Review ESSD-2017-66

Canadian soils data set

Summary: Perhaps and probably a notable and skilful effort by Canadian soil scientists. However, persistent and pervasive sloppiness about terminology substantially diminishes the quality of the work as described. The manuscript needs a section on uncertainties; some information exists in individual data processing sections but the overall description lacks a clear message to potential users about strengths and weaknesses of this approach and of the data as presented. This reviewer found, surprisingly, no indication of larger impact, e.g. to the community of SWAT users outside of Canada. Based on inherent deficiencies and absence of demonstrated impact this reviewer feels inclined to recommend that the editors reject this dataset and associated description for ESSD. However, considering the ESSD mission to promote open data access and the evident and positive motivation of these data providers to meet that goal, I would accede to an editorial decision to accept subject to very major revisions.

Data easy to access via Pangaea. Authors should please use the full doi citation, e.g. https://doi.pangaea.de/10.1594/PANGAEA.877298, to allow one-click access? But the downloads on the Pangaea landing page only lead to metadata. One needs to open the metadata files to find actual access, e.g. at http://store.pangaea.de/Publications/CordeiroM-etal_2017/SLCsoils_Canada_SWAT_Cordeiro-etal_2017.csv. The .csv files download successfully but one needs to go through 2 or 3 steps to access them. Could the authors work with Pangaea to reduce the number of steps required for access?

Already in the abstract we encounter this phrase: "14,063 unique soils". I don't believe the authors can document much less distinguish more than 14 thousand unique soils. I believe they might have 14k separate samples or profiles, but by any classification system (and one wonders why this paper did not follow the standard Canadian soil classification system, e.g. http://sis.agr.gc.ca/cansis/taxa/cssc3/index.html) these 14,000 individual profiles must fall into far fewer soil types?

Page 2, lines 44 to 46: Here the authors have grossly underestimated the use and application of SWAT. A Google Scholar search on SWAT model use and validation shows more than 7000 entries since 2013, covering agricultural soils in Africa, Asia, Europe and the Americas. If the authors hope to convince users that this effort has widespread value, e.g. for other agricultural systems across Ukraine, Hungary, China, or Australia, they need to first demonstrate that they understand the breadth of SWAT use.

Page 3, lines 51 to 53. Finally, mention of global impact of SWAT but tied to one old (2010) reference? No evidence from this group to show they understand widespread use nor anticipate widespread interest in their work.

Page 3, lines 59 to 62. Data only for USA? Researchers in no other country have undertaken a similar effort to the one described here? This reviewer doubts very much that the work presented here represents the only non-USA effort.

Page 3, lines 63 to 70. Here the authors list roughly 15 SWAT uses by the ag community of Canada alone. So SWAT does have wider impact among Canadian researchers, but what about its wider global impact? We don't need an inventory of international SWAT applications but we do need recognition and evidence that this work has impact beyond Canada's borders.

Page 4, lines 82 to 85 and Figure 1. Even if these areas represent all or even the majority of Canadian ag soils, this delineation falls far short of the phase in the title "country-wide". A more accurate description would say many of the agriculturally-relevant soils along the southern fringe of the country. Whatever the final wording, the title must change.

Page 4, line 88: "all of Canada's agricultural soils". We need evidence of the reliability of this statement. Figure 1 casts substantial doubt.

Page 5, line 98: "a country-level soils data set". This phrase seems more accurate than earlier phrases.

Page 5, lines 102, 103: "Such a laborious pre-processing exercise is widely, but inconsistently adopted in SWAT simulations reported in the literature." And? So? Does the effort described here have relevance to prior and future SWAT simulations outside of Canada? One hopes so, but the authors fail to make that point either as a preface or as an outcome.

Terminology confusion: "unique soils", "soil types", "soil series", "individual soil and landscape features", "soil component", "soil", "soil variables", soil properties", "unique soil records", "soils", "soils have been identified, but their properties have not yet been characterized", "soil profile and layer information", "inactive soil properties", "soil profile information", "different soils". The authors throw these various terms around casually, but without consistent intention? 'Soil types' used perhaps most frequently among all these terms and phrases but never defined. Even less definition for the other terms. Again, why have we apparently wandered so far from Canada's own soil classification system and why do we find little or no consistency with global (WoSIS) efforts described below? If, in the end, these authors focus on the hydrological properties (e.g. HSG: hydrological soil group), they should establish a clearer or more consistent terminology for the less essential (in this context) details? The manuscript could start from the goal of determining hydologic erodability and then proceed through a series a assembly and compilation steps starting with and adhering to a careful definition of terms?

This inconsistent terminology develops into acute confusion on page 11 (section 6.1). Here the SWAT model contains 200 predefined "soils". In order to avoid non-recognition of additional imported data, the Canadian researchers have manipulated their ID codes to allow ingest of more than 11k additional "unique soils". But we get zero information to confirm that the soil categorisation (level of detail of description) of the extant SWAT soil data matches the level of detail of the added 11k "unique" soils. One hopes and suspects that these researchers have matched apples with apples here, but they give us no evidence to show that they have not matched SWAT apples with SLC oranges. The 11k number invites concern. I used STATSGO2 for Montana, a state that borders and in fact shares watersheds with a substantial portion of the SLC regions. If I assume that the field MUKEY equates to what one might consider in terms of this activity "soil type", I find more than 4300 entries (soil samples?, soil profiles?) but only approximately 700 distinct MUKEY values. Just over the boarder to the north, we find 11,000 'entries' or 11,000 distinct soil 'types'?

The World Soil Information Service (WoSIS) records more than 90000 **profiles** (https://doi.org/ 10.5194/essd-9-1-2017) consisting (with the inclusion of depth / layer information) of more than 4 million records. They manage this international data deluge by adhering to analytical and physical properties described in the *GlobalSoilMap* (http://www.globalsoilmap.net) specifications. A reader wishes that this Canadian effort had adopted the same approach! Or, if they have followed the same approach despite their confusing terminology, they need to inform users and readers!

Page 13, line 284: "more than half of the soils in Canada". Not even close to true! Half of the soils under agricultural use for which profile data exist in a small southern portion of Canada!

Page 17, Section 7. The information conveyed here works very poorly or even not at all as a narrative. Make this into a flow chart or a bulleted list of steps?

Page 18, line 398: "Among the major contributions of this dataset, the calculation and/or estimation of variables not reported in the SLC database are of special importance." Perhaps true and perhaps this represents a substantial outcome of the skill and effort of this group. But does it have any relevance beyond this relatively small subset of Canadian soils? If the authors contend that it does have larger relevance, they have not made the case! If it does not have larger broader global relevance, does it belong in ESSD?