https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

Open Access Earth System Science Data Discussions

Spatially explicit re-harmonized terrestrial carbon densities for calibrating Integrated human-Earth System Models

Authors- Kanishka B. Narayan[i], Alan V. Di Vittorio[ii], Evan Margiotta[i], Seth A. Spawn-Lee[iii, iv], Holly K. Gibbs[iii, iv]*

(i) Joint Global Change Research Institute (JGCRI), Pacific Northwest National Lab (PNNL), Washington DC, USA
(ii) Lawrence Berkeley National Lab (LBNL), Berkeley, CA, USA
(iii)Department of Geography, University of Wisconsin-Madison, Madison, WI, USA
(iv) Center for Sustainability and the Global Environment (SAGE), Nelson Institute for Environmental Studies, University of Wisconsin-Madison, Madison, WI, USA

**Correspondence to*: Kanishka B. Narayan (kanishka.narayan@pnnl.gov)

Abstract- Soil and vegetation carbon densities play a critical role in global and regional human-Earth system models. These densities affect variables such as land use change emissions and also influence land use change pathways under climate forcing scenarios where terrestrial carbon is assigned a carbon price. Recently, more spatially explicit, fine resolution data have become available for both soil and vegetation carbon. However, for models to effectively use these data the fine resolution data need to be reharmonized to the initial land use and land cover conditions represented by these models. Without such reharmonization the carbon values may be inaccurate for particular land types and places where the source data and the model disagree on the land use/cover type. Here we present reharmonized soil and vegetation carbon densities both at the 5-arcmin resolution grid cell level and also aggregated to 235 water sheds for 4 land use types and 15 land cover types. These data are particularly useful as initial land carbon conditions for global Multisectoral Dynamic Models (MSD). Moreover, these data include six different statistical states calculated using distinct resampling methods for each of the land use and land cover types. These statistical states are used to define a range of possible carbon values for each land classification, and any state can be used for defining initial conditions of soil and vegetation carbon in MSD models. Users can also estimate any percentile of the carbon distribution defined by these six summary states. We make use of these statistical states to calculate spatially distinct uncertainties in the carbon densities by land type. We have implemented these data in a state-of-the-art multi sector dynamics model, namely the Global Change Analysis Model (GCAM), and show that these new data improve several land use responses in the model, especially when terrestrial carbon is assigned a carbon price. The statistical states in our data are validated against similar estimates in the literature both at a grid cell level and at a regional level.

1. Introduction

Soil and vegetation carbon densities play a critical role in global and regional models such as Earth system models (ESMs), multisector dynamics models (MSDs) and integrated human-Earth system models. These densities influence the predicted productivity of land types (e.g., forest yields, pasture yields, and crop yields) and directly influence land use change emissions. Moreover, these densities affect land use change pathways under climate forcing scenarios implemented in these models (Thomson et al., 2010; Wise et al., 2009). Many models make use of carbon density data that are differentiated by land type but are not spatially explicit. For example, models have previously used estimates of carbon values on undisturbed land from Houghton et al. (Houghton, 1999) and the IPCC (Jackson et al., 2017), among others. Recently,

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

spatially explicit soil carbon density data have been made available by the FAO (Nachtergaele et al., 2010) at a 1 km resolution and at 250 m resolution by the SoilGrids team at the International Soil Reference and Information Centre (ISRIC) (Batjes et al., 2017; Hengl et al., 2014). Similarly, spatially explicit data on vegetation carbon differentiated by above and below ground biomass and spanning several vegetation types have been made available by Spawn et al.(Spawn et al., 2020). Use of spatially distinct, fine resolution data has the potential to significantly improve results from global and regional models by better capturing the geographies of soil and vegetation carbon stocks (Jungkunst et al., 2022). These data can also be used to explore and validate effects of different carbon parameterizations in models (Wieder et al., 2014).

However, these carbon data need to be transformed significantly in order to be used in a robust manner by regional and global models. This is because each of these fine resolution datasets utilizes its own assumptions of land use and land cover which may be distinct from the land use and land cover definitions used by the models in question. For example, many of these fine resolution data use land cover definitions from the European Space Agency Climate Change Initiative (ESA CCI) dataset (Li et al., 2018; Liu et al., 2018) while models may use land use definitions from the Historical Database of the Global Environment (HYDE) dataset (Klein Goldewijk et al., 2017) and/or land cover definitions from the Moderate Resolution Imaging Spectrometer (MODIS)(Barnes et al., 2003; Justice et al., 2002).

Resolution mismatch between data and models provides an additional challenge. The new, spatially distinct carbon densities are available at a very fine resolution (250m / 300 m) while models are often configured to use coarser data that better match their working resolution. For example, consistent land datasets that have frequently been used for climate modelling are available at a resolution of 5 arcmins (i.e. ~ 10km at the equator) (van Asselen & Verburg, 2012), and many regional models operate on land units defined by geopolitical and/or geophysical boundaries. Given the difference in resolutions, and the above-mentioned differences in land classifications, a consistent harmonization method is required to appropriately match the fine resolution carbon data with the appropriate land uses and land cover types within a model.

Here, we prepare and present an aggregated dataset of fine resolution carbon density for soil and vegetation biomass in MgC/ha that has been aligned with the land use and land cover definitions and distribution in the Global Change Analysis Model (GCAM) (Calvin et al., 2019). GCAM represents the interactions between five major systems – energy, water, land, climate, and the economy – at global and regional scales. The soil data are based on the 250 m-resolution SoilGrids dataset and represent a depth of 0-30 cms (Hengl et al., 2014). The aboveground and below ground biomass are based on the 300 m-resolution Spawn et al. dataset (Spawn et al., 2020). The land data integration for GCAM is performed at a resolution of 5 arcmin using a grid-based land data system (Di Vittorio et al., 2020). Hence the final outputs are available as rasters at 5 arcmin resolution. These outputs include rasters corresponding to each of the land use and land cover types in GCAM.

In addition to the reharmonization, we calculated six different data driven statistical states for each carbon pool and each 5 arcmin grid cell using different resampling methods (area weighted

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

average, minimum, maximum, median, Q1, Q3) when re-gridding the data. We also present an easy-to-use tabular output summarizing the six carbon density states for each carbon pool within each of the 235 watersheds intersected with 207 country (ISO) boundaries that are modelled by GCAM. Using our six statistical states, users can calculate any percentile within the distribution of carbon both at a pixel level and at a land region/ water basin level for any land type (For example, the 90th percentile). Such a calculation would not be time intensive given that the six summary states are already available at multiple scales.

GCAM needs to be initialized with densities that represent long term potential maximum carbon values since these values are used to spin-up the model in historical years. In particular, the density values are used to spin up the carbon cycle from 1700-1975. Various studies have found that the long term potential carbon densities are much higher than the contemporary values due to ongoing land use and cover change (Erb et al., 2018; Walker et al., 2022).Moreover, studies have also highlighted difficulties in estimating long term potential carbon densities, since these estimations require long spin up periods themselves (Fang et al., 2014). Here we addressed this issue by deriving a more data driven long term potential carbon state from our new dataset. By analyzing the distribution of carbon values within each land type-watershed combination we found two potential options for initializing GCAM. The first is the Q3 state which would represent a low carbon initialization in 1700 (This state results in 2144 PgC of terrestrial carbon in 1700) and the second, the 90th percentile state which would represent a high carbon initialization (which results in an initial terrestrial carbon stock of 3028 PgC in 1700). We find that utilizing this new carbon dataset for the spin-up improved several responses in GCAM, especially under forcing scenarios where the value of terrestrial carbon is priced using a carbon tax.

We also compared the Q3 and the 90th percentile carbon state in our dataset (which are intended to represent a pre-industrial carbon state) with estimates of potential pre-industrial top-soil carbon by grid cell from Sanderman et al.(Sanderman et al., 2017) and with similar estimates of vegetation carbon from Walker et al. (Walker et al., 2022). We also perform global-level validation of our carbon data, respecting that there is a high degree of uncertainty in carbon estimates from different datasets (Scharlemann et al., 2014; Tifafi et al., 2018). We compare the global estimates of carbon from our reharmonized data with similar estimates of soil and vegetation carbon from other sources in the literature, with different meta-analyses of carbon inventories (Scharlemann et al., 2014) and with modelled estimates of contemporary and historical soil carbon (Sanderman et al., 2017).

The available dataset includes raster files for the six different statistical states for each land use type (Cropland, Urban land, Pasture and Unmanaged land) and each carbon pool, bringing the total to 72 distinct raster files. We also provide a thematic file that labels each cell with the dominant biome for Unmanaged land (out of 15). We also present a tabulated text file with the six carbon state values for each land type and carbon pool aggregated to 699 land regions (235 water basins intersected with 207 country boundaries). Making the data available at these different resolutions should help facilitate effective multiscale modelling of terrestrial carbon. We implemented this carbon reharmonization programmatically in a land data system, *moirai*

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

(Di Vittorio et al., 2020), which can further be used to update the data, validate the data, and reharmonize the data to any other land use and land cover types required by models other than GCAM.

2. Description of data processing

Our carbon data processing method can be organized into three stages:

1. Stage 1- Resampling source datasets based on fine resolution land cover
2. Stage 2- Re-mapping the carbon to Moirai land use and land cover
3. Stage 3- Aggregating raster carbon data to 699 land regions

Figure 1 below summarizes our processing approach from start to finish

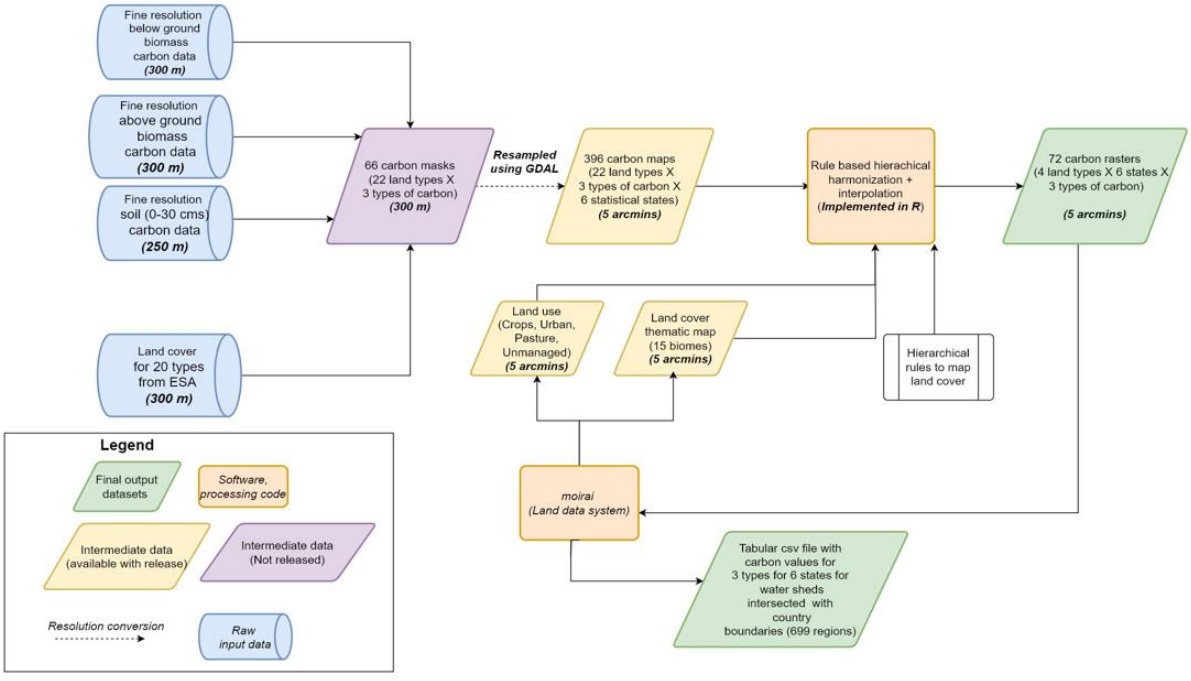


Figure 1: Description of data processing implementation to generate carbon datasets

2.1 Stage 1 – resampling source data

This stage combines the soil and vegetation carbon data (Mg/ha) both at 300 m resolution with the input land cover assumptions from the ESA CCI dataset that correspond to these data. Note that since the land cover dataset from ESA CCI is at a 300m resolution, we resample both our carbon datasets to 300m before this stage.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

Open Access Earth System Science Data Discussions

We first generate land cover masks (1=respective land type present, 0=otherwise) for each of 22 aggregated ESA CCI land cover types (**Table 1**). We combine the land cover masks with the carbon data to create 66 rasters (22 land types X 3 carbon pools), each representing a carbon data mask for an ESA land type. The resulting rasters are calculated as follows:

$$Carbon\_LT\_300m_{pool,j,LT} = Carbon\_300m_{pool,j} * LT\_mask\_300m_{j} \quad (1)$$

Where,

j is the index of a 300m grid cell,

pool is the carbon pool (soil, aboveground biomass, belowground biomass),

LT is the ESA land type.

Next we use six distinct resampling methods to re-grid these data to a 5 arcmin resolution. Each method is applied to each of the land types and thus we derive 6 statistical states for each 5 arcmin grid cell. These aggregated rasters are calculated as follows:

$$Carbon\_LT\_5arcmin_{pool,i,state} = state\begin{pmatrix} Carbon\_LT\_300m_{pool,j} & Carbon\_LT\_300m_{pool,j+2} \\ Carbon\_LT\_300m_{pool,j+1} & Carbon\_LT\_300m_{pool,j+n} \end{pmatrix} \quad (2)$$

Where,

i is the index of a 5 arcmin grid cell,
pool is the carbon pool (soil, aboveground biomass, belowground biomass),
state is the resampling method (weighted average, median, min, max, q1, q3),
j is the index of each 300 m grid cell within aggregated cell i,
n is the total number of 300 m cells that are aggregated into cell i.

Thus, we generate 366 (22 land cover types X 3 types of carbon X 6 states) layers of carbon that correspond to the aggregated ESA CCI land cover types. This processing is largely conducted through the GDAL software (Warmerdam, 2008) and implemented using bash scripts.

ESA_entry code	**ESA_classes**
10;Cropland rainfed;255;255;100	Cropland
11;Herbaceous cover;255;255;100	Unknown_Herb
12;Tree or shrub cover;255;255;0	Unknown_Tree
20;Cropland irrigated or post-flooding;170;240;240	Cropland

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

30;Mosaic cropland (>50%) / natural vegetation (tree shrub herbaceous cover) (<50%);220;240;100	Mosiac_Crop
40;Mosaic natural vegetation (tree shrub herbaceous cover) (>50%) / cropland (<50%) ;200;200;100	Mosaic_tree
50;Tree cover broadleaved evergreen	Broadleaf_Evergreen
60;Tree cover broadleaved deciduous	Broadleaf_Decidous
61;Tree cover broadleaved deciduous	Broadleaf_Decidous
62;Tree cover broadleaved deciduous	Broadleaf_Decidous
70;Tree cover needleleaved evergreen	Needleleaved_Evergreen
71;Tree cover needleleaved evergreen	Needleleaved_Evergreen
72;Tree cover needleleaved evergreen	Needleleaved_Evergreen
80;Tree cover needleleaved deciduous	Needleleaved_decidous
81;Tree cover needleleaved deciduous	Needleleaved_decidous
82;Tree cover needleleaved deciduous	Needleleaved_decidous
90;Tree cover mixed leaf type (broadleaved and needleleaved);120;130;0	Mixed_Forests
100;Mosaic tree and shrub (>50%) / herbaceous cover (<50%);140;160;0	Mosaic_tree
110;Mosaic herbaceous cover (>50%) / tree and shrub (<50%);190;150;0	Mosaic_Herb
120;Shrubland;150;100;0	Shrubland
121;Shrubland evergreen;120;75;0	Shrubland
122;Shrubland deciduous;150;100;0	Shrubland
130;Grassland;255;180;50	Grasslands
140;Lichens and mosses;255;220;210	Grasslands
150;Sparse vegetation (tree shrub herbaceous cover) (<15%);255;235;175	Sparse_Tree
151;Sparse tree (<15%);255;200;100	Sparse_Tree
152;Sparse shrub (<15%);255;210;120	Sparse_Shrub
153;Sparse herbaceous cover (<15%);255;235;175	Sparse_Shrub
160;Tree cover flooded fresh or brakish water;0;120;90	Flood_Tree_Cover
170;Tree cover flooded saline water;0;150;120	Flood_Tree_Cover
180;Shrub or herbaceous cover flooded fresh/saline/brakish water;0;220;130	Flooded_Shrub
190;Urban areas;195;20;0	Urbanland
200;Bare areas;255;245;215	Desert
201;Consolidated bare areas;220;220;220	Desert
202;Unconsolidated bare areas;255;245;215	Desert
220;Permanent snow and ice;255;255;255	Polar_desert_rock_ice

Table 1: Raw ESA codes mapped to ESA land types

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

2.2 Stage 2 – remapping the carbon data to Moirai land use/cover

2.2.1 Reharmonization of ESA land cover with Moirai land cover at 5 arcmins using a prioritization matrix

Next, the 366 layers described above are aligned with the default initial land use/cover for GCAM (2010) at a 5 arcmin resolution. These initial land use/cover data are generated by the Moirai land data system based on land use data from the HYDE (Klein Goldewijk et al., 2017) database and a one-half degree land cover product (Meiyappan & Jain, 2012). Moirai can generate land use and land cover maps for any year based on the these datasets combined with a potential vegetation dataset from Ramankutty et al. (1999). The Moirai land use and land cover types are listed in table 2. It is important to note that carbon values are independently assigned to each of the four Moirai land use types in each cell, and that the unmanaged land use type can be only one of the Moirai land cover types in a given cell. Moirai is described in more detail in Di Vittorio et al.(Di Vittorio et al., 2020).

Land use	**Land cover**
Cropland	Cropland
Pasture	Pasture
Urbanland	Urbanland
Unmanaged	TropicalEvergreenForest/Woodland
	TropicalDeciduousForest/Woodland
	TemperateBroadleafEvergreenForest/Woodland
	TemperateNeedleleafEvergreenForest/Woodland
	TemperateDeciduousForest/Woodland
	BorealEvergreenForest/Woodland
	BorealDeciduousForest/Woodland
	Evergreen/DeciduousMixedForest/Woodland
	Savanna
	Grassland/Steppe
	DenseShrubland
	OpenShrubland
	Tundra
	Desert
	Polardesert/Rock/Ice

Table 2: land use, land cover types for Moirai/GCAM. Total of 4 land use types, 15 types of land cover tracked for Unmanaged land type

The carbon for each Moirai land type in a cell needs to be selected from the 366 rasters generated in Stage 1 described above. We use a rule-based harmonization approach where we select the

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

appropriate carbon values by matching the Moirai land type with the corresponding ESA land cover type (Table 3). We assign 6 possible ESA land cover types to each Moirai land type and rank them according to their similarity with the Moirai land type. This means that carbon values for a particular moirai land type can come from any of six ESA land cover types, as long as they are present in a given cell. For example, a Tropical Evergreen Forest cell in Moirai, may be assigned carbon values from the Evergreen_Combined, Mixed_Forests, Mosaic_Tree, Flood_Tree_Cover, Unknown_Tree_Cover, or Sparse_Treecover ESA land cover types. The similarity ranking both maximizes the number of Moirai land type assignments and ensures that the most appropriate carbon values are selected. The first ESA land cover in the ranked list that is present in a given cell provides the carbon values for the corresponding Moirai land type in the same cell (Table 3). In the example above, The Evergreen_Combined carbon data would be chosen first over all other ESA land covers if it existed in a given cell and the Sparse_Treecover carbon data would be chosen if it were the only ESA land cover from the list that existed in a given cell. These prioritization rules are designed such that carbon data from one biome is not assigned to a different biome when reharmonizing and re-gridding the carbon. The ESA land cover selection is done once for each cell and Moirai land type, and then the data from the corresponding carbon pool and state rasters are assigned to the Moirai land type in the target cell. This results in 72 rasters that become input files for Moirai.

We used expert judgement when developing the matrix so as to best represent the Moirai land types when selecting from the ESA land types. For certain land types we constrain the choices by allowing less than six choices. For example, carbon for a Moirai Desert cell can only be chosen from a corresponding desert cell in the ESA masks. On the other hand, Moirai Tundra includes eight ESA land covers because ESA does not have an explicit Tundra class. The increased number of options aims to provide adequate data coverage for Tundra[1]. Furthermore, certain biome types that are not modelled by GCAM or represented explicitly in Moirai receive low priority rankings. For example, Flooded land types are never included as a first priority choice for any land type since Moirai does not explicitly include flooded land types. Conversely, the ESA land cover data do not include any explicit representation of pastures or rangeland. Our rules assign pasture carbon values based on proximate grassland or shrubland carbon values. Grasslands in particular are prioritized for Pasture carbon selection because the pasture definition in GCAM corresponds to grasslands used for grazing.

[1] Tundra data selection prioritizes polar desert rock ice pixels. The location of these pixels coincides with the Tundra land cover and they also represent pixels with high values for soil carbon densities.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

	Corresponding ESA LAND COVER prioritized from Primary to 8th							
moirai land type name	**Primary**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
TropicalEvergreenForest/Woodland	Evergreen_Combined	Mixed_Forests	Mosaic_Tree	Flood_Tree_Cover	Unknown_Tree	Sparse_Tree	-	-
TropicalDeciduousForest/Woodland	Decidous_Combined	Mixed_Forests	Mosaic_Tree	Flood_Tree_Cover	Unknown_Tree	Sparse_Tree	-	-
TemperateBroadleafEvergreenForest/Woodland	Broadleaf_Evergreen	Mixed_Forests	Mosaic_Tree	Flood_Tree_Cover	Unknown_Tree	Sparse_Tree	-	-
TemperateNeedleleafEvergreenForest/Woodland	Needleleaved_Evergreen	Mixed_Forests	Mosaic_Tree	Flood_Tree_Cover	Unknown_Tree	Sparse_Tree	-	-
TemperateDeciduousForest/Woodland	Decidous_Combined	Mixed_Forests	Mosaic_Tree	Flood_Tree_Cover	Unknown_Tree	Sparse_Tree	-	-
BorealEvergreenForest/Woodland	Evergreen_Combined	Mixed_Forests	Mosaic_Tree	Flood_Tree_Cover	Unknown_Tree	Sparse_Tree	-	-
BorealDeciduousForest/Woodland	Combined_Decidous	Mixed_Forests	Mosaic_Tree	Flood_Tree_Cover	Unknown_Tree	Sparse_Tree	-	-
Evergreen/DeciduousMixedForest/Woodland	Mixed_Forests	Mosaic_Tree	Flood_Tree_Cover	Unknown_Tree	Sparse_Tree	Sparse_Tree	-	-
Savanna	Mosaic_Herb	Grasslands	Unknown_Herb	Flood_Shrub	-	-	-	-
Grassland/Steppe	Grasslands	Unknown_Herb	Mosaic_Herb	Flood_Shrub	-	-	-	-
DenseShrubland	Shrubland	Unknown_Tree	Flooded_Shrub	Mosaic_Herb	-	-	-	-
OpenShrubland	Sparse_Shrub	Mosaic_Herb	Flooded_Shrub	Unknown_Herb	-	-	-	-
Tundra	Polar_Desert_Rock_Ice	Sparse_Shrub	Mosaic_Herb	Unknown_Herb	Unknown_Tree	Sparse_Tree	Shrubland	Mosaic_Tree
Desert	Desert	-	-	-	-	-	-	-
Polardesert/Rock/Ice	Polar_desert_rock_ice	-	-	-	-	-	-	-
Cropland	Cropland	Mosaic_Cropland	-	-	-	-	-	-
Pasture	Grasslands	Mosaic_Herb	Unknown_Herb	Sparse_Tree	Sparse_Shrub	-	-	-
Urbanland	Urbanland	-	-	-	-	-	-	-

Table 3: Prioritization matrix to match ESA land cover with moirai land types

2.2.2 Implementation of nearest neighbor algorithm to increase data coverage

After implementing the prioritization rules there remain 5 arcmin cells with no carbon data coverage for a given land type and carbon pool. This is expected since the land cover data used to generate the carbon masks (ESA CCI land cover data) may be different from the land cover data used in HYDE, SAGE. We therefore implement a nearest neighbor algorithm to interpolate data to each 'no data' cell based on availability in 40 neighboring cells. This algorithm fills the target cell and land type with the corresponding carbon data of the closest cell with matching land type. If no matches are found within the prescribed window then the target cell remains without data for that particular carbon pool and that particular land type.

Carbon data coverage after interpolation is reasonable with the exception of a few land types. Table 4 shows the data coverage by land type after implementation of the nearest neighbor algorithm. All but three land types have over 80% data coverage for soil and vegetation carbon.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

At least 25% of Tundra and Polar desert cells remain without carbon data. This is likely a result of differences in way Tundra land cover is defined by different datasets.

There have been more recent efforts to collect soil carbon data specifically for the permafrost and Tundra regions such as that by Hugelius et al.(Hugelius et al., 2014). This suggests that a future area of work would be to incorporate these more detailed datasets into either the source data or our processing workflow. Along with Tundra and Polar deserts, over 20% of the Urban land cells do not have carbon data. This is once again likely due to the different definitions of Urban land cover indifferent datasets. Our data coverage suggests that there exists more uncertainty in the Tundra, Polar, and Urban carbon values purely based on limited data availability. Recognizing and quantifying data availability by land type enables users to utilize their own judgement when using the carbon values for these land types.

Land type	Total 5arcmin grid cells	Vegetation carbon Percentage unfound (NO DATA cells after interpolation)	Soil carbon Percentage unfound (NO DATA cells after interpolation)
Pasture	1195396	2.3	2.3
Cropland	952850	17.0	17.0
Grassland/Steppe	498404	15.0	14.6
OpenShrubland	274296	16.0	16.0
Desert	195579	1.0	1.1
TropicalEvergreenForest/Woodland	190780	0.0	0.3
Savanna	173776	8.0	7.6
BorealEvergreenForest/Woodland	148756	0.0	0.0
Polardesert/rock/ice	132021	29.0	24.9
Urban	119597	22.3	22.3
TemperateDeciduousForest/Woodland	86922	1.0	1.1
DenseShrubland	78065	10.0	9.5
TemperateNeedleleafEvergreenForest/Woodland	71600	1.0	0.5
BorealDeciduousForest/Woodland	65824	0.0	0.4

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

TropicalDeciduousForest/Woodland	56377	1.0	1.4
Tundra	25000	29.0	24.9
TemperateBroadleafEvergreenForest/Woodland	14395	0.0	0.3

Table 4: Details of NODATA cells after nearest neighbor interpolation

2.3 Stage 3 - Aggregating raster carbon data to 699 land regions

As a final step, we pass the 72 rasters generated in Stage 2 and use those as inputs to the Moirai land data system. The land data system uses the inputs to aggregate the values to 699 land regions from the 5 arcmin grid cell level. The 699 land regions are the intersection of 235 water basins and 207 countries and are shown as a map in **SI figure 1**. The final carbon state values for each land type are aggregated to each land region for each carbon pool (aboveground biomass, belowground biomass, soil 0-30 cms). These outputs are available as a tabular text file. The moirai land data system performs this aggregation using the same land masks for the year 2010 which are used in the Stage 2 processing. The basic aggregation performed by moirai is summarized in equation 3 below

$$Carbon\_tabular_{pool,GLU,state,LT} = state\begin{pmatrix} Carbon\_5arcmin\_LT_{pool,j} & Carbon\_5arcmin\_LT_{pool,j+2} \\ Carbon\_5arcmin\_LT_{pool,j+1} & Carbon\_5arcmin\_LT_{pool,j+n} \end{pmatrix} \quad (3)$$

Where,
pool is the carbon pool (aboveground biomass, belowground biomass, topsoil (0-30 cms)),
state is the aggregation method (area-weighted average, median, min, max, q1, q3),
GLU represents a land region which is an intersection of 207 country boundaries and 235 watershed boundaries,
j is the grid cell index for each 5 arcmin grid cell in a basin with land type LT,
n is the total number of cells in a basin for a given land type,
and LT is the land type.

2.4 Stage 4 – Deriving any other percentile using our six statistical states

Using our six summary states, users can calculate any percentile for the carbon value in any pixel for each of our 19 land types and three carbon pools (soil, above ground biomass, below ground biomass). These values can also be calculated directly for a land region/water basin. The percentile values can be calculated assuming that carbon values are lognormally distributed (this is established in our analysis below- See section 3.1) The steps to calculate any percentile are as follows,

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

1. Compute a mean value as a natural log of the median state. Since the distribution of carbon values is lognormal, the natural log of our median would be an estimated mean for the lognormal distribution.
2. Compute an estimated standard deviation using a natural log of the Q3 and the mean value in step 1, specifically we use the formula- (LN(Q3)- LN(mean))/0.675.
3. Estimate the percentile value from the mean and standard deviation above. Since the logged distribution is normal, users can compute this value using a z table for a normal distribution.
4. Calculate the exponent of the value in step 3.
5. Constrain this value to the max observed value in our dataset.

This method would enable a timely calculation of percentiles and would be much faster than re-running the code to derive individual percentiles using re-sampling.

3. Analysis, Uncertainty and data validation

We first evaluate our main data products, namely the maps of soil and vegetation carbon across gridcells by land type (e.g., Figure 2 and Figure 3), with the goal of identifying the most appropriate carbon state for GCAM modeling, and then take a closer look at data uncertainty and spatial variability. Note that the authors of the source data on soil (Hengl et al.) and vegetation (Spawn et al.) did a detailed spatial validation of the data in their resperctive papers. Our validation will focus on uncertainties that have been introduced through our re-harmonization process. We will also compare our Q3 and 90th percentile (determined as described above) estimates with similar estimates from the literature since these estimates will be used to initialize GCAM.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

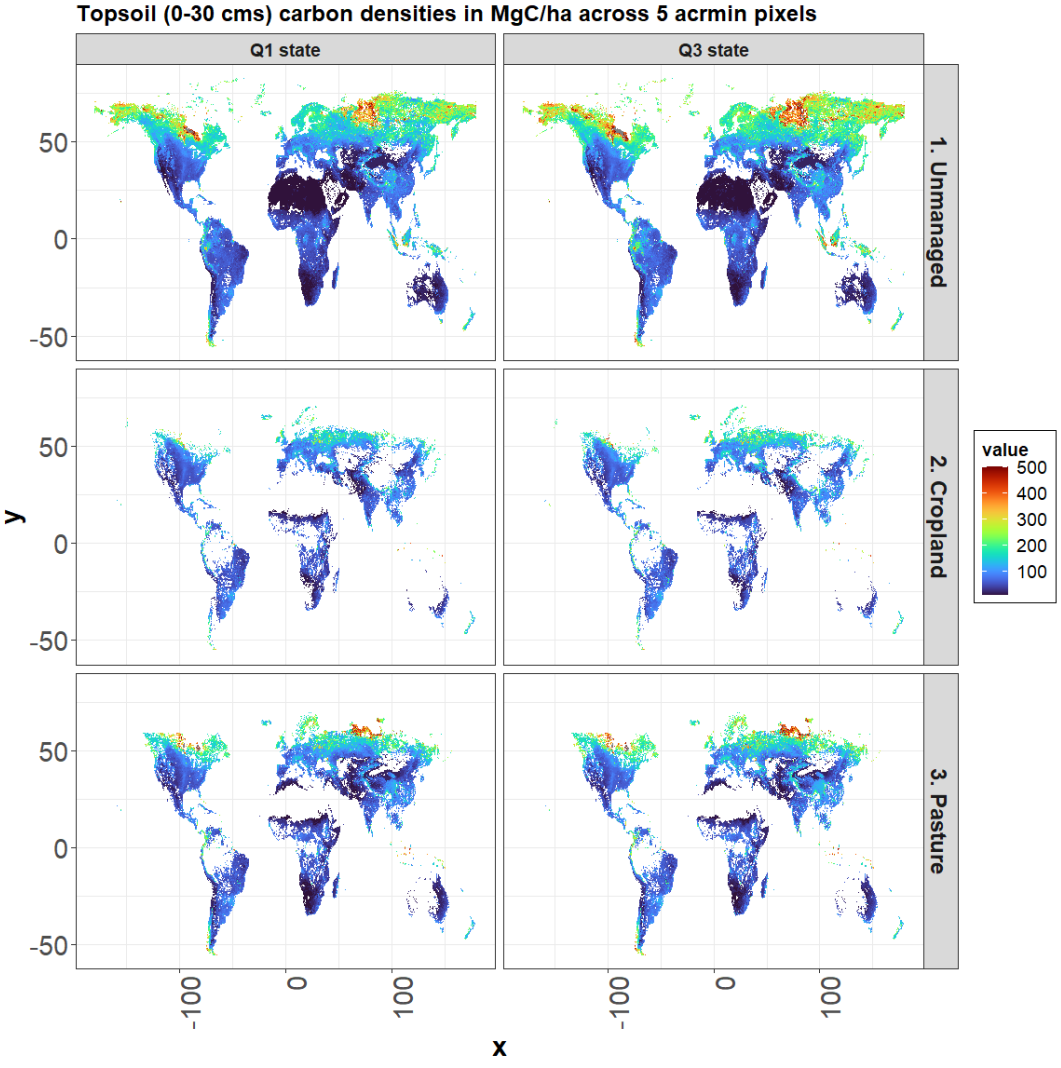


Figure 2: Soil carbon for each 5 arcmin grid cell in MgC/ha. Values are shown for two statistical states, namely the Q1 and the Q3.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

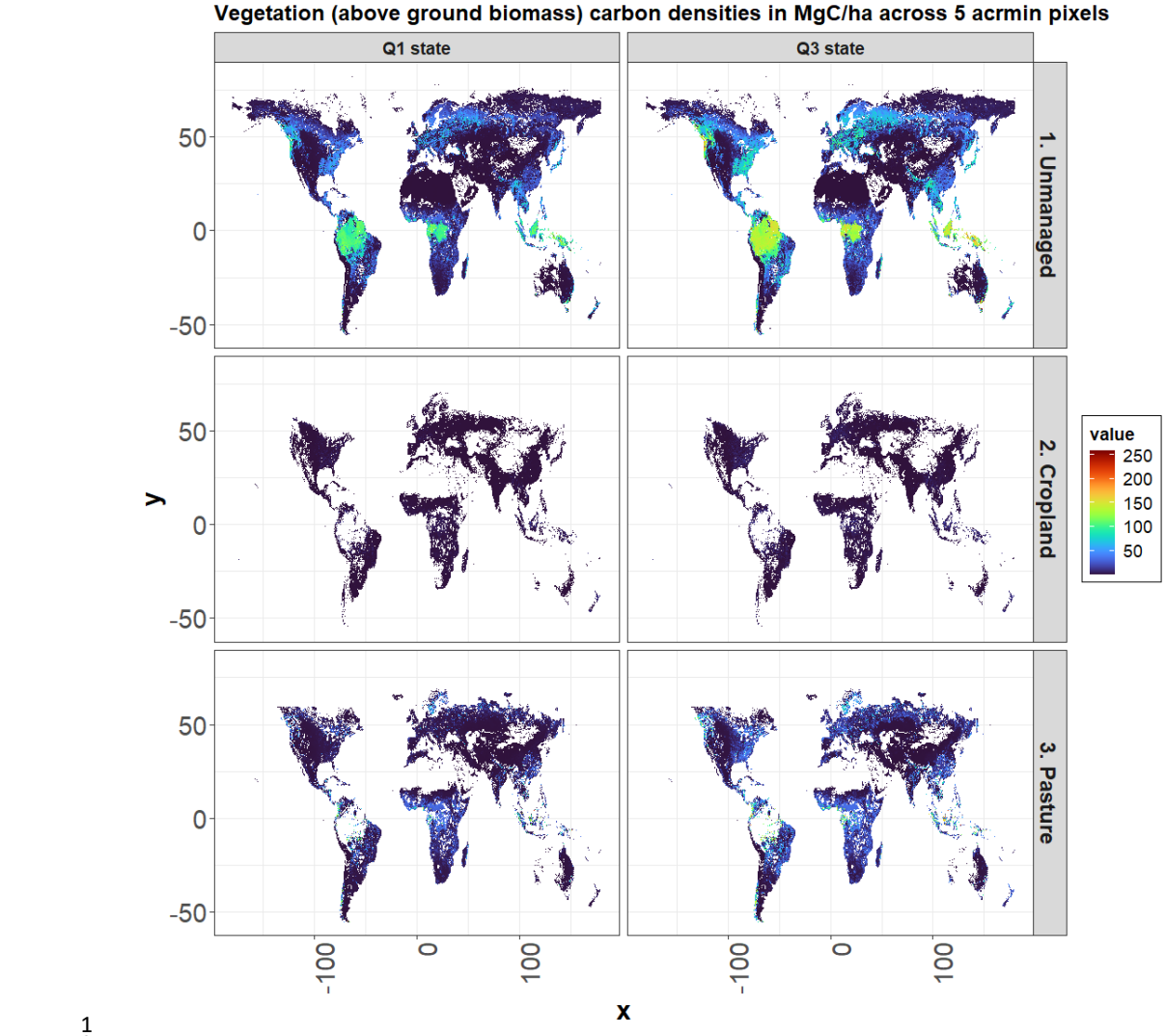


Figure 3 Veg carbon (aboveground) in MgC /ha across 5 arcmin grid cells for aggregate land types for the Q3 state. Values are shown for two statistical states, namely the Q1 and the Q3

3A Comparison of harmonized carbon values to estimates of historical values

As mentioned above, this dataset was generated to initialize GCAM with spatially explicit carbon values for spin up and further simulation of the land system. This initialization requires the carbon values to represent a maximum potential carbon density because these values determine the limiting parameters for vegetation growth and soil carbon accumulation curves. The pre industrial carbon density has been estimated to be much higher than the contemporary carbon stored in land (Erb et al., 2018) due to a long

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

history of land use. Moreover, various studies have highlighted the difficulties with the calculation of the long term potential maximum(Fang et al., 2014). However, our development of the statistical states has allowed us to adopt a more systematic approach to selecting a data driven maximum value that we use to initialize GCAM.

To select a data driven pre industrial equilibrium state, we compared the frequency distributions of carbon by pool within each land region for each land type with the final statistical states calculated. The frequency distributions represent a heterogeneous landscape at different stages of growth and management. The average or median values may be representative of the contemporary landscape, but not of an undisturbed landscape that has been allowed to equilibrate its carbon stocks. The maximum value in a land region may be an extreme outlier and likewise would not be representative of the undisturbed landscape. Our goal then is to find a value in between the contempory average and the maximum that is representative of a long term potential maximum value. Fortunately, most distributions of soil carbon generally follow a log-normal shape with a long tail. For example, we analyzed the distribution of soil carbon in the Amazon basin (Figure 4) for different land types.

One possible option for initialization is the Q3 statistical state. The Q3 statistical state value from these distributions does fall between the average and the maximum, as expected. Given the lognormal shape, the observations above the Q3 value are infrequent and can stretch to extremely high values. We also find that most vegetation carbon distributions follow a log-normal shape within each basin for each land type. However, forests have distributions that are more bimodal (Figure 5). Nonetheless, in these distributions the Q3 value provides an estimate of carbon that is reasonably higher than the contemporary average or median value. Using the Q3 values to initialize GCAM, we found that the initial carbon stock in the year 1700 would be approx 2144 PgC (1553 PgC of carbon in top soil and 591 PgC of vegetation biomass). This estimate is still on the lower end of other similar estimates from Walker et al. ,Erb et al, Houghton and Sanderman et al. Table 5 below summarizes our initial terrestrial carbon stock in 1700 calculated from different sources-

Data source	Topsoil (0-30 cms) carbon in PgC	vegetation (above+ below ground biomass) in PgC
Erb et al 2019		916
moirai (Q3)	1553	591.7
moirai (90th percentile)	2063	966
Walker 2022		715
Sanderman et al. 2017	2119	
Houghton 1999	1462	662

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

Table 5:Initial potential terrestrial carbon stock calculated from different sources. Sources from moirai are calculated using land maps in 1700. Sanderman et al. represents a carbon stock in 1800 given no land use. Walker and Erb et al. are based on potential vegetation carbon estimations.

In addition to the Q3 value, we also use the estimated 90th percentile state in order to represent a higher initialization of carbon in 1700. This 90th percentile is estimated from our six summary states using the methodology outlined in section 2.4. This 90th percentile provides an initial carbon stock of 3028 PgC (2063 PgC of carbon from topsoil and 966 PgC of vegetation biomass). Using these two states for initialization helps us understand the sensitivity of the model to the initial value.

In the case of forests, we note that we derive carbon values for Forests as a whole and do not differentiate between Primary Forests and Secondary Forests. This is a result of lack of availability of fine resolution land masks that differentiate between primary and secondary forests. This likely means that our long term potential maximum forest carbon densities include the impact of harvesting especially in regions with high levels of forest harvests. As more fine resolution data on different types of forests become available, a logical next step would be to derive separate carbon densities for this particular land type.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

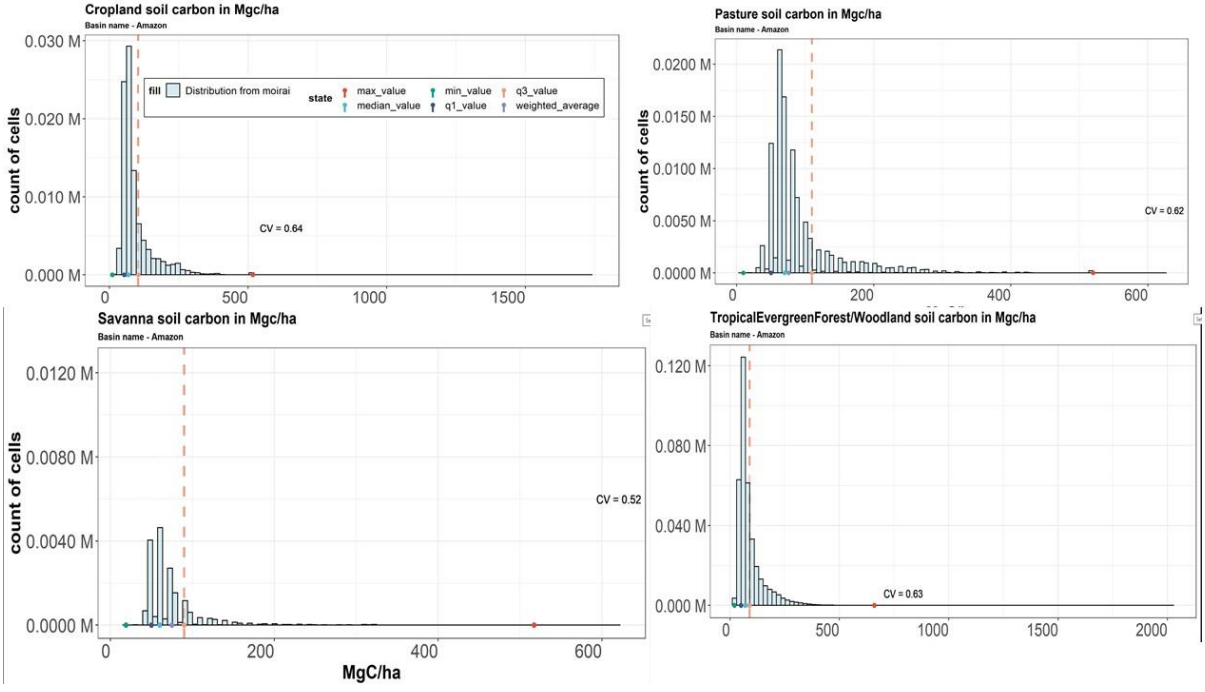
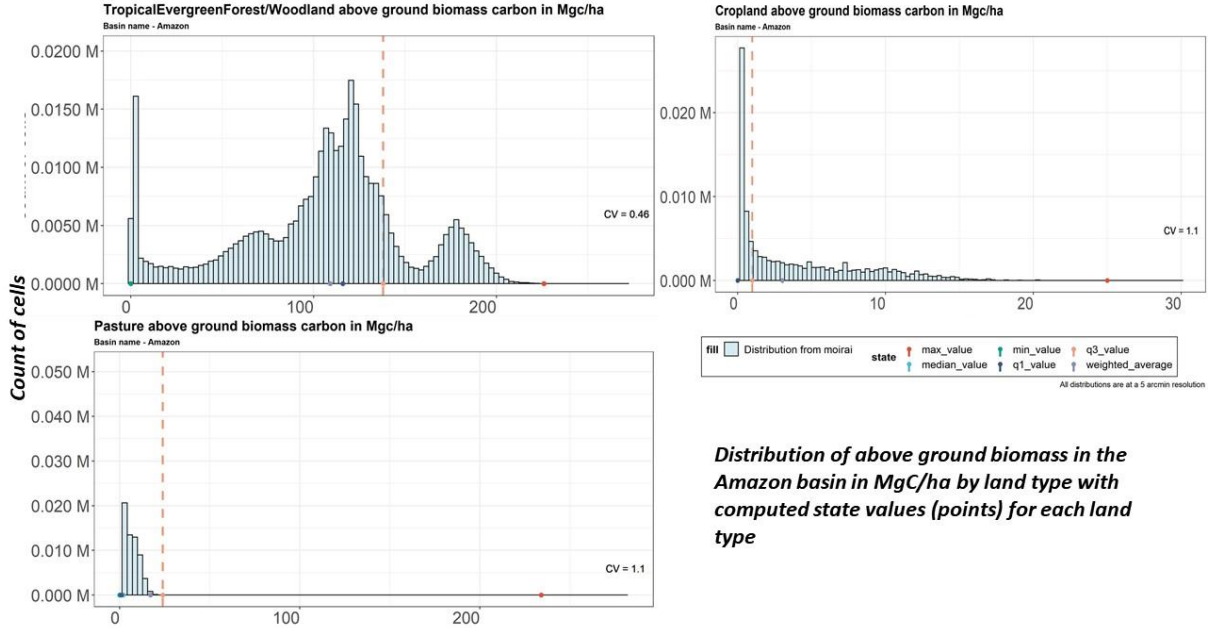


Figure 4: Within basin distributions of soil carbon in Mgc/ha for the Amazon basin. Each facet shows a distribution for a land type. The final basin level statistical states are shown as dots with the Q3 state shown as the orange line.



Distribution of above ground biomass in the Amazon basin in MgC/ha by land type with computed state values (points) for each land type

Figure 5: Within basin distributions of aboveground biomass carbon in Mgc/ha for the Amazon basin. Each facet shows a distribution for a land type. The final basin level statistical states are shown as dots with the Q3 state shown as the orange line.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

Open Access Earth System Science Data Discussions

3B Comparison of carbon values to other estimates of long term potential carbon by grid cell

Since the Q3 and 90th percentile values in our dataset will be used to represent pre-industrial carbon densities in GCAM, we compared our values to similar estimates in the literature. Specifically we compare the 90th percentile values at the pixel level with other estimates, since this statistical state represents a high carbon initialization compared to the Q3 and produces a global terrestrial carbon stock that is in line with other estimates of potential terrestrial carbon.

Sanderman et al. 2017 generated a pre-industrial soil carbon map for top soil in the year 1800. This map assumed no land use in that year. Similarly Walker et al. (Walker et al., 2022) generated a similar map for potential carbon in above and below ground vegetation. For a valid comparison we compared only our unmanaged land carbon values with these estimates (Figure 6 and 7).

We found that in the case of soil carbon, even though our maps track well with the maps from Sanderman et al. in terms of the overall spatial distribution, the mean error (moirai 90th percentile – Sanderman et al.) across gridcells that is close to -23%. There are some upper latitude pixels from the Sanderman et al. dataset that show almost 100% higher values compared to our data.

In case of aboveground vegetation carbon, the mean percen error (moirai 90th percentile – Walker et al. 2022) is -17%, which is lower than for soil carbon. The largest errors were observed for forest pixels. This is likely due to the combination of Primary and Secondary forests into a single forest category in our dataset (as described above). The highest differences between datasets are observed in forest pixels with high level of forest harvesting (Central and West Africa and South and East Asia).

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

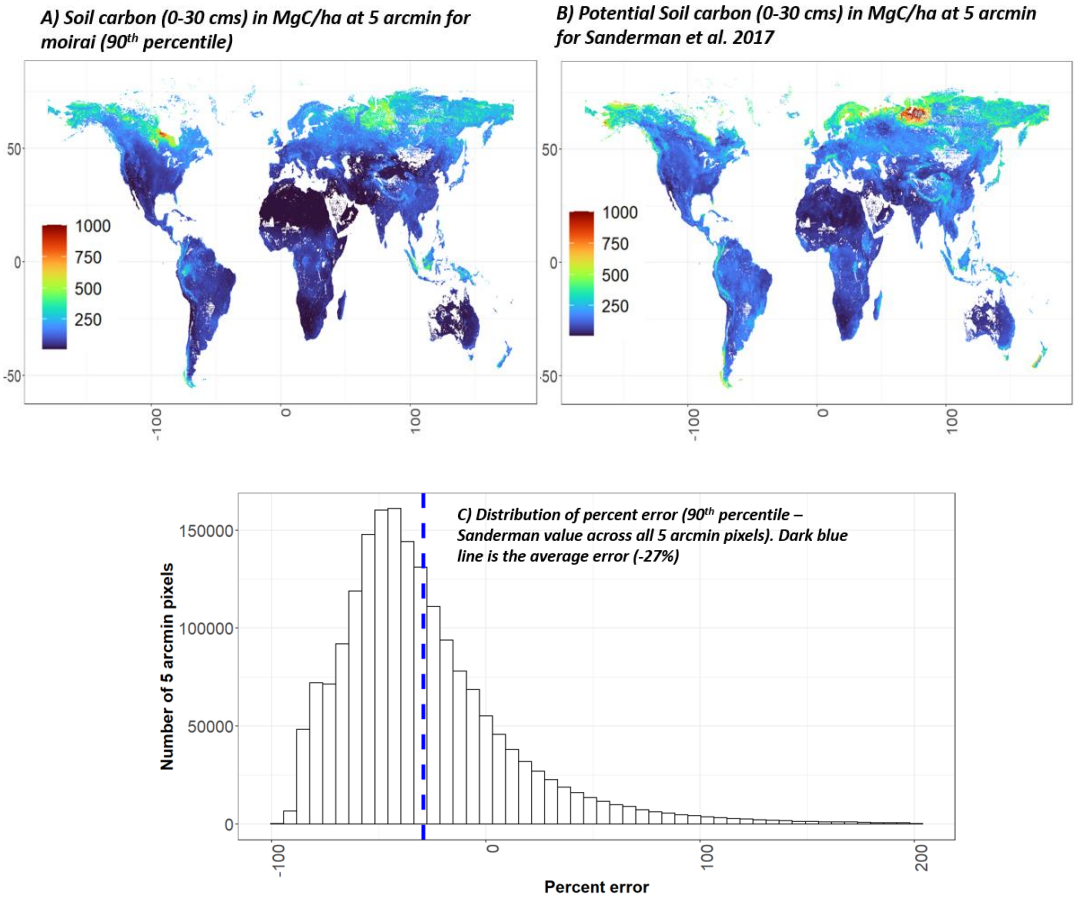


Figure 6: A.) Topsoil (0-30 cms) carbon in MgC/ha for 5 arcmin pixels using moirai 90th percentile B.) Top soil (0-30cms) carbon in MgC/ha from Sanderman et al. assuming a no land use condition. C.) Histogram showing percent error between A and B. Dark blue dashed line represents mean error across all pixels which is at -27%.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

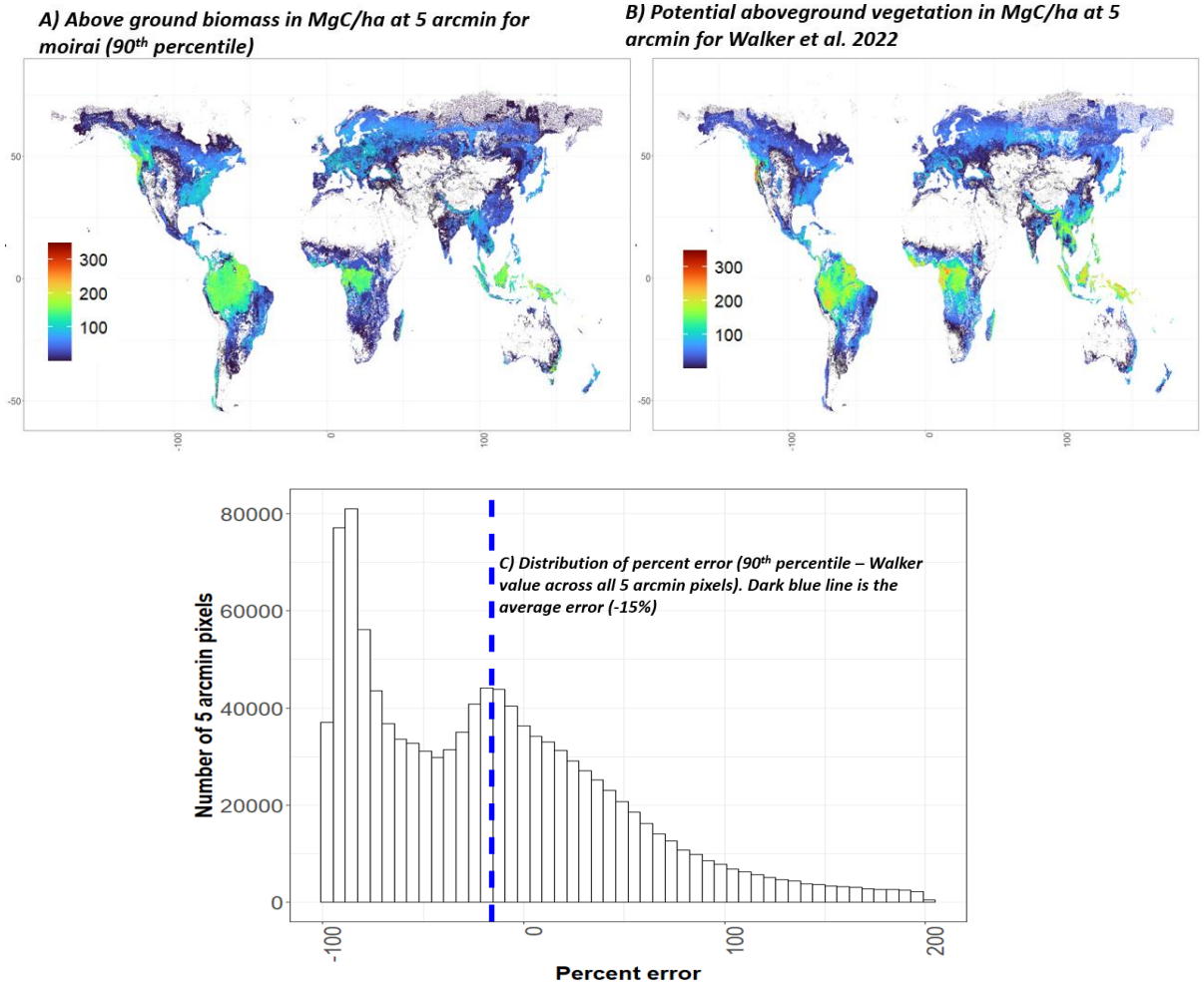


Figure 7: A.) Vegetation (aboveground) carbon in MgC/ha for 5 arcmin pixels using moirai 90th percentile B.) Vegetation (aboveground) carbon in MgC/ha from Walker et al. constrained for initial land use C.) Histogram showing percent error between A and B. Dark blue dashed line represents mean error across all pixels which is at -15%.

3C Comparison to C values to previously used in GCAM by land type and aggregate contemporary estimates

We compared the moirai densities by land type globally with similar carbon densities from Houghton (1999) (See **SI Table 2** and Figure 8 for a comparison of soil carbon estimates and **SI Table 3** and Figure 9 for aboveground biomass comparison). The Houghton carbon densities represent carbon values on undisturbed land differentiated by biome. We specifically compare against the Houghton carbon densities since those values were previously used for the spin up in GCAM. We also compared our statistical states with contemporary values where available (e.g. Jackson et al. for soil carbon and Vlek et al. for contemporary vegetation carbon).

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

Open Access Earth System Science Data Discussions

For soil carbon (Figure 8), we found that our values (Q3, 90th percentile) are generally higher than the Houghton values globally for most land types. The values are especially higher for Shrublands which are located in Boreal regions where the difference is approx 80 Mgc/ha. This is likely because the SoilGrids dataset shows high carbon values at high latitudes and includes peat soils in its estimates (e.g., figure 3). The high values of soil carbon at the upper latitudes may also be driven by high levels of predicted bulk density at those locations(Tifafi et al., 2018).Another version of soil grids has recently produced lower values in these regions to reduce the effects of peat soil estimates(Poggio et al., 2021).

For cropland, our Q3 estimates of carbon are as high as forest soil carbon. This is investigated in more detail in the sections below. Similarly the soil carbon under Urban land cover is extremely high. This is likely due to how the samples were collected for Urban land cover (these samples are collected in parks as opposed to built up areas). As expected, the values in our range are higher than the contemporary values from Jackson et al., especially in the Boreal regions. However, the Q1 values from our range are closest to contemporary values for soils.

For vegetation carbon (Figure 9), in the case of forests, the carbon densities are significantly scaled down across moirai states when compared to the literature (Houghton for the pre-industrial values and Vlek et al. for contemporary values). This is not surprising since in the Houghton inventory numbers, the spatial distribution of forest carbon is an unknown especially for tropical forests (Houghton, 2005). Also, as noted above our moirai values for forest carbon densities are a combination of Primary and Secondary Forests and are therefore underestimate the long term potential maximum carbon that can be stored in forests.

For grasslands and pastures, the moirai Q3 and 90th percentile estimate is higher than the literature values. However, for this land type, the overall distribution of carbon values is not very dispersed across basins. For forests, there is significant variation in values across basins , likely since some forests may be more intensively managed or harvested compared to others. Note that the reduction in carbon values compared to the global values from Houghton will likely reduce the afforestation response in GCAM under scenarios where carbon in forests is priced.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

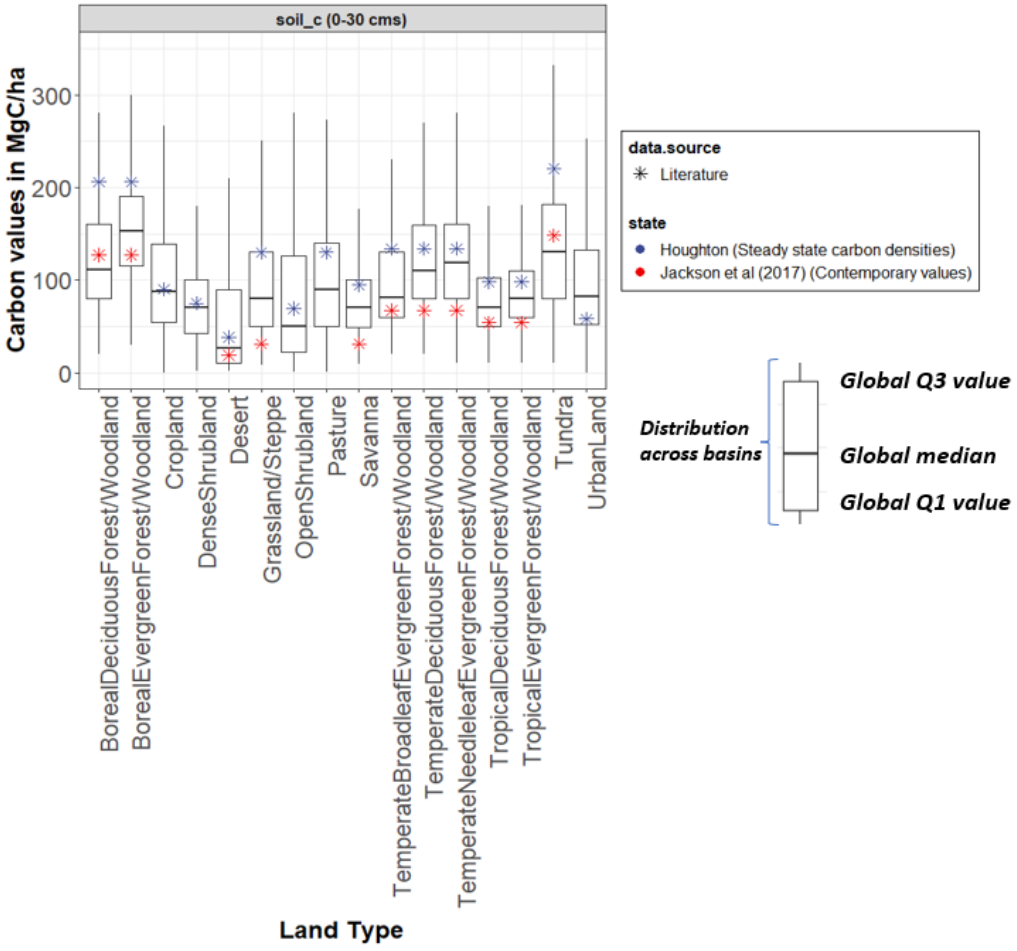


Figure 8: Carbon densities globally by land type across basins and Houghton pre-industrial state. Also shown is the contemporary carbon density value by land type from Jackson et al.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

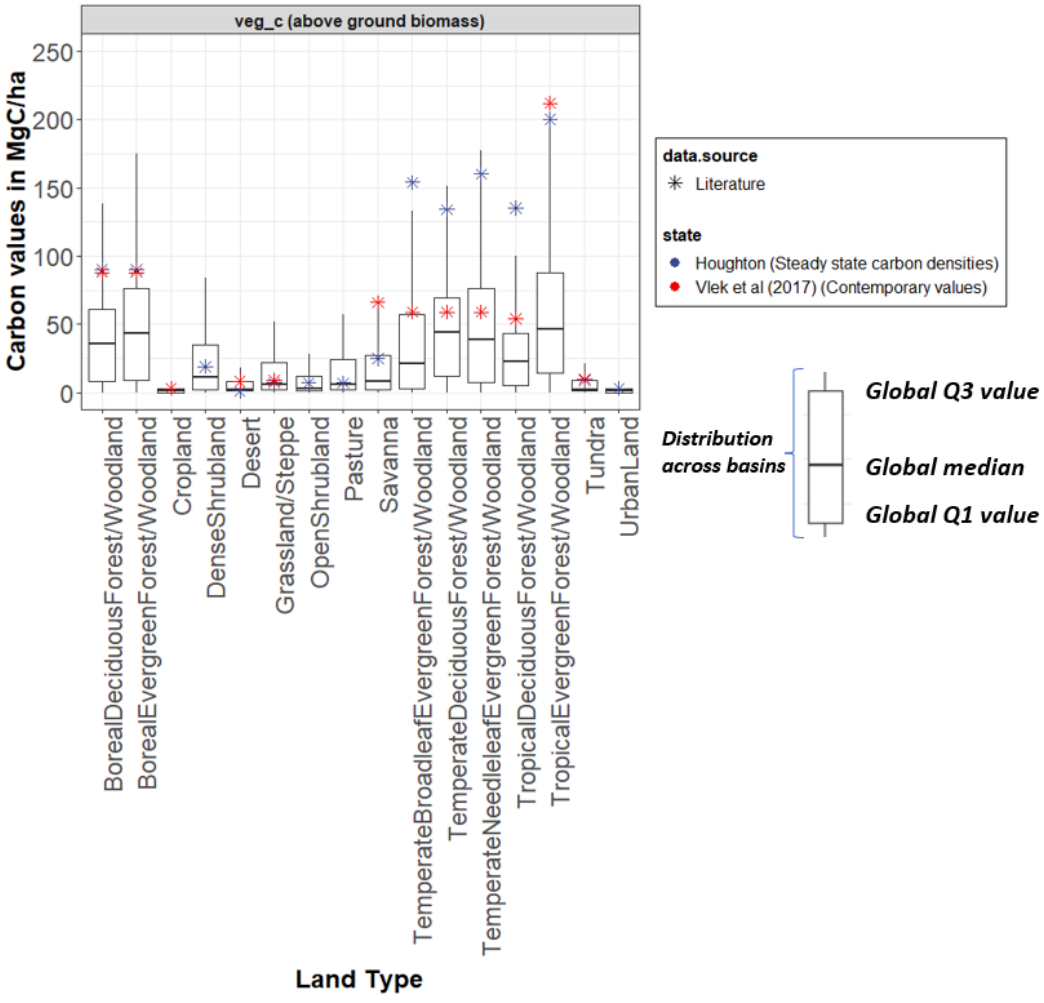


Figure 9: Carbon densities globally by land type for vegetation carbon across basins and Houghton values (Undisturbed carbon in 1850). Also shown is the contemporary carbon density value by land type from Vlek et al.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

3.1 Uncertainties in re-harmonized carbon data (spatially and across land types)

Here we explore uncertainty in the available data by further examing spatial distributions, aggregation statistics, and land type considerations.

i.) Do managed land types show a deprecation in carbon compared to unmanaged land types?

Studies show that managed land (i.e., Cropland, Pasture, Urbanland) has depleted carbon stocks in relation to undisturbed land (Cooper et al., 2021; Sanderman et al., 2017; Wei et al., 2014). The aim of processing the spatial managed land carbon data and adding it to Moirai is to obtain contemporary estimates for these lands that can be used in modeling rather than assuming a global value or that managed lands have a fixed fraction of unmanaged land carbon. Carbon data values do not correspond with a long term potential maximum for these managed land types by definition, as these land are actively disturbed. However, we still want higher than average carbon values for the parameters that define the limits of carbon accumulation for these land types. We expect that the carbon data reflect the effects of these managed land types and that our desired values would be lower than those for the surrounding unmanaged land types. We checked this expectation by first comparing Q3 carbon values for soil and aboveground biomass for Cropland with the corresponding values for Unmanaged land cover in each of our land regions (Figure 10). We found that Cropland soil carbon values do not show a consistent depletion for soil carbon compared to Unmanaged land. The reason for these differences among carbon pools is rooted in the source data sampling and processing methodologies. In case of the soilgrids dataset, the authors state that cropland soil carbon samples were largely collected in the US. In case of the vegetation carbon dataset from Spawn et al., the vegetation carbon was calculated for each crop type based on yields which explains the low values on cropland compared to unmanaged land.

For cropland, yields are determined from harvested area and production data, while the carbon data are used for land use emissions and when valuing carbon in forcing scenarios. To address the relatively high cropland soil carbon data in our modeling experiments we reduce these data by 30% before using them in GCAM. Previous studies have a found a similar loss of soil carbon through agricultural practices and land conver conversion from unmanaged land types to cropland(Cooper et al., 2021; Wei et al., 2014).

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

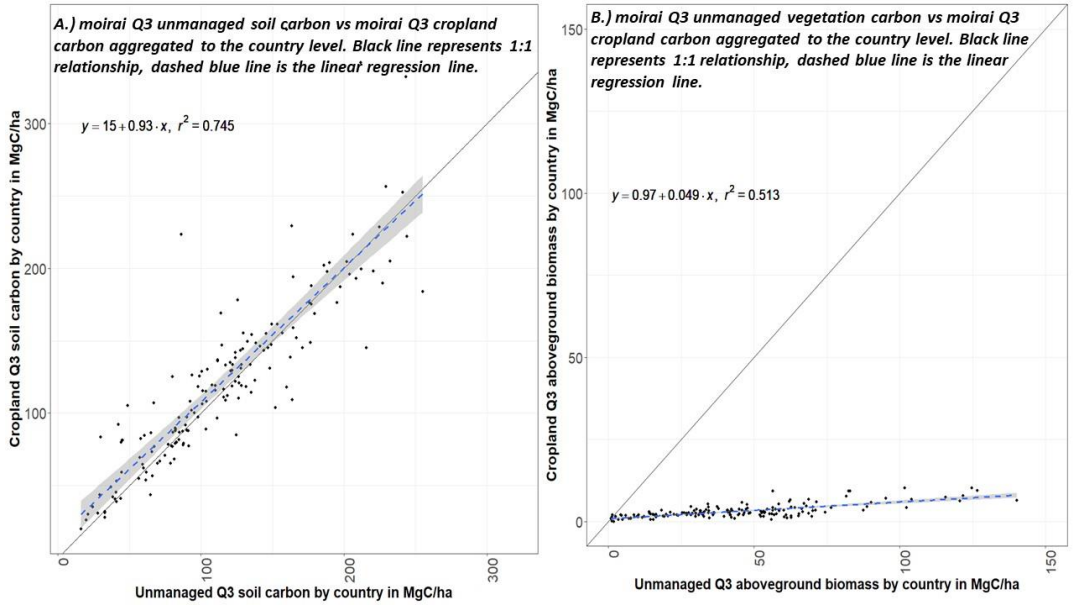


Figure 10: Comparison of unmanaged Q3 carbon densities and Cropland carbon densities for A.) soil carbon and B.) aboveground biomass. Values here are aggregated to individual countries

We performed a similar analysis for pasture carbon densities and found that Pasture carbon shows depletion or lower values compared to unmanaged land cover both for both soil and vegetation. In this case, it is reasonable to use the Q3 soil and vegetation carbon values for Pasture in GCAM without adjustment.

ii.) ***Assesing spatial uncertainties in soil and vegetation carbon within and across basins***

We have established that carbon distributions within land regions generally follow a lognormal pattern for soil carbon and for vegetation carbon for most land types while vegetation carbon for Forests has a more bimodal distribution. However, there may be more dispersion across values in some basins for some land types compared to others. To assess this systematically, we computed a quartile coefficient of dispersion (QCD) for each basin and land type as:

$$QCD_{GLU,LT,pool} = (Q3_{GLU,LT,pool} - Q1_{GLU,LT,pool})/(Q3_{GLU,LT,pool} + Q1_{GLU,LT,pool}) \quad (5)$$

Where,
pool is the carbon pool (aboveground biomass, belowground biomass, topsoil (0-30 cms)),
GLU represents a land region which is an intersection of basin boundaries and country boundaries,
and LT is the land type.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

The QCD values ranges from 0 to 1 where a value towards zero indicates less dispersion within a region-land type-carbon pool combination and a value towards 1 indicates more dispersion.

The QCD values for soil carbon (Figure 11) are generally similar across most basins across land types. This is expected since the distributions of soil carbon are generally lognormal. However, in some basins the QCD value is consistently high and similar across land types. This mainly occurs in individual basins in Russia and Indonesia which have high levels of peat soils which would mean that the level of dispersion across cells would be high since some cells would contain peat soils whereaes others would not.

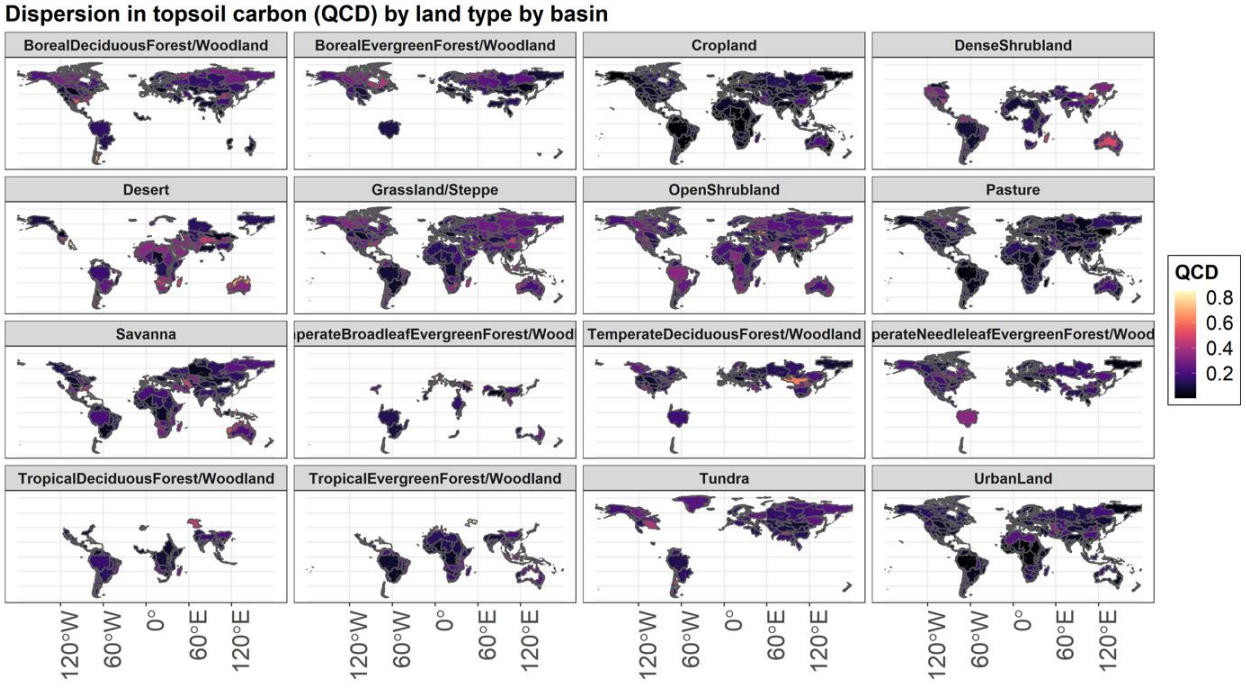


Figure 11: QCD values for topsoil carbon across basins and land types

Based on QCD values across basins and land types for vegetation carbon (Figure 12), we observe that there is significant variation in the QCD values within and across basins for Tundra (with values ranging from 0-1). This is likely due to the way Tundra pixels are defined in our dataset (they encompass different vegetation types). Similarly, there is significant variation within and across basins for grasslands, savannah and pastures, which is once again likely due to the definitions of what constitutes grasslands in the base land cover dataset. While there are also variations in vegetation carbon values for cropland and urbanland, the overall range of values for these land types when it comes to vegetation carbon is low (Figure 7). QCD values for forests

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

across and between basins is lower. This may be due to the more narrow definitions for what constitutes forests across datasets.

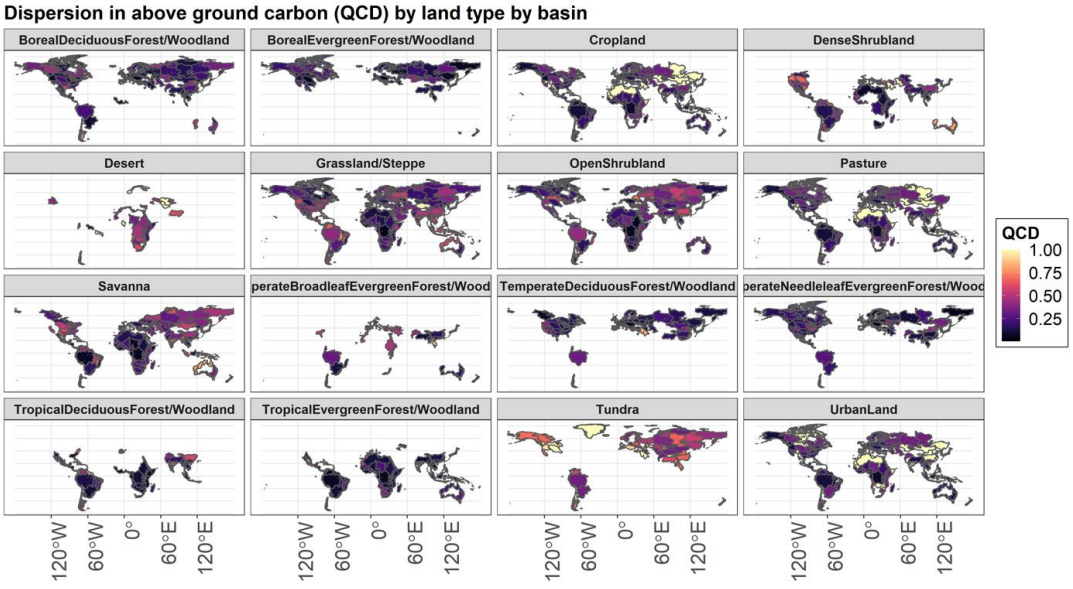


Figure 12: QCD values for aboveground vegetation across basins and land type

4. Results from implementation of spatiually explicit carbon in GCAM

We make the following assumptions when implementing the carbon densities in GCAM, based on the analyses above:

a.) The Q3 carbon values and the 90th percentile values are used throughout to reflect two potential options for a long term potential maximum state of carbon in 1700

b.) Cropland soil carbon is reduced by a factor of 0.3 (30% reduction) for all basins to reflect the effects of management

c.) Tundra, Urban, Desert, and Polar desert/rock/ice do not change in GCAM and so the assigned carbon values do not influence model simulations. If a model does include dynamics for these land types, then the associated uncertainties should be addressed.

4.1 Results from historical spin up

We initialize GCAM using our two options identified above This results in a pre-spin up carbon stock of 2144 PgC (1553 PgC in soil and 591.7 PgC in vegetation) when using the Q3 state and a carbon stock of 3028 PgC (2063 PgC in top soil and 965 PgC in vegetation) when using the 90th percentile. Note that these initialization values are calculated using the land cover in 1700, which does include some managed area, and the spatially explicit carbon.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

Open Access Earth System Science Data Discussions

During the spin up (Table 6), this carbon is reduced to 1967 PgC in 2015 when using the Q3 state(1481 PgC of topsoil carbon and 486 PgC of vegetation carbon) as a result of historical land transitions. Similarly during the spin up, this carbon is reduced to 2758 PgC when using the 90th percentile values (1965 PgC of topsoil carbon and 793 PgC of vegetation carbon).

An important point to note is that while the 90th percentile generates results more in line with independent pre-industrial estimates (e.g. Walker Sanderman, Erb), the Q3 state results in more realistic contemporary values in 2015 during the GCAM spin up. For example, the Q3 state results in a contemporary value of 486 PgC of vegetation carbon in 2015, which is closer to contemporary vegetation carbon stock estimates. Whereaes, the 90th percentile results in a global vegetation carbon stock of 793 PgC. Using the 90th percentile would effectively result in an unrealistically high initial vegetation carbon stock that is close to equilibrium in 2015. Furthermore, when running a carbon price scenario using the 90th percentile densities (Figure 13), the model would add another 54 PgC of vegetation carbon through afforestation that would result in a very unrealistic value of vegetation carbon in 2100- Close to 847 PgC which is higher than undistrurbed carbon stocks in 1700. When using the Q3 densities, this vegetation carbon stock in 2100 is close to 515 PgC (Additional 30 PgC of carbon added through planted forests).

Another point to note is that the amount of global historical emissions (1700-2015) produced by the Q3 initialization is 176 PgC which is much lower than the global historial emissions using the 90th percentile of 270 PgC. For context, the Global Carbon Project (as of 2021) produced an estimate of annual LUC emissions from 1700-2015 of 196 PgC(Friedlingstein et al., 2022). Figure 14 below shows the LUC emissions for the historical periods for the GCP and our two initialization options. As seen in the figure, the 90th percentile produces consistently higher annual LUC emissions.

Given the above results from the spin up, we found that the Q3 value from our dataset is appropriate for initialization and use in GCAM when using the model to estimate contemporary C dynamics. While the 90th percentile better resembles independent estimates of pre-settlmenet stocks, it results in substantial overestimation when used to estimate contemporary C fluxes. This is a result of assumptions and processes within GCAM pertaining to carbon dynamics. As such, What is appropriate for other models would likely be different and would require a similar analysis.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

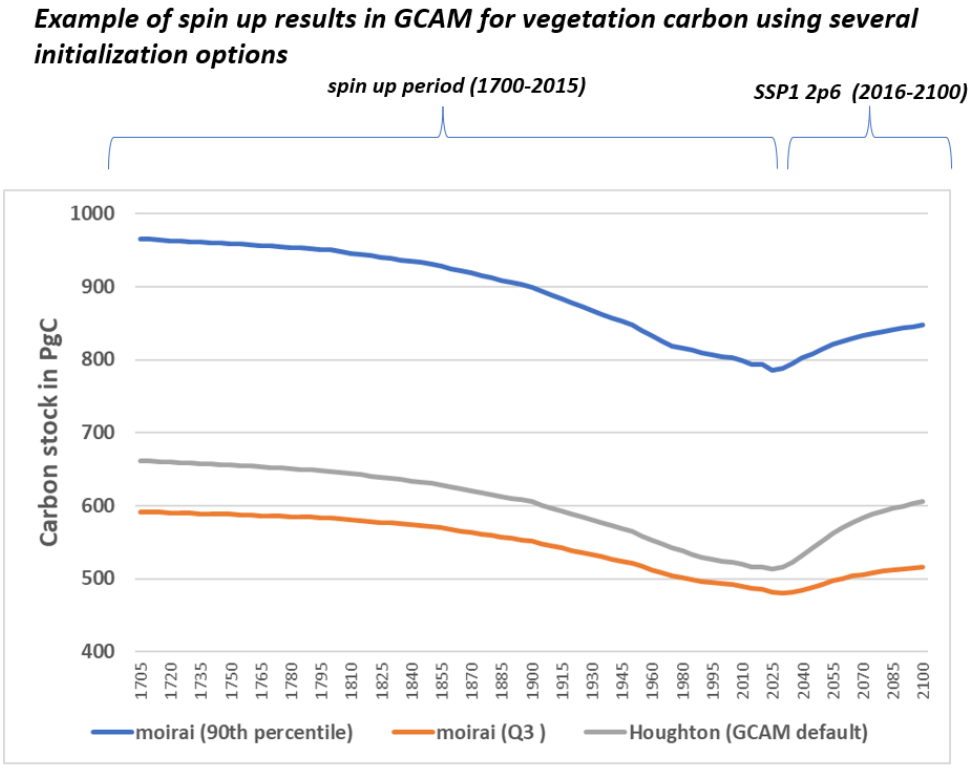


Figure 13: Descriptions of the results of the spin up process. Global vegetation carbon during spin up (1700-2015) and the SSP1 2p6 climate forcing scenario (2016-2100) for our initialization options.

Table 6: Results from the historical spin up

Initialization	carbon pool	Initial value in PgC In the year 1700)	Contemporary value after spin up in PgC (2015)	Historical emissions (PgC) (between 1700 and 2015)	Value in 2100 under SSP1 2p6	Additional carbon sequestered during afforestation scenario (2100 value- 2015 value)
Houghton	vegetation carbon	662.0	516.1	145.9	605.3	89.2
moirai (Q3 value)	vegetation carbon	591.7	486.3	105.4	515.9	29.6

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

moirai (90th percentile)	vegetation carbon	965.8	793.2	172.6	847.9	54.7
Houghton	soil carbon (top-soil)	1243.5	1181.4	62.1	1220.0	38.6
moirai (Q3 value)	soil carbon (top-soil)	1320.5	1249.1	71.4	1274.6	25.5
moirai (90th percentile)	soil carbon (top-soil)	1753.0	1655.2	97.8	1700.6	45.4
Houghton	**Total terrestrial carbon**	**1905.5**	**1697.5**	**208.0**	**1825.3**	**127.8**
moirai (Q3 value)	**Total terrestrial carbon**	**1912.3**	**1735.4**	**176.9**	**1790.6**	**55.2**
moirai (90th percentile)	**Total terrestrial carbon**	**2718.8**	**2448.3**	**270.5**	**2548.4**	**100.1**

1
2

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

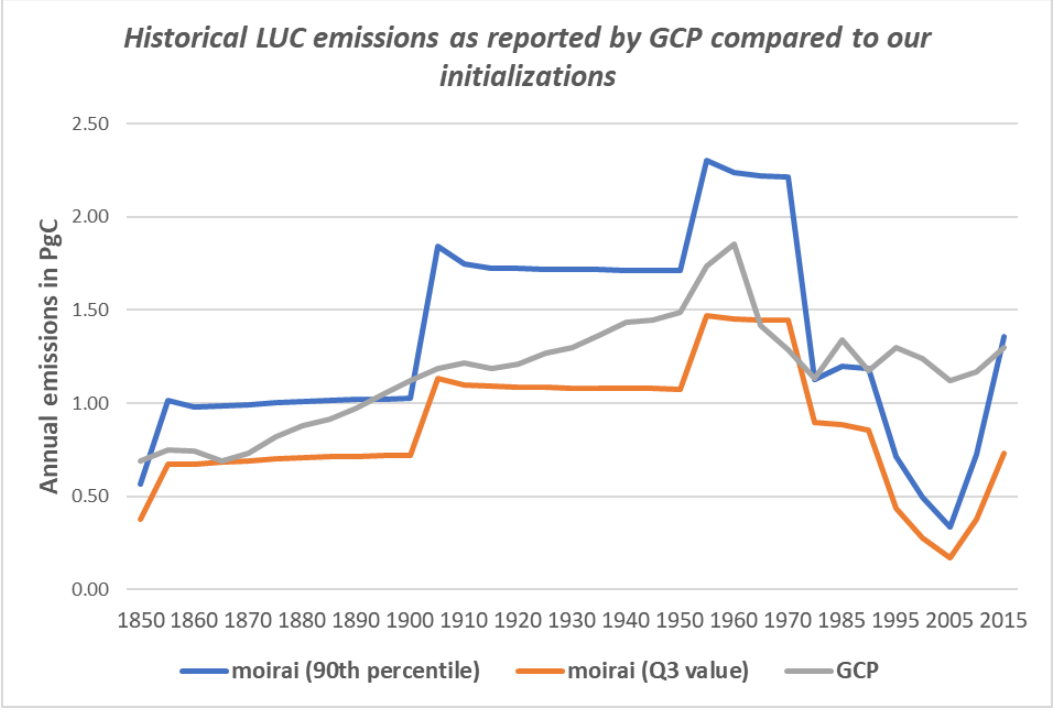


Figure 14: Annual Global LUC emissions from GCP 2021 and our two initialization options

4.2 Results from climate forcing scenario

We use one climate forcing socioeconomic scenario (SSP1 2p6) with a maximum radiative forcing level of 2.6 watts per square meter by 2100 to assess how the new carbon data influence land projection in GCAM. Under this scenario land carbon prices are implemented to assign value to terrestrial carbon at the same rate as carbon is valued in the energy system. GCAM by default uses carbon densities from Houghton et al. (1999) which are described in SI Table 5 (soil) and SI Table 6 (vegetation). Note that the changes in land cover under the climate forcing scenario are driven by relative levels of carbon across land types rather than absolute levels of carbon. Therefore, even if forest carbon in some tropical regions are lower than other estimates, forests still sequester much more carbon compared to other land types in these regions. Below, we will compare results for the climate forcing scenario when using values from Houghton, the moirai Q3 value and the moirai 90th percentile.

The global land allocation comparison under SSP1 2p6 scenario in GCAM (Figure 15) shows that the afforestation/reforestation response is greatly reduced as a result of the spatially explicit carbon (the increase in forest cover from 2020 to 2100 globally is only 3.2 million km^2 when using the moirai Q3 as opposed to 7 million km^2 with the

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

Houghton carbon). IAMs (Including GCAM) generally show a very optimistic afforestation response for this scenario that ranges from 0.5 to 12 million km^2 of trees planted as part of a nature based carbon sequestration strategy under SSP1 2p6 (Popp et al., 2017). The afforestation response in IAMs has been considered too optimistic in some studies (Pongratz et al., 2021). The reduced forest expansion in GCAM with the new carbon data is largely driven by lower vegetation carbon densities in the new data. When using the 90^{th} percentile, the afforestation values are the lowest at 0.1 million km2. This is expected given that the 90^{th} percentile carbon values are much higher so much lower increases in forest cover are required to meet additional afforestation targets. Usage of the 90^{th} percentile values leads to minimal changes across land types under a climate forcing scenario. This would be expected since the carbon pool generated by this scenario in 2015 is already close to equilibrium. This further illustrates why the Q3 state from our dataset is a better choice for initialization in a model like GCAM.

However, the afforestation responses are diverse by regions. In case of tropical forests, there is an increase in the afforestation response with our updated carbon densities. In case of boreal forests, the opposite is true.

Global Cropland and Shrubland dynamics show a more complicated response. The reduced emphasis on forest expansion reduces the need for Cropland abandonment. Cropland also sequesters more soil carbon in some regions (even with the 30% reduction factor), which also reduces abandonment. This Shrubland response is also enhanced by higher Shrubland vegetation carbon densities.

Regional responses are dictated by their respective land type distrubutions (e.g., SI Figure 5). For example, in Russia the afforestation strategy is completely replaced with a shrubland and grassland preservation strategy. This is expected since the region has a relatively high amount of boreal forests. In South Asia however, where non-forest land types dominate, forest expansion persists and is supplemented by shrubland expansion.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

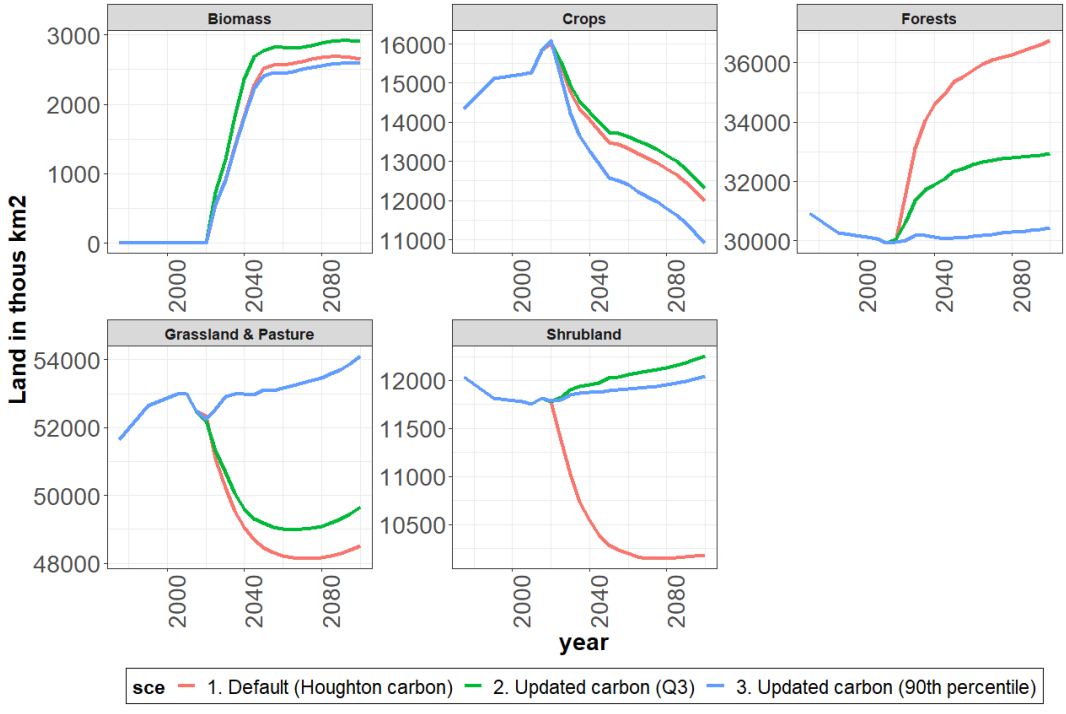


Figure 15: Global land allocation in GCAM under the SSP1 2p6 scenario by land type

The implementation of the spatially explicit carbon clearly improves land use responses and also suggests that high carbon sequestering shrubs can also be preserved as a part of nature based solutions to mitigate climate change. The robustness of these responses across other radiative forcing scenarios (implemented for more SSPs for example) and across other models need to be studied and is a subject worthy of exploration in a future paper.

5. Discussion and conclusion

In this paper we present a new dataset of grid cell level spatially explicit carbon harmonized with Moriai/GCAM land types. Our harmonized dataset presents carbon values for 3 pools (topsoil, above ground biomass and below ground biomass) for six statistical states for various land use types. Our dataset is available both at a 5 arcmin resolution and aggregated to 699 land regions. This dataset is specifically designed to enable initialization of spatially explicit carbon in IAMs and MSD models. This dataset was specifically designed to generate carbon values for GCAM, but can and should be extensible to other models. In the future, this dataset can be extended to include deeper soil (beyond 0-30 cms) so that land use responses in models can account for an additional deep soil carbon pool.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

We noted that there are some limitations with respect to the carbon observations (both for soil and vegetation) for the Tundra region. For example, we could find no data for 29% of the 5 arcmin gridcells for this land type. The biome mapping also needed to include several source land types to enable an increase in data coverage for Tundra. This issue was likely caused by the different definitions of Tundra land cover in different datasets. Recently, there have been efforts dedicated to collecting carbon data specifically for this land type. These data should be integrated in future releases of our data to address the current lack of data coverage.

As a part of our analysis, we observed that SoilGrids soil carbon values for cropland do not show a depletion when compared to SoilGrids soil carbon values in unmanaged land. As discussed, this is likely due the locations of sampling for cropland soil carbon. As a result, we reduced Cropland soil carbon by 30% when we applied it to GCAM. If better/improved data on crop soil carbon become available, our data should be updated with the same.

We have also noted that our current estimates of forest vegetation carbon are based on both primary and secondary forests. This is due to the lack of availability of fine resolution (300 m) land masks that distinguish between primary and secondary forests. As more data become available related to forest cover types, a logical next step would be to break out different forest types in our dataset.

Finally, our analysis showed that using the Q3 statistical state was most appropriate for GCAM even though it resulted in an initialization of pre-industrial carbon value that was lower than other estimates. Selection of the Q3 results in more accurate historical LUC emissions and the model therefore spins up to a value that is close to other estimates in the literature in 2015. What the correct initialization value is will differ from model to model and would require a similar analysis.

6. Data availability statement

Final data are available for download here- https://zenodo.org/record/7884615 (Narayan et al., 2023) The data repository contains the following-

1. 72 rasters (4 land use types X 6 states X 3 carbon pools) at a 5 arcmin resolution representative of carbon in 2010
2. 1 thematic raster which tracks 15 vegetation biomes for Unmanaged land use type (from 1. above)
3. Tabular data file showing aggregated carbon densities for 6 states of carbon for 699 land regions for soil (0-30cm), aboveground biomass and belowground biomass.

7. Code availability statement

As mentioned above, the data can be generated programmatically with scripts that are hosted on GitHub (https://github.com/JGCRI/moirai/tree/master/ancillary/carbon_harmonization).

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

The process has been split into two steps where the computationally intensive stage 1 (approx.. 6 hours of processing) is optional with outputs made available in the repository. The Stage 1 processing is performed using bash scripts which use the GDAL software (Warmerdam, 2008). The second stage processing uses an R script and can be completed for all carbon pools in approx. 15 minutes to generate the final 72 rasters and the final tabular output file. We have also made available optional diagnostic functions in the R script which can be used to validate results.

Competing Interests declaration- The authors have declared that none of the authors has any competing interests.

Acknowledgements
This research was supported by the U.S. Department of Energy, Office of Science, as part of research in Multi Sector Dynamics, Earth and Environmental System Modeling Program. The Pacific Northwest National Laboratory is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830

References

Barnes, W. L., Xiong, X., & Salomonson, V. V. (2003). Status of terra MODIS and aqua MODIS. *Advances in Space Research*, *32*(11), 2099-2106.

Batjes, N. H., Ribeiro, E., Van Oostrum, A., Leenaars, J., Hengl, T., & Mendes de Jesus, J. (2017). WoSIS: providing standardised soil profile data for the world. *Earth System Science Data*, *9*(1), 1-14.

Calvin, K., Patel, P., Clarke, L., Asrar, G., Bond-Lamberty, B., Cui, R. Y., Di Vittorio, A., Dorheim, K., Edmonds, J., & Hartin, C. (2019). GCAM v5. 1: representing the linkages between energy, water, land, climate, and economic systems. *Geoscientific Model Development*, *12*(2), 677-698.

Cooper, H., Sjögersten, S., Lark, R., & Mooney, S. (2021). To till or not to till in a temperate ecosystem? Implications for climate change mitigation. *Environmental Research Letters*, *16*(5), 054022.

Di Vittorio, A. V., Vernon, C. R., & Shu, S. (2020). Moirai version 3: a data processing system to generate recent historical land inputs for global modeling applications at various scales. *Journal of Open Research Software*, *8*(PNNL-SA-142149).

Erb, K.-H., Kastner, T., Plutzar, C., Bais, A. L. S., Carvalhais, N., Fetzel, T., Gingrich, S., Haberl, H., Lauk, C., & Niedertscheider, M. (2018). Unexpectedly large impact of forest management and grazing on global vegetation biomass. *Nature*, *553*(7686), 73-76.

Fang, Y., Liu, C., Huang, M., Li, H., & Leung, L. R. (2014). Steady state estimation of soil organic carbon using satellite-derived canopy leaf area index. *Journal of Advances in Modeling Earth Systems*, *6*(4), 1049-1064.

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C., Hauck, J., Le Quéré, C., Peters, G. P., Peters, W., & Pongratz, J. (2022). Global carbon budget 2021. *Earth System Science Data*, *14*(4), 1917-2005.

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G., & Walsh, M. G. (2014). SoilGrids1km—global soil information based on automated mapping. *PloS one*, *9*(8), e105992.

Houghton, R. (2005). Aboveground forest biomass and the global carbon balance. *Global Change Biology*, *11*(6), 945-958.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

Houghton, R. A. (1999). The annual net flux of carbon to the atmosphere from changes in land use 1850–1990. *Tellus B*, *51*(2), 298-313.

Hugelius, G., Strauss, J., Zubrzycki, S., Harden, J. W., Schuur, E., Ping, C.-L., Schirrmeister, L., Grosse, G., Michaelson, G. J., & Koven, C. D. (2014). Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps. *Biogeosciences*, *11*(23), 6573-6593.

Jackson, R. B., Lajtha, K., Crow, S. E., Hugelius, G., Kramer, M. G., & Piñeiro, G. (2017). The ecology of soil carbon: pools, vulnerabilities, and biotic and abiotic controls. *Annual Review of Ecology, Evolution, and Systematics*, *48*(1), 419-445.

Jungkunst, H. F., Göpel, J., Horvath, T., Ott, S., & Brunn, M. (2022). Global soil organic carbon–climate interactions: Why scales matter. *Wiley Interdisciplinary Reviews: Climate Change*, e780.

Justice, C., Townshend, J., Vermote, E., Masuoka, E., Wolfe, R., Saleous, N., Roy, D., & Morisette, J. (2002). An overview of MODIS Land data processing and product status. *Remote sensing of Environment*, *83*(1-2), 3-15.

Klein Goldewijk, K., Beusen, A., Doelman, J., & Stehfest, E. (2017). Anthropogenic land use estimates for the Holocene–HYDE 3.2. *Earth System Science Data*, *9*(2), 927-953.

Li, W., MacBean, N., Ciais, P., Defourny, P., Lamarche, C., Bontemps, S., Houghton, R. A., & Peng, S. (2018). Gross and net land cover changes in the main plant functional types derived from the annual ESA CCI land cover maps (1992–2015). *Earth System Science Data*, *10*(1), 219-234.

Liu, X., Yu, L., Si, Y., Zhang, C., Lu, H., Yu, C., & Gong, P. (2018). Identifying patterns and hotspots of global land cover transitions using the ESA CCI Land Cover dataset. *Remote Sensing Letters*, *9*(10), 972-981.

Meiyappan, P., & Jain, A. K. (2012). Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years. *Frontiers of earth science*, *6*(2), 122-139.

Nachtergaele, F., van Velthuizen, H., Verelst, L., Batjes, N., Dijkshoorn, K., van Engelen, V., Fischer, G., Jones, A., & Montanarela, L. (2010). The harmonized world soil database. Proceedings of the 19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia, 1-6 August 2010,

Kanishka B. Narayan, Alan Di Vittorio, Evan Margiotta, Seth A Spawn, & Holly Gibbs. (2023). Spatially explicit re-harmonized terrestrial carbon densities for calibrating Integrated Multisectoral Models (1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7884615

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil*, *7*(1), 217-240.

Pongratz, J., Schwingshackl, C., Bultan, S., Obermeier, W., Havermann, F., & Guo, S. (2021). Land use effects on climate: current state, recent progress, and emerging topics. *Current Climate Change Reports*, 1-22.

Popp, A., Calvin, K., Fujimori, S., Havlik, P., Humpenöder, F., Stehfest, E., Bodirsky, B. L., Dietrich, J. P., Doelmann, J. C., & Gusti, M. (2017). Land-use futures in the shared socio-economic pathways. *Global Environmental Change*, *42*, 331-345.

Ramankutty, N., & Foley, J. A. (1999). Estimating historical changes in land cover: North American croplands from 1850 to 1992: GCTE/LUCC RESEARCH ARTICLE. *Global Ecology and Biogeography*, *8*(5), 381-396.

Sanderman, J., Hengl, T., & Fiske, G. J. (2017). Soil carbon debt of 12,000 years of human land use. *Proceedings of the National Academy of Sciences*, *114*(36), 9575-9580.

Scharlemann, J. P., Tanner, E. V., Hiederer, R., & Kapos, V. (2014). Global soil carbon: understanding and managing the largest terrestrial carbon pool. *Carbon Management*, *5*(1), 81-91.

Spawn, S. A., Sullivan, C. C., Lark, T. J., & Gibbs, H. K. (2020). Harmonized global maps of above and belowground biomass carbon density in the year 2010. *Scientific Data*, *7*(1), 1-22.

https://doi.org/10.5194/essd-2023-251
Preprint. Discussion started: 17 July 2023

Thomson, A. M., Calvin, K. V., Chini, L. P., Hurtt, G., Edmonds, J. A., Bond-Lamberty, B., Frolking, S., Wise, M. A., & Janetos, A. C. (2010). Climate mitigation and the future of tropical landscapes. *Proceedings of the National Academy of Sciences*, *107*(46), 19633-19638.

Tifafi, M., Guenet, B., & Hatté, C. (2018). Large differences in global and regional total soil carbon stock estimates based on SoilGrids, HWSD, and NCSCD: Intercomparison and evaluation based on field data from USA, England, Wales, and France. *Global Biogeochemical Cycles*, *32*(1), 42-56.

van Asselen, S., & Verburg, P. H. (2012). AL and S ystem representation for global assessments and land-use modeling. *Global Change Biology*, *18*(10), 3125-3148.

Walker, W. S., Gorelik, S. R., Cook-Patton, S. C., Baccini, A., Farina, M. K., Solvik, K. K., Ellis, P. W., Sanderman, J., Houghton, R. A., & Leavitt, S. M. (2022). The global potential for increased storage of carbon on land. *Proceedings of the National Academy of Sciences*, *119*(23), e2111312119.

Warmerdam, F. (2008). Open source approaches in spatial data handling. *by Hall, GB & Leahy, MG Berlin, Heidelberg: Springer Berlin Heidelberg*, 87-104.

Wei, X., Shao, M., Gale, W., & Li, L. (2014). Global pattern of soil carbon losses due to the conversion of forests to agricultural land. *Scientific reports*, *4*(1), 1-6.

Wieder, W. R., Boehnert, J., & Bonan, G. B. (2014). Evaluating soil biogeochemistry parameterizations in Earth system models with observations. *Global Biogeochemical Cycles*, *28*(3), 211-222.

Wise, M., Calvin, K., Thomson, A., Clarke, L., Bond-Lamberty, B., Sands, R., Smith, S. J., Janetos, A., & Edmonds, J. (2009). Implications of limiting CO2 concentrations for land use and energy. *Science*, *324*(5931), 1183-1186.