Reply to Anonymous Referee #1:

Ralf Loritz (RL): We would like to thank the Anonymous Referee #1 for her/his insight and thoughtful comments. We are thankful for the effort the referee put into this review in the form of the high quality of his comments. In the revised manuscript we will follow many of the reviewer's recommendations, because this will definitely improve our study. Furthermore we see from the comments that some parts of our study need better or more detailed explanations which we aim to provide in a revised version of our manuscript.

Reviewer: Summary and Recommendation:

In this paper the authors address two basic questions:

1) If you have a lot of spatially-distributed information about the geology and soil-hydraulic properties in a catchment, can you parameterize a high-dimensional, spatially-distributed model (without any calibration or inverse optimization) to accurately represent water flow within a single 2-d hillslope, based on that existing knowledge?

2) If your knowledge-based (not optimized) model domain and parameterization prove reasonably representative, can you then extrapolate this representative 2-d hillslope across the 3-d volume of the entire catchment, to simulate hydrograph dynamics and the annual water balance for the entire catchment?

To address these questions the authors employ a Richards-equation-based model with evapotranspiration module and overland flow routing modules. They apply the model to simulate hillslope-scale soil moisture dynamics and water-balance partitioning and after extrapolation, whole-catchment streamflow dynamics from two catchments in Luxembourg with varying geology, topography, soil, and vegetation. In addition to the modeling, their analysis includes extensive and impressive data sets representing spatially distributed soil-hydraulic properties, geologic features, plant transpiration, and topography. These questions, the observations, and the methodological approach adopted here are of interest in scientific hydrology and would be received with interest by readers of HESS. The writing is mostly clear in a grammatical sense, though somewhat desultory and technically ambiguous in many areas. The model setup is technically sophisticated in many ways, though not so in others. The graphics are of very high quality. The organization of the paper is logical, though as I suggest below, there are significant portions that might be omitted to maintain a consistent focus of the paper throughout. Overall there were too many competing objectives in the paper. As such, any salient result or conclusion is hard to discern. The results and discussion could be greatly improved.

I recommend that this paper could be accepted for publication in HESS, but I believe some major revisions are needed beforehand, possibly including revised simulations and results. Those suggested revisions are outlined in the following g General Comment's section, with references to more specific Technical Comments.

RL: We thank the reviewer for this comment. Additionally to the two questions, the paper addresses a third question. It deals with the problem of how to find the most important information or data source

that is needed for setting up hillslope models. This is the basic idea of the virtual experiments. We agree with the reviewer that this part of the study can be much improved, but we do not agree that the reported results are not of interest.

Reviewers: General Comments:

Reviewer: *The manuscript can be improved upon significantly by reducing total length, omitting or clarifying the use of excessive jargon, and possibly removing sections of the paper to minimize superfluous and unfocused commentary. Some specific instances are noted in Technical Comments 5-10, 24, 27, and others. The manuscript may be much improved by omitting much of the commentary about modeling evapotranspiration, along with the virtual experiment 3 (VE3) and associated results, and rather focusing on the importance of explicitly representing (or not) landscape heterogeneity for the purpose of simulating hydrographs. The authors talk quite a lot about all the uncertainties associated with ET modeling in hillslope/catchment hydrology, but their modeling approach does not reflect the state of the science (e.g. as presented in disciplines such as hydrometeorology and plant biophysics), so this discussion does not seem warranted. See Technical Comments 12, 29, and 56. The methodological approach is inadequately described in many instances, with some revision being needed. See Technical Comments 16-26, 36, 39, and others. There are some aspects of the model domain and parameters that seem very unrealistic. For example, the bedrock for both modeled catchments is parameterized to have porosity of 40-45% (Table 1). That's comparable to, or greater than, the porosity of many soils. I can't imagine how Schist a metamorphic crystalline rock can be composed of 40% air space. This is certainly not consistent with most reported porosities for Schist, which are typically 10% or less. This will have a large impact on the flow simulations, since about half of the hillslope domain is bedrock. If there is some justification for this, and some other aspects of the model, then perhaps the only revision that is needed is to provide that justification. Otherwise, many of the simulations may need to be repeated with more appropriate parameterization. See Technical Comments 19-26, and others.*

I strongly suggest that the authors consider refocusing this paper on two subject areas. First, concentrate on the modeling of spatially-distributed soil moisture dynamics and the temporal dynamics in the hydrograph. You use a spatially distributed model but nowhere do you assess the models ability to accurately represent any spatial pattern. You have an enormous amount of interesting spatially distributed data, so this could become one of the most rigorous tests of the Richards equation model at the hillslope scale ever published. See Technical Comments 33-34, 47-48, and 54. I recommend the second focal point to be the argument that extrapolating the parameterization of a single 2-d hillslope to an entire 3-d catchment may, or may not, be defensible. At present this is stated as an objective, but the results and discussion are ambiguous, or possibly in disagreement, about this point. Showing these outcomes would be challenging enough, and a good contribution to contemporary scientific hydrology. Toward that aim, I suggest omitting the virtual experiments and associated discussion. I found the virtual experiments and the associated discussion around them to be desultory and vastly oversimplified. It was not clear to me how they relate to any previously stated objective. Those virtual experiments mainly consisted of changing a single variable (e.g. total relief, timing of bud break for plants, or hydraulic

conductivity), and speculating broadly about the implications of the resulting simulations. See Technical Comments 50, 52-53, and 55-56.

RL: We agree with the reviewer that the manuscript can be streamlined, particularly with respect to the discussion of the ET approach. We will further elaborate this in the technical comments.

The reviewer is also right that there are too many related objectives in the paper. We thought and still think that an exhaustive hillslope model study should address the two questions summarized by the reviewer and our proposed third question about the importance of different data sources. We agree that an exhaustive treatment of all these three questions is difficult to be digested within a single paper. A natural point to streamline the study is thus to remove most of the virtual experiments and treat them in a more exhaustive manner in a separate study - in this respect we look forward to the editors' advice.

We will focus our revised manuscript with respect to the two below summarized question of the reviewer and explain in more detail how we parameterized our two models. Our goal is to make our model structure and the choice of parameters as transparent as possible. Hence we will set up our vegetation parameters as far as this is possible with observed values (*technical comment 12*) and will rework and better explain the way how we chose our macropore parametrization (*technical comment 19 & 20*). Furthermore we will show that the large heterogeneity in the soil samples cannot be grouped by landscape characteristics in a simple manner (*technical comment 16*). Finally, we will try to better account for the soil moisture variability along the hillslope and be more careful with the term "distributed" (*technical comment 34*).

""

1) If you have a lot of spatially-distributed information about the geology and soil-hydraulic properties in a catchment, can you parameterize a high-dimensional, spatially-distributed model (without any calibration or inverse optimization) to accurately represent water flow within a single 2-d hillslope, based on that existing knowledge?

2) If your knowledge-based (not optimized) model domain and parameterization prove reasonably representative, can you then extrapolate this representative 2-d hillslope across the 3-d volume of the entire catchment, to simulate hydrograph dynamics and the annual water balance for the entire catchment?

""

***Reviewer:** Last, the authors seem to be advocating that high-dimensional, spatially-distributed models will always be wrong for a variety of reasons, but that they're still important for helping us learn about which state-variables and flow processes dominantly control the emergent streamflow dynamics at the catchment outlet. I think this is a relevant and worthwhile argument, but I encourage the authors to better focus their writing on this topic. This effort may be aided by omitting some other sections of the paper, as noted above. At present the authors present some seemingly conflicting (at least to me) statements about what exactly is the merit of taking this approach, and just how well it did or did not work out for them. See Technical Comments 34, 46, 47-48, and 54.*

RL: We will restructure our discussion in the revised manuscript. Although we would rather use the term "approximation" than "wrong" – the clue is to get the "best" approximation which is as simple as possible but not oversimplified. We come back to this point when dealing with the related technical comments.

General comment: Bedrock parametrization

Reviewer: *There are some aspects of the model domain and parameters that seem very unrealistic. For example, the bedrock for both modeled catchments is parameterized to have porosity of 40-45% (Table 1). That's comparable to, or greater than, the porosity of many soils. I can't imagine how Schist a metamorphic crystalline rock can be composed of 40% air space. This is certainly not consistent with most reported porosities for Schist, which are typically 10% or less. This will have a large impact on the flow simulations, since about half of the hillslope domain is bedrock. If there is some justification for this, and some other aspects of the model, then perhaps the only revision that is needed is to provide that justification. Otherwise, many of the simulations may need to be repeated with more appropriate parameterization.*

RL: You are right that Schist does not have a porosity of 40%. But this Schist formation is highly fractured and additionally covered by periglacial deposits. We are only modelling the upper 2 m of the subsurface, and hence only the upper meter of the weathered bedrock. Citing Wrede et al. (2015), the periglacial deposits "*may store significant amounts of water, more than expected from an analysis of soil mapping information alone".* Furthermore we assume a saturated hydraulic conductivity K_s (m s^{-1}) of $5x10^{-9}$. We chose such a low value because we expect no major groundwater body beneath the hillslope (Bos et al., 1996). Hence there is almost no water flow through the bedrock. Deep percolation (water leaving the hillslope through the lower boundary) in our reference model makes up around 0.001%. To follow up on the reviewer's comment we repeated the simulation for the Colpach with a reduced bedrock porosity of 0.1 (Figure 1 & 2). As you can see there bedrock porosity is not a sensitive parameter. We will stress this in the revised manuscript and explain the low sensitivity of this parameter because of the low permeability to avoid further confusion about this point.

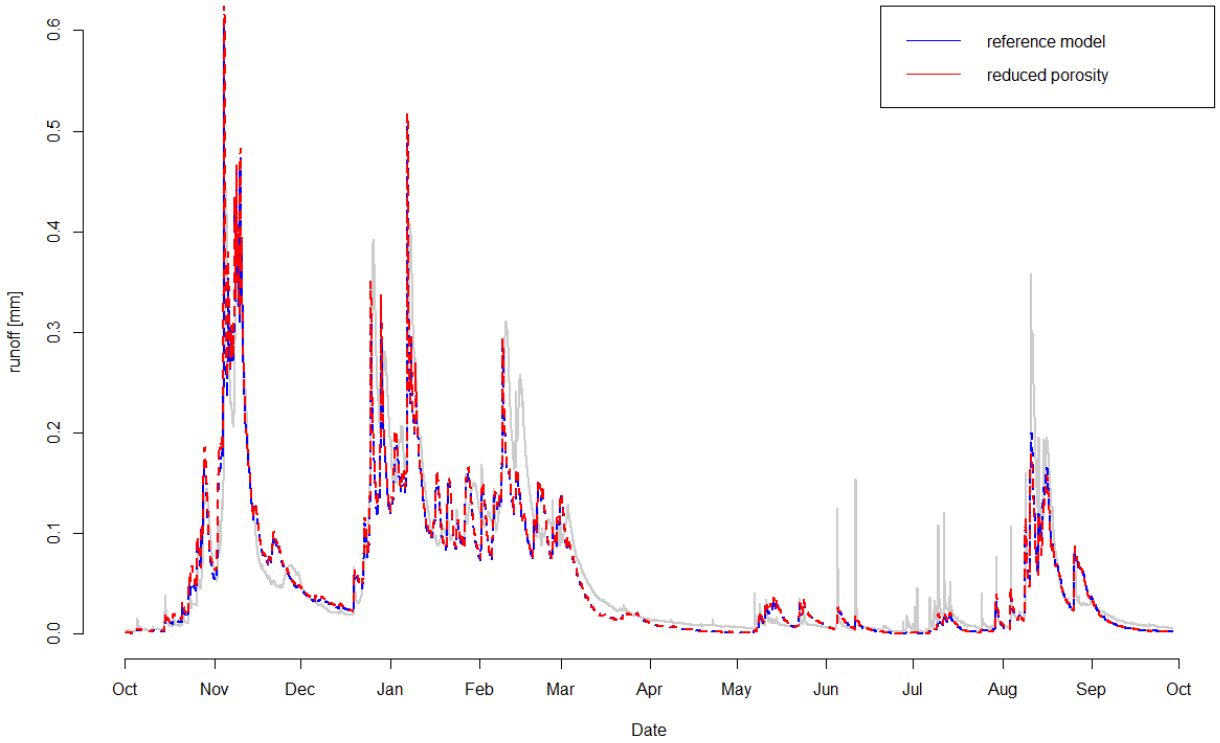

Figure 1 Observed (grey) and simulated discharge of the reference model (blue) and the reference model with the reduced bedrock porosity of 0.1 (red).

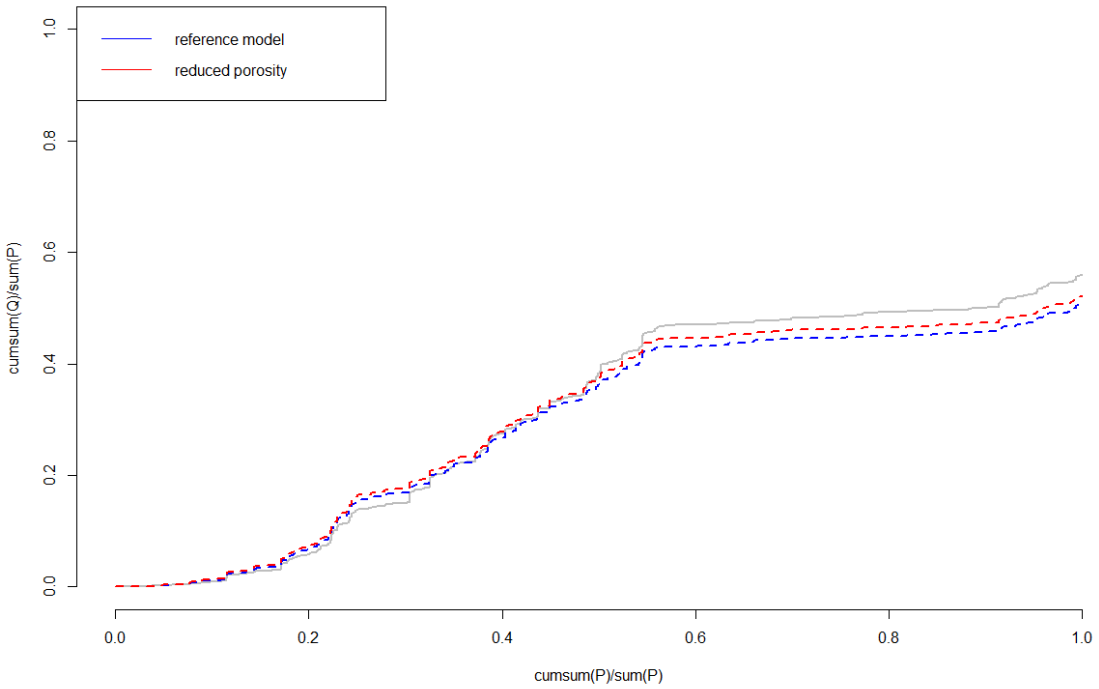

Figure 2 Double mass curves of the observed (grey) and the simulated discharge of the reference model (blue) and of the reference model with a reduced bedrock porosity of 0.1 (red).

Technical comments:

Reviewer: *Comments 1, 2, 3, 4, 8, 13, 21, 25, 27, 28, 38, 40, 42, 43, 45, 49.*

RL: We agree with the reviewer in these points. These mistakes/statements will be corrected and/or rephrased.

Reviewer: 5) Line 67: *Not immediately clear to the reader what distinguishes a "conceptual model" from a "perceptual model".*

RL: We apologize we thought these terms need no further specification. We will give a short definition to improve the clarity of the presentation. A conceptual model is a bucket-style model, for instance the HBV beta store for soil moisture accounting or a linear reservoir. A perceptional model reflects our imagination on the dominant processes and structures that govern for instance runoff formation in a catchment. Personally, we consider the classification of hydrological models into conceptual and physically-based models inappropriate; since it implies that conceptual models are not based on physics and that physically based models do not have an underlying concept. We actually prefer the definitions of Gupta et al. (2012) but we used the terms nevertheless, because we thought it is well-established hydrological jargon and to avoid confusion.

Reviewer 6) Lines 85-88*: Consider rephrasing or deleting. Not clear what is the message of the sentence. The result of any mathematical model must be compared to observations in the modeled system. This goes without saying, and certainly doesn't require a supporting citation. Perhaps I am just not clear what you mean by "benchmarking".*

RL: With this we mean that physically-based models can be evaluated against observations of stream flow, soil moisture states and even tracer data in a straightforward manner. This is not the case for conceptual models and thus an advantage. We will replace the term "benchmarking".

Reviewer 7) Lines 49-97: *This is very clearly written, but for sake of making your paper as concise as possible and therefore more likely to be read in full. You might consider abbreviating the section, or deleting. It doesn't assert much, or highlight some problem with the status quo in catchment modeling. It mainly states that model-based analysis are useful for learning about hydrological systems, which I think is already acknowledged ubiquitously in the community of hydrological scientists. It's your call, but there is no shortage of papers on hillslope modeling, and I always prefer to read a more concise one than a longer one.*

RL: You are right. We will definitely streamline this paragraph, though it contains a key point – the picture/image idea. By using a suitable color code for plotting for instance soil parameter of a two-dimensional hillslope, this plot resembles a perceptional sketch of a hillslope or catchment. In fact this "structural setup" can be tested against augers or ERT images. *Is the bedrock in our model in the same depth as observed or consistent with available ERT images?* Moreover, such information could be used for setting up the hillslope model. This is not possible with "conceptual models" because they are not spatially explicit and not thermodynamically consistent (fluxes are not driven by gradients). Hence, we can use much more information which is independent from our target data for an a-priori setup of the hillslope model.

Reviewer 9) Lines 113: *"functional behavior of catchments of organized complexity" is hard to interpret. Caution in using too much ambiguous jargon.*

RL: This formulation will be removed from a revised manuscript. However the term "organized complexity" was coined by Jim Dooge (Dooge, 1986) to characterize catchments that already exhibit too much heterogeneity to be treated in a deterministic, physically based way but a are yet too small for a conceptual treatment.

Reviewer 10) Lines 128-130*: Consider rephrasing using the plainest language possible, since this is seemingly an important part of your rationale statement for the study*

RL: This is indeed an important part of our rationale, thank you for pointing out that our text is not easy to read and interpret. We agree that this should be improved and we have tried to rewrite these lines to convey the message better. We have changed the sentence to: "We propose to start with the perceptual models (Top panel of Figure 3), which provide qualitative information (such as impermeable bedrock with shallow periglacial, highly porous soils on top, with a network of vertical and lateral flow paths) and to transfer this into a first-guess parametrization of the hillslope using the available data or literature values."

Reviewer 11) Lines 137-140: *Rephrase "behavioral physically model structures". Also, comparing model outputs to observed data sets (like tracer time series) doesn't inherently reduce the number of degrees of freedom in the modeling procedure. It might help constrain parameter values. If that observed data set somehow informs the modeler that a particular parameter is unnecessary, or that a spatially-distributed domain can be adequately represented in a lumped way, then the degrees of freedom might be reduced. Are the works you cite here examples of the latter? For example, in the application of Richards equation with the van Genuchten-Mualem soil-hydraulic model, you can't just decide based on some observation that you no longer need the shape parameter in the hydraulic model – it still has to be there. If the observational data set leads you to a coarser grid resolution for the domain, then that would be a reduction in degrees of freedom, since the number of spatially distributed elements where the equation is solved/averaged is reduced. But really, for multi-parameter models that are spatially distributed, the degrees of freedom are always grossly high, not even considering the fact that we most often ignore anisotropy and hysteresis in soil hydraulics in hillslope to catchment-scale applications. That's the whole motivation for lumped models, right?*

RL: Here we refer to studies of Klaus and Zehe (2011) and Wienhöfer and Zehe (2014) and how they simulated flow and tracer transport at a tile drained field as well as at a forested hillslope in a two-step procedure. In a first set they used a basic two 2d hillslope model and represented vertical and lateral preferential flow paths by different spatial densities and hydraulic conductivities. From the 400/120 trials several networks reproduced the observed till drain outflow/hillslope runoff in acceptable manner, though the networks were quite different. These hillslope structure were hence equally likely since they all produced the same amount of fast subsurface flow, either through a higher number of less conductive macropores or though "fewer more conductive macropores". In second step they simulated tracer transport through these behavioral hillslope structures (or architectures if you wish) and which reduced the number of behavioral ones to 4 or even to zero. Tracer data impose an additional constrained as not only the flow (the filter velocity) must be reproduced but also the transport velocity in the porous medium. This is what we mean with reducing the degrees of freedom in the "model space".

The sentence will be rephrased to: But the number of model structures that are physically meaningful can be reduced by using complementary observations such as tracers.

Reviewer 12) Lines 166-184*: Evapotranspiration is represented in a rudimentary way in many hydrological models precisely because those models have the primary aim of predicting hydrographs. For*

modelers with this primary interest, there will inevitably be a greater effort spent on representing such processes as non-equilibrium flow than on ET, because the former is of more interest. Models aimed at predicting streamflow use long-standing, and possibly antiquated ET models, because they're convenient, not because enhanced knowledge of stomatal dynamics and plant phenology is absent. Tree physiologists and hydrometeorologists have highly advanced understanding of these phenomena, and their discipline-specific models reflect that. These arguments are worth keeping in mind when you're noting the "uncertainty in the community on how to represent plant physiological controls on transpiration in hydrological and land surface models." Is it uncertainty, or just lack of interest/effort to study and implement models that reflect contemporary knowledge in plant physiology and boundary-layer biophysics? As an example, you go to great lengths here to incorporate small-scale non-uniformities into the subsurface flow domain, but you use a pretty standard version of the PM approach for ET that's been around for over 30 years now. You assume homogenous land cover in Colpach catchment, and you use vegetation parameters from a non-local catchment in Germany, with phenology assumed invariant from year to year (I assume that's what you mean by "fixed"). With that model setup, it's really not appropriate for you to be talking about how uncertain the community is in how to represent the complexity of these processes in models. It wouldn't be very complex for you to considerably improve this standard version of the PM model just by using site-specific data, a dynamic representation of phenology, and to accurately reflect the relative abundance of forest versus other vegetation cover in the catchment. I don't want to be overly critical, because some assumptions and simplifications are always made in modelling. My main point is that you don't want to go on philosophizing about how we learn from still-uncertain models, if the model you're employing is not nearly state-of-the-art, or not parameterized nearly as carefully as it could be.

RL: The reviewer is more than right with this point, the land surface model community uses far more sophisticated approaches for ET catchment modelers usually do. In fact it was not our goal to highlight our evapotranspiration routine as sophisticated, but to stress the deficiencies of the approach. Our opinion is that hydrological modeling does not end with a successful simulation of runoff especially if ET is the major outward flux of the catchment half of the year. Furthermore, ET is one of the main controls of the fill level of the storages in hydrological model. Since in almost all hydrological models the bulk of runoff is produced as a function of the model state, ET becomes increasingly important in long-term simulations.

The reason why more sophisticated approaches are ignored in the catchment modeler community might be, as the reviewer states, laziness or disinterest because simple models "seem to do the job" when the focus is on streamflow. We hence agree that the transfer of an annual phenological cycle from one catchment to another and the assumption of temporal invariance is crude. However, it is not at all unusual practice and often not mentioned in model studies at all. But the reviewer is right we did not give our ET parametrization the same attention as we have done it with other parts of our model. We will follow the reviewer's advice and set up our ET model in a revised manuscript as far as this is possible with observed values. In both catchments we will use the temperature index model from Menzel et al. (2003) to define the start and end of the vegetation period. We use this model because it could successfully identify the tipping points between the summer and winter season in both double mass curves. In the Colpach catchment we have access to observed LAI values for August and September and

we will use them within future simulations. Unfortunately we have no LAI values in the Wollefsbach and hence need to take values from the literature. We will use the reported values for corn by Breuer et al. (2003), which are in fact close to the values we already used. Additionally we will add a table to a revised manuscript were we will list the vegetation parameters from our ET routine.

Reviewer 14) Lines 240-242*: This concept certainly precedes the work of Zehe 2014. For example, consider these important papers, which are notably absent from works cited in your introduction:*

RL: We are sorry that we missed the above mentioned two studies (probably we missed even more) and will consider adding them to our revised introduction. Thank you for these references.

The term “functional units” was first proposed by Zehe et al. (2014) as an as advancement of the HRU concept. The latter is of course much older. The core idea of the functional unit concept is that similarity with respect to the energy balance and runoff formation emerges at different scales (the small field and the hillslope scale), because the land surface and subsurface characteristics controlling the related gradients and “resistance terms” have different characteristic length. The concept that hillslopes are key elements controlling runoff generation in the landscape is of course much older, and goes back to Troch’s Hillslope Boussinesq model (the work of Berne deals with this). However, the term “functional unit” states clearly that one can learn in a representative fashion about runoff formation by targeted clustering of multiple observations at the hillslope scale and that the hillslopes are key building blocks for setting up hydrological models. In fact this is shown within our study. We revise this passage along these lines.

Reviewer 15) Lines 323-324*: Please provide some explanation of what you mean by the onset of vegetation period? Presumably that would be timing of leaf development for crops and deciduous plants, but evergreen plants will be physiologically active even through winter and spring, albeit at lower rates than in summer. Then there is of course an extended period when crops and deciduous plants transition from no foliage to the maximum leaf area they will obtain that year. Than transition can span weeks to more than a month.*

RL: With the term "onset of the vegetation period" we mean the bud break of the deciduous trees as a main seasonal influence on evapotranspiration. We refer here to Beech trees as these are dominant in the Colpach. In the “VE3” we mainly adjusted the day of the first flushing, which of course implies that time to full coverage is the same as in the German Weiherbach.

Reviewer 16) Lines 355-368, and Figure 7: *The variability among measured moisture retention curves for your soil samples is remarkable. Can they be logically grouped in any way, for example, by landscape position or soil depth? If so, a color scheme to illustrate that would be very interesting. Some additional detail about where, and at what depths, the soil samples were taken is needed.*

RL: Yes, the variability is remarkable and this is exactly what we intended to show. Young soils on periglacial slope deposits prevail in the headwater. They exhibit large heterogeneity which cannot be grouped in a simple manner as detailed in Jackisch (2015) and Jackisch et al. (2016). This is due to a) the general mismatch of the scale of 250 mL undisturbed core samples with the relevant flow paths and b) the high content of gravel and voids, which affect the retention curve especially above field capacity and concerning its scaling with available pore space. For the study at hand the focus does not lie on this analysis but on the implications coming from it: A representative retention curve from measurements can be directly and successfully employed in the physically-based model Catflow, even for a heterogeneous headwater catchment like ours. We think that the latter is a noteworthy and non-intuitive finding. To clarify this, we will update the graph adding information about sampling depth and relative distance to the stream network. Moreover we will revise the description accordingly.

Reviewer 17) ***Lines 385-388****: Are the sapflow sensors collocated with rainfall and soil moisture measurements? Please elaborate on the exact type of sensors, depth of installation into trees, and other important details about their operation.*

RL: We will add further information about the sap flow measurements in our manuscript: "Furthermore we use sap flow measurements from 61 trees at 24 of the sensor cluster sites. The measurement technique is based on the heat ratio method (Burgess et al., 2001), sensors are East30Sensors 3-needle sap flow sensors. As a proxy for (the volume of) sap flow we use the maximum sap velocity of the measurements from three xylem depths 5, 18 and 30 mm as recorded by each sensor. To represent the daytime flux, we use 12-h daily means between 8am and 8pm."

Additionally we would like to highlight that the detailed spatial and temporal analyses of this sap flow dataset are currently being revised in a manuscript by Sibylle Hassler. Here we want to use the sap flow data mainly as an additional possibility to test whether our model represents a major hydrological flux which is clearly difficult to measure realistically.

Reviewer 18) Lines 424-426: *Any consideration of the soil-moisture dependence of vapor diffusivity in soil? It varies as a power-law function of air-filled porosity, over about 4 orders of magnitude.*

RL: Catflow doesn't account for movement of water vapor in the pore space. This might be a shortcoming particular in arid areas. We are aware of the work of Chris Milly on this issue.

Reviewer 19) Lines 432-446: *How are the probabilities of the Poisson process determined? Is this based on some knowledge that informs your perceptual model, or are though chosen arbitrarily? Or, do you determine some best-performing parameters based on a sensitivity/optimization process? Please describe in a little more detail, and consider providing an image of what these structures look like in your final model domain. Is this what we're seeing in Figure 3C, D? If so, please just allude to that figure here in the text.*

RL: We again apologize this was not written clearly enough. Normally the values for the probability of the Poison process can be estimated for instance based on spatial mapping of worm burrows (Zehe and Blöschl, 2004). However the parameterization in this study was chosen rather arbitrary since only qualitative information's were available for this parameter. In our discussion paper we followed the studies of Klaus and Zehe (2011) for the Colpach and the parametrization of Wienhöfer and Zehe (2014) for the Wollefsbach. The depth of the vertical macropores of around 1m with a standard deviation of 0.3 m is based on the results of Jackisch et al. (2016). By incidence (or not) this worked well that there was no further sensitivity or optimization process necessary. We will add a reference to Figure 3 C, D in the revised manuscript.

Reviewer 20) Lines 453-454: *Considering the horizontal resolution of model elements is 1-m, I'm wondering how realistically the vertically-oriented, preferential-flow zones can be represented? Certainly the macropores in your photographs are not 1-m wide. Representing their tortuosity by vertically-offset grid cells of 1-m width seems like a gross distortion as well. Can you discuss how you rationalize this model domain and horizontal grid resolution, especially with regard to those grid cells that are imposed to represent preferential flow structures? Also, can you provide detail about the dye-irrigation studies from which the photograph was derived? The pictures are very insightful. However, it is well known that irrigation studies often impose exceptionally high input fluxes, and with sprinkler systems that exhibit enormous spatial variability. Can you comment specifically on the irrigation rates and the spatial uniformity of the irrigation system, and how those irrigation rates compare to the frequency distribution of rainfall intensities that occur at these field sites?*

RL: This is a good point that needs to be explained better. At a grid size of 1 m the Poisson process does often generate several macropores in the same grid cell as the generation process is carried out at a much smaller grid (1-2cm). All macropores contribute to the enhanced conductance of the model element; the enlarged conductance is hence an effective representation of the subscale macropores. This infiltrability is sufficient to allow a fast flow to the bedrock in the Colpach and drainage in vertical and lateral direction in the Wollefsbach. To improve this we usually work with an adaptive grid size of a few centimeters, generate the macropores, and reduce the grid resolution in areas without macropores. This was done in the Wollefsbach and will be done in a revised manuscript in the Colpach.

We employed the same approach in the Colpach in Figure 3, and simulated the runoff of the hydrological year 2014. The result is basically the same as with the coarser grid (Figure 4). This corroborates that the spatial extent of the macropores is not too important as long as their combination represents the total amount of fast flow. In a revised manuscript we will reduce the size of macroporous grid cells in the Colpach from 1 m to 10 cm (Figure 3) similar to the hillslope model used in the Wollefsbach catchment to avoid confusion. We again use the macroporous medium proposed by Wienhöfer and Zehe (2014) which corresponds well with reported maximum velocities from Angermann et al. (2016) in the Colpach catchment. We will further use a fixed distances of 2 m for the lateral distance of the vertical macropores in the Colpach and of 3 m in the Wollefsbach instead of the Poison process to make our study more transparent. We chose this value again rather arbitrary with respect to create an image of the perceptual model and on qualitative information on macropore flow from field experiments (Jackisch et al., 2016).

We can add various model runs where we change the lateral distance of the vertical macropores to our study. But we would like to highlight that such macropore sensitivity studies with Catflow are already published (Klaus and Zehe, 2011; Wienhöfer and Zehe, 2014).

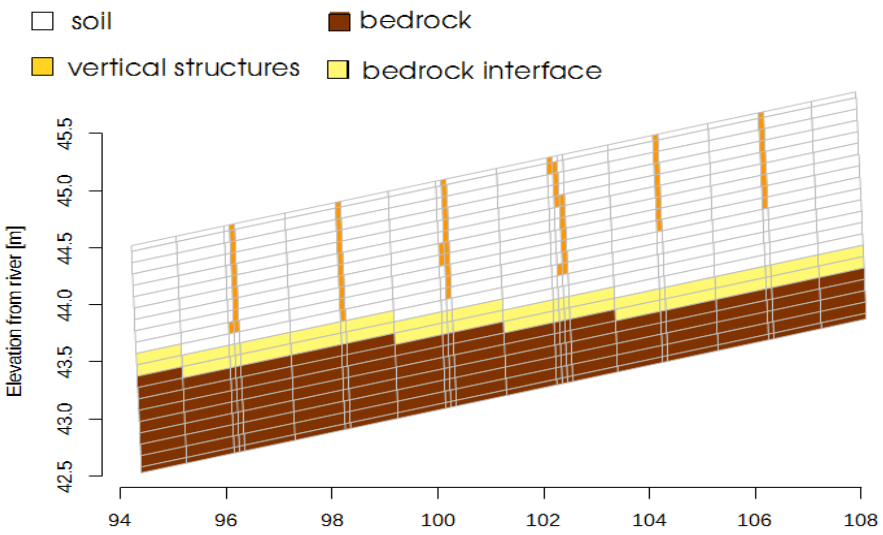

Figure 3 Section of the Colpach hillslope with a reduced grid size for macropores.

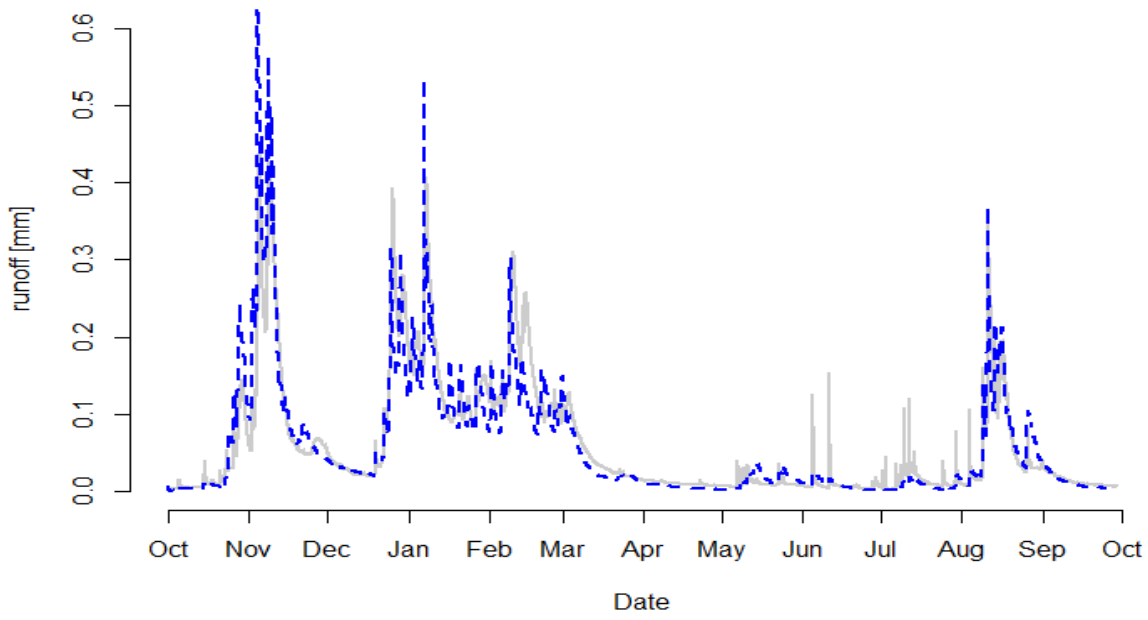

Figure 4 Observed (grey) and simulated discharge (blue) of the reference model with the reduced grid size for macropores.

We will add the following information to the article. The dye tracer images, in Figure 3B and D, were obtained with high rainfall intensities (Jackisch et al., 2016). The aim of these rainfall simulations was to visualize the macropore networks in the topsoil. For actual rainfall events, both in reality as well as in the model the degree of preferential infiltration depend not only on the structure of the macropores but also on the rainfall intensity and the antecedent moisture content.

Reviewer 23) Line 479: *Here again, it would be good to know about the depth distribution of the soil samples collected for hydraulic characterization. Did any actually come from that depth?*

RL: We will provide information about their location and the depth distribution (See technical comment 16).

Reviewer 24) Lines 486-490: *This sounds wonderfully sophisticated. I have no idea what it means. I'm probably not alone in that regard. It seems like important information about how spatial heterogeneity of hydraulic properties are generated in the model domain, so maybe a sentence or two in plain language to build the intuition of the reader/s that are not intimately familiar with this jargon.*

RL: We have adapted the article as follows: **"**We added correlated noise to the hydraulic conductivity**.** To this end we generated a random field of ln(ks) with the observed mean and a variance of 2 using a turning band generator. As we had no local information on the correlation length, we used a range of 5 m which corresponds to the range of soil moisture observations (Zehe et al., 2010) found for a distributed soil moisture network in a forested site in the Ore Mountains. We used a spherical variogram function with a nugget of 0.5 and a sill of 1.5."

Since the variability in the soil retention properties cannot be easily explained by their position and depth it is not straight forward to implement different soil layers in the hillslope model (See technical comment 16). But we do know from numerous experiments that the skeleton fraction in the Colpach is rather high with values above 50% in deeper soil layers. In the revision process we will try a model run with reduced porosity in the deeper soil layers.

Reviewer 26) Line 527-530: *You're simulating flow through a single 2-dimensional hillslope profile, so how can you compare the composite discharge (overland flow, subsurface flow, and deep drainage) to the measured streamflow for the whole 3-d catchment? Are you integrating the hillslope response over the entire 3rd dimension of the catchment? Please explain this. Also, do you think it is appropriate to include the deep drainage flux in this composite outflow when comparing to the stream hydrograph? It could conceivably travel through an aquifer system that discharges outside the boundaries of your catchment, no?*

RL: We compare specific discharge observations and specific flow simulated with Catflow, by normalizing the former with the catchment area and the latter with the hillslope area mm ((l*h)/m^2). We will add a short explanation. Deep percolation is included as stated above but does not change the result significantly since it is quite low (0.001% of the overall discharge). We agree with the second point.

Reviewer 29) Line 566-568: *The important trend for the catchment water balance is the timing of the leaf area expansion, maximum, and decline (in fall). The leaf area is the dominant control on transpiration and net radiation. So, I don't understand how or why you change the timing of phenology without changing the temporal dynamics of leaf area. Please explain the rationale for this*

RL: Sorry for being unprecise here. We shifted the start and the end of the LAI cycle from the original values to the values predicted by the temperature index, while not changing the LAI values. If both the start and endpoint are shifted, the cycle within remains the same. Since this is a virtual experiment, this part will be removed in a revised manuscript and the vegetation will be parameterized differently in our model (see technical comment 12).

Reviewer 30) Line 594: *Figure 9B, rather than 10B?*

RL: We refer to the simulated saturation patterns in 10 b showing the 2 d pattern in summer and winter. We will stress this.

Reviewer 31) Lines 614-615: *Are you talking about Weierbach catchment in Germany? It's irrelevant. I suggest you delete that and stay focused on your catchment*

RL: We are sorry. It is a little confusing. We talk about the Weierbach catchment which is a headwater of the Colpach in Luxembourg and thereby in the same hydrological landscape. This transferability is hence relevant and corroborates the hypothesis on functional units. The Weiherbach in Germany is mostly written with an "h". We will clarify this in our manuscript and apologize again.

Reviewer 32) Line 619-622: *Run-on sentence that is very hard to interpret. Please rephrase. Also, please explain what inference you think is made possible by comparing NSE with log(NSE).*

RL: We will rephrase this passage- The NSE was chosen because it is a common measure for the quality of model results in hydrology with regard to high flows; logNSE a better quality measure for low flows.

Reviewer 33) Lines 622-627*: These statements are questionable. Are you claiming that infiltration-excess overland flow is occurring, or overland flow due to saturation excess? You say that the model erroneously generates overland flow in the summer in Wollefsbach due to convective storms. The*

saturated-hydraulic-conductivity parameter you report in Table 1 is 2.9 x 10-4 m/s, or about 1.8 cm/minute. This is quite a high hydraulic conductivity what one might expect for coarse sand. Are your surficial soils sandy? Are those convective storms producing rainfall flux greater than 1.8 cm/minute? Seems doubtful storms like that would occur frequently. How is the model generating so much overland flow if the rainfall rates are (I assume) always considerably lower than the saturated conductivity? In Figure 9D it looks like your simulated-average-soil moisture is in excess of almost all the measured time series. Maybe you're way overestimating soil-water storage and generating saturation-excess overland flow, rather than infiltration excess overland flow. If that's the case, and if it's saturation-excess overland flow, then you can't immediately assume that enhancing Ksat to represent soil cracks is the next, necessary step to improving the model. You need to get the soil moisture dynamics correct before you can go off exploring that speculation. You're Ksat value is already pretty high, is it based on measurements? Also, it's somewhat a shame that you have all those soil moisture observations, and a spatially-distributed model, but you only compare the mean-simulated soil moisture (assume it's the mean) to the observations. The spatially-distributed model gives you lots of spatially-distributed results to compare to spatially-distributed observations. If you're just going to look at the average-simulated soil moisture, you're giving up all that detail that is provided by the model, and one has to ask why not just use a lumped model? You should compare simulated soil moisture at specific points in the landscape to the observed soil moisture at those same points

RL: Good point, we agree that ksat is large, though it has been measured like that. Hortonian overland flow occurs in desert catchments, because of wetting problems due to the extreme dryness, although ksat is even larger there. So speaking about the real system, Hortonian overland flow definitely occurs in the Wollefsbach catchment, which is also visible in frequent erosion events (Martínez-Carreras et al., 2012). With respect to the generation of Hortonian overland flow within the model is not ksat that counts but ku(theta). Note that the observed rolling mean of the topsoil moisture and the average of the simulation in summer are between 0.3 and 0.2 (Figure 9 c in the Discussion paper). This implies an unsaturated hydraulic conductivity of 3.5 10^{-11} m/s (compare Figure 5). With such a value the system definitely develops Hortonian overland flow (both the model and the real system). The key question (in real systems) is whether locally generated infiltration excess reaches the stream or not. The latter depends on the question whether a connected flow path of low infiltrability exists, or not. We admit that a proper investigation of overland flow path connectivity is not within the scope of a 2 d hillslope model.

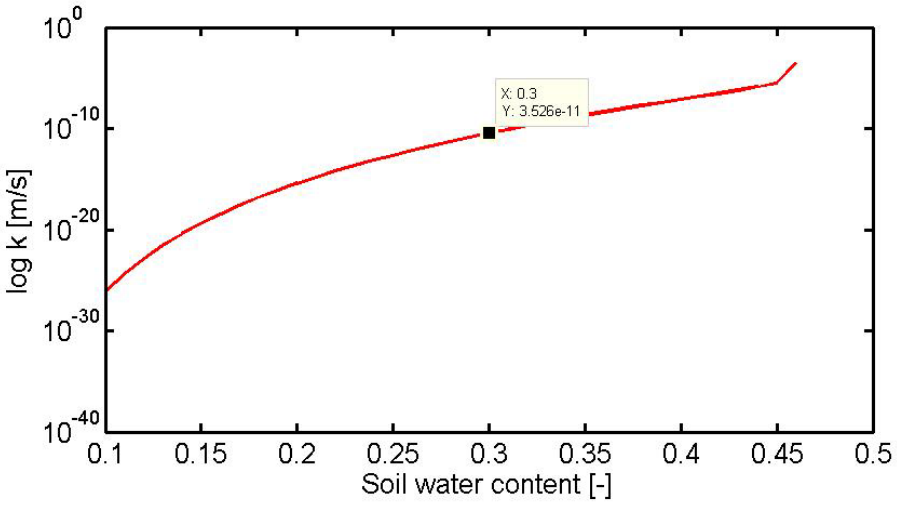

Figure 5 : Unsaturated hydraulic conductivity of the Marl soil at a value of 0.3 volumetric water content.

Reviewer 34) Lines 631-635: *I am quite confused by this statement. You are using a spatially-distributed model, so why are you claiming that it's unrealistic to expect the model to accurately represent the spatially-distributed nature of soil-moisture dynamics? It should be able to represent at least coarsely the spatial distribution of soil moisture, for example, differences in upslope versus downslope positions, or differences in areas overlying saturated bedrock depressions versus those areas where the bedrock roughly parallels the land surface. Again, if you don't expect your spatially-distributed model to accurately represent any of these spatial patterns (which are important for runoff), then why are you using a spatially-distributed model to begin with?*

RL: Thanks for this important point and sorry for being unprecise. We have of course a spatially distributed model along the hillslope. We believe that the full spatial variability of the soil moisture data is caused by 1.) spatial heterogeneity of rainfall, 2.) heterogeneity of soil properties and 3.) variability along the hillslope. Although our model cannot account yet for 1.) and 2.), we already account for the variability along the hillslope. In the revised manuscript we will use virtual observations along the hillslope in 10 and 50 cm depth (Figure 6). By doing so we can better account for the variability of the soil moisture observations in our 2 d profile (Figure 7). As we are not modelling one particular hillslope, we think a site-specific comparison is out of scope. Nevertheless we will try some color coding for the position of the soil moisture observation as well as for the virtual observation. But we would like to stress that the soil moisture observation similar to the soil water retention properties are not easy to classify by different landscape positions (For example by: up- or downslope).

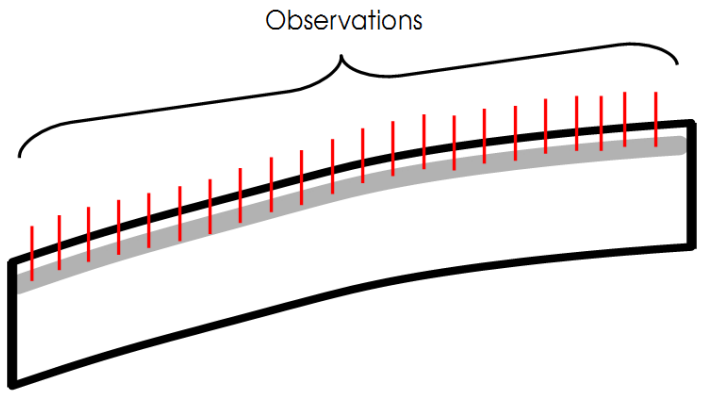

Figure 6 Sketch of our 2d representative hillslope with 20 virtual observation points. The position in lateral direction was chosen randomly the position in vertical direction is 10 cm.

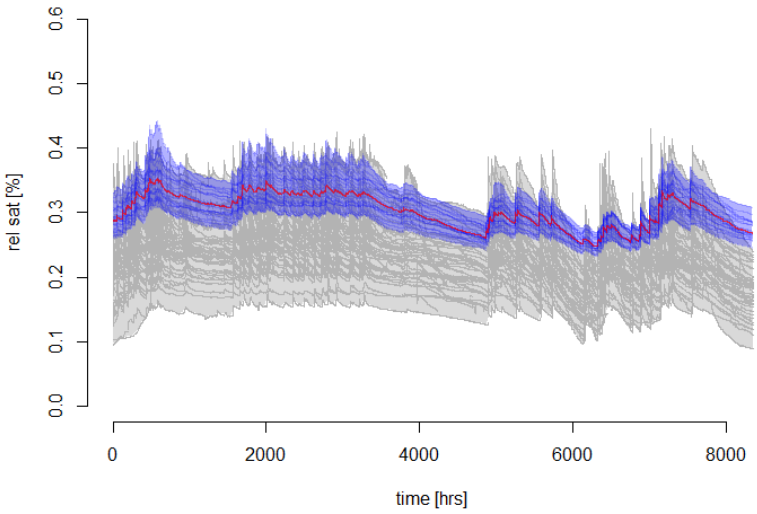

Figure 7 Top soil moisture observation in 10 cm depth against an ensemble of 20 virtual observations in the respective depth in vertical direction and randomly chosen position in lateral direction.

In principle it is of course possible to account for the spatial heterogeneity of soil properties by means of random fields, as proposed by the reviewer. To illustrate this we generated a random field of porosities with an unconditional, sequential gaussian simulation using the R package “gstat”. One may also optionally account for a reduced porosity of deeper soils layers by reducing the mean porosity with depth. Figure 8 corroborates that a simulation within such a heterogeneous domain could resemble most of the variance of the soil moisture observations.

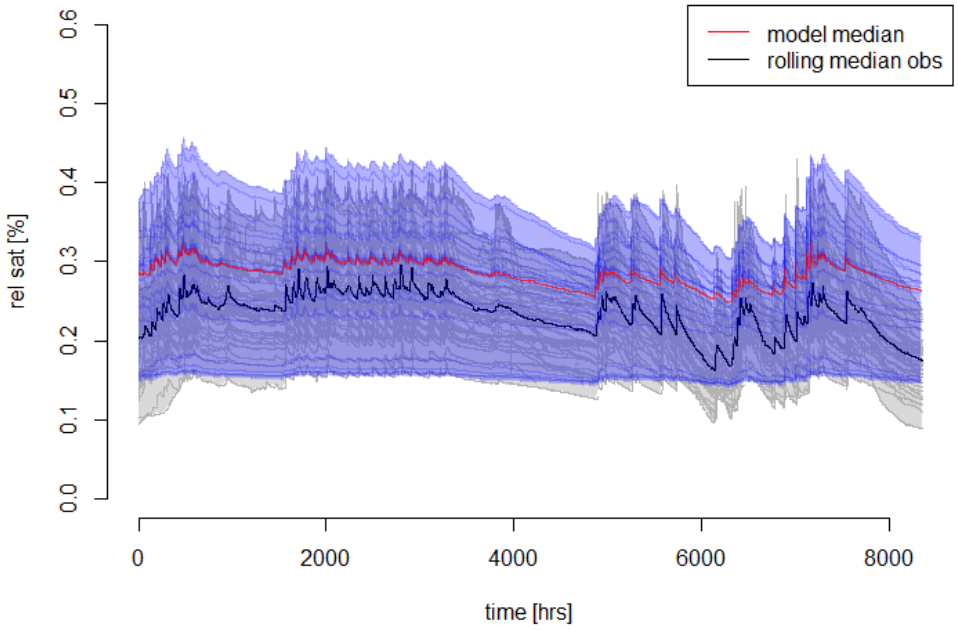

Figure 8 Observed top soil moisture observation in 10 cm depth as well as an ensemble of 20 virtual observations in the respectively depth in vertical direction and randomly chosen position in lateral direction. The soil parameter were varied in the observed range using random fields with a linear trend.

Reviewer 35) Lines 641-642: *This statement is inevitably true for every catchment in the world, and hence does not rely on any measured or simulated soil moisture dynamics. Maybe just delete.*

RL: Sorry for being unprecise. We meant that runoff generation in the Marls (Wollefsbach) is partly intensity-controlled due to the occurrence of Hortonian overland flow (as outlined above). We remove this statement to stay brief.

Reviewer 36) Lines 671-684, and Figure 10A: *This material needs much improvement. First, measuring sapflow in trees is a delicate business, with major discrepancies existing between methodologies, and significant errors arising from inexact application of methods (e.g. due to radially-varying flux rates within the sapwood, a well-documented phenomenon in the tree physiology literature). There are several reviews of this topic in the plant physiology literature (e.g. Steppe et al. 2010, Agricultural and Forest Meteorology, 150). You have provided essentially no detail about the nature of your field-based sapflow measurements. Also, how are you normalizing the measurements? Second, you use a version of the PM model to simulate water vapor flux from the plant canopy, not sapflow (L3 T-1). The two cannot be assumed to be equal. If you consider the tree as a system spanning the point of your sapflux measurement (breast height on the stem) to the canopy, then the sapflow (input to the system) only equals the volumetric flow out of the leaves (outflow from the system) if the system is in steady state (i.e. inflow = outflow and storage is constant). Water storage in the tree stems and canopy foliage is dynamic. I am not sure how good or bad is the assumption of steady state in your system, but it is*

certainly an assumption you should carefully consider and provide some justification for why this comparison (between measured sapflow and modeled canopy vapor flux) is valid. Without that, you should probably omit this text and figure 10A from the manuscript.

RL: We agree with the reviewer that sap flow measurements are delicate and not directly comparable with simulated ET from PM models. We will stress this in the revised manuscript. But we want to highlight that an exhaustive discussion of all related uncertainties that go along with all the observations (soil moisture, discharge, rainfall, soil water retention functions, sap flow, ERT measurements, etc.) we use might be out of scope.

If our sap flow observations were trustworthy, they cannot be directly compared to PM simulations results, as the former is a velocity and the latter is a normalized flow. This is in fact why we a) normalized both observed sap flow and ET by dividing their values by their range and do in fact only discuss the correlation among the normalized values. We still think that this comparison provides added value because it yields the models deficiencies and capabilities to match sap flow dynamics, and whether the maximum and minimum values coincide. We will better explain this in the revised manuscript. Furthermore we will update Figure 10 in the manuscript and show the ensemble of all 61 sap flow observation similar to our soil moisture plots.

Reviewer 37) Lines 691-692: *You might be careful in projecting your own expectations, surprises, and uncertainties onto your readers. The result you describe here is not counterintuitive to me; it's exactly what I would expect, for the exactly the reason you state. Higher gradient = more rapid drainage = less persistent storage (assuming all else is equal, which is what you've assumed for this virtual experiment)*

RL: We partly disagree with the reviewer. When presenting our results at the latest EGU conference, people were wondering if our hillslope model would also work acceptable if we change the hillslope topography. In fact this is not the case, as we show in the virtual experiment number 1. We will avoid the term astonishing and remove this part from our revised manuscript since it is a virtual experiment.

Reviewer 39) Table 2*: Please provide some rationale for why you use multiple error metrics (e.g. NSE, KGE, logNSE) instead of just one. It's just confusing to the reader when you talk about quality of results in one case using KGE, and in another case using NSE. Also, the different metrics show different sensitivities to the domain-changes utilized in the virtual experiments. Why? Which one is most appropriate in light of those differences? You're using these various metrics to make inference about the relative importance of different model features, so you need to argue why one or the other metric is better. Or just use one metric for clarity.*

RL: We believe that it is good practice to calculate multiple error metrics when showing model results. Every error metric or objective function has its advantages and disadvantages. There is a long discussion about this in hydrology (e.g. Kling and Gupta, 2009; Schaefli and Gupta, 2007). In the revised manuscript

we will shortly explain the advantages and different sensitivities of the different metrics and stress why we use different ones (see also Comment 32).

Reviewer 44) Lines 780-781*: Why is that remarkable? It's a predominantly upland catchment with forested hillslopes. Was it your initial expectation (null hypothesis) that the model would be incapable of simulating streamflow?*

RL: At least we found it remarkable that a catchment of 20 km^2 can be represented using a single hillslope (at least to some degree).

Generally, it is an interesting point why we were surprised. We used a physically-based hillslope model and parameterize the model with measured data when possible and take values from other studies and regions when we had no measurements. By doing so we are able to simulate the water balance and the streamflow of two lower mesoscale catchments to a certain extent. It is always subjective if and when a model simulation performs well and when it doesn't, especially considering that hydrological modelling studies are often rather data mining approaches that show only tables of objective functions to demonstrate that their model is working well. But keeping the numerous studies in hydrology in mind which state that a hydrological model cannot be parameterized by measurements, our results were at least for us surprising. We were also surprised that our model can mimic the dynamic of the sap flow observations even if they are not too accurate and our evapotranspiration routine is not state of the art. And yes, we were also surprised that our soil moisture simulations are within the margin of observation and not too far off the 12-hour rolling median of the soil moisture observations.

We are sorry for using the word 'surprised' but we had issues to find papers where the authors tested their models in such an extensive data driven way? We had the feeling that the catchment hydrology community agreed that we cannot use measurements to set up physically-based models and actually use them because of the well documented limitations.

Reviewer 47) Lines 802-804: *Are you so sure this can be concluded? It seems to me that if you want to advocate the use of highly parameterized, spatially-distributed models for the sake of learning about catchments, you need to illustrate that the model is accurately representing some of the spatial dynamics in the hillslope (or catchment that is a composite of your hillslopes). In those cases the matching between simulated and observed averages is not that great for soil moisture there are systematic errors in all cases (Figure 9A-D). By comparing average-simulated soil moisture for the whole hillslope to the average-observed soil moisture for the whole catchment, you're failing to rigorously test the spatially-explicit predictions made possible by the model. You should try to show that the model actually properly represents spatial variability in soil moisture, saturated-zone expansion, hydraulic gradients, etc. If not, then it's hard to argue that the distributed model teaches us anything more than we would learn from a lumped model.*

RL: We agree that this needs to be further specified. In fact we "only" showed that the model works well for runoff. The setup can obviously be improved to better reproduce soil moisture dynamics, by perturbing for instance porosity to match the observed variability of soil moisture. We will revise the conclusions accordingly and carry out a more spatially distributed comparison of simulated and observed soil moisture.

Reviewer 48) Lines 812-816: *I fully agree with this statement. You don't need a high-dimensional, spatially-distributed model if all you want to do is predict runoff at an annual timescale, or even at shorter time scales. Use a transfer function, maybe even a time variable transfer function you will still have vastly fewer degrees of freedom than in the spatially-distributed Richards equation model. But doesn't this statement contradict the overall message of your paper, that those more complex models are needed for learning about catchment functioning?*

RL: No, again we believe that a successful simulation of the catchment functioning does not end with successful simulation of the runoff. It is quite difficult to estimate soil moisture and evapotranspiration with a transfer function. The second point is that we use an a priori model setup that was not calibrated automatically, but based on several observations which are independent of discharge. Particularly the latter is not possible with a transfer function approach, which can only be calibrated using discharge data. But again you are right our language is not precise and we need to clarify this in the revised manuscript.

Reviewer 49)*: The spatial variability of soil-hydraulic properties may be quite important, in fact, for properly simulating all those runoff peaks in the summer, where your model does quite poorly (Figure 8B,D, and Figure 12B)*

RL: This might be the case. But it could also be that steep runoff peaks are generated by forest roads and paved areas in the catchment. Furthermore, most of the runoff is produced in winter. That means, around 90% of the overall runoff within a hydrological year is not primarily controlled by soil heterogeneity of soil-hydraulic properties since our model does well with respect to runoff in winter. Please note that we do not consider spatial variability of soil properties unimportant. We simply want to stress that the proposed approach to define representative soil hydraulic functions works acceptable. We will rewrite this in a revised manuscript.

Reviewer 50) Line 823-844: *I would suggest you delete all of this. It seems wildly speculative and I have no idea how, based on the analyses performed in this paper, you conclude that equifinality and the concept of a representative hillslope is rather more a blessing than a curse since there is an infinite number of possible macropore setups which yield the same runoff characteristics. If this were not the case, we could not transfer macropore setups from the literature across system borders and successfully simulate two distinct runoff regimes which are strongly influenced by preferential flow." An infinite*

number of macropore setups that yield the same runoff characteristic. What are you talking about here? You tested 2 such scenarios (Figure 3C, D). Your model does a fairly poor job at matching runoff peaks at many times of year. Those peaks are the hydrological attribute most likely to be influenced by the activation or latency of preferential flow paths. When you say you that you "successfully simulate two runoff regimes" I presume you are talking about the double-mass curves, because your simulated hydrographs show significant errors at many times of the years.

RL: The reviewer is right that the section is not entirely supported by the evidence provided in this paper and will be removed. However, the above mentioned studies by Klaus and Zehe (2010) and Wienhöfer and Zehe (2014) show that more than a single macropore network architecture is capable to yield the same simulate runoff (Equifinality). This is because the total amount of fast subsurface flow is jointly determined by the macropore density and their hydraulic conductance. Note that Klaus and Zehe (2010) varied both macropore density and their conductance within the range of observed values, and found 13 out of 420 setups that simulated tile train discharge in the same manner. This is of course not an infinite number.

The points we wanted to stress here is that the macroporous medium and the setup Wienhöfer and Zehe (2014) used in their study in Austria also improved the model performance significantly in the Wollefsbach and in fact also in the Colpach catchment. We think that the fact that several parameterizations of the macropore network work equally well is maybe an advantage rather than a problem. Simply because we cannot measure the real macropore network in the catchment nor could we map the real setup in our model or any model. That's what we mean when we write that equifinality is rather a blessing than a curse in the case of physically-based models.

Reviewer 51) Line 864-866: *Do you mean, "below which", instead of "above which"? I wouldn't spend much time on that. By changing slope and nothing else, you're vastly oversimplifying how soils, geology, and geomorphology affect streamflow, and how all those variables are related to topography in naturally evolving landscapes.*

RL: In the virtual experiment we doubled the gradient, in fact we picked on of the steepest slopes, with no effect. When changing the gradient to slopes much less steep, we expect that this will have an effect, because water flows slower through the fast lateral flow path. Hence we expect a threshold below which a reduction of the gradient starts to matter. This will not be part of a revised manuscript.

Reviewer 52) Lines 888-889*: How do you justify this statement? Did all of your virtual experiments where you manipulate the bedrock topography have an equal volume of depressions? And if so, how do you go about quantifying the volume of depressions in an undulating rock surface? What constitutes a depression or a high point, versus a portion of the rock surface that is part of a datum plane?*

RL: The reference slope, the slope without bedrock interface and the slope with the riparian zone (VE 2.3) had the same volume of depressions. The number of depressions is equal to the number of local

maxima in the bedrock topography. The depression volume is then simply counting the number of grid cells upslope a local minimum with an elevation above the local bedrock interface but below the next downslope maximum. As the reference hillslope and the VE2.3 have the storage volume (Figure 9), but differ with respect to its distribution, we think this statement is supported. This part will be removed in a revised manuscript.

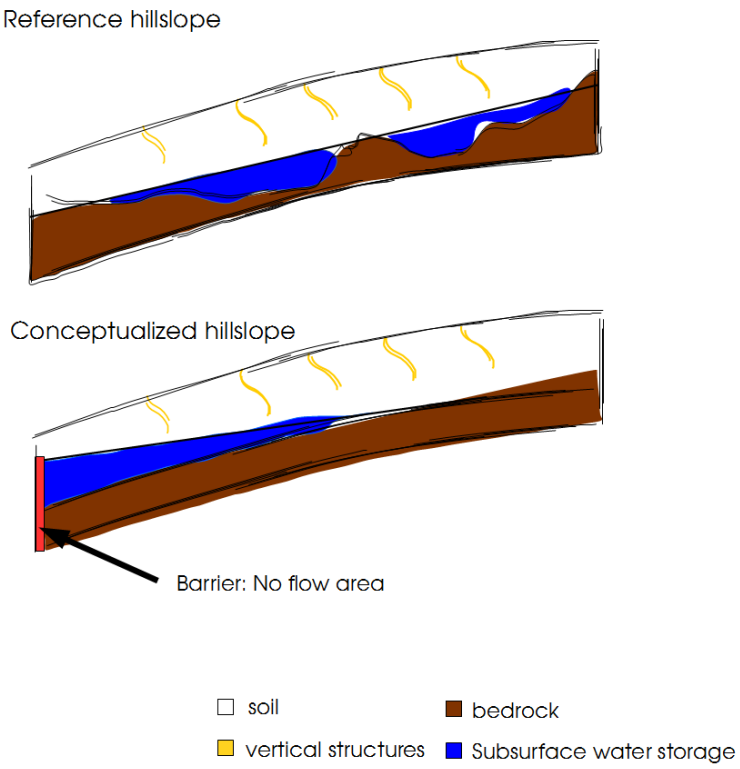

Figure 9 Sketch of the reference hillslope and of the conceptualized hillslope with the barrier (no flow area) at the left hillslope border.

Reviewer 53) Lines 867 and 895: *You're using the questions as section headers, but the subsequent content does not answer the questions. First of all, define what you mean by "first order control". Do you just mean that the response variable (annual runoff ratio, or other?) is a linear function of the independent variable (bedrock topography, or vegetation)? If so, do your data corroborate such a linear relationship? I'm not sure you can say, based on the limited scenarios of bedrock topography you tried. You would have to come up with some quantitative metric distinguishing one bedrock scenario from another. In terms of vegetation, all you did was try 2 different times for bud break. Can you discern a linear relationship between "vegetation" and some response variable based on these tests?*

RL: The term first-order control is often used to specify the most sensitive parameter or information source – or which of the parameter or information sources contributes most to the explained variance and thus to an error function. Simulations without macropores and the fast bedrock interface (using the same bedrock topographie) reduces the KGE by 0.12, while an additional removal of the bedrock topography reduces the KGE to 0.59. This supports that bedrock topography is a first-order control in

respect to the runoff formation. We will leave this out in a revised manuscript since it belongs to the virtual experiment.

Reviewer: 54) Lines 912-923: *Of course you can! You can setup heterogeneous rainfall inputs at the soil surface in your model domain. You can setup different scenarios of incoming solar radiation along the hillslope domain to emulate aspect-related differences in the radiation balance, water budget, and possibly soil hydraulic/or geological characteristics. I'm really struggling to understand how you continually advocate for spatially distributed models, but continually state that one can't expect them to accurately represent spatially-explicit hydrological processes. If you don't expect a spatially distributed model to accurately represent spatially-explicit hydrological processes, then why use it, instead of a lumped model?*

RL: Here we discuss the limitations of using a single hillslope for a catchment which consists of numerous hillslopes. Much confusion arises from our fuzzy use of the term "distributed" and "single hillslope". The representative hillslope is distributed along the slope line. Of course we can add distributed rainfall to the hillslope and allow for perturbations of soil parameters. However, the characteristic length of rainfall variability in the Colpach is larger than the total extent of our hillslope, similarly the variability of radiation depends not only on slope but also on aspect and landuse. A representation of the full sources of variability requires a model setup consisting of all the hillslopes in the catchment and their interconnecting river network. Catflow allows for this, as for instance shown in Zehe et al., (2001), but this was not the goal in our study here. We will better explain this part in the revised manuscript.

The reviewer is generally right that we waste too much effort on discussing the limitations of the Richards approach. We will change this tenor in the revised manuscript.

Rewiewer 55) Lines 937-969: *I would suggest deleting this to shorten and focus your discussion. There is not much reference to your analysis in this section, it's a little bit of a ramble, and you don't say anything about the Richards equations that hasn't already been said many times before over the last 70-80 years.*

RL: Good point, we will remove most of this part, but keep the part on how to deal with emergent soil structures.

Reviewer 56) Lines 971-993: *I recommend you remove this from the manuscript. You're pontificating about all the ways the land surface models must be fundamentally improved for hydrological modeling, and in doing so you're demonstrating that you have no awareness of the related disciplines of hydrometeorology, plant physiology, and biophysics. All the phenomena that you imply are important, for example, "implies that phenology evolves in response to climate and hydrological controls, thereby creating feedbacks" are in fact known to be important by people in those fields, and others. The upgrades to our model representations that you suggest should happen, have in fact happened, and continue to be upgraded, for example, models that link plant metabolism and water use, or that utilize spatially- and*

temporally-dynamic root uptake schemes. You conclude by saying that the literature is full of more realistic models for parameterizing stomatal conductance, but you still use a fairly standard version of the PM model with non-local parameters. So I don't think your analyses are very relevant to the state of practice in evapotranspiration modeling. I think you should stay focused on the representation of runoff processes.

RL: We will remove this passage. The reviewer is right, that there are much more sophisticated approaches for ET available.

Reference

Angermann, L., Jackisch, C., Allroggen, N., Sprenger, M., Zehe, E., Tronicke, J., Weiler, M., Blume, T., 2016. In situ investigation of rapid subsurface flow: Temporal dynamics and catchment-scale implication. Hydrol. Earth Syst. Sci. Discuss. 2016, 1–34. doi:10.5194/hess-2016-189

Bos, R. van den, Hoffmann, L., Juilleret, J., Matgen, P., Pfister, L., 1996. Conceptual modelling of individual HRU ' s as a trade-off between bottom-up and top-down modelling , a case study ., in: Conf. Environmental Modelling and Software. Proc. 3rd Biennal Meeting of the International Environmental Modelling and Software Society. Vermont, USA.

Breuer, L., Eckhardt, K., Frede, H.-G., 2003. Plant parameter values for models in temperate climates. Ecol. Modell. 169, 237–293. doi:10.1016/S0304-3800(03)00274-6

Dooge, J., 1986. Looking for hydrologic laws. Water Resour. Res. 22, 46S–58S. doi:10.1029/WR022i09Sp0046S

Gupta, H. V., Clark, M.P., Vrugt, J. a., Abramowitz, G., Ye, M., 2012. Towards a comprehensive assessment of model structural adequacy. Water Resour. Res. 48, 1–16. doi:10.1029/2011WR011044

Jackisch, C., Angermann, L., Allroggen, N., Sprenger, M., Blume, T., Weiler, M., Tronicke, J., Zehe, E., 2016. In situ investigation of rapid subsurface flow: Identification of relevant spatial structures beyond heterogeneity. Hydrol. Earth Syst. Sci. Discuss. 2016, 1–32. doi:10.5194/hess-2016-190

Klaus, J., Zehe, E., 2011. A novel explicit approach to model bromide and pesticide transport in connected soil structures. Hydrol. Earth Syst. Sci. 15, 2127–2144. doi:10.5194/hess-15-2127-2011

Klaus, J., Zehe, E., 2010. Modelling rapid flow response of a tile-drained field site using a 2D physically based model: assessment of "equifinal" model setups. Hydrol. Process. 24, 1595–1609. doi:10.1002/hyp.7687

Kling, H., Gupta, H., 2009. On the development of regionalization relationships for lumped watershed models: The impact of ignoring sub-basin scale variability. J. Hydrol. 373, 337–351. doi:10.1016/j.jhydrol.2009.04.031

Martínez-Carreras, N., Krein, A., Gallart, F., Iffly, J.-F., Hissler, C., Pfister, L., Hoffmann, L., Owens, P.N., 2012. The Influence of Sediment Sources and Hydrologic Events on the Nutrient and Metal Content of Fine-Grained Sediments (Attert River Basin, Luxembourg). Water, Air, Soil Pollut. 223, 5685–5705. doi:10.1007/s11270-012-1307-1

Menzel, A., Jakobi, G., Ahas, R., Scheifinger, H., Estrella, N., 2003. Variations of the climatological growing season (1951-2000) in Germany compared with other countries. Int. J. Climatol. 23, 793–812. doi:10.1002/joc.915

Schaefli, B., Gupta, H. V, 2007. Do Nash values have value? Hydrol. Process. 21, 2075–2080.

Wienhöfer, J., Zehe, E., 2014. Predicting subsurface stormflow response of a forested hillslope – the role of connected flow paths. Hydrol. Earth Syst. Sci. 18, 121–138. doi:10.5194/hess-18-121-2014

Wrede, S., Fenicia, F., Martínez-Carreras, N., Juilleret, J., Hissler, C., Krein, A., Savenije, H.H.G., Uhlenbrook, S., Kavetski, D., Pfister, L., 2015. Towards more systematic perceptual model development: a case study using 3 Luxembourgish catchments. Hydrol. Process. 29, 2731–2750. doi:10.1002/hyp.10393

Zehe, E., Blöschl, G., 2004. Predictability of hydrologic response at the plot and catchment scales: Role of initial conditions. Water Resour. Res. 40, 1–21. doi:10.1029/2003WR002869

Zehe, E., Ehret, U., Pfister, L., Blume, T., Schröder, B., Westhoff, M., Jackisch, C., Schymanski, S.J., Weiler, M., Schulz, K., Allroggen, N., Tronicke, J., Dietrich, P., Scherer, U., Eccard, J., Wulfmeyer, V., Kleidon, A., 2014. HESS Opinions: Functional units: a novel framework to explore the link between spatial organization and hydrological functioning of intermediate scale catchments. Hydrol. Earth Syst. Sci. Discuss. 11, 3249–3313. doi:10.5194/hessd-11-3249-2014

Zehe, E., Graeff, T., Morgner, M., Bauer, A., Bronstert, A., 2010. Plot and field scale soil moisture dynamics and subsurface wetness control on runoff generation in a headwater in the Ore Mountains. Hydrol. Earth Syst. Sci. 14, 873–889. doi:10.5194/hess-14-873-2010

Zehe, E., Maurer, T., Ihringer, J., Plate, E., 2001. Modeling water flow and mass transport in a loess catchment. Phys. Chem. Earth, Part B Hydrol. Ocean. Atmos. 26, 487–507. doi:10.1016/S1464-1909(01)00041-7