thompson M. V. (1996). Substrate limitations for heterotrophs: Implications for models that estimate the seasonal cycle of atmospheric CO2. *Global*.

Schmidt M. W. I., Torn M. S., Abiven S., Dittmar T., Guggenberger G., Janssens I. A., … Trumbore S. E. (2011). Persistence of soil organic matter as an ecosystem property. *Nature*, *478*(7367), 49–56. https://doi.org/10.1038/nature10386.

Shao P., Zeng X., Moore D. J. P. & Zeng X. (2013). Soil microbial respiration from observations and Earth System Models. *Environmental research letters: ERL [Web site]*, *8*(3), 034034. https://doi.org/10.1088/1748-9326/8/3/034034.

Shi Z., Allison S. D., He Y., Levine P. A., Hoyt A. M., Beem-Miller J., … Randerson J. T. (2020). The

554 age distribution of global soil carbon inferred from radiocarbon measurements. *Nature geoscience*,
555 *13*(8), 555–559. https://doi.org/10.1038/s41561-020-0596-z.

556 Stocker T. F., Qin D., Plattner G.-K., Tignor M., Allen S. K., Boschung J., … Wuebbles, D. (2013).
557 Climate change 2013: The physical science basis. *Contribution of working group I to the fifth*
558 *assessment report of the intergovernmental panel on climate change*, *1535*.

559 Sulman B. N., Moore J. A. M., Abramoff R., Averill C., Kivlin S., Georgiou K., … Classen A. T. (2018).
560 Multiple models and experiments underscore large uncertainty in soil carbon dynamics.
561 *Biogeochemistry*, *141*(2), 109–123. https://doi.org/10.1007/s10533-018-0509-z.

562 Sulman B. N., Phillips R. P., Oishi A. C., Shevliakova E. & Pacala S. W. (2014). Microbe-driven
563 turnover offsets mineral-mediated storage of soil carbon under elevated CO2. *Nature climate change*,
564 *4*(12), 1099–1102.

565 Sulman B. N., Brzostek E. R., Medici C., Shevliakova E., Menge D. N. L. & Phillips R. P. (2017).
566 Feedbacks between plant N demand and rhizosphere priming depend on type of mycorrhizal
567 association (ed Cleland E.). *Ecology letters*, *20*(8), 1043–1053. https://doi.org/10.1111/ele.12802.

568 Tang X., Fan S., Du M., Zhang W., Gao S., Liu S., … Yang W. (2020). Spatial-and temporal-pattern of
569 global soil heterotrophic respiration in terrestrial ecosystems. *earth-syst-sci-data-discuss.net*, *12*(2),
570 1037–1051. https://doi.org/10.5194/essd-12-1037-2020.

571 Tan Z., Leung L. R., Li H.-Y., Tesfa T., Zhu Q. & Huang M. (2020). A substantial role of soil erosion in
572 the land carbon sink and its future changes. *Global change biology*.
573 https://doi.org/10.1111/gcb.14982.

574 Todd-Brown K. E., Randerson J. T. & Post W. M. (2013). Causes of variation in soil carbon simulations
575 from CMIP5 Earth system models and comparison with observations. *Biogeosciences*.

576 Todd-Brown K. E. O., Hopkins F. M., Kivlin S. N. & Talbot J. M. (2012). A framework for representing
577 microbial decomposition in coupled climate models. *Biogeochemistry*.

578 Tramontana G., Jung M., Schwalm C. R., Ichii K., Camps-Valls G., Ráduly B., … Papale D. (2016).
579 Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression
580 algorithms. *Biogeosciences* , *13*(14), 4291–4313.

581 Wang Y. P., Law R. M. & Pak B. (2010). A global model of carbon, nitrogen and phosphorus cycles for
582 the terrestrial biosphere. *Biogeosciences* , *7*(7).

583 Warner D. L., Bond-Lamberty B., Jian J., Stell E. & Vargas R. (2019). Spatial predictions and associated
584 uncertainty of annual soil respiration at the global scale. *Global biogeochemical cycles*, *7*, 983.

https://doi.org/10.1029/2019GB006264.

Werf G. R. van der, van der Werf G. R., Randerson J. T., Giglio L., van Leeuwen T. T., Chen Y., … Kasibhatla P. S. (2017). Global fire emissions estimates during 1997–2016. *Earth System Science Data*, *9*(2), 697–720.

Wieder W. R., Bonan G. B. & Allison S. D. (2013). Global soil carbon projections are improved by modelling microbial processes. *Nature climate change*, *3*(10), 909–912. https://doi.org/10.1038/nclimate1951.

Wieder W. R., Grandy A. S., Kallenbach C. M. & Bonan G. B. (2014). Integrating microbial physiology and physio-chemical principles in soils with the MIcrobial-MIneral Carbon Stabilization (MIMICS) model. *Biogeosciences* , *11*(14), 3899.

Wieder W. R., Grandy A. S., Kallenbach C. M., Taylor P. G. & Bonan G. B. (2015a). Representing life in the Earth system with soil microbial functional traits in the MIMICS model. *Geoscientific Model Development Discussions*, *8*(2).

Wieder W. R., Hartman M. D., Sulman B. N., Wang Y.-P., Koven C. D. & Bonan G. B. (2018). Carbon cycle confidence and uncertainty: Exploring variation among soil biogeochemical models. *Global change biology*, *24*(4), 1563–1579. https://doi.org/10.1111/gcb.13979.

Wieder W. R., Sulman B. N., Hartman M. D., Koven C. D. & Bradford M. A. (2019). Arctic Soil Governs Whether Climate Change Drives Global Losses or Gains in Soil Carbon. *Geophysical research letters*, *46*(24), 14486–14495. https://doi.org/10.1029/2019GL085543.

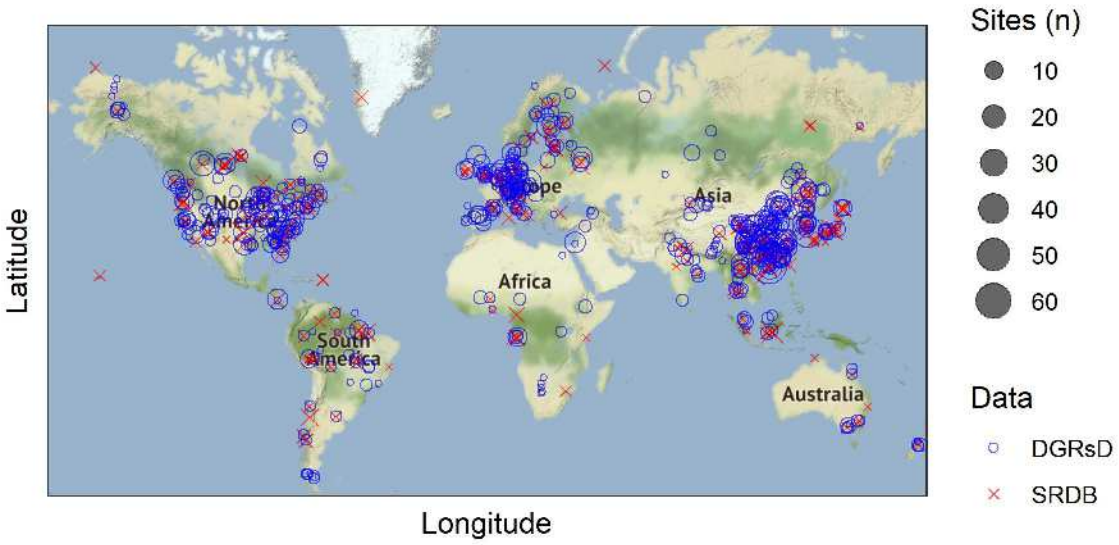
Wieder W. R., Allison S. D., Davidson E. A., Georgiou K., Hararuk O., He Y., … Xu X (2015b). Explicitly representing soil microbial processes in Earth system models. *Global biogeochemical cycles*, *29*(10), 1782–1800.

Yang J. M., Yang J. Y., Liu S. & Hoogenboom G. (2014). An evaluation of the statistical methods for testing the performance of crop models with observed data. *Agricultural systems*, *127*, 81–89. https://doi.org/10.1016/j.agsy.2014.01.008.

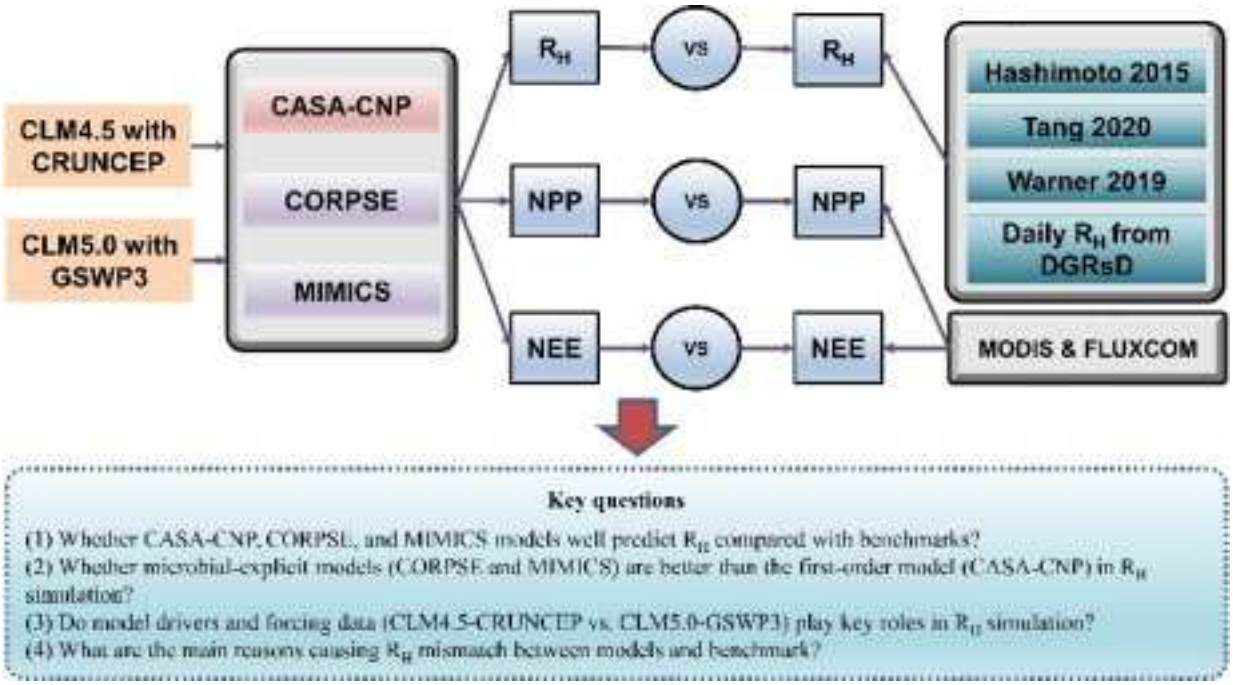
Yoshimura K. & Kanamitsu M. (2013). Incremental Correction for the Dynamical Downscaling of Ensemble Mean Atmospheric Fields. *Monthly Weather Review*, *141*(9), 3087–3101. https://doi.org/10.1175/MWR-D-12-00271.1.

Zhang H., Goll D. S., Wang Y.-P., Ciais P., Wieder W. R., Abramoff R., … Tang X. (2020). Microbial dynamics and soil physicochemical properties explain large-scale variations in soil organic carbon. *Global change biology*, *26*(4).

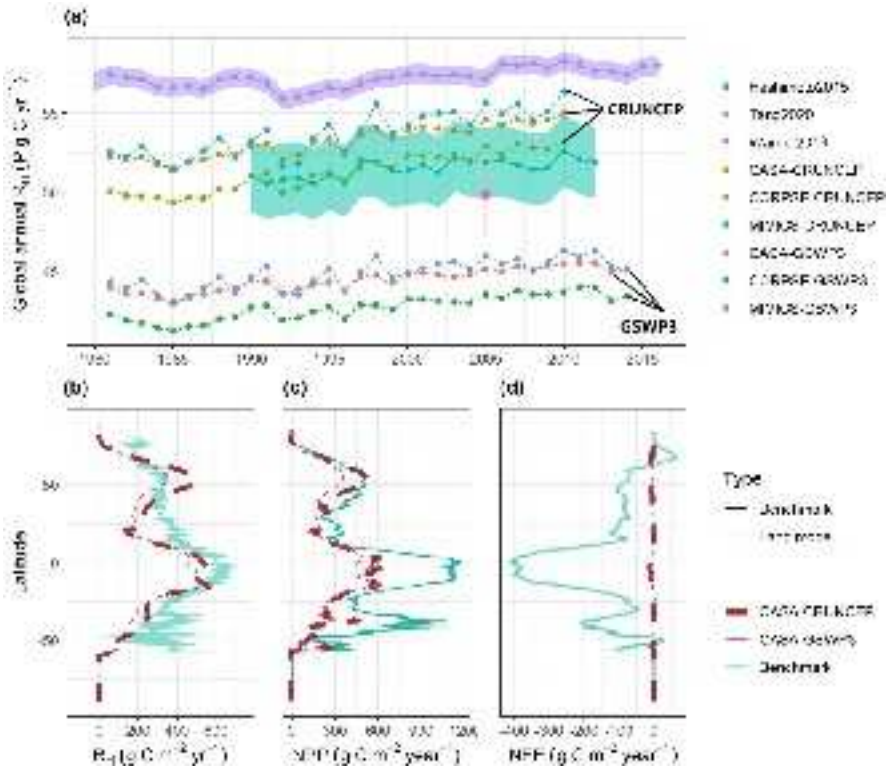
Zhao M., Heinsch F. A., Nemani R. R. & Running S. W. (2005). Improvements of the MODIS terrestrial gross and net primary production global data set. *Remote sensing of environment*, *95*(2), 164–176. https://doi.org/10.1016/j.rse.2004.12.011.



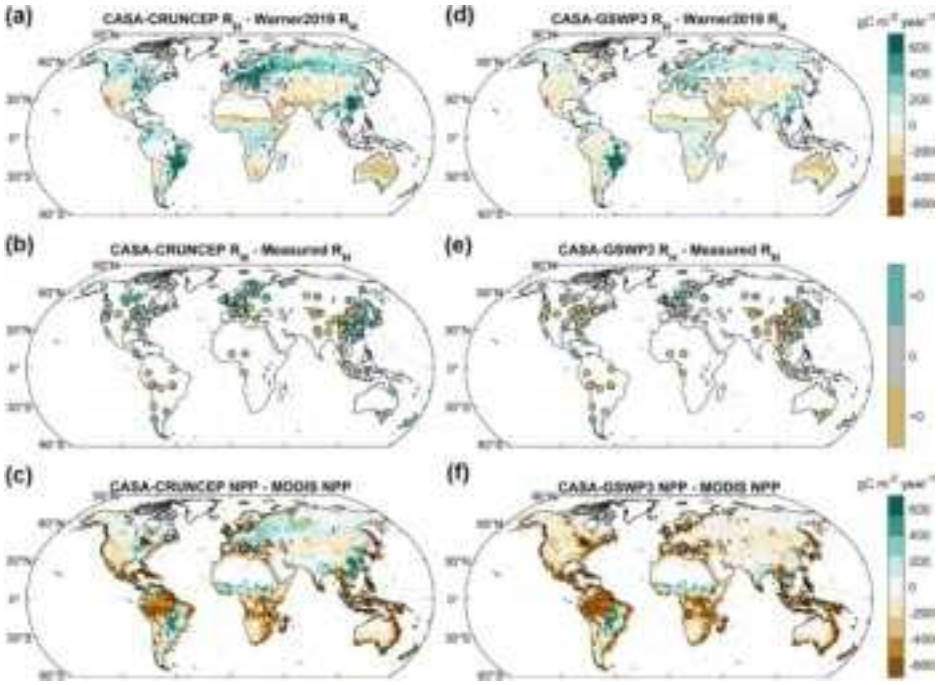
gcb_15795_f1.jpg



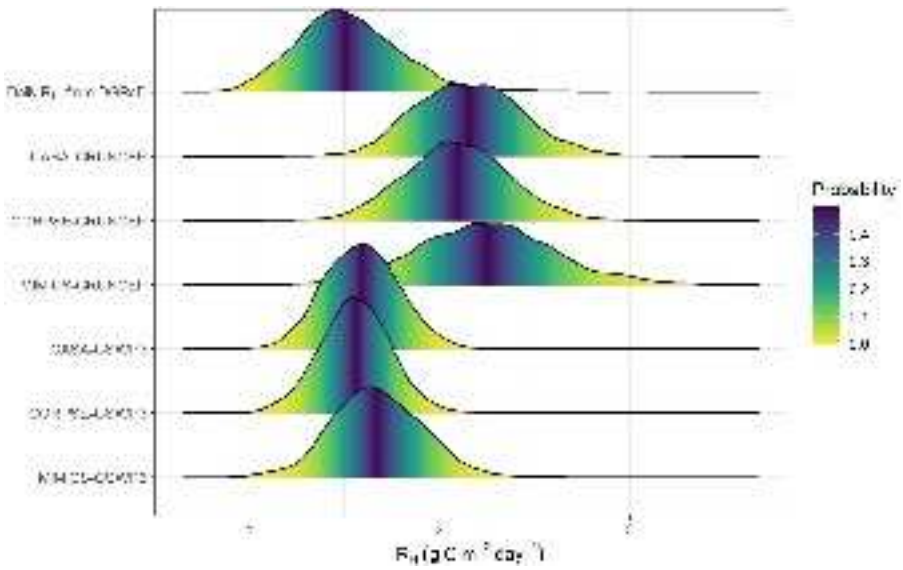
gcb_15795_f2.jpg



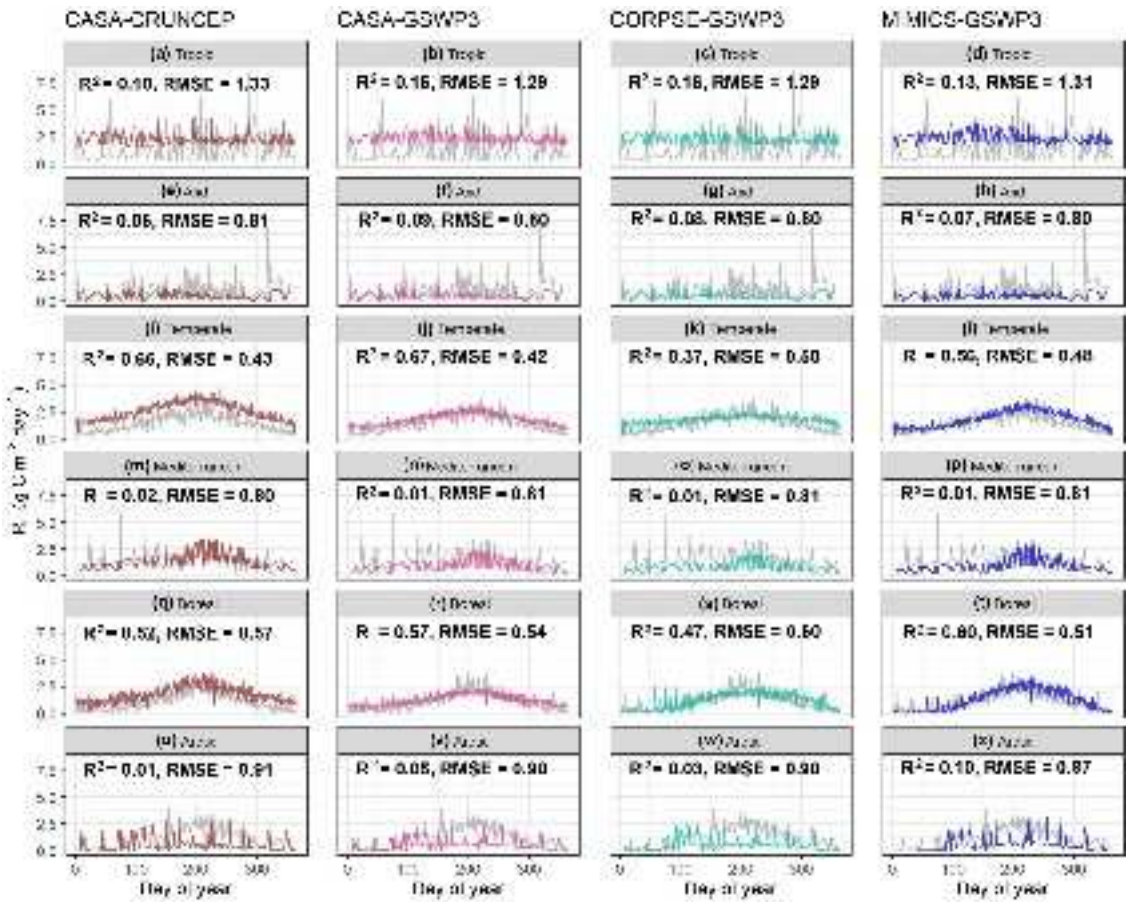
gcb_15795_f3.jpg



gcb_15795_f4.jpg



gcb_15795_f5.jpg



gcb_15795_f6.jpg