

Dear Dr. Sandu,

Thank you very much for handling our manuscript. We apologize for the delay in submitting a response. We have revised the manuscript according to the helpful suggestions by the referees. We hope that you will find the revised manuscript suitable for publication in Geoscientific Model Development. Thank you for your consideration.

Sincerely,

Shoji Hashimoto, Ph.D.

**Contact information:**

Shoji Hashimoto

Department of Forest Site Environment, Forestry and Forest Products Research Institute (FFPRI), 1 Matsunosato, Tsukuba, Ibaraki, 305-8687, Japan

Tel: +81-29-829-8227, Fax: +81-29-874-3720

E-mail: shojih@ffpri.affrc.go.jp

## Referee #1

We greatly appreciate your thoughtful and constructive comments. We have revised the manuscript on the basis of your comments, and our responses to the Major and Specific comments can be found below. According to the editorial instructions, our response is structured as follows: (1) comments from Referees; (2) author's response; and (3) author's changes to the manuscript. Thank you very much.

#####

### Major comments

#####

Comment 1: Need for improved texts and figures.

Response: We have revised the text and figures on the basis of your specific comments. Please see our responses to your specific comments.

Changes to the manuscript: Please see our responses to your specific comments.

#####

Comment 2: More information for reproducibility

Response: We have added more details about our methodology (e.g., software details and procedural details) along with the code that we used. We cannot attach the CMIP5 data because of their terms of use (<http://cmip-pcmdi.llnl.gov/cmip5/terms.html>), but we have attached the other data (data for observational databases) in the Supplement.

Changes to the manuscript: Please see our responses to your specific comments 6. Additionally, please see the Supplement.

#####

Comment 3: “The authors need to expand on the genuinely interesting parts, and offer more interesting, novel insights on how their results apply, and will be useful, to future work. The lack of reproducibility (above) means that it’s also not clear how any of this would inform or be useful for modelers seeking to improve their software and science.”

Response: This type of the manuscript was “Methods for assessment of models” (please see [http://www.geoscientific-model-development.net/about/manuscript\\_types.html](http://www.geoscientific-model-development.net/about/manuscript_types.html)), and the main purpose of the paper was to demonstrate the possibility of using a machine learning algorithm to compare model outputs with observational databases; it was not aimed at identifying new mechanisms of the soil carbon cycle. We have improved the reproducibility of our analyses by adding more information about the methods along

with the code and applied data in the Supplement. We have modified the discussion to clearly convey our message.

Changes to the manuscript: Please see our responses to your specific comments, and Discussion in the revised manuscript.

#####

Specific comments

#####

Comment 1. Page 1, line 1: I'd suggest either "Data-mining analysis of the global..." or "Factors affecting the global..."

Response: We have changed the title to "Data-mining analysis of the global..."

Changes to the manuscript: "Data-mining analysis of the global distribution of soil carbon in observational databases and Earth system models"

#####

Comment 2. P. 1, lines 9, 20, and 26-27: these three short sentences could be deleted with no real loss

Response: We have deleted the first two sentences, but we prefer to retain the third sentence (lines 26-27) because, as mentioned above, we would like to emphasize this point.

Changes to the manuscript: We have deleted the first two sentences.

#####

Comment 3. P. 1, l. 25: "elucidate the nature" of the databases? Confusing

Response: We have rewritten the sentence.

Changes to the manuscript: (Page 1, line 24–25) "The results of this study should aid in identifying the causes of mismatches between observational SOC databases and ESM outputs and improve the modelling of terrestrial carbon dynamics in ESMs."

#####

Comment 4. P. 2, l. 4: what recent study?

Response: This is Todd-Brown et al. 2013. We have added "(Todd-Brown et al., 2013)".

Changes to the manuscript: (Page 2, line 7–8) “a recent study (Todd-Brown et al., 2013) has found that...”

#####

Comment 5. 5.P. 3, l. 8-9: divided over what spatial scale? Some more detail in this entire para-graph would be useful

Response: A resolution was added.

Changes to the manuscript: (Page 3, line 30–31) “The wetland ratio was calculated by dividing the number of wetland grids at 30 seconds by the total grids at 1°.”

#####

Comment 6. Methods: need to give version numbers CDO, R, and all packages used. Also, I’m shocked at the complete lack of any mention of data or code availability (no, that one sentence on p. 7 doesn’t count). It’s 2016, and I expect all code and data (at least that backing the main results) to be included as Supplementary info, or posted in a repository. It’s not acceptable to produce results from a black box; see also [http://www.geoscientific-model-development.net/about/code\\_and\\_data\\_policy.html](http://www.geoscientific-model-development.net/about/code_and_data_policy.html)

Response: We agree with the importance of openness. We have added details about the software used in this study and about the main code that we used. We have also attached the data used in this study. We cannot attach the CMIP5 output (please see the terms of use of CMIP5: <http://cmip-pcmdi.llnl.gov/cmip5/terms.html>), but our results are easily reproduced by pasting CMIP5 data into the attached data and applying the data codes.

Changes to the manuscript: (Page 3, line 38–Page 4, line 2) “All global databases, except for the databases with 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, version 1.6.9, provided by the Max Plank Institute for Meteorology (<https://code.zmaw.de/projects/cdo>). A bilinear interpolation, which is one of the most widely used algorithms, was used (remapbil in CDO).”

(Page 4, line 15–23) “We used the open-source BRT package (brt.functions.R) in R software version 3.2.1 and 3.2.2 (R Core team, 2013) developed by Elith et al. (2008). The gbm package was used (version 2.1.1) to run the BRT package. The calculations were performed in Mac OS X (version 10.9.5 and version 10.10.5). To do so, the “windows” function in the “brt.functions.R” needed to be replaced with the “quartz” function in R. In practice, three parameters in the BRT package—the 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 at 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.”

Please see Supplement.

#####

Comment 7. P. 4, l. 17: “Relationships with a mean annual temperature were relatively close to each other” – what does this mean? Clarify

Response: We have rephrased the sentence.

Changes to the manuscript: (Page 5, line 1) “Relationships with the mean annual temperature were similar.”

#####

Comment 8. P. 4, l. 34: “The contribution of each variable varied between ESMs” ?

Response: We have rephrased the sentence.

Changes to the manuscript: (Page 5, line 19) “The contributions of some variables varied among ESMs”

#####

Comment 9. P. 4, l. 36: “large inconsistencies. . .demonstrated low contributions” – what?

Response: We have modified the sentence.

Changes to the manuscript: (Page 5, line 20–22) “Large inconsistencies between the observational databases and ESMs were found in the low contributions of clay content and the CN ratio and in the high contributions of NPP in ESMs (Figs. 5a and 5b)”

#####

Comment 10. P. 5, l. 23-24: this is an interesting point, and should be expanded upon. What are the implications, if the seemingly wide variety of CMIP5 models in fact uses a much smaller number of fundamental assumptions or modeling approaches? I’m pretty sure that Kathe Todd-Brown made this point in one of her papers; see also Alexander et al. (2015), 10.5194/gmd-8-1221-2015

Response: We have expanded this part.

Changes to the manuscript: (Page 6, line 4–21) “Analyses of the ESM outputs showed large variability, but the influential factors were predominantly similar among the ESMs (Fig. 5). This similarity most probably indicates that the structures of the models that describe SOC dynamics in the ESMs are similar. One reason for the similarity is probably because some ESMs share common code (Alexander and Easterbrook, 2015). Another reason may be rooted in the basic structure of the soil carbon model: SOC is calculated as the balance between dead organic matter input to soil and carbon emissions from the decomposition of organic matter in soil, and these processes are influenced by temperature and water conditions. The SOC pool is characterized by its turnover time (decomposition constant). In general, decomposition exhibits an exponential response to temperature, which is more severe than its response to water. As a result, modelled SOC is strongly influenced by NPP (litter input), temperature, and turnover time, which have been demonstrated by previous studies (Exbrayat et al., 2014; Todd-Brown et al., 2013) and were also confirmed in our analyses. As shown in Table 2, SOC submodels in ESMs differ in the number of SOC pools and function types of temperature and moisture. Todd-Brown et al. (2013) have reported the absence of any pattern of agreement between ESM outputs and observational SOC databases with soil carbon pools, temperature and moisture sensitivity functions, and Exbrayat et al. (2014) have found that turnover times of SOC in ESM outputs are not affected by the number of SOC pools. Our analyses also indicated that a match or mismatch of major contributing factor between ESM outputs and observational databases are not strongly related to these properties of SOC submodels. Thus, it is likely that the spatial pattern of SOC from ESMs are more strongly affected by the basic structure, driving variables (NPP and temperature) and parameterisations (turnover time and influential parameters of temperature and moisture sensitivity) than by the number of pools and the function types of temperature and moisture sensitivity.”

#####

Comment 11. P. 6, l. 13-14: “The use of temperature sensitivity. . .” - ?

Response: We have deleted the sentence.

Changes to the manuscript: Deleted.

#####

Comment 12. P. 7, l. 2-3: would such model-data fusion ever be possible, given the extremely long running time of modern ESMs?

Response: I agree. As you point out, model-data fusion of the whole ESM system is very difficult because of the long running time. In practice, I think that applying model-data fusion to a limited part of ESMs (e.g., ecosystem carbon cycle models) would be realistic. We have added the above points to the main text.

Changes to the manuscript: (Page 8, line 4– 7) “Constraining model parameters with observational databases through data assimilation, such as a Bayesian approach, would improve the performance of ESMs. Applying such model-data fusion to whole ESMs, however, would require a very long running time; therefore, model-data fusion to a part of an ESM (e.g., ecosystem carbon cycle model) would be realistic.”

#####

Comment 13. Table A1: an URL or reference for each model would be useful

Response: We have added a URL for each model.

Changes to the manuscript: Please see Table 2 in the revised manuscript.

#####

Comment 14. Table A2: this classification was applied to. . .? Where is it from?

Response: We used soil texture data in the ISLSCPII database (Table 1), and a classification in the database was used (Table A2). Because the contribution of soil texture was not high, the relationships between the SOC and soil texture are not shown. To clarify that this soil texture classification is for the soil texture shown in Table 1, we have modified the caption of Table A1 and A2 in the revised manuscript.

Changes to the manuscript: (Page 15, Table A1) “Classification of soil texture in ISLSCPII (see Table 1).”

#####

Comment 15. Figures 1 and 2: these are so tiny I’m not sure they convey any information, really

Response: The maps have been redrawn.

Changes to the manuscript: Please see Fig. 1 and Fig. 2 in the revised manuscript.

#####

Comment 16. Figure 6 should be the central, most important figure of the entire paper—showing how variable importance compares between observational databases and ESMs—but it’s very difficult to see what’s going on. I’d suggest re-thinking this, and carefully considering the most effective way to show this

Response: We have redrawn Fig. 6 using a box plot, and we have added a new figure to clearly show the results from each ESM. We have redrawn other figures, too.

Changes to the manuscript: Please see Fig. 5 in the revised manuscript.

Dear Dr. Todd-Brown,

We greatly appreciate your constructive comments and suggestions. We have revised the manuscript on the basis of your comments, and the responses to the Major and Specific comments are found below. According to the editorial instructions, our response is structured as follows: (1) comments from Referees; (2) author's response; and (3) author's changes to the manuscript. Thank you very much.

#####

General comments:

#####

Comment 1: Need for considering model structures

Response: We have added information about model structure to Table 2 in the revised manuscript. As reported in previous studies, we did not see clear influences of model structure on our results. We have also added a discussion.

Changes to the manuscript: (Page 6, line 4– 21) “Analyses of the ESM outputs showed large variability, but the influential factors were predominantly similar among the ESMs (Fig. 5). This similarity most probably indicates that the structures of the models that describe SOC dynamics in the ESMs are similar. One reason for the similarity is probably because some ESMs share common code (Alexander and Easterbrook, 2015). Another reason may be rooted in the basic structure of the soil carbon model: SOC is calculated as the balance between dead organic matter input to soil and carbon emissions from the decomposition of organic matter in soil, and these processes are influenced by temperature and water conditions. The SOC pool is characterized by its turnover time (decomposition constant). In general, decomposition exhibits an exponential response to temperature, which is more severe than its response to water. As a result, modelled SOC is strongly influenced by NPP (litter input), temperature, and turnover time, which have been demonstrated by previous studies (Exbrayat et al., 2014; Todd-Brown et al., 2013) and were also confirmed in our analyses. As shown in Table 2, SOC submodels in ESMs differ in the number of SOC pools and function types of temperature and moisture. Todd-Brown et al. (2013) have reported the absence of any pattern of agreement between ESM outputs and observational SOC databases with soil carbon pools, temperature and moisture sensitivity functions, and Exbrayat et al. (2014) have found that turnover times of SOC in ESM outputs are not affected by the number of SOC pools. Our analyses also indicated that a match or mismatch of major contributing factor between ESM outputs and observational databases are not strongly related to these properties of SOC submodels. Thus, it is likely that the spatial pattern of SOC from ESMs are more strongly affected by the basic structure, driving variables (NPP and temperature) and parameterisations (turnover time and influential parameters

of temperature and moisture sensitivity) than by the number of pools and the function types of temperature and moisture sensitivity.”

Please see Table 2 in the revised manuscript.

#####

Comment 2: Lack of other studies included in detail.

Response: We have added more detail regarding citation of other studies. Please see the specific response to comment 4.

Changes to the manuscript: (Page 2, line 14– 21) “Todd-Brown et al. (2013) have analysed soil carbon outputs from 11 ESMs from the fifth phase of the Coupled Model Intercomparison Project (CMIP5) and data from HWSD, and have found that the spatial variation of SOC from ESMs can be explained by net primary productivity (NPP) and temperature but that the spatial variation in HWSD cannot be explained by NPP and temperature. They have also found that the differences in SOC from ESMs are driven by differences in the simulated NPP and the parameterisation of soil heterotrophic respiration, not by differences in soil model structure in ESMs. The important influence of parameterisation of soil heterotrophic respiration (e.g., turnover time) on SOC in CMIP5 ESMs has also been suggested by Exbrayat et al. (2013).”

(Page 6, line 9–16) “The SOC pool is characterized by its turnover time (decomposition constant). In general, decomposition exhibits an exponential response to temperature, which is more severe than its response to water. As a result, modelled SOC is strongly influenced by NPP (litter input), temperature, and turnover time, which have been demonstrated by previous studies (Exbrayat et al., 2014; Todd-Brown et al., 2013) and were also confirmed in our analyses. As shown in Table 2, SOC submodels in ESMs differ in the number of SOC pools and function types of temperature and moisture. Todd-Brown et al. (2013) have reported the absence of any pattern of agreement between ESM outputs and observational SOC databases with soil carbon pools, temperature and moisture sensitivity functions, and Exbrayat et al. (2014) have found that turnover times of SOC in ESM outputs are not affected by the number of SOC pools.”

#####

Specific comments

#####

Comment 1: Please make it clearer that the ESMs are regressed against data products not other ESM output. While the modern ESM NPP and temperature distributions

match better with current data products. There are some notable differences in modeled NPP in particular in the CMIP5 models and this could be a source of bias in the analysis.

Response: In this study, we downloaded the SOC output of ESMs from CMIP5 and examined the relationships between many variables in various data products listed in Table 1. We have revised the sentences to make this clearer. Furthermore, we have added the code to the manuscript Supplement, which should be helpful for understanding what we did. We have stated the cause for the variation in global SOC as the variation of modelled NPP.

Changes to the manuscript: (Page 2, line 28– 30) “We combined the potentially influential variables from many data products and SOC data from both observational databases with those by ESMs, and we examined the factors influencing the distribution of SOC and the relationships between these factors and SOC stocks.”,

(Page 7, line 20– 21) “Todd-Brown et al. (2013) have found that one of the major causes of variations in SOC among ESMs is differences in simulated NPP and that the strong control by NPP is not present in HWSD.”

#####

Comment 2: P1L23-24 C:N ratio and clay content are in most ESMs in the allocation scheme. While it is intractable to investigate each modeling code directly, much of the documentation for these ESMs includes Lignin:N ratio (similar to C:N ratios) and clay content mediating decomposition. CENTURY Parton et al 1988 use Lignin:N ratio and clay content for allocation parameters (IPSL-CM5 Krinner et al 2005 cite CENTURY: Parton et al 1988)

Response: As you noted, it is intractable to investigate each modelling code directly, and hence we believe that collecting and investigating each model code here is beyond the scope of this study. We have modified the sentences in the abstract and discussion.

However, we fully agree with the importance of investigating each model code, and we are very curious about how many processes in site-scale process-based models are incorporated in land ecosystem models in ESMs. For instance, the CENTURY model simulates C and N, and the dynamics of C and N influence each other. Hence, the CENTURY model SOC submodels without N do not fully capture the processes in the CENTURY model (for example, please see Figure S12 in Tupek et al. 2016 Biogeosciences). These detailed comparisons should be required to identify the source of variation of SOC dynamics by ESMs in the future. To do so, full descriptions of the model structure and parameters are needed.

Changes to the manuscript: (Page 6, line 28– 35) “The SOC increased with increasing CN ratio in the observational databases (Fig. 4c), whereas the outputs of the ESM were

insensitive to the CN ratio. Our results support the importance of properly incorporating the N cycle (e.g., control over decomposition, soil fertility, nutrient availability, and plant litter quality) into SOC models (Berg et al., 2001; Cotrufo et al., 2013; Fernández-Martínez et al., 2014; Liski et al., 2005; Tuomi et al., 2009; Tupek et al., 2016). All of the ESMs, except for the CESM1 and NorESM in CMIP5, do not include 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.”

(Page 6, line 36– Page 7, line 2) “Clay content is also often used as a regulator of the decomposability of organic matter in the soil (e.g., CENTURY and RothC). Generally, high clay content inhibits organic matter decomposition in the soil. Furthermore, high clay content often results in low drainage and anaerobic soil conditions, which also inhibit organic matter decomposition. For IGBP-DIS, the 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) and may be incorporated in some ESMs, because soil carbon submodels in some ESMs are based on the CENTURY model (see the soil model history reported in Todd-Brown et al., 2014). However, regardless of incorporation of the control in decomposability by clay, our results suggest that the influence of clay on the carbon cycle is not well captured in present ESMs.”

#####

Comment 3: P2L4-8 Should there be a citation here?

Response: This is Todd-Brown et al. 2013. We have added “(Todd-Brown et al., 2013)”.

Changes to the manuscript: (Page 2, line7– 8) “a recent study (Todd-Brown et al., 2013) has found that...”

#####

Comment 4: P2 L8-11 A more in depth treatment of past attempts to disentangle drivers of data- model differences is called for here. Please expand on each of these treatments with particular attention to the ones that looked at the same models and data products the authors are using in this study. In addition, add something to the discussion to contrast your results with these studies.

Response: We have expanded the description of previous studies, added sentences and also omitted some sentences to contrast our results with previous work.

Changes to the manuscript: Please see the response to the General comment 1 and 2.

#####

Comment 5: Section 2.1 There needs to be some discussion about model structure in the ESM vs data products. These data products are typically constructed using correlation to the local environment (climate + land cover + geology) where the pedon was collected.

Response: We have included more sentences about the difference in the Introduction.

Changes to the manuscript: (Page 1, line 38– Page 2, line5) “These databases incorporate observed data points with global coverage, although there are biases in the spatial distribution or densities of the data points. In these databases, gridded SOC have been generated on the basis of inter-extrapolation of model outputs derived from analysis of observed SOC data points.

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. In ecosystem carbon cycle models of ESMs, SOC is calculated as the balance between dead organic matter input into soil and carbon emission from the decomposition of organic matter in soil, and these processes are influenced by temperature and water conditions. Compared with the observational estimation of SOC, the SOC distribution in ESMs involves more process-oriented simulations.”

#####

Comment 6: Section 2.1: Please summarize the methods used for each specific data product. For ESMs a discussion of their sensitivity functions and pool structure is appropriate (note that BCC was incorrectly stated to have their N-cycle turned on for CMIP5 in Todd-Brown et al 2013).

Response: We summarized the methods used for observational SOC databases. “BCC” was omitted from the list of ESMs with an N cycle. We have added a discussion about the model structure.

Changes to the manuscript: (Page 2, line 35–Page 3, line 9) “We used SOC data from two global and one northern observational database. The first global database was the HWSD (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012). The HWSD is a global database of soil physiochemical properties that has been developed by the International Institute for Applied Systems Analysis (IIASA) and the Food and Agriculture Organization of the United Nations (FAO) in collaboration with the International Soil Reference and Information Centre (ISRIC) -World Soil Information, the European Commission Joint Research Centre (JRC), and the Institute of Soil Science, Chinese Academy of Sciences (ISSCAS). The database was constructed by compiling the European Soil Database (ESDB), a 1:1 million soil map of China, various regional SOTER databases (SOTWIS Database), and a soil map of the world from the FAO. We used an SOC stock database

obtained with HWSD from the Joint Research Centre (JRC) (Hiederer and Köchy, 2011) (Fig. 1a). The second database included global gridded surfaces of selected soil characteristics (IGBP-DIS) (Global Soil Data Task Group, 2000) (Fig. 1b), which contains gridded soil physiochemical properties. The database has been developed by the Global Soil Data Task Group of the international Geosphere Biosphere Programme's (IGBP) Data and Information System (DIS), and the database was generated by linking the pedon records in the Global Pedon Database to the FAO/UNESCO digital soil map of the world. 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. The soil map data were obtained from different regions/countries (e.g., USA, Canada, Russia etc.) and were harmonized. The NCSCD were based on 1778 pedon data points.”

Please see the Table 2.

(Page 6, line 4– 21)“Analyses of the ESM outputs showed large variability, but the influential factors were predominantly similar among the ESMs (Fig. 5). This similarity most probably indicates that the structures of the models that describe SOC dynamics in the ESMs are similar. One reason for the similarity is probably because some ESMs share common code (Alexander and Easterbrook, 2015). Another reason may be rooted in the basic structure of the soil carbon model: SOC is calculated as the balance between dead organic matter input to soil and carbon emissions from the decomposition of organic matter in soil, and these processes are influenced by temperature and water conditions. The SOC pool is characterized by its turnover time (decomposition constant). In general, decomposition exhibits an exponential response to temperature, which is more severe than its response to water. As a result, modelled SOC is strongly influenced by NPP (litter input), temperature, and turnover time, which have been demonstrated by previous studies (Exbrayat et al., 2014; Todd-Brown et al., 2013) and were also confirmed in our analyses. As shown in Table 2, SOC submodels in ESMs differ in the number of SOC pools and function types of temperature and moisture. Todd-Brown et al. (2013) have reported the absence of any pattern of agreement between ESM outputs and observational SOC databases with soil carbon pools, temperature and moisture sensitivity functions, and Exbrayat et al. (2014) have found that turnover times of SOC in ESM outputs are not affected by the number of SOC pools. Our analyses also indicated that a match or mismatch of major contributing factor between ESM outputs and observational databases are not strongly related to these properties of SOC submodels. Thus, it is likely that the spatial pattern of SOC from ESMs are more strongly affected by the basic structure, driving variables (NPP and temperature) and parameterisations (turnover time and influential parameters of temperature and moisture sensitivity) than by the number of pools and the function types of temperature and moisture sensitivity.”

#####

Comment 7: P2 L33-35 Be more convincing about averaging models from the same center, there is some clustering analysis that is in the Supplemental of Todd-Brown et al 2013 that could support this.

Response: We have cited Todd-Brown et al. 2013, who have found that ESMs from the same climate centre generate very similar distributions of SOC.

Changes to the manuscript: (Page 3, line 17–18) “: Todd-Brown et al. (2013) showed through a hierarchical cluster analysis that SOC distributions were very similar among ESMs from the same climate centre.”

#####

Comment 8: P2 L34-35 Todd-Brown et al 2013 averaged all ensembles that were available at the time, this statement is incorrect. Please either provide a different justification for only considering one ensemble or, preferably, go back and re-analyze the data with the multi-ensemble mean (even better if you can incorporate the modeled uncertainty).

Response: We apologize for the incorrect description and have corrected it. We have cited other references that use only r1i1p1; most ESMs have this ensemble member output (Dirmeyer et al. Journal of Hydrometeorology 2013.; Chang et al. Journal of Geophysical Research 2012; Kumar et al. Climate Dynamics 2014; Jiang et al. Journal of Climate 2015).

Changes to the manuscript: (Page 3, line 19–21) “The notation “r1i1p1” is an identifier of the model simulation and is an ensemble member that is often used for analyses (Chang et al., 2012; Dirmeyer et al., 2013; Jiang et al., 2015; Kumar et al., 2014).”

#####

Comment 9: Section 2.4 What regridding algorithm did you use? There are several options in CDO, not all are appropriate for soil data, temperature and NPP. Please discuss which algorithm was used and why.

Response: We used “remapbil” (a bilinear interpolation) in CDO for the soil data. We used this algorithm simply because this is one of the most widely used algorithms for regridding. This study focuses only on the spatial pattern of SOC, and the total amounts of SOC were beyond the scope of this study. We believe that the difference in regridding algorithms would not affect the conclusions of this study, but we will willingly conduct a recalculation if you strongly recommend a specific algorithm.

Changes to the manuscript: (Page 4, line 1– 2) “A bilinear interpolation, which is one of the most widely used algorithms, was used (remapbil in CDO).”

#####

Comment 10: P5 L1 Describe the results here in addition to referencing the figure.

Response: The description of the results was after the sentence P5 L1 in the former manuscript. We agree with your comment that this sentence was unnecessary. We have modified this paragraph by inserting figure numbers after descriptions.

Changes to the manuscript: (Page 5, line 24– 30) “The relationships between SOC and certain variables substantially varied among the ESM databases (Fig. 6a–e), particularly in the mean annual temperature (Fig. 6a). The SOC decreased with increasing mean annual temperature (Fig. 6a) but increased with increasing precipitation (Fig. 6b) and NPP (Fig. 6e). The mean of the relationship with mean annual temperature for ESMs was highly consistent with that in the HWSD and IGBP-DIS databases of the temperature range –5–15 °C (Fig. 6a). The increasing trend with increasing NPP in ESMs was consistent with that of the HWSD, particularly below approximately 500 g C m<sup>-2</sup> of NPP (Fig. 6e). Although the wetland ratio did not contribute to the ESMs (Fig. 6a) with respect to land cover, permanent wetlands had higher SOC (Fig. 6d).”

#####

Comment 11: P7 L15 A BRT tutorial is not appropriate to cite under ‘Code availability’. Please either link or reference as SI to the actual code used in this analysis (preferred) or remove this section.

Response: We have added the codes and data for the observational databases to the Supplement.

Changes to the manuscript: (Page 8, line 18– 20) “The R code, with a tutorial for BRT, is available in the supplementary material of Elith et al. (2008) (<http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2656.2008.01390.x/full>). The codes and data for the observational databases are available in the Supplement.”

Please see the Supplement.

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

Shoji Hashimoto<sup>1</sup>, Kazuki Nanko<sup>1</sup>, Boris Tupek<sup>2</sup>, Aleksi Lehtonen<sup>2</sup>

5 <sup>1</sup>Forestry and Forest Products Research Institute (FFPRI), Tsukuba, 305-8687, Japan

<sup>2</sup>Natural Resources Institute Finland, P.O. Box 18, Vantaa, 01301, Finland

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

**Abstract.** Future climate change will dramatically change the carbon balance in the 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 project future climate conditions, 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 these factors that govern the SOC stock, remain unclear; the inclusion of such missing information would improve the agreement between modelled and observational data. In this study, we sought 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 identify 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). Globally, the contributions of mean annual temperature, clay content, CN ratio, wetland ratio, and land cover were high in observational databases, whereas the contributions of the mean annual temperature, land cover, and NPP were predominant in the SOC distribution in ESMs. A comparison of the influential factors in observational databases and ESMs, at a global scale, revealed that the most distinct differences among the observational databases and the outputs of ESMs were the low contributions of clay content and the CN ratio as well as the high contributions of NPP in ESMs. The results of this study should aid in identifying the causes of mismatches between observational SOC databases and ESM outputs and improve the modelling of terrestrial carbon dynamics in ESMs. This study indicates how 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 inputs 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 global soil carbon balance (Bond-Lamberty and Thomson, 2010; Friedlingstein et al., 2006; Hashimoto et al., 2011, 2015), and this change will affect terrestrial carbon and, consequently, the climate itself (Cox et al., 2000; Zaehle, 2013).

In the past two decades, several global soil databases have been developed, and some are undergoing further improvement (Scharlemann et al., 2014). These databases describe the global distribution of soil physiochemical properties, enabling calculation of the global distribution of SOC stocks (e.g., Harmonized World Soil Database (HWSD)), and some databases provide SOC stocks by default (e.g., IGBP-DIS). These databases incorporate observed data points with global coverage, although there are biases in the spatial distribution or densities of the data points. In these databases, gridded SOC have been generated on the basis of inter-extrapolation of model outputs derived from analysis of observed SOC data points.

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. In ecosystem carbon cycle models of ESMs, SOC is calculated as the balance between dead organic matter input into soil and carbon emission from the decomposition of organic matter in soil, and these processes are influenced by temperature and water conditions. Compared with the observational estimation of SOC, the SOC distribution in ESMs involves more process-oriented simulations. The above-mentioned observational global soil databases are often used as benchmarks 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 (Todd-Brown et al., 2013) has found that the results of ESMs are moderately consistent at the biome level, whereas the correlation between the distribution of soil carbon stock simulated by ESMs and that of observational databases is poor when the two are compared at fine scales (e.g., a 1° scale). Furthermore, estimates of SOC by ESMs and terrestrial biosphere models exhibit high uncertainty (Nishina et al., 2014, 2015; Tian et al., 2015). Several studies that have examined the cause of the mismatch between observational databases and ESM outputs and the cause of the high variation of SOC outputs from ESMs (Exbrayat et al., 2013; Todd-Brown et al., 2013; Wieder et al., 2013). For example, including microbial processes in an ESM has resulted in better reproducibility of the spatial distribution of SOC in HWSD (Wieder et al., 2013). Todd-Brown et al. (2013) have analysed soil carbon outputs from 11 ESMs from the fifth phase of the Coupled Model Intercomparison Project (CMIP5) and data from HWSD, and have found that the spatial variation of SOC from ESMs can be explained by net primary productivity (NPP) and temperature but that the spatial variation in HWSD cannot be explained by NPP and temperature. They have also found that the differences in SOC from ESMs are driven by differences in the simulated NPP and the parameterisation of soil heterotrophic respiration, not by differences in soil model structure in ESMs. The important influence of parameterisation of soil heterotrophic respiration (e.g., turnover time) on SOC in CMIP5 ESMs has also been suggested by Exbrayat et al. (2013). As stated above, some potential factors (e.g., net primary production or temperature) have been suggested; however, the key processes and factors, as well as the relationships among factors that govern the SOC stock, remain unclear, and the appropriate inclusion of these processes/factors would improve the agreement between the model and observational data.

In this study, we sought 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 stocks (Elith et al., 2008). BRT is a method based on regression trees and boosting. We combined the potentially influential variables from many data products and SOC data from both observational databases with those by ESMs, and we examined the factors influencing the distribution of SOC and the relationships between these factors and SOC stocks. 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 database. The first global database was the HWSD (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012). The HWSD is a global database of soil physiochemical properties that has been developed by the International Institute for Applied Systems Analysis (IIASA) and the Food and Agriculture Organization of the United Nations (FAO) in collaboration with the International Soil Reference and Information Centre (ISRIC) -World Soil Information, the European Commission Joint Research Centre (JRC), and the Institute of Soil Science, Chinese Academy of Sciences (ISSCAS). The database was constructed by compiling the European Soil Database (ESDB), a 1:1 million soil map of China, various regional SOTER databases (SOTWIS Database), and a soil map of the world from the FAO. We used an

SOC stock database obtained with HWSD from the Joint Research Centre (JRC) (Hiederer and Köchy, 2011) (Fig. 1a). The second database included global gridded surfaces of selected soil characteristics (IGBP-DIS) (Global Soil Data Task Group, 2000) (Fig. 1b), which contains gridded soil physiochemical properties. The database has been developed by the Global Soil Data Task Group of the International Geosphere Biosphere Programme's (IGBP) Data and Information System (DIS), and the database was generated by linking the pedon records in the Global Pedon Database to the FAO/UNESCO digital soil map of the world. 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. The soil map data were obtained from different regions/countries (e.g., USA, Canada, Russia etc.) and were harmonized. The NCSCD were based on 1778 pedon data points.

We used the HWSD and IGBP-DIS to analyse the global distribution of SOC stocks; then, we extracted a database of northern circumpolar regions from the three above databases and analysed the SOC stocks in the northern region. The relationships among 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 stocks estimated by ESMs was obtained from CMIP5. We examined the results of 15 ESMs (Fig. 2) (Table 2). 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): Todd-Brown et al. (2013) showed through a hierarchical cluster analysis that SOC distributions were very similar among ESMs from the same climate centre. The mean values from 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 the model simulation and is an ensemble member that is often used for analyses (Chang et al., 2012; Dirmeyer et al., 2013; Jiang et al., 2015; Kumar et al., 2014). The overviews of SOC submodels in the ESMs have been previously described (Exbrayat et al., 2014; Todd-Brown et al., 2013, 2014) and are also shown in Table 2. In general, each soil submodel consisted of 1 to 9 pools and incorporated the effects of temperature and moisture. Some ESMs have litter carbon pools; these were excluded from this study. A comparison between the mean of ESMs and global observational databases in a 1° grid is 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. The mean annual temperature, and annual precipitation were used as the climate variables, and the clay content, CN ratio, and texture (Appendix Table A1) were used as the soil variables (0–30 cm). The compound topographic index, elevation, slope, and wetland ratio were used as the topographic indices. 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 at 30 seconds by the total grids at 1°. The lake, reservoir, and river were not quantified as wetlands and were excluded from the total grids. The land cover type (Appendix Table A2) and NPP were adopted as 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 the indices of land-use history. The average human appropriation of the 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 with 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, version 1.6.9, provided by the Max Plank Institute for

Meteorology (<https://code.zmaw.de/projects/cdo>). A bilinear interpolation, which is one of the most widely used algorithms, was used (remapbil in CDO).

## 2.5 Boosted regression trees (BRT)

To identify the influential factors and their relationships with SOC stocks, BRT were used in this study (Elith et al., 2008).

- 5 This technique involves a 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 classify data through 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 this method are that BRT can analyse different types of variables and interaction effects among variables, and are applicable to nonlinear relationships. In recent years, the BRT technique has been  
10 used to examine the distribution of soil characteristics at a regional scale (Aertsen et al., 2011; Cools et al., 2014; Martin et al., 2011). Major outputs from BRT analyses can identify the following: (1) the relative importance (percentage of influence or contribution) of predictor variables (explanatory variables), on the basis of the weighted and scaled number of times a variable is selected for splitting (Elith et al., 2008) and (2) the relationships among variables and the explained variable shown in partial dependence plots.
- 15 We used the open-source BRT package (brt.functions.R) in R software version 3.2.1 and 3.2.2 (R Core team, 2013) developed by Elith et al. (2008). The gbm package was used (version 2.1.1) to run the BRT package. The calculations were performed in Mac OS X (version 10.9.5 and version 10.10.5). To do so, the “windows” function in the “brt.functions.R” needed to be replaced with the “quartz” function in R. In practice, three parameters in the BRT package—the 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 at 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.
- 20

## 2.6 Model performance

- 25 The goodness of fit between the BRT model and data was assessed by using the linear relationship between the predicted and observed values, the coefficient of determination ( $R^2$ ), and the root mean square error (RMSE); it is shown in Tables S1 and S2. For both the observational databases and ESM databases, the BRT models exhibited 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 stocks to the observational databases are shown in Fig. 3a and 3b. In HWSD, the contributions of land cover, mean annual temperature, CN ratio, and wetland ratio were high. For IGBP-DIS, the mean annual temperature, followed by clay content, CN ratio, and land cover also highly contributed. In  
35 particular, the mean annual temperature was very influential. The contribution of elevation to 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. 4a–e. In general, the two databases showed similar relationships. For example, the SOC decreased with increasing mean annual temperature, particularly at sites with a mean annual temperature  $> 0^\circ\text{C}$  (Fig. 4a), but increased with increasing clay content and CN ratio (Fig. 4b and 4c). The SOC

increased rapidly with an increasing CN ratio. Relationships with the mean annual temperature were similar (Fig. 4a). The relationship with clay was steeper in IGBP-DIS than in HWSD, but the opposite was true for the CN ratio (Fig. 4b and 4c). With respect to land cover, evergreen needleleaf forests and permanent wetlands had higher SOC (Fig. 4e).

### 3.1.2 Northern soils

- 5 In the northern region, the dominant contributors differed among northern soil databases and from those identified in the global database analyses described above (Fig. 3c–e). In 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 of the variables except for the cropland ratio and HANPPct contributed 5–15%. The mean annual temperature was not as influential as the global databases.
- 10 The relationships between variables and SOC stock varied more among the databases for northern soils than those of global databases (Fig. 4f–k). Furthermore, because the northern regions were extracted, the ranges of variables were narrower than the global databases. In NCSCD, the SOC decreased with increasing temperature (Fig. 4f) and increased with increasing precipitation (Fig. 4g). The SOC increased with increasing clay content and CN ratio in HWSD and IGBP-DIS (Fig. 4h and 4i), which was consistent with the findings obtained from the global databases. The increasing trend with increasing CN ratio 15 was also observed in NCSCD. The SOC decreased with increasing elevation in all databases but showed considerable variability at low elevations (Fig. 4j).

## 3.2 Earth system models

### 3.2.1 Global soil

The contributions of some variables varied among ESMs, but the mean of the results of the ESMs showed that the mean annual 20 temperature, land cover, and NPP clearly contributed to SOC distribution (Figs. 5a and 5b). Large inconsistencies between the observational databases and ESMs were found in the low contributions of clay content and the CN ratio and in the high contributions of NPP in ESMs (Figs. 5a and 5b). The contribution of NPP to ESMs was greater than in the observational databases.

The relationships between SOC and certain variables substantially varied among the ESM databases (Fig. 6a–e), particularly 25 in the mean annual temperature (Fig. 6a). The SOC decreased with increasing mean annual temperature (Fig. 6a) but increased with increasing precipitation (Fig. 6b) and NPP (Fig. 6e). The mean of the relationship with mean annual temperature for ESMs was highly consistent with that in the HWSD and IGBP-DIS databases of the temperature range –5–15 °C (Fig. 6a). The increasing trend with increasing NPP in ESMs was consistent with that of the HWSD, particularly below approximately 30 500 g C m<sup>-2</sup> of NPP (Fig. 6e). Although the wetland ratio did not contribute to the ESMs (Fig. 6a) with respect to land cover, permanent wetlands had higher SOC (Fig. 6d).

### 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) 35 were mainly the same as in the ESM global outputs (Fig. 5c and Fig. 5d). The contribution of the mean annual temperature was lower than that of the global results of the ESMs (mean of 14% for the northern and 29% for the global temperatures). The relatively large discrepancy between the observational databases and ESMs included the lower contribution of clay content, CN ratio, and elevation and the higher contribution of the mean annual temperature, land cover, and NPP in the ESMs. The relationship between SOC and variables in ESMs as well as the results of the observational databases are shown in Fig. 6f–i. The mean of the ESMs indicated that the SOC in the northern region increased with increasing NPP, and the relationship

was similar to that in HWSD (Fig. 6i), although the contribution of NPP in the ESMs differed from those of the observational database (Fig. 5c). The decreasing trend with elevation was not replicated in the ESMs (Fig. 6g).

#### 4 Discussion and concluding remarks

Analyses of the ESM outputs showed large variability, but the influential factors were predominantly similar among the ESMs (Fig. 5). This similarity most probably indicates that the structures of the models that describe SOC dynamics in the ESMs are similar. One reason for the similarity is probably because some ESMs share common code (Alexander and Easterbrook, 2015). Another reason may be rooted in the basic structure of the soil carbon model: SOC is calculated as the balance between dead organic matter input to soil and carbon emissions from the decomposition of organic matter in soil, and these processes are influenced by temperature and water conditions. The SOC pool is characterized by its turnover time (decomposition constant).

In general, decomposition exhibits an exponential response to temperature, which is more severe than its response to water. As a result, modelled SOC is strongly influenced by NPP (litter input), temperature, and turnover time, which have been demonstrated by previous studies (Exbrayat et al., 2014; Todd-Brown et al., 2013) and were also confirmed in our analyses. As shown in Table 2, SOC submodels in ESMs differ in the number of SOC pools and function types of temperature and moisture. Todd-Brown et al. (2013) have reported the absence of any pattern of agreement between ESM outputs and observational SOC databases with soil carbon pools, temperature and moisture sensitivity functions, and Exbrayat et al. (2014) have found that turnover times of SOC in ESM outputs are not affected by the number of SOC pools. Our analyses also indicated that a match or mismatch of major contributing factor between ESM outputs and observational databases are not strongly related to these properties of SOC submodels. Thus, it is likely that the spatial pattern of SOC from ESMs are more strongly affected by the basic structure, driving variables (NPP and temperature), and parameterisations (turnover time and influential parameters of temperature and moisture sensitivity) than by the number of pools and the function types of temperature and moisture sensitivity.

Using the data-mining technique, our BRT analyses revealed the 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. We examined the contributions of wider variations of factors to SOC distributions than examined in previous studies. 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 content (Fig. 5). For both global observational databases, the CN ratio made substantial contributions (Figs. 3a and 3b). The important contribution of the CN ratio was the same in the northern databases (Fig. 3c–e). The SOC increased with increasing CN ratio in the observational databases (Fig. 4c), whereas the outputs of the ESM were insensitive to the CN ratio. Our results support the importance of properly incorporating the N cycle (e.g., control over decomposition, soil fertility, nutrient availability, and plant litter quality) into SOC models (Berg et al., 2001; Cotrufo et al., 2013; Fernández-Martínez et al., 2014; Liski et al., 2005; Tuomi et al., 2009; Tupek et al., 2016). All of the ESMs, except for the CESM1 and NorESM in CMIP5, do not include 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.

Clay content is also often used as a regulator of the decomposability of organic matter in the soil (e.g., CENTURY and RothC). Generally, high clay content inhibits organic matter decomposition in the soil. Furthermore, high clay content often results in low drainage and anaerobic soil conditions, which also inhibit organic matter decomposition. For IGBP-DIS, the 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) and may be incorporated in some ESMs, because soil carbon submodels in some ESMs are based on the CENTURY model (see the soil model history reported in Todd-Brown et al., 2014). However,

regardless of incorporation of the control in decomposability by clay, our results suggest that the influence of clay on the carbon cycle is not well captured in present ESMs.

The mean annual temperature was identified as an influential factor in global databases (Fig. 3a and 3b) but not in northern soils (Fig. 3c–e). Temperature is a main factor controlling 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, the  $Q_{10}$  values, of soil organic matter decomposition in ESMs have been 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 probably exists because temperature sensitivity is an exponential process and the magnitude of changes with changing temperature is relatively small at a low temperature range. The relationships between SOC and temperature obtained in this study include 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 ESM outputs and observational databases.

The mean annual precipitation made a moderate contribution in both global observational databases and outputs from ESMs, probably because NPP and temperature were strongly correlated with moisture and also because the temperature sensitivity of decomposition is generally more dominant than soil moisture sensitivity. Similar results for outputs from ESMs have been reported in Todd-Brown et al. (2013). However, it should be noted that precipitation does not necessarily represent the actual moisture conditions in soil, as illustrated by the wetland ratio being identified as one of the influential factors (described below).

In ESMs, NPP was selected as an influential factor in ESM analyses for global and northern SOC (Fig. 5) but not in observational databases (Fig. 3), a result consistent with findings obtained in a previous study (Todd-Brown et al., 2013).

Todd-Brown et al. (2013) have found that one of the major causes of variations in SOC among ESMs is differences in simulated NPP and that the strong control by NPP is not present in HWSD. 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 proportionate 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 N cycle, the effect of NPP may be diluted in ESMs. It should also be emphasized that the large variations in the total amount of SOC from ESMs are partly due to the variation of modelled NPP in each ESM (Todd-Brown et al., 2013). 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. Wetlands are one of the influential land cover types with high carbon content. In general, wetland soils store more carbon per unit area than upland soils. Incorporating the hydrology and the resulting carbon dynamics in wetlands would be an important improvement for ESMs.

Elevation was found to be an influential factor, particularly in northern observational databases (Fig. 3d and 3e). 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 that in observational databases (Fig. 5). We estimated that the effect of elevation might automatically increase if the other aforementioned 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 in SOC, as shown in our study, taking land-use history into consideration may be essential. Furthermore, 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 with observational databases**  
5 **through data assimilation, such as a Bayesian approach, would improve the performance of ESMs.** Applying such model-data fusion to whole ESMs, however, would require a very long running time; therefore, model-data fusion to a part of an ESM (e.g., ecosystem carbon cycle model) would be realistic. Another uncertainty of this analysis is the issue of scale; if the analysis were applied at a much finer resolution, such as 1 km, then the influential factors might differ.

In this study, the same data-mining BRT algorithm was applied to observational databases of SOC stock and ESM outputs. By  
10 comparing the outputs from both analyses, we revealed the similarities and differences among the observational databases and ESMs. On a global scale, **in addition to improving parameterisation of temperature sensitivity and NPP, properly incorporating the influence of the nitrogen cycle 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 should help to identify the causes of mismatches between observational SOC databases and ESM outputs and improve the terrestrial carbon dynamics  
15 **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 and Data availability**

The R code, with a tutorial for BRT, is available in the supplementary material of Elith et al. (2008)  
[\(<http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2656.2008.01390.x/full>\)](http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2656.2008.01390.x/full). The codes and data for the observational  
20 databases are available in the Supplement.

### **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. **The constructive comments by an anonymous reviewer and Katherine Todd-Brown as reviewers greatly improved the manuscript.** We thank Dr. Tomohiro Hajima for helping to improve our  
25 understanding of the CMIP5 models, **and we also thank Dr. Atsushi Obata for information about the MRI ESM model.**

We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups (listed in Table 2 of this paper) for producing and making their model output available. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provided coordinating support and led development of software infrastructure in partnership with the Global Organization for  
30 Earth System Science Portals.

## 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.
- 5 Alexander, K. and Easterbrook, S. M.: The software architecture of climate models: a graphical comparison of CMIP5 and EMICAR5 configurations, *Geosci. Model Dev.*, 8, 1221–1232, doi:10.5194/gmd-8-1221-2015, 2015.
- 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.1365-1038.1996.tb01386.x, 1996.
- 10 Berg, B., McClaugherty, C., De Santo, A. V. and Johnson, D.: Humus buildup in boreal forests: effects of litter fall and its N concentration, *Canadian Journal of Forest Research*, 31, 988–998, 2001.
- Bond-Lamberty, B. and Thomson, A.: Temperature-associated increases in the global soil respiration record., *Nature*, 464, 579–582, doi:10.1038/nature08930, 2010.
- 15 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.
- 20 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.
- 25 Chang, E. K. M., Guo, Y. and Xia, X.: CMIP5 multimodel ensemble projection of storm track change under global warming, *J. Geophys. Res.-Atmos.*, 117, 1–19, doi:10.1029/2012JD018578, 2012.
- 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.
- 30 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.
- 35 Davidson, E. A. and Trumbore, S. E.: Gas diffusivity and production of CO<sub>2</sub> in deep soils of the eastern Amazon, *Tellus B*, 47, 550–565, doi:10.1034/j.1600-0889.47.issue5.3.x, 1995.
- Dirmeyer, P. a, Jin, Y., Singh, B. and Yan, X.: Trends in Land–Atmosphere Interactions from CMIP5 Simulations, *J. Hydrometeorol.*, 14, 829–849, doi:10.1175/JHM-D-12-0107.1, 2013.
- 40 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.
- 45 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/ORNLDAAAC/968>, last access: 3 February 2016, doi:10.3334/ORNLDAAAC/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 C4MIP 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](http://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=569), last access: 13 May 2015, doi:10.3334/ORNLDAAAC/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.-Biogeo.*, 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<sub>2</sub> concentration and CO<sub>2</sub> 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.

Jiang, L., Yan, Y., Hararuk, O., Mikle, N., Xia, J., Shi, Z., Tjiputra, J., Wu, T. and Luo, Y.: Scale-dependent performance of CMIP5 earth system models in simulating terrestrial vegetation carbon, *J. Climate*, 28, 5217–5232, doi:10.1175/JCLI-D-14-00270.1, 2015.

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.

Kumar, D., Kodra, E. and Ganguly, A. R.: Regional and seasonal intercomparison of CMIP3 and CMIP5 climate model ensembles for temperature and precipitation, *Clim. Dynam.*, 1–28, doi:10.1007/s00382-014-2070-3, 2014.

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/ORNLDAAAC/1015>, last access: 3 February 2016, doi:10.3334/ORNLDAAAC/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., Warszawski, 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., Warszawski, 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/ORNLDAAAC/1023>, last access: 3 February 2016, doi:10.3334/ORNLDAAAC/1023, 2011.
- R Core team: R: A language and environment for statistical computing, BOOK, 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/ORNLDAAAC/966>, last access: 3 February 2016, doi:10.3334/ORNLDAAAC/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/ORNLDAAAC/1004>, last access: 3 February 2016, doi:10.3334/ORNLDAAAC/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.
- Tupek, B., Ortiz, C. A., Hashimoto, S., Stendahl, J., Dahlgren, J., Karlton, E. and Lehtonen, A.: Underestimation of boreal soil carbon stocks by mathematical soil carbon models linked to soil nutrient status, *Biogeosciences*, 13, 4439–4459, doi:10.5194/bg-13-4439-2016, 2016.
- Verdin, K. L.: ISLSCP II HYDRO1k Elevation-derived Products, available at: <http://dx.doi.org/10.3334/ORNLDAAAC/1007>, last access: 3 February 2016, doi:10.3334/ORNLDAAAC/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.

**Table**

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

Variable	Abbreviation	Source (database)	Original resolution	Reference
Mean annual temperature* <sup>1</sup>	MAT	ISLSCPII (CRU05)	1 °	New et al., 2011
Mean annual precipitation* <sup>1</sup>	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)* <sup>2</sup>	CN ratio	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* <sup>3</sup>	CTI	ISLSCPII	1 °	Verdin, 2011
Elevation* <sup>3</sup>	Elev.	ISLSCPII	1 °	Verdin, 2011
Slope* <sup>3</sup>	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	HANPPct	HANPP collection	0.25°	Imhoff et al., 2004

\*<sup>1</sup> The original database provides monthly data. Annual means were calculated by the authors.

5 \*<sup>2</sup> The CN ratio was calculated by dividing the carbon density by the nitrogen density.

\*<sup>3</sup> The native database is hydro1k, and its resolution is 1 km. The mean value of 1 km was used in this study.

Table 2: ESMs used as outputs in this study. The term “ensemble” indicates the ensemble of outputs from the same families. The number of soil pools, types temperature sensitivity function, types of moisture sensitivity function, and link to nitrogen cycling. URLs of model/modelling group/model description paper are also shown.

ID	ESM	Number of Pool* <sup>1</sup>	Temperature* <sup>1</sup>	Moisture* <sup>1</sup>	Nitrogen* <sup>1</sup>	URL of model or modelling group or model description paper
1	BCC-ensemble	6	Hill	Hill	No	<a href="http://forecast.bcccsrm.cma.gov.cn/htm/">http://forecast.bcccsrm.cma.gov.cn/htm/</a>
2	BNU-ESM	2	Arrhenius	Increasing	No	<a href="http://esg.bnu.edu.cn/BNU_ESM_webs/htmls/index.html">http://esg.bnu.edu.cn/BNU_ESM_webs/htmls/index.html</a>
3	CanESM2	1	Q <sub>10</sub>	Hill	No	<a href="http://ec.gc.ca/ccmac-ccma/default.asp?lang=En&amp;n=4596B3A2-1">http://ec.gc.ca/ccmac-ccma/default.asp?lang=En&amp;n=4596B3A2-1</a>
4	CCSM4	3	Arrhenius	Increasing	Yes	<a href="http://www.cesm.ucar.edu/models/ccsm4.0/">http://www.cesm.ucar.edu/models/ccsm4.0/</a>
5	CESM1-ensemble	3	Arrhenius	Increasing	Yes	<a href="http://www.cesm.ucar.edu/models/cesm1.0/">http://www.cesm.ucar.edu/models/cesm1.0/</a>
6	CMCC-CESM	3* <sup>2</sup>	Unknown	Unknown	Unknown	<a href="http://www.cmcc.it/models/cmcc-esm-earth-system-model">http://www.cmcc.it/models/cmcc-esm-earth-system-model</a>
7	GFDL-ESM2M	2	Hill	Increasing	No	<a href="https://www.gfdl.noaa.gov/earth-system-model/">https://www.gfdl.noaa.gov/earth-system-model/</a>
8	GISS-ensemble	9	Increasing	Increasing	No	<a href="http://www.giss.nasa.gov/tools/modelE/">http://www.giss.nasa.gov/tools/modelE/</a>
9	HadGEM2-CC	4	Q <sub>10</sub>	Hill	No	<a href="http://www.metoffice.gov.uk/research/modelling-systems/unified-model/climate-models/hadgem2">http://www.metoffice.gov.uk/research/modelling-systems/unified-model/climate-models/hadgem2</a>
10	INMCM4	1	Q <sub>10</sub>	Hill	No	-
11	IPSL-ensemble	4	Q <sub>10</sub>	Increasing	No	<a href="http://icmc.ipsl.fr/index.php/icmc-models">http://icmc.ipsl.fr/index.php/icmc-models</a>
12	MIROC-ensemble	2	Arrhenius	Increasing	No	<a href="http://www.geosci-model-dev.net/4/845/2011/">http://www.geosci-model-dev.net/4/845/2011/</a>
13	MPI-ensemble	3	Q <sub>10</sub>	Increasing	No	<a href="http://www.mpimet.mpg.de/en/science/models/mip-esm.html">http://www.mpimet.mpg.de/en/science/models/mip-esm.html</a>
14	MRI-ESM1	2* <sup>2</sup>	Arrhenius* <sup>3</sup>	Increasing* <sup>3</sup>	No* <sup>3</sup>	<a href="http://www.mri-jma.go.jp/Publish/Technical/DATA/VOL_64/index_en.html">http://www.mri-jma.go.jp/Publish/Technical/DATA/VOL_64/index_en.html</a>
15	NorESM1-ensemble	3	Arrhenius	Increasing	Yes	<a href="https://wiki.met.no/noresm/start">https://wiki.met.no/noresm/start</a>

5 \*<sup>1</sup>Adopted from Todd-Brown et al. 2013, 2014. \*<sup>2</sup> Technical reports. \*<sup>3</sup>Personal communications

## Appendix

Table A1: Classification of soil texture in ISLSCPII (see Table 1).

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

Table A2: Classification of land cover in ISLSCPII (see Table 1).

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

## Figures

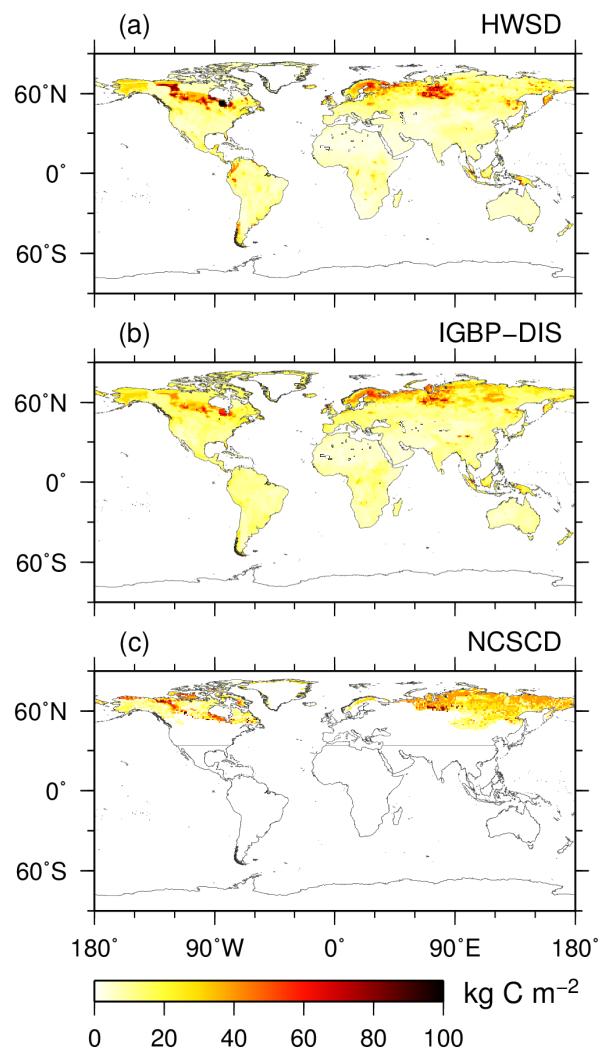


Figure 1. Soil carbon stock in the upper 100 cm ( $\text{kg C m}^{-2}$ ) from the observational databases (HWSD, IGBP-DIS, and NCSCD).

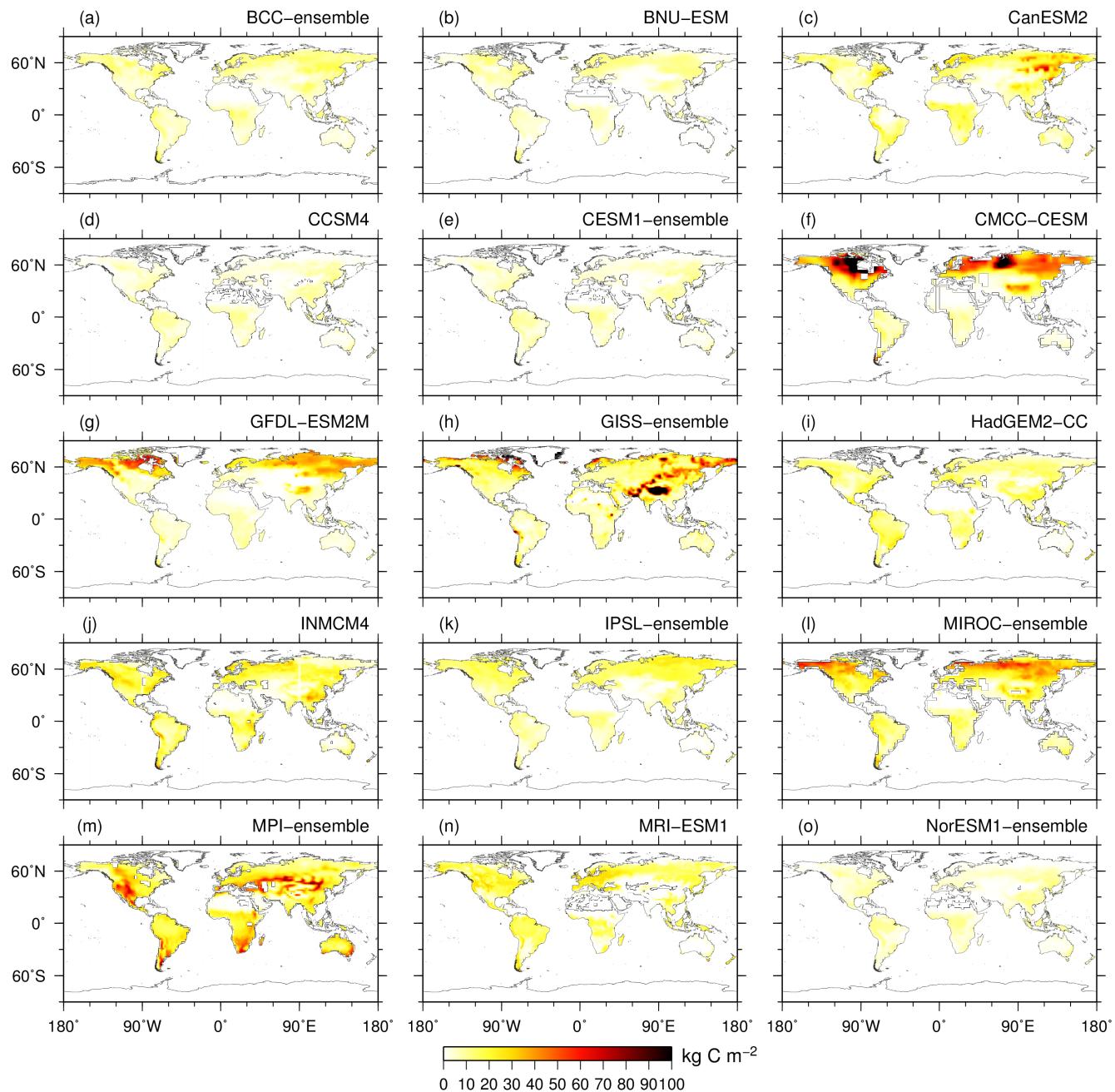


Figure 2. Soil carbon stocks ( $\text{kg C m}^{-2}$ ) from Earth system models (CMIP5). The term “ensemble” indicates the result of an ensemble of family members.

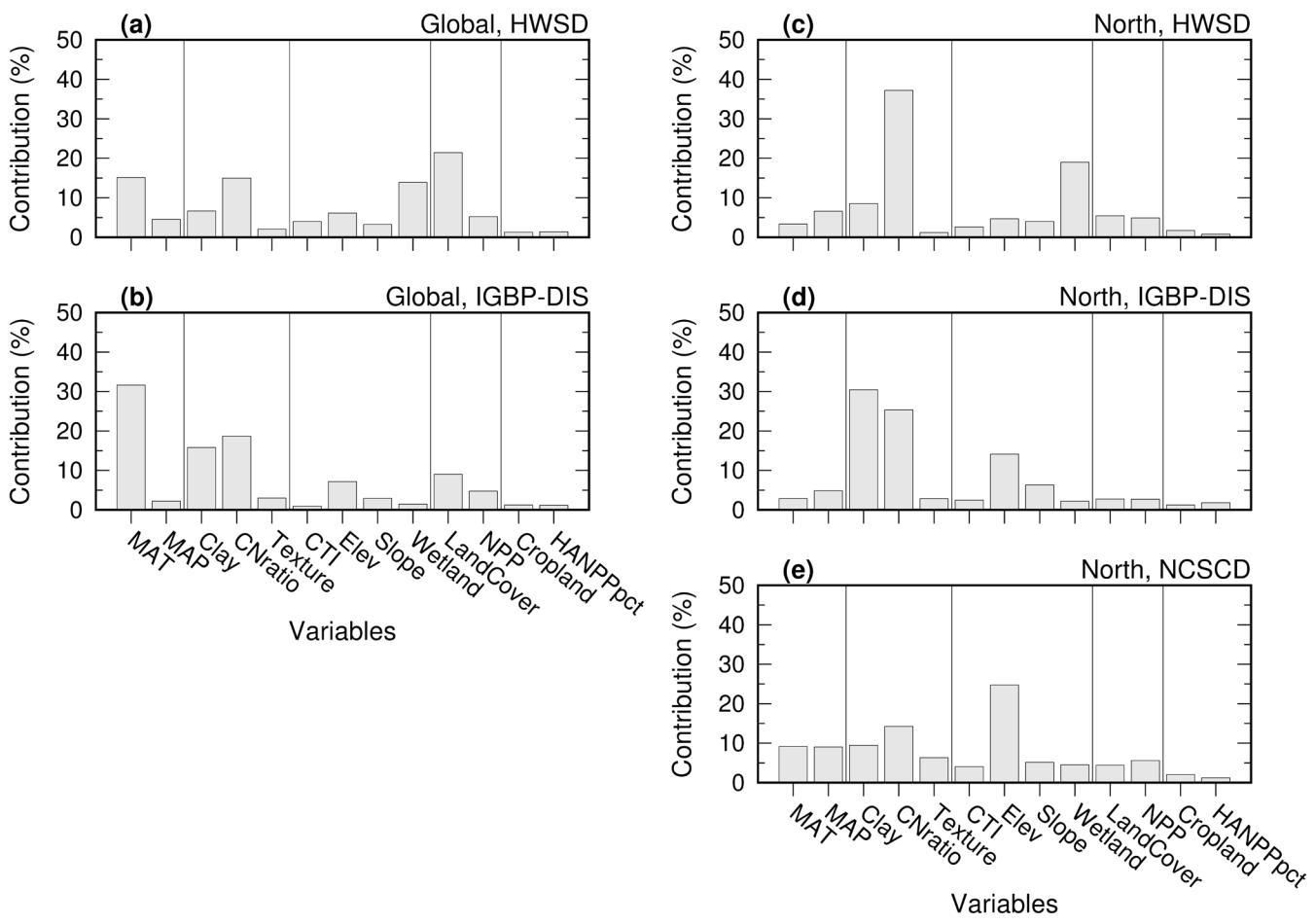


Figure 3. Relative contribution (influence) of predictive variables for the model of soil carbon stocks in the global observational databases (left) and northern observational databases (right).

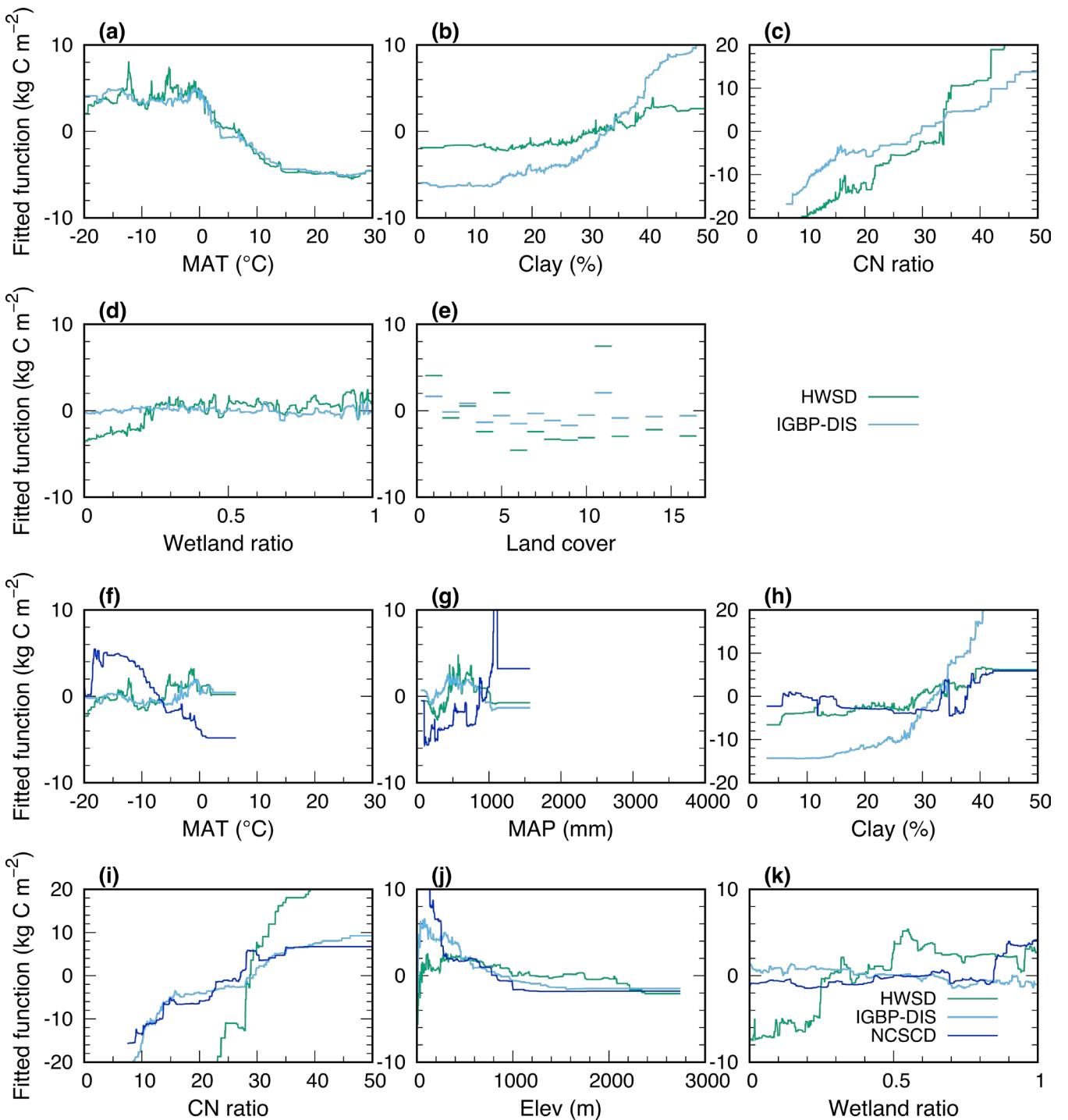
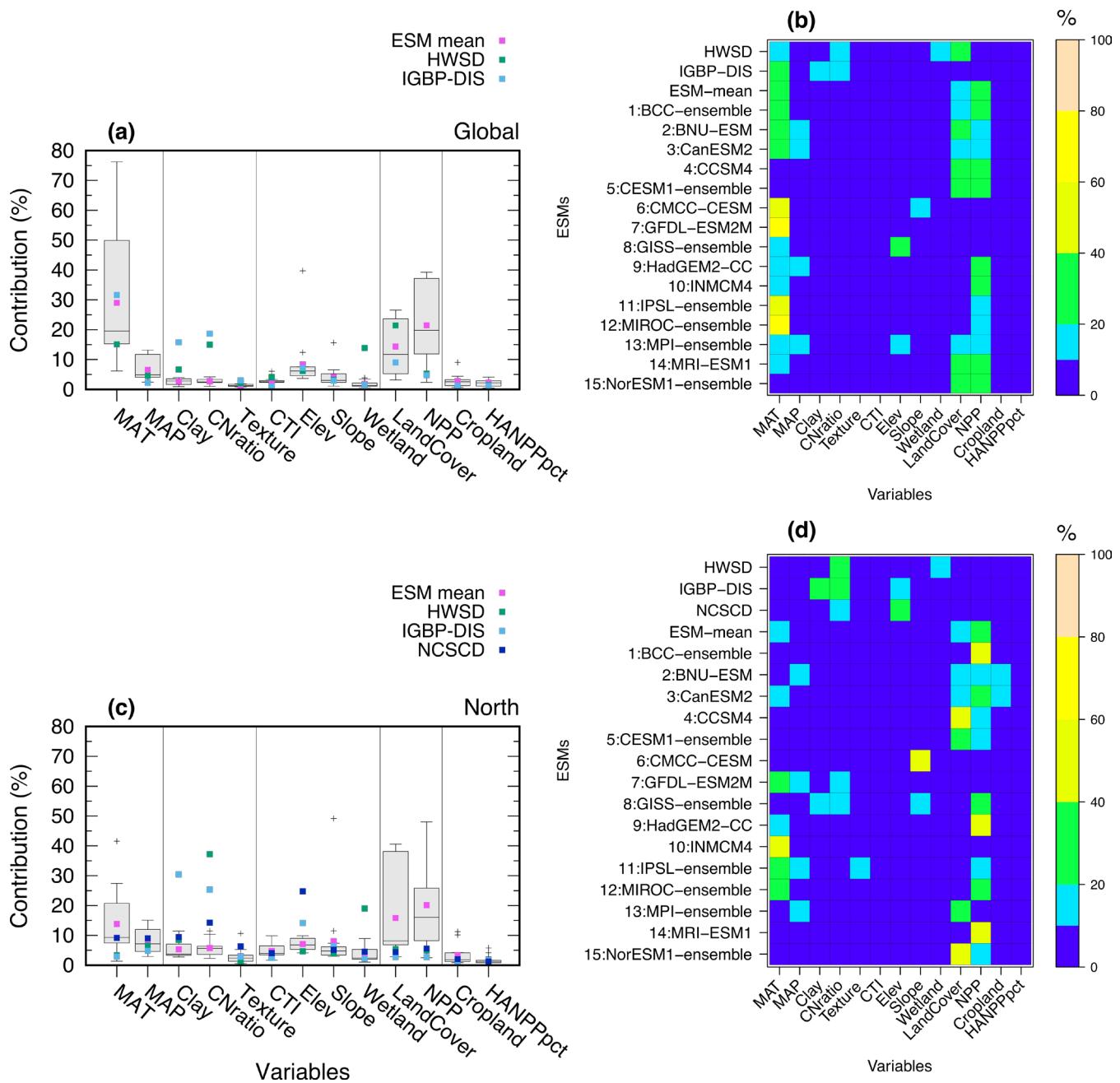


Figure 4. Effects of the most influential variables in the model of the soil carbon stock for each global (a–e) and northern (f–k) observational databases. The fitted functions were centred by subtracting their means. See Table A2 for land cover classifications. Because of the small number of data points, the results for “15, Permanent Snow and Ice” are not shown (e). The y-axis scales for clay and the CN ratio are different from those of other factors (c, h, i).



5

Figure 5. Relative contribution (influence) of predictive variables for the model of the soil carbon stock from ESMs and a comparison with those of observational databases. Box plots show the results of ESM, and the purple, green, light blue, and blue marks indicate the mean of the ESMs and results from observational databases (a: global; c: north). Mosaic plots of detailed relative contributions for each ESM (b: global; d: north).

10

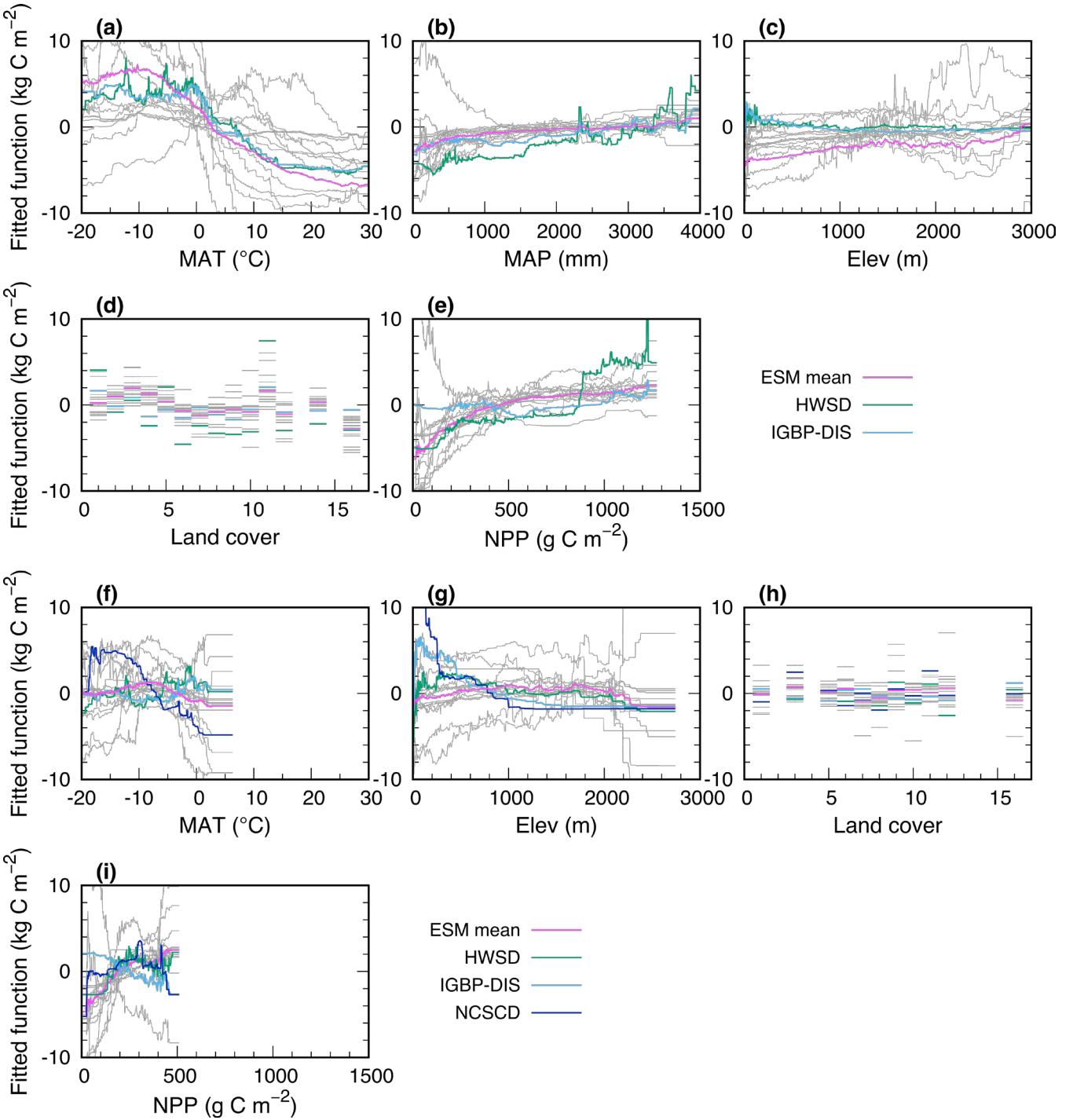


Figure 6. Effect of the most influential variables in the model for global (a–e) and northern (f–i) outputs from ESMs and a comparison with those of observational databases. Grey lines show the results of each ESM, and the purple line indicates the mean of the ESMs. The fitted functions were centred by subtracting their means. See Table A2 for land cover classifications. Because of the small number of data points, the results for “15, Permanent Snow and Ice” are not shown (d, h).