Reply to review by Anonymous Referee #1

We thank the reviewer for his/her careful read to the manuscript and insightful suggestions.

After reading carefully the comments of both the reviewers, in order to address some of their most substantial comments, there will be some major changes in the revised version of the manuscript. These changes will affect most significantly the data analysis part of the paper (section 3), and they are summarized hereafter:

- We will extend the selection of hydrological signatures in order to have a more comprehensive picture of the hydrological behavior of the catchments.
- We will select only signatures that are not correlated (reducing substantially their number) for the subsequent analyses.
- We will select a reduced number of catchment and meteorological indices in order to reduce the problem of correlated features.
- All the correlation analysis will be based on the Spearman's rank score since, as pointed out by Dr. Melsen, the usage of Pearson may not be adequate.
- Following this decision, we will remove the regression analysis (table 3) since (1), as stated by the Anonymous Referee #1, " it does not add any new knowledge to the literature occupies a lot of space" and because (2) the usage of linear regression (that looks for linear dependencies in the data) is not coherent with the choice of basing the analysis on the Spearman's rank score (which accounts also for nonlinear correlations).

Below, we answer in detail the various comments, and illustrate our intended approach to address them in the revised version. The original comments of the reviewer are reported in *black and italics*, our replies in blue.

The authors propose to infer the structure of a hydrological model based on landscape and process characteristics (signatures) of the catchment. In the first of a two-stage process different landscape and catchment characteristics are compared to different streamflow signatures to identify the most important controls on runoff formation. In the second step this information is used "as an inspiration for model structure design" (p17, l. 32) as the authors put it.

Inferring structure from function (or vice versa) is at the core of hydrological model building and subject to numerous studies. The topic is hence highly relevant for the hydrological community. The manuscript is well structured and well-written. Accordingly the manuscript is suitable for a publication on HESS. However, I cannot recommend publishing to current version of the manuscript due to several major points:

1. *The purpose of the modelling exercise is not clear. Model requirements for flood forecasting are e.g. totally different from model requirements to simulate climate change. The relevant*

signatures, temporal and spatial model discretisation, model evaluation metrics and also the degree of model conceptualisation differ accordingly. Please specify more clearly the purpose of you modelling study. Otherwise it is not possible to evaluate the study meaningfully.

We will complement the objectives of the studies, already summarized in the introduction (lines 6-10 page 3) with the objectives of the modeling exercise. In the introduction, we already specified the main objectives of the study, which in summary consist of proposing a model building strategy which starts from an analysis of the data, which provide a basis for motivating the various model decisions.

However, as the reviewer noted, the purpose of the model itself has remained unclear. Although one of the main objectives of models is making predictions to address some practical issues, here we do not have such an immediate applied objective. The model exercise is mainly an instrument to help understand and interpret catchment scale processes. In particular, we are interested in identifying the landscape properties and associated processes that dominate catchment response, and that mostly influence the observed spatial variability in streamflow behavior, as characterized by the set of suggested signatures.

The reviewer is right in the fact that aspects such as "signatures, temporal and spatial model discretization, model evaluation metrics and also the degree of model conceptualization differ accordingly". We will clarify that there the objective is explaining the observed spatial diversity of streamflow characteristics, with the minimum possible complexity, while trying to maintain a process based interpretation.

2. *I consider the selection, evaluation and identification of landscape characteristics as fairly weak due to a number of different reasons:*
    a. *The authors provide no information about why certain characteristics were selected (and why others were not). Catchment characteristics (or signatures) can only provide information on the underlying processes if they have some kind of diagnostic potential or causal relationship. It is clear that these relationships are often unknown and difficult to obtain; nevertheless the selection of appropriate characteristics is vital for the identification of underlying processes and mechanisms. I miss a clear and elaborate description on the selection of catchment descriptors and on their expected diagnostic potential (both in space and time): E.g. why or how can the different land cover ratios or aspects help to derive information on hydrological processes? Are the same characteristics suitable for all catchments (independent of size, altitude, geology)? Please also comment on the importance of the time step e.g. you calculated the flashiness index based on daily streamflow data, although you state that streamflow can change two orders of magnitude in a few hours (p. 3 l. 18). If this is true please explain why you consider a daily-data based flashiness index as a meaningful variable? Please do also explain why you think that "half streamflow period" is a suitable parameter to discriminate to importance of snow. I expect that there are much simpler and more meaningful variables such as temperature and rain, temperature sums or snow data itself to describe the importance of snow. The results in Fig 7-9 also show that streamflow, runoff coefficient and half streamflow period are pretty identically in all*

cases. Do you consider them being suitable signatures? Please also provide more signature papers in the introduction as the number of up to date references is small.

The reviewer correctly points out that we "miss a clear and elaborate description on the selection of catchment descriptors and on their expected diagnostic potential (both in space and time)". We will complement the selection of catchment characteristics with their expected diagnostic potential. For example, vegetation characteristics are typically assumed to affect evaporation, soil characteristics are typically assumed to influence the partitioning of water between retention and runoff. In general, we tried to select a broad class of characteristics, to be as inclusive as possible. However, it is also true that these characteristics can be represented through a large class of indices, and in order to reduce the size of the problem, some choices had to be made. We will motivate some of our decisions based on how other models have dealt with similar issues, so that some of the choices will appear less ad hoc.

We used a daily data resolution, and this choice clearly affects some of the signatures. As the reviewer points out, the flashiness index is one of such signatures. The values of the flashiness index reduce with increasing time step due to a smoothing effect. In this paper, we did not experiment with varying data resolution, as it was outside our scope. However, we will comment that this choice is expected to affect some of the signatures, and consequently their diagnostic power on some of the associated processes.

We experimented with several signatures to account for the effect of snow on streamflow seasonality, and we ended up selecting the "half streamflow period". The reason is that 1) this signature was used in previous publications to quantify streamflow seasonality, and we did not want to invent our own signature if something was already existing, and 2) this signature captured well the difference between streamflow regimes, because we have seen (figure 5) that all the catchments receive similar precipitation input (in terms of monthly variability) but the snow-affected ones show the peaks during late spring/beginning of summer while the rain dominated ones show their peaks during the winter and the spring.

Figures 7-9 show that all the model configurations represent well the yearly streamflow, the runoff coefficient, and the half streamflow period and this is a result of our study that is also coherent with our assumption that only distributing the inputs (precipitation, PET, and temperature) is sufficient in order to have a model that captures the water balance and the snow dynamics. In the revised paper we will show the analysis of the correlations between the signatures in order to select only the not redundant ones.

We will add more references in the introduction to signatures papers.

b. *In your study you included several (fairly easy to derive) landscape characteristics that are obviously highly correlated and describe in great detail how you identify and select appropriate ones based on regression and correlation. In my opinion a rather trivial part which does not add any new knowledge to the literature occupies a lot of space. I hence suggest shortening and streamlining the entire section. If you want to derive structure from function than the first goal must be to derive a (comprehensive) matrix of*

uncorrelated catchment characteristics that have some kind of diagnostic potential. In my opinion this should be the source of the story and not a result.

We will try in the revised version to streamline this part. We will remove some of catchment characteristics before carrying out the correlation. This can apply, for example, to variables that are relatively uniform over the catchment, or to catchment characteristics that occupy very small areas of the catchment.

We could also remove variables that are strongly correlated to others, and do not have any perceived influence on the selected signatures.

3. *The approach for informing model structure does not appear very elaborate to me. Since this is the core of "model building for understanding catchment process" I particularly miss a clear and elaborate discussion on how the identified landscape characteristics help in the model building process. More specifically:*
    a. *In chapter 3.1.3. you state that the results of the regression analysis were used to build the hydrological model e.g. the subdivision of the catchment in HRUs (p. 7 l. 32). Later, in 4.1.1 you state that subdivisions were defined by gauge locations (p. 11 l 26). I did not find information on how you derived the number of HRUs and the role of catchment characteristics in this context? Chapter 4.1.1. should be more comprehensive in this regard.*

    We will clarify this aspect in the revised paper. Our intention was to present chapter 4.1.1 as a general overview of the model structure in order to make the following clearer. The information from the regression analysis are used to derive the HRUs is described in chapters 3.3 and 4.1.5. It is important to make clear the difference between the division in subcatchments (areas that have uniform inputs) and HRUs (areas that have the same hydrological response). The former are defined by the presence of gauging stations (and this division is not negotiable) while the latter reflect our understanding of the catchment functioning (and, in this study, of the regression analysis).

    b. *The argument that "the regression analyses have indicated that precipitation is a dominant control on average streamflow" (p. 12 l. 4) is trivial. I don't think you need this and particular not as a justification for using spatially distribution rainfall as a model input. From your manuscript it appears to me that the spatial discretization of your model was based on the definition of the subcatchments (which are in turn defined by the location of gauges) and according the definition of fields (definition not clear). In consequence I don't see that landscape characteristics played an important role in this process. Please clarify?*

    We will clarify that, although it may be a priori clear that precipitation needs to be distributed per subcatchment, it may be not as taken for granted that this is sufficient to capture the water balance of the subcatchments, as many other aspects could in principle play a role (e.g. regional groundwater flow). Here we show that considering distributed precipitation over the subcatchments (defined by the presence of the gauging stations) could by itself be sufficient. Other landscape characteristics play a role in the definition of the HRUs (section 4.1.5, see previous comment).

We will make the definition of subcatchments, HRUs and fields clearer and comment on the role of precipitation.

c. *You also mention that "the parameters were motivated by the results of the regression analysis" (p. 8 l. 1). Please omit or explain more detailed. A matrix to illustrate the relationships between model parameters and catchment descriptors would be good. I would for instance be interested in how one could use catchment descriptors to derive (or at least constrain) model storage (kFR or kSR) or network lag (trise,IL trise,OL) parameters. Please comment on that*

We will clarify this part, which was obviously misleading. All the process of building the model was motivated by the results of the regression analysis (in particular the decisions on the division in HRUs). The parameters are just calibrated using streamflow data (section 4.1.3). No inference of the parameters from catchment characteristics was done.

We will improve the text (section 3.1.3) to make it clearer.

d. *Chapter 4.1.5 is difficult to me due to different reasons: i) your analysis does not VERIFY that "models that account for influencing factors … lead to an improved representation". Essentially it only shows that a complicated model (with a larger number of degrees of freedom) outperforms a simpler model (with a smaller amount of degrees of freedom). Please use precise wording. It addresses the question of adequate model complexity. If a lumped representation (M1) is not adequate than also the comparison of M2 to M1 is not adequate. Please explain in more detail why you consider M1 a suitable reference? ii) Please explain why unconsolidated areas receive an individual HRU and why consolidated and alluvial areas can be lumped together (what are your expectation on the underlying processes)? iii) The parameterization of M3 is based on land use, which is not considered to have a causal relationship to the streamflow signatures (Table 2). Please explain why a model which is derived from non-causal properties can be a considered a meaningful reference? Why did you group based on geology and not on elevation, slope or the aridity index which you considered to have a causal relationship? This would maybe be a more appropriate benchmark? iv) Essentially chapter 4.1.5 addresses the questions of optimal degree of model complexity and optimal degree of spatial discretization - which are both very important. However, these aspects are treated together and not separated from each other. Moreover, potential answers to these questions miss a clear link to catchment descriptors. Essentially only differences in geology were considered in the model building. Please clarify to novelties of your study more clearly.*

We will more clearly explain the reasoning behind the choice of the model configurations done in this study. Essentially the two main model configurations are M1 and M2: the first is the baseline and it is a semidistributed model (in the sense that the inputs are spatially distributed and the routing between subcatchments is explicitly addressed in the model) with only one HRU (meaning that all the catchment responds in the same way to the forcings); the second extends the first providing a subdivision of the subcatchments in two HRUs. M3 is used to show that the subdivision in HRUs has to be

carefully made otherwise a more complex model doesn't imply automatically better results. Answering to the specific points:

i) M2 is indeed more complex than M1 but our thesis is that its better performance is not just due to the fact that it is more complex but to the fact that it incorporates the right catchment characteristics. This is also demonstrated by M3 that is as complex as M2 but it has the same deficiencies of M1. M1 is already a quite complex model since it already considers the spatial distribution of the inputs and incorporates information about the routing between subcatchments. The real baseline would have been a lumped model, with uniform input and no information about the catchment characteristics but it was too simple for the comparison.

ii) There is an error in the text: the two HRUs are unconsolidated and alluvial (HRU1) vs consolidated (HRU2). Alluvial and unconsolidated geology were put together because they show a similar behavior in terms of water dynamics in the sense that they both represent areas with high storage capacity, especially if compared with HRU2 that is quite impermeable.

iii) M3 was designed to demonstrate that M2 outperforms M1 not just because it is more complex but only because it incorporates characteristics that actually have an impact on the response of the catchment. For this reason we used a model with the same complexity of M2 but based on characteristics that don't correlate with the streamflow signatures. Also topography was considered in the modeling experiment (but not reported in the paper), experimenting with a subdivision in HRUs based on the slope, but the model resulted similar to M2 (in terms of spatial discretization) but slightly worse in terms of signatures representation. The meteorological characteristics are known at subcatchment scale and therefore, due to the configuration of the model, they are not suitable for the subdivision in HRUs.

iv) We will consider this comment when we will rewrite paragraph 4.1.5.

e. *the whole structure of the model building story is a bit complicated as aspects are described in chapters 3.1.3, 3.3, 4.1.4 und 4.1.5 which makes it difficult to follow. I suggest combining them into a single chapter. Therein start with the theory e.g. snow is important followed by the surrogate you considered it e.g. half stream flow period. Or geology is important due to… –> different HRUs.*

We will improve the readability of the paper making clearer the process of model building. It was divided in different paragraphs along the paper in order to emphasize the connection between data analysis and modeling choices but we understand that it makes more difficult to follow the story. We will keep it in mind when we will rewrite chapter 4.1.5.

4. *Several conclusions are not appropriate: e.g. "the proposed approach is useful in the fact that modelling can be used to test specific hypotheses on dominant processes resulting from regression analysis" (p. 19 l. 4). This has not be shown. More over aspects related to the event scale are mentioned in the first three bullet points but not subject to the manuscript. In the third bullet point you state: "Higher proportion of consolidated material has an influence on the baseflow vs quickflow portioning, causing lower baseflow and higher peaks" (p 19 l 14). Does the*

study provide evidence for this statement or does it support this hypothesis? I expect the latter and missed this statement in the chapter 3.1.1. I suggest re-writing of the entire section and to differentiate concisely between hypothesis, results and conclusions.
We will be more precise in our conclusions. With respect to the first point, we think that the model comparisons have been useful to confirm the interpretations of the regression analysis. Clearly the regression between variables is also a model, but the hydrological model is an integrated model that is meant to explain all dependencies at once, whereas the regression model provides a separate model (regression) for each of the dependencies. Therefore there is an added value in the hydrological model, compared to the regression model.
With respect to the use of "event" in the first three bullet points, we will be more precise and refer directly to the signatures rather than to the time scale.
In terms of the baseflow vs quickflow, we will clarify that we refer to the baseflow index, and to how the model tries to mimic it by varying the partitioning of water between fast and slow reservoirs.

5. *The model performance evaluation (chapter 4.1.4) is complicated but of minor importance in this context. I suggest shorting the evaluation section and to focus on a single, interpretable metric e.g. the Kling-Gupta Efficiency as the NASH has several limitations and the normalized log-likelihood is difficult to interpret. But this is a minor point and a matter of taste.*
We agree that the Nash-Sutcliffe efficiency has several limitations, as any individual index is somehow limited. This is why we have introduced several signatures to evaluate model results. Indeed, we could see that a significant improvement in some of the signatures could result into a negligible improvement in the Nash- Sutcliffe efficiency.

Technical corrections (figures and tables only) I only provide technical corrections for the figures and tables as I expect that several parts of the manuscript will be subject to major revisions.

Thank you for the comments for improving the quality of our figures and tables; we will address them bellow.

- *Figure 1: A: I suggest to remove the colour code and to provide notations (abbreviation) in or around the map. This would help improve the readability of the stream network and the location of the gauges. If you want to keep the legend please add catchment abbreviations to it, order it according to Fig 2. and use a meaningful colour code (e.g. mean annual precipitation, elevation or geology), B: Try a discrete legend like in atlases, will improve readability. C: Forest and pasture are hardly distinguishable both on my screen and in a printed version.*
Figure 1A: We agree that there are some problems with the readability of the river network but they are mainly due to the poor resolution. In the final version we will upload the figures separately with an higher resolution. The presence of the legend doesn't make the figure smaller since the constraint is the height and not the width of the box; we will then keep the legend sorting the names according to the other figures and using abbreviations. The colors used for the single catchments were chosen from a "categorical" color scale in order to be as different as possible. Linking them to some characteristic would mean using a "sequential" color scale, with

little difference between subcatchments, and this would be problematic in the other figures (assuming that we want to be consistent) where we want to clearly see the behavior of the single catchments.
Figure 1B: we will improve the figure according to the suggestion
Figure 1C: we will change the colors to improve the readability and the figures will have better resolution.

- *Figure 2: Please repeat the variables and their abbreviation in the caption such that the figure can be read independent from the text. Maybe add another row and provide grouping indices based on the results in chapter 3.2.1*
  As the names are relatively long, they would not fit on the y axes. Instead, we have opted to place them in the title of the subplots.
- *Figure 3: Please repeat the variables and their abbreviation in the caption such that the figure can be read independent from the text.*
  See reply at earlier point.
- *Figure 4: Please repeat the variables and their abbreviation in the caption such that the figure can be read independent from the text. Information on the range of the different variables would be pretty helpful as well. If possible include it otherwise please mention the ranges in the text or add the information to table 1.*
  In order to make the figure clearer, we will report the meaning of the variables in the caption and group them according to the category that they represent. The range of the variables is always between 0 and 1: all the variables plotted are percentage of the area of the subcatchment occupied by a certain characteristic. The characteristics that don't belong to the category "part of the catchment occupied by …" are reported in table 1.
- *Figure 5: I'm not sure if this figure is required since B and C show very little variation. The only important message from A is that there a catchments that are stronger controlled by snow than others. I suggest removing it. If you decide to keep it update the colour code according to the suggestion for Figure 1.*
  Although the plots B and C show little variability across the catchment, it is still interesting to present the seasonal dynamics. Moreover, we consider that it is useful to show that the monthly variability in streamflow (plot A) is not directly ascribed to variability in precipitation or potential evaporation (plot B or C).
- *Figure 6: I cannot find a description of the symbols and abbreviations in the Appendix. Please specify at least the meaning of the capital letters in the caption (as in 4.1.1) and provide a more comprehensive description in the appendix.*
  We will put a detailed description either in the caption or in the appendix.
- *Figure 7: Order according to Figure 2 or 3. Line type and colour code are redundant.*
  We will reorder plot B according to figure 2. We will substitute the dashed line with a continuous thin line (same for all colors).
- *Figure 8, 9, 10: Nice figures! Suggestions: Combine all three figures in one (each model setup as an individual row). This would improve readability. Streamflow, runoff coefficient and half streamflow period have no or little variation (two out of these could be omitted such that all results would fit in one figure). Remove the correlation coefficients due to their distracting nature*

(correlation (alone) is pretty meaningless in this context). Update colour code according to Fig 1 A.

Point taken. We acknowledge that is more meaningful to put the different models together in order to facilitate the comparison. Since we will reduce the number of signatures in the analysis, there will be enough space to host all of them in the same figure.

- *Table 1: Order according to fig 1 A. Index column is not relevant, omit Code or put it to the very right. Rounding is not yet meaningfully and consistent.*
  We will reorder the table according to the figures. The Index column is used in the "upstream catchments" column to define the river network. The "code" column is present to avoid ambiguity with the naming of the gauging stations providing the reader with the code of the gauging station used by the Federal Office for the Environment FOEN. The rounding is meant to maintain the same number of significant digits, and it is consistent with how the FOEN gives the values in the website (no decimal digits if the area is greater than 100 km^2, one digit if it is between 10 km^2 and 100 km^2, two digits if it is lower than 10 km^2).
- *Table 2: This table includes variables with spurious correlations (Brett 2004, Kenny 1982). This includes variables that are considered statistically significant and where causality was assumed e.g. the correlation between aridity index AI and the runoff coefficient RC which are both are derived from precipitation. The same applies for P and RC. Since P and Q are highly correlated and AI is based on P I also wonder about the significance of AI and RC, BFI, FI and HDP. Please clarify. Please also explain why you assumed causality among LP and BFI and among LP and FI? Differences among the geological fractions are small as well. Why do you consider causality in some of the individual relationships and in others not?*
  This table will be modified during the review of the paper taking into account this comment and the changes in the text.
- *Table 3: This analysis also includes variables with spurious correlations. Please comment on that.*
  This table will be modified during the review of the paper taking into account this comment and the changes in the text.
- *Table A1: Please provide a brief explanation on parameters and components. Where does the range of variability come from?*
  We will describe the meaning of the columns in the caption and add more description in the appendix.
- *Table A2: Explain component*
  We will describe the meaning of the columns in the caption and add more description in the appendix.

Literature

Brett, M. T. (2004). When is a correlation between non-independent variables "spurious"?

Oikos, 105(3), 647–656. Kenney, B. C. (1982). Beware of spurious self-correlations! Water Resources Research, 18(4), 1041–1048. https://doi.org/10.1029/WR018i004p01041