Response to comments of anonymous Referee #2

We thank the anonymous referee #2 for reviewing our manuscript and the comments concerning mainly statistical issues. We answer below to each comment in a point-by-point reply. For clarity, the comments of the referee were copied in black and our comments are in blue.

General Comments

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.

We will add a flow chart in the method section for a clear description of the analysis and clarify the grouping of the soil moisture profiles for each analysis step. We will also better connect the different sections to the objectives and hypotheses and therefore improve structure and readability.

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.

It is true that different sample sizes can lead to problems in the interpretation of test statistics (as shown in e.g. Amrhein et al. 2019. Nature 567:305-307) and that the p-value should not be used as a rigorous yes/no criterion. However, test statistics can give additional insights, especially while comparing large samples. None of our analysis and interpretations is purely based on p-value statistics and they were just added to give additional information (such as error bars, distributions etc.). Since sample size is a critical issue we added the sample size where it was missing in the manuscript to support interpretation.

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.

The linear model (LM) was fitted to the mean NSR percentage of the 45 sites. The generalized linear model (GLM) was fitted to the infiltration events of the 135 profiles since it covers the temporal domain. Indeed pseudo-replications are an issue for the GLM. We think the general intention of the model is still important (see response RC1 Heye Bogena) and therefore modify the GLM to a generalized linear mixed effect model (GLMM) with a logit link function to account for the binomial data (NSR/no NSR) (see suggestions referee #3). By using a GLMM for the 45 sites with the individual sites as the random effect of the model we avoided pseudo-replications and treated the individual spatial landscape effects (geology, land-cover, slope, aspect etc.) together as one random landscape effect. Since the results of the GLM did not show clear landscape properties that had an effect on NSR occurrence (see RC1 Heye Bogena) we think treating it as a random factor is appropriate.

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.

The GLM was fitted to all events of the 9 profiles of the Sandstone grassland sites, which are 698 data points. Furthermore, some predictors were removed by stepwise AIC, so that only four predictors were used (not 13). Hence, the model does not result in an underdetermined solution. We will clarify that the temporal model (GLMM) was fitted to all the events of each site providing a good basis for such a statistical model.

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.

We will add a diagram to the supplement that shows how many profiles in which landscape units were active or met the quality criteria over the entire time period (~2012-2017).

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.

The within-site/profile variability is indeed an interesting topic. We will add a sentence about the within-profile variability to the results. Table 3 already gives some information about the within-profile variability. However, the site or profile-level variability was not the main aim of this study. Other studies have already focused on that topic (see e.g. Wiekenkamp et al. (2016) or Liu and Lin (2015) in the reference list of the manuscript). The aim was to show the effect and variation of larger-scale landscape units with different properties. Furthermore, we wanted to identify potential temporal differences and similarities among their reactions. We will clarify the aim in the revised manuscript.

Many results in the manuscript include graphical approaches (see Fig. 3, 4, 6, 7, 8). We do not think that adding the test statistics weakens our findings (see response to General Comments).

Specific Comments

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

We incorporated the section into the sentence.

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.

We removed the sentence.

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.

We will revise the paragraph and summarize the studies.

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?

Indeed we did not clearly identify processes. The research question will be rephrased. By "underlying controls" we meant spatial and temporal influences or drivers of preferential flow occurrence on a larger scale (e.g. landscape units).

"Temporally stable" refers to the preferential flow occurrence (if it is changing over time or not).

We will clarify both.

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?

The orifice diameter of the rain gauges is 16.5 cm (collection area 214 cm^2). The rain gauges were randomly placed on the 29 forest sites. The information will be added to the sentence. The experimental design of placing the five throughfall gauges aimed at covering the variability in canopy cover at each site, and variability in measured throughfall between the gauges was expected.

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

Hood infiltrometer measurements are often time consuming and we were not able to measure all sites during the same field campaigns.

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

During the reconnecting of the loggers following a logger error (no power etc.), the rain gauges sometimes produced this kind of implausible events. Furthermore, clogging and release of the clogged water could be a reason of the unrealistic rainfall events. Out of 10675 rainfall events (sum of rainfall events at all 45 sites), 464 rainfall events were excluded by this quality criteria. The number of rejected events was added to the sentence.

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

From a total of 15721 soil moisture events (sum of soil moisture event observations at all 135 soil moisture profiles), 7470 soil moisture event observations were rejected because of this criterion. If a single soil moisture sensor of a given sensor profile fails during an event, response analysis for the profile is not possible for this event and hence it is rejected. The high number of rejected events mainly results from long-term sensor failures of single sensors in different sensor profiles. Therefore, using a different quality criterion, for example 95% of usable data points of each soil moisture sensor per event, would lead to similar results (7243 soil moisture events rejected). We included the number of rejected soil moisture event and the explanation into the section.

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 geological and cover class combinations?

Table 1 is just an overview of the study sites. The different sensor responses (SR, NSR) were calculated for every single soil moisture profile. We think it is appropriate to give the number of profiles, since they determine the number of observations. We added the full textural information and the standard deviation to the supplements.

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.

Please see our response under General Comments. The GLM was fitted to the individual landscape units to test for differences in the predictors on this scale. Furthermore, we compared

it with one GLM for the whole catchment to see the potential for such an approach. We will add the sample size (number of events at each site) for the new GLMM.

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

A slope < 10% is relatively flat and the orientation is not strongly pronounced. To focus on the main findings of this study (temporal variability of PF occurrence between landscape units), the analysis of the aspect was removed.

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 ~ 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?

It is correct that the events are not identical and necessarily sequential due to the rainfall heterogeneity, quality criteria, and the length of the time series. Therefore, single sites can show high deviation even within the same landscape unit. The motivation was to show that there is no systematic difference between the landscape units. We think that Fig. 2b) provides enough information and we will remove Fig. 2a).

p11,l24: Why only 2014 and 2015?

These were the first two years with all sites installed. Since the calculated proportions of these two years (and also the other years) do not add additional relevant information for interpretation, we removed the section to focus on the main findings (see also response to RC1).

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

Some authors found a relationship on hillslope position (see references P22L32-P23L4). We agree that an explanation should be mentioned earlier in the manuscript and is currently missing. However, the analysis of the small scale spatial patterns will be removed to focus on the main findings (temporal patterns of soil moisture and rainfall).

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. 4 is based on the same data as all other diagrams. It shows the mean NSR of all events that were measured in the twelve individual months independent of the landscape unit. Hence, the diagram averages across different years. We will modify this diagram and separate between forest and grassland sites and add additional lines to show the variability among years. The number of events used for this analysis (number of events per month for all years and the individual years) will be added to a table in the supplements.

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.

We totally agree, the fit is statistically significant, but explains only little variation (we wrote this on P18L11-12). Furthermore, on P24L28-30 we note: "The θ-v_{max} relationship shows that even though the decrease of v_{max} with [decreasing] θ is significant, it has little explanatory power and fast flow (>1000 cm day^{-1}) can occur at any θ." We included the fit to highlight the strong variation that not simply follows the trend. We specified this in the discussion.

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

The range of hillslope position was determined by the distance to stream with a range between 4 and 251 m from the different sites to the stream. Please see the table in Appendix A.

p24,l10: Speculations on why this is?

Texture seems not to be the main driver of water flow velocity during infiltration in the classical manner that fine grained texture corresponds to slow flow. Infiltration seems to be strongly controlled by PF phenomena, which are dependent on soil structure (influenced by a high clay content), biotic macropores (roots channels, earthworm borrows) and initiation processes (hydrophobicity, rain intensity). The high heterogeneity of the landscape and its temporal variation leads to PF that is caused by different drivers that are partly independent of texture (e.g. organic carbon content, number and species of soil organisms, vegetation type, rainfall characteristics). We added this as a potential explanation to the discussion.