Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2019-80-RC2, 2019

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Relevance and controls of preferential flow at the landscape scale" *by* Dominic Demand et al.

Anonymous Referee #2

Received and published: 25 March 2019

The manuscript by Demand et al examines the occurrence of preferential flow during infiltration for three geological regions and two land cover classes in a catchment in Luxembourg. Field measurements consisted of soil moisture measurements made at 45 sites distributed across geological and land cover classes, as well as hillslope position, slope and aspect. Each of the 45 sites consisted of three soil moisture profiles with sensors installed at three depths (10, 30 and 50 cm below the surface). Micrometeorological measurements were also collected at each site. Data were collected from 2012/2013 to 2017. Based on two proxies for preferential flow (1. non-sequential soil moisture response and 2. sequential soil moisture response velocities that exceed predicted capillary flow velocities), they found that preferential flow dominated across sites. The greatest differences between theoretical predictions of capillary flow velocities and

observed flow velocities occurred in soils with high clay content. These results suggest that preferential flow is the primary way that water infiltrates into soil in this region and that clay soils should not be considered as low hydraulic conductivity mediums.

The study tackles an important topic and contains an impressive amount of field measurements. I am a physical hydrologist, but my training and research is primarily on stream channel processes, so I am not aware of the current literature on soil infiltration and plot-scale preferential flow dynamics. Therefore, it is difficult for me to assess the novelty of the study, but it appears that the data collection and methods used are well established and that the uniqueness of this study is that they sample different geological and land cover settings for a catchment with generally uniform climate.

The introduction does a good job of reviewing the literature and setting the context for the study. My main issue with the manuscript (echoing Dr. Bogena's comments) concerns the rationale and presentation of the results. The results are very difficult to follow and it is not always clear why certain analyses were done and how they link back to the main objectives of the study. It is often not clear how and why soil profiles were grouped for certain analyses. I encourage the authors to provide more clarity on the analyses and highlight how the analyses address the research objectives.

Below, I outline two general comments, followed by some specific comments.

General comments:

a) Statistical issues

Some of the key conclusions made in this study rely on frequentist statistical testing (e.g. p-values), which, as the authors acknowledge (p23,l5-11), can be highly sensitive to sample size issues. There has also been considerable discussion recently about the major limitations to this approach (see Amrhein et al. 2019. Nature 567:305-307 for a very recent example, also Wasserstein and Lazar 2016. The American Statistician 70:129-133). It might be valuable if the authors discuss some of the inferential uncertainties and limitations of their approach. In particular, I see three topics worth discussing:

1) pseudo-replication: It seems to me that the statistical models should be fit to the 45 sites, not the 135 soil profiles, since the grouping of three profiles within each site cannot be treated as independent. Focusing on the 135 soil profiles could be done within the GLMs if within-site variability were accounted for, but it's not clear to me that this was done.

2) sample size vs number of predictor variables used in the models: Although there is an impressive amount of data collected for this study, I'm concerned that some of the results (e.g., identification of statistically significant predictors) are simply the product of small sample sizes and noise in the model fits. For example, the GLM for Grassland-Sandstone was fit to 9 soil moisture profiles (so really, just three sites), but 13 predictor variables were used in the model fitting, which will result in an underdetermined solution. If the authors decide to keep the statistical analyses, I would suggest some sort of cross-validation exercise be done to assess the rigor of the models.

3) data exclusion. It was suggested in the methods that there is some incompleteness to the time series for each soil profile (due to logger failures and criteria for including data in the analysis). How many sites and profiles were excluded and for what time periods? This is important to know as it relates to the sample size issue outlined above.

b) Within-site and temporal variability

Instead of focusing on statistical significance, I think the authors could make an excellent contribution by focusing more on the within-site and temporal variability of their field measurements. My understanding is that the grouping of the sampling approach can be organized as: geology - land cover - site - profile. Most of the analysis focuses at the geological and land cover levels; however, throughout the manuscript I found myself constantly wanting to know more about the within-site variability in terms

of both infiltration event characteristics and soil properties. Also, at the profile level, I wanted to know more about the temporal variability. Did profiles that exhibit NSR only exhibit NSR or did they shift between NR, SR and NSR? If so, why? Instead of generalizing the results using p-values, I suggest focusing on graphical approaches to show evidence to support the research objectives.

Some specific comments:

p1,l27-28: Consider incorporating the parenthetical into the sentence - as is, this makes for a weak opening.

p2,l17: Consider removing this last sentence or expand on it to clarify to the reader what is meant by hotspots and hot moments of PF.

p2,l29-p3,l18: Consider revising these paragraphs. Right now these feel like simply a list of results from other studies. I suggest trying to better synthesize these results and identify key findings and knowledge gaps.

p4,l9-11: I think the research questions could be improved. What is meant by 'underlying controls'? Has this actually been done in this study? It seems like the PF proxies are linked to precipitation, landscape, and soil characteristics through statistical modeling. 'Underlying controls' suggests to me a more process-based approach (e.g., soil physics modelling), which isn't done in this study - outside of the predicted matrix flow velocities. What is meant by temporally stable?

p5,l25-26: What is the orifice diameter of the rainfall gauges? How was the placement of the forest gauges determined? Was variability in canopy cover and throughfall a concern?

p6,l3: Why weren't infiltrometer measurements available for the grassland/Sandstone sites?

p7,l5: Why would the sensors log these kinds of 'implausible' events? How many events were rejected because of these criteria?

p7,l29-31: How many times were data from a profile rejected because of these criteria?

Table 1: The first row highlights to me the potential issue of pseudo-replication in this study. It seems more appropriate to report the number of sites, not profiles. Also, for the soil texture and mean clay content, how variable were these values within geologic-land cover class combinations?

p9,l18-26: This paragraph is unclear. Why were some GLMs fit for individual landscape units and one GLM fit to all profiles? What was the sample size used in each of these models? If I understand this correctly, the GLM model for Sandstone-Grassland was fit to just 3 sites (9 profiles)? This seems like much too small sample size to fit models with up to 13 predictor variables.

p9,l27: Why restrict to slopes > 10%?

Fig 2a: I'm not sure what conclusions to draw from this graph? The text suggests that it shows no clear difference between landscape units; however, there appear to be considerable differences in cumulative precipitation (e.g., for event numbers $\sim$ 150, we see a range of P[sum] of almost 750 mm). Also, I don't think the various landscape units share the same event records (i.e., the x-axis is not sequential for each landscape unit), so are they comparable?

p11,l24: Why only 2014 and 2015?

p15,l1: Would a relationship between NSR and distance to stream be expected? Perhaps a provide a rationale.

Fig 4: Where do these data come from? Are these averages across all years in the study? How much did this vary between years? Does this figure account for differing number of events each year or exclusion of profiles due to logger failure or selection criteria?

Fig 8: The fit is statistically significant (but see general comment (a) above), but is this relationship practically meaningful? If you were to remove the line of best fit, I'm not

sure someone would identify a relationship in these data.

p22,l32: How much range in hillslope position was sampled? Was this the distance to stream metric?

p24,l10: Speculations on why this is?

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-80, 2019.