Most root-derived carbon inputs do not contribute to the global bulk soil carbon pool

Authors: Guocheng Wang1,2†, Liujun Xiao2†, Xiali Mao2, Xiaowei Guo2, Annette Cowie3, Shuai Zhang2, Mingming Wang2, Songchao Chen4, Ganlin Zhang5, Zhou Shi2, Zhongkui Luo2*

Affiliations: 1State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China; 2College of environmental and resource sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China; 3New South Wales Department of Primary Industries/University of New England, Armidale, NSW 2351, Australia; 4INRAE Unité InfoSol, Orléans 45075, France; 5State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, Jiangsu 210008, China; †These authors contributed equally; *email: luozk@zju.edu.cn

Plant root-derived carbon (C) inputs (I_root) are the primary source of C in mineral bulk soil. However, a fraction of I_root may lose directly (I_loss, e.g., via rhizosphere microbial respiration, leaching and fauna feeding) without contributing to bulk soil C pool. This loss has never been quantified, particularly at global scale, inhibiting reliable estimation of soil C dynamics. Here we integrate three observational global datasets including radiocarbon content, allocation of photosynthetically assimilated C, and root biomass distribution in 2,034 soil profiles to quantify I_root and its contribution to the bulk soil C pool. We show that global average I_root in the 0-200 cm soil profile is 3.5 Mg ha⁻¹ yr⁻¹, ~80% of which (i.e., I_loss) is lost rather than entering bulk soil. If ignoring I_loss, bulk soil C turnover will be incorrectly estimated to be four times faster. This can explain why Earth system models (in which all I_root enters bulk soil C pools) predict much faster soil C turnover than radiocarbon-constrained estimates. I_root decreases exponentially with soil
depth, and the top 20 cm soil contains >60% of total $I_{\text{root}}$. Actual C input to bulk soil (i.e., $I_{\text{root}} - I_{\text{loss}}$) shows a similar depth distribution to $I_{\text{root}}$. We also map $I_{\text{loss}}$ and its depth distribution across the globe. Our results demonstrate the global significance of direct C losses which limit the contribution of $I_{\text{root}}$ to bulk soil C storage; and provide spatially explicit data to facilitate reliable soil C predictions via separating direct C losses from total root-derived C inputs.

Globally, mineral soil is the largest terrestrial carbon (C) reservoir, mediating C exchanges among other C reservoirs particularly the atmosphere and water bodies (e.g., rivers, lakes, groundwater and oceans)\textsuperscript{1,2}. Soil receives photosynthetically assimilated C (i.e., organic carbon inputs) and releases C mainly via microbial decomposition (i.e., heterotrophic respiration outputs). Both C inputs and outputs to the soil system, the balance of which determines soil C stocks, are influenced by various environmental factors\textsuperscript{1,3,4}, resulting in great spatial and temporal variability of soil C stocks down the soil profile as well as across the globe\textsuperscript{5,6}. While bulk soil C stock is relatively easy to measure via sampling soil cores, the determination of C inputs and outputs is not that straightforward. Indeed, \textit{in situ} quantification of C outputs in the form of CO$_2$ from the bulk soil is a grand challenge because CO$_2$ emissions from the soil are comprised of heterotrophic respiration by decomposers and root autotrophic respiration, which are difficult, if not impossible, to separate\textsuperscript{7}. An even harder question is how to partition CO$_2$ emissions to different soil depths where CO$_2$ is produced. Our poor knowledge of C inputs to and outputs from different soil depths inhibits understanding of whole-soil profile C dynamics in response to environmental change factors\textsuperscript{8,9} and management\textsuperscript{10,11}.

Compared to challenges of quantifying C outputs and their depth-origin, C inputs to the soil can be represented by root-derived C and reasonably estimated using state-of-the-art tracer (e.g., C isotopes) techniques\textsuperscript{8} or by root growth measurements\textsuperscript{12-14}. Numerous relevant data
have been collected in the last decades\textsuperscript{15}, and used to estimate soil C turnover based on the ratio of soil C stocks to C inputs under the assumption of steady state\textsuperscript{15,16}. Alternatively, radiocarbon-inferred C age can be used to estimate soil C turnover time\textsuperscript{8,10,17}, which is equivalent to C age\textsuperscript{18}. In theory, soil C turnover times estimated by the two approaches should be the same or similar at least, but, surprisingly, have nearly an order of magnitude difference\textsuperscript{8,10,15,17,19,20}. For example, the average turnover time of soil C in the 30-100 cm soil layer across the globe is estimated to be \textasciitilde 1,000 years based on belowground C allocation\textsuperscript{15}, and \textasciitilde 8,000 years through radiocarbon\textsuperscript{17}. An immediate question is: why? A possible explanation for this contrast is the steady state assumption adopted by both approaches, because real systems are transient, particularly in this anthropogenic era with significant perturbations. However, both empirical and modelling studies have demonstrated that the consequences of the steady state assumption on the millennial C turnover are minimum\textsuperscript{8,21}. Another possibility is that plant-derived C inputs only participate in the cycling of a small fraction of total soil C pool\textsuperscript{10,17}. This explanation is mechanistically possible but has not been tested at global scale, and conflicts with the common assumption that both methods estimate the average turnover time of soil C pool.

Here we propose that the apparent root-derived C inputs ($I_{\text{root}}$) does not equal actual C inputs to bulk soil ($I_{\text{bulk}}$) because there is a “leaky bridge” connecting $I_{\text{root}}$ and $I_{\text{bulk}}$. This “bridge” may be the rhizosphere or other direct C loss pathways, where direct C efflux (e.g., volatile root exudates\textsuperscript{22}) and/or transformation (e.g., microbial decomposition and root herbivore consumption\textsuperscript{23}) processes are active, reducing the transfer of $I_{\text{root}}$ to bulk soil\textsuperscript{24}. The importance of direct C loss pathways such as rhizosphere processes in controlling soil C dynamics has been acknowledged\textsuperscript{24-27}. However, in practice the extent to which the pathways modulate C inputs to bulk soil is rarely considered, mainly due to the difficulty of \textit{in situ} measurement. Earth system models, the common tool for soil C projections at global scale, simply partition $I_{\text{root}}$ into
one or several conceptual C pools in bulk soil\textsuperscript{10,11}, which may underestimate immediate turnover rates of $I_{\text{root}}$ before entering the bulk soil\textsuperscript{28} thereby overestimating the contribution of $I_{\text{root}}$ to the bulk soil C pool. In this study, we first quantify the global pattern of $I_{\text{root}}$ and $I_{\text{bulk}}$ and their drivers, and then estimate the “leaked” amount of $I_{\text{root}}$ [i.e., the direct C loss via various pathways ($I_{\text{loss}}$)] as: $I_{\text{root}} - I_{\text{bulk}}$. At last, taking advantage of the data sources, we map $I_{\text{loss}}$ across global terrestrial soils at the resolution of 1 km.

**Root-derived C inputs ($I_{\text{root}}$)**

The input of root-derived C to soil ($I_{\text{root}}$) can be estimated by combining plant net primary productivity (NPP, i.e., the difference between photosynthesized and respired C) and its belowground allocation ($f_{\text{BNPP}}$): $I_{\text{root}} = \text{NPP} \times f_{\text{BNPP}}$. We compiled a comprehensive dataset of NPP enabling the calculation of $f_{\text{BNPP}}$ from field measurements at 725 sites across the globe(i.e., NPP\_data hereafter; Methods, Extended Data Fig. 1) to assess the global pattern and underlying drivers of $I_{\text{root}}$. Averaged across 725 sites in the 0-200 cm soil profile, $I_{\text{root}}$ is 3.50 Mg ha\textsuperscript{-1} yr\textsuperscript{-1} ranging from 0.09 [2.5% confidence intervals (CI)] to 13.38 Mg ha\textsuperscript{-1} yr\textsuperscript{-1} (97.5% CI, Fig. 1a). The proportional belowground allocation of total NPP (i.e., $f_{\text{BNPP}}$) averages 37.7%, ranging from 7.6% to 82.7% (Fig. 1c). Among biomes, $I_{\text{root}}$ is significantly different ($P < 0.05$, Fig. 1a). Mediterranean/montane shrublands have the highest $I_{\text{root}}$ (5.42 Mg ha\textsuperscript{-1} yr\textsuperscript{-1}), followed by cropland (4.42 Mg ha\textsuperscript{-1} yr\textsuperscript{-1}); tundra has the lowest (0.90 Mg ha\textsuperscript{-1} yr\textsuperscript{-1}), followed by boreal forests (2.36 Mg ha\textsuperscript{-1} yr\textsuperscript{-1}, Fig. 1a). $f_{\text{BNPP}}$ is $>50\%$ in temperate grasslands, deserts and Mediterranean/montane shrublands, but only $\sim30\%$ in tropical/subtropical forests, temperate forests and boreal forests (Fig. 1c and Supplementary Fig. 1).

With respect to depth distribution of $I_{\text{root}}$, the allocation to the 0-20 cm soil layer is 2.43 Mg ha\textsuperscript{-1} yr\textsuperscript{-1} on average across all sites (Fig. 1b), accounting for 60% of total $I_{\text{root}}$ (Fig. 1d), and $>80\%$ $I_{\text{root}}$ is allocated to the top 60 cm soil layer (Fig. 1d). Among biomes, depth
distribution of $I_{\text{root}}$ differs significantly ($P < 0.05$, Supplementary Fig. 2). For example, boreal forests and tundra have much higher allocation to top layers (e.g., 0-20 cm) than deserts and temperate forests. The different depth distribution of $I_{\text{root}}$ reflects the general soil moisture pattern among biomes. That is, roots go to deeper to find moisture in drier environments\(^{29}\). In tundra, permafrost inhibits root growth into the subsoil\(^{30}\), resulting in large fraction of $I_{\text{root}}$ allocated to surface layers.

Focusing on the variability of total $I_{\text{root}}$ across the 725 sites, we assess the underlying environmental drivers including climatic, edaphic and topographic variables, and biome type (Supplementary Table 1), using a multivariate linear mixed regression treating biome type as a random effect (Methods). The regression explains 50% of the variance of global $I_{\text{root}}$ (Fig. 2). Soil and climate show similar overall importance, but are more important than topography. The regression coefficients indicate that the drivers differ within each biome (Fig. 2), reflecting the strong dependence of environmental controls on biome type. These results demonstrate that $I_{\text{root}}$ is driven by complex interplay among edaphic, climatic and topographic variables, which is further modulated by biome type as different biomes may adopt distinct NPP allocation strategies\(^{31-33}\).

**Carbon inputs to bulk soil ($I_{\text{bulk}}$)**

Radiocarbon ($\Delta^{14}\text{C}$) measurements in bulk soil provide the opportunity to inversely estimate actual C inputs to bulk soil ($I_{\text{bulk}}$, Methods). We first used the ISRaD\(^{34}\) radiocarbon dataset ($\Delta^{14}\text{C}_{\text{data}}$) to calculate soil C ages\(^{35-37}\) ($C_{\text{age}}$) in 750 soil profiles across the globe (Extended Data Fig. 1) using a radiocarbon model (Methods) together with historical atmospheric radiocarbon concentration since 1950 (Supplementary Fig. 5a). The results indicate that the average $C_{\text{age}}$ in the 0-200 cm soil profile across the 750 soil profiles is 3,862 years (Supplementary Fig. 5b). Using a similar dataset, Shi et al\(^{17}\) estimated a global average $C_{\text{age}}$ of 1,390 years in the 0-30
cm soil layer, which is generally in line with our estimation of 982 years for that soil depth. In the 30-100 cm soil layer, however, their estimation is 8,280 years, while ours is much younger at an average of 3,869 years. This discrepancy could be due to the estimation by Shi et al being the global average based on global mapping product of $\Delta^{14}C$, while our results are based on original $\Delta^{14}C$ observations in the 750 soil profiles which include few samples in tundra and permafrost regions (Extended Fig. 1) where soil C usually has much longer turnover times than other regions.

Based on the $\Delta^{14}C$-derived $C_{age}$ and soil C stock ($C_{stock}$) measured with $\Delta^{14}C$ (Methods), we can estimate C inputs to bulk soil ($I_{bulk}$) in each soil layer for $\Delta^{14}C_{data}$ profiles as $C_{stock}/C_{age}$ under the steady state assumption (Methods). Averaging across the 750 soil profiles in the 0-200 cm soil profile, $I_{bulk}$ is 0.60 Mg ha$^{-1}$ yr$^{-1}$ ranging from 0.06 (2.5% CI) to 3.59 Mg ha$^{-1}$ yr$^{-1}$ (97.5% CI, Fig. 1a); and is significantly different among global biomes (Fig. 1a). Temperate forests have the highest $I_{bulk}$ (1.20 Mg ha$^{-1}$ yr$^{-1}$), followed by boreal forests (0.77 Mg ha$^{-1}$ yr$^{-1}$); and deserts have the lowest $I_{bulk}$ (0.16 Mg ha$^{-1}$ yr$^{-1}$), followed by tundra (0.26 Mg ha$^{-1}$ yr$^{-1}$, Fig. 1a). The depth distribution of $I_{bulk}$ is generally consistent with that of $I_{root}$ (Fig. 1b and d). On average, the results indicate 0.26, 0.06 and 0.03 Mg ha$^{-1}$ yr$^{-1}$ of $I_{bulk}$ in the 0-20 cm, 20-40 cm and 40-60 cm soil layers, respectively (Fig. 1b), accounting for 68.9%, 14.1%, and 5.2% of total $I_{bulk}$, respectively (Fig. 1d). Across all biomes, $I_{bulk}$ is significantly lower than $I_{root}$ (Fig. 1a). As expected, depth is the dominant control on $I_{bulk}$ in different soil layers (Fig. 2). The multivariate linear mixed regression considering soil depth, edaphic, climatic and topographic variables, and biome type (Methods) explains 74% of the variance of $I_{bulk}$ (Fig. 2). Compared to the effects of environmental factors on $I_{root}$, their effects on $I_{bulk}$ are more consistent among biome types. In addition, the three groups of environmental factors (i.e., edaphic, climatic and topographic) show similar overall importance for controlling $I_{bulk}$ (Fig. 2).
Loss of root-derived C inputs (I_{loss})

Using the estimates of I_{root} and I_{bulk}, we can calculate the difference between I_{root} and I_{bulk} (i.e., I_{root} – I_{bulk}), which represents the loss of root-derived C inputs (I_{loss}) that leaves the soil via any direct loss pathways without entering the bulk soil and participating in the C cycle there.

Considering potential uncertainties in NPP thus I_{root} and its allocation to soil depths, we conducted six independent estimates of I_{loss} based on four independent datasets (Methods, Extended Data Table 1) to provide confidence in the estimate of I_{loss} (Fig. 3, Supplementary Table 2 and Table 3).

Across the globe, the six independent estimates of I_{loss} are generally comparable, with an average of 2.31 Mg ha\(^{-1}\) yr\(^{-1}\) in the 0-200 cm soil profile (Fig. 3a), ranging from 1.80 to 3.36 Mg ha\(^{-1}\) yr\(^{-1}\) across different estimates (Supplementary Table 2). This average I_{loss} accounts for 77.9\% of I_{root} ranging from 65.3\% to 90.9\% among the six estimates (Fig. 3c and Supplementary Table 2), demonstrating that most I_{root} does not enter bulk soil. Among biome types, I_{loss} is significantly different (P < 0.05, Fig. 3a and Supplementary Table 2). On average, tropical/subtropical forests have the largest I_{loss} of 3.30 Mg ha\(^{-1}\) yr\(^{-1}\), followed by crops (2.74 Mg ha\(^{-1}\) yr\(^{-1}\)); and tundra has the smallest I_{loss} (0.66 Mg ha\(^{-1}\) yr\(^{-1}\)), followed by boreal forests (1.35 Mg ha\(^{-1}\) yr\(^{-1}\), Fig. 3a). With respect to depth distribution, the average I_{loss} is 1.45, 0.44 and 0.16 Mg ha\(^{-1}\) yr\(^{-1}\) in the 0-20 cm, 20-40 cm and 40-60 cm soil layer respectively (Fig. 3b), accounting for 61.8\%, 19.1\% and 7.2\% of total I_{loss} in the 0-200 cm soil profile, respectively (Fig. 3d). Depth distributions of I_{loss} quantified by the six independent estimates are also generally comparable (Figs. 3b and d), verifying the robustness of the estimation of I_{loss}. We note that I_{loss} is negative in some situations (Figs. 3a and b), which is possible if there is significant downward movement of C in the soil profile\(^{38}\). For example, the top mineral soil layer would receive C from organic matter on soil surface due to downward movement of
dissolved and particulate C, or soil mixing through bioturbation. For this reason, $I_{\text{bulk}}$ in that layer may include illuvial C or other non-root-derived C.

We developed ensemble models based on five machine learning algorithms to predict $I_{\text{root}}$, $I_{\text{bulk}}$ and their depth distributions (Methods). Independent validation of the derived machine learning models indicates that the $I_{\text{root}}$ and $I_{\text{bulk}}$ models are accurate (Extended Data Figs. 2 and 3). Using the independently tested models together with 54 global layers of soil, climate, biome and topography covariates (Supplementary Table 1), we digitally mapped $I_{\text{root}}$, $I_{\text{bulk}}$ and their depth distribution (Methods, Extended Figs. 4 and 5) and their uncertainty across the globe at the resolution of 30 arc-second or approximately 1 km (Supplementary Figs. 6 and 7). Then, these $I_{\text{root}}$ and $I_{\text{bulk}}$ maps were used to estimate global $I_{\text{loss}}$ (i.e., $I_{\text{root}} - I_{\text{bulk}}$, Fig. 4 and Extended Data Fig. 6) and its uncertainty (Methods, Extended Data Fig. 7). Across the globe, for each 1 km cell, $I_{\text{root}}$ in the 0-200 cm soil profile ranges from 0.19 to 3.98 Mg ha$^{-1}$ yr$^{-1}$ (lower and upper limits of the 95% CI, Extended Data Fig. 4h). The maps reveal notable spatial patterns of $I_{\text{root}}$, with the highest $I_{\text{root}}$ in tropical regions and lowest in Arctic regions. For $I_{\text{bulk}}$, it ranges from 0.15 to 2.58 Mg ha$^{-1}$ yr$^{-1}$ (Extended Data Fig. 5). $I_{\text{loss}}$ ranges from $-0.44$ to 3.18 Mg ha$^{-1}$ yr$^{-1}$ across the globe (Fig. 4a). Higher $I_{\text{loss}}$ generally occurs in lower latitudinal regions, and vice versa (Figs. 4a and c). In contrast, larger uncertainties in quantified $I_{\text{loss}}$ generally occur in higher latitudes (Figs. 4b and d). These maps illustrate the global significance of direct C losses and provide spatial insights into prediction uncertainties.

**Implications and conclusions**

The connections and interactions between belowground and aboveground processes remain elusive and unquantified. One of the key issues is our poor understanding of the plant-soil interactions in which the rhizosphere may play a core mediating role. Under elevated atmospheric CO$_2$, for example, numerous field manipulation experiments have found enhanced
photosynthetic C assimilation and thus stimulated C inputs to soil, but soil C stock shows no or marginal response to such enhanced belowground C inputs and remains relatively stable\textsuperscript{40,41}. The reason for this imbalance is widely debated. One explanation proposed is that microbial decomposition of organic matter in bulk soil is accelerated as microbes “mine” nutrients from soil organic matter in order to utilize the enhanced input of newly assimilated C, particularly in nutrient-poor environments. But this ignores the fact that plants also need additional nutrients to support their stimulated growth\textsuperscript{4}. Our results shed light on another mechanism, i.e., only a minor fraction of additional I\textsubscript{root} goes to bulk soil and the remaining majority has left the system via direct loss pathways such as rhizosphere microbial respiration. Consistent with this mechanism, a recent study of a mature forest under elevated CO\textsubscript{2} indeed found that half of the extra C fixed under elevated CO\textsubscript{2} was quickly emitted through soil respiration, and did not result in bulk soil C accumulation\textsuperscript{41}. Our results indicate that it may be general across the globe that the rhizosphere provides a buffer zone before environment change-induced plant responses propagate into bulk soil. It is an open question how changes in plant communities and thus their root systems under climate change\textsuperscript{42,43} interact with rhizosphere C dynamics to influence soil C turnover.

At the global scale, our results demonstrate that most root-derived C inputs to soil does not directly contribute to the bulk soil C pool. That is, around 80% of root-derived C inputs leaves the soil directly via various pathways potentially mediated by the rhizosphere. Here, we propose three such pathways (Fig. 5). First, the majority of released C by roots as exudates may be respired as CO\textsubscript{2} by microbes in the rhizosphere\textsuperscript{24} instead of moving to the C pool in bulk soil. Root-released C (in the form of exudates and root detritus) may mainly contribute to the bulk soil C pool via by-products (e.g., necromass) of microbes which utilize root-derived C in the rhizosphere. Indeed, there is evidence that microbial necromass accounts for >50% of soil organic C in temperate cropping, grassland and forest topsoils\textsuperscript{44}. Second, soil fauna feeds
on roots and microbes, and thus removes a fraction of root-derived C\textsuperscript{23,45}. Third, root-derived C may pass directly to other systems beyond the soil layer such as bedrock or groundwater. For example, plant roots may penetrate to the bedrock\textsuperscript{30}, and thus the relevant root-derived C does not contribute to bulk soil C at all. It is also intriguing to note that the loss of root-derived C relative to inputs in each soil layer decreases with soil depth except the top 0-20 cm soil layer (Fig. 5). This may due to decreasing activities of soil microbes and fauna with soil depth; and in the top layer, bulk soil C inputs may include marked amount of C from soil surface\textsuperscript{38}.

Earth system models usually have a litter soil C pool with fast turnover rates\textsuperscript{10,17}, which may conceptually capture the direct C losses. However, the size and turnover rate of this pool have been rarely explicitly tested or verified. The spatially explicit maps for $I_{\text{root}}$, $I_{\text{bulk}}$ and $I_{\text{loss}}$ provide benchmark global layers to force and verify these models and explore the relevant implications for long-term C dynamics. The significant loss of root-derived C inputs provides insights into C sequestration by capturing the “leaked” C, such as through the application of biochar which can stabilize new root-derived C by formatting microaggregates via organo-mineral interactions\textsuperscript{46}. We suggest that the oxidation of fixed C that occurs in the rhizosphere must be considered for reliable soil C predictions and land management to simulate C sequestration.

References

1 Jackson, R. B. \textit{et al.} The ecology of soil carbon: Pools, vulnerabilities, and biotic and abiotic controls. \textit{Annual Review of Ecology, Evolution, and Systematics} \textbf{48}, 419-445 (2017).

2 Regnier, P. \textit{et al.} Anthropogenic perturbation of the carbon fluxes from land to ocean. \textit{Nature Geoscience} \textbf{6}, 597-607 (2013).

3 Crowther, T. W. \textit{et al.} Quantifying global soil carbon losses in response to warming. \textit{Nature} \textbf{540}, 104-108 (2016).
Terrer, C. et al. Nitrogen and phosphorus constrain the CO₂ fertilization of global plant biomass. *Nature Climate Change* **9**, 684-689 (2019).

Luo, Z. K., Feng, W. T., Luo, Y. Q., Baldock, J. & Wang, E. L. Soil organic carbon dynamics jointly controlled by climate, carbon inputs, soil properties and soil carbon fractions. *Global Change Biology* **23**, 4430-4439 (2017).

Jobbágy, E. G. & Jackson, R. B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological Applications* **10**, 423-436 (2000).

Werth, M. & Kuzyakov, Y. C-13 fractionation at the root-microorganisms-soil interface: A review and outlook for partitioning studies. *Soil Biology & Biochemistry* **42**, 1372-1384 (2010).

Balesdent, J. et al. Atmosphere–soil carbon transfer as a function of soil depth. *Nature* **559**, 599-602 (2018).

Hicks Pries, C. E., Castanha, C., Porras, R. C. & Torn, M. S. The whole-soil carbon flux in response to warming. *Science* **355**, 1420-1423 (2017).

He, Y. et al. Radiocarbon constraints imply reduced carbon uptake by soils during the 21st century. *Science* **353**, 1419-1424 (2016).

Luo, Y. et al. Toward more realistic projections of soil carbon dynamics by Earth system models. *Global Biogeochemical Cycles* **30**, 40-56 (2016).

Garnier, E. Resource capture, biomass allocation and growth in herbaceous plants. *Trends in Ecology & Evolution* **6**, 126-131 (1991).

Raich, J. W. & Nadelhoffer, K. J. Belowground carbon allocation in forest ecosystems: Global trends. *Ecology* **70**, 1346-1354 (1989).

Davidson, E. A. et al. Belowground carbon allocation in forests estimated from litterfall and IRGA-based soil respiration measurements. *Agricultural and Forest Meteorology* **113**, 39-51 (2002).
Luo, Z., Wang, G. & Wang, E. Global subsoil organic carbon turnover times dominantly controlled by soil properties rather than climate. *Nature Communications* **10**, 3688 (2019).

Fan, N. *et al.* Apparent ecosystem carbon turnover time: uncertainties and robust features. *Earth System Science Data Discuss* **2020**, 1-25 (2020).

Shi, Z. *et al.* The age distribution of global soil carbon inferred from radiocarbon measurements. *Nature Geoscience*, 555-559 (2020).

Sierra, C. A., Müller, M., Metzler, H., Manzoni, S. & Trumbore, S. E. The muddle of ages, turnover, transit, and residence times in the carbon cycle. *Global Change Biology* **23**, 1763-1773 (2017).

Carvalhais, N. *et al.* Global covariation of carbon turnover times with climate in terrestrial ecosystems. *Nature* **514**, 213-217 (2014).

Braakhekke, M. C. *et al.* The use of radiocarbon to constrain current and future soil organic matter turnover and transport in a temperate forest. *Journal Geophysical Research-Biogeosciences* **119**, 372-391 (2014).

Wu, D. *et al.* Accelerated terrestrial ecosystem carbon turnover and its drivers. *Global Change Biology* **20**, 5052-5062 (2020).

Massalha, H., Korenblum, E., Tholl, D. & Aharoni, A. Small molecules below-ground: the role of specialized metabolites in the rhizosphere. *Plant Journal* **90**, 788-807 (2017).

Osler, G. H. R. & Sommerkorn, M. Toward a complete soil C and N cycle: Incorporating the soil fauna. *Ecology* **88**, 1611-1621 (2007).

Finzi, A. C. *et al.* Rhizosphere processes are quantitatively important components of terrestrial carbon and nutrient cycles. *Global Change Biology* **21**, 2082-2094 (2015).

Drigo, B. *et al.* Shifting carbon flow from roots into associated microbial communities in response to elevated atmospheric CO$_2$. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10938-10942 (2010).
26 Kuzyakov, Y. & Razavi, B. S. Rhizosphere size and shape: Temporal dynamics and spatial
stationarity. *Soil Biol. Biochem.* **135**, 343-360 (2019).

27 Schimel, J. P. & Schaeffer, S. M. Microbial control over carbon cycling in soil. *Front.
Microbiol.* **3**, 11 (2012).

28 Strand, A. E., Pritchard, S. G., McCormack, M. L., Davis, M. A. & Oren, R. Irreconcilable
differences: fine-root life spans and soil carbon persistence. *Science* **319**, 456-458 (2008).

29 Ledo, A. *et al.* Tree size and climatic water deficit control root to shoot ratio in individual
trees globally. *New Phytol.* **217**, 8-11 (2018).

30 Canadell, J. *et al.* Maximum rooting depth of vegetation types at the global scale. *Oecologia*
**108**, 583-595 (1996).

31 McCarthy, H. R. *et al.* Re-assessment of plant carbon dynamics at the Duke free-air CO₂
enrichment site: interactions of atmospheric CO₂ with nitrogen and water availability over
stand development. *New Phytol.* **185**, 514-528 (2010).

32 Wieder, W. R., Cleveland, C. C., Smith, W. K. & Todd-Brown, K. Future productivity and
carbon storage limited by terrestrial nutrient availability. *Nature Geoscience* **8**, 441-444
(2015).

33 Malhi, Y. *et al.* The variation of productivity and its allocation along a tropical elevation
gradient: a whole carbon budget perspective. *New Phytol.* **214**, 1019-1032 (2017).

34 Lawrence, C. R. *et al.* An open-source database for the synthesis of soil radiocarbon data:
International Soil Radiocarbon Database (ISRaD) version 1.0. *Earth Syst. Sci. Data* **12**, 61-
76 (2020).

35 Negron-Juarez, R. I., Koven, C. D., Riley, W. J., Knox, R. G. & Chambers, J. Q. Observed
allocations of productivity and biomass, and turnover times in tropical forests are not
accurately represented in CMIP5 Earth system models. *Environmental Research Letters* **10**
(2015).
36 Todd-Brown, K. E. O. et al. Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences* **10**, 1717-1736 (2013).

37 Bradford, M. A. et al. Managing uncertainty in soil carbon feedbacks to climate change. *Nature Climate Change* **6**, 751-758 (2016).

38 Kaiser, K. & Kalbitz, K. Cycling downwards - dissolved organic matter in soils. *Soil Biol. Biochem.* **52**, 29-32 (2012).

39 Haichar, F. E., Santaella, C., Heulin, T. & Achouak, W. Root exudates mediated interactions belowground. *Soil Biol. Biochem.* **77**, 69-80 (2014).

40 Kuzyakov, Y., Horwath, W. R., Dorodnikov, M. & Blagodatskaya, E. Review and synthesis of the effects of elevated atmospheric CO\(_2\) on soil processes: No changes in pools, but increased fluxes and accelerated cycles. *Soil Biology and Biochemistry* **128**, 66-78 (2019).

41 Jiang, M. et al. The fate of carbon in a mature forest under carbon dioxide enrichment. *Nature* **580**, 227-231 (2020).

42 Lange, M. et al. Plant diversity increases soil microbial activity and soil carbon storage. *Nature Communications* **6**, 8 (2015).

43 Sokol, N. W. & Bradford, M. A. Microbial formation of stable soil carbon is more efficient from belowground than aboveground input. *Nature Geoscience* **12**, 46-53 (2019).

44 Liang, C., Amelung, W., Lehmann, J. & Kästner, M. Quantitative assessment of microbial necromass contribution to soil organic matter. *Global Change Biology* **25**, 3578-3590 (2019).

45 van den Hoogen, J. et al. Soil nematode abundance and functional group composition at a global scale. *Nature* **572**, 194-198 (2019).

46 Weng, Z. et al. Biochar built soil carbon over a decade by stabilizing rhizodeposits. *Nature Climate Change* **7**, 371-376 (2017).
Methods

**Soil profile data acquisition.** Three datasets including net primary productivity (NPP) containing its belowground allocation (hereafter NPP$_{data}$), root biomass (Root$_{data}$) and soil radiocarbon measurement ($\Delta^{14}$C$_{data}$) from 725, 559 and 750 soil profiles, respectively, were used in this study. These measurement sites are distributed across the globe and cover the main biome types including tropical/subtropical forests, tropical/subtropical grasslands/savannas, temperate forests, temperate grasslands, Mediterranean/montane shrublands, boreal forests, tundra, deserts, and croplands (Extended Data Fig. 1). NPP$_{data}$ is a comprehensive database of field observation of above- (ANPP) and belowground NPP (BNPP, i.e., root NPP), which was obtained through a thorough literature search and data synthesis from 54 published peer-reviewed papers (Supplementary references) and the ORNL DAAC NPP data collection (https://daac.ornl.gov/cgi-bin/dataset_lister.pl?p=13). Here we note that most NPP data only include root production but root exudates are difficult to measure and not included, which may result in underestimation of NPP, particularly BNPP.

Root$_{data}$ was compiled by Schenk and Jackson$^{47}$ (publicly obtainable from: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=660). We calculated the depth distribution of root biomass in seven standard layers in the top 200 cm of soil profiles (i.e., 0-20, 20-40, 40-60, 60-80, 80-100, 100-150 and 150-200 cm), using the approach of Luo et al$^{15}$. $\Delta^{14}$C$_{data}$ were obtained up to 7 July 2020 (ref.$^{34}$). It is an open-source database for the synthesis of soil radiocarbon data using ‘ISRaD.getdata’ function in R package ‘ISRaD’. From this dataset, the column ‘lyr_14c_fill_extra’ was used because it merges radiocarbon measurements reported as either $\Delta^{14}$C or “fraction of modern” which was subsequently converted to $\Delta^{14}$C (personal communication with the organizer of the dataset). We focused on the $\Delta^{14}$C in mineral soils from 0 to 200 cm soil depth with clear records of observation year and soil layer (upper and lower depths). In total, we used 3,128 unique measurements of $\Delta^{14}$C from 750 profiles.
Estimation of root-derived C inputs to soil ($I_{\text{root}}$). Using the NPP$_{\text{data}}$ dataset, belowground C allocation, i.e., root-derived C inputs to soil ($I_{\text{root}}$), can be directly estimated as belowground NPP (i.e., BNPP) recorded by NPP$_{\text{data}}$ (Extended Data Fig. 1a). It is a significant challenge to determine the allocation of $I_{\text{root}}$ to different soil layer depths using current technologies. However, it is reasonable to assume that $I_{\text{root}}$ to a specific layer is proportional to root biomass in that layer. So we can infer C inputs to different soil layer depths by combining depth distribution of root biomass and BNPP$^{15}$. The depth distribution of root biomass based on Root$_{\text{data}}$ (Extended Data Fig. 1) was used to estimate the depth distribution of $I_{\text{root}}$. In order to quantify the absolute $I_{\text{root}}$ to each soil layer, ideally we need to know BNPP at the corresponding NPP$_{\text{data}}$ location, but such data are unavailable for Root$_{\text{data}}$ sites. Here, we used a machine learning-based model for BNPP trained by NPP$_{\text{data}}$ dataset to estimate $I_{\text{root}}$ (i.e., BNPP) at Root$_{\text{data}}$ sites (see the section on Machine learning models for $I_{\text{root}}$ and $I_{\text{bulk}}$), and then allocated $I_{\text{root}}$ to seven standard layers (i.e., 0-20, 20-40, 40-60, 60-80, 80-100, 100-150 and 150-200 cm) using the depth distribution of root biomass recorded by Root$_{\text{data}}$ using the same approach as Luo et al$^{15}$. At $\Delta^{14}$C$_{\text{data}}$ locations, we also used the model developed for BNPP at NPP$_{\text{data}}$ sites to predict $I_{\text{root}}$, which is further allocated to different soil layers according to the vertical distribution of $I_{\text{bulk}}$, assuming these two variables (i.e., depth distribution predicted by Root$_{\text{data}}$ and $\Delta^{14}$C$_{\text{data}}$) follow a similar pattern of depth distribution (Fig. 1d, see the section on Machine learning models for $I_{\text{root}}$ and $I_{\text{bulk}}$).

Estimation of actual C inputs to bulk soil ($I_{\text{bulk}}$). Under the assumption of steady state, C inputs to bulk soil ($I_{\text{bulk}}$) can be estimated as bulk soil C stock divided by C age. Using the $\Delta^{14}$C$_{\text{data}}$, a soil radiocarbon model$^{48}$ was adopted to estimate C$_{\text{age}}$:

$$\frac{dA^{\text{soil}}}{dt} = k \cdot A^{\text{input}} - (k + \gamma) \cdot A^{\text{soil}},$$  \hspace{1cm} (1)

(Extended Data Fig. 1).
where $A_{\text{input}}$ is the $^{14}\text{C}/^{12}\text{C}$ ratio of C inputs into bulk soil and assumed to be equal to the $^{14}\text{C}/^{12}\text{C}$ ratio of the atmosphere, $A_{\text{soil}}$ is the $^{14}\text{C}/^{12}\text{C}$ ratio of bulk soil C, $k$ is the decay rate of bulk soil C, and $\gamma$ is the $\beta$-decay rate of $^{14}\text{C}$ and equal to 1/8267 per year. Before the nuclear testing in the 1950s and 1960s, atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio was relatively stable. Under the steady state assumption, $A_{\text{soil}}$ can be estimated by Equation (2), and with $A_{\text{soil}}$ measurements at the steady state, $k$ can be estimated by Equation (3) as follow:

$$A_{\text{soil}} = \frac{k}{k + \gamma} \cdot A_{\text{input}},$$  \hspace{1cm} (2)

$$k = \gamma \cdot \frac{A_{\text{soil}}}{A_{\text{input}} - A_{\text{soil}}},$$  \hspace{1cm} (3)

If $A_{\text{soil}}$ is measured after the nuclear testing, Equation (3) cannot be directly used to estimate $k$ because the atmospheric $^{14}\text{C}/^{12}\text{C}$ ratio is significantly increased. However, the estimated $A_{\text{soil}}$ by Equation (3) provides a starting point as the pre-nuclear test status. Combining Equations (3) and (2), we can start the modelling from any pre-nuclear test time (from 1950 in this study), and estimate $k$ using time-series measurements of $A_{\text{input}}$ (i.e., the $^{14}\text{C}$ activity of the atmosphere in each year after the nuclear testing) and $A_{\text{soil}}$ measurements at any time points after the testing. Once $k$ is obtained, the reciprocal of the decay rate $k$ defines $C_{\text{age}}$. Atmospheric radiocarbon data for each year during the period 1950-2010 were extracted from Hua et al\textsuperscript{49} for the northern and southern hemispheres, separately (Supplementary Fig. 5a). To utilise the soil radiocarbon measurements conducted in more recent years, we interpolated the atmospheric radiocarbon data during the period of 2011-2019 using random forest imputation method in the ‘mice’ function in R 3.6.1 (ref.\textsuperscript{50}). In these databases, however, radiocarbon data are reported as $\Delta^{14}\text{C}$ – the per mille ($\%\text{o}$) deviation from a standard of fixed isotopic composition:

$$\Delta^{14}\text{C} = \left(\frac{A_{\text{sample}}}{A_{\text{standard}}} - 1\right) \cdot 1000.$$  \hspace{1cm} (4)
where $A_{\text{sample}}$ and $A_{\text{standard}}$ are the $^{14}\text{C}/^{12}\text{C}$ ratio of measured samples and a common international standard respectively. Hence, we can estimate $A_{\text{input}}$ and $A_{\text{soil}}$ based on the corresponding reported $\Delta^{14}\text{C}$ as:

$$A_{\text{sample}} = \left( \frac{\Delta^{14}\text{C}}{1000} + 1 \right) \cdot A_{\text{standard}}. \quad (5)$$

Replacing $A_{\text{input}}$ and $A_{\text{soil}}$ in Equations (2-4) with Equation (5), we find that the model is independent of $A_{\text{standard}}$, enabling us to calculate $k$ thus $C_{\text{age}}$ (i.e., $1/k$) using $\Delta^{14}\text{C}$ in the bulk soil and atmosphere. Finally, $I_{\text{bulk}}$ is estimated as:

$$I_{\text{bulk}} = \frac{C_{\text{stock}}}{C_{\text{age}}}. \quad (6)$$

where $C_{\text{stock}}$ is the soil C stock in a typical layer for which $\Delta^{14}\text{C}$ is recorded. For $\Delta^{14}\text{C}_{\text{data}}$ soil profiles soil C stock is not recorded, we retrieved C stock from the WISE30sec database using the coordinates of the soil measurement locations. Using this approach, a total of 3,128 estimates of $I_{\text{bulk}}$ in different soil layer depths across 750 soil profiles were obtained. To facilitate the comparison with $I_{\text{root}}$ at standard soil layers, we harmonized $I_{\text{bulk}}$ at the $\Delta^{14}\text{C}_{\text{data}}$ profiles to above-mentioned seven layers (i.e., 0-20, 20-40, 40-60, 60-80, 80-100, 100-150 and 150-200 cm) using a machine learning model, which on average explained 95% of the variation of $I_{\text{bulk}}$ in different soil depths for the $\Delta^{14}\text{C}_{\text{data}}$ profiles (Supplementary Fig. 8f, see the section on Machine learning models for $I_{\text{root}}$ and $I_{\text{bulk}}$).

**Estimation of direct loss of root-derived C inputs ($I_{\text{loss}}$).** Direct C losses of root-derived C inputs ($I_{\text{loss}}$, i.e., root-derived C that does not enter bulk soil C pool) can be inferred as the difference between $I_{\text{root}}$ and $I_{\text{bulk}}$ (i.e., $I_{\text{root}} - I_{\text{bulk}}$). However, $I_{\text{root}}$ and $I_{\text{bulk}}$ cannot be simultaneously and directly determined using observed data at the sites of the NPP$_{\text{data}}$, Root$_{\text{data}}$, or $\Delta^{14}\text{C}_{\text{data}}$ data sets. We needed to predict $I_{\text{bulk}}$ at NPP$_{\text{data}}$ and Root$_{\text{data}}$ sites and also predict $I_{\text{root}}$.
at $\Delta^{14}\text{C}_{\text{data}}$ and Root$_{\text{data}}$ sites (Extended Data Table 1). We integrated NPP$_{\text{data}}$, $\Delta^{14}\text{C}_{\text{data}}$ and Root$_{\text{data}}$ datasets to estimate I$_{\text{loss}}$ at sites in each dataset and conducted cross-comparisons to verify I$_{\text{loss}}$ estimations as well as quantify uncertainties in I$_{\text{loss}}$ based on different datasets (Extended Data Table 1).

At sites for each dataset, we estimated I$_{\text{root}}$ and I$_{\text{bulk}}$ using the observed data as a priority. If the observed data were unavailable, we used machine learning-based statistical models (see the section on Machine learning models for I$_{\text{root}}$ and I$_{\text{bulk}}$) to predict the missing values. Each estimation of I$_{\text{loss}}$ utilized at least one observed dataset (Extended Data Table 1). As a comparison to BNPP estimation based on NPP$_{\text{data}}$, additionally we included the MODIS NPP$^{52}$ in our estimation of I$_{\text{loss}}$. Specifically, we retrieved NPP data from the MODIS NPP product using the geographical coordinates of each site location for the three observed datasets. To determine the quantity of MODIS NPP allocated belowground (i.e., BNPP), we additionally fit a machine learning-based regression model of f$_{\text{BNPP}}$ (i.e., BNPP/NPP) using the observed NPP and BNPP at NPP$_{\text{data}}$ locations (see the section on Machine learning models for I$_{\text{root}}$ and I$_{\text{bulk}}$). In this way, we obtained six independent estimations of I$_{\text{loss}}$ (Extended Data Table 1), which enabled us to make cross-comparisons and assess uncertainties in I$_{\text{loss}}$. Details of the quantifications of each I$_{\text{loss}}$ estimate are provided in Extended Data Table 1.

Here, it should be noted that the observed NPP is somewhat inconsistent with MODIS NPP for the NPP$_{\text{data}}$ sites (Extended Data Fig. 8). In all biomes assessed, on average MODIS NPP is an underestimation of field observed NPP$_{\text{data}}$ (Extended Data Fig. 8b). This phenomenon may result from the mismatch of spatiotemporal scales of the two datasets. That is, MODIS NPP represents the annual average of a ~1 km$^2$ pixel, while NPP$_{\text{data}}$ is measured at a specific location and time. Further, MODIS NPP has uncertainties$^{53}$.

Drivers of I$_{\text{root}}$ and I$_{\text{bulk}}$. We used linear mixed regression to examine the relationship of a
suite of predicting variables with \( I_{\text{root}} \) and \( I_{\text{bulk}} \). We assume \( I_{\text{root}} \) is associated with soil properties (i.e., 20 soil physiochemical properties), climatic factors (i.e., 19 bioclimatic variables), topographic conditions (i.e., 13 topographical variables; Supplementary Table 1). Here we only considered the whole-profile \( I_{\text{root}} \) due to data availability. For \( I_{\text{bulk}} \), however, apart from these predictor variables, soil layer including top and bottom depths were also included as driving factors. For both \( I_{\text{root}} \) and \( I_{\text{bulk}} \), biome type was treated as a random effect, that is, the coefficients of other predictor variables vary among biome types. The multivariate linear mixed-effects model is expressed as:

\[
y_i = N(X_i \beta, \sigma_y^2), \text{ for } i = 1, ..., n,
\]

and \( \beta \) is further modelled within each biome type:

\[
\beta_j = N(\mu_\alpha, \sigma_\alpha^2), \text{ for } j = \text{each of the 9 biome types},
\]

where \( n \) is the sample size of \( I_{\text{root}} \) and \( I_{\text{bulk}} \), respectively, \( y_i \) the \( i \)th variable of interest (i.e., \( I_{\text{root}} \) and \( I_{\text{bulk}} \)), \( X_i \) the vector of predictor variables; \( \beta \) a vector of coefficients for the predictor variables; \( \mu_\alpha \) the estimation of \( \beta \); \( \sigma_y^2 \) and \( \sigma_\alpha^2 \) the error terms for \( y_i \) and \( \beta_j \), respectively, which follow a normal distribution (\( N \)). Before fitting the linear mixed-effects regression, all variables were standardized to unit variance therefore the absolute magnitude of the coefficients for the predictor variables reflect their relative importance. A principal component analysis (PCA) was applied to eliminate potential correlations in the 20 edaphic variables, 19 climatic variables and 13 topographic variables, respectively. The most important PCs with variances greater than 1 were retained for the regression. The PCA and linear mixed-effect model regression were conducted using the ‘procomp’ function in the ‘stats’ package and the ‘lmer’ function in the ‘arm’ package, respectively, in R 3.6.1 (ref. 50).

Environmental covariates. We obtained 54 global layers of soil, climate, topography and biome type. A total of 20 soil physical and chemical properties was obtained from ISRIC-WISE
soil profile database$^{51}$ with a spatial resolution of 1 km. We obtained 19 climatic attributes with the same resolution as the WISE database from WorldClim$^{56}$, which quantifies biologically meaningful variables using monthly temperature and precipitation. In addition, we derived spatial layers of biome type using the approach of Luo et al$^{15}$. Specifically, we aggregated two land cover maps (i.e., the MODIS land cover map$^{57}$ and the Terrestrial Ecoregions of the World$^{58}$) to generate a map consisting of nine biome types. We also retrieved the soil order data from Global Soil Regions Map database (https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/?cid=nrcs142p2_054013). Finally, we calculated 13 topographic attributes from SRTM-DEM at 90 m resolution (http://srtm.csi.cgiar.org) using ‘elevatr’, ‘spatialEco’ and ‘dynatopmodel’ packages in R 3.6.1 (ref.$^{50}$) and SAGA$^{59}$. More details of these global spatial covariate layers are given in Supplementary Table 1.

**Machine learning models for predicting $I_{\text{root}}$ and $I_{\text{bulk}}$.** To extrapolate $I_{\text{root}}$ and $I_{\text{bulk}}$ spatially, we generated spatial predictive models of these two variables, by treating the edaphic, climatic, topographic attributes, and biome type (i.e., the 54 global layers of environmental covariates) as potential drivers (Supplementary Table 1). Using the geographical coordinates of NPP$_{\text{data}}$ and $\Delta^{14}$C$_{\text{data}}$ locations, we retrieved the above-mentioned environmental covariates. For the $I_{\text{bulk}}$ model, the radiocarbon-derived $I_{\text{bulk}}$ of the 3,128 soil radiocarbon measurements was treated as a dependent variable. For $I_{\text{bulk}}$, soil layers (the top and bottom depths of a layer) were also included as predictors. Using the model for $I_{\text{bulk}}$, we quantified $I_{\text{bulk}}$ at NPP$_{\text{data}}$ and Root$_{\text{data}}$ sites and at the global scale (Extended Data Fig. 5). Because soil layers were treated as independent drivers in the model of $I_{\text{bulk}}$, we could directly predict the depth distribution of $I_{\text{bulk}}$. It is also important to highlight that the depth distribution of $I_{\text{bulk}}$ is comparable with that of root biomass (Fig. 1d). This gave us confidence to use $I_{\text{bulk}}$ depth distribution to infer $I_{\text{root}}$ depth distribution, where this information was unavailable (e.g., at NPP$_{\text{data}}$ and $\Delta^{14}$C$_{\text{data}}$ locations).
Treating the observed BNPP (i.e., $I_{\text{root}}$) and $f_{\text{BNPP}}$ (the fraction of BNPP to total NPP) data at the NPP$_{\text{data}}$ sites as dependent variables, we fitted two machine learning-based models for $I_{\text{root}}$ and $f_{\text{BNPP}}$, respectively. The purpose of fitting the model for $f_{\text{BNPP}}$ is to estimate the allocation of the NPP retrieved from the MODIS NPP product to belowground, and further calculate $I_{\text{root}}$ derived from MODIS NPP (Extended Data Table 1). Here, the 52 environmental covariates (Supplementary Table 1) at the NPP$_{\text{data}}$ sites were used as independent variables. For all these three models (i.e., $I_{\text{bulk}}$, $I_{\text{root}}$ and $f_{\text{BNPP}}$), we trained a suite of machine learning algorithms including random forest (RF), extreme gradient boosting (XGBoost), Cubist, support vector machines (SVM) and multivariate adaptive regression splines (MARS). Before fitting these machine learning models, we first converted the categorical variables (e.g., biome type) to dummy variables. Then, the function ‘findCorrelation’ and ‘findLinearCombo’ in R package ‘caret’ was used to exclude those attributes with high multicollinearities. The remaining variables were further used in model calibration (80% stratified samples) and validation (the remaining 20% stratified samples). It should be noted that, when fitting the model for $I_{\text{bulk}}$ used to harmonize $I_{\text{bulk}}$ at the $\Delta^{14}C_{\text{data}}$ profiles to the seven standard soil layers, all the $I_{\text{bulk}}$ data were used to constrain the model (Supplementary Fig. 8). We also generated an ensemble model using a regression method of principal component analysis (PCA) with R packages ‘pls’ and ‘caret’ based on the prediction outputs from the top three models with the lowest RMSE (rooted mean squared error). Specifically, the better an individual model’s performance is (with a smaller root mean squared error), the more the model’s output contributes to the ensemble model’s final predictions. The results showed that the ensemble model and random forest model can generally better simulate $I_{\text{root}}$ (Extended Data Fig. 2), $I_{\text{bulk}}$ (Extended Data Fig. 3) and $f_{\text{BNPP}}$ (Extended Data Fig. 9) compared with other four algorithms (i.e., XGBoost, SVM, Cubist and MARS). Using these fitted machine learning models (ensemble models) and retrieving the predicting variables at other locations, we predicted $I_{\text{bulk}}$
at Root\textsubscript{data} and NPP\textsubscript{data} sites, and also calculated $I\text{\textsubscript{root}}$ and $f\text{\textsubscript{BNPP}}$ at Root\textsubscript{data} and $\Delta^{14}\text{C}\text{\textsubscript{data}}$ locations.

**Global mapping and prediction uncertainty.** Using the fitted machine learning models, we mapped $I\text{\textsubscript{root}}$ (Extended Data Fig. 4), $I\text{\textsubscript{bulk}}$ (Extended Data Fig. 5) and $I\text{\textsubscript{loss}}$ (i.e., $I\text{\textsubscript{root}} - I\text{\textsubscript{bulk}}$) in the seven standard layers across the globe at the resolution of 30 arc-second (~1 km) grid. Here, we used the random forest model, rather than the ensemble model, due mainly to limited computing resources and the overall strong performance of RF models compared to model ensembles (Extended Data Figs. 2, 3 and 9).

In each grid, the model was run using environmental covariates extracted from the global data layers. Prediction uncertainty was quantified using a Monte Carlo approach proposed for digital soil mapping with machine learning by Viscarra Rossel et al\textsuperscript{60} and Coulston et al\textsuperscript{61}. First, 500 bootstrap samples of soil profiles were randomly drawn with replacement to construct 500 random forest models. For each bootstrap resample, 80\% of the soil profile-specific data (randomly selected without replacement) were used to train a random forest model with 500 individual trees, while the other 20\% were reserved for model testing. For each derived random forest model, a scaling factor $\tau$ for each observation in the hold-out testing data was estimated to approximate prediction uncertainty\textsuperscript{61}:

$$
\tau = \sqrt{\frac{(y - \tilde{y})^2}{\text{var}(\tilde{y})}},
$$

where $y$ is the observed value, $\text{var}(\tilde{y})$ the variance of the predictions of individual trees (i.e., 500 trees) in the random forest model, $\tilde{y}$ the average of $\tilde{y}$, i.e., the prediction of the random forest model. As such, we obtained a vector of $\tau$ with length equal to the number of observations in the hold-out data multiplied by bootstrapping 500 times, which represents the distribution of model prediction errors. Finally, a new random forest model based on all available data was trained for mapping. For mapping in each grid, a variable of interest $\theta$ (e.g.,
I_{bulk} was estimated as:

\[ \theta = \bar{y} \pm \hat{\tau} \cdot sd(\hat{y}), \]  

(10)

where \( \bar{y} \) is the prediction of the new random forest model, \( sd(\hat{y}) \) the standard deviation of the predictions of individual trees in the new random forest model, \( \hat{\tau} \) the scaling factor for desired prediction interval width estimated based on the derived distribution of model prediction errors.

Using a Monte Carlo approach, we took the 95th percentile of the distribution as \( \tau_{95} \) to quantify the 95% prediction interval\(^{61}\). In each grid, uncertainty of \( I_{root} \) and \( I_{bulk} \) was standardized as the absolute value of corrected margin of the deviation [i.e., \( \tau_{95} \cdot sd(\hat{y}) \)] divided by the mean (i.e., \( \bar{y} \)). \( I_{loss} \) for each grid was estimated by \( I_{root} - I_{bulk} \) with their uncertainties propagated by conducting 500 Monte Carlo simulations. The median and average of \( I_{loss} \) were estimated based on the 500 simulations, and the uncertainty of \( I_{loss} \) was estimated as the margin of deviation of the 500 simulations divided by their mean.

**Data and code availability**

All data supporting the findings of this study are available on https://figshare.com/articles/dataset/Datasets_of_NPP_soil_radiocarbon_and_root_biomass/12840050. Code used to generate the results is available from the corresponding author (luozk@zju.edu.cn).

**References**

47 Jochen, S. H. & B., J. R. The global biogeography of roots. *Ecological Monographs* 72, 311-328 (2002).

48 Cherkinsky, A. & Brovkin, V. Dynamics of radiocarbon in soils. *Radiocarbon* 35, 363-367 (1993).
Hua, Q., Barbetti, M. & Rakowski, A. Z. Atmospheric radiocarbon for the period 1950–2010. *Radiocarbon* **55**, 2059-2072 (2013).

50 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (2020).

51 Batjes, N. H. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma* **269**, 61-68 (2016).

52 Zhao, M. & Running, S. W. Drought-induced reduction in global terrestrial net primary production from 2000 through 2009. *Science* **329**, 940-943 (2010).

53 Zhao, M., Heinsch, F. A., Nemani, R. R. & Running, S. W. Improvements of the MODIS terrestrial gross and net primary production global data set. *Remote Sensing of Environment* **95**, 164-176 (2005).

54 Gelman, A. & Hill, J. *Data analysis using regression and multilevel/hierarchical models*. (Cambridge University Press, 2006).

55 Kaiser, H. F. The application of electronic computers to factor analysis. *Educational and Psychological Measurement* **20**, 141-151 (1960).

56 Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* (2017).

57 Channan, S., Collins, K. & Emanuel, W. Global mosaics of the standard MODIS land cover type data. *University of Maryland and the Pacific Northwest National Laboratory, College Park, Maryland, USA* **30** (2014).

58 Olson, D. M. *et al*. Terrestrial ecoregions of the world: a new map of life on earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *Bioscience* **51**, 933-938 (2001).

59 Conrad, O. *et al*. System for automated geoscientific analyses (SAGA) v. 2.1.4. *Geosci. Model Dev.* **8**, 1991-2007 (2015).
Viscarra Rossel, R. A., Webster, R., Bui, E. N. & Baldock, J. A. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. *Global Change Biology* **20**, 2953-2970 (2014).

Coulston, J. W., Blinn, C. E., Thomas, V. A. & Wynne, R. H. Approximating prediction uncertainty for random forest regression models. *Photogrammetric Engineering & Remote Sensing* **82**, 189-197 (2016).

**Authors contributions** Z.L. conceived the study; G.W., S.C., Z.L. compiled the data; G.W., L.X. and Z.L. led the data assessment with the contributions of X.M., X.G., S.Z. and M.W.; L.X. conducted global mapping; G.W. and Z.L. interpreted the results with the contribution of A.C., G.Z., Z.S., S.Z and M.W.; Z.L. led manuscript writing with substantial contributions of G.W. and A.C.

**Competing interests** The authors declare no competing interests.
Fig. 1. Root-derived carbon inputs (I$_{\text{root}}$) and actual carbon inputs to bulk soil (I$_{\text{bulk}}$). a: the probability density distribution of I$_{\text{root}}$ (blue color) and I$_{\text{bulk}}$ (green color); b: the depth distribution of I$_{\text{root}}$ and I$_{\text{bulk}}$; c: the probability density distribution of the percentage fraction of I$_{\text{root}}$ and I$_{\text{bulk}}$ relative to net primary production (NPP); d: the percentage depth distribution of I$_{\text{root}}$ and I$_{\text{bulk}}$ relative to total I$_{\text{root}}$ and I$_{\text{bulk}}$, respectively. Inset tables in a and c show the median, mean and sample size across the globe and in nine biome types (1: tropical/subtropical forests, 2: tropical/subtropical grasslands/savannas, 3: temperate forests, 4: temperate grasslands, 5: Mediterranean/montane shrublands, 6: boreal forests, 7: tundra, 8: deserts, and 9: croplands), and different lowercase and capital letters indicate significant ($P < 0.05$) difference between I$_{\text{root}}$ and I$_{\text{bulk}}$ and between biomes, respectively. Boxplots (b and d) show the median and interquartile range with whiskers extending to 1.5 times of the interquartile range, and red dots show averages.
Fig. 2. Drivers of root derived carbon input ($I_{\text{root}}$) and actual carbon input to bulk soil ($I_{\text{bulk}}$). Color grids show the coefficients and intercepts of linear mixed-effect regression models for $I_{\text{root}}$ (up panel) and $I_{\text{bulk}}$ (bottom panel) treating biome type as a random effect. Blank grids indicate that the predictor variable was not included; grey grids indicate that the effect is insignificant (i.e., $p>0.05$). Biome 1-9 show the nine biome types (1: tropical/subtropical forests, 2: tropical/subtropical grasslands/savannas, 3: temperate forests, 4: temperate grasslands, 5: Mediterranean/montane shrublands, 6: boreal forests, 7: tundra, 8: deserts, and 9: croplands). PC1-5 are the most important principal components (PCs) of different groups of driving factors (i.e., soil, climate, topography and depth). Top and Bottom show the top and bottom depths of soil layer, which are treated as driving factors for $I_{\text{bulk}}$. The $R^2$ for the models of $I_{\text{root}}$ and $I_{\text{bulk}}$ are 0.50 and 0.74, respectively. Detailed principal component analyses on the predictor variables are presented in Supplementary Figs. 3 and Fig. 4.
Fig. 3. Direct soil carbon loss ($I_{loss}$, which is estimated as $I_{root} - I_{bulk}$) without contributing to bulk soil carbon pool. **a**, the probability density distribution of $I_{loss}$ quantified using six independent data sources and/or methods ($I_{loss 1-6}$), which are described in Extended Data Table 1; **b**, depth distribution of $I_{loss}$; **c**, the probability density distribution of the percentage fraction of $I_{loss}$ relative to $I_{root}$; **d**, the percentage $I_{loss}$ in each soil layer relative to total $I_{loss}$. Inset tables in **a** and **c** show the median, mean and sample size across the globe and in nine biome types. Different capital letters indicate significant ($P < 0.05$) difference between biomes. See Fig. 1 for details of the 9 biome types. Boxplots (**b** and **d**) show the median and interquartile range with whiskers extending to 1.5 times of interquartile range, and red dots show averages.
Fig. 4. Global pattern of direct soil carbon loss (i.e., $I_{loss}$) in the 0-200 cm soil profile. a, $I_{loss}$; b, uncertainty of $I_{loss}$ [in terms of the margin of the deviation of 500 Monte Carlo simulations divided by their mean, taking into account uncertainties in both observational data sets and prediction models]; c, the proportion of $I_{loss}$ relative to root-derived C inputs ($I_{root}$); d, uncertainty of the proportion of $I_{loss}$ relative to $I_{root}$.
Fig. 5. Schematic representation of fluxes of root-derived carbon inputs. The majority of root-derived carbon inputs is lost directly via three potential pathways: rhizosphere microbial consumption, fauna feeding and leaching; and does not contribute to bulk soil carbon pool.
Extended Data Table 1. Estimation of root-derived carbon inputs ($I_{\text{root}}$), carbon inputs to bulk soil ($I_{\text{bulk}}$) and direct carbon loss of $I_{\text{root}}$ ($I_{\text{loss}}$) in soil profiles at NPP$_{\text{data}}$, $\Delta^{14}$C$_{\text{data}}$ and Root$_{\text{data}}$ sites, respectively. BNPP is the belowground net primary production (NPP), $f_{\text{BNPP}}$ is the fraction of BNPP to NPP.

| Sites         | Variables to calculate $I_{\text{root}}$ (Mg ha$^{-1}$ yr$^{-1}$) | $I_{\text{root}}$ (Mg ha$^{-1}$ yr$^{-1}$) | $I_{\text{bulk}}$ (Mg ha$^{-1}$ yr$^{-1}$) | $I_{\text{loss}}$ (Mg ha$^{-1}$ yr$^{-1}$) |
|--------------|---------------------------------------------------------------|------------------------------------------|------------------------------------------|------------------------------------------|
| NPP$_{\text{data}}$ | $\mathbf{1}^{\text{OBS}}$: observed BNPP | $\mathbf{1} = \mathbf{1}^{\text{OBS}} \times \mathbf{2}^{\text{PRE}}$ | $\mathbf{1}^{\text{PRE}}$: predicted using $\Delta^{14}$C$_{\text{data}}$-derived models | $\mathbf{1} = \mathbf{1}^{\text{PRE}}$ ($I_{\text{loss1}}$) |
|              | $\mathbf{2}^{\text{PRE}}$: predicted BNPP depth distribution using $\Delta^{14}$C$_{\text{data}}$-derived models | $\mathbf{2} = \mathbf{5}^{\text{PRE}} \times \mathbf{4}^{\text{OBS}} \times \mathbf{2}^{\text{PRE}}$ | $\mathbf{2}^{\text{PRE}}$: predicted using $\Delta^{14}$C$_{\text{data}}$-derived models | $\mathbf{2} = \mathbf{1}^{\text{PRE}}$ ($I_{\text{loss2}}$) |
|              | $\mathbf{3}^{\text{MOD}}$: MODIS NPP | | | |
|              | $\mathbf{4}^{\text{OBS}}$: observed $f_{\text{BNPP}}$ | | | |
| $\Delta^{14}$C$_{\text{data}}$ | $\mathbf{5}^{\text{PRE}}$: predicted BNPP using NPP$_{\text{data}}$-derived models | $\mathbf{3} = \mathbf{5}^{\text{PRE}} \times \mathbf{6}^{\text{OBS}}$ | $\mathbf{2}^{\text{OBS}}$: observed | $\mathbf{3} = \mathbf{2}^{\text{OBS}}$ ($I_{\text{loss3}}$) |
|              | $\mathbf{6}^{\text{OBS}}$: observed BNPP depth distribution | $\mathbf{4} = \mathbf{7}^{\text{MOD}} \times \mathbf{8}^{\text{PRE}} \times \mathbf{6}^{\text{OBS}}$ | | $\mathbf{4} = \mathbf{2}^{\text{OBS}}$ ($I_{\text{loss4}}$) |
|              | $\mathbf{7}^{\text{MOD}}$: MODIS NPP | | | |
|              | $\mathbf{8}^{\text{PRE}}$: predicted $f_{\text{BNPP}}$ using NPP$_{\text{data}}$-derived models | | | |
| Root$_{\text{data}}$ | $\mathbf{9}^{\text{PRE}}$: predicted BNPP using NPP$_{\text{data}}$-derived models | $\mathbf{5} = \mathbf{9}^{\text{PRE}} \times \mathbf{10}^{\text{OBS}}$ | $\mathbf{3}^{\text{PRE}}$: predicted using $\Delta^{14}$C$_{\text{data}}$-derived models | $\mathbf{5} = \mathbf{3}^{\text{PRE}}$ ($I_{\text{loss5}}$) |
|              | $\mathbf{10}^{\text{OBS}}$: observed BNPP depth distribution | $\mathbf{6} = \mathbf{1}^{\text{MOD}} \times \mathbf{12}^{\text{PRE}} \times \mathbf{10}^{\text{OBS}}$ | $\mathbf{6}^{\text{PRE}}$: predicted using $\Delta^{14}$C$_{\text{data}}$-derived models | $\mathbf{6} = \mathbf{3}^{\text{PRE}}$ ($I_{\text{loss6}}$) |
|              | $\mathbf{11}^{\text{MOD}}$: MODIS NPP | | | |
|              | $\mathbf{12}^{\text{PRE}}$: predicted $f_{\text{BNPP}}$ using NPP$_{\text{data}}$-derived models | | | |
Extended Data Fig. 1. Location distribution of measurements of above- and belowground NPP (NPP_{data}), root biomass distribution (Root_{data}) and soil radiocarbon content (\Delta^{14}C_{data}). Values in parentheses show the number of profiles for each of the three data sets.
Extended Data Fig. 2. Performance of fitted machine learning models to predict $I_{\text{root}}$. Data are log-transformed. a, random forest (RF); b, extreme gradient boosting (XGBoost); c, Cubist; d, support vector machines (SVM); e, multivariate adaptive regression splines (MARS); f, the ensemble model of a-e. For each individual model, 80% of the stratified samples of observations were used for model calibration, with the other 20% for validation.
Extended Data Fig. 3. Performance of fitted machine learning models to predict $I_{\text{bulk}}$. Data are log-transformed. a, random forest (RF); b, extreme gradient boosting (XGBoost); c, Cubist; d, support vector machines (SVM); e, multivariate adaptive regression splines (MARS); f, the ensemble model of a-e. For each individual model, 80% of the stratified samples of observations were used for model calibration, with the other 20% stratified data for validation.
Extended Data Fig. 4. Global pattern of root-derived C inputs (i.e., $I_{\text{root}}$, Mg ha$^{-1}$ yr$^{-1}$) in seven standard soil layers and in the whole soil profile (0-200 cm).
Extended Data Fig. 5. Global pattern of actual carbon inputs to bulk soil (i.e., $I_{\text{bulk}}$, Mg ha$^{-1}$ yr$^{-1}$) in seven standard soil layers and in the whole soil profile (0-200 cm).
Extended Data Fig. 6. Global pattern of direct loss of root-derived C inputs before entering bulk soil carbon pool (i.e., $I_{\text{loss}} = I_{\text{root}} - I_{\text{bulk}}$, Mg ha$^{-1}$ yr$^{-1}$).
Extended Data Fig. 7. Global pattern of the uncertainty of I_{loss} in different soil layers.

Uncertainty is estimated as the margin of the deviation of 500 Monte Carlo simulations divided by their mean, considering uncertainties in both observational data sets and prediction models.
Extended Data Fig. 8. Comparison between observed net primary production (NPP) and MODIS-retrieved NPP at NPP data sites. a: site-by-site comparison; b: comparison by aggregating data into nine biomes. Dashed lines show the 1:1 line.
Extended Data Fig. 9. Performance of fitted machine learning models to predict the belowground fraction of NPP (i.e., $f_{BNPP}$). a, random forest (RF); b, extreme gradient boosting (XGBoost); c, Cubist; d, support vector machines (SVM); e, multivariate adaptive regression splines (MARS); f, the ensemble model of a-e. For each individual model, 80% of the stratified samples of observations were used for model calibration, with the other 20% for validation.
Supplementary information:

Supplementary Data 1: The observed above- and belowground NPP data at the 725 NPP\textsubscript{data} sites (Extended Data Fig. 1).

Supplementary Data 2: The depth distribution of root biomass at the 559 Root\textsubscript{data} sites (Extended Data Fig. 1).

Supplementary Data 3: Estimated carbon age (C\textsubscript{age}) and I\textsubscript{bulk} at the 750 Δ\textsuperscript{14}C\textsubscript{data} sites (Extended Data Fig. 1).

Supplementary Data 4: The 54 studies from which the NPP\textsubscript{data} were derived.

Supplementary Tables 1-3

Supplementary Figures 1-8