Reply to Referees (ESSDD)

Interactive comment on “SoilKsatDB: global soil saturated hydraulic conductivity measurements for geoscience applications” by Surya Gupta et al.

Anonymous Referee #1

Review of SoilKsatDB: global soil saturated hydraulic conductivity measurements for geoscience applications by Surya Gupta et al. The saturated hydraulic conductivity Ksat dataset that the authors compiled is extremely useful and highly needed. The paper describes the datatset clearly and is well written and easy to follow. The initial analyses done with the new dataset are interesting as well. Some of the figures in the paper can easily be used in lectures on soil hydrology. I checked the csv file of the database (from the website given at the end of the paper) and it contains more columns than described in the paper. This is a bit confusing. I have very few comments on the paper itself and highly recommend publication of the paper after some minor revisions.

RE: We thank the Reviewer for the positive assessment of our manuscript and for the numerous comments and suggestions. In the revised version, we have clarified the methodology and the database description we explain now all columns in the paper that are shown in the database. We have also modified the manuscript based on your feedback/edits on the manuscript. In addition, we have provided answers to your questions as listed below (in red color).

Dataset:

Q: I checked the csv file of the database and found the use of the ? to indicate missing data a bit annoying (even though it can be easily replaced by NaN or some other identifier). In column “hzn_desgn” both “no data” and “?” are used for no data. This is a bit confusing. Also, there are columns that seem to only have missing data and aren’t defined in Table 2a: “usiteid”, “labsampnum”, “layer_sequence”, “db_13b”, “COLEws”, “w15bfm”, “adod wrd_ws13”, “cec7_cly”, “w15cly”, “ph_kcl”, “cec_sum”, “cec_nh4”. The column “site_obsdate” isn’t defined and explained in Table 2a and it isn’t clear what this is as it clearly isn’t a date. Similarly, the columns “hzn_desgn”, “w15bfm”, wpg2” are not described in Table 2a, nor shown in Table 2b.

RE: We made several changes of the database. We have now removed all empty columns from the KSatDB. "?" and "No data" was replaced by "NA". All columns sown in the database are now explained in Table 2a in the paper.

Q: I would find it very useful if the database also contained a column with your classification of the climate and the calculated texture % based on the Nemes et al. method. This would mean less double work for other researchers who want to use the data (and possible errors).

RE: Thank you for this suggestion! We have now added an overlay of points and majority of www.OpenLandMap.org layers in the table "sol_hydro.pnts_horizons_rm.rds" (https://doi.org/10.5281/zenodo.3752721). This will be continuously maintained and extended with other layers. A complete list of layers and their codes is available at: https://gitlab.com/openlandmap/global-layers. Following the Reviewer's suggestion, we have now incorporated climate classification information in the database: file name "sol_ksat.pnts_cl_pedo.csv" (see version 0.3, https://doi.org/10.5281/zenodo.3752721). Regarding the calculation of texture % based on the Nemes et al. method, we calculated the texture % only for the UNSODA dataset (as this was not provided in first place) using the methodology by Nemes and co-authors. All other datasets already have texture % information. We have modified the methodology section to make this point clear (see P3L28-30).

Q: I would find it useful if the headers contained not only the name but also the units but this is just a personal preference that helps to avoid errors when reusing the data.

RE: We prefer to separate the units from column names. Instead we recommend the user to refer to the documentation/metadata which is now listed both on the dataset repository (https://gitlab.com/openlandmap/compiled-ess-point-data-sets/-/tree/master/themes/sol/SoilHydroDB) and in the paper (see Table 2a and 2b). We believe no users should have a problem locating the metadata and using the data correctly.

Paper:

Q: In the introduction, the authors argue that it is important to have accurate information on the location of the data points but this argument is not clearly supported by examples. The authors invested a lot of effort in obtaining this data for sites that were already included in other databases but for which the database didn't have the location information. I think that this is highly useful but the argument could be stronger. The PTF example for the use of the database doesn't use any detailed information on the location of the measurements. The paper would be stronger if examples were given or if there was (more) discussion of applications for which this spatial precision is indeed necessary.

RE: We give some specific examples in section 4.4.

Q: In addition to the compilation of existing (national-scale) databases, the authors also actively searched for data from underrepresented areas. This is very useful but it is, however, not fully clear how underrepresented areas were defined or how exactly they searched for these additional data points. Was there a certain cutoff in terms of publication date? Did they search for data from specific countries or was it based solely on soil type or climate? A bit more information on how they searched for these studies and thus which studies were included (and which were not included) would be useful.

RE: Based on the global map shown in Figure 1, we looked for countries and regions without values reported in the existing databases. We made then specific literature research on "Ksat" values for a specific country (or region like 'arid regions in Africa). In some cases, we also contacted colleagues that worked in these regions to ask for data support. We have better clarified this in the revised manuscript (P4L10-15).

Q: The paper contains several very useful figures that compare Ksat values for different soil types. It would be useful if it was indicated on these figures for which soil types the mean values are statistically significantly different.

RE: We thank the reviewer for this suggestion. We have prepared a table to show the significant differences between each soil texture class for table 5 as well as Figure 3b. (See Tables ST1 and ST2 in the Supplementary Files).

Q: On P3L26, it is mentioned that the sand silt clay fractions were estimated based on the method of Nemes et al. but from the text and Venn diagram in Figure 2, it appears that these data were available for most of the papers/databases. Were they only estimated when they were not available already from the database? This is not so clear. How well did the Nemes method estimate the fractions when data were available?

RE: We apologize for the imprecision. We computed texture % using the method by Nemes et al. (2001) only for the UNSODA database, for which soil texture information was not directly available. We have now clarified this point in the manuscript (see P3L28-30).

Q: The authors develop a subjective accuracy score based on the location accuracy and the method. They state (P9L8) that they consider lab measurements more accurate than field measurements. Even though I understand what they mean, this was still a bit surprising to me as samples may be disturbed, suffer from compaction or smearing and are generally too small to contain a network of macropores. This is partly addressed in the discussion but some discussion (perhaps with a focus on accuracy vs precision?) and acknowledgements of the issues with soil samples in this part of the paper would be useful.

RE: We agree with the Reviewer. In the revised version, we removed the confidence degree based on the measurement method and only provided the positional accuracy based on the location. See subsection "Standardization and quality assignment".

Q: I know that there are different ways to use the word "sample" but here it is confusing to use the word for different things. I therefore suggest not to use the word sample for a datapoint, and to only use it to mean a soil sample (and thus for the laboratory measurements). In particular, "field measured soil samples" is a confusing use of the word sample. Also "temperate soil samples" seems to be used to indicate both field and lab (sample based) values from sites in a temperate climate. It would be better to reword these types of sentences to avoid any confusion.

RE: We have modified the text accordingly and used "field measured Ksat value" or "temperate-climate based Ksat values".

Q: The annotated pdf contains some additional suggestions (all minor) and highlights where the text can perhaps be improved a bit (these are just suggestions though, the paper is well written as it is).

RE: Thank you so much for the additional helpful suggestions. We have modified the text accordingly.

Reply to Referee (ESSDD)

Interactive comment on “SoilKsatDB: global soil saturated hydraulic conductivity measurements for geoscience applications” by Surya Gupta et al.

Anonymous Referee #2

Gupta and co-workers present an interesting dataset about pedo-hydrological properties. They collected data with a focus on saturated hydraulic conductivity from various publications and repositories around the globe. Such a dataset highly deserves publication and has strong potential to contribute to the advance of pedo-hydrological sciences. However, I see quite some room for improvement of the manuscript to really stretch out for this potential and to meet the standards of ESSD. Some of the co-authors are my "idol-pedologists", who are always inspiring my own research. I feel slightly humble and confused to find this manuscript in such a sloppy and imprecise setting about methods, scale, pedometrics and functional soil description. I have not found any methodological reference about the steps taken to derive, compile and evaluate the data. Instead, the amount of time to digitise and compile the data is emphasised. I am full of confidence that the authors can and will rework their study to a more coherent and scientifically founded state. I hope my comments can constructively guide this process.

RE: We thank the Reviewer for the critical assessment of our manuscript and for the numerous comments and suggestions. In the revised version, we have clarified the methodology and the database description and included additional analyses. We have improved in the dataset and removed the typos as you mentioned in your feedback. We have provided answers to your questions as listed below (in red).

Major comments:

Q: Clarification of the conceptual and methodological meta-information:

Throughout the manuscript the authors are not very shy in promoting the central role of Ksat for hydrological applications. Quite to the contrary, there is no word about the conceptual framing and implicit assumptions of Ksat and the respective methods to measure it. Ksat (saturated hydraulic conductivity) is commonly understood as the invert of the Darcy filter resistance (as implicitly argued in most of the manuscript). Ksat is also interpreted as infiltration capacity (as claimed in the abstract). The methods to measure saturated hydraulic conductivity and infiltration capacity differ strongly with respect of their conceptual assumptions. Infiltration capacity is even more under debate, since it has to account for surface conditions too. I clearly see that resolving this debate is not in the scope of this data publication. However, I recommend to be much more clear about your conceptual setting in general and to avoid overrating Ksat measurements. Moreover, I strongly assume that mixing different measurement techniques will inevitably introduce biases to the data. To my experience, each method has limits which lead to different estimates of Ksat. In addition, the repeatability of "free drain" experiments (i.e. ring infiltrometers and to some degree also Amoozemeter) is very limited. Tension-controlled measurements have a much better performance, which can really be repeated with similar readings. In the lab, the situation is much more controlled. But the difference between 100 ml and 250 ml ring samples can be substantial. Also here, different techniques and procedures might introduce biases. Such methodological biases cannot be recovered in the final dataset if they are not reported (at least where possible).

RE: We agree that it is essential to provide information related to methods. Now we have included the information of Ksat method (and soil texture, bulk density, and organic carbon methods) in the CSV file “sol_ksat.pnts_metadata.csv” (see version 0.3, https://doi.org/10.5281/zenodo.3752721) to recover the methodological bias in the final dataset. We also added a new table 4 listing all the methods used for the measurements and gave references to the methods.

Q: Global coverage and number of samples:

The authors have done a phantasmic job in compiling all the data. However,I am under the impression that there is little thought given to well-known scale issues. I understand that the authors try to leave this to the interpretation by the users of the database by reporting the geographic location. However, the manuscript holds several examples where coverage, data density and similar are referred to countries, continents or studies. I cannot really judge the value of the dataset based on the presented accumulation. Maybe defining a site as some pedological unit would be helpful. Alternatively, at least main textural and climatic classes could guide the overview? Tab. 1 lists the data sources. Half of the datasets contribute only 10% of the data points. Half of the data stems from one publication about Florida soils. Moreover, it is obvious that there is a substantial amount of data still out there, which has not been published in a way that you could locate it. This gives rise to three questions: How does the skewed distribution of data sources influence the final product? How does the skewed distribution of data points in general imprint on the final product? How could colleagues add their data to the dataset? I am also under the impression that the mere number of samples does not give me much insight with the necessary meta information about location, site conditions and method. 1000 double ring measurements at one sight might weigh little over 50 precise analyses with tension hood infiltrometers or lab measurements…

In addition, the dataset you describe actually contains 152042 entries for soil hydraulic properties. Ksat is only reported in 13267 entries. So why do you emphasise Ksat so much?

RE: We thank the Reviewer for appreciating our effort. With regards to specific points raised:

a) We have now included the information related to climate zones based on Köppen-Geiger climate zones map (Rubel and Kottek, 2010, Hamel et al., 2017) and pedological units based on openlandmap.org in the CSV file "sol_ksat.pnts_cl_pedo.csv" (see version0.3, https://doi.org/10.5281/zenodo.3752721)

b) The Reviewer is right, in the manuscript we mentioned that 50% of the Ksat data is from Florida and we agree that this would impact statistically the final product. We would like to give this liberty to the users to use this dataset as per their requirement.

C) Our database and code is publicly available (https://gitlab.com/openlandmap/compiled-ess-point-data-sets/-/tree/master/themes/sol/SoilHydroDB) and users can contribute new data by either opening a new issue or directly by adding code and doing a pull request (https://docs.gitlab.com/ee/user/project/merge_requests/creating_merge_requests.html).

Q: Confidence index:

Using a subjective confidence index about location and overall method appears rather unnecessarily sloppy to me. First of all, I suspect location much less of an issue than the reported values - especially at the scope of the dataset. The authors appear to emphasise otherwise. Second, I see quite easy to implement ways reducing subjectivity: For the location one could instead give some sort of standard deviation (e.g. if you only know the basin than the location is the centroid±half the basin's extent). For the actual value, I find it of dramatic importance to report the used method whenever possible. Simply assuming field measurements to be less trustworthy than lab ones has weak reasons. Understandably, the authors do not analyse any coherence with neighbouring measurements or possible biases in different labs. However, this essential meta information needs to be conveyed to allow others to make use of the data. This also holds for the analyses of texture and Corg.

RE: Thank you for this suggestion. We do not use a confidence index anymore and just list the location accuracy (as shown in Table 3). We have also emphasized in the text that the actual measurement errors are usually unknown and digitized legacy soil data from scientific reports and similar should be used with caution (P17L9-10).

Q: Pedo transfer functions:

I recommend to drop the topic of PTFs. The way it is introduced in the manuscript and the methods applied open hundreds of questions which I do not consider in the scope of the data publication. The current form does not adhere to the state of science in this field.

RE: Thank you for your feedback. Our main objective was to show pedotransfer functions as a way to use the dataset (although we understand that there is a plethora of additional applications for the dataset). To better convey this, we modified the text and better explained the purpose for this application (P7L3-7and P8L1-3). Moreover, we removed the section on multilinear polynomial regression, focused on PTFs derived from random forest (as state of the art approach), and better described the importance of different variables in the result section. We have now showed how we derived the PTF (https://github.com/ETHZ-repositories/Ksat_database_2020/blob/master/Ksat_data_PTF_supplimentry_code.pdf)

Minor comments:

Q: P1L2: Isn't the infiltration capacity controlling this partioning and it is due to the commonly used models that ksat is considered a key parameter? I suggest to avoid overly strong claims but to emphasise on the value of the data in its own realm.

RE: We modified the text in the abstract (P1L1).

Q: P1L2f.: Again, this is the concept but the physical processes are taking place in the soil pores. As some of the co-authors pioneer research in this domain, I can surely assume that we do not disagree about this. Hence, I think it is important to be precise about the conceptual underpinnings of the data.

RE: We modified the text in the abstract (P1L2).

Q: P1L4: There is substantial literature about the scope- and scale-dependency of transferring measured ksat values to model applications. Using many data points obtained from a rather difficult to control measurement procedure (i.e. ring infiltrometers, and amoozemeters) might end up in more blur due to the method than insight about infiltration capacity. In the same lines of thought, lab measurements of ksat in differently sized ring samples and under different methods are prone to generate unknown biases on the recorded values for different soil situations. Moreover, it is well known that different landscape settings (e.g. forest vs. agricultural lands) have substantial impact. Hence, I am a little reluctant to follow your argumentation and to be impressed by the mere number of records here.

RE: In the modified manuscript, we refer to the scale dependency (P15L14-15 and P16L1). We modified the text in the manuscript (P1L3)

Q: P1L6: "global database": How does your study relate to other globally available soil data products? How many classes are covered with how many samples? In which respect has standardisation been applied?

RE: In a new figure (Figure 3d), we list the number of samples per soil textural class. In this work, standardization refers to make units of datasets identical (this has been clarified in the manuscript - We modified the text in the abstract (P1L5).

Q: P1L7: "data density": Again, how does your data density relate to globally available soil maps/classes? I do not understand why the ranking of a country and continents shall be of importance. Most cover a broad range of climates and landscapes which might not be unique...

RE: Thank you for this question. Data density was provided to give an overview to the users about the compilation of data from different continents. In the revised version, we also provide information on distribution of samples across different climatic regions (P11L21-22). We have made some modifications to the text (P1L6-7). We also agree that it might be important to relate this data with soil maps/classes. Therefore, we overlaid the Ksat values on the openlandmap.org layer and extracted the values of soil classes (please see sol_ksat.pnts_cl_pedo.csv (see version 0.3, https://doi.org/10.5281/zenodo.3752721))

Q: P1L8: "other soil variables": Again, I cannot judge from the numbers given if and to what degree the samples are comparable. E.g. soil texture can be measured by quite a spectrum of methods with known biases. The retention properties are not fully covered by these more agronomically motivated references...

RE: We agree with the Reviewer. We have now provided the method for these properties as much as we could extract from the respective papers. Please have a look at the CSV file “sol_ksat.pnts_metadata.csv” (see version0.3 ’https://doi.org/10.5281/zenodo.3752721). We have also modified the text (P12L1-3)

Q: P1L11 "temperate climatic regions": Does this mean that your dataset mainly covers this climatic region? If so, maybe the title should include this.

RE: Dataset covers all climatic regions (this is quantified in the revised manuscript). Here, we extracted the Ksat values belonging to the temperate climate region by overlaying the climate zone map. Further, the PTF was derived using these points. Then, PTF was tested for Ksat values belonging to the tropical climate region.

Q: P1L12 "random forest": This statement appears rather generic to me. Given some data, a random forest is known to produce very good fits. Moreover, I do not understand the reference to temperate and lab based measurements. You mean that one subset refers to the climate region and the other subset to all climate regions but excluding field measurements? This is difficult to get and set into perspective. How can I differentiate between methodological and conceptual effects here? I mean, could it be that PTFs based on the given variables have been developed in and for lab samples in temperate regions and thus apply well for these but that for field measurements and other climatic regions, the PTFs miss an important predictor?

RE: Sorry for the confusion. In the manuscript, our goal was to address two different aspects.

1. Firstly, we overlaid the 13,267 points on the climate zone map (now explained in the method section) and extracted only those points where information on sand, clay, and bulk density was available. Then, we extracted only points in the temperate climate zone ksat values and fitted the model to 80% of these measurements using the Random forest approach. The fitted model was tested on the remaining temperate data points (20 %) and on tropical Ksat values. In this case, we mixed both lab and field measurements.
2. In the second case, we separated 13267 points based on lab and field methods (9162 and 4133, respectively). For lab data, we fitted the model based on 80% of the lab-based ksat values and tested it on the remaining (20%) lab-based data values and on all the field-based Ksat values. In this application, we did not differentiate between different climatic regions.

We have now clarified this in the methods (P7L3-7and P8L1-3).

Q: P1L18 "data license": I am not a fan of Zenodo to publish such valuable data. Why don’t you use a more geoscience specific, long-term available repository like Pangeae or GFZ-dataservice etc.?

RE: Thank you for your suggestions. We will consider these options in the future.

Q: P2L16f.: I do not understand this. https://esdac.jrc.ec.europa.eu/content/ 3d-soil-hydraulic-database-europe-1-km-and-250-m-resolution I assume that this is the respective data product and it is public. Do you mean the raw data behind the product? Since one of the co-authors is also author of the data product, why is it omitted?

RE: The publically available maps show the **predictions** of Ksat. However, the underlying measured data are not publicly available. We tried initially to ask this data from the authors, but due to government restrictions, they could not share.

Q: P2L21f.: Please specify the spectrum of methods for Ksat derivation.

RE: We have modified the text (P2L25-27). Please see also Table 4.

Q: P2L23ff: ESMs operate at scales where even topography is highly aggregated. RS products are very quick in claiming surface properties which only show weak coherence with soil water dynamics. The scale of RS products varies greatly but is well below the scale of ESMs. Honestly, I do not get your point here. It appears to me that you follow a quite classic but maybe not very contemporary conceptual model of soils as static filters which can be easily predicted once the filter resistance (or Ksat as the invert) is defined. This approach has its merits and does not counteract the value of your dataset. However, I would suggest to precisely clarify this conceptual setting and to refer these assumptions to the set of methods to derive the values of Ksat in the database.

RE: We have modified the sentence (P2L28-29) but we are not sure if we understood the reviewer correctly. In advanced Earth System models, the spatial resolution (~1km) also for very large regions is comparable for many RS-based products.

Q: P2L26ff.: I agree. In my opinion, this is a discussion topic on how to define a standard for pedohydrological data to ease data processing. I came across several rather generic formulations so far which I strongly suggest to revise and recompile in a discussion section - or simply omit.

RE: We agree. We have now removed the sentence.

Q: P3L9: If I am not mistaken the only methodological citation goes to machine learning, which you do not at all tackle in the manuscript. Please strongly rework the manuscript to refer to the state of pedological and hydrological sciences.

RE: Thanks for this suggestion. Now, we have modified the text. Please see the subsection "statistical modelling of Ksat".

Q: P3L27f.: Sorry, but coordinate conversion is not an issue any longer as long the geographic system / EPSG code is given. You can directly use https://proj.org with the software of your choice... or https://espg.io online.

RE: The Reviewer is right. However, to facilitate the user, we have standardized the geographic system.

Q: P3L33f.: I thank you very much for doing this work and providing the data. However, I do not expect digitising to be an issue worth debating here. There are many ways including automated processing. Definitely MS Word is not a necessary step but your choice of processing...

RE: We agree with the Reviewer. We removed this sentence from the manuscript.

Q: Table 2a: The README in the dataset gives slightly different entries. Please make coherent.

RE: The README file has been corrected.

Q: Table 2b: I do not see why table 2b is given. All information is or can be provided in table 2a already.

RE: The table provides a glimpse of the CSV file and its inclusion was recommended by the editor.

Q: Table 3: As stated above, I suggest to fully rework the matter of confidence measures. Your proposed subjective index can only obscure the data – Especially since you combine spatial precision with lab/field method assumptions.

RE: We have modified this part of the manuscript. We have provided the Ksat method for each study and separated it from location accuracy (please have a look at table 3).

Q: P7 Sec. 2.3 "Standardization and quality assignment": I do not see if or how this has been performed. Despite agreeing to your judgement about very small Ksat values, I would be interested why the colleagues did not perform such "cleaning" in the original data. How can they possibly measure 10e-14m/day? I suspect some strange averaging with small numbers behind this. What do you mean with "cross-checking"?

RE: In the SWIG database, 1845 Ksat measurements were extracted from the literature, and Ksat for other samples were computed using the infiltration database, fitting infiltration data series to Ksat. Some Ksat values computed using infiltration database were less than 10^{-14} m/day, which seems unreasonable, so these values were not included in the database. We have modified in the text (P6L14-16).

Cross-checking: Here cross-checking means that we crosschecked all the datasets to avoid the mistakes considering the same dataset two times. For example, SWIG database included the database from Zhao et al. (2018) in the Tibetan plateau. We removed Zhao et al (2018) from SWIG and presented the data of Zhao as separated database.

Q: Table 4 bottom row: I do not understand 32*. You report 11635 Samples for texture but 32 without texture class? Once you know the composition, the texture class is defined.? How many of the Ksat_lab samples have been measured in the field, too? I think this table is not very helpful. Maybe once the main topics and questions are clarified, a couple of easy plots would be more helpful to understand the dataset?

RE: We thank the Reviewer for noticing this - These 32 values in the soil texture class are errors. It means that the total of sand, silt, and clay % is more or less than 102 or 98%. However, after reanalyzing the data, we found that 75 values have the same problem. Hence, we provided soil texture class as “Error”. We have modified in the text (P12L8-9).

Q: P9L1"SWIG": Am I right assuming that this dataset holds 65 samples? If this is roughly 1% of the total number, I am not quite sure why this is highlighted here. Again, I would strongly recommend to include such specific metadata in the final table/database – especially because I suspect many other samples to suffer from similar issues.

RE: The SWIG dataset holds 3637 samples. No we have added the Ksat methods. Please have a look at “sol_ksat.pnts_metadata.csv” (see version 0.3, https://doi.org/10.5281/zenodo.3752721) file and table 4)

Q: P9L4f.: Why? Are the methods mostly unknown? I suspect this to be of dramatic importance to report the used method whenever possible.

RE: Now we have provided the method information for each sample (Please see "sol_ksat.pnts_metadata.csv" (see version 0.3, https://doi.org/10.5281/zenodo.3752721) file and table 4)

Q: P9L6: See above about the index.

RE: As stated above, we don't use the confidence index anymore.

Q: P9L9: I strongly disagree. Why should a sample carried to the lab have a better depth precision than an experiment in the field? The procedure to measure the depth is one of the most simple ones in pedology. The issue might be about the actual measurement though. E.g. if I use an Amoozemeter, I can precisely position the water supply probe but the recorded value might not reflect Ksat in the sense of hydrological models...

RE: We agree with the Reviewer that it might not be the correct way to provide a subjective confidence degree based on the measurement method. Hence, in the revised version, we removed the confidence degree based on the measurement method and only provided the positional accuracy based on the location (P6L20-24).

Q: P9L10f.: This points right into the essence of whether Ksat reflects infiltration capacity (as claimed in the abstract)or if it is the invert of the Darcy filter resistance (as implicitly argued thoughout the manuscript). I recommend to be much more clear about your conceptual setting again. With respect to the air entry and/or full saturation (which I see as two distinct issues) there is clear reference in the respective measurement procedures. Hence I would not agree that lab and field mostly differ in this respect but in the definition of the sample boundaries. In the lab, the sample is (more or less) well contained in a ring (with all known issues about it). In the field, the lateral component of capillary water movement is mostly unknown. In addition, there is little control about the vertical extent of the sampled location and conductive macropores and/or less permeable cross-sections... (to name one example).

RE: We agree with the Reviewer and modified in the text (P6L20-24).

Q: P9L12: Why should spatial accuracy (I suspect something like numbers of digits) be a quality attribute? Sec. 2.4: I can not at all follow your method here. What kind of PTF, what predictors, what training sets etc. pp. As stated above, I suggest to remove the PTFs.

RE: We have modified the text accordingly to make it more clear to the users (P7L3-7and P8L1-3).

Q: P10L15 "13,267 values": Please clarify this number (which I see is the count in the file). In Table 4 you report 11,727 from field and lab (13,294 including those without texture classes).

RE: It is because; there are 4 studies in the dataset which have both field and lab measurements. We mentioned this in the metadata CSV file “sol_ksat.pnts_metadata.csv” (see version 0.3, https://doi.org/10.5281/zenodo.3752721).

Q: P10L15 "sites": What is counted as one site?

RE: One site is equal to one location id (Combination of latitude and longitude).

Q: P10L17: I find this list very difficult. You mix countries and continents. What is the information in it? Maybe it would be better to define the distribution of sites? Next line you refer to the state of Florida with half the samples…

RE: We thank the Reviewer for pointing this out. Now we have given the Ksat points distribution based on continents and climate region (P11L21-23).

Q: P10L21: Sorry, but the numbers in table 4 are slightly different... Moreover, I do not gain any insight from them

RE: We thank the Reviewer for pointing this out. In the revised version, we rechecked the numbers and fixed typos. It is important to show the mean values of soil properties under various soil texture classes for the users.

Q: P10L24: What are statistical properties?

RE: We have modified the subsection from “Statistical properties” to “Statistical properties of SoilKsatDB”.

Q: Fig 2: I find this plot not only superfluous but reporting incorrect proportions. Please drop.

RE: We modified the captions to highlight that the proportions are not correct. However, we prefer to keep the figure because it is illustrative to show for how many samples the different soil properties are measured.

Q: Fig 3b: I do not understand this. A) Table 1 gives far more than 9 databases. B) Why should I look at a distribution of Ksat per database (holding an unspecified ensemble of sites) instead of any other site attribute?

RE: Thank you for your feedback. It is illustrative to show that databases with many field data and from different regions show the highest spread of data. Now we have also added the violin plot for soil texture classes (see figure 3).

Q: Fig 4: Please keep the colour coding static! Maybe convert the counts to percentages of the data? How about plotting all plots in one line with the respective marginal distributions? This is one of the most insightful plots and deserved far more description in the caption and text.

RE: Thanks for noticing this. We have now made the color coding static for figure 4 and revised the captions.

Q: P12L6f.: This does not surprise me. However, you address this topic later. Why do you refer to it here?

RE: We incorporated figure 7 as new panel in Figure 3 to present the statistics of measured value in one concise Figure. In this section, we just report the key differences and discuss the origin of the differences later on.

References

Hamel, P., Falinski, K., Sharp, R., Auerbach, D. A., Sánchez-Canales, M., & Dennedy-Frank, P. J. (2017). Sediment delivery modeling in practice: Comparing the effects of watershed characteristics and data resolution across hydroclimatic regions. Science of the Total Environment, 580, 1381-1388.

Rubel, F., & Kottek, M. (2010). Observed and projected climate shifts 1901–2100 depicted by world maps of the Köppen-Geiger climate classification. Meteorologische Zeitschrift, 19(2), 135-141.

Kutílek, M., & Krejca, M. (1987). Three-parameter infiltration equation of Philip type. Vodohosp. ˇCas, 35, 52-61.

Haverkamp, R., Ross, P. J., Smettem, K. R. J., & Parlange, J. Y. (1994). Three-dimensional analysis of infiltration from the disc infiltrometer: 2. Physically based infiltration equation. Water Resources Research, 30(11), 2931-2935.

Rahmati, M., Weihermüller, L., Vanderborght, J., Pachepsky, Y. A., Mao, L., Sadeghi, S. H., ... & Schütt, B. (2018). Development and analysis of the Soil Water Infiltration Global database. Earth system science data, 10, 1237-1263.

Reply to Dr. Attila Nemes (ESSDD)

Short Comments (Dr. Attila Nemes):

Interactive comment on "SoilKsatDB: global soil saturated hydraulic conductivity measurements for geoscience applications" by Surya Gupta et al.

After reading the paper I take the liberty of submitting a few uninvited recommendations – not a full review - to the authors while this paper is still in the review phase. I give a lot of credit to Reviewer #2's remarks and I strongly encourage the authors to clarify a substantial number of issues around the database in order to prevent avoidable criticism later. I congratulate the authors on the initiative and effort–assembling any large and heterogeneous database of the like is a never-ending fight. Yet, I think the documentation of the data currently stops short of where it should be and leaves too many doubts about the actual contents and its meta-information. I try to add rather that repeat earlier comments by the Reviewers.

RE: We thank Dr. Nemes for appreciating our effort, for the positive assessment of our work, and for the additional suggestions and comments. We agree with Dr. Nemes and the Reviewers that some parts of the manuscript needed improvement (in terms of clarity in the data description, analyses made and in the discussion part).

Q: In terms of the data and the database, my first focus is primarily but not solely on Table 2a. It is cited that the'codes', which Iinterpret as the field names that are adopted from the USDA NCSS database. I can recognize some of that, yes. However, I need to warn that most of the larger data sources taken advantage here will not hold data that adhere to many of those codes and the definitions behind them in the USDA NCSS database. Just as examples, those fields that have 'clod' in their names will likely not be possible to match due to methodological differences (i.e. clod vs core measurements), and therefore this documentation will be misleading and infuses confusion for later users. Ever since the first such international databases were published – including those with my involvement – the need and quest remains to be clear and specific about such details as methods, definitions, and the like. The USDA NCSS Soil Characterization Database sets some great example in that sense, but it cannot be unconditionally followed when the data in question are either mixed or do not adhere to those definitions/standards. I strongly suggest revising the documentation accordingly. This is better done now than later exploited by users and/or potentially hindering advancement in science.

RE: For practical purposes, we have tried to avoid creating yet another soil standard and have used instead some well-documented soil laboratory data standard such as the NCSS Soil Characterisation database (https://ncsslabdatamart.sc.egov.usda.gov/). Having said that, we also agree that this might create confusion, as computation methods are different. Hence, we have changed the headers name for most of the variables.

Q: Some additional specifics based on Table 2a, which does not cover the entire extent of the database (38 columns of information/data): - hzn_top/bottom appears to refer to horizon/layer designation, and not sample depths as suggested by the description and as also suggested by the examples in Table 2b - db_od: are all the data surely from oven dried samples? - Water retention data (w6, w10, w3, w15): Please clarify the methods and change the code/field names to the appropriate ones, once USDA NCSS is emulated. They have multiple data columns for several of those, differing in methodology. - Particle-size data: were all the data really given in the FAO/USDA format, and if not, then possible to interpolate with no specific challenges? Please confirm. - OC: this has been a source of grand confusion in more than one past database, and the language used throughout the paper is soft about it (at some point only calling it (OC – organic content). Please be explicit about handling this variable – to what extent conversion was needed from the publications and how it was done. - Ksat: Was Ksat always published in the source? Did it have to be calculated from infiltration data? Please be explicit about the methods, I do not recall seeing it.

RE: Now, we have provided the methods for soil texture, OC, bulk density, and Ksat (as much information as we could extract from the papers). Please look at CSV file "sol_ksat.pnts_metadata.csv" (see version 0.3, https://doi.org/10.5281/zenodo.3752721).

Q: Some comments/questions with respect to the pedotransfer part of the paper: With respect to the PTF comparisons, I think the authors left a lot on the table and stripped themselves from greater potential impact. The temperate-tropical comparison is well known, and the field-lab aspect could have been explored much deeper with not too excessive work.

RE: Thanks for this suggestion. We have tried to discuss the field-lab comparison in more detail in the "Discussion" section.

I invite the authors to include discussion on any locations/data for which field and lab Ksat was co-existent and whether those were handled/explored in some specific way. It is rare to have that capability.

RE: Thank you for the good suggestion. Unfortunately, only 28 Ksat values are available that have both field and lab values. Therefore, it is not possible to conduct this test with such a few data.

I think excluding 15% of the data in exchange for OC to be part of the models could have been an affordable loss – but the authors will likely offer a big-picture response to that. Bad correlation with Ksat does not seem to be unique for this variable. Could the authors include a third metric for a measure of bias? I can see greater spread in Figure 5 b and d than in a and c, but I cannot readily comprehend the claimed 'bias' from those two plots.

RE: Now we have added the bias in the manuscript (P13L24-28)

With respect to the offered discussion on lab vs. field results: I can accept the offered reasons as part of the big picture but lack any mention of e.g. measurement scale. Let me simply refer to the work by Ghanbarian et al. (2016) (10.1016/j.catena.2016.10.015) who explored the effect of sample dimensions on Ksat measurements – and that is only the laboratory part of this question. The presence of top-to-bottom connected (macro-)pores in a soil sample can also go both ways! Yes the taker of the sample may be tempted to avoid marcopores/cracks, but a short sample has greater chances for top-bottom connected clusters than at all one. I just wanted to indicate that there is much more that could/should be added here. In terms of field measurements, methodology may matter a lot as well.

RE: We agree with Dr. Nemes. Now, we have discussed in the manuscript (P15L13-14 and P16L1).

With respect to the offered discussion on temperate vs. tropical findings: Again, I can accept the offered points here, but there is likely more to the differences, and the authors could profit from expanding on this, in case PTFs remain part of this data paper. To mention one – a well-known one – the min-max range of particle-size metrics typically does not allow one to appreciate the differences in textural distribution between prevailing soils of those two climate regions. That very simply makes the tropical soils – and potentially their pore network types un- or underrepresented in any temperate PTF.

RE: We rephrased the paragraph and referred to different clay mineralogy and different soil formation processes.

And finally two short comments on the text: I suggest rewriting/reorganizing lines 1-15 of page 13 a bit. I found it very difficult to comprehend it because of the order of values and the many subsequent mentions of CCC and RMSE. Many values are very similar, and for CCC high value is good, for RMSE it is the opposite. Are any of the metrics significantly different between the MPR and RF methods?

RE: We have now modified this section (P13L23-28).

I suggest including an explicit warning to the user about the scale of applicability, especially where the assigned quality metric is high (meaning location is uncertain). A difference of 10km looks small at the world scale but may not serve any smaller scale work too well. The true point may almost fall into a different country in some cases.

RE: In the revised version, we highlight such ‘warning’ in the section on limitations of the database (P17L9-10).

SoilKsatDB: global soil saturated hydraulic conductivity measurements for geoscience applications

Surya Gupta[1], Tomislav Hengl[2], Peter Lehmann[1], Sara Bonetti[1,3], and Dani Or[1,4]
[1]Soil and Terrestrial Environmental Physics, Department of Environmental Systems Science, ETH, Zürich, Switzerland
[2]OpenGeoHub foundation / EnvirometriX, Wageningen, the Netherlands
[3]Institute for Sustainable Resources, University College London, London, UK
[4]Division of Hydrologic Sciences, Desert Research Institute, Reno, NV, USA

Correspondence: Gupta S.
surya.gupta@usys.ethz.ch

Abstract. Saturated soil hydraulic conductivity (Ksat) is a key parameter in many hydrological and climatic modeling applications. Ksat values are primarily determined from soil textural properties and may vary over several orders of magnitude. Despite availability of Ksat datasets in the literature, significant efforts are required to import and combine the data before it can be used for specific applications. In this work, a total of 13,267 Ksat measurements from 1,910 sites were assembled from published literature and other sources, standardized (units made identical), and quality-checked in order to provide a global database of soil saturated hydraulic conductivity (SoilKsatDB). The SoilKsatDB covers most regions across the globe, with the highest number of Ksat measurements from North America, followed by Europe, Asia, South America, Africa, and Australia. In addition to Ksat, other soil variables such as soil texture (11,591 measurements), bulk density (11,269 measurements), soil organic carbon (9,787 measurements), field capacity (7,389) and wilting point (7,418) are also included in the dataset. To show an application of SoilKsatDB, we fit Ksat pedotransfer functions (PTFs) for temperate regions and laboratory-based soil properties (sand and clay content, bulk density). Accurate models can be fitted using a Random Forest machine learning algorithm (best concordance correlation coefficient (CCC) = 0.70 and CCC = 0.73 for temperate and laboratory-based measurements, respectively). However, when these temperate and laboratory based Ksat PTFs are applied to soil samples from tropical climates and field measurements, respectively, the model performance is significantly lower (CCC = 0.52 for tropical and CCC = 0.10 for field samples). These results indicate that there are significant differences between Ksat data collected in temperate and tropical regions and measured in lab or the field. The SoilKsatDB dataset is available at *'version 0.3'* https://doi.org/10.5281/zenodo.3752721 (Gupta et al., 2020) and the code used to extract the data from the literature, for the quality control and applied random forest machine learning approach is publicly available under an open data license.

1 Introduction

Soil saturated hydraulic conductivity (Ksat) describes the rate of water movement through water saturated soils and is defined as the ratio between water flux and hydraulic gradient (Amoozegar and Warrick, 1986). It is a key variable in a number of hydrological, geomorphological, and climatological applications, such as rainfall partitioning into infiltration and runoff

(Vereecken et al., 2010), optimal irrigation design (Hu et al., 2015), as well as the prediction of natural hazards including catastrophic floods and landslides (Batjes, 1996; Gliński et al., 2000; Zhang et al., 2018). Accurate measurements of Ksat in the laboratory and field are laborious and time consuming and most samples are taken from agricultural soils (Romano and Palladino, 2002).

Efforts to produce reliable and spatially refined datasets of hydraulic properties date back to the 1970's with the proliferation of distributed hydrologic and climatic modeling. Some of these early notable works also provided basic databases (some of which are used in this study) for Australia (McKenzie et al., 2008; Forrest et al., 1985), Belgium (Vereecken et al., 2017; Cornelis et al., 2001), Brazil (Tomasella et al., 2000, 2003; Ottoni et al., 2018), France (Bruand et al., 2004), Germany (Horn et al., 1991; Krahmer et al., 1995), Hungary (Nemes, 2002), the Netherlands (Wösten et al., 2001), Poland (Glinski et al., 1991), and USA (Rawls et al., 1982). Nemes (2011) discussed the available datasets on Ksat and other hydro-physical properties in detail. Collaborative efforts have resulted in the compilation of multiple databases, including the Unsaturated Soil Hydraulic Database (UNSODA) (Nemes et al., 2001), the Grenoble Catalogue of Soils (GRIZZLY) (Haverkamp et al., 1998), and the Mualem cataloge (Mualem, 1976) - these, however, focused on soil types and not on the spatial context of Ksat mapping. In an effort to provide spatial context, Jarvis et al. (2013), Rahmati et al. (2018) and Schindler and Müller (2017) published global databases for soil hydraulic and soil physical properties. Likewise, the European soil data center also started projects such as SPADE (Hiederer et al., 2006) and HYPRES (Wösten et al., 2000), for generating spatially referenced databases for several countries. Since HYPRES represents only western European countries, Weynants et al. (2013) gathered data from 18 countries and developed the European Hydropedological Data Inventory (EU-HYDI) database - this dataset is, however, not publicly available and was not included in this compilation. The datasets mentioned above cover almost all climatic zones except tropical regions, where Ksat values can be significantly different due to the strong local weathering processes and different clay mineralogy (Hodnett and Tomasella, 2002). Recently, Ottoni et al. (2018) published a dataset named HYBRAS (Hydrophysical Database for Brazilian Soils) improving the coverage of South American tropical regions. In addition, Rahmati et al. (2018) recently published the Soil Water Infiltration Global database (SWIG) collecting information on Ksat for the whole globe. In SWIG database, some Ksat values were extracted from literature and other Ksat values were deduced from infiltration time series. In contrast to lab measurements that determine Ksat as ratio of flux density to gradient, infiltration-based methods determine Ksat by fitting infiltration dynamics to parametric models (using three-parameter infiltration equation of Philip (Kutílek and Krejca, 1987) or simplified form of Haverkamp et al. (1994)).

The ever increasing demand for highly resolved description of surface processes require commensurate advances in Ksat representation for modern Earth System Model (ESM) applications. Several existing Ksat datasets miss either coordinates or these have been recorded with unknown accuracy thus limiting their applications for spatial modeling. For example, the SWIG dataset misses information on soil depth and assigns a single coordinate for entire watersheds. Similarly, the UNSODA dataset does not provide coordinates and soil texture information for all samples. For a few locations, HYBRAS uses a different coordinate system. Taken together, these limitations highlight that, to prepare spatially referenced global Ksat datasets for large scale applications, a serious effort to compile, standardize and quality check all literature (available publicly) is often required.

The objective of the work here is to provide a new global standardized Ksat database (SoilKsatDB) that can be used for geoscience applications. To do so, a total of 13,267 Ksat measurements have been collected, standardized, and cross-checked to produce a harmonized compilation which is analysis-ready (i.e., it can directly be used for model fitting and spatial analysis). We compiled data from existing datasets and, to improve the spatial coverage in regions with sparse data, we further conducted a literature search to include Ksat measurements in geographic areas that were not yet covered in other existing databases. In the manuscript, we first describe the data compilation process and then describe methodological steps used to spatially reference, filter, and standardize the existing datasets. As an illustrative application of the dataset, we derive PTFs for different regions and measurement methods and discuss their transferability to other regions/measurement methodologies. We fully document all importing, standardization and binding steps using the R environment for statistical computing (R Core Team, 2013), so that we can collect feedback from other researchers and increase the speed of further updates and improvements. The newly created data set (SoilKsatDB) can be accessed via *'version 0.3'* https://doi.org/10.5281/zenodo.3752721 and directly used to test various Machine Learning algorithms (Casalicchio et al., 2017).

2 Methods and materials

2.1 Data sources

To locate and obtain all compatible datasets for compilation, a literature search was conducted using different search engines, including Science Direct (https://www.sciencedirect.com/), Google Scholar (https://scholar.google.com/) and Scopus (https://www.scopus.com). We searched soil hydraulic conductivity datasets using keywords such as *"saturated hydraulic conductivity database"*, *"Ksat"*, and similar. The collected datasets are listed in Table 1 together with number of Ksat observations for each study, and can be classified into three main categories, namely: i) Existing datasets (in form of tables) published and archived with a DOI in a peer-review publication; ii) legacy datasets in paper/document format (e.g., legacy reports, PhD theses, and scientific studies), iii) on-line materials.

Existing datasets include published datasets such as HYBRAS (Ottoni et al., 2018), UNSODA (Nemes et al., 2001), SWIG (Rahmati et al., 2018), and the soil hydraulic properties over the Tibetan Plateau (Zhao et al., 2018), from which we extracted the required information as described in Table 2a. The major challenge with making the existing datasets compatible for binding (standardization, removing redundancy), was to obtain the locations for a particular sample as well as the corresponding measurement depths. For instance, the UNSODA database completely lacks geographical locations. To fill the gaps and make the data suitable also for spatial analysis, we used Google Earth to find the coordinates based on the given location (generally an address or a location name). We separated the UNSODA data based on laboratory and field measurements and we computed sand, silt and clay contents based on the particle diameters between 0-2 µm (clay), 2-50 µm (silt), and >50 µm (sand) from the available particle-size data, assuming a log-normal distribution as described in Nemes et al. (2001). We further note that, in some datasets, the coordinates were missing or reported in diverse coordinate systems. For example, in the HYBRAS database, the locations needed to be converted from UTM to a decimal degrees. In the SWIG database, the information related to location (coordinates for each point), soil depth and measurement method (laboratory or field) was completely missing, so we went through each publication referenced in Rahmati et al. (2018) (except the unpublished literature) and added coordinates and applied the necessary conversions.

In the case of legacy datasets (paper or document format, data from journals, theses, and legacy reports with and without peer-reviewed publications), we invested a significant effort to digitize tabular data, clean it and make it analysis-ready. After the digitization process, all data values were cross-checked one more time with the original PDFs to avoid any artifacts or error in the final database.

Two datasets were also collected directly from project websites that might be peer reviewed such as the NASA project based on hydraulic and thermal conductivity (retrieved from https://daac.ornl.gov/FIFE/guides/Soil_Hydraulic_Conductivity_Data.html and described in Kanemasu (1994)) and the Florida database from Grunwald (2020).

There are many biomes and climatic regions, such as desert dunes, peatlands and frozen soils, where very few data of Ksat were publicly available. Because it is essential for global modeling to provide some values or range to reduce the uncertainty in the spatial maps, we have also intensively searched for these areas and, in addition to the major datasets (SWIG, UNSODA HYBRAS), we have also found several minor studies (that contain less than 5 Ksat measurements) to cover these regions. We thus digitized Ksat values from these studies (shown either in bar charts or line plots), georeferenced the maps where necessary, and then converted the data into tabular form. All these datasets are also listed in Table 1. In some cases, we also contacted colleagues that worked in these regions to ask for data support.

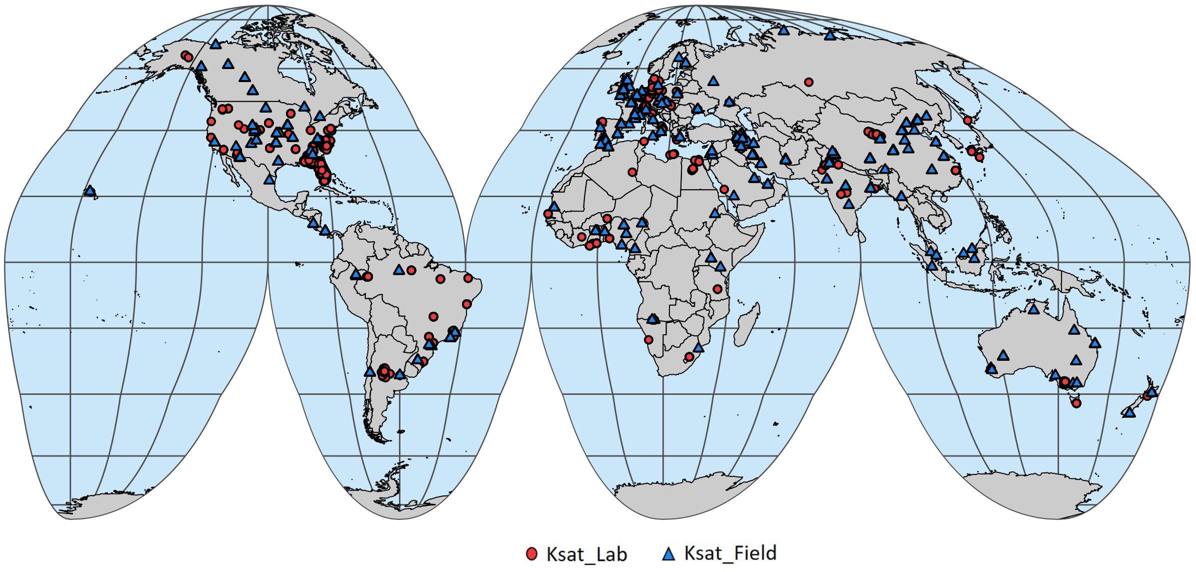

Figure 1. Spatial distribution of Ksat points (red and blue for laboratory and field measurements, respectively) in the SoilKsatDB. A total of 1,910 spatial locations are on this map.

Table 1. List of reference articles and digitized Ksat datasets, and number of points (N) per data set used to generate the new SoilKsatDB product.

Reference	N	Reference	N	Reference	N
Rycroft et al. (1975)	1	Abagandura et al. (2017)	3	Jabro (1992)	18
Waddington and Roulet (1997)	1	Habel (2013)	3	Greenwood and Buttle (2014)	18
Takahashi (1997)	1	Nyman et al. (2011)	3	Wang et al. (2008)	19
Katimon and Hassan (1997)	1	Bhattacharyya et al. (2006)	4	Deshmukh et al. (2014)	19
El-Shafei et al. (1994)	1	Lopes et al. (2020)	4	Price et al. (2010)	20
Lopez et al. (2015)	1	Yasin and Yulnafatmawita (2018)	4	Bonsu and Masopeh (1996)	24
Kramarenko et al. (2019)	1	Daniel et al. (2017)	6	Bambra (2016)	24
Zakaria (1992)	1	Anapalli et al. (2005)	7	Verburg et al. (2001)	26
Ramli (1999)	1	Arend (1941)	7	Southard and Buol (1988)	27
Singh et al. (2011)	1	Helbig et al. (2013)	7	Chang (2010)	30
Campbell et al. (1977)	1	Gwenzi et al. (2011)	7	Yao et al. (2013)	33
Chief et al. (2008)	1	Päivänen et al. (1973)	9	Becker et al. (2018)	34
Conedera et al. (2003)	1	Mahapatra and Jha (2019)	9	Baird et al. (2017)	50
Ebel et al. (2012)	1	Amer et al. (2009)	9	Keisling (1974)	56
Ferreira et al. (2005)	1	Radcliffe et al. (1990)	10	Rahimy (2011)	56
Imeson et al. (1992)	1	Vogeler et al. (2019)	10	Hao et al. (2019)	57
Johansen et al. (2001)	1	Singh et al. (2006)	10	Kanemasu (1994)	60
Lamara and Derriche (2008)	1	Kelly et al. (2014)	10	Tete-Mensah (1993)	60
Parks and Cundy (1989)	1	Elnaggar (2017)	11	Zhao et al. (2018)	65
Ravi et al. (2017)	1	Ganiyu et al. (2018)	12	Hinton (2016)	77
Smettem and Ross (1992)	1	Cisneros et al. (1999)	12	Vieira and Fernandes (2004)	86
Helbig et al. (2013)	2	Niemeyer et al. (2014)	12	Houghton (2011)	88
Boike et al. (1998)	2	Sharratt (1990)	14	Tian et al. (2017)	91
Andrade (1971)	2	Habecker et al. (1990)	14	Li et al. (2017)	118
Beyer et al. (2015)	2	Nielsen et al. (1973)	14	Forrest et al. (1985)	118
Blake et al. (2010)	2	Robbins (1977)	15	Richard and Lüscher (1983/87)	121
Bonell and Williams (1986)	2	Sonneveld et al. (2005)	15	Sanzeni et al. (2013)	127
Kutiel et al. (1995)	2	Quinton et al. (2008)	16	Vereecken et al. (2017)	145
Martin and Moody (2001)	2	Simmons (2014)	16	Coelho (1974)	176
Mott et al. (1979)	2	Ouattara (1977)	17	Kool et al. (1986)	240
Rab (1996)	2	Hardie et al. (2011)	17	Nemes et al. (2001)	283
Soracco et al. (2010)	2	Baird (1997)	17	Ottoni et al. (2018)	326
Varela et al. (2015)	2	Kirby et al. (2001)	17	Rahmati et al. (2018)	3637
Sayok et al. (2007)	3	Yoon (2009)	17	Grunwald (2020)	6532

2.2 Georeferencing Ksat values

Georeferencing of Ksat measurements is important for using data for local, regional or global spatial modeling. Once georeferenced, points can be directly used in hydrological and land surface models. Although many studies provided information on the geographical location of the measurements, the studies conducted in the 70's and 80's only provided the name of the locations and approximate distance from the exact location. Therefore, we extracted the latitude and longitude of the location using Google maps for some datasets (which did not provide the spatial locations). We digitized provided maps or sketches with locations of the points. We first georeferenced these maps using ESRI ArcGIS software (v10.3) and then digitized the coordinates from georeferenced images. Some of the documents we digitized (e.g. Nemes et al. (2001)) provided the names specific locations, and hence we used Google Earth to obtain the coordinates. We estimate that the spatial location accuracy of these points is roughly between 0 to 5 km. Similarly, spatial maps in jpg format (e.g. Becker et al. (2018)) were geo-referenced with 100–500 m location accuracy. In contrast, few studies (e.g. Yoon (2009)) provided the extract location of the sampling with assumed location accuracy of 10–20 m.

2.3 Standardization and quality assignment

The database was cleaned to remove unrealistic low values. For example, In the SWIG database, Ksat values computed using infiltration time series were less than 10^{-14} m/day, which seems unreasonable, so they were not included in the database. All datasets were cross-checked to avoid redundancy. For example, UNSODA data consist of Vereecken et al. (2017) and Richard and Lüscher (1983/87) datasets and SWIG database used Zhao et al. (2018). Hence we removed these datasets from UNSODA and SWIG database and used the original sources. Moreover, in the SWIG database, soil depth information was not available, so we assumed that infiltration experiments were conducted close to the surface and assigned a depth of 0–20 cm.

To describe position accuracy of each dataset, we assigned each Ksat value to one of seven 'accuracy classes' ranging from highest (0 - 100 m) to lowest accuracy (more than 10000 m or non available information (NA)). For example, Forrest et al. (1985), Zhao et al. (2018) and Ottoni et al. (2018) provided detailed site coordinates, thus we assigned a location accuracy of 0-100 m (i.e., highly accurate) (see Table 3 for more details). After data extraction from literature, geo-referencing and standardization, all information was collected in tabulated form in the new data base SoilKsatDB *'version 0.3'* (https://doi.org/10.5281/zenodo.3752721). The database consists of 22 columns (various sample properties) and 13,268 rows (a header and 13,267 samples). An excerpt of the database with some key properties is shown in Table 2b.

2.4 Statistical modeling of Ksat

To show a possible application of the database, we computed various pedotransfer functions (PTFs). The PTF models were fitted using a random forest (RF) machine learning algorithm (Breiman, 2001) in the R environment for statistical computing (R Core Team, 2013). We tested fitting the RF model for log-transformed (log_{10}) Ksat values as function of primary soil properties. For 15% of samples with information on bulk density and soil texture, the value of organic content (OC) was not

Table 2a. Description and units of some key variables listed in the database. The complete list can be found in the link to the data base *'version 0.3'* (https://doi.org/10.5281/zenodo.3752721) in the readme-file. We used the same codes adopted in the National Cooperative Soil Survey (NCSS) Soil Characterization Database (National Cooperative Soil Survey, 2016).

Headers	Description	Dimension
`site_key`	Data set identifier	—
`longitude_decimal_degrees`	Ranges up to +180 degrees down to -180 degrees	Decimal degree
`latitude_decimal_degrees`	Ranges up to +90 degrees down to -90 degrees	Decimal degree
`location_accuracy_min`	Minimum value of location accuracy	m
`location_accuracy_max`	Maximum value of location accuracy	m
`hzn_top`	Top of soil sample	cm
`hzn_bot`	Bottom of soil sample	cm
`hzn_desgn`	Designation of soil horizon	—
`db`	Bulk density	g cm^{-3}
`w3cld`	Soil water content at 33 kPa (field capacity)	vol %
`w15l2`	Soil water content at 1500 kPa (wilting point)	vol %
`tex_psda`	Soil texture classes based on USDA	—
`clay_tot_psa`	Mass of soil particles, $<$ 0.002 mm	%
`silt_tot_psa`	Mass of soil particles, $>$ 0.002 and $<$ 0.05 mm	%
`sand_tot_psa`	Mass of soil particle, $>$ 0.05 and $<$ 2 mm	%
`oc_v`	Soil organic carbon content	%
`ph_h2o_v`	Soil acidity	—
`Ksat_lab`	Soil saturated hydraulic conductivity from lab	cm day^{-1}
`Ksat_field`	Soil saturated hydraulic conductivity from field	cm day^{-1}
`source_db`	Sources of the datasets	—
`location_id`	Combination of latitude and logitude	—
`hzn_depth`	Mean depth of soil horizon	—

reported. Therefore, we expressed the PTF for Ksat as a function of bulk density, clay and sand content only. We derived two PTFs for Ksat:

1. *PTFs for temperate regions*: the map of Ksat locations were overlaid on the Köppen-Geiger climate zone map (Rubel and Kottek, 2010; Hamel et al., 2017) and then divided based on climatic regions (temperate, tropical, boreal, and arid) to account for differences in climate and related weathering processes (Hodnett and Tomasella, 2002). A total of 8,296 temperate-climate based Ksat values that contain information on sand, clay, and bulk density were used to develop PTF. The data set was randomly divided into a training (6,637 samples, 80%) and testing dataset (1,659 samples, 20%).

Table 2b. Example of Ksat database structure with key variables (from left to right: reference, longitudinal and latitudinal coordinates (decimal degree), top and bottom of soil sample (cm), bulk density (g cm^{-3}), soil textural class, clay, silt and sand content (%) and saturated hydraulic conductivity measured in lab or field (cm day^{-1})). NA is 'no value'. Column names are explained in Table 2a.

`site_key`	`longitude_decimal_degrees`	`latitude_decimal_degrees`	`hzn_top`	`hzn_bot`	`db`	`tex_psda`	`clay_tot_psa`	`silt_tot_psa`	`sand_tot_psa`	`ksat_lab`	`ksat_field`
Saseendran_2005	-103.15	40.15	15	30	1.33	Loam	23.4	44.3	32.3	232.08	NA
Saseendran_2005	-103.15	40.15	30	60	1.32	Loam	22.3	40.7	37.0	232.08	NA
Saseendran_2005	-103.15	40.15	60	90	1.36	Loam	17.6	36.7	45.7	337.92	NA
Saseendran_2005	-103.15	40.15	90	120	1.40	Loam	12.0	42.3	45.7	284.88	NA
Saseendran_2005	-103.15	40.15	120	150	1.42	Loam	10.0	41.7	48.3	259.20	NA
Saseendran_2005	-103.15	40.15	150	180	1.42	Loam	10.0	41.7	48.3	259.20	NA
Becker_2018	-110.13	31.73	0	15	NA	Sandy loam	NA	NA	NA	NA	26.40
Becker_2018	-110.09	31.72	0	15	NA	Sandy loam	NA	NA	NA	NA	27.84
Becker_2018	-110.09	31.69	0	15	NA	Sandy loam	NA	NA	NA	NA	21.60
Becker_2018	-110.05	31.74	0	15	NA	Loam	NA	NA	NA	NA	23.76
Becker_2018	-110.04	31.72	0	15	NA	Sandy loam	NA	NA	NA	NA	39.12
Becker_2018	-110.04	31.69	0	15	NA	Sand	NA	NA	NA	NA	102.96

Table 3. Number of samples (N) assigned to each class of spatial accuracy. A minimum and maximum accuracy is defined for each class. NA are samples without information on spatial accuracy.

Minimum location error	Maximum location error	N
0 m	100 m	9937
100 m	250 m	1422
250 m	500 m	959
500 m	1000 m	516
1000 m	5000 m	163
5000 m	10000 m	128
10000 m	NA	142
Total		**13,267**

2. *PTFs from laboratory-based Ksat values*: In a second application, the dataset (total 13,267) was divided into laboratory and field based Ksat values. The laboratory dataset (8,498 soil samples) was used for training (6,798) and testing (1,700) following the same method as used for the temperate climate PTF (i.e., 80% for training and 20% for testing).

Table 4. Instruments and methods used to estimate Ksat. A key reference with further details is given for all methods. In some cases, ’ponding’ or ’permeameter’ methods were listed in original studies without specification (18 samples in total).

Lab Ksat methods	*N*	Field Ksat methods	*N*
Constant head method (Klute and Dirksen, 1986)	8014	Mini-infiltrometer (Leeds-Harrison et al., 1994)	739
Falling head method (Klute, 1965)	766	Tension infiltrometer (Reynolds et al., 2000)	705
Triaxial cell (ASTM D 5084) (Purdy and Suryasasmita, 2006)	99	Double ring infiltrometer (Bodhinayake et al., 2004)	625
Cylinder method or soil core method (Reynolds et al., 2000)	27	Disc infiltrometer (Soracco et al., 2010)	584
Hydraulic head (Robbins, 1977)	15	Single ring (Bagarello and Sgroi, 2004)	467
Pressure plate (Sharratt, 1990)	14	Guelph Permeameter (Reynolds and Elrick, 1985)	156
Oedometer test (UNI CEN ISO/TS 17892-5) (Terzaghi, 2004)	9	BEST method (Bagarello and Sgroi, 2004)	147
Oedometer test (ASTM D2435-96) (Sutejo et al., 2019)	12	Aardvark permeameter (Hinton, 2016)	142
		Guelf Infiltrometer (Gupta et al., 1993)	87
		Piezometer slug test (Baird et al., 2017)	72
		Tensiometers (Nielsen et al., 1973)	70
		Rainfall simulator (Gupta et al., 1993)	55
		Hood infiltrometer (Schlüter et al., 2020)	40
		Micro-infiltrometer (Sepehrnia et al., 2016)	35
		Mini Disc infiltrometer (Naik et al., 2019)	32
		Disc permeameter (Mohanty et al., 1994)	27
		Constant head permeameter (Amoozegar, 1989)	22
		Steady infiltration (Scotter et al., 1982)	16
		Permeameter	10
		Ponding	8
		Philip–Dunne permeameter (Muñoz-Carpena et al., 2002)	6
		Augur method (Mohsenipour and Shahid, 2016)	5
Unknown	206	Unknown	83
Total	**9162**		**4133**

Table 5. Mean values of soil hydro-physical properties for each soil textural class. The number of samples (N) is given in parenthesis under each soil variable for each soil texture classes. N values marked with * correspond to undefined soil texture classes. BD = bulk density (g/cm^3), OC = organic carbon (%), FC = field capacity (% vol), WP = wilting point (% vol), Ksat$_l$, Ksat$_f$ = laboratory and field Ksat (cm/day). For Ksat the geometric mean is reported (due to the sensitivity on few extreme values). For all other properties the arithmetic mean is provided.

Texture Classes	Clay (N)	Silt (N)	Sand (N)	BD (N)	OC (N)	FC (N)	WP (N)	Ksat$_l$ (N)	Ksat$_f$ (N)
Clay	56.3	23.6	20.0	1.27	2.00	43.2	30.0	8.22	110.07
	(830)	(830)	(830)	(639)	(448)	(447)	(449)	(499)	(331)
Silty Clay	45.2	45.1	9.6	1.18	3.83	49.9	30.2	3.63	196.65
	(181)	(181)	(181)	(175)	(116)	(46)	(46)	(85)	(96)
Sandy Clay	39.3	8.1	52.5	1.52	0.23	34.7	23.4	14.16	——
	(176)	(176)	(176)	(172)	(140)	(158)	(158)	(172)	(4)
Clay Loam	31.4	38.6	29.9	1.27	2.49	37.2	22.1	13.34	60.56
	(544)	(544)	(544)	(382)	(360)	(76)	(76)	(127)	(417)
Silty Clay loam	33.1	57.1	9.7	1.24	2.67	46.2	23.9	1.57	48.45
	(335)	(335)	(335)	(283)	(227)	(57)	(56)	(113)	(222)
Sandy Clay Loam	26.3	12.1	61.6	1.53	1.26	28.7	17.1	19.43	14.23
	(1148)	(1148)	(1148)	(966)	(950)	(805)	(759)	(876)	(272)
Silt	7.7	84.6	7.6	1.16	1.65	51.4	7.5	13.27	—-
	(25)	(25)	(25)	(19)	(11)	(12)	(11)	(25)	
Silt Loam	15.2	66.8	17.9	1.34	3.65	35.2	15.6	5.87	44.63
	(810)	(810)	(810)	(618)	(498)	(148)	(138)	(447)	(364)
Loam	19.0	39.1	41.7	1.29	2.16	32.07	14.2	45.62	34.21
	(692)	(692)	(692)	(600)	(561)	(101)	(104)	(226)	(466)
Sandy Loam	13.5	16.8	69.7	1.49	1.33	24.2	11.0	39.71	74.57
	(1601)	(1601)	(1601)	(1492)	(1337)	(806)	(792)	(1078)	(523)
Loamy Sand	7.3	8.5	84.0	1.55	1.13	17.3	6.5	95.37	132.33
	(736)	(736)	(736)	(711)	(674)	(582)	(586)	(637)	(99)
Sand	2.2	3.1	94.6	1.51	0.62	8.2	2.5	488.46	209.55
	(4513)	(4513)	(4513)	(4437)	(4179)	(4063)	(4062)	(4409)	(106)
Total	**11,591**	**11,591**	**11,591**	**10,494**	**9,501**	**7,301**	**7,236**	**8,694**	**2,900**
	(17*)		(38*)	(775*)	(286*)	(88*)	(182*)	(468*)	(1,233*)

The *'ranger'* package version 0.12.1 (Wright and Ziegler, 2015) was implemented to process the large dataset. The PTFs developed for temperate regions and for laboratory data were then applied to test their applicability in tropical climate (1,111 samples) and for field measurements (1,998 samples), respectively. The code for generating and testing the PTFs is provided in the supplementary file.

2.5 Evaluation of Ksat PTFs

The relative importance of the covariates to determine the PTF was assessed by the increase in node purity. It is calculated using the Gini criterion from all the splits (in our case 3 splits) in the forest based on a particular variable (Rodrigues and de la Riva, 2014). Furthermore, the accuracy of the predictions was evaluated using bias, root mean square error (RMSE, in log-transformed Ksat measurement) and concordance correlation coefficient (CCC) (Lawrence and Lin, 1989).

Bias and RMSE are defined as:

$$bias = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)}{n} \tag{1}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{2}$$

where y and $\widehat{y}$ are observed and predicted Ksat values, respectively, and n is the total number of cross-validation points.

In addition, Concordance Correlation Coefficient (CCC) (as measure of the agreement between observed and predicted Ksat values) of cross validation (Lawrence and Lin, 1989) is defined as:

$$CCC = \frac{2 \cdot \rho \cdot \sigma_{\hat{y}} \cdot \sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \tag{3}$$

where $\mu_{\hat{y}}$ and μ_y are predicted and observed means, $\sigma_{\hat{y}}$ and σ_y are are predicted and observed variances and ρ is the Pearson correlation coefficient between predicted and observed values. CCC is equal to 1 for a perfect model.

3 Results

3.1 Data coverage of SoilKsatDB

Based on the literature search and data compilation, we have assembled a total of 13,267 values of Ksat from 1,910 sites (one site is equal to one location ’id’) across the globe. Figure 1 shows the global distribution of the sites used in this study. Most data originate from North America, followed by Europe, Asia, South America, Africa, and Australia. With respect to climatic regions, 10,093 Ksat values belong to the temperate region and 1,443, 1,113, 582, and 36 to tropical, arid, boreal, and polar regions, respectively. The points are often spatially clustered with the biggest cluster of points (1,103 site locations with 6,532 Ksat values) in Florida (Grunwald, 2020). Ksat data include 4,133 values from field measurement and 9,162 values from

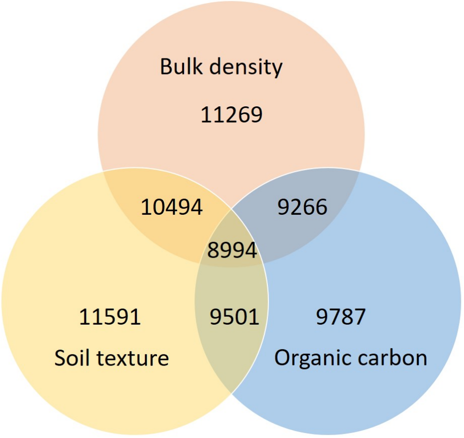

Figure 2. Venn diagram illustrating the number of samples containing information on bulk density, soil texture, and organic carbon. Out of 13,267 samples, 11,269, 11,591 and 9,787 samples have values of bulk density, soil texture and organic carbon, respectively. Furthermore, 10,494, 9,266 and 9,501 samples have information of bulk density and soil texture, bulk density and organic carbon and soil texture and organic carbon, respectively. 8,994 samples have information of all three soil properties. Note that the size of the intersecting areas does not represent the correct fractions (otherwise the intersection with 8,994 would be much bigger).

laboratory measurements. In particular, different types of infiltrometers (e.g., Mini-infiltrometer, Tension infiltrometer, double ring infiltrometer) and permeaters (e.g., Guelf permeameter, Aardwark permeameter) were used for Ksat field measurements, whereas constant or falling head methods were predominantly used in laboratory analyses, as shown in Table .4.

Out of the 13,267 Ksat measurements, 11,591, 11,269, 9,787, 7,389 and 7,418 points had information on soil texture, bulk density, organic carbon, field capacity and wilting point, respectively, while 8,994 samples had information for all soil basic properties (bulk density, soil texture and organic carbon) (Figure 2). The methods used to compute these soil properties (as much as we could extract from the literature and existing databases) were listed in the supplementary CSV file sol_ksat.pnts_metadata.csv available at *'version 0.3'* https://doi.org/10.5281/zenodo.3752721. Note that in addition to 11,591 soil texture values, 75 samples have soil texture information with total (sand+silt+clay) less than 98% or greater than 102%. We did not use these values in the PTF development. Moreover, the database contains total of 13,295 Ksat values because few studies have reported both field and lab measurements for the same sampling point.

3.2 Statistical properties of SoilKsatDB

The distribution of soil samples based on soil texture classes is shown on the USDA soil texture triangle in Figure 3a. The database covers all textural classes, with a high clustering in sandy soils due to the numerous samples from Florida (Grunwald, 2020). The violin distribution plot in Figure 3c shows the range of Ksat values for the different databases. Most of the datasets report Ksat values between $\approx 10^{-2}$ and $10^{2.5}$ cm/day, with a wider range of Ksat values observed in measurements from theses and reports (including studies with extreme values from sandy desert soils and low conductive clay soils) and from the SWIG database (databases 9 and 6 in Figure 3c, respectively). Likewise, Figure 3d shows the violin distribution of Ksat based on soil texture classes. Sand and loamy sand soils showed the highest arithmetic mean (i.e., 2.68 and 1.99, respectively), while the lowest mean values were found for silt and silty loam (i.e., 1.12 and 1.15, respectively). The significance between each soil texture class was also tested using a t-test (Kim, 2015) and results are presented in the supplementary file. Table ST1 shows that the Ksat values under sand and loamy sand soil texture class are significantly different from all other soil texture classes, however, silt, silty clay, and silty clay loam class are not significantly different from clay, sandy clay, and sandy clay loam Ksat values.

Average values of Ksat and other hydro-physical properties are shown in Table 5. Higher average organic carbon and bulk density values were observed in clayey and loamy soils compared to sandy soils. Ksat values obtained from field measurements were on average higher (depending on the type of instrument used) than those obtained from laboratory Ksat values. Particularly, for the clay texture class much lower Ksat values were observed for laboratory (mean Ksat $\approx$ 8 cm/day) compared to field (mean Ksat $\approx$ 110 cm/day) measurements (Table 5). Figure 3b further illustrates the higher range of Ksat values obtained for finer texture soils (clay and loam) compared to coarser soils (sand).

3.3 Ksat PTFs derivation

As a test application of SoilKsatDB, two PTFs were derived for Ksat (i.e., for temperate regions and based on laboratory measurements) using basic soil properties as covariates. Such basic soil properties are plotted against Ksat in Figure 4, showing that Ksat decreases with increasing clay content and bulk density, and increases with sand content. The observed correlation between these soil properties and Ksat motivates their use as key variables for the estimation of PTFs. In this application, PTFs for Ksat were built on bulk density and sand and clay content. Organic carbon (OC) was not used to build the PTFs because (i) this information was missing for 15% of samples and (ii) the correlation between OC and Ksat was poor (i.e. 0.005).

Figure S1 shows the list of relative importance of the covariates the PTFs models obtained for temperate regions and laboratory-based measurements. Clay content was found to be the most important variable followed by sand and bulk density for temperate climate PTF. On the other hand, sand content was found to be the most important variable followed by clay and bulk density for the laboratory-based Ksat PTF. CCC, bias, and RMSE were respectively equal to 0.70, -0.002, and 0.69, for the temperate region based PTF, and to 0.73, 0.0004, and 0.65 for laboratory-based PTF.

PTF models derived for temperate and laboratory-based Ksat values overestimate Ksat for tropical and field-based Ksat values, respectively (see Figure 6b and Figure 5b). CCC, bias, and RMSE values were respectively equal to 0.52, 0.2, and 0.90 for tropical Ksat values, and to 0.10, 0.21, and 1.2 for field measured Ksat values.

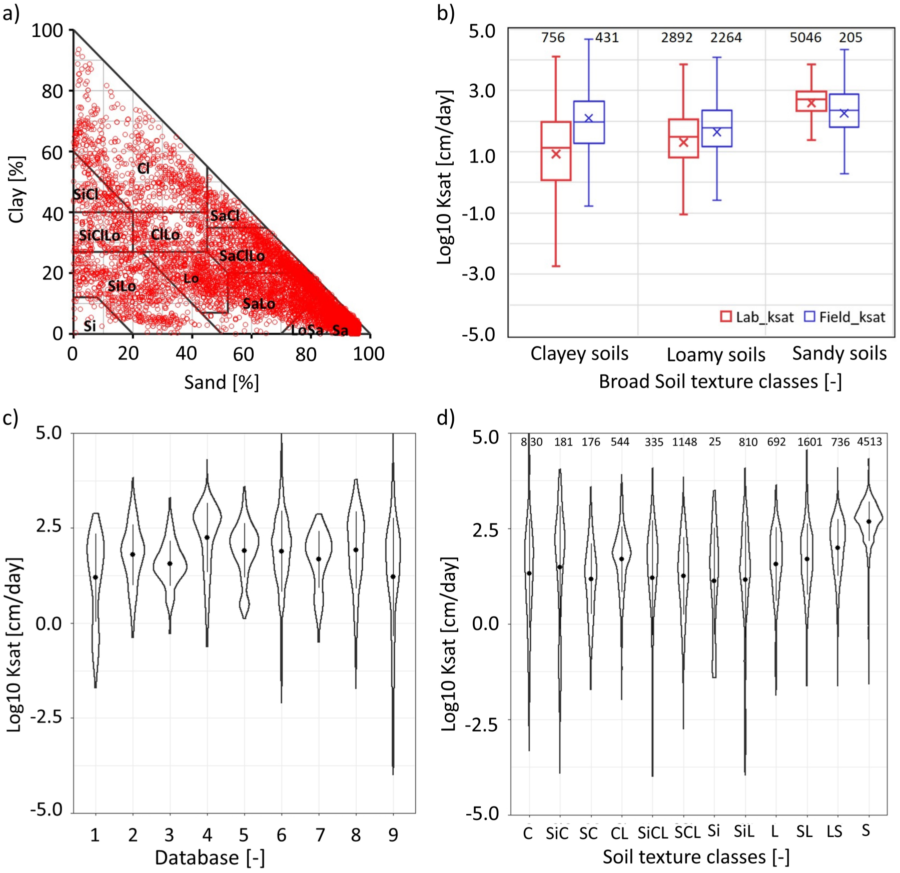

Figure 3. Characterization of collected Ksat values. (a) Distribution of soil samples on the USDA soil texture triangle. The data points cover all soil textural classes and only few samples belong to the silt textural class. b) Distribution of Ksat values using broad soil texture classes (sandy soils: sand, loamy sand; loamy soils: sandy loam, loam, silt loam, silt, clay loam, sandy clay loam; clayey soils: sandy clay, silty clay, clay) based on laboratory and field methods. The number of samples provided on the top of the figure. The increase in Ksat values in clayey and loamy soils under field methods is likely due to the effect of soil structure. A t-test showed that all broad soil texture classes are significantly different from each other except clayey soils field Ksat values and sandy soils field Ksat values (see Table ST2). The violin plot (c) represents the range of Ksat values spanned by each data source. The dot represents the mean value, and the line represents the standard deviation for each data set. The numbers 1–9 refer to different sources and databases: 1 = Australia (Forrest et al., 1985), 2 = Belgium (Vereecken et al., 2017), 3 = China (Tian et al., 2017; Li et al., 2017), 4 = Florida (Grunwald, 2020), 5 = HYBRAS (Ottoni et al., 2018), 6 = SWIG (Rahmati et al., 2018), 7 = Tibetan Plateau (Zhao et al., 2018), 8 = UNSODA (Nemes et al., 2001), 9 = all other databases in Table 1. d) Distribution of Ksat based on soil textural classes with the number of samples shown on the top of the figure. The significance was also tested for each class using a t-test (Kim, 2015) and results are presented in the supplementary file.

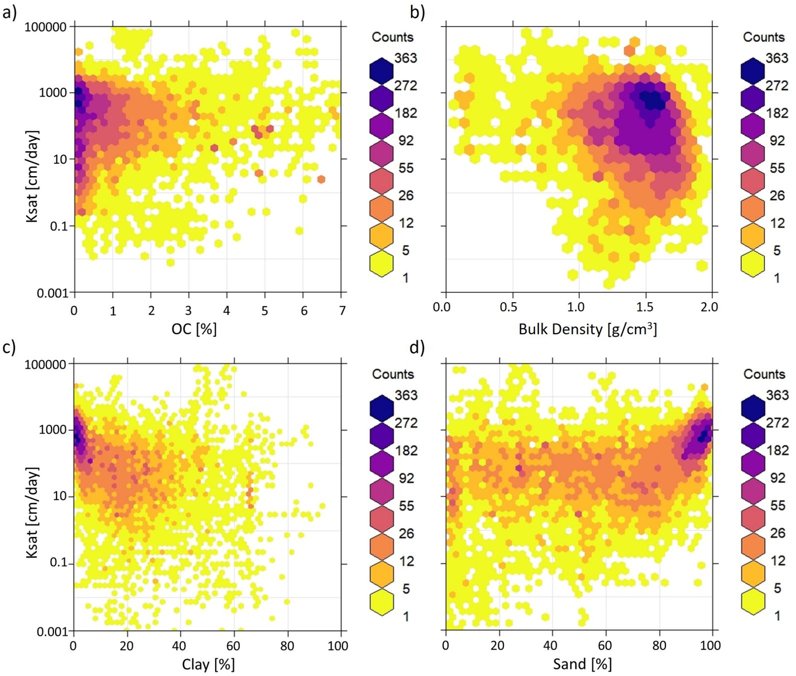

Figure 4. Partial correlation between Ksat and a) organic carbon (%), b) bulk density (g/cm^3), c) clay (%) and d) sand content (%). Ksat decreases with increasing clay content and bulk density, and increases with sand content. The color of each hexagonal cell shows the number of the counts in each cell.

4 Discussion

4.1 Laboratory vs field estimated Ksat: effect of soil structure

The Ksat values were, on average, higher for samples measured using field methods compared to laboratory methods for most soil texture classes (Table 5 and Figures 3b and 5). The difference in laboratory and field based Ksat values and higher range of Ksat values in fine textured soil is probably related to the effect of biologically-induced soil structure that might be neglected in laboratory measurements. The omission of soil structures in many laboratory samples limits the possibility to properly reproduce field observations that are likely to be more affected by the presence of biopores (Fatichi et al., 2020). In other words, variability in the Ksat values depends on the consideration (and existence) of soil structural pores by the measurement methods. Soil structural pores change the pore size distribution and subsequently affect Ksat values (Tuller and Or, 2002). Such an effect is more likely to be neglected more in laboratory measurements compared to field studies. Presence or absence of large structural pores also depends on the scale of measurements (that is usually larger in the field). Mohanty et al. (1994), for example, compared three field methods and one laboratory method and found that the sample size affects the measurement of Ksat due to the presence and absence of open-ended pores. Similarly, Ghanbarian et al. (2017) showed that the sample dimensions (e.g., internal diameter and height) also impact Ksat. The authors further developed a sample dimension-dependent

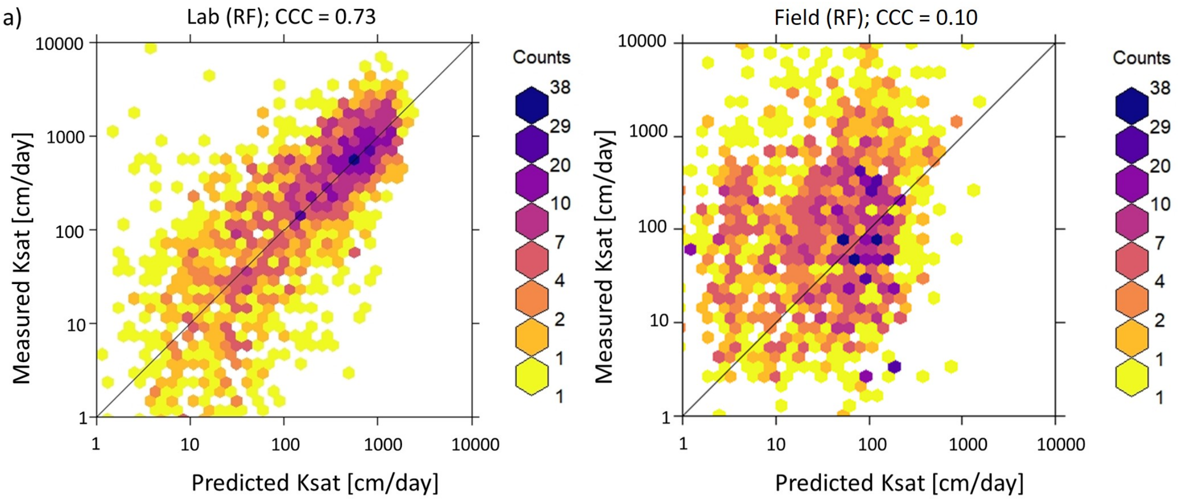

Figure 5. The correlation between observed and predicted Ksat values obtained from (a, b) random forest (RF) models. The RF-based Pedotransfer function (PTF) model was fitted using data for laboratory measurements of Ksat and tested on both laboratory (a) and field (b) measurements. Results showed reasonable agreement (CCC = 0.73) using RF algorithms for laboratory measurements, but low CCC (0.10) for field measurements. PTFs developed based on laboratory measurements do not provide accurate estimates of Ksat measured in the field.

PTF and showed a better performance compared to other available PTFs in the literature. Likewise, Braud et al. (2017) used three field methods for Ksat measurements and found significant variation between these methods of measurements. Davis et al. (1996) presents the necessity to choose the most appropriate scale of measurement for a particular soil when undertaking conductivity measurements. The authors tested small cores (73 mm wide and 63 mm high) and large cores (22 mm wide and 300 mm high) using the constant head method in the laboratory and found the difference of 1 to 3 orders of magnitude.

4.2 Temperate vs tropical soils: effect of clay mineralogy

Results showed that PTFs obtained for temperate soils performed poorly for tropical soils (Figure 6), with Ksat being underestimated by the temperate-based PTFs. This result is in agreement with Tomasella et al. (2000) who derived PTFs using data from tropical Brazilian soils, which did not properly capture observations in temperate soils. We argue that the significant differences in the models validated for tropical and temperate soils are due to the differences in the soil-forming processes defining the clay type and mineralogy. In fact, Oxisols (highly weathered clay minerals in tropical regions) are turned into inactive (non-swelling) clay minerals as a result of high rainfall and temperatures. On the other hand, in the temperate regions, active (smectite) and moderately active clay minerals (illite) are the dominant clay minerals. These swelling clay minerals retain the water within internal structures with very low hydraulic conductivity. Therefore, such a difference in clay mineralogy is likely responsible for the underestimation of Ksat in tropical soils from PTFs obtained in temperate ones. In addition, soil

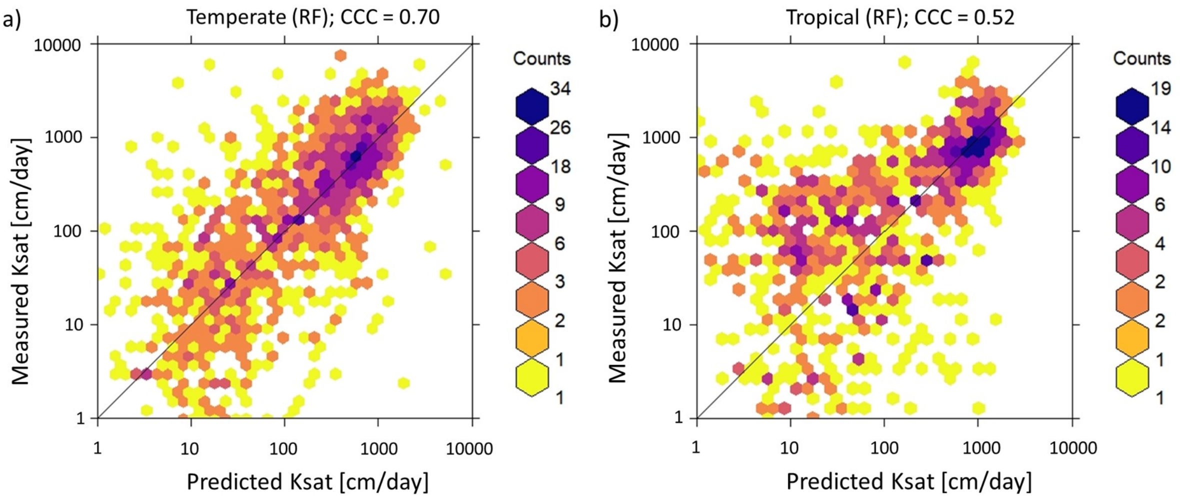

Figure 6. Correlation between observed and predicted Ksat values obtained from random forest (RF) model. The RF-based Pedotransfer function (PTF) model was obtained by fitting 6,637 training points obtained in a temperate-climate and tested on (a) temperate (1,659 samples) and (b) tropical testing points (1,111 samples). CCC is the concordance correlation coefficient. PTFs showed good performance (CCC = 0.70) for the temperate soil samples (including both laboratory and field measurements), but lower CCC values were obtained for tropical soil samples (0.52 for RF). PTFs determined for temperate regions cannot be easily transferred to tropical regions due to different soil forming processes.

structure formation processes may be different in tropical and temperate regions and intensify the differences between Ksat values measured in the two different climatic regions.

4.3 Limitations of SoilKsatDB

We have put an effort to combine laboratory and field data from most global regions.. However, we acknowledge that there are still gaps in some regions such as Russia and higher northern latitudes in general, which may produce uncertainties in Ksat estimations in such regions. The SoilKsatDB could also be of limited use for fine-resolution applications because many data points were characterized by limited spatial accuracy and missing soil depth information. Specifically, the spatial accuracy of many points is between tens of meters to several kilometers (see the methodology sections regarding the extraction of the spatial locations using Google Earth). Many of the records in the SoilKsatDB come from legacy scientific reports and the original authors can not be traced and contacted, hence we advise to use this data with caution. In addition, in the SWIG database, the soil depth and measurement method information were not provided, and often one location was used to represent an entire watershed. We tried to revisit each publication and extract the most accurate coordinates of assumed sampling locations. In addition, we assumed that most of the samples were obtained from field measurements as authors used different infiltrometers to compute Ksat, so there might be few points in our SoilKsatDB that belong to laboratory measurements and that we have incorrectly assigned to field measurements.

For each measurement, a location accuracy (0-100 m = highly accurate, >10000 m = least accurate) was assigned based on the sampling location accuracy. The location accuracy can be used as a weight or probability argument in Machine Learning for Ksat mapping. We acknowledge that this was a rather subjective decision and a more objective way to assign weights would be to use the actual spatial positioning errors. Because these were not available for most of the datasets, we have opted for the definition of a location accuracy estimated from the available documentation.

4.4 Further developments

The advancement in remote sensing technology opens the doors to link the hydraulic properties with global environmental features. Using satellite-based maps of environmental properties, local information on vegetation, climate, and topography for specific areas, which are often ignored by basic PTFs, can be incorporated. For example, Sharma et al. (2006) developed PTFs using environmental variables such as topography and vegetation and concluded that these attributes, at finer spatial scales, were useful to capture the observed variations within the soil mapping units. Likewise, Szabó et al. (2019) used the random forest machine learning algorithm for mapping soil hydraulic properties and incorporated local environmental variable information.

5 Data availability

All collected data and related soil characteristics are provided online for reference and are available at *'version 0.3'* https://doi.org/10.5281/zenodo.3752721 (Gupta et al., 2020).

6 Summary and conclusions

We prepared a comprehensive global compilation of measured Ksat training point data ($N = 13,267$) by importing, quality controlling, and standardizing tabular data from existing soil profile databases and legacy reports. The produced SoilKsatDB covers a broad range of soil types and climatic regions and hence is useful for global soil modeling. A higher variation in Ksat values was observed in fine-textured soil compared to coarse-textured soils, indicating the effect of soil structure on Ksat. Moreover, Ksat values obtained from field measurements were generally higher than those from laboratory measurements, likely due to impact of soil structural pores at larger scale in field measurements.

The new database was applied to develop pedotransfer functions (PTFs) for Ksat using measurements in temperate climates and laboratory based soil samples using RF algorithms. PTFs developed for a certain climatic region (temperate) or measurement method (laboratory) could not be satisfactorily applied to estimate Ksat for other regions (tropical) or measurement method (field) due to the role of different soil forming processes (inactive clay minerals in tropical soils and impact of biopores in field measurements).

There are still some gaps in the geographical representation of sampling points, especially in Russia and the higher northern latitudes, that could induce uncertainty in global modeling. Therefore, the data set can be further improved by covering the missing areas and achieve better accuracy in the hydrological applications.

The SoilKsatDB was developed in R software and is available via *'version 0.3'* https://doi.org/10.5281/zenodo.3752721. We have made code and data publicly available to enable further developments and improvements as a collective effort.

Acknowledgements. The SoilKsatDB is a compilation of numerous existing datasets from which the most significant are: SWIG dataset (Rahmati et al., 2018), UNSODA (Leij et al., 1996; Nemes et al., 2001), and HYBRAS (Ottoni et al., 2018). The study was supported by ETH Zurich (Grant ETH-18 18-1). OpenGeoHub maintains an global repository of Earth System Science datasets at www.openlandmap.org. We thank Zhongwang Wei for helping in collecting the datasets and for insightful discussions. We acknowledge Samuel Bickel (ETH Zurich) for the help with High Performance Computing. We would also like to thank two anonymous reviewers and Dr. Attila Nemes for their constructive feedback to improve the manuscript.

References

Abagandura, G. O., Nasr, G. E.-D. M., and Moumen, N. M.: Influence of tillage practices on soil physical properties and growth and yield of maize in jabal al akhdar, Libya, Open Journal of Soil Science, 7, 118–132, 2017.

Amer, A.-M. M., Logsdon, S. D., and Davis, D.: Prediction of hydraulic conductivity as related to pore size distribution in unsaturated soils, Soil science, 174, 508–515, 2009.

Amoozegar, A.: A compact constant-head permeameter for measuring saturated hydraulic conductivity of the vadose zone, Soil Science Society of America Journal, 53, 1356–1361, 1989.

Amoozegar, A. and Warrick, A.: Hydraulic conductivity of saturated soils: field methods, Methods of Soil Analysis: Part 1 Physical and Mineralogical Methods, 5, 735–770, 1986.

Anapalli, S. S., Nielsen, D. C., Ma, L., Ahuja, L. R., Vigil, M. F., and Halvorson, A. D.: Effectiveness of RZWQM for simulating alternative Great Plains cropping systems, Agronomy journal, 97, 1183–1193, 2005.

Andrade, R. B.: The influence of bulk density on the hydraulic conductivity and water content-matric suction relation of two soils, 1971.

Arend, J. L.: Infiltration rates of forest soils in the Missouri Ozarks as affected by woods burning and litter removal, J. For., 39, 726–728, 1941.

Bagarello, V. and Sgroi, A.: Using the single-ring infiltrometer method to detect temporal changes in surface soil field-saturated hydraulic conductivity, Soil and Tillage research, 76, 13–24, 2004.

Baird, A. J.: Field estimation of macropore functioning and surface hydraulic conductivity in a fen peat, Hydrological Processes, 11, 287–295, 1997.

Baird, A. J., Low, R., Young, D., Swindles, G. T., Lopez, O. R., and Page, S.: High permeability explains the vulnerability of the carbon store in drained tropical peatlands, Geophysical Research Letters, 44, 1333–1339, 2017.

Bambra, A.: Soil loss estimation in experimental orchard at Nauni in Solan district of Himachal Pradesh, Ph.D. thesis, Dr. Yashwant Singh Parmar, University of horticulture and forestry, 2016.

Batjes, N. H.: Total carbon and nitrogen in the soils of the world, European journal of soil science, 47, 151–163, 1996.

Becker, R., Gebremichael, M., and Märker, M.: Impact of soil surface and subsurface properties on soil saturated hydraulic conductivity in the semi-arid Walnut Gulch Experimental Watershed, Arizona, USA, Geoderma, 322, 112–120, 2018.

Beyer, M., Gaj, M., Hamutoko, J. T., Koeniger, P., Wanke, H., and Himmelsbach, T.: Estimation of groundwater recharge via deuterium labelling in the semi-arid Cuvelai-Etosha Basin, Namibia, Isotopes in environmental and health studies, 51, 533–552, 2015.

Bhattacharyya, R., Prakash, V., Kundu, S., and Gupta, H.: Effect of tillage and crop rotations on pore size distribution and soil hydraulic conductivity in sandy clay loam soil of the Indian Himalayas, Soil and Tillage Research, 86, 129–140, 2006.

Blake, W. H., Theocharopoulos, S. P., Skoulikidis, N., Clark, P., Tountas, P., Hartley, R., and Amaxidis, Y.: Wildfire impacts on hillslope sediment and phosphorus yields, Journal of Soils and Sediments, 10, 671–682, 2010.

Bodhinayake, W., Si, B. C., and Noborio, K.: Determination of hydraulic properties in sloping landscapes from tension and double-ring infiltrometers, Vadose Zone Journal, 3, 964–970, 2004.

Boike, J., Roth, K., and Overduin, P. P.: Thermal and hydrologic dynamics of the active layer at a continuous permafrost site (Taymyr Peninsula, Siberia), Water Resources Research, 34, 355–363, 1998.

Bonell, M. and Williams, J.: The two parameters of the Philip infiltration equation: their properties and spatial and temporal heterogeneity in a red earth of tropical semi-arid Queensland, Journal of Hydrology, 87, 9–31, 1986.

Bonsu, M. and Masopeh, B.: Saturated hydraulic conductivity values of some forest soils of Ghana determined by a simple method, Ghana Journal of Agricultural Science, 29, 75–80, 1996.

Braud, I., Desprats, J.-F., Ayral, P.-A., Bouvier, C., and Vandervaere, J.-P.: Mapping topsoil field-saturated hydraulic conductivity from point measurements using different methods, Journal of Hydrology and Hydromechanics, 65, 264–275, 2017.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Bruand, A., Duval, O., and Cousin, I.: Estimation des propriétés de rétention en eau des sols à partir de la base de données SOLHYDRO: Une première proposition combianant le type d'horizon, sa texture et sa densité apparente., 2004.

Campbell, R. E., Baker, J., Ffolliott, P. F., Larson, F. R., and Avery, C. C.: Wildfire effects on a ponderosa pine ecosystem: an Arizona case study, USDA For. Serv. Res. Pap. RM-191. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experimental Station. 12 p., 191, 1977.

Casalicchio, G., Bossek, J., Lang, M., Kirchhoff, D., Kerschke, P., Hofner, B., Seibold, H., Vanschoren, J., and Bischl, B.: OpenML: An R package to connect to the machine learning platform OpenML, Computational Statistics, pp. 1–15, 2017.

Chang, Y.-J.: Predictions of saturated hydraulic conductivity dynamics in a midwestern agricultural watershed, Iowa, 2010.

Chief, K., Ferré, T., and Nijssen, B.: Correlation between air permeability and saturated hydraulic conductivity: Unburned and burned soils, Soil Science Society of America Journal, 72, 1501–1509, 2008.

Cisneros, J., Cantero, J., and Cantero, A.: Vegetation, soil hydrophysical properties, and grazing relationships in saline-sodic soils of Central Argentina, Canadian Journal of Soil Science, 79, 399–409, 1999.

Coelho, M. A.: Spatial variability of water related soil physical properties., 1974.

Conedera, M., Peter, L., Marxer, P., Forster, F., Rickenmann, D., and Re, L.: Consequences of forest fires on the hydrogeological response of mountain catchments: a case study of the Riale Buffaga, Ticino, Switzerland, Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group, 28, 117–129, 2003.

Cornelis, W. M., Ronsyn, J., Van Meirvenne, M., and Hartmann, R.: Evaluation of pedotransfer functions for predicting the soil moisture retention curve, Soil Science Society of America Journal, 65, 638–648, 2001.

Daniel, S., Gabiri, G., Kirimi, F., Glasner, B., Näschen, K., Leemhuis, C., Steinbach, S., and Mtei, K.: Spatial distribution of soil hydrological properties in the Kilombero floodplain, Tanzania, Hydrology, 4, 57, 2017.

Davis, S. H., Vertessy, R. A., Dunkerley, D. L., Mein, R. G., et al.: The influence of scale on the measurement of saturated hydraulic conductivity in forest soils, in: National Conference Publication-Institution of Engineers Australia NCP, vol. 1, pp. 103–108, Institution of Engineers, Australia, 1996.

Deshmukh, H., Chandran, P., Pal, D., Ray, S., Bhattacharyya, T., and Potdar, S.: A pragmatic method to estimate plant available water capacity (PAWC) of rainfed cracking clay soils (Vertisols) of Maharashtra, Central India, Clay Res, 33, 1–14, 2014.

Ebel, B. A., Moody, J. A., and Martin, D. A.: Hydrologic conditions controlling runoff generation immediately after wildfire, Water Resources Research, 48, 2012.

El-Shafei, Y., Al-Darby, A., Shalaby, A., and Al-Omran, A.: Impact of a highly swelling gel-forming conditioner (acryhope) upon water movement in uniform sandy soils, Arid Land Research and Management, 8, 33–50, 1994.

Elnaggar, A.: Spatial Variability of Soil Physiochemical Properties in Bahariya Oasis, Egypt, Egyptian J. of Soil Sci. (EJSS), 57, 313–328, https://doi.org/10.21608/EJSS.2017.4438, 2017.

Fatichi, S., Or, D., Walko, R., Vereecken, H., Young, M. H., Ghezzehei, T. A., Hengl, T., Kollet, S., Agam, N., and Avissar, R.: Soil structure is an important omission in Earth System Models, Nature Communications, 11, 2020.

Ferreira, A., Coelho, C., Boulet, A., and Lopes, F.: Temporal patterns of solute loss following wildfires in Central Portugal, International Journal of Wildland Fire, 14, 401–412, 2005.

Forrest, J., Beatty, H., Hignett, C., Pickering, J., and Williams, R.: Survey of the physical properties of wheatland soils in eastern Australia, 1985.

Ganiyu, S., Rabiu, J., and Olatoye, R.: Predicting hydraulic conductivity around septic tank systems using soil physico-chemical properties and determination of principal soil factors by multivariate analysis, Journal of King Saud University-Science, 2018.

Ghanbarian, B., Taslimitehrani, V., and Pachepsky, Y. A.: Accuracy of sample dimension-dependent pedotransfer functions in estimation of soil saturated hydraulic conductivity, Catena, 149, 374–380, 2017.

Glinski, J., Ostrowski, J., Stepniewska, Z., and Stepniewski, W.: Soil sample bank representing mineral soils of Poland, Problemy Agrofizyki (Poland), 1991.

Gliński, J., Stępniewski, W., Stępniewska, Z., Włodarczyk, T., Brzezińska, M., et al.: Characteristics of aeration properties of selected soil profiles from central Europe., International agrophysics, 14, 17–31, 2000.

Greenwood, W. and Buttle, J.: Effects of reforestation on near-surface saturated hydraulic conductivity in a managed forest landscape, southern Ontario, Canada, Ecohydrology, 7, 45–55, 2014.

Grunwald, S.: Florida soil characterization data, Soil and water science department, IFAS-Instituite of food and agriculture science, University of Florida, http://soils.ifas.ufl.edu, 2020.

Gupta, R., Rudra, R., Dickinson, W., Patni, N., and Wall, G.: Comparison of saturated hydraulic conductivity measured by various field methods, Transactions of the ASAE, 36, 51–55, 1993.

Gupta, S., Hengl, T., Lehmann, P., Bonetti, S., and Or, D.: SoilKsatDB: a global compilation of soil saturated hydraulic conductivity measurements, Zenodo, https://doi.org/10.5281/zenodo.3752721, 2020.

Gwenzi, W., Hinz, C., Holmes, K., Phillips, I. R., and Mullins, I. J.: Field-scale spatial variability of saturated hydraulic conductivity on a recently constructed artificial ecosystem, Geoderma, 166, 43–56, 2011.

Habecker, M., McSweeney, K., and Madison, F.: Identification and genesis of fragipans in Ochrepts of north central Wisconsin, Soil Science Society of America Journal, 54, 139–146, 1990.

Habel, A. Y.: The role of climate on the aggregate stability and soil erodibility of selected El-Jabal Al-Akhdar soils-Libya, Alexandria Journal of Agricultural Research, 58, 261–271, 2013.

Hamel, P., Falinski, K., Sharp, R., Auerbach, D. A., Sánchez-Canales, M., and Dennedy-Frank, P. J.: Sediment delivery modeling in practice: Comparing the effects of watershed characteristics and data resolution across hydroclimatic regions, Science of the Total Environment, 580, 1381–1388, 2017.

Hao, M., Zhang, J., Meng, M., Chen, H. Y., Guo, X., Liu, S., and Ye, L.: Impacts of changes in vegetation on saturated hydraulic conductivity of soil in subtropical forests, Scientific reports, 9, 8372, 2019.

Hardie, M. A., Cotching, W. E., Doyle, R. B., Holz, G., Lisson, S., and Mattern, K.: Effect of antecedent soil moisture on preferential flow in a texture-contrast soil, Journal of Hydrology, 398, 191–201, 2011.

Haverkamp, R., Ross, P., Smettem, K., and Parlange, J.: Three-dimensional analysis of infiltration from the disc infiltrometer: 2. Physically based infiltration equation, Water Resources Research, 30, 2931–2935, 1994.

Haverkamp, R., Zammit, C., Bouraoui, F., Rajkai, K., Arrúe, J., and Heckmann, N.: GRIZZLY: Grenoble catalogue of soils: Survey of soil field data and description of particle-size, soil water retention and hydraulic conductivity functions, Lab. d'Etude des Transferts en Hydrol. et Environ., Grenoble, France, 1998.

Helbig, M., Boike, J., Langer, M., Schreiber, P., Runkle, B. R., and Kutzbach, L.: Spatial and seasonal variability of polygonal tundra water balance: Lena River Delta, northern Siberia (Russia), Hydrogeology Journal, 21, 133–147, 2013.

Hiederer, R., Jones, R. J., and Daroussin, J.: Soil Profile Analytical Database for Europe (SPADE): reconstruction and validation of the measured data (SPADE/M), Geografisk Tidsskrift-Danish Journal of Geography, 106, 71–85, 2006.

Hinton, H.: Land Management Controls on Hydraulic Conductivity of an Urban Farm in Atlanta, GA, 2016.

Hodnett, M. and Tomasella, J.: Marked differences between van Genuchten soil water-retention parameters for temperate and tropical soils: a new water-retention pedo-transfer functions developed for tropical soils, Geoderma, 108, 155–180, 2002.

Horn, A., Stumpfe, A., Kues, J., Zinner, H.-J., and Fleige, H.: Die Labordatenbank des Niedersächsischen Bodeninformationssystems (NIBIS)-. Teil: Fachinformationssystem Bodenkunde, Geologisches Jahrbuch. Reihe A, Allgemeine und regionale Geologie BR Deutschland und Nachbargebiete, Tektonik, Stratigraphie, Paläontologie, pp. 59–97, 1991.

Houghton, T. B.: Hydrogeologic characterization of an alpine glacial till, Snowy Range, Wyoming, Ph.D. thesis, Colorado State University. Libraries, 2011.

Hu, W., She, D., Shao, M., Chun, K. P., and Si, B.: Effects of initial soil water content and saturated hydraulic conductivity variability on small watershed runoff simulation using LISEM, Hydrological Sciences Journal, 60, 1137–1154, 2015.

Imeson, A., Verstraten, J., Van Mulligen, E., and Sevink, J.: The effects of fire and water repellency on infiltration and runoff under Mediterranean type forest, Catena, 19, 345–361, 1992.

Jabro, J.: Estimation of saturated hydraulic conductivity of soils from particle size distribution and bulk density data, Transactions of the ASAE, 35, 557–560, 1992.

Jarvis, N., Koestel, J., Messing, I., Moeys, J., and Lindahl, A.: Influence of soil, land use and climatic factors on the hydraulic conductivity of soil, Hydrology and Earth System Sciences, 17, 5185–5195, 2013.

Johansen, M. P., Hakonson, T. E., and Breshears, D. D.: Post-fire runoff and erosion from rainfall simulation: contrasting forests with shrublands and grasslands, Hydrological processes, 15, 2953–2965, 2001.

Kanemasu, E.: Soil Hydraulic Conductivity Data (FIFE), ORNL Distributed Active Archive Center, https://doi.org/10.3334/ORNLDAAC/107, 1994.

Katimon, A. and Hassan, A. M. M.: Field hydraulic conductivity of some Malaysian peat, Malaysian Journal of Civil Engineering, 10, 1997.

Keisling, T. C.: Precision with which selected physical properties of similar soils can be estimated, Ph.D. thesis, Oklahoma State University, 1974.

Kelly, T. J., Baird, A. J., Roucoux, K. H., Baker, T. R., Honorio Coronado, E. N., Ríos, M., and Lawson, I. T.: The high hydraulic conductivity of three wooded tropical peat swamps in northeast Peru: measurements and implications for hydrological function, Hydrological Processes, 28, 3373–3387, 2014.

Kim, T. K.: T test as a parametric statistic, Korean journal of anesthesiology, 68, 540, 2015.

Kirby, J., Kingham, R., and Cortes, M.: Texture, density and hydraulic conductivity of some soils in San Luis province, Argentina, Ciencia del suelo, 19, 20–28, 2001.

Klute, A.: Laboratory measurement of hydraulic conductivity of saturated soil, Methods of Soil Analysis: Part 1 Physical and Mineralogical Properties, Including Statistics of Measurement and Sampling, 9, 210–221, 1965.

Klute, A. and Dirksen, C.: Hydraulic conductivity and diffusivity: Laboratory methods, Methods of Soil Analysis: Part 1 Physical and Mineralogical Methods, 5, 687–734, 1986.

Kool, J., Albrecht, K. A., Parker, J., Baker, J., et al.: Physical and chemical characterization of the Groseclose soil mapping unit, 1986.

Krahmer, U., Hennings, V., Müller, U., and Schrey, H.-P.: Ermittlung bodenphysikalischer Kennwerte in Abhängigkeit von Bodenart, lagerungsdichte und Humusgehalt, Zeitschrift für Pflanzenernährung und Bodenkunde, 158, 323–331, 1995.

Kramarenko, V., Brakorenko, N., and Molokov, V.: Hydraulic conductivity of peat in Western Siberia, in: E3S Web of Conferences, vol. 98, p. 11003, EDP Sciences, 2019.

Kutiel, P., Lavee, H., Segev, M., and Benyamini, Y.: The effect of fire-induced surface heterogeneity on rainfall-runoff-erosion relationships in an eastern Mediterranean ecosystem, Israel, Catena, 25, 77–87, 1995.

Kutílek, M. and Krejca, M.: Three-parameter infiltration equation of Philip type, Vodohosp. ˇCas, 35, 52–61, 1987.

Lamara, M. and Derriche, Z.: Prediction of unsaturated hydraulic properties of dune sand on drying and wetting paths, Electron. J. Geotech. Eng, 13, 1–19, 2008.

Lawrence, I. and Lin, K.: A concordance correlation coefficient to evaluate reproducibility, Biometrics, pp. 255–268, 1989.

Leeds-Harrison, P., Youngs, E., and Uddin, B.: A device for determining the sorptivity of soil aggregates, European Journal of Soil Science, 45, 269–272, 1994.

Leij, F., Alves, W., Van Genuchten, M. T., and Williams, J.: The UNSODA Unsaturated Soil Hydraulic Database; User's Manual, Version 1.0, Rep. EPA/600/R-96, 95, 103, 1996.

Li, X., Liu, S., Xiao, Q., Ma, M., Jin, R., Che, T., Wang, W., Hu, X., Xu, Z., Wen, J., et al.: A multiscale dataset for understanding complex eco-hydrological processes in a heterogeneous oasis system, Scientific data, 4, 170 083, 2017.

Lopes, V. S., Cardoso, I. M., Fernandes, O. R., Rocha, G. C., Simas, F. N. B., de Melo Moura, W., Santana, F. C., Veloso, G. V., and da Luz, J. M. R.: The establishment of a secondary forest in a degraded pasture to improve hydraulic properties of the soil, Soil and Tillage Research, 198, 104 538, 2020.

Lopez, O., Jadoon, K., and Missimer, T.: Method of relating grain size distribution to hydraulic conductivity in dune sands to assist in assessing managed aquifer recharge projects: Wadi Khulays dune field, western Saudi Arabia, Water, 7, 6411–6426, 2015.

Mahapatra, S. and Jha, M. K.: On the estimation of hydraulic conductivity of layered vadose zones with limited data availability, Journal of Earth System Science, 128, 75, 2019.

Martin, D. A. and Moody, J. A.: Comparison of soil infiltration rates in burned and unburned mountainous watersheds, Hydrological Processes, 15, 2893–2903, 2001.

McKenzie, N., Jacquier, D., and Gregory, L.: Online soil information systems–recent Australian experience, in: Digital soil mapping with limited data, pp. 283–290, Springer, 2008.

Mohanty, B., Kanwar, R. S., and Everts, C.: Comparison of saturated hydraulic conductivity measurement methods for a glacial-till soil, Soil Science Society of America Journal, 58, 672–677, 1994.

Mohsenipour, M. and Shahid, S.: Estimation OF saturated hydraulic conductivity: A Review, Malasia: Academia Edu. Recuperado de http://bit. ly/2WShxfW, 2016.

Mott, J., Bridge, B., and Arndt, W.: Soil seals in tropical tall grass pastures of northern Australia, Soil Research, 17, 483–494, 1979.

Mualem, Y.: Catalogue of the hydraulic properties of unsaturated soils, Technion Israel Institute of Technology, Technion Research & Development, 1976.

Muñoz-Carpena, R., Regalado, C. M., Álvarez-Benedi, J., and Bartoli, F.: Field evaluation of the new Philip-Dunne permeameter for measuring saturated hydraulic conductivity, Soil Science, 167, 9–24, 2002.

Naik, A. P., Ghosh, B., and Pekkat, S.: Estimating soil hydraulic properties using mini disk infiltrometer, ISH Journal of Hydraulic Engineering, 25, 62–70, 2019.

National Cooperative Soil Survey: National cooperative soil survey characterization database, United States Department of Agriculture, Natural Resoucres Conservation, Lincoln, NE, 2016.

Nemes, A.: Unsaturated soil hydraulic database of Hungary: HUNSODA, Agrokémia és Talajtan, 51, 17–26, 2002.

Nemes, A.: Databases of soil physical and hydraulic properties, Encyclopedia of agrophysics, pp. 194–199, 2011.

Nemes, A. d., Schaap, M., Leij, F., and Wösten, J.: Description of the unsaturated soil hydraulic database UNSODA version 2.0, Journal of Hydrology, 251, 151–162, 2001.

Nielsen, D., Biggar, J., and Erh, K.: "Spatial variability of field-measured soil water properties. Hilgardia, 42 (7), 215–259., 1973.

Niemeyer, R., Fremier, A. K., Heinse, R., Chávez, W., and DeClerck, F. A.: Woody vegetation increases saturated hydraulic conductivity in dry tropical Nicaragua, Vadose Zone Journal, 13, 2014.

Nyman, P., Sheridan, G. J., Smith, H. G., and Lane, P. N.: Evidence of debris flow occurrence after wildfire in upland catchments of south-east Australia, Geomorphology, 125, 383–401, 2011.

Ottoni, M. V., Ottoni Filho, T. B., Schaap, M. G., Lopes-Assad, M. L. R., and Rotunno Filho, O. C.: Hydrophysical database for Brazilian soils (HYBRAS) and pedotransfer functions for water retention, Vadose Zone Journal, 17, 2018.

Ouattara, M.: Variation of saturated hydraulic conductivity with depth for selected profiles of Tillman-Hollister soil, Ph.D. thesis, Oklahoma State University, 1977.

Päivänen, J. et al.: Hydraulic conductivity and water retention in peat soils., Suomen metsätieteellinen seura, 1973.

Parks, D. S. and Cundy, T. W.: Soil hydraulic characteristics of a small southwest Oregon watershed following high-intensity wildfires, in: In: Berg, Neil H. tech. coord. Proceedings of the Symposium on Fire and Watershed Management: October 26-28, 1988, Sacramento, California. Gen. Tech. Rep. PSW-109. Berkeley, Calif.: US Department of Agriculture, Forest Service, Pacific Southwest Forest and Range Experiment Station: 63-67, vol. 109, 1989.

Price, K., Jackson, C. R., and Parker, A. J.: Variation of surficial soil hydraulic properties across land uses in the southern Blue Ridge Mountains, North Carolina, USA, Journal of Hydrology, 383, 256–268, 2010.

Purdy, S. and Suryasasmita, V.: Comparison of hydraulic conductivity test methods for landfill clay liners, in: Advances in Unsaturated Soil, Seepage, and Environmental Geotechnics, pp. 364–372, 2006.

Quinton, W. L., Hayashi, M., and Carey, S. K.: Peat hydraulic conductivity in cold regions and its relation to pore size and geometry, Hydrological Processes: An International Journal, 22, 2829–2837, 2008.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 2013.

Rab, M.: Soil physical and hydrological properties following logging and slash burning in the Eucalyptus regnans forest of southeastern Australia, Forest Ecology and Management, 84, 159–176, 1996.

Radcliffe, D., West, L., Ware, G., and Bruce, R.: Infiltration in adjacent Cecil and Pacolet soils, Soil Science Society of America Journal, 54, 1739–1743, 1990.

Rahimy, P.: Effects of Soil Depth and Saturated Hydraulic Conductivity Spatial Variation on Runoff Simulation by the Limburg Soil Erosion Model, LISEM: A Case Study in Faucon Catchment, France, University of Twente Faculty of Geo-Information and Earth Observation (ITC), 2011.

Rahmati, M., Weihermüller, L., Vanderborght, J., Pachepsky, Y. A., Mao, L., Sadeghi, S. H., Moosavi, N., Kheirfam, H., Montzka, C., Van Looy, K., et al.: Development and analysis of the Soil Water Infiltration Global database, 2018.

Ramli, M.: Management of Groundwater Resources from Peat in Sarawak, 1999.

Ravi, S., Wang, L., Kaseke, K. F., Buynevich, I. V., and Marais, E.: Ecohydrological interactions within “fairy circles” in the Namib Desert: Revisiting the self-organization hypothesis, Journal of Geophysical Research: Biogeosciences, 122, 405–414, 2017.

Rawls, W. J., Brakensiek, D. L., and Saxtonn, K.: Estimation of soil water properties, Transactions of the ASAE, 25, 1316–1320, 1982.

Reynolds, W. and Elrick, D.: In situ measurement of field-saturated hydraulic conductivity, sorptivity, and the α-parameter using the Guelph permeameter, Soil science, 140, 292–302, 1985.

Reynolds, W., Bowman, B., Brunke, R., Drury, C., and Tan, C.: Comparison of tension infiltrometer, pressure infiltrometer, and soil core estimates of saturated hydraulic conductivity, Soil Science Society of America Journal, 64, 478–484, 2000.

Richard, F. and Lüscher, P.: Physikalische Eigenschaften von Böden der Schweiz. Lokalformen. Eidg. Anstalt für das forstliche Versuchswesen. Sonderserie., 1983/87.

Robbins, C. W.: Hydraulic conductivity and moisture retention characteristics of southern Idaho’s silt loam soils, 1977.

Rodrigues, M. and de la Riva, J.: An insight into machine-learning algorithms to model human-caused wildfire occurrence, Environmental Modelling & Software, 57, 192–201, 2014.

Romano, N. and Palladino, M.: Prediction of soil water retention using soil physical data and terrain attributes, Journal of Hydrology, 265, 56–75, 2002.

Rubel, F. and Kottek, M.: Observed and projected climate shifts 1901–2100 depicted by world maps of the Köppen-Geiger climate classification, Meteorologische Zeitschrift, 19, 135–141, 2010.

Rycroft, D., Williams, D., and Ingram, H.: The transmission of water through peat: I. Review, The Journal of Ecology, pp. 535–556, 1975.

Sanzeni, A., Colleselli, F., and Grazioli, D.: Specific surface and hydraulic conductivity of fine-grained soils, Journal of Geotechnical and Geoenvironmental Engineering, 139, 1828–1832, 2013.

Sayok, A., Ayob, K., Melling, L., Goh, K., Uyo, L., and Hatano, R.: Hydraulic conductivity and moisture characteristics of tropical peatland-preliminary investigation, Malaysian Society of Soil Science (MSSS), 2007.

Schindler, U. G. and Müller, L.: Soil hydraulic functions of international soils measured with the Extended Evaporation Method (EEM) and the HYPROP device, Open Data Journal for Agricultural Research, 3, 2017.

Schlüter, S., Albrecht, L., Schwärzel, K., and Kreiselmeier, J.: Long-term effects of conventional tillage and no-tillage on saturated and near-saturated hydraulic conductivity–Can their prediction be improved by pore metrics obtained with X-ray CT?, Geoderma, 361, 114 082, 2020.

Scotter, D., Clothier, B., and Harper, E.: Measuring saturated hydraulic conductivity and sorptivity using twin rings, Soil Research, 20, 295–304, 1982.

Sepehrnia, N., Hajabbasi, M. A., Afyuni, M., and Lichner, L.: Extent and persistence of water repellency in two Iranian soils, Biologia, 71, 1137–1143, 2016.

Sharma, S. K., Mohanty, B. P., and Zhu, J.: Including topography and vegetation attributes for developing pedotransfer functions, Soil Science Society of America Journal, 70, 1430–1440, 2006.

Sharratt, B. S.: Water retention, bulk density, particle size, and thermal and hydraulic conductivity of arable soils in interior Alaska, 1990.

Simmons, L. A.: Soil hydraulic and physical properties as affected by logging management, Ph.D. thesis, University of Missouri–Columbia, 2014.

Singh, I., Awasthi, O., Sharma, B., More, T., Meena, S., et al.: Soil properties, root growth, water-use efficiency in brinjal (Solanum melongena) production and economics as affected by soil water conservation practices, Indian Journal of Agricultural Sciences, 81, 760, 2011.

Singh, R., Van Dam, J., and Feddes, R. A.: Water productivity analysis of irrigated crops in Sirsa district, India, Agricultural Water Management, 82, 253–278, 2006.

Smettem, K. and Ross, P.: Measurement and prediction of water movement in a field soil: The matrix-macropore dichotomy, Hydrological processes, 6, 1–10, 1992.

Sonneveld, M., Everson, T., and Veldkamp, A.: Multi-scale analysis of soil erosion dynamics in Kwazulu-Natal, South Africa, Land Degradation & Development, 16, 287–301, 2005.

Soracco, C. G., Lozano, L. A., Sarli, G. O., Gelati, P. R., and Filgueira, R. R.: Anisotropy of saturated hydraulic conductivity in a soil under conservation and no-till treatments, Soil and Tillage Research, 109, 18–22, 2010.

Southard, R. and Buol, S.: Subsoil saturated hydraulic conductivity in relation to soil properties in the North Carolina Coastal Plain, Soil Science Society of America Journal, 52, 1091–1094, 1988.

Sutejo, Y., Saggaff, A., Rahayu, W., et al.: Hydraulic conductivity and compressibility characteristics of fibrous peat, in: IOP Conference Series: Materials Science and Engineering, vol. 620, p. 012053, IOP Publishing, 2019.

Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., and Pásztor, L.: Mapping soil hydraulic properties using random-forest-based pedotransfer functions and geostatistics, Hydrology and Earth System Sciences, 23, 2615–2635, 2019.

Takahashi, H.: Studies on microclimate and hydrology of peat swamp forest in Central Kalimantan, Indonesia, in: Biodiversity and Sustainability of Tropical peatlands, Samara Publishing Limited, 1997.

Terzaghi, K.: Geotechnical investigation and testing-Laboratory testing of soil-Part 5: Incremental loading oedometer test 2, W3C XML, 1, 2006, 2004.

Tete-Mensah, I.: Evaluation of Some Physical and Chemical Properties of Soils Under two Agroforestrv Practices, Ph.D. thesis, University of Ghana, 1993.

Tian, J., Zhang, B., He, C., and Yang, L.: Variability in soil hydraulic conductivity and soil hydrological response under different land covers in the mountainous area of the Heihe River Watershed, Northwest China, Land degradation & development, 28, 1437–1449, 2017.

Tomasella, J., Hodnett, M. G., and Rossato, L.: Pedotransfer functions for the estimation of soil water retention in Brazilian soils, 2000.

Tomasella, J., Pachepsky, Y., Crestana, S., and Rawls, W.: Comparison of two techniques to develop pedotransfer functions for water retention, Soil Science Society of America Journal, 67, 1085–1092, 2003.

Tuller, M. and Or, D.: Unsaturated Hydraulic Conductivity of Structured Porous MediaA Review of Liquid Configuration–Based Models, Vadose Zone Journal, 1, 14–37, 2002.

Varela, M., Benito, E., and Keizer, J.: Influence of wildfire severity on soil physical degradation in two pine forest stands of NW Spain, Catena, 133, 342–348, 2015.

Verburg, K., Bridge, B. J., Bristow, K. L., and Keating, B. A.: Properties of selected soils in the Gooburrum–Moore Park area of Bundaberg, CSIRO Land and Water Technical Report, 9, 77, 2001.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M., Genuchten, M. T., et al.: Using pedotransfer functions to estimate the van Genuchten–Mualem soil hydraulic properties: A review, Vadose Zone Journal, 9, 795–820, 2010.

Vereecken, H., Van Looy, K., Weynants, M., and Javaux, M.: Soil retention and conductivity curve data base sDB, link to MATLAB files, 2017.

Vieira, B. C. and Fernandes, N. F.: Landslides in Rio de Janeiro: the role played by variations in soil hydraulic conductivity, Hydrological Processes, 18, 791–805, 2004.

Vogeler, I., Carrick, S., Cichota, R., and Lilburne, L.: Estimation of soil subsurface hydraulic conductivity based on inverse modelling and soil morphology, Journal of Hydrology, 574, 373–382, 2019.

Waddington, J. and Roulet, N.: Groundwater flow and dissolved carbon movement in a boreal peatland, Journal of Hydrology, 191, 122–138, 1997.

Wang, T., Zlotnik, V. A., Wedin, D., and Wally, K. D.: Spatial trends in saturated hydraulic conductivity of vegetated dunes in the Nebraska Sand Hills: Effects of depth and topography, Journal of Hydrology, 349, 88–97, 2008.

Weynants, M., Montanarella, L., Toth, G., Arnoldussen, A., Anaya Romero, M., Bilas, G., Borresen, T., Cornelis, W., Daroussin, J., Gonçalves, M. D. C., et al.: European HYdropedological Data Inventory (EU-HYDI), EUR Scientific and Technical Research Series, 2013.

Wösten, J., Pachepsky, Y. A., and Rawls, W.: Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics, Journal of hydrology, 251, 123–150, 2001.

Wösten, J. et al.: The HYPRES database of hydraulic properties of European soils., Advances in GeoEcology, pp. 135–143, 2000.

Wright, M. N. and Ziegler, A.: Ranger: a fast implementation of random forests for high dimensional data in C++ and R, arXiv preprint arXiv:1508.04409, 2015.

Yao, S., Zhang, T., Zhao, C., and Liu, X.: Saturated hydraulic conductivity of soils in the Horqin Sand Land of Inner Mongolia, northern China, Environmental monitoring and assessment, 185, 6013–6021, 2013.

Yasin, S. and Yulnafatmawita, Y.: Effects of Slope Position on Soil Physico-chemical Characteristics Under Oil Palm Plantation in Wet Tropical Area, West Sumatra Indonesia, AGRIVITA, Journal of Agricultural Science, 40, 328–337, 2018.

Yoon, S. W.: A measure of soil structure derived from water retention properties: A kullback-Leibler distance approach, Ph.D. thesis, Rutgers University-Graduate School-New Brunswick, 2009.

Zakaria, S.: Water management in deep peat soils in Malaysia, Ph.D. thesis, Cranfield University, 1992.

Zhang, S., Xiahou, Y., Tang, H., Huang, L., Liu, X., and Wu, Q.: Study on the spatially variable saturated hydraulic conductivity and deformation behavior of accumulation reservoir landslide Based on surface nuclear magnetic resonance survey, Advances in civil engineering, 2018.

Zhao, H., Zeng, Y., Lv, S., and Su, Z.: Analysis of soil hydraulic and thermal properties for land surface modeling over the Tibetan Plateau, Earth system science data, 10, 1031, 2018.

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 1

```
%%
 https://www.earth-system-science-data.net/f
or_authors/manuscript_preparation.html

%% Copernicus Publications Manuscript
 Preparation Template for LaTeX Submissions
%% ---------------------------------
%% This template should be used for the
 following class files: copernicus.cls,
 copernicus2.cls, copernicus_discussions.cls
%% The class files, the Copernicus LaTeX
 Manual with detailed explanations regarding
 the comments
%% and some style files are bundled in the
 Copernicus Latex Package which can be
 downloaded from the different journal
 webpages.
%% For further assistance please contact the
 Publication Production Office
 (production@copernicus.org).
%% http://publications.copernicus.org
%% copernicus.cls

\documentclass[essd,manuscript]{copernicus}
%\documentclass[essd]{copernicus}
\bibliographystyle{copernicus}
\usepackage{natbib}
\usepackage{amssymb,amsmath}
\usepackage{graphicx}
\usepackage{textcomp}
\usepackage{array, rotating}
\usepackage{url}

%\usepackage[noae]{Sweave}
\usepackage{lineno}
\usepackage{caption}
\usepackage{subcaption}
\usepackage{multicol}
\usepackage{hyperref}
\usepackage{siunitx,booktabs}
\usepackage[T1]{fontenc}
\usepackage[utf8]{inputenc}
%\hypersetup{draft}
\hypersetup{colorlinks=false,
 linkcolor=black, citecolor=black,
 bookmarksnumbered=true,
        urlcolor=black, bookmarksopen=true,
 pdfview=FitH, pdfstartview=FitH,
        pdftitle={SoilKsatDB: global soil
 saturated hydraulic conductivity
 measurements for geoscience applications},
        pdfauthor={Gupta et al.}} %%
%\hypersetup{colorlinks=true,
 linkcolor=blue, citecolor=red,
 bookmarksnumbered=true,
%        urlcolor=blue, bookmarksopen=true,
 pdfview=FitH, pdfstartview=FitH,
%        pdftitle={SoilKsatDB: global soil
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 1

```
%%
 https://www.earth-system-science-data.net/f
or_authors/manuscript_preparation.html

%% Copernicus Publications Manuscript
 Preparation Template for LaTeX Submissions
%% ---------------------------------
%% This template should be used for the
 following class files: copernicus.cls,
 copernicus2.cls, copernicus_discussions.cls
%% The class files, the Copernicus LaTeX
 Manual with detailed explanations regarding
 the comments
%% and some style files are bundled in the
 Copernicus Latex Package which can be
 downloaded from the different journal
 webpages.
%% For further assistance please contact the
 Publication Production Office
 (production@copernicus.org).
%% http://publications.copernicus.org
%% copernicus.cls

\documentclass[essd,manuscript]{copernicus}
%\documentclass[essd]{copernicus}
\bibliographystyle{copernicus}
\usepackage{natbib}
\usepackage{amssymb,amsmath}
\usepackage{graphicx}
\usepackage{textcomp}
\usepackage{array, rotating}
\usepackage{url}

%\usepackage[noae]{Sweave}
\usepackage{lineno}
\usepackage{caption}
\usepackage{subcaption}
\usepackage{multicol}
\usepackage{hyperref}
\usepackage{siunitx,booktabs}
\usepackage[T1]{fontenc}
\usepackage[utf8]{inputenc}
%\hypersetup{draft}
\hypersetup{colorlinks=false,
 linkcolor=black, citecolor=black,
 bookmarksnumbered=true,
        urlcolor=black, bookmarksopen=true,
 pdfview=FitH, pdfstartview=FitH,
        pdftitle={SoilKsatDB: global soil
 saturated hydraulic conductivity
 measurements for geoscience applications},
        pdfauthor={Gupta et al.}} %%
%\hypersetup{colorlinks=true,
 linkcolor=blue, citecolor=red,
 bookmarksnumbered=true,
%        urlcolor=blue, bookmarksopen=true,
 pdfview=FitH, pdfstartview=FitH,
%        pdftitle={SoilKsatDB: global soil
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 37

```
 saturated hydraulic conductivity
 measurements %for geoscience applications},
%        pdfauthor={Gupta et al.}} %%
%\renewcommand{\ttdefault}{cmcr}
%\renewcommand{\sfdefault}{cmss}

\begin{document}
\title{SoilKsatDB: global soil saturated
 hydraulic conductivity measurements for
 geoscience applications}
\author[1]{Surya~Gupta}
\author[2]{Tomislav~Hengl}
\author[1]{Peter~Lehmann}
\author[1]{Sara~Bonetti}
\author[1]{Dani~Or}
\affil[1]{Soil and Terrestrial Environmental
 Physics, Department of Environmental
 Systems Science, ETH, Z\"urich,
 Switzerland}
\affil[2]{OpenGeoHub foundation /
 EnvirometriX, Wageningen, the Netherlands}

\runningtitle{A compilation of soil
 saturated hydraulic conductivity
 measurements}
\runningauthor{Gupta~S.}
\correspondence{Gupta~S.\\
 surya.gupta@usys.ethz.ch}

\received{}
\revised{}
\accepted{}
\published{}

%% These dates will be inserted by the
 Publication Production Office during the
 typesetting process.

\firstpage{1}

\maketitle

\begin{abstract}
Saturated soil hydraulic conductivity (Ksat)
 is a key parameter in many hydrological anc
 climatic modeling applications as it
 controls the partitioning between
 precipitation, infiltration and runoff.
 Ksat values are primarily determined from
 soil textural properties and soil forming
 processes, and may vary over several orders
 of magnitude. Despite availability of Ksat
 datasets at catchment or regional scale,
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 38

```
 saturated hydraulic conductivity
 measurements %for geoscience applications},
%        pdfauthor={Gupta et al.}} %%
%\renewcommand{\ttdefault}{cmcr}
%\renewcommand{\sfdefault}{cmss}

\begin{document}
\title{SoilKsatDB: global soil saturated
 hydraulic conductivity measurements for
 geoscience applications}
\author[1]{Surya~Gupta}
\author[2]{Tomislav~Hengl}
\author[1]{Peter~Lehmann}
\author[1,3]{Sara~Bonetti}
\author[1,4]{Dani~Or}
\affil[1]{Soil and Terrestrial Environmental
 Physics, Department of Environmental
 Systems Science, ETH, Z\"urich,
 Switzerland}
\affil[2]{OpenGeoHub foundation /
 EnvirometriX, Wageningen, the Netherlands}
\affil[3]{Institute for Sustainable
 Resources,  University College London,
 London, UK}

\affil[4]{Division of Hydrologic Sciences,
 Desert Research Institute, Reno, NV, USA}
\runningtitle{A compilation of soil
 saturated hydraulic conductivity
 measurements}
\runningauthor{Gupta~S.}
\correspondence{Gupta~S.\\
 surya.gupta@usys.ethz.ch}

\received{}
\revised{}
\accepted{}
\published{}

%% These dates will be inserted by the
 Publication Production Office during the
 typesetting process.

\firstpage{1}

\maketitle

\begin{abstract}
Saturated soil hydraulic conductivity (Ksat)
 is a key parameter in many hydrological anc
 climatic modeling applications Ksat values
 are primarily determined from soil textural
 properties and may vary over several orders
 of magnitude. Despite availability of Ksat
 datasets in the literature, significant
 efforts are required to import and combine
 the data before it can be used for specific
 applications. In this work, a total of
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 68

```
significant efforts are required to import
and bind the data before it could be used
for modeling. In this work, a total of
1,910 sites with 13,267 Ksat measurements
were assembled from published literature
and other sources, standardized, and
quality-checked in order to provide a
global database of soil saturated hydraulic
conductivity (SoilKsatDB). The SoilKsatDB
covers most global regions, with the
highest data density from the USA, followed
by Europe, Asia, South America, Africa, and
Australia. In addition to Ksat, other soil
variables such as soil texture (11,667
measurements), bulk density (11,151
measurements), soil organic carbon (9,787
measurements), field capacity (7,389) and
wilting point (7,418) are also included in
the dataset. The results of using the
SoilKsatDB to fit Ksat pedotransfer
functions (PTFs) for temperate climatic
regions and laboratory based soil samples
based on soil properties (sand and clay
content, bulk density) show that reasonably
accurate models can be fitted using Random
Forest (best CCC = 0.70 and CCC = 0.73 for
temperate and lab based measurements,
respectively). However when temperate and
laboratory based Ksat PTFs are applied to
soil samples from tropical climates and
field measurements, respectively, the model
performance is significantly lower (CCC =
0.51 for tropical and CCC = 0.13 for field
samples). PTFs derived for temperate soils
and laboratory measurements might not be
suitable for estimating Ksat for tropical
regions or field measurements,
respectively. The SoilKsatDB dataset is
available at
\url{https://doi.org/10.5281/zenodo.375272
} \citep{surya_gupta_2020_3752722} and the
code used to produce the compilation is
publicly available under an open data
license.
\end{abstract}

\begin{pagewiselinenumbers}

\introduction
Soil saturated hydraulic conductivity (Ksat)
describes the water movement through water
saturated soils and is defined as ratio
between water flux and hydraulic gradient
\citep{amoozegar1986hydraulic}. It is a ke
variable in a number of hydrological,
geomorphological, and climatological
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 71

```
13,267 Ksat measurements from 1,910 sites
were assembled from published literature
and other sources, standardized (units made
identical), and quality-checked in order to
provide a global database of soil saturated
hydraulic conductivity (SoilKsatDB). The
SoilKsatDB covers most regions across the
globe, with the highest number of Ksat
measurements from North America, followed
by Europe, Asia, South America, Africa, and
Australia. In addition to Ksat, other soil
variables such as soil texture (11,591
measurements), bulk density (11,269
measurements), soil organic carbon (9,787
measurements), field capacity (7,389) and
wilting point (7,418) are also included in
the dataset. To show an application of
SoilKsatDB, we fit Ksat pedotransfer
functions (PTFs) for temperate regions and
laboratory-based soil properties (sand and
clay content, bulk density). Accurate
models can be fitted using a Random Forest
machine learning algorithm (best
concordance correlation coefficient (CCC) =
0.70 and CCC = 0.73 for temperate and
laboratory-based measurements,
respectively). However, when these
temperate and laboratory based Ksat PTFs
are applied to soil samples from tropical
climates and field measurements,
respectively, the model performance is
significantly lower (CCC = 0.52 for
tropical and CCC = 0.10 for field samples).
These results indicate that there are
significant differences between Ksat data
collected in temperate and tropical regions
and measured in lab or the field. The
SoilKsatDB dataset is available at
\emph{'version 0.3'}
\url{https://doi.org/10.5281/zenodo.375272
} \citep{surya_gupta_2020_3752722} and the
code used to extract the data from the
literature, for the quality control and
applied random forest machine learning
approach is publicly available under an
open data license.
\end{abstract}

\begin{pagewiselinenumbers}

\introduction
Soil saturated hydraulic conductivity (Ksat)
describes the rate of water movement
through water saturated soils and is
defined as the ratio between water flux and
hydraulic gradient
\citep{amoozegar1986hydraulic}. It is a ke
variable in a number of hydrological,
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 74

```
applications, such as rainfall partitioning into infiltration and runoff \citep{vereecken2010using}, optimal irrigation design \citep{hu2015effects}, as well as the prediction of natural hazards including catastrophic floods and landslides \citep{batjes1996total,glinski2000characteristics, zhang2018study}. Accurate measurements of Ksat in the laboratory and field are laborious and time consuming and most samples are taken from agricultural soils \citep{romano2002prediction}.

Efforts to produce reliable and spatially refined datasets of hydraulic properties date back to the 1970's with the proliferation of distributed hydrologic and climatic modeling. Some of these early notable works also providedsome of the basic databases (some of which are used in this study) for Australia \citep{mckenzie2008online,forrest1985survey}, Belgium \citep{vereecken2017soil,cornelis2001evaluation}, Brazil \citep{tomasella2000pedotransfer,tomasella2003comparison,ottoni2018hydrophysical}, France \citep{bruand2004estimation}, Germany \citep{horn1991labordatenbank,krahmer1995ermittlung}, Hungary \citep{nemes2002unsaturated}, the Netherlands \citep{wosten2001pedotransfer}, Poland \citep{glinski1991soil}, and USA \citep{rawls1982estimation}. \citet{nemes2011databases} discussed the available datasets on Ksat and hydro-physical properties in detail. Collaborative efforts have resulted in the compilation of multiple databases, including the Unsaturated Soil Hydraulic Database (UNSODA) \citep{nemes2001description}, the Grenoble Catalogue of Soils (GRIZZLY) \citep{haverkamp1998grizzly}, and the Mualem cataloge \citep{mualem1976catalogue} - these however focused on soil types and not on spatially context mapping of Ksat. In an effort to provide spatial context, \citet{jarvis2013influence}, \citet{rahmati2018development} and \citet{schindler2017soil} published global databases for soil hydraulic and soil physical properties. Likewise, the European soil data center also started projectsfor generating spatially referenced databases
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 77

```
geomorphological, and climatological applications, such as rainfall partitioning into infiltration and runoff \citep{vereecken2010using}, optimal irrigation design \citep{hu2015effects}, as well as the prediction of natural hazards including catastrophic floods and landslides \citep{batjes1996total,glinski2000characteristics, zhang2018study}. Accurate measurements of Ksat in the laboratory and field are laborious and time consuming and most samples are taken from agricultural soils \citep{romano2002prediction}.

Efforts to produce reliable and spatially refined datasets of hydraulic properties date back to the 1970's with the proliferation of distributed hydrologic and climatic modeling. Some of these early notable works also provided basic databases (some of which are used in this study) for Australia \citep{mckenzie2008online,forrest1985survey}, Belgium \citep{vereecken2017soil,cornelis2001evaluation}, Brazil \citep{tomasella2000pedotransfer,tomasella2003comparison,ottoni2018hydrophysical}, France \citep{bruand2004estimation}, Germany \citep{horn1991labordatenbank,krahmer1995ermittlung}, Hungary \citep{nemes2002unsaturated}, the Netherlands \citep{wosten2001pedotransfer}, Poland \citep{glinski1991soil}, and USA \citep{rawls1982estimation}. \citet{nemes2011databases} discussed the available datasets on Ksat andother hydro-physical properties in detail. Collaborative efforts have resulted in the compilation of multiple databases, including the Unsaturated Soil Hydraulic Database (UNSODA) \citep{nemes2001description}, the Grenoble Catalogue of Soils (GRIZZLY) \citep{haverkamp1998grizzly}, and the Mualem cataloge \citep{mualem1976catalogue} - these, however, focused on soil types and not on the spatial context of Ksat mapping. In an effort to provide spatial context, \citet{jarvis2013influence}, \citet{rahmati2018development} and \citet{schindler2017soil} published global databases for soil hydraulic and soil physical properties. Likewise, the European soil data center also started projects such as SPADE \citep{hiederer2006soil}and
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 76

for several countries such as SPADE \citep{hiederer2006soil} and HYPRES \citep{wosten2000hypres}. Since HYPRES represents only western European countries, \citet{weynants2013european} gathered the data from 18 countries and developed the European Hydropedological Data Inventory (EU-HYDI) database - this dataset is, however, not publicly available and was not included in this compilation. The datasets mentioned above cover almost all climatic zones except tropical regions, where Ksat values could be significantly different due to the strong local weathering processes \citep{hodnett2002marked}. Recently, \citet{ottoni2018hydrophysical} published a dataset named HYBRAS (Hydrophysical Database for Brazilian Soils) improving the coverage of South American tropical regions. In addition, \cite{rahmati2018development} recently published the Soil Water Infiltration Global database (SWIG) collecting information on Ksat for the whole globe as deduced from infiltration experiments.

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 79

HYPRES \citep{wosten2000hypres}, for

generating spatially referenced databases for several countries. Since HYPRES represents only western European countries, \citet{weynants2013european} gathered data from 18 countries and developed the European Hydropedological Data Inventory (EU-HYDI) database - this dataset is, however, not publicly available and was not included in this compilation. The datasets mentioned above cover almost all climatic zones except tropical regions, where Ksat values can be significantly different due to the strong local weathering processes and different clay mineralogy \citep{hodnett2002marked}. Recently, \citet{ottoni2018hydrophysical} published a dataset named HYBRAS (Hydrophysical Database for Brazilian Soils) improving the coverage of South American tropical regions. In addition, \cite{rahmati2018development} recently published the Soil Water Infiltration Global database (SWIG) collecting information on Ksat for the whole globe. In SWIG database, some Ksat values were extracted from literature and other Ksat values were deduced from infiltration time series. In contrast to lab measurements that determine Ksat as ratio of flux density to gradient, infiltration-based methods determine Ksat by fitting infiltration dynamics to parametric models (using three-parameter infiltration

The increased observation of various surface properties using satellite based imaging capability as well as the ever increasing demand for highly resolved description of surface processes require commensurate advances in Ksat representation for modern Earth System Model (ESM) applications. Despite availability of datasets at catchment or regional scale, to be able to use the various soil datasets listed above for global modeling, a significant amount of time is required to import and bind data. In addition, several existing Ksat datasets miss either coordinates of points or these have been recorded with unknown accuracy thus limiting their applications for spatial modeling. For example the SWIG dataset misses information on soil depth and assigns a single coordinate for entire watersheds. Similarly, UNSODA dataset does not provide coordinates and soil texture information for all samples. For a few locations, HYBRAS uses a different coordinate system. Taken together, these limitations highlight that, to prepare spatially referenced global Ksat datasets for large scale applications, a serious effort to compile, standardize and quality check all literature (available publicly) is often required.

The objective of the work here is to provide a new global standardized Ksat database (SoilKsatDB) that can be used for geoscience applications. To do so, a total of 13,267 Ksat measurements have been collected, standardized, and cross-checked to produce a harmonized compilation which is analysis-ready (i.e., it can directly be used for model fitting and spatial analysis). We collected data from existing datasets and, to improve the spatial coverage in regions with sparse data, we have further conducted a literature search to include Ksat measurements in geographic areas that were not yet covered in other existing databases.

In the manuscript, we first describe the data collection process and then describe methodological steps used to spatially reference, filter, and standardize existing datasets. As an illustrative application of the dataset we derive pedotransfer functions (PTFs) for different regions and

equation of Philip \citep{kutilek1987three or simplified form of \citet{haverkamp1994three}).

The ever increasing demand for highly resolved description of surface processes require commensurate advances in Ksat representation for modern Earth System Model (ESM) applications. Several existing Ksat datasets miss either coordinates or these have been recorded with unknown accuracy thus limiting their applications for spatial modeling. For example, the SWIG dataset misses information on soil depth and assigns a single coordinate for entire watersheds. Similarly, the UNSODA dataset does not provide coordinates and soil texture information for all samples. For a few locations, HYBRAS uses a different coordinate system. Taken together, these limitations highlight that, to prepare spatially referenced global Ksat datasets for large scale applications, a serious effort to compile, standardize and quality check all literature (available publicly) is often required.

The objective of the work here is to provide a new global standardized Ksat database (SoilKsatDB) that can be used for geoscience applications. To do so, a total of 13,267 Ksat measurements have been collected, standardized, and cross-checked to produce a harmonized compilation which is analysis-ready (i.e., it can directly be used for model fitting and spatial analysis). We compiled data from existing datasets and, to improve the spatial coverage in regions with sparse data, we further conducted a literature search to include Ksat measurements in geographic areas that were not yet covered in other existing databases.

In the manuscript, we first describe the data compilation process and then describe methodological steps used to spatially reference, filter, and standardize the existing datasets. As an illustrative application of the dataset, we derive PTFs for different regions and measurement

```
     measurement methods and discuss their
     transferability to other regionsand
     measurement methodologies.
082 We fully document all importing,
     standardization and binding steps using R
     environment for statistical computing
     \citep{Rbook}, so that we can collect
     feedback from other researchers and
     increase the speed of further updates and
     improvements. The newly created data set
     (SoilKsatDB) can be accessed via
     \url{https://doi.org/10.5281/zenodo.3752721
    } and directly used to test various Machine
     Learning algorithms
     \citep{casalicchio2017openml}.
083
084 \section{Methods and materials}
085 \subsection{Data sources}
086 To locate and obtain all compatible datasets
     for compilation, a literature search was
     conducted using different search engines,
     including Science Direct
     (\url{https://www.sciencedirect.com/}),
     Google Scholar
     (\url{https://scholar.google.com/}) and
     Scopus (\url{https://www.scopus.com}). We
     searched soil hydraulic conductivity
     datasets using keywords such as
     \emph{``saturated hydraulic conductivity
     database''}, \emph{``Ksat''}, and similar.
     The collected datasets are listed in
     Table~\ref{tab:my_label} together with
     number of Ksat observations for each study,
     and  can be classified into three main
     categories, namely:
087 i)  Existing datasets (in form of tables)
     published and archived with a DOI in a
     peer-review publication; ii) legacy
     datasets in paper/document format
     (e.g.,legacy reports, PhD theses, and
     scientific studies), iii) on-line
     materials.
088
089 Existing datasets include published datasets
     such as HYBRAS
     \citep{ottoni2018hydrophysical}, UNSODA
     \citep{nemes2001description}, SWIG
     \citep{rahmati2018development}, and the
     soil hydraulic properties over the Tibetan
     Plateau \citep{zhao2018analysis}, from
     which we extracted the required information
     as described in Table~\ref{tab:list_names}.
     The major challenge with making the
     existing datasets compatible for binding
     (standardization, removing redundancy), was
     to obtain the locations for a particular
     sample as well as the corresponding
```

```
     methods and discuss their transferability
     to other regions/measurement methodologies.
086
    We fully document all importing,
     standardization and binding steps usingthe
     R environment for statistical computing
     \citep{Rbook}, so that we can collect
     feedback from other researchers and
     increase the speed of further updates and
     improvements. The newly created data set
     (SoilKsatDB) can be accessed via
     \emph{'version 0.3'}
     \url{https://doi.org/10.5281/zenodo.3752721
    } and directly used to test various Machine
     Learning algorithms
     \citep{casalicchio2017openml}.
087
088 \section{Methods and materials}
089 \subsection{Data sources}
090 To locate and obtain all compatible datasets
     for compilation, a literature search was
     conducted using different search engines,
     including Science Direct
     (\url{https://www.sciencedirect.com/}),
     Google Scholar
     (\url{https://scholar.google.com/}) and
     Scopus (\url{https://www.scopus.com}). We
     searched soil hydraulic conductivity
     datasets using keywords such as
     \emph{``saturated hydraulic conductivity
     database''}, \emph{``Ksat''}, and similar.
     The collected datasets are listed in
     Table~\ref{tab:my_label} together with
     number of Ksat observations for each study,
     and  can be classified into three main
     categories, namely:
091 i)  Existing datasets (in form of tables)
     published and archived with a DOI in a
     peer-review publication; ii) legacy
     datasets in paper/document format (e.g.,
     legacy reports, PhD theses, and scientific
     studies), iii) on-line materials.
092
093 Existing datasets include published datasets
     such as HYBRAS
     \citep{ottoni2018hydrophysical}, UNSODA
     \citep{nemes2001description}, SWIG
     \citep{rahmati2018development}, and the
     soil hydraulic properties over the Tibetan
     Plateau \citep{zhao2018analysis}, from
     which we extracted the required information
     as described in Table~\ref{tab:list_names}.
     The major challenge with making the
     existing datasets compatible for binding
     (standardization, removing redundancy), was
     to obtain the locations for a particular
     sample as well as the corresponding
```

```
/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 89
        measurement depths. For instance, the
        UNSODA database completely lacks
        geographical locations. To fill the gaps
        and make the data suitable also for spatial
        analysis, we used Google Earth to find the
        coordinates based on the given location
        (generally an address or a location name).
        We separated the data based on laboratory
        and field measurements and we computed
        sand, silt and clay contents based on the
        algorithm described in
        \citet{nemes2001description}.We further
        note that, in some datasets, the
        coordinates were missing or reported in
        diverse coordinate systems. For example, in
        the HYBRAS database, the locations needed
        to be converted from UTM to a decimal
        degrees. In the SWIG database, the
        information related to location
        (coordinates for each point), soil depth
        and measurement method (laboratory or
        field) was completely missing, so we went
        through each publication referenced in
        \citet{rahmati2018development}(except the
        unpublished literature) and added
        coordinates and applied the necessary
        conversions.
090
091   \begin{table} [htbp]
092       \centering
093       \caption{List of reference articles and
        digitized Ksat datasets, and number of
        points (N) per data set used to generate
        the new SoilKsatDB product.}
094       \small\addtolength{\tabcolsep}{0pt}
095       \begin{tabular}{lc|lc|lc}
096       \hline
097        Reference & $N$  & Reference & $N$ &
```

```
/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 93
        measurement depths. For instance, the
        UNSODA database completely lacks
        geographical locations. To fill the gaps
        and make the data suitable also for spatial
        analysis, we used Google Earth to find the
        coordinates based on the given location
        (generally an address or a location name).
        We separated the UNSODA data based on
        laboratory and field measurements and we
        computed sand, silt and clay contents basec
        on the particle diameters between
094   0-2 µm (clay), 2-50 µm (silt), and >50 µm
        (sand)
095   from the available particle-size data,
        assuming a log-normal distribution as
        described in \citet{nemes2001description}.
        We further note that, in some datasets, the
        coordinates were missing or reported in
        diverse coordinate systems. For example, ir
        the HYBRAS database, the locations needed
        to be converted from UTM to a decimal
        degrees. In the SWIG database, the
        information related to location
        (coordinates for each point), soil depth
        and measurement method (laboratory or
        field) was completely missing, so we went
        through each publication referenced in
        \citet{rahmati2018development} (except the
        unpublished literature) and added
        coordinates and applied the necessary
        conversions.
096
097   \begin{table} [htbp]
098       \centering
099       \caption{List of reference articles and
        digitized Ksat datasets, and number of
        points (N) per data set used to generate
        the new SoilKsatDB product.}
100       \small\addtolength{\tabcolsep}{0pt}
101       \begin{tabular}{lc|lc|lc}
102       \hline
103        Reference & $N$  & Reference & $N$ &
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 97

```
      Reference & $N$  \\
098         \hline
099    \citet{rycroft1975transmission}&       1
        \citet{abagandura2017influence}&      3 &
        \citet{jabro1992estimation}&  18\\
100    \citet{waddington1997groundwater}&     1&
        \citet{habel2013role}&3 &
        \citet{greenwood2014effects}&18       \\
101    \citet{takahashi1997studies}&  1&
        \citet{nyman2011evidence}&3 &
        \citet{wang2008spatial}&      19\\
102    \citet{katimon1997field}&      1&
        \citet{habel2013role}&3 &
        \citet{deshmukh2014pragmatic}& 19\\
103    \citet{el1994impact}&
              1&\citet{bhattacharyya2006effect}&4&
        \citet{price2010variation}&   20\\
104    \citet{lopez2015method}&       1&
        \citet{lopes2020establishment}&4&
        \citet{bonsu1996saturated}&24 \\
105    \citet{kramarenko2019hydraulic}&       1&
        \citet{yasin2018effects}&4&
        \citet{bambra2016soil} &      24      \\
106    \citet{zakaria1992water}&      1&
        \citet{daniel2017spatial}&6&
        \citet{verburg2001properties}& 26\\
107    \citet{ramli1999management}&   1&
        \citet{anapalli2005effectiveness}&7 &
        \citet{southard1988subsoil} & 27\\
108    \citet{singh2011soil}&  1&
        \citet{arend1941infiltration}&7&
        \citet{chang2010predictions}&  30
109    \citet{campbell1977wildfire}&1 &
        \citet{helbig2013spatial}&7 &
        \citet{yao2013saturated}&     33      \\
110    \citet{chief2008correlation}&1 &
        \citet{gwenzi2011field}&7 &
        \citet{becker2018impact} & 34\\
111    \citet{conedera2003consequences}&1
        \citet{paivanen1973hydraulic}&9&
        \citet{baird2017high} &       50\\
112    \citet{ebel2012hydrologic}&1  &
        \citet{mahapatra2019estimation} &9&
        \citet{keisling1974precision}& 56     \\
113    \citet{ferreira2005temporal}&1 &
        \citet{amer2009prediction}&    9&
        \citet{rahimy2011effects}&    56      \\
114    \citet{imeson1992effects}&1  &
        \citet{vogeler2019estimation} &       10 &
        \citet{hao2019impacts} &      57      \\
115        \citet{johansen2001post}&1 &
        \citet{singh2006water}&10&
        \citet{Kanemasu1994} &60      \\
116    \citet{lamara2008prediction}&1 &
        \citet{kelly2014high}&10 &
        \citet{tete1993evaluation}& 60 \\
117    \citet{parks1989soil}&1 &
        \citet{article}&11&
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 103

```
      Reference & $N$  \\
104         \hline
105    \citet{rycroft1975transmission}&       1
        \citet{abagandura2017influence}&      3 &
        \citet{jabro1992estimation}&  18\\
106    \citet{waddington1997groundwater}&     1&
        \citet{habel2013role}&3 &
        \citet{greenwood2014effects}&18       \\
107    \citet{takahashi1997studies}&  1&
        \citet{nyman2011evidence}&3 &
        \citet{wang2008spatial}&      19\\
108    \citet{katimon1997field}&      1& \citet
        {bhattacharyya2006effect}&4&
        \citet{deshmukh2014pragmatic}& 19\\
109    \citet{el1994impact}&
              1&\citet{lopes2020establishment}&4 &
        \citet{price2010variation}&   20\\
110    \citet{lopez2015method}&       1&
        \citet{yasin2018effects}&4&
        \citet{bonsu1996saturated}&24 \\
111    \citet{kramarenko2019hydraulic}&       1&
        \citet{daniel2017spatial}&6&
        \citet{bambra2016soil} &      24      \\
112    \citet{zakaria1992water}&      1&
        \citet{anapalli2005effectiveness}&7 &
        \citet{verburg2001properties}& 26\\
113    \citet{ramli1999management}&   1&
        \citet{arend1941infiltration}&7&
        \citet{southard1988subsoil} & 27\\
114    \citet{singh2011soil}&  1&
        \citet{helbig2013spatial}&7 &
        \citet{chang2010predictions}&  30
115    \citet{campbell1977wildfire}&1 &
        \citet{gwenzi2011field}&7 &
        \citet{yao2013saturated}&     33      \\
116    \citet{chief2008correlation}&1 &
        \citet{paivanen1973hydraulic}&9&
        \citet{becker2018impact} & 34\\
117    \citet{conedera2003consequences}&1
        \citet{mahapatra2019estimation} &9&
        \citet{baird2017high} &       50\\
118    \citet{ebel2012hydrologic}&1  &
        \citet{amer2009prediction}&    9&
        \citet{keisling1974precision}& 56     \\
119    \citet{ferreira2005temporal}&1 &
        \citet{radcliffe1990infiltration}&10&
        \citet{rahimy2011effects}&    56      \\
120    \citet{imeson1992effects}&1  &
        \citet{vogeler2019estimation} &       10 &
        \citet{hao2019impacts} &      57      \\
121        \citet{johansen2001post}&1 &
        \citet{singh2006water}&10&
        \citet{Kanemasu1994} &60      \\
122    \citet{lamara2008prediction}&1 &
        \citet{kelly2014high}&10 &
        \citet{tete1993evaluation}& 60 \\
123    \citet{parks1989soil}&1 &
        \citet{article}&11&
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 117

```
     \citet{zhao2018analysis}&     65      \\
118 \citet{ravi2017ecohydrological}&1 &
     \citet{ganiyu2018predicting} &12 &
     \citet{hinton2016land}&       77      \\
119 \citet{smettem1992measurement}&1       &
     \citet{cisneros1999vegetation}&12&
     \citet{vieira2004landslides}&86\\
120 \citet{helbig2013spatial}&     2&
     \citet{niemeyer2014woody} &12 &
     \citet{houghton2011hydrogeologic} &   88\\
121 \citet{boike1998thermal}&      2 &
     \citet{sharratt1990water} & 14 &
     \citet{tian2017variability}&  91      \\
122 \citet{andrade1971influence}&
            2&\citet{habecker1990identification}
     \citet{li2017multiscale}&     108     \\
123 \citet{beyer2015estimation}&2 &
     \citet{nielsen1973spatial}&14 &
     \citet{forrest1985survey}&    120 \\
124 \citet{blake2010wildfire}&2&
     \citet{robbins1977hydraulic} &15 &
     \citet{Richard1987Schweiz} &  121     \\
125 \citet{bonell1986two}&2&
     \citet{sonneveld2005multi}&15&
     \citet{sanzeni2013specific} & 127     \\
126 \citet{kutiel1995effect}&2 &
     \citet{quinton2008peat}&16&
     \citet{vereecken2017soil}& 145 \\
127     \citet{martin2001comparison}&2 &
     \citet{simmons2014soil}&16&
     \citet{coelho1974spatial}& 17 \\
128 \citet{mott1979soil}&2 &
     \citet{ouattara1977variation} &17&
     \citet{kool1986physical}&     240     \\
129 \citet{rab1996soil}&2 &
     \citet{hardie2011effect}& 17&
     \citet{nemes2001description}& 283     \\
130 \citet{soracco2010anisotropy}&2&
     \citet{baird1997field}&       17&
     \citet{ottoni2018hydrophysical}&      326\\
131 \citet{varela2015influence}& 2&
     \citet{kirby2001texture}&17 &
     \citet{rahmati2018development}&3637\\
132 \citet{sayok2007hydraulic}&    3&
     \citet{yoon2009measure}&      18 &
     \citet{Floridadatabase}&6532\\
133  \hline
134     \end{tabular}
135     \label{tab:my_label}
136 \end{table}
137
138 In the case of legacy datasets (paper or
     document format, data from journals,
     theses, and legacy reports with and withou
     peer-reviewed publications), we invested a
     significant effort to digitize tabular
     data, clean it and make it analysis-ready.
     In some cases we had to convert PDF
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 123

```
     \citet{zhao2018analysis}&     65      \\
124 \citet{ravi2017ecohydrological}&1 &
     \citet{ganiyu2018predicting} &12 &
     \citet{hinton2016land}&       77      \\
125 \citet{smettem1992measurement}&1       &
     \citet{cisneros1999vegetation}&12&
     \citet{vieira2004landslides}&86\\
126 \citet{helbig2013spatial}&     2&
     \citet{niemeyer2014woody} &12 &
     \citet{houghton2011hydrogeologic} &   88\\
127 \citet{boike1998thermal}&      2 &
     \citet{sharratt1990water} & 14 &
     \citet{tian2017variability}&  91      \\
128 \citet{andrade1971influence}&
            2&\citet{habecker1990identification}
     \citet{li2017multiscale}&     118     \\
129 \citet{beyer2015estimation}&2 &
     \citet{nielsen1973spatial}&14 &
     \citet{forrest1985survey}&    118 \\
130 \citet{blake2010wildfire}&2&
     \citet{robbins1977hydraulic} &15 &
     \citet{Richard1987Schweiz} &  121     \\
131 \citet{bonell1986two}&2&
     \citet{sonneveld2005multi}&15&
     \citet{sanzeni2013specific} & 127     \\
132 \citet{kutiel1995effect}&2 &
     \citet{quinton2008peat}&16&
     \citet{vereecken2017soil}& 145 \\
133     \citet{martin2001comparison}&2 &
     \citet{simmons2014soil}&16&
     \citet{coelho1974spatial}& 176 \\
134 \citet{mott1979soil}&2 &
     \citet{ouattara1977variation} &17&
     \citet{kool1986physical}&     240     \\
135 \citet{rab1996soil}&2 &
     \citet{hardie2011effect}& 17&
     \citet{nemes2001description}& 283     \\
136 \citet{soracco2010anisotropy}&2&
     \citet{baird1997field}&       17&
     \citet{ottoni2018hydrophysical}&      326\\
137 \citet{varela2015influence}& 2&
     \citet{kirby2001texture}&17 &
     \citet{rahmati2018development}&3637\\
138 \citet{sayok2007hydraulic}&    3&
     \citet{yoon2009measure}&      17 &
     \citet{Floridadatabase}&6532\\
139  \hline
140     \end{tabular}
141     \label{tab:my_label}
142 \end{table}
143
144 In the case of legacy datasets (paper or
     document format, data from journals,
     theses, and legacy reports with and withou
     peer-reviewed publications), we invested a
     significant effort to digitize tabular
     data, clean it and make it analysis-ready.
     After the digitization process, all data
```

	/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 138		/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 144
	documents to Microsoft Word files, after that to tabular data. Some documents had to be digitized manually due to the low resolution of PDFs. After the digitization process, all data values were cross-checked one more time with the original PDFs to avoid any artifacts or gross error in the final database.		values were cross-checked one more time with the original PDFs to avoid any artifacts or error in the final database.
		145	
		146	
139		147	
140	Two datasets were also collected directly from project websites that might be peer reviewed such as the NASA project based on hydraulic and thermal conductivity (retrieved from \url{https://daac.ornl.gov/FIFE/guides/Soil_Hydraulic_Conductivity_Data.html} and described in \citet{Kanemasu1994}) and the Florida database from \citet{Floridadatabase}.	148	Two datasets were also collected directly from project websites that might be peer reviewed such as the NASA project based on hydraulic and thermal conductivity (retrieved from \url{https://daac.ornl.gov/FIFE/guides/Soil_Hydraulic_Conductivity_Data.html} and described in \citet{Kanemasu1994}) and the Florida database from \citet{Floridadatabase}.
141		149	
142	Besides these, there are many locations, such as desert dunes, peatlands, frozen soils, and similar, in the world, where very few data of Ksat were available publicly. Because it is essential for global modeling to provide some values or range to reduce the uncertainty in the spatial maps, we have also intensively searched for these areas and found several minor studies providing Ksat values in these locations. We then digitized the Ksat values from these studies (shown either in bar charts and line plots), georeferenced the maps where necessary, and then converted the data into tabular form. All these datasets are also listed in Table~\ref{tab:my_label}.	150	There are many biomes and climatic regions, such as desert dunes, peatlandsand frozen soils, where very few data of Ksat were publicly available. Because it is essential for global modeling to provide some values or range to reduce the uncertainty in the spatial maps, we have also intensively searched for these areas and, in addition to the major datasets (SWIG, UNSODA HYBRAS), we have also found several minor studies (that contain less than 5 Ksat measurements) to cover these regions. We thus digitized Ksat values from these studies (shown either in bar charts or line plots), georeferenced the maps where necessary, and then converted the data into tabular form. All these datasets are also listed in Table~\ref{tab:my_label}. In some cases, we also contacted colleagues that worked in these regions to ask for data support.
143		151	
144	\begin{figure*}[!hbt]	152	\begin{figure*}[!hbt]
145	\centering	153	\centering
146	\begin{subfigure}[b]{\textwidth}	154	\begin{subfigure}[b]{\textwidth}
147	\includegraphics[width=.9\textwidth]{Ksat_points.pdf}	155	\includegraphics[width=.9\textwidth]{Global_points1.pdf}
148	\end{subfigure}	156	\end{subfigure}
149	\includegraphics[width=.2\textwidth]{12.jp }	157	\includegraphics[width=.2\textwidth]{121.jpg}
150	\caption{Spatial distribution of Ksat points (red and blue for field and laboratoratory measurements, respectively) in the SoilKsatDB. A total of 1,910 spatial	158	\caption{Spatial distribution of Ksat points (red and blue for laboratoy and field measurements, respectively) in the SoilKsatDB. A total of 1,910 spatial

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 150

```
 locations are on this map.}
    \label{Fig:points_map}
\end{figure*}

\subsection{Georeferencing Ksat values}

Georeferencing of Ksat measurements is
 important for using data for local,
 regional or global spatial modeling. Once
 georeferenced, points can be directly used
 in hydrological and land surface models.
 Although many studies providedthe
 information of spatial locations, however,
 the studies conducted in the 70's and 80’s
 only provided the name of the locations and
 approximate distance from the exact
 location. Therefore, we extracted the
 latitude and longitude of the location
 using Google maps for some datasets (which
 did not provide the spatial locations).
 Most of the studies we digitized provide
 maps or sketches with locations of the
 points. We first georeferenced these maps
 using ESRI ArcGIS software (v10.3) and then
 digitized the coordinates from
 georeferenced images. Some of the documents
 we digitized (e.g.
 \citet{nemes2001description})provided the
 names of the places, and hence we used
 Google Earth to obtain the coordinates. We
 estimate that the spatial location accurac
 of these points is roughly between 0 to
 5~km. Similarly, spatial maps in jpg forma
 (e.g. \citet{becker2018impact}) were
 geo-referenced with 100--500~m location
 accuracy. In contrast, few studies (e.g.
 \citet{yoon2009measure}) provided the
 extract location of the sampling with
 assumed location accuracy of 10--20~m.
\begin{table*}[ht]
  \renewcommand{\thetable}{\arabic{table}a}
%\begin{table*}[!hbt]
\begin{center}
\caption{Description and units of some key
 variables listed in the database. The
 complete list can be found in the link to
 the data base
 (\url{https://doi.org/10.5281/zenodo.37527
1}) in the readme-file. We used the same
 codes adopted in the National Cooperative
 Soil Survey (NCSS) Soil Characterization
 Database \citep{national2016national}.}
\addtolength{\tabcolsep}{0pt}
               \begin{tabular}{m{5cm} m{7cm
 m{2.1cm}}
               \hline
             Headers   &  Description  &
       Dimension\\
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 158

```
 locations are on this map.}
    \label{Fig:points_map}
\end{figure*}

\subsection{Georeferencing Ksat values}

Georeferencing of Ksat measurements is
 important for using data for local,
 regional or global spatial modeling. Once
 georeferenced, points can be directly used
 in hydrological and land surface models.
 Although many studies provided information
 on the geographical location of the
 measurements, the studies conducted in the
 70's and 80’s only provided the name of the
 locations and approximate distance from the
 exact location. Therefore, we extracted the
 latitude and longitude of the location
 using Google maps for some datasets (which
 did not provide the spatial locations). We
 digitized provided maps or sketches with
 locations of the points. We first
 georeferenced these maps using ESRI ArcGIS
 software (v10.3) and then digitized the
 coordinates from georeferenced images. Some
 of the documents we digitized (e.g.
 \citet{nemes2001description})provided the
 names specific locations, and hence we used
 Google Earth to obtain the coordinates. We
 estimate that the spatial location accurac
 of these points is roughly between 0 to
 5~km. Similarly, spatial maps in jpg forma
 (e.g. \citet{becker2018impact}) were
 geo-referenced with 100--500~m location
 accuracy. In contrast, few studies (e.g.
 \citet{yoon2009measure}) provided the
 extract location of the sampling with
 assumed location accuracy of 10--20~m.

\begin{table*}[ht]
  \renewcommand{\thetable}{\arabic{table}a}
%\begin{table*}[!hbt]
\begin{center}
\caption{Description and units of some key
 variables listed in the database. The
 complete list can be found in the link to
 the data base \emph{'version 0.3'}
 (\url{https://doi.org/10.5281/zenodo.37527
1}) in the readme-file. We used the same
 codes adopted in the National Cooperative
 Soil Survey (NCSS) Soil Characterization
 Database \citep{national2016national}.}
\addtolength{\tabcolsep}{0pt}
               \begin{tabular}{m{5cm} m{7cm
 m{2.1cm}}
               \hline
             Headers   &  Description  &
       Dimension\\
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 166

```
166       \hline
167     \verb"site_key"     &   Data set identif
    & ---  \\
168     \verb"longitude_decimal_degrees"    & Ran
    up to +180 degrees down to -180 degrees
    & Decimal degree       \\
169     \verb"latitude_decimal_degrees"     &
    Ranges up to +90 degrees down to -90
    degrees   & Decimal degree     \\




170     \verb"hzn_top" &   Top of soil sample   &
    cm     \\
171     \verb"hzn_bot" &   Bottom of soil sample
     & cm  \\
172     \verb"db_od" &   Bulk density  & g
    cm$^{-3}$      \\
173     \verb"w6clod"

174 &    Soil water content at 6 kPa   & vol \%
         \\
175 \verb"w10clod"
176 &    Soil water content at 10 kPa   & vol
    \%     \\
177 \verb"w3cld"
178 &    Soil water content at 33 kPa (field
    capacity)   & vol \%  \\
179 \verb"w15l2"
180 &    Soil water content at 1500 kPa
    (wilting point)   & vol \%    \\
181 \verb"tex_psda" &    Soil texture classes
    based on USDA  & ---   \\
182 \verb"clay_tot_psa" &     Mass of soil
    particles, < 0.002 mm  & \%   \\
183 \verb"silt_tot_psa" &     Mass of soil
    particles, > 0.002 and < 0.05 mm  & \% \\
184 \verb"sand_tot_psa" &    Mass of soil
    particle, > 0.05 and < 2 mm  & \%     \\
185 \verb"oc" &    Soil organic carbon content
    & \%   \\
186 \verb"ph_h2o" &  Soil acidity  & ---  \\
187 \verb"Ksat_lab" &    Soil saturated
    hydraulic conductivity from lab  & cm
    day$^{-1}$\\
188 \verb"Ksat_field" &    Soil saturated
    hydraulic conductivity from field  & cm
    day$^{-1}$ \\
189 \verb"source_db" & Sources of the datasets
       & ---       \\
190 \verb"confidence_degree" &  Reliability on
    the data set based on spatial locations  &
    ---    \\
191 \verb"location_id" & Combination of
    latitude and logitude   & --- \\
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 174

```
174       \hline
175     \verb"site_key"     &   Data set identif
    & ---  \\
176     \verb"longitude_decimal_degrees"    & Ran
    up to +180 degrees down to -180 degrees
    & Decimal degree       \\
177     \verb"latitude_decimal_degrees"     &
    Ranges up to +90 degrees down to -90
    degrees   & Decimal degree     \\
178     \verb"location_accuracy_min" &   Minimum
    value of location accuracy  & m       \\
179     \verb"location_accuracy_max" &   Maximum
    value of location accuracy  & m       \\
180        \verb"hzn_top" &   Top of soil sample
     & cm  \\
181     \verb"hzn_bot" &   Bottom of soil sample
     & cm  \\
182     \verb"hzn_desgn" &   Designation of soil
    horizon  & --- \\
183     \verb"db" &   Bulk density  & g cm$^{-3}$
            \\




184      \verb"w3cld"
185 &    Soil water content at 33 kPa (field
    capacity)   & vol \%  \\
186 \verb"w15l2"
187 &    Soil water content at 1500 kPa
    (wilting point)   & vol \%    \\
188 \verb"tex_psda" &    Soil texture classes
    based on USDA  & ---   \\
189 \verb"clay_tot_psa" &     Mass of soil
    particles, < 0.002 mm  & \%    \\
190 \verb"silt_tot_psa" &     Mass of soil
    particles, > 0.002 and < 0.05 mm  & \% \\
191 \verb"sand_tot_psa" &    Mass of soil
    particle, > 0.05 and < 2 mm  & \%     \\
192 \verb"oc_v" &    Soil organic carbon
    content  & \%  \\
193 \verb"ph_h2o_v" &  Soil acidity  & --- \\
194 \verb"Ksat_lab" &    Soil saturated
    hydraulic conductivity from lab  & cm
    day$^{-1}$\\
195 \verb"Ksat_field" &    Soil saturated
    hydraulic conductivity from field  & cm
    day$^{-1}$ \\
196 \verb"source_db" & Sources of the datasets
       & ---       \\



197 \verb"location_id" & Combination of
    latitude and logitude   & --- \\
198 \verb"hzn_depth" &  Mean depth of soil
    horizon  & --- \\
```

```
/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 192
192        \hline
193    \end{tabular}
194    \label{tab:list_names}
195    \end{center}
196    \end{table*}
197
198    \begin{table}[!hbtp]
199      \addtocounter{table}{-1}
200      \renewcommand{\thetable}{\arabic{table}b}
201      %\renewcommand{\theHtable}{\thetable B}%
        To keep hyperref happy
202      \caption{Example of Ksat database
        structure with key variables (from left to
        right: reference, longitudinal and
        latitudinal coordinates (decimal degree),
        top
203    and bottom of soil sample (cm), bulk density
        (g cm$^{-3}$), soil textural class, clay,
        silt  and sand content (\%) and saturated
        hydraulic conductivity measured in lab or
        field (cm day$^{-1}$). NA is ‘no value’.
        Note that the titles of the columns are
        explained in Table 2a.
204    }\label{second}
205      \addtolength{\tabcolsep}{0pt}
206                  \begin{tabular}{m{2.3cm}
        m{1.7cm}
        m{1.7cm}m{0.7cm}m{0.7cm}m{0.7cm}m{1.5cm}m{0
       .8cm}m{0.8cm}m{0.8cm}m{0.7cm}m{0.7cm}}
207                  \hline
208                 \texttt{site\_key}\par    &
        \texttt{longitude\_
209                 decimal\_
210                 degrees} & \texttt{latitude\_
211                 decimal\_
212                 degrees}& \texttt{hzn\_
213                 top}\par & \texttt{hzn\_
214                 bot}\par & \texttt{db\_

215                 od}\par  & \texttt{tex\_
216                 psda}\par  &      \texttt{clay\
217                 tot\_
218                 psa} &    \texttt{silt\_
219                 tot\_
220                 psa} &    \texttt{sand\_
221                 tot\_
222                 psa}&     \texttt{ksat\_
223                 lab}\par & \texttt{ksat\_
224                 field}\par\\
225                 \hline
226                 Saseendran\_2005  &  -103.15 &
             40.15&  15&     30&     1.33&Loam&
             232.08& NA\\
227                  Saseendran\_2005 &  -103.15 &
             40.15&  30&     60&     1.32&Loam&
             232.08& NA\\
228                   Saseendran\_2005 &  -103.15
             40.15&  60&     90&     1.36&Loam&
```

```
/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 199
199        \hline
200    \end{tabular}
201    \label{tab:list_names}
202    \end{center}
203    \end{table*}
204
205    \begin{table}[!hbtp]
206      \addtocounter{table}{-1}
207      \renewcommand{\thetable}{\arabic{table}b}
208      %\renewcommand{\theHtable}{\thetable B}%
        To keep hyperref happy
209      \caption{Example of Ksat database
        structure with key variables (from left to
        right: reference, longitudinal and
        latitudinal coordinates (decimal degree),
        top
210    and bottom of soil sample (cm), bulk density
        (g cm$^{-3}$), soil textural class, clay,
        silt  and sand content (\%) and saturated
        hydraulic conductivity measured in lab or
        field (cm day$^{-1}$)). NA is ‘no value’.
        Column names are explained in Table 2a.

211    }\label{second}
212      \addtolength{\tabcolsep}{0pt}
213                  \begin{tabular}{m{2.3cm}
        m{1.7cm}
        m{1.7cm}m{0.7cm}m{0.7cm}m{0.7cm}m{1.5cm}m{0
       .8cm}m{0.8cm}m{0.8cm}m{0.7cm}m{0.7cm}}
214                  \hline
215                 \texttt{site\_key}\par    &
        \texttt{longitude\_
216                 decimal\_
217                 degrees} & \texttt{latitude\_
218                 decimal\_
219                 degrees}& \texttt{hzn\_
220                 top}\par & \texttt{hzn\_
221                 bot}\par & \texttt{db}\par  &
        \texttt{tex\_

222                 psda}\par  &      \texttt{clay\
223                 tot\_
224                 psa} &    \texttt{silt\_
225                 tot\_
226                 psa} &    \texttt{sand\_
227                 tot\_
228                 psa}&     \texttt{ksat\_
229                 lab}\par & \texttt{ksat\_
230                 field}\par\\
231                 \hline
232                 Saseendran\_2005  &  -103.15 &
             40.15&  15&     30&     1.33&Loam&
             232.08& NA\\
233                  Saseendran\_2005 &  -103.15 &
             40.15&  30&     60&     1.32&Loam&
             232.08& NA\\
234                   Saseendran\_2005 &  -103.15
             40.15&  60&     90&     1.36&Loam&
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 228

```
        337.92& NA\\
               Saseendran\_2005 &  -103.15 &
        40.15&  90&120& 1.40&Loam&      12.0&
        284.88& NA\\
               Saseendran\_2005 &  -103.15 &
        40.15&  120&    150&    1.42&Loam&
        48.3&   259.20& NA\\
               Saseendran\_2005 &  -103.15 &
        40.15&  150&180&        1.42&Loam&
        48.3&   259.20& NA\\
               Becker\_2018    &  -110.13 &
        31.73&0 & 15&   NA&Sandy loam&  NA&
        NA& 26.40\\
               Becker\_2018    &  -110.09 &
        31.72&0 & 15&   NA&Sandy loam&  NA&
        NA& 27.84\\
               Becker\_2018    &  -110.09 &
        31.69&0 & 15&   NA&Sandy loam&  NA&
        NA& 21.60\\
               Becker\_2018    &  -110.05 &
        31.74&0 & 15&   NA& Loam&       NA&
 23.76\\
               Becker\_2018    &  -110.04 &
        31.72&0 & 15&   NA&Sandy loam&  NA&
        NA& 39.12\\
               Becker\_2018    &  -110.04 &
        31.69&0 & 15&   NA&Sand&        NA&
 102.96\\
             \end{tabular}
\label{tab:database_str}
\end{table}


\begin{table}[!hbt]
\begin{center}
\caption{Confidence weights provided to each
 sample based on location accuracy and
 method used: LM = laboratory method, FM =
 field method.}

              \begin{tabular}{r{2cm} r{1.5cm}
 r{2cm} r{1.5cm}}
               \hline
             \parbox{1.8cm}{\centering
 Location errors (LM)}  &
 \parbox{1.8cm}{\centering Confidence index}
  &      \parbox{1.8cm}{\centering Location
 errors (FM)} & \parbox{1.8cm}{\centering
 Confidence index} \\
     \hline
     0 -- 100~m &1 & 0 -- 100~m & 3\\
     100 -- 250~m &3 & 100 -- 250~m & 6\\
     250 -- 500~m &5 & 250 -- 500~m & 9\\
    0.5 -- 1~km &7 & 0.5 -- 1~km &  12\\
    1 -- 5~km &9 & 1 -- 5~km & 15\\
     5 -- 10~km &20 & 5 -- 10~km &  30\\
    >10~km &40 & >10~km & 40\\
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 234

```
        337.92& NA\\
               Saseendran\_2005 &  -103.15 &
        40.15&  90&120& 1.40&Loam&      12.0&
        284.88& NA\\
               Saseendran\_2005 &  -103.15 &
        40.15&  120&    150&    1.42&Loam&
        48.3&   259.20& NA\\
               Saseendran\_2005 &  -103.15 &
        40.15&  150&180&        1.42&Loam&
        48.3&   259.20& NA\\
               Becker\_2018    &  -110.13 &
        31.73&0 & 15&   NA&Sandy loam&  NA&
        NA& 26.40\\
               Becker\_2018    &  -110.09 &
        31.72&0 & 15&   NA&Sandy loam&  NA&
        NA& 27.84\\
               Becker\_2018    &  -110.09 &
        31.69&0 & 15&   NA&Sandy loam&  NA&
        NA& 21.60\\
               Becker\_2018    &  -110.05 &
        31.74&0 & 15&   NA& Loam&       NA&
 23.76\\
               Becker\_2018    &  -110.04 &
        31.72&0 & 15&   NA&Sandy loam&  NA&
        NA& 39.12\\
               Becker\_2018    &  -110.04 &
        31.69&0 & 15&   NA&Sand&        NA&
 102.96\\
             \end{tabular}
\label{tab:database_str}
\end{table}


\begin{table}[!hbt]
\begin{center}
\caption{Number of samples (N) assigned to
 each class of spatial accuracy. A minimum
 and maximum accuracy is defined for each
 class. NA are samples without information
 on spatial accuracy.}

              \begin{tabular}{r{2cm} r{2cm}
 r{1.5cm} }
               \hline
             {\centering Minimum location
 error} &{\centering Maximum location
 error}& { N}  \\


     \hline
     0~m  & 100~m &9937 \\
     100~m  & 250~m &1422 \\
     250~m  & 500~m &959 \\
    500~m  & 1000~m &516 \\
    1000~m  & 5000~m &163 \\
     5000~m  & 10000~m &128 \\
    10000~m  & NA &142\\
```

```
258
259  \hline
260  \end{tabular}
261  \label{Table:weights}
262  \end{center}
263  \end{table}
264
265  \begin{table}[hbt!]
266  \caption{Mean values of soil hydro-physical properties for each soil texture class. The number of samples (N) is given in parenthesis under each soil variable for each soil texture classes. $N$ values marked with $^*$ correspond to undefined soil texture classes. BD = bulk density (g/cm$^{3}$), OC = organic carbon (\%), FC = field capacity (\% vol), WP = wilting point (\% vol), Ksat$_{l}$,  Ksat$_{f}$ = laboratory and field Ksat (cm/day). For Ksat the geometric mean is reported (due to the sensitivity on few extreme values). For all other properties the arithmetic mean is provided.}
267  \addtolength{\tabcolsep}{1mm}
268                \begin{tabular}{m{2.5cm} c{2mm} c{2mm} c{2mm} c{2mm} c{2mm} c{2mm} r{2mm} cc}
269
270                      \hline
```

```
264       \hline
265         \textbf{Total} & & \textbf{13,267}\\
266       \hline
267
268  \end{tabular}
269  \label{Table:weights}
270  \end{center}
271  \end{table}
272
273  \begin{table} [htbp]
274      \centering
275      \caption{Instruments and methods used to estimate Ksat. A key reference with further details is given for all methods. In some cases, 'ponding' or 'permeameter' methods were listed in original studies without specification (18 samples in total).}
276      \small\addtolength{\tabcolsep}{0pt}
277      \begin{tabular}{lc|lc}
278      \hline
279       Lab Ksat methods
280   & $N$  & Field Ksat methods
281   & $N$  \\
282       \hline
283       Constant head method \citep{klute1986hydraulic}&   8014      & Mini-infiltrometer \citep{leeds1994device}
284  & 739\\
285  Falling head method \citep{klute1965laboratory}&  766      & Tension  infiltrometer \citep{reynolds2000comparison}
286  & 705\\
287  Triaxial cell (ASTM D 5084) \citep{purdy2006comparison}&  99 & ring infiltrometer \citep{bodhinayake2004determination} &       625\\
288  Cylinder method or soil core method \citep{reynolds2000comparison}&      27   Disc infiltrometer \citep{soracco2010anisotropy}
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 271

```

               {Texture Classes\par} &
 \parbox{0.5cm}{Clay\par (N)}   &
 \parbox{0.5cm}{Silt\par(N)}&
        \parbox{0.5cm}{Sand \par(N)}&
        \parbox{0.5cm}{BD \par (N)} &
        \parbox{0.5cm}{OC \par (N)} &
        \parbox{0.5cm}{FC \par (N)} &
 \parbox{0.5cm}{WP \par (N)} &
        \parbox{1.2cm}{\centering Ksat$_{l}$
 (N)}  &        \parbox{0.5cm}{Ksat$_{f}$
 \par(N)}\\

     \hline
      {Clay} & 56.3& 23.8 &    19.9 & 1.27 &
 1.98 & 45.0& 30.9& 8.17& 110.33\\
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 289

```
& 584\\
Hydraulic head \citep{robbins1977hydraulic}&
        15      & Single ring
 \citep{bagarello2004using}& 467\\
Pressure plate \citep{sharratt1990water}&
      & Guelph Permeameter
 \citep{reynolds1985situ}& 156\\
Oedometer test (UNI CEN ISO/TS 17892-5)
 \citep{terzaghi2004geotechnical}&     9
 BEST method \citep{bagarello2004using}&
 147\\
Oedometer test (ASTM D2435-96)
 \citep{sutejo2019hydraulic}&  12      &
 Aardvark permeameter
 \citep{hinton2016land}& 142\\
 &           & Guelf Infiltrometer
 \citep{gupta1993comparison}
& 87\\
&        & Piezometer slug test
 \citep{baird2017high}& 72\\
&        & Tensiometers
 \citep{nielsen1973spatial}& 70\\
&        & Rainfall simulator
 \citep{gupta1993comparison}& 55\\
&        & Hood infiltrometer
 \citep{schluter2020long}& 40\\
&        & Micro-infiltrometer
 \citep{sepehrnia2016extent}& 35\\
&        & Mini Disc infiltrometer
 \citep{naik2019estimating}& 32\\
&        & Disc permeameter
 \citep{mohanty1994comparison}& 27\\
&        & Constant head permeameter
 \citep{amoozegar1989compact}& 22\\
&        & Steady infiltration
 \citep{scotter1982measuring}& 16\\
&        & Permeameter& 10\\
&        & Ponding& 8\\
&        & Philip–Dunne permeameter
 \citep{munoz2002field} & 6\\
&        & Augur method
 \citep{mohsenipour2016estimation} & 5\\

Unknown& 206     & Unknown& 83\\

\hline
      \textbf{Total} &   \textbf{9162} & &
 \textbf{4133}\\
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 276

```
276
277        &       (835)& (835) &(835)     & (6
 (454)&(452)& (454)& (507)& (331)\\
278
279       Clay Loam &      31.4&   38.6&   30.0
        2.49&   39.7&   24.1&   12.25&  59.96
280
281        &       (543)&  (543)&  (543)&  (382)
 (360)&(76)&    (76)&   (139)&  (423)\\
282
283       Loam &   19.1&   39.3&   41.6&   1.28
        32.6&   14.0&   43.49&35.59\\
284
285       &        (699)&(699)     &(699)  &
        (102)&  (106)&  (206)&  (504)\\
286
287       Loamy Sand &      7.5&    8.5 &   84.0
 1.14&  17.5&   6.6&    96.49&  127.06\\
288
289        &       (742)& (742) & (742)    &
        (712)&(680)&    (558)&(592)&    (633)
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 313

```
313
314
315
316 \hline
317     \end{tabular}
318     \label{tab:Ksat_methods}
319 \end{table}
320
321
322
323 \subsection{Standardizationand quality
     assignment}
324 The database was cleaned to remove
     unrealistic low values. For example, In the
     SWIG database, Ksat values computed using
     infiltration time series were less than
     $10^{-14}$ m/day, which seems unreasonable,
     so they were not included in the database.
     All datasets were cross-checked to avoid
     redundancy. For example, UNSODA data
     consist of \citet{vereecken2017soil} and
     \citet{Richard1987Schweiz} datasets and
     SWIG database used
     \citet{zhao2018analysis}. Hence we removed
     these datasets from UNSODA and SWIG
     database and used the original sources.
     Moreover, in the SWIG database, soil depth
     information was not available, so we
     assumed that infiltration experiments were
     conducted close to the surface and assigne
     a depth of 0--20~cm.
325
326 To describe position accuracy of each
     dataset, we assigned each Ksat value to one
     of seven 'accuracy classes' ranging from
     highest (0 - 100 m) to lowest accuracy
     (more than 10000 m or non available
     information (NA)). For example,
     \citet{forrest1985survey},
     \citet{zhao2018analysis}and
     \citet{ottoni2018hydrophysical}provided
     detailed site coordinates, thus we assigned
     a location accuracy of 0-100 m (i.e.,
     highly accurate) (see
     Table~\ref{Table:weights}for more
     details). After data extraction from
     literature, geo-referencing and
     standardization, all information was
     collected in tabulated form in the new data
```

```
290
291          Sand &    2.2&    3.1&    94.7&   1.51&
              8.2&    2.5&    501.08& 252.31\\
292
293          &         (4526)& (4526)  & (4526)
     (4193)&(4077)& (4074)& (4218)& (320)\\
294
295          Sandy Clay &      39.3&   8.1&    52.6&
              0.23&   34.7&   23.4&    14.02& -----
296
297           &        (179)& (179) & (179)    &
              (143)&  (161)&  (161)&  (175)&(4)
298
299          Sandy Clay Loam &         26.3&   12.2&
              1.54&   1.25&   28.9&   17.3&19.28&
300
301         & (1149)&(1149)   &(1149) &        (941)
              (959)&  (806)&  (760)&  (869)&  (288)
```

```
     base SoilKsatDB \emph{'version 0.3'}
     (\url{https://doi.org/10.5281/zenodo.37527
    1}). The database consists of 22 columns
     (various sample properties) and 13,268 rows
     (a header and 13,267 samples). An excerpt
     of the database with some key properties is
     shown in Table~\ref{tab:database_str}.
327
328  \begin{table}[hbt!]

329 \caption{Mean values of soil hydro-physical
     properties for each soil textural class.
     The number of samples (N) is given in
     parenthesis under each soil variable for
     each soil texture classes. $N$ values
     marked with $^*$ correspond to undefined
     soil texture classes. BD = bulk density
     (g/cm$^{3}$), OC = organic carbon (\%), FC
     = field capacity (\% vol), WP = wilting
     point (\% vol), Ksat$_{l}$,  Ksat$_{f}$ =
     laboratory and field Ksat (cm/day). For
     Ksat the geometric mean is reported (due t
     the sensitivity on few extreme values). Fo
     all other properties the arithmetic mean i
     provided.}
330 \addtolength{\tabcolsep}{1mm}
331                \begin{tabular} {m{2.5cm}
     c{2mm} c{2mm} c{2mm} c{2mm} c{2mm} c{2mm}
     r{2mm} cc}
332
333                  \hline
334
335                {Texture Classes\par} &
     \parbox{0.5cm}{Clay\par(N)}   &
     \parbox{0.5cm}{Silt\par(N)}&
            \parbox{0.5cm}{Sand\par(N)}&
            \parbox{0.5cm}{BD\par (N)} &
            \parbox{0.5cm}{OC\par (N)} &
            \parbox{0.5cm}{FC\par (N)} &
     \parbox{0.5cm}{WP\par (N)} &
            \parbox{1.2cm}{\centeringKsat$_{l}$
     (N)}  &        \parbox{0.5cm}{Ksat$_{f}$
     \par(N)}\\
336
337      \hline

338         {Clay} & 56.3& 23.6 &   20.0 & 1.27 &
     2.00 & 43.2& 30.0& 8.22& 110.07\\
339
340           &        (830)& (830) &(830)     & (63
     (448)&(447)& (449)& (499)& (331)\\
341 Silty Clay &    45.2&   45.1&   9.6&    1.18
            49.9&   30.2&3.63&      196.65\\
342
343            &       (181)& (181)& (181)&    (175)
           (116)&  (46)&   (46)&(85)&      (96)\
344 Sandy Clay &    39.3&   8.1&    52.5&   1.52&
```

```
/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESS
D_v1.tex, Top line: 302

        Sandy Loam &      13.5&   16.7&   69.8&
         1.33&   24.2&   11.0&   34.53&  85.31

          &       (1610)& (1610)  &(1610) &
         (1352)& (815)&  (801)&  (999)&  (636)

        Silt &   7.5&    84.7 &  7.8 &   1.17&
         51.43&  7.5&    13.27 & ----    \\

         &       (25)& (25)       & (25)  &
         (11)&   (751)&  (25)&    \\

         Silt Loam &      15.2&   67.0&   17.8&
         3.65&   35.3&   15.6&   5.76&   43.64

          &       (813)& (813)    & (813) &
 (500)& (148)&  (138)&  (444)&  (383)\\

        Silty Clay &      45.5&   45.5&   10.0&
         3.83&   49.9&   30.2&1.22&      217.6

           &      (181)& (181)& (181)&    (175)
         (116)&  (46)&   (46)&(69)&      (112)

        Silty Clay loam &         33.1&   57.2&
         1.24&   2.67&   46.2&   23.9&   1.45&

           &      (333)& (333) & (333) &  (282)
         (226)&  (57)&   (56)&   (110)&  (232)


         \hline
        \textbf{Total} &  \textbf{11,635}\par
 &  \textbf{11,635}\par &
 \textbf{11,635}\par &  \textbf{10,464} &
 \textbf{9,555} & \textbf{7,340} &
 \textbf{7,275} & \textbf{8,394} &
 \textbf{3,333}\\
```

```
/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESS
D_v2.tex, Top line: 344
         34.7&   23.4&    14.16& -----\\

         &       (176)& (176) & (176)    &
         (140)&  (158)&  (158)&  (172)&(4)
        Clay Loam &      31.4&   38.6&   29.9
         2.49&   37.2&   22.1&   13.34&  60.56

          &       (544)&  (544)&  (544)&  (382)
 (360)&(76)&    (76)&   (127)&  (417)\\
Silty Clay loam &       33.1&   57.1&   9.7&
         2.67&   46.2&   23.9&   1.57&   48.45

          &       (335)& (335) & (335) &  (283)
         (227)&  (57)&   (56)&   (113)&  (222)
Sandy Clay Loam &        26.3&   12.1&   61.6&
         1.26&   28.7&   17.1&19.43&     14.23

      & (1148)&(1148)   &(1148) &        (966)
         (950)&  (805)&  (759)&  (876)&  (272)

        Silt &   7.7&    84.6 &  7.6 &   1.16&
         51.4&   7.5&    13.27 & ----    \\

        &       (25)& (25)       & (25)  &
         (12)&   (11)&   (25)&   \\
         Silt Loam &      15.2&   66.8&   17.9&
         3.65&   35.2&   15.6&   5.87&   44.63

         &       (810)& (810)    & (810) &
 (498)& (148)&  (138)&  (447)&  (364)\\
 Loam & 19.0&   39.1&   41.7&   1.29&   2.16&
         32.07&  14.2&   45.62&34.21\\

        &       (692)&(692)     &(692)  &
         (101)&  (104)&  (226)&  (466)\\
Sandy Loam &     13.5&   16.8&   69.7&   1.49&
         24.2&   11.0&   39.71&  74.57\\

         &       (1601)& (1601)  &(1601) &
         (1337)& (806)&  (792)&  (1078)& (523)
Loamy Sand &     7.3&    8.5 &   84.0 &  1.55&
         17.3&   6.5&    95.37&  132.33\\

          &       (736)& (736) & (736)    &
         (711)&(674)&    (582)&(586)&    (637)
Sand &  2.2&     3.1&    94.6&   1.51&   0.62&
         2.5&    488.46& 209.55\\

        &        (4513)& (4513)  & (4513)
 (4179)&(4063)& (4062)& (4409)& (106)\\
        \hline
       \textbf{Total} &  \textbf{11,591}\par
 &  \textbf{11,591}\par &
 \textbf{11,591}\par &  \textbf{10,494} &
 \textbf{9,501} & \textbf{7,301} &
 \textbf{7,236} & \textbf{8,694} &
 \textbf{2,900}\\
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 326

```
326        &  (32$^*$)  &(32$^*$)   &(32$^*$)  &
   (687$^*$) &  (232$^*$) & (49$^*$) &
   (143$^*$) & (413$^*$) &         (1,154$^*$)\
327          \hline
328
329              \end{tabular}
330              \label{tab:Average}
331
332   \end{table}
333
334   \subsection{Standardization and quality
       assignment}
335
336   The database was cleaned on the basis of
       highest and lowest values of saturated
       hydraulic conductivity. In SWIG database,
       some values of Ksat were less than
       $10^{-14}$ m/day, that seem unreasonable,
       so they were not included in the database.
       All datasets were cross-checked to avoid
       redundancy. For example, UNSODA data
       consist of \citet{vereecken2017soil} and
       \citet{Richard1987Schweiz} datasets and
       SWIG database used
       \citet{zhao2018analysis}. Hence we removed
       these datasets from UNSODA and SWIG
       database and used the original source
       datasets. Moreover, in the SWIG database,
       soil depth information was not available,
       so we assumed that data were obtained from
       field measurements and assumed it was
       obtained at a depth of 0--20~cm.
337
338   To describe the accuracy and reliability of
       each dataset, a quality flag (or confidence
       degree) was assigned to each data set based
       on (a) positional accuracy of the site, and
       (b) methodology used (i.e. only
       differentiating between field and
       laboratory measurements, not accounting for
       different laboratory and field methods) for
       measuring Ksat. Here, we separated each
       study based on the measurement of Ksat and
       subjectively selected a range from 1 to 50
       (i.e., 1 = highly accurate, 50 = least
       accurate) to describe the level of accuracy
       of each dataset. Table~\ref{Table:weights}
       shows the allocation of different weights
       for laboratory and field methods. Here, we
       assigned a slightly higher confidence to
       laboratory methods (compared to field ones)
       because the analyzed soil depth is well
       defined in lab samples but unclear in field
       infiltration measurements. In contrast,
       field methods are representative of larger
       areas. The other main difference is the
       entrance of atmospheric air into the soil.
       It is, in fact, more difficult in field
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 378

```
378        &  (17$^*$)  &    &(38$^*$)  &  (775$^*$)
    &  (286$^*$) & (88$^*$) & (182$^*$) &
    (468$^*$) &     (1,233$^*$)\\
379          \hline
380
381              \end{tabular}
382              \label{tab:Average}
383
384   \end{table}
385
386
387
388   \subsection{Statistical modeling of Ksat}
```

```
methods to reach a saturated state because
of the interference of atmospheric air and
fast infiltration velocities at beginning
of the process
\citep{faybishenko1997comparison}.In
addition, a higher confidence was assigned
to measurements with higher spatial
accuracy. For example, laboratory
measurements at high spatial accuracy were
given the highest confidence degree. Among
these, \citet{forrest1985survey}and/or
\citet{ottoni2018hydrophysical}measured
Ksat in the laboratory and provided
detailed site coordinates, thus we assigned
a confidence degree of 1 (i.e., highly
accurate).  \citet{zhao2018analysis}
measured Ksat using field methods and
provided the exact locations of the field
sites thus we assigned 3 as a confidence
degree. If the spatial accuracy was be
between 100--250~m, then we would have
assigned a value of 6 (see
Table~\ref{Table:weights}for more
details). After data extraction from
literature (and data bases),
geo-referencing and standardization, all
information was collected in tabulated form
in the new data base SoilKsatDB
(\url{https://doi.org/10.5281/zenodo.375272
1}). The database consists of a 38 columns
(various sample properties) and 13,268 rows
(for column titles and 13,267 samples). An
excerpt of the data base with some key
properties is shown in
Table~\ref{tab:database_str}.
```

```
To show a possible application of the
 database, we computed various pedotransfer
 functions (PTFs). The PTF models were
 fitted using a random forest (RF) machine
 learning algorithm
 \citep{breiman2001random} in the R
 environment for statistical computing
 \citep{Rbook}. We tested fitting the RF
 model for log-transformed ($log_{10}$) Ksat
 values as function of primary soil
 properties. For 15\% of samples with
 information on bulk density and soil
 texture, the value of organic content (OC)
 was not reported. Therefore, we expressed
 the PTF for Ksat as a function of bulk
 density, clay and sand content only. We
 derived two PTFs for Ksat:
\begin{enumerate}
\item \textit {PTFs for temperate regions}:
 the map of Ksat locations were overlaid on
 the Köppen-Geiger climate zone map
 \citep{rubel2010observed,hamel2017sediment}
 and then divided based on climatic regions
```

```latex

\subsection{Statistical modeling}

The PTF models were fitted using multivariate polynomial regression (MPR) and random forest (RF) in the R environment for statistical computing \citep{Rbook}. We tested fitting the MPR model for Ksat values as function of primary soil properties. For 15\% of samples with information on bulk density and soil texture, the value of organic content (OC) was not reported. Therefore, we expressed the PTF for Ksat as function of bulk density, clay and sand content (without OC). To test if PTFs for different climatic regions or measurement types are different, we have split fitting the PTF using (1) temperate-climate soil samples (including both laboratory and field measurements), and (2) laboratory based measured samples (including all climates). To develop PTFs with temperate climate soil samples, the dataset (total 13,267 points) was divided based on climatic regions (temperate, tropical, boreal, and arid) to account for differences in climate and related weathering processes \citep{hodnett2002marked}. A total of 8,333 temperate-climate soil samples were used that contain information on sand, clay, and bulk density. The data set was randomly divided into training (6,666 samples, 80\%) and testing dataset (1,667 samples, 20\%). Likewise, MPR was also applied to develop a PTF for laboratory measurements. In a second application, the dataset (total 13,267) was divided into laboratory and field based soil Ksat samples. The laboratory dataset (8,055 soil samples) was used for training (6,444) and testing (1,611) following the same method as used for the temperate climate PTF (i.e., 80\% for training and 20\% for testing).

The following equation was fitted using MPR:

```

```latex
 (temperate, tropical, boreal, and arid) to account for differences in climate and related weathering processes \citep{hodnett2002marked}.  A total of 8,296 temperate-climate based Ksat values that contain information on sand, clay, and bulk density were used to develop PTF. The data set was randomly divided into a training (6,637 samples, 80\%) and testing dataset (1,659 samples, 20\%).

\item \textit {PTFs from laboratory-based Ksat values}:
In a second application, the dataset (total 13,267) was divided into laboratory and field based Ksat values. The laboratory dataset (8,498 soil samples) was used for training (6,798) and testing (1,700) following the same method as used for the temperate climate PTF (i.e., 80\% for training and 20\% for testing).
 \end{enumerate}

 The \emph{'ranger'} package version 0.12.1 \citep{wright2015ranger} was implemented
```

```
%
\begin{equation}\label{Eq:Ksat_ptf}
 \begin{split}
    \log({ \mathrm{Ksat} }) = b_0 + b_1
 \cdot \mathrm{BD} + b_2 \cdot \mathrm{BD}^2
 + b_3 \cdot \mathrm{CL} + b_4 \cdot
 \mathrm{BD}\cdot \mathrm{CL} + b_5 \cdot
 \mathrm{CL}^2 + b_6 \cdot\mathrm{SA} +b_7
 \cdot \mathrm{BD}\cdot \mathrm{SA}+b_8
 \cdot \mathrm{CL}\cdot \mathrm{SA} + b_9
 \cdot \mathrm{SA}^2
    \end {split}
    \end{equation}
%
where Ksat is in cm/day, clay
 ($\mathrm{CL}$) and sand ($\mathrm{SA}$)
 are expressed in $\%$ and bulk density
 ($\mathrm{BD}$) is in g/cm$^3$.

Likewise, PTFs were also developed using a
 RF algorithm both for temperate-climate and
 laboratory based soil samples. The same
 soil variables (sand, clay and, bulk
 density) were fitted with Ksat values and
 we used the same number of points as for
 MPR for training and testing the models
 (i.e., 80\% and 20\%, respectively). The
 \emph{'ranger'} package
 \citep{wright2015ranger} was implemented
 to process the large data. The PTFs
 developed for temperate regions and for
 laboratory data were then applied to
 estimate Ksat in tropical climate (1,122
 samples) or field measurements (2,396
 samples), respectively. Root mean square
 error (RMSE) and concordance correlation
 coefficient (CCC)
```

```
 to process the large dataset. The PTFs
 developed for temperate regions and for
 laboratory data were then applied to test
 their applicability in tropical climate
 (1,111 samples) and for field measurements
 (1,998 samples), respectively. The code for
 generating and testing the PTFs is provided
 in the supplementary file.

 \subsection{Evaluation of Ksat PTFs}
 The relative importance of the covariates
 to determine the PTF was assessed by the
 increase in node purity. It is calculated
 using the Gini criterion from all the
 splits (in our case 3 splits) in the forest
 based on a particular variable
 \citep{rodrigues2014insight}.Furthermore,
 the accuracy of the predictions was
 evaluated using
 bias, root mean square error (RMSE, in
 log-transformed Ksat measurement) and
 concordance correlation coefficient (CCC)
 \citep{lawrence1989concordance}.


 Bias and RMSE are defined as:

\begin{equation}
bias = {\sum_{i=1}^{n}
 \frac{(y_{i}-\hat{y}_{i})}{n}}
\end{equation}

\begin{equation}
```

```
\citep{lawrence1989concordance}were computed to assess the accuracy of the models.

\section{Results}

\subsection{Data coverage}

Based on the intensive literature search and data collection, we have assembled a total of 13,267 values of Ksat from 1,910 sites across the globe. Figure~\ref{Fig:points_map}shows the global distribution of the sites locations used in this study. Most data originate from the USA, followed by Europe, Asia, South America, Africa, and Australia. The points are often spatially clustered with the biggest cluster of points (1,103 site locations with 6,532 Ksat values) in Florida \citep{Floridadatabase}.Ksat data include 4,460 values from field measurement and 8,807 values from laboratory measurements. In particular, different types of infiltrometers were used for Ksat field measurements, whereas constant or falling head methods were predominantly
```

```
RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_{i}-y_{i})^2}{n}}
\end{equation}

\noindent where $y$ and $\widehat{y}$ are observed and predicted Ksat values, respectively,  and n is the total number o cross-validation points.

In addition, Concordance Correlation Coefficient (CCC) (as measure of the agreement between observed and predicted Ksat values) of cross validation \citep{lawrence1989concordance} is defined as:

\begin{equation}\label{Eq:CCC}
        CCC  = \frac{2\cdot\rho\cdot\sigma_{\hat{y}}\cdot\sigma_{y}}{\sigma_{\hat{y}}^2 +\sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2}
\end{equation}

\noindet where $\mu_{\hat{y}}$ and $\mu_y$ are predicted and observed means, $\sigma_{\hat{y}}$ and $\sigma_y$ are are predicted and observed variances and $\rho is the Pearson correlation coefficient between predicted and observed values. CCC is equal to 1 for a perfect model.

\section{Results}

\subsection{Data coverage of SoilKsatDB}

Based on the literature search and data compilation, we have assembled a total of 13,267 values of Ksat from 1,910 sites(one site is equal to one location 'id') across the globe. Figure~\ref{Fig:points_map} shows the global distribution of the sites used in this study.  Most data originate from North America, followed by Europe, Asia, South America, Africa, and Australia. With respect to climatic regions, 10,093 Ksat values belong to the temperate region and 1,443, 1,113, 582, and 36 to tropical, arid, boreal, and polar regions, respectively.   The points are often spatially clustered with the biggest cluster of points (1,103 site locations with 6,532 Ksat values) in Florida \citep{Floridadatabase}.Ksat data include 4,133 values from field measurement and
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 359

```
 used in laboratory analyses.

Out of the 13,267 Ksat measurements, 11,667,
 11,151, 9,787, 7,389 and 7,418 points had
 information on soil texture, bulk density,
 organic carbon, field capacity and wilting
 point, respectively, and 8,947 samples had
 information for all soil basic properties
 (bulk density, soil texture and organic
 carbon) as shown in
 Figure~\ref{Fig:Venn_diagram}

\begin{figure}
    \centering
    \includegraphics[width=0.35\columnwidth]
 {Venn_diagram.jpg}
    \caption{Venn diagram illustrating the
 number of samples containing information on
 bulk density, soil texture, and organic
 carbon. Out of 13,267 samples, 11,151,
 11,667 and 9,787 samples have values of
 bulk density, soil texture and organic
 carbon, respectively. Furthermore, 10,452,
 9,150 and 9,570 samples have information of
 bulk density and soil texture, bulk density
 and organic carbon and soil texture and
 organic carbon, respectively. 8,947 samples
 have information of all three soil
 properties}
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 428

```
 9,162 values from laboratory measurements.
 In particular, different types of
 infiltrometers (e.g., Mini-infiltrometer,
 Tension infiltrometer, double ring
 infiltrometer) and permeaters (e.g., Guelf
 permeameter, Aardwark permeameter) were
 used for Ksat field measurements, whereas
 constant or falling head methods were
 predominantly used in laboratory analyses,
 as shown in Table~.\ref{tab:Ksat_methods}.

Out of the 13,267 Ksat measurements, 11,591,
 11,269, 9,787, 7,389 and 7,418 points had
 information on soil texture, bulk density,
 organic carbon, field capacity and wilting
 point, respectively, while 8,994 samples
 had information for all soil basic
 properties (bulk density, soil texture and
 organic carbon)
 (Figure~\ref{Fig:Venn_diagram}). The
 methods used to compute these soil
 properties (as much as we could extract
 from the literature and existing databases)
 were listed in the supplementary CSV file
 sol\textunderscore ksat.pnts\textunderscore
 metadata.csv available at \emph{'version
 0.3'}
 \url{https://doi.org/10.5281/zenodo.3752721
}. Note that in addition to 11,591 soil
 texture values, 75 samples have soil
 texture information with total
 (sand+silt+clay) less than 98\% or greater
 than 102\%. We did not use these values in
 the PTF development. Moreover, the database
 contains total of 13,295 Ksat values
 because few studies have reported both
 field and lab measurements for the same
 sampling point.

\begin{figure}
    \centering
    \includegraphics[width=0.5\columnwidth]
 {Venn_diagram.jpg}
    \caption{Venn diagram illustrating the
 number of samples containing information on
 bulk density, soil texture, and organic
 carbon. Out of 13,267 samples, 11,269,
 11,591 and 9,787 samples have values of
 bulk density, soil texture and organic
 carbon, respectively. Furthermore, 10,494,
 9,266 and 9,501 samples have information of
 bulk density and soil texture, bulk density
 and organic carbon and soil texture and
 organic carbon, respectively. 8,994 samples
 have information of all three soil
 properties. Note that the size of the
 intersecting areas does not represent the
 correct fractions (otherwise the
```

```
    \label{Fig:Venn_diagram}
\end{figure}

\begin{figure*}[!htb]

    \centering

 \includegraphics[width=0.7\textwidth]{trian
gle_box_plot.jpg}
    \caption{Distribution of collected Ksat
 values: (a) distribution of soil samples on
 the USDA soil texture triangle. The bulk of
 the samples were from Florida (cluster of
 sandy soil samples). The Ksat values covers
```

```
 intersection with 8,994 would be much
 bigger). }
    \label{Fig:Venn_diagram}
\end{figure}


\subsection{Statistical properties of
 SoilKsatDB}

The distribution of soil samples based on
 soil texture classes is shown on the USDA
 soil texture triangle in
 Figure~\ref{Fig:texture_triangle}a. The
 database covers all textural classes, with
 a high clustering in sandy soils due to the
 numerous samples from Florida
 \citep{Floridadatabase}. The violin
 distribution plot in
 Figure~\ref{Fig:texture_triangle}c shows
 the range of Ksat values for the different
 databases. Most of the datasets report Ksat
 values between $\approx$ $10^{-2}$ and
 $10^{2.5}$~cm/day, with a wider range of
 Ksat values observed in measurements from
 theses and reports (including studies with
 extreme values from sandy desert soils and
 low conductive clay soils) and from the
 SWIG database (databases 9 and 6 in
 Figure~\ref{Fig:texture_triangle}c,
 respectively). Likewise,
 Figure~\ref{Fig:texture_triangle}d shows
 the violin distribution of Ksat based on
 soil texture classes. Sand and loamy sand
 soils showed the highest arithmetic mean
 (i.e., 2.68 and 1.99, respectively), while
 the lowest mean values were found for silt
 and silty loam (i.e., 1.12 and 1.15,
 respectively). The significance between
 each soil texture class was also tested
 using a t-test \citep{kim2015t} and results
 are presented in the supplementary file.
 Table ST1 shows that the Ksat values under
 sand and loamy sand soil texture class are
 significantly different from all other soil
 texture classes, however, silt, silty clay,
 and silty clay loam class are not
 significantly different from clay, sandy
 clay, and sandy clay loam Ksat values.
\begin{figure*} []
    \centering

 \includegraphics[width=0.7\textwidth]{trian
gle_box_plot1.jpg}
    \caption{Characterization of collected
 Ksat values. (a) Distribution of soil
 samples on the USDA soil texture triangle.
 The data points cover all soil textural
 classes and only few samples belong to the
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 373

```
all soil textural classes and only few
samples belong to the silt textural class.
The histogram plot (b) represents the range
of Ksat values spanned by each data source.
The dot represents the mean value, and the
line represents the standard deviation for
each data set. The numbers 1--9 refer to
different sources and databases: 1 =
Australia \citep{forrest1985survey},2 =
Belgium \citep{vereecken2017soil}, 3 =
China \citep{tian2017variability,
li2017multiscale}, 4 = extracted from
thesis and reports (see
Table~\ref{tab:my_label}),5 = Florida
\citep{Floridadatabase},6 = HYBRAS
\citep{ottoni2018hydrophysical},7 = SWIG
\citep{rahmati2018development},8 = Tibetan
Plateau \citep{zhao2018analysis},9 =
UNSODA \citep{nemes2001description}.}
    \label{Fig:texture_triangle}
\end{figure*}

\subsection{Statistical properties}
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 446

```
silt textural class. b) Distribution of
Ksat values using broad soil texture
classes (sandy soils: sand, loamy sand;
loamy soils: sandy loam, loam, silt loam,
silt, clay loam, sandy clay loam; clayey
soils: sandy clay, silty clay, clay) based
on laboratory and field methods. The number
of samples provided on the top of the
figure. The increase in Ksat values in
clayey and loamy soils under field methods
is likely due to the effect of soil
structure. A t-test showed that all broad
soil texture classes are significantly
different from each other except clayey
soils field Ksat values and sandy soils
field Ksat values (see Table ST2). The
violin plot (c) represents the range of
Ksat values spanned by each data source.
The dot represents the mean value, and the
line represents the standard deviation for
each data set. The numbers 1--9 refer to
different sources and databases: 1 =
Australia \citep{forrest1985survey},2 =
Belgium \citep{vereecken2017soil},3 =
China \citep{tian2017variability,
li2017multiscale}, 4 = Florida
\citep{Floridadatabase},5 = HYBRAS
\citep{ottoni2018hydrophysical},6 = SWIG
\citep{rahmati2018development},7 = Tibetan
Plateau \citep{zhao2018analysis},8 =
UNSODA \citep{nemes2001description},9 =
all other databases in
Table~\ref{tab:my_label}.d) Distribution
of Ksat based on soil textural classes with
the number of samples shown on the top of
the figure. The significance was also
tested for each class using a t-test
\citep{kim2015t} and results are presented
in the supplementary file. }
    \label{Fig:texture_triangle}
\end{figure*}

Average values of Ksat and other
hydro-physical properties are shown in
Table~\ref{tab:Average}.Higher average
organic carbon and bulk density values were
observed in clayey and loamy soils compared
to sandy soils. Ksat values obtained from
field measurements were on average higher
(depending on the type of instrument used)
than those obtained from laboratory Ksat
values. Particularly, for the clay texture
class much lower Ksat values were observed
for laboratory (mean Ksat $\approx$
8~cm/day) compared to field (mean Ksat
$\approx$ 110~cm/day) measurements
(Table~\ref{tab:Average}).
Figure~\ref{Fig:texture_triangle}bfurther
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 377

```
The distribution of soil samples based on soil texture classes is shown on the USDA soil texture triangle in Figure~\ref{Fig:texture_triangle}a. The database covers all textural classes, with a high clustering in sandy soils due to the numerous samples from Florida. The violin distribution plot in Figure~\ref{Fig:texture_triangle} shows the range of Ksat values for the different databases. Most of the datasets showed Ksat values between $\approx$ $10^{-2}$ and $10^{2.5}$~cm/day, with a wider range of Ksat values observed in measurements from theses and reports (including studies with extreme values from sandy desert soils and low conductive clay soils) and from the SWIG database (databases 4 and 7 in Figure~\ref{Fig:texture_triangle}b, respectively).

Average values of Ksat and other hydro-physical properties are shown in Table~\ref{tab:Average}. Higher average organic carbon and bulk density values were observed in clayey and loamy soils compared to sandy soils. Ksat values obtained from field measurements were on average higher (depending on the type of instrument used) than those obtained from laboratory samples. Particularly, for the clay texture class much lower Ksat values were observed for laboratory (mean Ksat $\approx$ 8~cm/day) compared to field (mean Ksat $\approx$ 110~cm/day) measurements.

\begin{figure*}
    \centering
    \includegraphics[width=0.6\textwidth]{Partialplots1.jpg}
    \caption{Partial correlation between Ksat and a) organic carbon (\%), b) bulk density (g/cm$^3$), c) clay (\%) and d) sand (\%). }
    \label{Fig:Partial_plots}
\end{figure*}

\subsection{PTFs derivation}
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 450

```
illustrates the higher range of Ksat values obtained for finer texture soils (clay and loam) compared to coarser soils (sand).

\begin{figure*}
    \centering
    \includegraphics[width=0.6\textwidth]{Partialplots1.jpg}
    \caption{Partial correlation between Ksat and a) organic carbon (\%), b) bulk density (g/cm$^3$), c) clay (\%) and d) sand content (\%). Ksat decreases with increasing clay content and bulk density, and increases with sand content. The color of each hexagonal cell shows the number of the counts in each cell.  }
    \label{Fig:Partial_plots}
\end{figure*}

\subsection{Ksat PTFs derivation}

As a test application of SoilKsatDB, two
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 391

```
As a test application of SoilKsatDB, PTFs
 were derived for temperate climate region
 and laboratory based samples using basic
 soil properties as covariates. Such basic
 soil properties (i.e., clay and sand
 fraction, organic carbon, and bulk density
 are plotted against Ksat in
 Figure~\ref{Fig:Partial_plots}, showing
 that Ksat decreases with increasing clay
 content and bulk density, and increases
 with sand content. The observed correlatio
 between these soil properties and Ksat
 motivates their use as key variables for
 the estimation of PTFs. Due to limiting
 data availability (15\% of samples without
 OC information) and the poor correlation
 between OC and Ksat
 (Figure~\ref{Fig:Partial_plots}), we built
 the PTF for Ksat using bulk density, clay
 and sand content (without OC).

Coefficients of Eq.~(\ref{Eq:Ksat_ptf}) wer
 fitted to values obtained from i) temperat
 sites and from ii) laboratory measurements
 The fitted model coefficients are listed i
 Table~\ref{Table:coefficents}. The fitting
 procedure provided $R^{2}$ of 0.47 and 0.5
 for temperate and laboratory values,
 respectively. Validation of the fitted
 equations against the testing data set
 provided CCC and RMSE for the temperate an
 laboratory based predictions equal to 0.64
 (CCC, temperate), 0.71 (RMSE, temperate)
 and 0.70 (CCC, lab), and 0.67 (RMSE, lab),
 respectively.

\begin{table}
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 461

```
PTFs were derived for Ksat (i.e., for
temperate regions and based on laboratory
measurements) using basic soil properties
as covariates. Such basic soil properties
are plotted against Ksat in
Figure~\ref{Fig:Partial_plots}, showing
that Ksat decreases with increasing clay
content and bulk density, and increases
with sand content. The observed correlatio
between these soil properties and Ksat
motivates their use as key variables for
the estimation of PTFs. In this
application, PTFs for Ksat were built on
bulk density and sand and clay content.
Organic carbon (OC) was not used to build
the PTFs because (i) this information was
missing for 15\% of samples and (ii) the
correlation between OC and Ksat was poor
(i.e. 0.005).
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 397

```
397  \caption{Pedotransfer function
      (Eq.~(\ref{Eq:Ksat_ptf})) coefficients
      obtained for  temperate and laboratory soil
      measurements.}
398  \begin{tabular}{S[table-format=-2.2]S[table-
     format=-2.2]S[table-format=-2.2]}%
399  \toprule
400  { Coefficient} & {Value (temp.)} & {Value
      (lab.)}  \\
401   \hline
402  $b_{0}$  & 2.17&1.44\\
403   $b_{1}$ & 0.9387&2.053\\
404   $b_{2}$ & -0.8026&  -1.256\\
405   $b_{3}$ & 0.0037& -0.0533\\
406   $b_{4}$ & -0.017&  -0.000051\\
407   $b_{5}$ & 0.000015&0.00055\\
408   $b_{6}$ & 0.0025&0.0079\\
409   $b_{7}$ & 0.00086 & -0.00080\\
410   $b_{8}$ & -0.00025& 0.000043\\
411   $b_{9}$ & 0.000073 & 0.000052\\
412  \hline
413  \end{tabular}
414   \label{Table:coefficents}
415  \end{table}
416
417   \begin{figure*}[!hbt]
418      \centering
419      \includegraphics[width = 07
      \textwidth]{MLR_RF_Temp_trop1.jpg}
420      \caption{Correlation between observed
      and predicted Ksat values obtained from (a,
      b) multivariate polynomial regression (MPR)
      and (c, d) random forest (RF) models.
      Models were obtained by fitting 6,666
      temperate-climate training points and
      tested on temperate (1,667 samples, panels
      a, c) and tropical testing points (1,122
      samples, panels b, d). The density of point
      pairs for Ksat is shown in logarithmic
      scale. CCC is the concordance correlation
      coefficient. PTFs showed reasonable
      agreement for both MPR (CCC = 0.64) and RF
      (CCC = 0.69) algorithms with temperate soil
      samples, while lower CCC values were
      obtained for tropical soil samples (0.53
      and 0.51 for MPR and RF, respectively).
      PTFs determined for temperate regions
      cannot be easily transferred to tropical
      regions due to different soil forming
      processes.}
421      \label{Fig:Temperate_Tropical}
422  \end{figure*}
423
424  \begin{figure*}[!hbt]
425      \centering
426      \includegraphics[width = 07
      \textwidth]{MLR_RF_lab_filed.jpg}
427      \caption{The correlation between
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 466

```
466
467  \begin{figure*}[!hbt]
468      \centering
469      \includegraphics[width = 09
      \textwidth]{RF_lab_field1.jpg}
470      \caption{The correlation between
      observed and predicted Ksat values obtaine
      from (a, b) random forest (RF) models. The
      RF-based Pedotransfer function (PTF) model
      was fitted using data for laboratory
      measurements of Ksat and tested on both
      laboratory (a) and field (b) measurements.
      Results showed reasonable agreement (CCC =
      0.73) using RF algorithms for laboratory
      measurements, but low CCC (0.10) for field
      measurements. PTFs developed based on
      laboratory measurements do not provide
      accurate estimates of Ksat measured in the
      field.}
471      \label{Fig:lab_field}
472  \end{figure*}
473
474  \begin{figure*}[!hbt]
475      \centering
476      \includegraphics[width = 09
      \textwidth]{MLR_RF_Temp_trop2.jpg}
477      \caption{Correlation between observed
```

observed and predicted Ksat values obtained from (a, b) multivariate polynomial regression (MPR) and (c, d) random forest (RF) models. The model was fitted using laboratory measurements and tested on both laboratory (a, c) and field (b, d) measurements. Results showed reasonable agreement (CCC = 0.70, CCC = 0.73) using both algorithms (RF and MPR) for laboratory measurements, but low CCC (0.16, 0.13) for field measurements. PTFs developed based on laboratory measurements do not provide accurate estimates of Ksat measured in the field.}

    \label{Fig:lab_field}
\end{figure*}

Results obtained from RF modeling using the same number of data points and the same independent variables (sand, clay, and bulk density) show a better accuracy. Specifically, the RF model performance based on CCC and RMSE was 0.69 (CCC, temperate region) and 0.70 (RMSE, temperate region), 0.73 (CCC, lab measurements), and 0.66 (RMSE, lab measurements), respectively.

Figure~\ref{Fig:Temperate_Tropical}(b and d) and Figure~\ref{Fig:lab_field}(b and d) indicates that both models underestimated Ksat for both tropical and field measured soil samples. In fact, for the RF model we obtained CCC and RMSE values equal to 0.51 and 0.90 for tropical and 0.13 and 1.1 for field measured samples, whereas CCC and RMSE values obtained from MPR were equal to 0.53 and 0.83, and 0.16 and 1.0 for tropical and field measurements, respectively.

\section{Discussion}

\subsection{Laboratory vs field estimated Ksat: effect of soil structure}

Results showed that Ksat values were, on average, higher for samples measured using field methods compared to laboratory

and predicted Ksat values obtained from random forest (RF) model. The RF-based Pedotransfer function (PTF) model was obtained by fitting 6,637 training points obtained in a temperate-climate and tested on (a) temperate (1,659 samples) and (b) tropical testing points (1,111 samples). CCC is the concordance correlation coefficient. PTFs showed good performance (CCC = 0.70) for the temperate soil samples (including both laboratory and field measurements), but lower CCC values were obtained for tropical soil samples (0.52 for RF). PTFs determined for temperate regions cannot be easily transferred to tropical regions due to different soil forming processes.}

    \label{Fig:Temperate_Tropical}
\end{figure*}

Figure~S1 shows the list of relative importance of the covariates the PTFs models obtained for temperate regions and laboratory-based measurements. Clay content was found to be the most important variable followed by sand and bulk density for temperate climate PTF. On the other hand, sand content was found to be the most important variable followed by clay and bulk density for the laboratory-based Ksat PTF. CCC, bias, and RMSE were respectively equal to 0.70, -0.002, and 0.69, for the temperate region based PTF, and to 0.73, 0.0004, and 0.65 for laboratory-based PTF.

PTF models derived for temperate and laboratory-based Ksat values overestimate Ksat for tropical and field-based Ksat values, respectively (see Figure~\ref{Fig:Temperate_Tropical}b and Figure~\ref{Fig:lab_field}b). CCC, bias, and RMSE values were respectively equal to 0.52, 0.2, and 0.90 for tropical Ksat values, and to

0.10, 0.21, and 1.2 for field measured Ksat values.

\section{Discussion}

\subsection{Laboratory vs field estimated Ksat: effect of soil structure}

The Ksat values were, on average, higher for samples measured using field methods compared to laboratory methods for most

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 438

```
methods for most  soil texture classes
(Table~\ref{tab:Average}).
Figure~\ref{Fig:boxplot_lab_field} further
illustrates the higher range of Ksat values
obtained for finer texture soils (clay and
loam) compared to coarser soils (sand). The
difference in laboratory and field based
Ksat values and higher range of Ksat values
in fine textured soil is probably related
to the effect of biologically-induced soil
structure that might be neglected in
laboratory measurements. In other words,
variability in the Ksat values depends on
the consideration of soil macropores by the
measurement methods. Soil macropores change
the pore size distribution and subsequently
affect Ksat values
\citep{tuller2002unsaturated}.Such an
effect is likely to be neglected more in
laboratory measurements compared to field
ones. \citet{mohanty1994comparison},for
example, compared the three field methods
and one laboratory method and found that
the sample size affects the measurement of
Ksat and maximum variability observed in
the Ksat values at shallow depth might be
due to the presence and absence of
open-ended pores. Likewise,
\citet{braud2017mapping}used three field
methods for Ksat measurements and found
significant variation between these methods
of measurements.


As shown in Figure~\ref{Fig:lab_field} Ksat
values measured in the field were
underestimated by PTFs derived from
laboratory measurements. The omission of
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 489

```
soil texture classes
(Table~\ref{tab:Average}and
Figures~\ref{Fig:texture_triangle}b and 5).
The difference in laboratory and field
based Ksat values and higher range of Ksat
values in fine textured soil is probably
related to the effect of
biologically-induced soil structure that
might be neglected in laboratory
measurements. The omission of soil
structures in many laboratory samples
limits the possibility to properly
reproduce field observations that are
likely to be more affected by the presence
of biopores \citep{Fatichi2020soil}. In
other words, variability in the Ksat values
depends on the consideration (and
existence) of soil structural pores by the
measurement methods. Soil structural pores
change the pore size distribution and
subsequently affect Ksat values
\citep{tuller2002unsaturated}.Such an
effect is more likely to be neglected more
in laboratory measurements compared to
field studies. Presence or absence of large
structural pores also depends on the scale
of measurements (that is usually larger in
the field). \citet{mohanty1994comparison},
for example, compared three field methods
and one laboratory method and found that
the sample size affects the measurement of
Ksat due to the presence and absence of
open-ended pores. Similarly,
\citet{ghanbarian2017accuracy}showed that
the sample dimensions (e.g., internal
diameter and height) also impact Ksat. The
authors further developed a sample
dimension-dependent PTF and showed a better
performance compared to other available
PTFs in the literature.
Likewise, \citet{braud2017mapping} used
three field methods for Ksat measurements
and found significant variation between
these methods of measurements.
\citet{davis1996influence} presents the
necessity to choose the most appropriate
scale of measurement for a particular soil
when undertaking conductivity measurements.
The authors tested small cores (73 mm wide
and 63 mm high) and large cores (22 mm wid
and 300 mm high) using the constant head
method in the laboratory and found the
difference of 1 to 3 orders of magnitude.
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 440

```
     soil structures in many laboratory samples
     limits the possibility to properly
     reproduce field observations that are
     likely to be more affected by the presence
     of biopores \citep{Fatichi2020soil}.
441
442 \begin{figure}[!hbt]
443     \centering
444     \includegraphics[width = 0.5
     \columnwidth]{Boxplot_ksat1.jpg}
445     \caption{The distribution of Ksat values
     based on laboratory and field methods.
     Field measurements gave higher values than
     laboratory ones in clayey and loamy soils
     likely due to the effect of structure.}
446     \label{Fig:boxplot_lab_field}
447 \end{figure}
448
449 \subsection{Temperate vs tropical soils:
     effect of clay mineralogy}
450
451 Results showed that PTFs obtained for
     temperate soils performed poorly for
     tropical soils
     (Figure~\ref{Fig:Temperate_Tropical}), with
     Ksat being underestimated by the
     temperate-based PTFs. This result is in
     agreement with
     \citet{tomasella2000pedotransfer} who
     derived PTFs using data from tropical
     Brazilian soils, which  did not properly
     capture observations in  temperate soils.
     We argue that the significant differences
     in the models fitted for tropical and
     temperate soils are due to the differences
     in the soil-forming processes defining the
     clay type and mineralogy. In fact, Oxisols
     (highly weathered clay minerals in tropical
     regions) are turned into inactive
     (non-swelling) clay minerals as a result of
     high rainfall and temperatures. On the
     other hand, in the temperate regions,
     active (smectite) and moderately active
     clay minerals (illite) are the dominant
     clay minerals. These swelling clay minerals
     retain the water within internal structures
     with very low hydraulic conductivity.
     Therefore, such a difference in clay
     mineralogy is likely responsible for the
     underestimation of Ksat in tropical soils
     from PTFs obtained in temperate ones.
452
453 \subsection{Limitations of SoilKsatDB}
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 492

```
492
493
494 \subsection{Temperate vs tropical soils:
     effect of clay mineralogy}
495
496 Results showed that PTFs obtained for
     temperate soils performed poorly for
     tropical soils
     (Figure~\ref{Fig:Temperate_Tropical}), with
     Ksat being underestimated by the
     temperate-based PTFs. This result is in
     agreement with
     \citet{tomasella2000pedotransfer} who
     derived PTFs using data from tropical
     Brazilian soils, which  did not properly
     capture observations in  temperate soils.
     We argue that the significant differences
     in the models validated for tropical and
     temperate soils are due to the differences
     in the soil-forming processes defining the
     clay type and mineralogy. In fact, Oxisols
     (highly weathered clay minerals in tropical
     regions) are turned into inactive
     (non-swelling) clay minerals as a result of
     high rainfall and temperatures. On the
     other hand, in the temperate regions,
     active (smectite) and moderately active
     clay minerals (illite) are the dominant
     clay minerals. These swelling clay minerals
     retain the water within internal structures
     with very low hydraulic conductivity.
     Therefore, such a difference in clay
     mineralogy is likely responsible for the
     underestimation of Ksat in tropical soils
     from PTFs obtained in temperate ones.In
     addition, soil structure formation
     processes may be different in tropical and
     temperate regions and intensify the
     differences between Ksat values measured in
     the two different climatic regions.
497
498 \subsection{Limitations of SoilKsatDB}
```

We have put an effort to collect laboratory and field data from all parts of the globe. However, we acknowledge that there are still gaps in some regions such as Russia and higher northern latitudes in general, which may produce uncertainties in Ksat estimations in such regions. The SoilKsatDB could also be of limited use for fine-resolution applications because many data points were characterized by limited spatial accuracy and missing soil depth information. Specifically, the spatial accuracy of many points is between tens of meters to several kilometers (see the methodology sections regarding the extraction of the spatial locations using Google Earth). In addition, in the SWIG database the soil depth and measurement method information were not provided, and often one location was used to represent an entire watershed. We tried to revisit each publication and extract the most accurate coordinates of assumed sampling locations and we assumed that most of the samples belonged to the field measurements as authors used different infiltrometers to compute Ksat. Hence, there might be few points in our SoilKsatDB that belong to laboratory measurements and that we have incorrectly assigned to field measurements.

For each measurement, a confidence index (1 = highest, 50 = lowest) was assigned based on the sampling location accuracy and measurement technique (laboratory or field), which can be used as a weight or probability argument in Machine Learning. We acknowledge that this was a rather subjective decision and a more objective way to assign weights would be to use the actual measurement and spatial positioning errors. Because these were not available for most of the datasets, we have opted for the definition of a confidence index estimated from the available documentation.

\subsection{Further developments}

We envisage several further developments of this database. The advancement in remote sensing technology opens the doors to link the hydraulic properties with global

We have put an effort to combine laboratory and field data from most global regions.. However, we acknowledge that there are still gaps in some regions such as Russia and higher northern latitudes in general, which may produce uncertainties in Ksat estimations in such regions. The SoilKsatDB could also be of limited use for fine-resolution applications because many data points were characterized by limited spatial accuracy and missing soil depth information. Specifically, the spatial accuracy of many points is between tens of meters to several kilometers (see the methodology sections regarding the extraction of the spatial locations using Google Earth). Many of the records in the SoilKsatDB come from legacy scientific reports and the original authors can not be traced and contacted, hence we advise to use this data with caution. In addition, in the SWIG database, the soil depth and measurement method information were not provided, and often one location was used to represent an entire watershed. We tried to revisit each publication and extract the most accurate coordinates of assumed sampling locations. In addition, we assumed that most of the samples were obtained from field measurements as authors used different infiltrometers to compute Ksat, so there might be few points in our SoilKsatDB that belong to laboratory measurements and that we have incorrectly assigned to field measurements.

For each measurement, a location accuracy (0-100 m = highly accurate, >10000 m = least accurate) was assigned based on the sampling location accuracy. The location accuracy can be used as a weight or probability argument in Machine Learning for Ksat mapping. We acknowledge that this was a rather subjective decision and a more objective way to assign weights would be to use the actual spatial positioning errors. Because these were not available for most of the datasets, we have opted for the definition of a location accuracy estimated from the available documentation.

\subsection{Further developments}

The advancement in remote sensing technology opens the doors to link the hydraulic properties with global environmental features. Using satellite-based maps of

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 461

```
 environmental features. Using
 satellite-based maps of environmental
 properties enables to incorporate local
 information on vegetation, climate, and
 topography for specific areas, which are
 often ignored by basic PTFs. For example,
 \citet{sharma2006including} developed PTFs
 using environmental variables such as
 topography and vegetation and concluded
 that these attributes, at finer spatial
 scales, were useful to capture the observec
 variations within the soil mapping units.
 Likewise, \citet{szabo2019mapping} used the
 random forest machine learning algorithm
 for mapping soil hydraulic properties and
 incorporated local environmental variable
 information.
\section{Data availability}
All collected data and related soil
 characteristics are provided online for
 reference and are available at
 \url{https://doi.org/10.5281/zenodo.3752721
} \citep{surya_gupta_2020_3752722}.

\section{Summary and conclusions}
We prepared a comprehensive global
 compilation of measured Ksat training point
 data ($N=13,267$) by importing, quality
 controlling, and standardizing tabular data
 from existing soil profile databases and
 legacy reports.

The produced SoilKsatDB covers a broad range
 of soil types and climatic regions and
 hence is applicable for global soil
 modeling. A higher variation in Ksat values
 was observed in fine-textured soil comparec
 to coarse-textured soils, possibly
 indicating the effect of soil structure on
 Ksat. Moreover, Ksat values obtained from
 field measurements were generally higher
 than those from laboratory measurements,
 likely due to impact of macropores at
 larger scale in field measurements.
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 506

```
 environmental properties, local information
 on vegetation, climate, and topography for
 specific areas, which are often ignored by
 basic PTFs, can be incorporated. For
 example, \citet{sharma2006including}
 developed PTFs using environmental
 variables such as topography and vegetatior
 and concluded that these attributes, at
 finer spatial scales, were useful to
 capture the observed variations within the
 soil mapping units. Likewise,
 \citet{szabo2019mapping} used the random
 forest machine learning algorithm for
 mapping soil hydraulic properties and
 incorporated local environmental variable
 information.
\section{Data availability}
All collected data and related soil
 characteristics are provided online for
 reference and are available at
 \emph{'version 0.3'}
 \url{https://doi.org/10.5281/zenodo.3752721
} \citep{surya_gupta_2020_3752722}.

\section{Summary and conclusions}
We prepared a comprehensive global
 compilation of measured Ksat training point
 data ($N=13,267$) by importing, quality
 controlling, and standardizing tabular data
 from existing soil profile databases and
 legacy reports. The produced SoilKsatDB
 covers a broad range of soil types and
 climatic regions and hence is useful for
 global soil modeling. A higher variation in
 Ksat values was observed in fine-textured
 soil compared to coarse-textured soils,
 indicating the effect of soil structure on
 Ksat. Moreover, Ksat values obtained from
 field measurements were generally higher
 than those from laboratory measurements,
 likely due to impact of soil structural
 pores at larger scale in field
 measurements.
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 470

```
470  The new database was applied to develop
      pedotransfer functions (PTFs) for Ksat
      using temperate and laboratory based soil
      samples using both MPR and RF algorithms.
      Both algorithms provided reasonable
      accuracy. However, PTFs developed for a
      certain climatic region (temperate) or
      measurement method (laboratory) could not
      be satisfactorily applied to estimate Ksat
      for other regions (tropical) or measurement
      method (field) due to the role of different
      soil forming processes (inactive clay
      minerals in tropical soils and impact of
      biopores in field measurements).
471
472  There are still some gaps in the
      geographical representation of sampling
      points, especially in Russia and the higher
      northern latitudes, that could induce
      uncertainty in global modeling. Therefore,
      the data set can be further improved by
      covering the missing areas and achieve
      better accuracy in the hydrological
      applications.
473
474  The SoilKsatDB was developed in R software
      and is available via
      \url{https://www.openml.org/d/42332}and
      \url{https://doi.org/10.5281/zenodo.3752721
     }. We have made code and data publicly
      available to enable further developments
      and improvements as a collective effort.
475
476  % \subsection
      %% Appendix A1, A2, etc.
477  \begin{acknowledgements}
478  The SoilKsatDB is a compilation of numerous
      existing datasets from which the most
      significant: SWIG dataset
      \citep{rahmati2018development}, UNSODA
      \citep{leij1996unsoda,nemes2001description}
     , and HYBRAS
      \citep{ottoni2018hydrophysical}. The study
      was supported by ETH Zurich (Grant ETH-18
      18-1). OpenGeoHub maintains an global
      repository of Earth System Science datasets
      at www.openlandmap.org. We thank Zhongwang
      Wei for helping in collecting the datasets
      and for insightful discussions. Wewould
      also want to thank Samuel Bickel (ETH
      Zurich) for boosting the leading author's
      confidence in High Performance Computing.
479  \end{acknowledgements}
480
481  \bibliography{soil_physics.bib}
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 513

```
513  The new database was applied to develop
      pedotransfer functions (PTFs) for Ksat
      using measurements in temperate  climates
      and laboratory based soil samples using RF
      algorithms. PTFs developed for a certain
      climatic region (temperate) or measurement
      method (laboratory) could not be
      satisfactorily applied to estimate Ksat for
      other regions (tropical) or measurement
      method (field) due to the role of different
      soil forming processes (inactive clay
      minerals in tropical soils and impact of
      biopores in field measurements).
514
515  There are still some gaps in the
      geographical representation of sampling
      points, especially in Russia and the higher
      northern latitudes, that could induce
      uncertainty in global modeling. Therefore,
      the data set can be further improved by
      covering the missing areas and achieve
      better accuracy in the hydrological
      applications.
516
517  The SoilKsatDB was developed in R software
      and is available via \emph{'version 0.3'}
      \url{https://doi.org/10.5281/zenodo.3752721
     }. We have made code and data publicly
      available to enable further developments
      and improvements as a collective effort.
518
519  % \subsection
      %% Appendix A1, A2, etc.
520  \begin{acknowledgements}
521  The SoilKsatDB is a compilation of numerous
      existing datasets from which the most
      significant are: SWIG dataset
      \citep{rahmati2018development}, UNSODA
      \citep{leij1996unsoda,nemes2001description}
     , and HYBRAS
      \citep{ottoni2018hydrophysical}. The study
      was supported by ETH Zurich (Grant ETH-18
      18-1). OpenGeoHub maintains an global
      repository of Earth System Science datasets
      at www.openlandmap.org. We thank Zhongwang
      Wei for helping in collecting the datasets
      and for insightful discussions. We
      acknowledge Samuel Bickel (ETH Zurich) for
      the help with High Performance Computing.
      We would also like to thank two anonymous
      reviewers and Dr. Attila Nemes for their
      constructive feedback to improve the
      manuscript.
522  \end{acknowledgements}
523
524  \bibliography{soil_physics.bib}
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v1.tex, Top line: 482

```

\end{pagewiselinenumbers}

\end{document}
```

/home/tomislav/Downloads/ksat_extra/Gupta_2019_ESSD_v2.tex, Top line: 525

```

\end{pagewiselinenumbers}

\end{document}
```