Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

Geoscientific Model Development Discussions

Data-mining analysis of factors affecting the global distribution of soil carbon in observational databases and Earth system models

Shoji Hashimoto[1], Kazuki Nanko[1], Boris Ťupek[2], Aleksi Lehtonen[2]

[1]Forestry and Forest Products Research Institute (FFPRI), Tsukuba, 305-8687, Japan
[2]Natural Resources Institute Finland, P.O. Box 18, Vantaa, 01301, Finland

Correspondence to: S. Hashimoto (shojih@ffpri.affrc.go.jp)

Abstract. Soil carbon dynamics are a key process in the terrestrial carbon cycle. Future climate change will dramatically change the carbon balance in soil, and this change will affect the terrestrial carbon stock and the climate itself. Earth system models (ESMs) are used to understand the current climate and to produce future climate projections, but the soil organic carbon (SOC) stock simulated by ESMs and those of observational databases are not well correlated when the two are compared at fine grid scales. However, the specific key processes and factors, as well as the relationships among factors, that govern the SOC stock, remain unclear, and the inclusion of such missing information would improve the agreement between modelled and observational data. In this study, we aimed to identify the influential factors that govern global SOC distribution in observational databases, as well as those simulated by ESMs. We used a data-mining (machine-learning) scheme (boosted regression trees: BRT) to reveal the factors affecting the SOC stock. We applied BRT to three observational databases and 15 ESM outputs from the fifth phase of the Coupled Model Intercomparison Project (CMIP5) and examined the effects of 13 variables/factors categorized into five groups (climate, soil property, topography, vegetation, and land-use history). These analyses revealed the influential variables and their correlations with SOC. Globally, the contributions of mean annual temperature, clay content, CN ratio, wetland ratio, and land cover were high in observational databases, whereas the contribution of mean annual temperature, land cover, and NPP governed the SOC distribution in ESMs. A comparison of the influential factors in observational databases and ESMs, at the global scale, revealed that the CN ratio and clay content were key processes to include in ESMs to reproduce the distribution of SOC in observational databases. The results of this study will help elucidate the nature of both observational SOC databases and ESM outputs and improve the modelling of terrestrial carbon dynamics with ESMs. This study shows that a data-mining algorithm can be used to assess model outputs.

1 Introduction

Soil is the largest organic carbon stock in terrestrial ecosystems (Batjes, 1996; IPCC, 2013; Köchy et al., 2015). The soil organic carbon (SOC) stock is the result of the balance between carbon input into soil and decomposition, and the soil carbon influx and efflux are controlled directly and indirectly by environmental conditions (Carvalhais et al., 2014; Schimel et al., 1994). Future climate change will dramatically affect the carbon balance in the soil cycle (Bond-Lamberty and Thomson, 2010; Friedlingstein et al., 2006; Hashimoto et al., 2011, 2015), and this change will affect the terrestrial carbon and, consequently, the climate itself (Cox et al., 2000; Zaehle, 2013).

In the last two decades, several global soil databases have been developed, and some are under further improvement (Scharlemann et al., 2014). These databases describe the global distribution of soil physiochemical properties, enabling us to calculate the global distribution of the SOC stock (e.g., Harmonized World Soil Database), and some databases provide the SOC stock by default (e.g., IGBP-DIS). These databases are based on observed data points with global coverage, although there are biases in the spatial distribution or densities of the data points.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

Earth system models (ESMs), which have been developed to understand the current climate and to provide future climate projections, incorporate the terrestrial carbon cycle, including SOC. The above-mentioned observational global soil databases are often used as benchmarks and to examine whether the ESMs successfully describe the global distribution of the soil carbon stock (Hararuk et al., 2014; Todd-Brown et al., 2013; Wieder et al., 2014). However, a recent study observed that the results of ESMs were moderately consistent at the biome levels, whereas the correlation between the distribution of soil carbon stock simulated by ESMs and that of observational databases was poor when the two were compared at fine scales (e.g., a 1° scale). In addition, estimates of SOC by ESMs and terrestrial biosphere models exhibit high uncertainty (Nishina et al., 2015; Tian et al., 2015). Although some potential factors (e.g., net primary production or temperature) have been suggested (Exbrayat et al., 2013; Nishina et al., 2014; Todd-Brown et al., 2013; Wieder et al., 2013), the key processes and factors, as well as the relationships among factors, that govern the SOC stock, and whose inclusion would improve the agreement between model and observational data, remain unclear.

In this study, we aimed to identify the key factors that govern the global SOC distribution in observational databases as well as those simulated by ESMs. We applied a data-mining (machine-learning) scheme (boosted regression trees: BRT) to identify the influential factors and how these factors relate to SOC stock (Elith et al., 2008). BRT is a method based on regression trees and boosting. We examined the influential factors on the distribution of SOC and their relationships between these factors and SOC stock. We assessed how the ESMs could match the influential factors and their relationships with factors from observational databases. By comparing the influential factors in observational databases with those in ESMs, we clarified the model-data discrepancies and the areas in which ESMs can be improved.

2 Materials and Methods

2.1 Observational global SOC database

We used SOC data from two global and one northern observational databases. The first global database is the Harmonized World Soil Database (HWSD) (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012). The HWSD is a global database of soil physiochemical properties. We used SOC stock database obtained using HWSD by Joint Research Centre (JRC) (Hiederer and Köchy, 2011) (Fig. 1a). The second database was the global gridded surfaces of selected soil characteristics (IGBP-DIS) (Global Soil Data Task Group, 2000) (Fig. 1b), which contains gridded soil physiochemical properties. The third database was the Northern Circumpolar Soil Carbon Database version 2 (NCSCD) (Hugelius et al., 2013; Tarnocai et al., 2009) (Fig. 1c). This database is a spatial database of SOC stock of the northern circumpolar permafrost region.

We used the HWSD and IGBP-DIS to analyse the global distribution of SOC stock, and then we extracted the database of northern circumpolar regions from the three above databases and analysed the SOC stocks in the northern region. The relationships between the databases are shown in Fig. S1. The SOC in the upper 100 cm in each database was used.

2.2 Global SOC estimated using Earth system models

The global distribution of SOC stock estimated by ESMs was obtained from the fifth phase of the Coupled Model Intercomparison Project (CMIP5). We examined the results of 15 ESMs (Fig. 2) (Appendix Table A1). When more than one result was obtained by the same model family (e.g., MIROC-ESM and MIROC-ESM-CHEM), we generated an ensemble average database for each family (e.g. average of MIROC-ESM and MIROC-ESM-CHEM). The mean values for 1980–2004 were calculated. The results of the historical and ensemble member r1i1p1 were used in this study. The notation "r1i1p1" is an identifier of model simulation and is an ensemble member that is often used for analyses (e.g. Todd-Brown et al., 2013). The overviews of SOC submodels in the ESMs have been previously described (Exbrayat et al., 2014; Todd-Brown et al., 2013). In general, each soil submodel consisted of 1 to 9 pools and incorporated the effects of temperature and moisture.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

Some ESMs have litter carbon pools, and those were excluded from this study. The comparisons between the mean of ESMs and global observational databases in a 1° grid are shown in Fig. S2.

2.3 Other databases

We used five groups of variables/factors to examine their effects on global SOC: climate, soil property, topography, vegetation, and land-use history. Detailed data sources for the databases are described in Table 1. For climate and soil property, the mean annual temperature and annual precipitation, and the clay content, CN ratio, and texture (Appendix Table A2) were used (0–30 cm), respectively. As topographical indices, the compound topographic index, elevation, slope, and wetland ratio were used. The CN ratio was calculated by dividing the carbon density by the nitrogen density. The wetland ratio was calculated by dividing the number of wetland grids with the total grids at 1°. Lake, reservoir, and river were not quantified as wetlands and were excluded from the total grids. The land cover type (Appendix Table A3) and net primary production (NPP) were adopted for vegetation indices, and the cropland ratio and human appropriation of net primary production percentage, which is a percentage of human consumption of NPP to local NPP (Imhoff and Bounoua, 2006), were used as indices of land-use history. The average human appropriation of NPP percentage was calculated at 1°. Histograms of the variables are shown in Fig. S3.

2.4 Database handling

All global databases, except for the databases of a spatial resolution of 1° by default, including observational and ESM model outputs, were regridded to a spatial resolution of 1° for the analyses. Regridding of data in the NetCDF format was performed using the Climate Data Operators (CDO) software provided by the Max Plank Institute for Meteorology (https://code.zmaw.de/projects/cdo).

2.5 Boosted regression trees (BRT)

To identify the influential factors and their relationships with SOC stock, BRT were used in this study (Elith et al., 2008). This technique involves a type of data-mining (machine-learning) algorithm that combines the advantages of a regression tree (decision tree) algorithm and boosting. Regression trees are a classification algorithm that classifies data by recursive binary splits, and boosting is a machine-learning algorithm that generates many rough models and combines them to improve their predictive capability. The main advantages of the method are that BRT can analyse different types of variables and interaction effects between variables and are applicable to nonlinear relationships. In recent years, the BRT technique has been used to examine the distribution of soil characteristics at the regional scale (Aertsen et al., 2011; Cools et al., 2014; Martin et al., 2011). Major outputs from the BRT analyses identify (1) the relative importance (percentage of influence or contribution) of predictor variables (explanatory variables), which is based on the weighted and scaled number of times a variable is selected for splitting (Elith et al., 2008), and (2) relationships between variables and the explained variable shown in partial dependence plots.

We used the open-source BRT package in R software (R Core team, 2013) developed by Elith et al. (2008). In practice, three parameters in the BRT package—learning rate (*lr*), tree complexity (*tc*), and bag fraction (*bg*)—control the BRT performance. The *lr* determines the contribution of each tree, the *tc* controls the number of splits, and the *bg* is the proportion of data selected in each step. The number of trees was determined using the cross-validation method in the R package. The maximum number of trees was set to 15,000. The *tc* value was set to 5. We tested different *lr* (0.001, 0.005, 0.01, 0.05, 0.1) and *bg* values (0.5, 0.6, 0.7) and used the best parameter set for each database, but the changes in parameter values had little effect on the model performance.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

2.6 Model performance

The goodness of fit between the BRT model and data was assessed using the linear relationship between the predicted and observed values, the coefficient of determination (R^2), and the root mean square error (RMSE) and is shown in Tables S1 and S2. For both the observational databases and ESMs databases, the BRT models showed good performance with high R^2 values in most of the databases, but the performance was relatively lower for NCSCD and CMCC (northern soils).

3 Results

3.1 Observational databases

3.1.1 Global soil

The relative contributions of variables in the BRT model of global SOC stock to the observational databases are shown in Fig. 3a. In HWSD, the contributions of land cover and mean annual temperature, CN ratio, and wetland ratio were high. For IGBP-DIS, the mean annual temperature, followed by clay content, CN ratio, land cover, also greatly contributed. In particular, the mean annual temperature was very important. The contribution of elevation for each HWSD and IGBP-DIS was 6% and 7%, respectively. The NPP contributed 5% in both databases.

The relationships between the influential variables and SOC are shown in Fig. 4. In general, the two databases showed similar relationships. For example, SOC decreased with an increasing mean annual temperature, particularly for sites with a mean annual temperature > 0 °C, but increased with increasing clay content and CN ratio. The SOC increased rapidly with an increasing CN ratio. Relationships with a mean annual temperature were relatively close to each other. The relationship with clay was steeper in IGBP-DIS than in HWSD, but the opposite was true for the CN ratio. With respect to land cover, evergreen needleleaf forest and permanent wetlands had higher SOC.

3.1.2 Northern soils

In the northern region, the dominant contributors differed among northern soil databases and from those identified in the global database analyses described above (Fig. 3b). For HWSD, the CN ratio was the dominant contributor, followed by the wetland ratio, clay content, and mean annual precipitation. In IGBP-DIS, clay content, CN ratio, and elevation were the most important contributors. For NSCD, elevation contributed the most (~25%), but all variables except for cropland ratio and HANPPpct contributed 5–15%. The mean annual temperature was not as influential as the global databases.

The relationships between variables and SOC stock varied more among the databases for northern soils than those of global databases (Fig. 5). In addition, because the northern regions were extracted, the ranges of variables were narrower than the global databases. In NCSCD, the SOC decreased with increasing temperature and increased with increasing precipitation. The SOC increased with increasing clay content and CN ratio in HWSD and IGBP-DIS, which was consistent with the findings obtained from the global databases. The increasing trend with increasing CN ratio was also observed in NCSCD. The SOC decreased with increasing elevation in all databases but showed considerable variability at low elevations.

3.2 Earth system models

3.2.1 Global soil

The contributions of each variable among ESMs highly varied, but the mean of the results of the ESMs showed that the mean annual temperature, land cover, and NPP distinctively contributed to SOC distribution, and large inconsistencies between the observational databases and ESMs demonstrated low contributions of clay content and CN ratio and high contributions of NPP in ESMs (Fig. 6a). The contribution of NPP in ESMs was greater than in the observational databases.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

The relationships between the SOC and variables in ESMs as well as the results of the observational databases are shown in Fig. 7. The relationships between SOC and certain variables largely varied among the ESMs databases, particularly for the mean annual temperature. The SOC decreased with increasing mean annual temperature but increased with increasing precipitation and NPP. The mean of the relationship with mean annual temperature for ESMs was very consistent with that in the HWSD and IGBP-DIS databases at temperature range −5–15 °C. The increasing trend with increasing NPP in ESMs was consistent with that of the HWSD, particularly below approximately 500 g C m^{-2} of NPP. Although the wetland ratio did not contribute to the ESMs (Fig. 6a), with respect to land cover, permanent wetlands had higher SOC (Fig. 7d).

3.2.2 Northern soils

The mean of the ESMs showed that for northern soils, the main contributors (mean annual temperature, land cover, and NPP) were mainly the same as in the ESMs' global outputs (Fig. 6b). The contribution of the mean annual temperature was lower than that for the global results of ESMs (mean of 14% for the northern and 29% for the global temperatures). The relatively large discrepancy between observational databases and ESMs included the lower contribution of clay content, CN ratio, and elevation and the higher contribution of mean annual temperature, land cover, and NPP in the ESMs.

The relationships between SOC and variables in ESMs as well as the results of the observational databases are shown in Fig. 8. The mean of ESMs showed that the SOC in the northern region increased with increasing NPP, and the relationship was similar to that in HWSD, although the contribution of NPP in ESMs differed from those of the observational database. The decreasing trend with elevation was not replicated in the ESMs.

4 Discussion and concluding remarks

Using the data-mining technique, our BRT analyses revealed influential variables for global and northern SOC in the observational databases and the output of ESMs. The influential factors differed between observational databases and between the global and northern databases. Analyses of the ESMs' output showed large variability, but the influential factors were predominantly similar among ESMs (Fig. 6). This similarity most likely indicates that the structures of the models that describe SOC dynamics in the ESMs are similar. By comparing the results from the ESMs' output to those of observational databases, we identified differences between the observational databases and the output of ESM, which suggest directions for the future improvement of SOC sub-models of ESMs.

Our analyses revealed that the most distinct differences between the observational databases and the outputs of ESMs were the effects of the CN ratio and clay contents (Fig. 6). For both the global observational databases, the CN ratio had substantial contributions (Fig. 3a). The important contribution of the CN ratio was the same in the northern databases (Fig. 3b). The SOC increased with increasing CN ratio in the observational databases (Fig. 4), whereas the outputs of the ESM were insensitive to the CN ratio. The relationships between the CN ratio and SOC may show "causal-resultant" relationship. The decomposability of organic matter generally decreases with increasing CN ratio (Berg et al. 2001; Zhang et al. 2008). In addition, when undecomposed organic matter accumulates, the CN ratio increases, as supported by the fact that soils with higher SOC have a higher CN ratio (Batjes, 1996). The high contribution of the CN ratio may suggest the importance of soil fertility (or nutrient availability) and plant litter quality in the carbon cycle and SOC accumulation (Cotrufo et al., 2013; Fernández-Martínez et al., 2014; Liski et al., 2005; Tuomi et al., 2009; Ťupek et al., 2016). All of the ESMs except for the BCC, CESM1, and NorESM in CMIP5 do not have terrestrial nitrogen processes (Todd-Brown et al., 2013). Including the nitrogen process has been suggested as an important improvement for the next model intercomparison (CMIP6) (Hajima et al., 2014; Zaehle et al., 2015). The results derived from our analysis support the importance of the appropriate inclusion of the N cycle in ESM models.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

Clay content is also often used as a regulator of the decomposability of organic matter in soil (e.g., CENTURY and RothC). Generally, high clay content inhibits organic matter decomposition in the soil. In addition, high clay content often results in low drainage and anaerobic soil conditions, which also inhibits organic matter decomposition. For IGBP-DIS, clay content had as high a contribution as the CN ratio. The control of decomposability by clay content has been previously incorporated in site-scale process-based models (Parton et al., 1987). Including the influence of clay on decomposition will be needed in ESMs.

The mean annual temperature was identified as an influential factor in global databases (Fig. 3a) but not in northern soils (Fig. 3b). Temperature is a main controller of both plant production (source of carbon input to soil) and the decomposition of soil organic matter, which are already incorporated in ESMs. The temperature sensitivities, Q_{10} values, of soil organic matter decomposition in ESMs were reported to be 1.4 to 2.2 (Todd-Brown et al., 2014), and our analyses showed the diverse relationships between the mean annual temperature and SOC. The lower contribution of mean annual temperature in northern soils most likely exists because temperature sensitivity is an exponential process and the magnitude of changes with changing temperature is relatively small in a low temperature range. The use of temperature sensitivity that captures the observational data will improve the performance of ESMs. The relationships between SOC and temperature obtained in this study are the integration of temperature sensitivity of both plant production and soil organic decomposition and thus do not provide the sensitivity of individual processes for ESMs. However, the results of this study can be used to examine the consistency between the ESM output and observational databases.

In ESMs, NPP was selected as an influential factor in ESM analyses for global and northern SOC (Fig. 6ab) but not in observational databases (Fig. 3ab), which is consistent with findings obtained in a previous study (Todd-Brown et al., 2013). This high NPP contribution in ESMs is understandable because in the terrestrial carbon balance modelled in ESMs, the SOC stock is calculated through NPP or plant litter input to soil and soil organic matter decomposition. Plant litter input is proportional to NPP. However, our analyses suggest that the influence of NPP on soil organic matter in observational soil databases was obscured by other factors. When ESMs incorporate the effects of other factors, for example, the CN ratio and clay contents, the effect of NPP may be diluted in ESMs. Furthermore, SOC storage results from organic matter accumulation over decades and even millennia. Thus, past NPP, land fires, and land-use change may still have an effect on current SOC (Carvalhais et al., 2008; Wutzler and Reichstein, 2007). Land cover was also an important factor. Wetland is one of influential land cover types with high carbon contents. In general, wetland soil stores more carbon per unit area than upland soils. Incorporating hydrology and the resulting carbon dynamics in wetlands would be an important improvement for ESMs.

Elevation was revealed as an influential factor particularly in northern observational databases (Fig. 3b). We speculate that elevation may serve as a comprehensive index of SOC in a limited area because other variables, such as temperature, NPP, soil texture and other factors, change with increasing elevation. The effect of elevation in ESMs was not as high as in observational databases (Fig. 6). We estimated that the effect of elevation might automatically increase if the aforementioned other processes are properly adjusted/included in ESMs.

We examined key factors from a wide variety of candidate properties, but some potentially important mechanisms that would improve the reproducibility of SOC by ESMs and process-based ecosystem models may be missing. For example, it has been suggested that including microbial dynamics in SOC models improves projections of global soil carbon by ESMs (Wieder et al., 2013). Mycorrhizae have been reported to play an important role in soil carbon storage (Averill et al., 2014). Because soil carbon accumulation and decomposition are slow processes and land cover is an important factor of SOC, as shown in our study, taking land-use history into consideration may be essential. In addition, because soil has depth and SOC and soil environments vary according to depth (Davidson and Trumbore, 1995; Hashimoto and Komatsu, 2006; Jobbágy and Jackson, 2000), vertical soil heterogeneity/processes are important (Braakhekke et al., 2013; Wieder et al., 2013). The importance of mineral reactivity has also been suggested (Doetterl et al., 2015). However, our results may suggest that the

performance of ESMs can be improved simply through the adequate re-evaluation/inclusion of well-known processes. Another approach would be model-data fusion (assimilation) (Hararuk et al., 2014). Constraining model parameters by observational databases through data assimilation such as a Bayesian approach would improve the performance of ESMs. Another uncertainty of this analysis is the issue of scale: if the analysis is applied at a much finer resolution, such as 1 km, then the influential factors may differ.

In this study, the same data-mining BRT algorithm was applied to observational databases of SOC stock and ESM outputs. By comparing the outputs from both analyses, we revealed the similarities and differences between the observational databases and ESMs. On the global scale, incorporating the influence of the CN ratio and clay content in ESMs was identified as a potential means to improve the ability of these models to reproduce the distribution of SOC in observational databases. The results of this study will help elucidate the nature of both observational SOC databases and ESM outputs and improve the terrestrial carbon dynamics modelled in ESMs. This study demonstrates that the data-mining scheme can be used to compare results from observational databases and ESMs in detail and to determine the key factors involved in the mismatches.

Code availability

The R code with a tutorial for BRT is available as a Supplementary material of Elith et al. (2008).

Acknowledgements

This study was supported by JSPS KAKENHI Grant Number 24510025. We also acknowledge the Academy of Finland and mobility funding (nr. 276300) for supporting this work. We thank Dr. Tomohiro Hajima for help with our understanding of the CMIP5 models.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

References

Aertsen, W., Kint, V., De Vos, B., Deckers, J., Van Orshoven, J. and Muys, B.: Predicting forest site productivity in temperate lowland from forest floor, soil and litterfall characteristics using boosted regression trees, Plant Soil, 354, 157–172, doi:10.1007/s11104-011-1052-z, 2011.

Averill, C., Turner, B. L. and Finzi, A. C.: Mycorrhiza-mediated competition between plants and decomposers drives soil carbon storage, Nature, 505, 543–545, doi:10.1038/nature12901, 2014.

Batjes, N. H.: Total carbon and nitrogen in the soils of the world, Eur. J. Soil Sci., 47, 151–163, doi:10.1111/j.1365-2389.1996.tb01386.x, 1996.

Bond-Lamberty, B. and Thomson, A.: Temperature-associated increases in the global soil respiration record., Nature, 464, 579–582, doi:10.1038/nature08930, 2010.

Braakhekke, M. C., Wutzler, T., Beer, C., Kattge, J., Schrumpf, M., Ahrens, B., Schöning, I., Hoosbeek, M. R., Kruijt, B., Kabat, P. and Reichstein, M.: Modeling the vertical soil organic matter profile using Bayesian parameter estimation, Biogeosciences, 10, 399–420, doi:10.5194/bg-10-399-2013, 2013.

Carvalhais, N., Forkel, M., Khomik, M., Bellarby, J., Jung, M., Migliavacca, M., Mu, M., Saatchi, S., Santoro, M., Thurner, M., Weber, U., Ahrens, B., Beer, C., Cescatti, A., Randerson, J. T. and Reichstein, M.: Global covariation of carbon turnover times with climate in terrestrial ecosystems, Nature, 514, 213–217, doi:10.1038/nature13731, 2014.

Carvalhais, N., Reichstein, M., Seixas, J., Collatz, G. J., Pereira, J. S., Berbigier, P., Carrara, A., Granier, A., Montagnani, L., Papale, D., Rambal, S., Sanz, M. J. and Valentini, R.: Implications of the carbon cycle steady state assumption for biogeochemical modeling performance and inverse parameter retrieval, Global Biogeochem. Cy., 22, GB2007, doi:10.1029/2007GB003033, 2008.

Cools, N., Vesterdal, L., De Vos, B., Vanguelova, E. and Hansen, K.: Tree species is the major factor explaining C:N ratios in European forest soils, Forest Ecol. Manag., 311, 3–16, doi:10.1016/j.foreco.2013.06.047, 2014.

Cotrufo, M. F., Wallenstein, M. D., Boot, C. M., Denef, K. and Paul, E.: The Microbial Efficiency-Matrix Stabilization (MEMS) framework integrates plant litter decomposition with soil organic matter stabilization: do labile plant inputs form stable soil organic matter?, Glob. Change Biol., 19, 988–995, doi:10.1111/gcb.12113, 2013.

Cox, P. M., Betts, R. a, Jones, C. D., Spall, S. a and Totterdell, I. J.: Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model., Nature, 408, 184–187, doi:10.1038/35041539, 2000.

Davidson, E. A. and Trumbore, S. E.: Gas diffusivity and production of CO_2 in deep soils of the eastern Amazon, Tellus B, 47, 550–565, doi:10.1034/j.1600-0889.47.issue5.3.x, 1995.

Doetterl, S., Stevens, A., Six, J., Merckx, R., Oost, K. Van, Pinto, M. C., Casanova-katny, A., Muñoz, C., Boudin, M., Venegas, E. Z. and Boeckx, P.: Soil carbon storage controlled by interactions between geochemistry and climate, Nat. Geosci., 8, 780–783, doi:10.1038/NGEO2516, 2015.

Elith, J., Leathwick, J. R. and Hastie, T.: A working guide to boosted regression trees., J. Anim. Ecol., 77, 802–813, doi:10.1111/j.1365-2656.2008.01390.x, 2008.

Exbrayat, J.-F., Pitman, A. J. and Abramowitz, G.: Response of microbial decomposition to spin-up explains CMIP5 soil carbon range until 2100, Geosci. Model Dev., 7, 2683–2692, doi:10.5194/gmd-7-2683-2014, 2014.

Exbrayat, J.-F., Pitman, A. J., Zhang, Q., Abramowitz, G. and Wang, Y.-P.: Examining soil carbon uncertainty in a global model: response of microbial decomposition to temperature, moisture and nutrient limitation, Biogeosciences, 10, 7095–7108, doi:10.5194/bg-10-7095-2013, 2013.

FAO/IIASA/ISRIC/ISSCAS/JRC: Harmonized World Soil Database (version 1.2), available at: http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/, last access: 13 May 2015, 2012.

Fernández-Martínez, M., Vicca, S., Janssens, I. a., Sardans, J., Luyssaert, S., Campioli, M., Chapin III, F. S., Ciais, P., Malhi, Y., Obersteiner, M., Papale, D., Piao, S. L., Reichstein, M., Rodà, F. and Peñuelas, J.: Nutrient availability as the key regulator of global forest carbon balance, Nat. Clim. Change, 4, 471–476, doi:10.1038/nclimate2177, 2014.

Friedl, M. A., Strahler, A. H. and Hodges, J.: ISLSCP II MODIS (Collection 4) IGBP land cover, 2000-2001, available at: http://dx.doi.org/10.3334/ORNLDAAC/968, last access: 3 February 2016, doi:10.3334/ORNLDAAC/968, 2010.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K.-G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C. and Zeng, N.: Climate–

carbon cycle feedback analysis: results from the C^{4}MIP model intercomparison, J. Climate, 19, 3337–3353, doi:10.1175/JCLI3800.1, 2006.

Global Soil Data Task Group: Global Gridded Surfaces of Selected Soil Characteristics (International Geosphere-Biosphere Programme - Data and Information System), available at: http://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=569, last access: 13 May 2015, doi:10.3334/ORNLDAAC/569, 2000.

Hajima, T., Kawamiya, M., Watanabe, M., Kato, E., Tachiiri, K., Sugiyama, M., Watanabe, S., Okajima, H. and Ito, A.: Modeling in Earth system science up to and beyond IPCC AR5, Prog. Earth Planet. Sci., 1, 1–25, doi:10.1186/s40645-014-0029-y, 2014.

Hararuk, O., Xia, J. and Luo, Y.: Evaluation and improvement of a global land model against soil carbon data using a Bayesian MCMC method, J. Geophys. Res., 119, 403–417, doi:10.1002/2013JG002535, 2014.

Hashimoto, S., Carvalhais, N., Ito, A., Migliavacca, M., Nishina, K. and Reichstein, M.: Global spatiotemporal distribution of soil respiration modeled using a global database, Biogeosciences, 12, 4121–4132, doi:10.5194/bg-12-4121-2015, 2015.

Hashimoto, S. and Komatsu, H.: Relationships between soil CO_2 concentration and CO_2 production, temperature, water content, and gas diffusivity: implications for field studies through sensitivity analyses, J. For. Res., 11, 41–50, doi:10.1007/s10310-005-0185-4, 2006.

Hashimoto, S., Morishita, T., Sakata, T. and Ishizuka, S.: Increasing trends of soil greenhouse gas fluxes in Japanese forests from 1980 to 2009, Sci. Rep., 1, doi:10.1038/srep00116, 2011.

Hiederer, R. and Köchy, M.: Global soil organic carbon estimates and the Harmonized World Soil Database. EUR 25225 EN, Publications Office of the European Union., 2011.

Hugelius, G., Tarnocai, C., Broll, G., Canadell, J. G., Kuhry, P. and Swanson, D. K.: The Northern Circumpolar Soil Carbon Database: spatially distributed datasets of soil coverage and soil carbon storage in the northern permafrost regions, Earth Syst. Sci. Data, 5, 3–13, doi:10.5194/essd-5-3-2013, 2013.

Imhoff, M. L. and Bounoua, L.: Exploring global patterns of net primary production carbon supply and demand using satellite observations and statistical data, J. Geophys. Res., 111, D22S12, doi:10.1029/2006JD007377, 2006.

Imhoff, M. L., Bounoua, L., Ricketts, T., Loucks, C., Harriss, R. and Lawrence, W. T.: HANPP Collection: Human Appropriation of Net Primary Productivity as a Percentage of net primary productivity, available at: http://sedac.ciesin.columbia.edu/es/hanpp.html, last access: 3 February 2016, 2004.

IPCC: Climate Change 2013 The Physical Science Basis, Cambridge University Press., 2013.

Jobbágy, E. G. and Jackson, R. B.: The vertical distribution of soil organic carbon and its relation to climate and vegetation, Ecol. Appl., 10, 423–436, doi:10.1890/1051-0761(2000)010[0423:TVDOSO]2.0.CO;2, 2000.

Köchy, M., Hiederer, R. and Freibauer, A.: Global distribution of soil organic carbon – Part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world, SOIL, 1, 351–365, doi:10.5194/soil-1-351-2015, 2015.

Lehner, B. and Döll, P.: Development and validation of a global database of lakes, reservoirs and wetlands, J. Hydrol., 296, 1–22, doi:10.1016/j.jhydrol.2004.03.028, 2004.

Liski, J., Palosuo, T., Peltoniemi, M. and Sievänen, R.: Carbon and decomposition model Yasso for forest soils, Ecol. Model., 189, 168–182, doi:10.1016/j.ecolmodel.2005.03.005, 2005.

Martin, M. P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L. and Arrouays, D.: Spatial distribution of soil organic carbon stocks in France, Biogeosciences, 8, 1053–1065, doi:10.5194/bg-8-1053-2011, 2011.

New, M., Jones, P. D. and Hulme, M.: ISLSCP II Climate Research Unit CRU05 Monthly Climate Data, available at: http://dx.doi.org/10.3334/ORNLDAAC/1015, last access: 3 February 2016, doi:10.3334/ORNLDAAC/1015, 2011.

Nishina, K., Ito, A., Beerling, D. J., Cadule, P., Ciais, P., Clark, D. B., Falloon, P., Friend, A. D., Kahana, R., Kato, E., Keribin, R., Lucht, W., Lomas, M., Rademacher, T. T., Pavlick, R., Schaphoff, S., Vuichard, N., Warszawaski, L. and Yokohata, T.: Quantifying uncertainties in soil carbon responses to changes in global mean temperature and precipitation, Earth Syst. Dynam., 5, 197–209, doi:10.5194/esd-5-197-2014, 2014.

Nishina, K., Ito, A., Falloon, P., Friend, A. D., Beerling, D. J., Ciais, P., Clark, D. B., Kahana, R., Kato, E., Lucht, W., Lomas, M., Pavlick, R., Schaphoff, S., Warszawaski, L. and Yokohata, T.: Decomposing uncertainties in the future terrestrial carbon budget associated with emission scenarios, climate projections, and ecosystem simulations using the ISI-MIP results, Earth Syst. Dynam., 6, 435–445, doi:10.5194/esd-6-435-2015, 2015.

Parton, W. J., Schimel, D. S., Cole, C. V and Ojima, D. S.: Analysis of factors controlling soil organic matter levels in great plains grasslands, Soil Sci. Soc. Am. J., 51, 1173–1179, 1987.

Prince, S. D. and Zheng, D. L.: ISLSCP II global primary production data initiative gridded NPP data, available at: http://dx.doi.org/10.3334/ORNLDAAC/1023, last access: 3 February 2016, doi:10.3334/ORNLDAAC/1023, 2011.

R Core team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna., 2013.

Ramankutty, N. and Foley, J. A.: ISLSCP II historical croplands cover, 1700-1992, available at: http://dx.doi.org/10.3334/ORNLDAAC/966, last access: 3 February 2016, doi:10.3334/ORNLDAAC/966, 2010.

Scharlemann, J. P., Tanner, E. V., Hiederer, R. and Kapos, V.: Global soil carbon: understanding and managing the largest terrestrial carbon pool, Cabon Manag., 5, 81–91, doi:10.4155/cmt.13.77, 2014.

Schimel, D. S., Braswell, B. H., Holland, E. a., McKeown, R., Ojima, D. S., Painter, T. H., Parton, W. J. and Townsend, A. R.: Climatic, edaphic, and biotic controls over storage and turnover of carbon in soils, Global Biogeochem. Cy., 8, 279–293, doi:10.1029/94GB00993, 1994.

Scholes, E. and Brown de Colstoun, E.: ISLSCP II global gridded soil characteristics, available at: http://dx.doi.org/10.3334/ORNLDAAC/1004, last access: 3 February 2016, doi:10.3334/ORNLDAAC/1004, 2011.

Tarnocai, C., Canadell, J. G., Schuur, E. A. G., Kuhry, P., Mazhitova, G. and Zimov, S.: Soil organic carbon pools in the northern circumpolar permafrost region, Global Biogeochem. Cy., 23, GB2023, doi:10.1029/2008GB003327, 2009.

Tian, H., Lu, C., Yang, J., Banger, K., Huntzinger, D. N., Schwalm, C. R., Michalak, A. M., Cook, R., Ciais, P., Hayes, D., Huang, M., Ito, A., Jain, A. K., Lei, H., Mao, J., Pan, S., Post, W. M., Peng, S., Poulter, B., Ren, W., Ricciuto, D., Schaefer, K., Shi, X., Tao, B., Wang, W., Wei, Y., Yang, Q., Zhang, B. and Zeng, N.: Global patterns and controls of soil organic carbon dynamics as simulated by multiple terrestrial biosphere models: Current status and future directions, Global Biogeochem. Cy., 29, 775–792, doi:10.1002/2014GB005021, 2015.

Todd-Brown, K. E. O., Randerson, J. T., Hopkins, F., Arora, V., Hajima, T., Jones, C., Shevliakova, E., Tjiputra, J., Volodin, E., Wu, T., Zhang, Q. and Allison, S. D.: Changes in soil organic carbon storage predicted by Earth system models during the 21st century, Biogeosciences, 11, 2341–2356, doi:10.5194/bg-11-2341-2014, 2014.

Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G. and Allison, S. D.: Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations, Biogeosciences, 10, 1717–1736, doi:10.5194/bg-10-1717-2013, 2013.

Tuomi, M., Thum, T., Järvinen, H., Fronzek, S., Berg, B., Harmon, M., Trofymow, J. A., Sevanto, S. and Liski, J.: Leaf litter decomposition—Estimates of global variability based on Yasso07 model, Ecol. Model., 220, 3362–3371, doi:10.1016/j.ecolmodel.2009.05.016, 2009.

Ťupek, B., Ortiz, C., Hashimoto, S., Stendahl, J., Dahlgren, J., Karltun, E. and Lehtonen, A.: Underestimation of boreal soil carbon stocks by mathematical soil carbon models linked to soil nutrient status, Biogeosciences Discussions, 1–32, doi:10.5194/bg-2015-657, 2016.

Verdin, K. L.: ISLSCP II HYDRO1k Elevation-derived Products, available at: http://dx.doi.org/10.3334/ORNLDAAC/1007, last access: 3 February 2016, doi:10.3334/ORNLDAAC/1007, 2011.

Wieder, W. R., Boehnert, J. and Bonan, G. B.: Evaluating soil biogeochemistry parameterizations in Earth system models with observations, Global Biogeochem. Cy., 28, 211–222, doi:10.1002/2013GB004665, 2014.

Wieder, W. R., Bonan, G. B. and Allison, S. D.: Global soil carbon projections are improved by modelling microbial processes, Nat. Clim. Change, 3, 909–912, doi:10.1038/nclimate1951, 2013.

Wutzler, T. and Reichstein, M.: Soils apart from equilibrium - consequences for soil carbon balance modelling, Biogeosciences, 4, 125–136, doi:10.5194/bg-4-125-2007, 2007.

Zaehle, S.: Terrestrial nitrogen-carbon cycle interactions at the global scale., Philos. T. R. Soc. B, 368, 20130125, doi:10.1098/rstb.2013.0125, 2013.

Zaehle, S., Jones, C. D., Houlton, B., Lamarque, J.-F. and Robertson, E.: Nitrogen availability reduces CMIP5 projections of twenty-first-century land carbon uptake, J. Climate, 28, 2494–2511, doi:10.1175/JCLI-D-13-00776.1, 2015.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

Table

Table 1. Variables used in the analyses and their sources.

Variables	Abbreviation	Source (database)	Original resolution	Reference
Mean annual temperature*[1]	MAT	ISLSCPII (CRU05)	1 °	New et al., 2011
Mean annual precipitation*[1]	MAP	ISLSCPII (CRU05)	1 °	New et al., 2011
Clay content (0–30 cm)	Clay	ISLSCPII	1 °	Scholes and Brown de Colstoun, 2011
CN ratio (0–30 cm)*[2]	CNratio	ISLSCPII	1 °	Scholes and Brown de Colstoun, 2011
Soil texture (0–30 cm)	Texture	ISLSCPII	1 °	Scholes and Brown de Colstoun, 2011
Compound Topographic index*[3]	CTI	ISLSCPII	1 °	Verdin, 2011
Elevation*[3]	Elev	ISLSCPII	1 °	Verdin, 2011
Slope*[3]	Slope	ISLSCPII	1 °	Verdin, 2011
Wetland ratio	Wetland	Global Lakes and Wetlands Database	30 sec	Lehner and Döll, 2004
Land cover	LandCover	ISLSCPII	1 °	Friedl et al., 2010
Net primary production	NPP	ISLSCPII	1 °	Prince and Zheng, 2011
Cropland ratio	Cropland	ISLSCPII	1 °	Ramankutty and Foley, 2010
Human appropriation of NPP percentage	HANPPpct	HANPP collection	0.25°	Imhoff et al., 2004

*[1] The original database provides monthly data. Annual means were calculated by the authors.

*[2] The CN ratio was calculated by dividing the carbon density by nitrogen density.

*[3] The native database is hydro1k, and its resolution is 1 km. The mean value of 1 km values was used in this study.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

Appendix

Table A1: ESMs we used as the outputs in this study. The term "ensemble" indicates the ensemble of outputs from the same families.

ID	ESM
1	BCC-ensemble
2	BNU-ESM
3	CanESM2
4	CCSM4
5	CESM1-ensemble
6	CMCC-CESM
7	GFDL-ESM2M
8	GISS-ensemble
9	HadGEM2-CC
10	INMCM4
11	IPSL-ensemble
12	MIROC-ensemble
13	MPI-ensemble
14	MRI-ESM1
15	NorESM1-ensemble

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

Geoscientific Model Development Discussions
Open Access
EGU

Table A2: Classification of soil texture.

ID	Texture
1	Sand
2	Loamy Sand
3	Sandy Loam
4	Silt Loam
5	Silt
6	Loam
7	Sandy Clay Loam
8	Silt Clay Loam
9	Clay Loam
10	Sandy Clay
11	Silty Clay
12	Clay

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

Geoscientific Model Development Discussions

Table A3: Classification of land cover.

ID	Land cover
1	Evergreen Needleleaf Forest
2	Evergreen Broadleaf Forests
3	Deciduous Needleleaf Forests
4	Deciduous Broadleaf Forests
5	Mixed Forests
6	Closed Shrublands
7	Open Shrublands
8	Woody Savannahs
9	Savannahs
10	Grasslands
11	Permanent Wetlands
12	Croplands
13	Urban and Built-Up
14	Cropland/Natural Vegetation Mosaic
15	Permanent Snow and Ice
16	Barren or Sparsely Vegetated

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

Geoscientific Model Development Discussions
Open Access
EGU

Figures

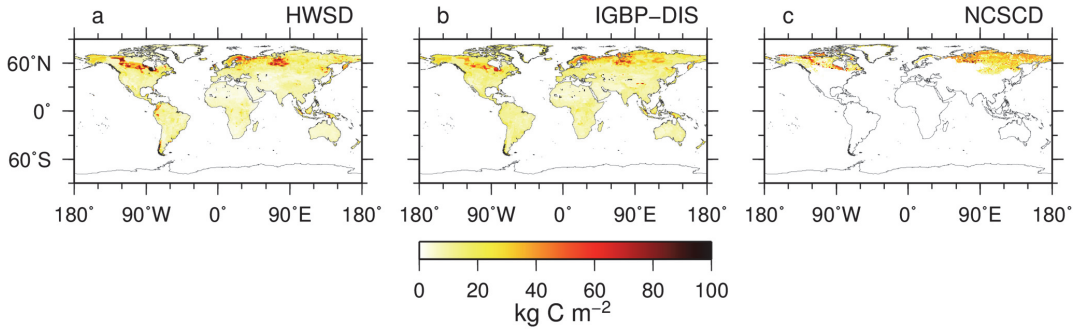

Figure 1. Soil carbon stock in the upper 100 cm (kg C m^{-2}) from the observational databases (HWSD, IGBP-DIS, and NCSCD).

5

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

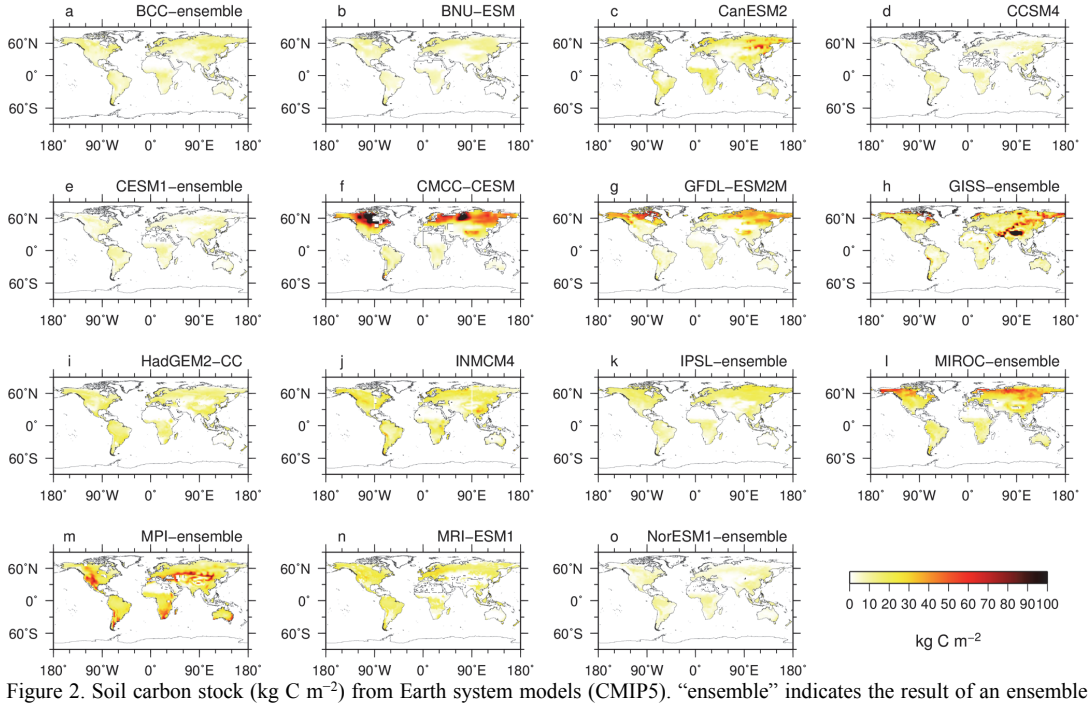

Figure 2. Soil carbon stock (kg C m^{-2}) from Earth system models (CMIP5). "ensemble" indicates the result of an ensemble of family members.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

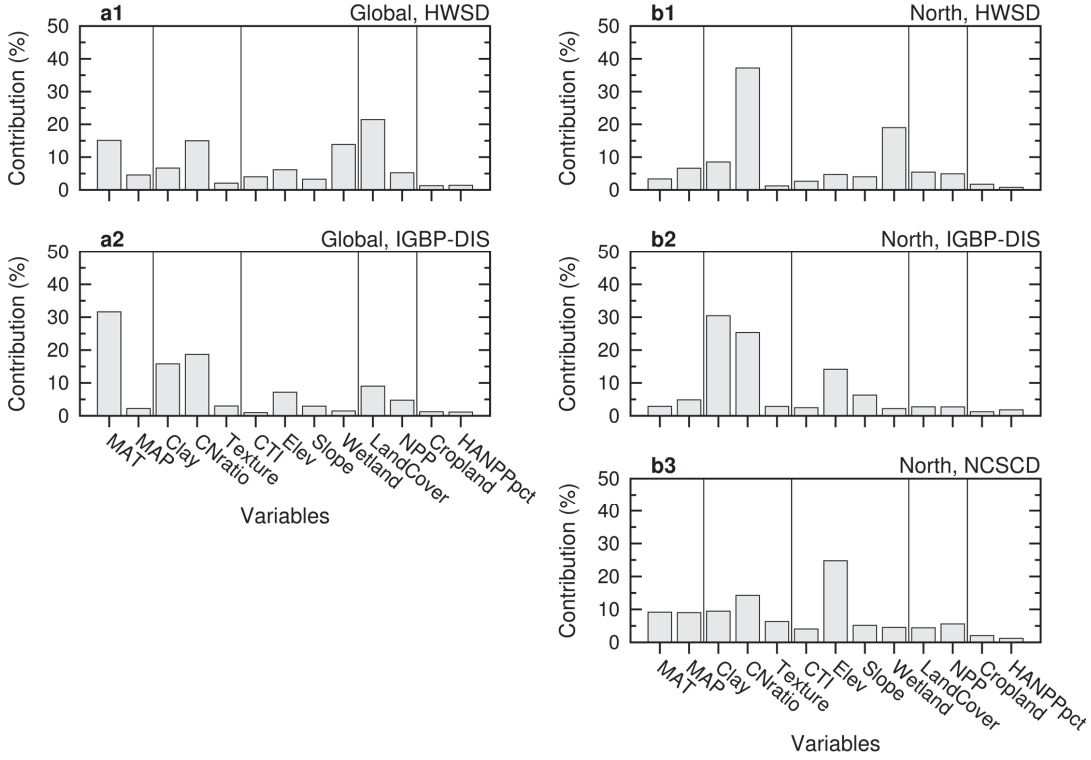

Figure 3. Relative contribution (influence) of predictive variables for the model of soil carbon stock in the global observational databases (left) and northern observational databases (right).

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

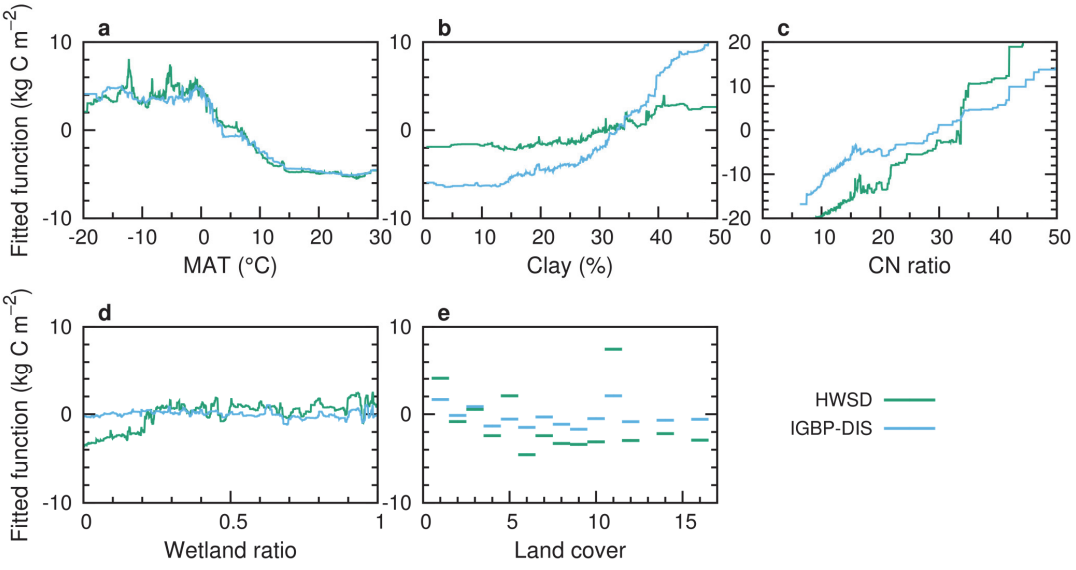

Figure 4. Effect of the most influential variables in the model of the soil carbon stock for each global observational database. The fitted functions were centred by subtracting their mean. See Table A3 for land cover classifications. Because of the small number of data points, the results for "15, Permanent Snow and Ice" were not shown (e).

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

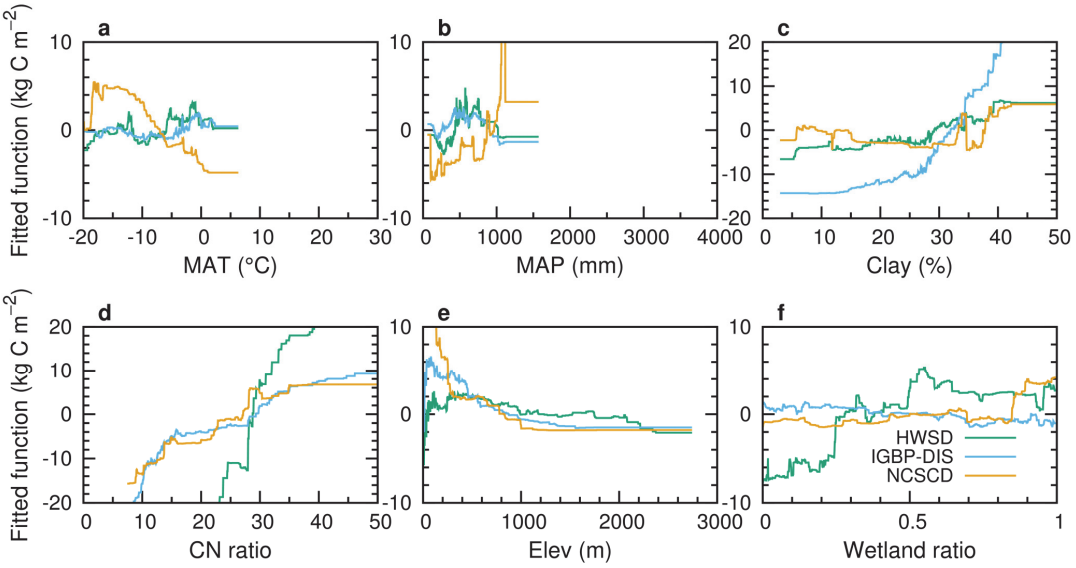

Figure 5. Effect of the most influential variables in the model of the soil carbon stock for each northern observational database. The fitted functions were centred by subtracting their mean. Note that the y-axis scales for clay and the CN ratio are different from other factors.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

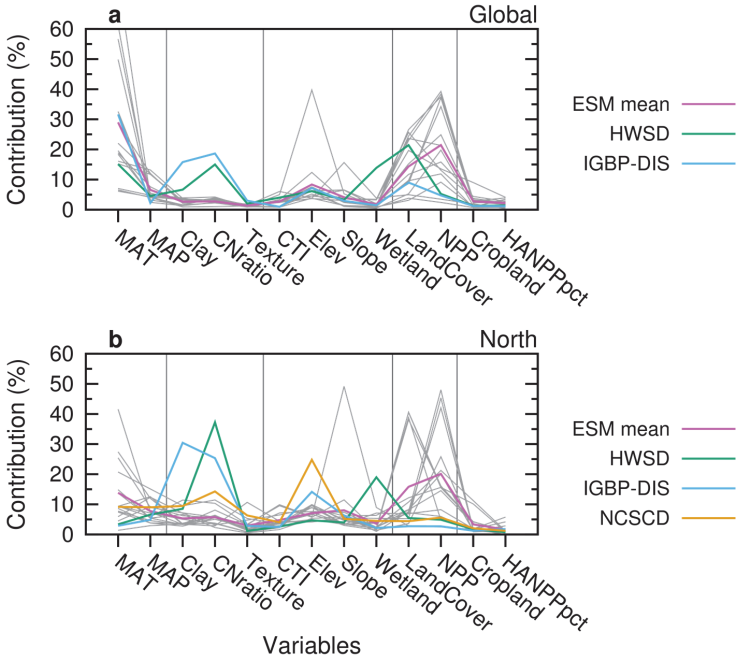

Figure 6. Relative contribution (influence) of predictive variables for the model of the soil carbon stock from ESMs and the comparison with those of observational databases. Grey lines show the results of each ESM, and the purple line indicates the mean of the ESMs.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

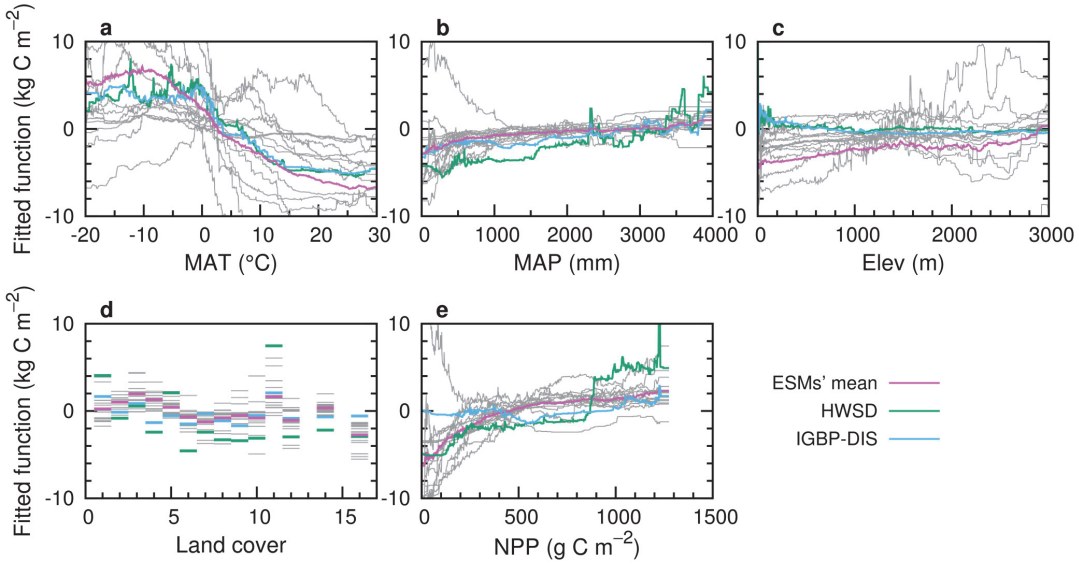

Figure 7. Effect of the most influential variables in the model for global outputs from ESMs and the comparison with those of observational databases. Grey lines show the results of each ESM, and the purple line indicates the mean of the ESMs. The fitted functions were centred by subtracting their mean. See Table A3 for land cover classifications. Because of the small number of data points, the results for "15, Permanent Snow and Ice" were not shown (d).

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-138, 2016
Manuscript under review for journal Geosci. Model Dev.
Published: 30 June 2016

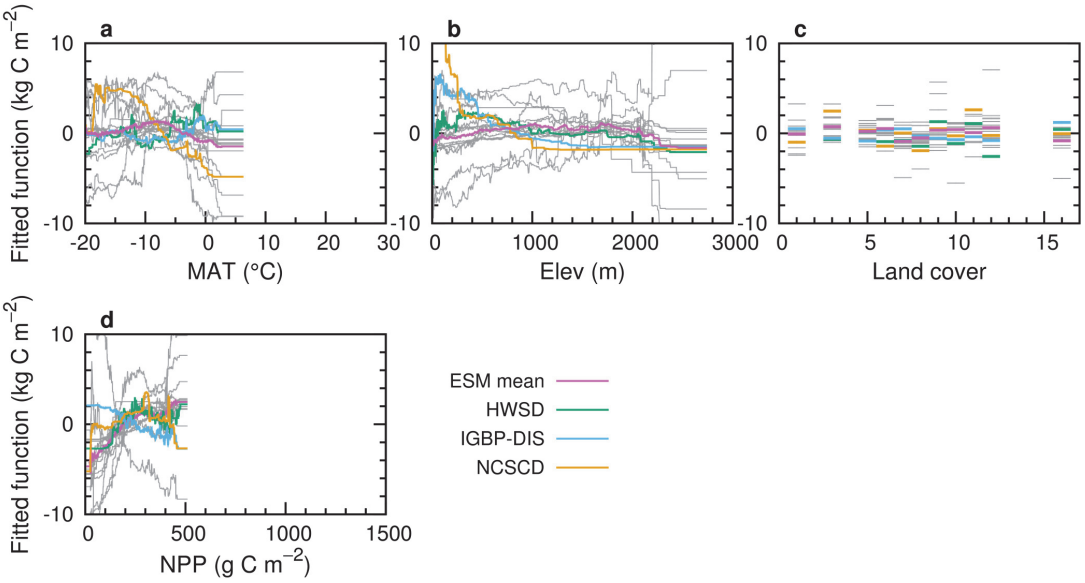

Figure 8. Effect of the most influential variables in the model for northern outputs from ESMs and the comparison with those of observational databases. Grey lines show results of each ESM, and the purple line indicates the mean of the ESMs. The fitted functions were centred by subtracting their mean. See Table A3 for land cover classifications. Because of the small number of data points, the results for "15, Permanent Snow and Ice" were not shown (c).