SOIL Discuss.,
https://doi.org/10.5194/soil-2018-32-AC1, 2018

Interactive comment on "A review on the global soil datasets for earth system modeling" *by* Yongjiu Dai et al.

Yongjiu Dai et al.

shgwei@mail.sysu.edu.cn

Received and published: 16 November 2018

1. General comments

This is a timely review of global scale soil data sets that are used to underpin Earth System Models, and the still numerous, associated uncertainties. Such soil data sets have evolved greatly since the coarse 1-degree resolution map generalised by Zobler (1986) resulting in a new generation of digital soil maps, and the underpinning soil point data sets and/or covariates layers. That being said, I have a number of queries and comments. For example, rather little attention is given to difficulties associated with the limited comparability of soil analytical data worldwide and uncertainty propagation. Further, several recent global soil databases of possible interest for ESM modelling have not been considered in the review and discussion.

Reply: Thanks for your valuable and detailed comments which help us a lot in improving our manuscript. The reviewer's comments have been addressed one by one in the following replies. This review was done from the perspective of ESMs and its users rather than that of soil data development. So we omitted some details about data development and associate uncertainty as pointed out by the reviewer. We are aware that many uncertainty sources exist in the derived soil dataset, which need attentions to be paid by ESM applications. After considering the comments of the reviewer, we added a paragraph concentrating on the uncertainty sources, uncertainty estimation and accuracy of soil data. And the comparability of soil analytical data, the covariates uncertainty and others are discussed in this paragraph. Some contents about the uncertainty in the original manuscript were also moved to this paragraph. As we see in the literature, ESMs usually do not consider much about uncertainty or even data quality of soil properties, which is not a good situation. ESM users should be more concerned about the uncertainty estimation rather than the uncertainty sources, while data developers need to know both aspects well. Further, we added more global soil databases as suggested by the reviewer (see the reply to table 2 and 3).

Here is the uncertainty paragraph we added:

Because soil property maps are derived products based on soil measurements of soil profiles (point observations) and spatial continuous covariates (including soil maps), it is necessary to discuss the uncertainty sources, uncertainty estimation and accuracy assessment of these derived data. More attention should be paid to this issue in ESM applications instead of taking soil property maps as observations without error. There are various uncertainty sources in deriving soil property maps, including uncertainty from soil maps, soil measurements, soil-related covariates and the linkage method itself (Shangguan et al., 2012; Batjes, 2016; Stoorvogel et al., 2016). The following may not be the complete list of uncertainty but the major ones. The uncertainty of soil maps is a major source of global dataset derived by the linkage methods. For these dataset, large sections of the world are drawn on the coarse FAO SMW map and the purity of soil maps is likely to be around 50 to 65% (Landon, 1991). Another important source of uncertainty is the limited comparability of different analytical methods of a given soil property in using soil profiles coming from various sources. A week correlation or even negative correlation were found between different analytical methods, though strong positive correlation are revealed in most cases (McLellan et al. 2013). Both dataset by the linkage method and those by digital soil mapping suffers this uncertainty. Though there are no straightforward mechanisms to harmonize the data, efforts are undertaken to address this issue and provide quality assess (Batjes, 2017; Pillar 5 Working Group, 2017). Another source of uncertainty comes from the geographic and taxonomic distribution of soil profiles, especially for the under-represented areas and soils (Batjes, 2016). The fourth source of uncertainty is from the linkage method itself. It does not represent the intra-polygon spatial variation and usually do not consider soil forming factors explicitly like digital soil mapping, though Stoorvogel et al. (2017) proposed a methodology to incorporate landscape properties in the linkage method. Finally, uncertainty from the covariates is minor because spatial perdition models such as machining learning in digital soil mapping can reduce its influences (Hengl et al., 2014). Uncertainty are estimated by different methods for the linkage method and digital soil mapping methods. For the linkage method, statistics such as standard derivation and percentiles can be used as uncertainty estimation, which are calculated for the population of soil profiles linked to a soil type (Batjes, 2016). This estimation has some limitations because soil profiles are not taken probabilistically but based on their availability, especially for the global soil datasets. Uncertainty will be underestimated when the sample size is not big enough to represent a soil type. For digital soil mapping, uncertainty could be estimated by methods such as geostatistical methods and quantile regression forest (Vaysse and Lagacherie, 2017), which make sense of statistic. The accuracy of soil dataset derived by digital soil mapping are estimated by cross-validation, but it is not trivial for those derived by the linkage method due to the global scale, the support of the data and independent data (Stoorvogel et al., 2017). Instead, some datasets, including WISE and GSDE, use some indictors such as linkage level of soil class and sample size to offer quality control information (Shangguan et al. 2014; Batjes, 2016). A simple way to compare the accuracy of datasets by both methods may be to use a global soil profile database as a validation dataset, though some of these profiles were used in deriving these datasets and questions will be raised.

The manuscript would benefit from a thorough English edit by a native speaker.

Reply: We will take a language service for the revised manuscript.

2. Specific comments L15-16: Rephrase this as e.g.: Soil is an important regulator of earth system processes, but remains one of the least well-described data layers in such models.

Reply: Modified as: Soil is an important regulator of earth system processes, but remains one of the least well-described data layers in Earth System Models (ESMs).

L17: Function as->provide

Reply: Modified.

L22: Abundant soil observations are not ‘enough’; these should have been analysed according to comparable analytical methods and quality-assessed (which is seldom the case, see Batjes et al. 2017). What about the geographical distribution, or possible clustering, of the available (i.e. shared) soil profile data?

Reply: We changed the expression as ‘with abundant, harmonized and quality controlled soil observations’. Corresponding contents are added accordingly. See the replies to related comments of the reviewer.

L24: By their nature, pedotransfer functions generally are not portable from one region to the other. Please add some discussion.

Reply: We add a sentence to the comments on Line 323.

L24-25: Speculative as written, provide some arguments for this.

Reply: See reply to comments on L451-452.

L27-28: What about uncertainty in the co-variates?

Reply: We put this as a part of the paragraph discussing uncertainty sources of the derive soil dataset

L35-36 / 45: You may consider the following reference here: http://dx.doi.org/10.1002/2015GB005239.

Reply: The reference was added. It is helpful to understand the role of soil information in ESMs.

L43: Remove available

Reply: Removed.

L45: How do you define ‘better’ here? Please clarify.

Reply: We changed the word to ‘more realistic’. This is in the following citations, Brunke et al. (2016); Luo et al. (2016); Oleson et al. (2010). We added an example here: ‘For example, Brunke et al., (2016) incorporated the depth to bedrock data in a land surface model using variable soil layers and instead of the previous constant depth.’

L47-48: Also other types of soil data, for example soil biology (see ref. line L35-36). See also discussion in https://doi.org/10.1111/gcb.13896.

Reply: We changed the sentence into ‘ESMs require detailed information on the soil physical and, chemical and biological properties’.

L56: Useful to say that the range of soil data collected during a soil survey, will vary with scale and projected applications of the data (i.e. type of soil survey, routine versus surveys/studies aimed at answering specific user demands).

Reply: We added a sentence to say this: The range of soil data collected during a soil survey, varies with scale, specifications of a country or a region, and projected applications of the data (i.e. type of soil surveys, routine versus specifically designed surveys). As a result, the availability of soil properties differs in different soil databases.

L72: How would you define reliable soil data? Remove from this sentence.

Reply: Removed

L76: Rather refer to measurements here.

Reply: Modified.

L87-88: Should add HWSD (FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012) as reference for this type of 'traditional' approach.

Reply: Added

L93: usually not ready for ...! ...not appropriately scaled or formatted for ...

Reply: modified

L113-114...: representing main soil types in a landscape unit characterised by soil profiles considered representative for the main component soils of the respective mapping units.

Reply: Here we are describing two kinds of data from soil survey, i.e., soil map and soil profiles. So We modified the sentence as: soil polygon maps representing distribution ofmain soil types in a landscape unit and soil profiles with observations of soil properties which are considered representative for the main component soils of the respective mapping units.

L124: Rephrase this...: (FAO, 2003b, Zobler 1986) and these products are known to be outdated. The information on the initial SMW and DSMW has since been updated for large sections of the world in the HWSD product (FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012), which has recently been revised in

WISE30sec (http://dx.doi.org/10.1016/j.geoderma.2016.01.034).

Reply: Added.

L124-125: Start new paragraph for the regional and national level data.

Reply: Modified.

L132: multiply –> multiple

Reply: Modified.

L133: soil properties are observed (e.g. site data) or measured (e.g. pH, sand, silt, clay content)

Reply: Modified.

L138-141: Important to mention here that data served through WoSIS have been standardised, with special attention for the description/comparability of soil analytical methods worldwide. See: http://dx.doi.org/10.17027/isric-wdcsoils.20180001. Also an important element for the discussion is that many countries, although having a large collection of soil profile data, are not yet sharing such data. See for example: https://doi.org/10.1016/j.grj.2017.06.001

Reply: Modified.

L141: The initial list of attributes corresponds with the GlobalSoilMap specifications, with additional properties added/considered later in WoSIS (see http://dx.doi.org/10.17027/isric-wdcsoils.20180001).

Reply: Modified by adding the number of soil properties as follows: The soil profiles database of World Soil Information Service (WoSIS) contains the most abundant profiles (about 118,400) from national and global databases including most of the databases mentioned below (Batjes, 2017), though only a selection of important soil properties (12) are included (Ribeiro et al., 2018).

L164: The linkage methods assigns a best-estimate for each soil property (and soil interval) under consideration to each component soil unit of a polygon (see e.g. HWSD). [see also 359-360].

Reply: modified as: Because the linkage method assigned only one value or a statistical distribution to a soil type in soil polygons (usually a polygon contains multiple soil types), the intra-polygonal spatial variation is not taken into account.

L171-173: For a more comprehensive review see also: http://dx.doi.org/10.1016/j.geoderma.2016.01.034 and http://dx.doi.org/10.1002/ldr.2656.

Reply: we added these two ref.

L178: FYI, WISE30sec considers seven layers up to 200 cm depth and 20 soil properties.

Reply: We added description of WISE30sec as one of the recent global datasets: WISE30sec is another improvement of HWSD incorporated more soil profiles with seven layers up to 200 cm depth and with uncertainty estimated by mean $\pm$ standard deviation. WISE30sec used the soil map from HWSD with minor corrections and climate zone maps as categorical covariate.

L201: Possibly, also mention the GSOC effort of the GSP here, see: https://doi.org/10.5194/soil-4-173-2018.

Reply: Added as: A third global soil mapping project is the Global SOC Map of the Global Soil Partnership, which focuses on country-specific soil organic carbon estimates (Guevara et al., 2018).

L205: . . . which is currently the most detailed, though not necessarily most accurate estimation of . . .

Reply: Modified.

L214: See also: Tifafi M, Guenet B and Hatté CCG 2018. Large differences in global and regional total soil carbon stock estimates based on SoilGrids, HWSD and NCSCD: Intercomparison and evaluation based on field data from USA, England, Wales and France. Global Biogeochemical Cycles, 42-56. http://dx.doi.org/10.1002/2017GB005678. Note: This paper is erroneously referred to as Marwa et al. 2018 in manuscript. This should be: Tifafi et al. 2018.

Reply: Corrected.

L214: Check if this is for 0-100 cm; likely these estimates are for 0-200 cm (see also recent sources mentioned above).

Reply: It is reported as 0-100 cm in the ref.

L224: Large sections of HWSDv1.2 still draw on the now outdated DSMW.

Reply: Modified as: Except GSDE, HWSD and STATSGO (Miller and White, 1998) for USA in Table 1, these datasets were derived from the Soil Map of the World (note that large sections of GSDE and HWSD still used this map as a base map because there are no available regional or national maps)

L295-296: See earlier comments.

Reply: We added WISE30sec in Table 2 and 3.

L299: WISE30sec presents estimations of uncertainty, unlike the HWSD and GSDE.

Reply: Modified as: Except WISE30sec, all these databases do not contain uncertainty estimation.

L300: Needs some discussion and references to publications on the subject.

Reply: we deleted this statement: Soilgrids is considered to be the most accurate one. Because there is not any evaluation of these four datasets, we added a discussion as: The accuracy of these datasets will need to be evaluated and compared for each

Printer-friendly version

Discussion paper

soil properties, especially for those frequently used in ESM, including sand, silt and clay content, coarse fragments, bulk density and organic carbon. Special attention should be paid to the data quality of each soil properties in the above datasets as their uncertainty and accuracy vary as discussed above. See also the following discussion.

L302: Larger number of soil properties for GSDE, but what about the accuracy of the predictions? (not given as indicated earlier).

Reply: Not only GSDE but also HWSD, WISE30sec do not provide a quantitative accuracy assessment. WISE30sec provides uncertainty estimation, and HWSD and GSDE could take similar way to estimate the uncertainty. But uncertainty estimation is different from accuracy assessment. As we discussed above, we may need further studies to evaluate them. Maybe use cross validation or independent soil profiles datasets to validate. But it seems like no one has done a cross validation for the datasets estimated by the linkage method like the digital soil mapping yet. GSDE did some quality assessment using some indicators like WISE, including linkage level of soil class, sample size, texture consideration, search radius and map unit level (see figure 6 of Shangguan et al., 2014). But it is only a reference of the accuracy and not straight forward for users, and most users may not even take a look at it. We add some discussions in the paragraph of uncertainty.

L303: Rephrase. ...SoilGrids products currently consider the list of attributes as defined by the GlobalSoilMap consortium.

Reply: modified: while Soilgrids currently contains ten primary soil properties defined by the GlobalSoilMap consortium.

L323: Most PTFs are not portable (i.e. locally or regionally validated).

Reply: we added: PTFs generally are not portable from one region to the other (i.e. locally or regionally validated).

L331-332: add database (word is missing in sentence)

Reply: modified. L360-361: ...component soil unit in most cases, and thus a one-to-many relationship exists between the SMU and the profile attributes of the respective soil units...

Reply: modified.

L397-398: Possibly, rephrase this sentence.

Reply: modified: However, some researches used the "aggregating after" method producing misleading results (Hiederer and Köchy, 2012).

L410: remove high from sentence

Reply: removed.

L441-442: Provide some justification (a sentence or two) for this statement.

Reply: added: because they provide spatial continuous estimations of soil properties using spatial prediction models with various soil-related covariates.

L451-452: Speculative as written. Please provide some evidence for this.

Reply: This issue is discussed extensively by Looy et al. 2017 at the end of section 7. For briefty, we added a sentence here instead of long discussions: because ensemble modeling carries a number of benefits and potential over the use of a single model (Looy et al., 2017).

For you reference, I copied the content from Looy et al. (2017) here: Another recent technique that has merits in this respect is ensemble modeling – i.e. the use of a number of models in combination. This technique is a natural part of weather and climate modeling today, yet it is less used in the prediction of soil properties [Baker and Ellison, 2008b]. Ensemble modeling carries a number of benefits and potential over the use of a single model. Models can differ in their theory and structure, but also in the information that they require. As a result, their sensitivity and scale of support may also differ. The use of ensemble modeling is easy to justify if it is difficult to determine which, if any, single model may be superior to others. In ensemble modeling, the main aim is not to make the single model perfect, but to capture the trend that multiple models agree on. The ensemble will amplify trends that are common among models, while by-chance predictions will be softened. The outputs, therefore, can be interpreted – qualitatively or quantitatively - as a measure of uncertainty. In the context of integrated Earth system models, the represented complex processes – integrating physical and biochemical processes typically – can be covered by a number of models with strongly varying concept and structure. Here lies an opportunity to construct ensemble models entering different PTF-based parameterizations.

L460: and quantifying uncertainty in the predictions

Reply: Added

L461: 'need to gain popularity in ...'. Basically, the "proof of the pudding is in the eating".

Reply: We provided some examples at regional scales, which shows products by digital soil mapping improved climate modelling results (Kearney and Maino, 2018; Trinh et al., 2018). But no global studies have been taken to compare digital soil mapping products and linkage method products in ESMs yet, which we are doing now. So we changed this sentence to a more conservative one: the new generation soil datasets derived by digital soil mapping need to be tested in ESMs, and some regional studies have shown that these datasets provided better modelling results than products by the linkage method (Kearney and Maino, 2018; Trinh et al., 2018). Moreover, many studies from digital soil mapping have identified that soil maps are not very important to predict soil properties and are usually not used as a covariate in most studies (eg. Hengl et al., 2014; Viscarra Rossel et al., 2015; Arrouays et al., 2018). However, the linkage method usually takes soil map as the major covariate, which essentially affect the accuracy of the derived soil property maps, especially for areas without detailed soil maps.

L462: What I miss in this paper, is a discussion of the inherent uncertainty attached to using soil profile data coming from various sources. Often, little consideration is given to differences in analytical methods used for analysing e.g. soil organic carbon content worldwide (see Shangguang et al 2014, who consider this as ‘a major imitation to their approach’). For a discussion of issues see e.g.: http://dx.doi.org/10.17027/isricwdcsoils.20180001

Reply: This was mentioned in L483-L484: Data compatibility of different analysis methods and different description protocols including soil classifications is also an important issue and data harmonization is necessary when the data are made available to public. Also, see the paragraph discussing uncertainty we added.

L463-464: More soil profiles is not necessarily the solution. More quality assessed data, analysed according to comparable analytical methods, are needed to support such efforts. Reference should be made to ’new’ types of data as derived from proximal sensing (e.g. http://dx.doi.org/10.5194/soil-2017-36), and associated limitations. Reference, in this respect, could also be made to the GLOSOLAN effort, initiated by the GSP (http://www.fao.org/global-soilpartnership/resources/events/detail/en/c/1037455/) and work of GSP Pillar 5 towards harmonisation (http://www.fao.org/3/a-bs756e.pdf). Also, importantly, the geographical distribution and possible clustering of the shared soil profiles.

Reply: these are added: More quality assessed data, analysed according to comparable analytical methods, are needed to support such efforts. The harmonization of soil data is undertaking by the work of GSP Pillar 5 (Pillar 5 Working Group, 2017) and WoSIS (Batjes et al., 2017). Data derived from proximal sensing, although with higher uncertainty than traditional soil measurements, can be used in soil mapping (England and Viscarra Rossel, 2018). To avoid spatial extrapolation, soil profiles should have a good geographical coverage.

L471-475: True, but how many of these profiles are actually being shared for the greater benefit of the international community? See paper by Arrouays et al. 2017 for a discussion.

Reply: We added: Arrouays et al. (2017) reported that about 800,000 soil profiles have been rescued in the selected countries.

L479: Some reference to the ongoing work of the Global Soil Partnership, Pillars 4 and 5, is needed here.

Reply: Added: (Pillar four Working Group, 2014; Pillar 5 Working Group, 2017)

L948: Table 2 is not complete; 'recent' datasets not yet considered in the review should be added here (http://dx.doi.org/10.1002/ldr.2656; http://dx.doi.org/10.1016/j.geoderma.2016.01.034). Idem for Table 3.

Reply: WISE30sec is added. The other one (Stoorvogel et al., 2017, which we cited in our paper) is more about proposing a new method which can improve HWSD results. 'The RMSD for SâĂŘWorld was considerably smaller (2.1% SOC) than the RMSD for HWSDweighted (2.9% SOC) .'But this method has some limitation for soil properties with limited samples and for those having week relationship with covariates. We don't find the dataset available online. And in the paper, they only tested 6 soil properties. i) topsoil thickness (cm), ii) soil depth (cm), iii) soil organic carbon (SOC) content in the top 30âĂĽcm (%), iv) SOC content in the subsoil (30 to 120âĂĽcm) (%), v) clay content in the soil profile (%), and vi) sand content in the soil profile (%) . So we did not add this citation as a dataset for now. Meanwhile, I have written email to the author to check the availability

L952: Table 3. Change title to "Derived soil properties considered in three global soil datasets". Essentially, this is a simple enumeration of derived soil properties. However, the fact that many different analytical methods have been used to derive a given soil property (e.g. soil organic carbon Walkley & Black method or LECO total analyses) or which CEC (e.g. measured at 'field pH' or in a buffer-solution at 'pH7' or 'pH8') has been considered is not mentioned here (in a footer perhaps). In their study, Shangguan et al. (2014) rightly indicate that this has not been the case and indicate that they see this an important limitation. However, there are still no straightforward mechanisms for harmonising the data (cf. GSP Pillar 5 and GLOSOLAN activities, as mentioned above).

Reply: title changed. We add a sentence in the footnote: It should be noted that many different analytical methods have been used to derive a given soil property, which is a major source of the dataset.

Potaasium –> Potassium

Reply: corrected

Interactive comment on SOIL Discuss., https://doi.org/10.5194/soil-2018-32, 2018.