First of all, we would like to thank the reviewer for his help improving the paper. Reviewer' comments are in italic, our answers are in bold.

Referee comment #3

CONTENT

The authors performed laboratory experiments on soil cores of 1 liter volume, packed with a clay soil and with stone or glass beads inclusions, to determine saturated hydraulic conductivity by constant-head method and unsaturated conductivities by evaporation experiments. A series of numerical 2D simulations were additionally performed to study the effects of volume fraction, shapes, and sizes of stone inclusions on effective hydraulic conductivity. For saturated conductivity, some existing simple predictive models exist which are used for comparison.

The main message of the paper is that for saturated conductivity, increasing fractions of inclusions into a fine textured matrix can lead in practice to an increase of conductivity, which is contrary to predicted effects, and is hypothesized to be due to the formation of a macropore system that drains water on preferential flow paths along the stone-fine earth interfaces. For unsaturated conductivity, inclusions caused a general decrease of conductivity as compared to the reference case, and experimental observation and numerical simulations agreed qualitatively. The decrease in the simulations depended on volume fractions of inclusions, number of inclusions, and shape of inclusions.

ASSESSMENT

Altogether, this study is well done and well written. The results are interesting and suitable for readers of SOIL. However, I do have some annotations, which will be listed below. Some more technical remarks are listed later in this review.

As major remarks, I address the following points:

1) The authors compare (i) results from predictive models for effective saturated conductivity (K_{Se}), (ii) results from numerical simulations, and (ii) results from physical experiments. In their paper they use sometimes a slightly confusing nomenclature, such as "virtual experiments", "numerical experiments", "numerical model", "virtual permeameter tests", "virtual constant-head permeameter experiments", "virtual permeameter and evaporation experiments" and so on. I suggest to STRICTLY address the results from the different sources as "predictive K_{Se} models", "[numerical] simulations", and "experiments". If this nomenclature is strictly kept throughout the paper (and also in the captions), it will be easier to understand the discussion. **It will be done as asked by the reviewer.**

2) Interestingly, the authors do not mention in any single word the problem of 2D vs. 3D flow fields. Whereas I understand that simulations in 3D are so demanding at this moment that one cannot request to repeat the simulations in 3D, I would expect at least a qualitative statement and a hint that this problem has been recognized an should be further addressed in the future. As a side note – I assume that the authors used an areal fraction in their 2D simulations that is equal to the volumetric fraction in the true 3D system (which is ok for a perfectly isotropic distribution), right? This information might be added in a side sentence. **The reviewer is right, we considered an areal fraction equal to the 3D volumetric fraction.**

3) In evaluating the simulations with respect to unsaturated conductivity, the authors used tensiometric values at the top and at the bottom of their soil columns I must say that this is quite a "dangerous" strategy, because the validity of the simplified evaporation method has not been shown to be valid for such an extreme setup (see Peters et al., 2015: Journal of Hydrology 527, 531-542, for more information on that issue). Note that once stage-2 evaporation is reached, the top of the *sample dries so much out that the pressure head drops extremely and it is doubtful to apply the SEM. Furthermore, it is quite unnecessary to take that risk, because from the numerical simulations any position of the tensiometers could be used (maybe, even averaging tensions along a line to circumvent the problem of differences due to the distorted flow field is applicable). On the other hand, the authors restrict the depicted K data to values < pF 2.5; for the clayey matrix, this might be still wet enough to allow their method to yield valid results (which can be checked easily for the sample without stones).* **In fact, as the reviewer notes it, the setup is extreme, but in the range of the ones considered by Peters et al. Besides, because of the presence of inclusions, it was difficult to consider observation nodes deeper in the sample. Results of our numerical simulations show numerical instability near inclusions. Moreover, as pointed by the reviewer, the texture of the soil studied allows taking into account values in the pF range considered here. Finally, we've checked that pressure head was linear in the sample, which was the case.**

4) Chapter 2.4: I am not happy with the overview over the experiments and believe it can be improved. In particular, I find the formulation misleading that says "the accuracy of the conductivity curve from the evaporation experiments in the near-saturated zone was improved by using real and virtual permeameter tests". In fact – you cannot improve the conductivity estimation in the near-saturated zone. You can just add a single saturated conductivity point as end point of the function. Considering the interesting finding that saturated conductivity might increase in reality with increasing stone fraction, whereas it will always decrease under unsaturated condition, it is not only unjustified to speak of an "improvement of the near-saturated conductivity estimation" by your methodology – it is even misleading, because rather the contrary is true: you just interpolate smoothly the suction range where you have no direct results. So just stick to the simple facts. You have unsaturated conductivity up to pF 1.6, and then the saturated value. **The text will be modified.**

5) I found the description of the packing procedure a bit meagre. Giving slightly more detail is to be considered, since the authors speculate that the compaction by the packing procedure causes voids along the stones and glass spheres. It is not easy to understand for me how this should happen. **As described in the text, we build the stony samples layer layer-by- layer: soil-stones-soil etc. Each layer of soil was compacted in order to obtain a specific bulk density (each layer had to have a specific depth). Even though the filling and compaction procedure was conducted with precision, it is probably impossible to avoid local bulk density heterogeneity as stones can move and/or soil between stones can be less compacted due to difficult access of the area close to the stone during compaction.**

6) As a personal view, I am not sure whether Figures 2 to 4 are required at all, since it is all contained in nice and concise form in Table A1 (which should be part of the main text anyway). But if the authors have enough money to pay the charges … the figures don't hurt.

SPECIFIC ANNOTATIONS

P 1105, line 16: please define your term "effective saturated hydraulic conductivity". To my understanding, it is the saturated conductivity of the sample, so "effective" is not really necessary and opens a realm of theoretical difficulties.

P 1106, line 5: There is no such thing as "decreasing hydraulic properties".

P 1106, line 25: "numerical experiments", "numerical permeability experiments" and so on: please refer to my remark (1). Furthermore, you "performed" (rather than "completed") simulations.

P 1108, line 6: "… based on the pore size distribution of Mualem" – Wrong. You refer to the pore-bundle model of Mualem.

Page 1108, line 10: In soil physics, we refer to S_e as "effective saturation" (not "saturation state").

Page 1108, line 21: "… rock fragments with a mean diameter between 1 and 2 cm" – a reader asks himself what is meant with "mean" diameter, and how the distribution might look like. I suggest to delete "mean" or else to define it.

All these comments will be addressed as the reviewer asks.

Page 1109, line 4: "… equal to the mean bulk density …" – from where do you know the mean bulk density if that material? Measured at the site where the samples are from? **Indeed, the bulk density was measured in situ. It will be indicated in the text.**

Page 1109, line 21: "… In order to avoid preferential flow due to the introduction of the tensiometers […] vis-à-vis the center of the tube" – This is hard to understand to me. I am sure you can express this more clearly, so that somebody who has not seen the experimental setup will understand it.

Page 1109, line 26: "Tensions beyond the consolidation point were not taken into account". I work my life long with tensiometers, but I never heard of a "consolidation point"! Even if you explain it afterwards, the terminology appears strange. Maybe, you can avoid the terminology or cite a proper source? **It is the term that is not adequate, it is the air-entry point. The text will be modified.**

Page 1113, line 21f: "The parameters …" – Please be more specific here –just a half sentence is not sufficient to explains what you did. Did you fit the hydraulic function to the SEM data, or did you do an inversion of the Richards-based numerical simulation of an evaporation experiment? If the latter applies: What was used in the object function, and how did you determine/set the weights of the different data types? **We fit the hydraulic function to the SEM data. The text will be modified to better express it.**

Page 1114, line 9ff: "As numerical errors occur…" – this passage is written in a very diffuse and non-specific manner. I see from Figure 5 that K-data are only determined above pF 1.6 – which is clearly not the threshold that can be reached by the numerical accuracy of simulations. It appears that the authors sacrificed some accuracy by deriving tensiometric gradients in a non-optimal manner? To me, this does not hurt the value of the paper, but they should specify more precisely how they numerically derived the conductivities. Either, the accuracy of the tensiometric values were limited by the number of digits in the output, or else they had problems with the numerical stability of the simulation that could be possible improved with some altered numerical parameters? **As the relative mass balance error was large at the beginning of the simulations, we considered values when this relative error was lower than 5%. This validation criterion was set arbitrarily, based on the comparison between evaluation points from the simulation of the evaporation experiment on stone-free samples and the expected values obtained from the inputs of the simulation.**

Page 1114, Chapter 2.4: As indicated in the major comments, this section can be improved. Also re-consider the organization of Table 2 and consider using symbols for the used soil shapes that are more suggestive: ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰filled circle, filled triangles, and vertical and horizontal slabs (available certainly in special fonts – but I had to look myself…).

Page 1116, line 15: "95 % confidence intervals" – I do understand how you calculate the median, but I do not understand how you calculate a 95% confidence interval for the median. Please specific in the methods section. **The interval is the 95% around the median of the predicted values. So it is not a confidence interval, but an interval of the variation between models results. This text will be modified.**

Page 1119, line 23: "… inclusion vicinity…" – consider re-phrasing. Page 1120, line 26: You cite "Gras, 1994", but I did not find it in the references.

Page 1121, line 22: "We also hypothesize …" – this comes out of the blue and was not mentioned nor discussed in the paper. So, you cannot bring it in the conclusion. I suggest to delete the sentence, or to specify what brings you to the conclusion.

Page 1126 and 1127: As indicated above, I suggest to use more suggestive symbols for the five shape variants.

Page 1128, Figure 1: Where are the error bars for the lab experiments? I understand all of them were performed in two replicates? Hence, the difference is equal to the estimate of the standard deviation (it makes no sense to show 95 % confidence intervals, since the t-value is huge). Furthermore: how did you calculate the 95% confidence intervals of the median? **At first we had only one replication for Ks measurements. Now we have 5 replications and error bars will be added. Concerning the IC, the interval is the 95% around the median of the predicted values. So it is not a confidence interval, but an interval of the variation between models results.** *Page 1132, Figure 5: Please scale the y-axis on a log scale, as is usually done for K plots (and as you do in Figure 6). Furthermore, consider to start* *the plot at pF 1 and indicate by symbols on the left Y-Axis the values for saturated conductivities (please note that saturated values are NOT related to pF 0, as is presently suggested). The same applies to Figure* 6.