

Dear Editor and authors,

Thank you for the invitation to review this manuscript. This study compiles a series of datasets of soil properties and pedotransfer functions available for catchment scale modeling, evaluate them and provide an R open script for soil data derivation. I believe this is a valuable piece of work, however, I have several major concerns about its suitability for publication in its current shape but I believe it could be significantly improved. The main concern is the fact that there is no data openness in this work, the main data used is not available to assess reproducibility and another smaller dataset (independent dataset provided by agricultural company) is not even described. If the authors are willing to share the data and correct the identified problems, I believe this could be an excellent contribution.

My comments for improvement follows:

Introduction

Lines 39 -58 – The introduction could approach more deeply the importance/availability of all the soil properties mentioned in the abstract (Lines 23-25), however the examples provided seemed disconnected between each other. Alternatively, further explanation could be provided to justify giving those examples and not other.

Lines 64-67 – What would this allow the scientific community to do? Support modeling studies? Assist researchers in the decision for methodological approaches? Please provide a statement.

Materials and Methods

I was hoping to read here how you determine the compilation in table 1. How to ensure this dataset is complete?

Lines 92 -109 – These are the “soil properties most frequently required by environmental models” based on what? How did you determine this list?

Lines 112 – Regarding the statement “Local and national datasets provide more accurate input information”, I would say depends. If is a spatially explicit modeled database I would say definitely yes, if is point data with precise coordinates of the source of the data (e.g. LUCAS) I would say no. Of course, in this condition is not possible to address quantity of information (i.e. density of points) but one would expect local data to present higher data availability, but this is not mentioned in the text. Please be more precise on this matter.

Line 114-119 – I read this text several times, I understand you want to make the case regarding the previous sentence, but is not entirely clear please rephrase it.

Line 119 – 121 – Perhaps there is an expert on soil cracking in this team, but I have to admit I never used such variable. I had to search SWAT documentation, and in the EU only a small amount of soils are classified as Vertisols. Therefore, I question the authors to justify the importance of soil cracking in comparison with all the other properties listed.

Lines 122 – 126 – These are suggestions and not materials and methods, please move this to an appropriate section, or rephrase it if means you took those considerations when analyzing data.

Lines 127 -134 – same as before.

Lines 139 – 141 – Not clear. I understand there is partial data availability for comparison, but I don't understand what is the consequence of such approach. Please clarify the text.

Line 138 -139 – All right so this EU-HYDI is you data for validation right? It is reasonable to say this dataset hasn't been updated since 2013, and considering the report is actual, should I ask if you only considered measured values? Because I see in the report that part of this dataset contains estimated values. In addition to that, could you point the readers to the data itself? The possibility to reproduce the same analysis is necessary.

Line 152 -333– I honestly don't understand why you wrote this as a protocol format, please reformulate to describe the data analysis that you produced. This is written as a textbook that is not the purpose right? Moreover, out of the sudden I realized that there are datasets that have not been used (e.g depth to water table) and others that have been, could you synthesize that information?

Line 312 – How can we know about local fertilization schemes?

Line 319 - Locally independent measured dataset?? Provided by an agricultural company? How many samples? When samples were taken? Laboratorial methods? Statistical analysis? This is clearly an insufficient methods description.

Results & discussion

Porosity – Lines 396-399 – so what have you done regarding this? If 43% of the samples presented errors, was this data excluded? (explain for all parameters)

Soil erodibility – I thought the Renard et al 1997 was the most used version of the RUSLE model (almost 5 thousand citations according to semantic scholar), but in case I am not right would you provide an information about it?

[Predicting Soil Erosion by Water: A Guide to Conservation Planning With the Revised Universal Soil Loss Equation \(RUSLE\) \(usda.gov\)](#) page 65.

Either way, why testing only one equation?

Field capacity – Would be worth to make a reference to table 4 early in the beginning. (for all parameters)

Wilting point – I don't understand why the number of points available to assess VG (table 11) differs from the ptf (table 12)

Saturated hydraulic conductivity – found strange you didn't use this database [ESSD - SoilKsatDB: global database of soil saturated hydraulic conductivity measurements for geoscience applications \(copernicus.org\)](#) for comparisons also, any justification for that?

Line 625 - So these workflow are the result of your analysis, whereas you described the most efficient workflow for better data quality. Right?

Conclusions

Line 653 – “Key findings underscored the significance of local soil data over global or large-scale datasets in environmental modelling”, I’m afraid you haven’t provided hard evidence on this. Besides this Figure 17, you presented zero information about the sampling, the timing of the year, the laboratorial analysis of this procedure, among many other details. Please provide the sufficient info in order to assess if such assessment is even comparable.

Data availability – I find this justification rather poor considering we are not talking about personal information, nor information that could reduce the value of the land. There is a report online, and this work was paid with taxpayer’s money already more than 10 years ago. In the meanwhile, many things changed in science, and the open data is the new reality. Making this data available, in an open data journal as this one, would help the scientific community to overcome many obstacles in the hydrological modelling.

Greetings

Diana Vieira