Final response

We sincerely thank the three anonymous referees for their detailed comments, as well as the editor for giving us the opportunity to revise our manuscript.

Here we provide our final response to all major comments made by the referees and the editor. Since our manuscript will be subject to major revisions we do not provide a point-by-point list, but rather a general reply to what we have identified as the most important issues related to our initial contribution.

For RC#1, but also for RC#2, RC#3 and the editor, the presentation of the material was a key point of concern. Among all referees was a consensus on the lack of a logical and smooth structure, ultimately making the manuscript difficult to read. We can see this point (now) and agree that several sections should be re-written and/or re-structured. This includes particularly the abstract, introduction & objectives, the part on "conceptual framework and candidate diagnostics", as well as the conclusions. Although we agree with RC1 that the paper could be shortened by such a revision, the comments also suggest that several aspects need to be clarified and explained in more detail. This concerns in particular the rationale of the normalization, the logic of the storage predictors and the whole concept of "intensity controlled" runoff production. We understand almost all of the related points of criticism and will revise our manuscript according to the suggestions of the referees/editor. However, we suspect that the clarification of these issues might further extend the paper – which is already very long. For this reason, we decided to split our contribution into a companion study. We suggest that the initially proposed manuscript becomes part 1 and to defer certain elements into a second manuscript, which we will submit separately. We intend to organize the companion study along the following avenues:

> Manuscript 1) Exploring the interplay between state, structure and runoff behavior of lower mesoscale catchments I: unraveling abiotic and biotic controls on the seasonal water balance.
>
> Manuscript 2) Exploring the interplay between state, structure and runoff behavior of lower mesoscale catchments II: storage versus intensity controlled event runoff production.

The overarching goal of our contribution is to better understand the interplay between state, structure and runoff behavior in lower mesoscale catchments. By distinguishing between seasonal and event runoff production we study catchment runoff dynamics at different time scales. At event scale we furthermore differentiate according to the nature of the runoff formation process. Our key intention is to isolate multiple impacts on runoff production, i.e. to separately study the impact of catchment structure, moisture state or that of biotic controls. To discriminate among the impacts of the rainfall forcing and that of the structural (terrestrial) properties of the catchment we use normalization. In essence we hypothesise that normalized state-response and forcing-response plots are suitable diagnostics to discriminate differences in catchment runoff production at different time scales. We further differentiate this general hypothesis (H) within the two contributions and test it through complementary research questions (Q):

Paper 1: Abiotic and biotic controls on the seasonal water balance

H1: Normalized forcing-response plots are suitable diagnostics to discriminate biotic and abiotic controls on seasonal catchment runoff production.

Q1: How to evaluate and normalize seasonal runoff production, so that the impact of biotic and abiotic controls can be evaluated independent from meteorological forcing and/or the structural setup of the catchment?

Q2: Which structural catchment characteristics explain the differences in seasonal runoff production during the dormant season and do any of them operate in groups?

Paper 2: Storage versus intensity controlled event runoff production

H2: Normalized state-response and forcing-response plots can discriminate differences in catchment runoff production at the event time scale.

Q1: How to estimate and evaluate the impact of different catchment water storages on event runoff production?

Q2: How to normalize both storage and response to (i) consider the site specific structural setup of the catchment and (ii) detect similarities in event scale runoff production in inter-comparison studies?

Q3: Is it possible to detect evidence for intensity controlled runoff formation, which refers to the activation of rapid flow paths, such as surface runoff or preferential flow, based upon commonly available (hourly aggregated) data of mesoscale catchments

In addition to our reply to the anonymous referee # 1 we provide answers to the major aspects raised by the reviewers/editor in the following:

What is the rationale of normalization?

To normalize means to relate a variable of interest to a reference, or to express it in terms of the latter. Such a normalization allows to separate multiple impacts on a single variable and facilitates studying the importance of different properties on a single quantity. Normalization hence allows to compare different sites, which makes it a promising tool for comparative analyses. Well-known and widely used examples of normalized quantities include specific discharge, typically expressed in [length/time] where the impact of the catchment size [length x length] is separated from the discharge measurement [volume/time], or runoff coefficients [-] where the response, i.e. again specific discharge [length/time] is normalized such that it is independent from the forcing, i.e. specific rainfall [length/time]. Normalized quantities are often used as "diagnostic" variables to study runoff generation (e.g. Merz et al., 2006 or Graeff et al., 2012).

In line with these studies we propose that the analysis of normalized runoff surrogates allows to separate the impact of the rainfall forcing from that of the terrestrial influences, i.e. catchment state and structure. The latter includes both abiotic and biotic components. In essence, we suggest that the inherent "functional behavior" of a catchment with respect to runoff generation can be described by normalized state-response and/or forcing-response diagrams.

In our study, we analyze runoff generation at different time scales. Specifically, we distinguish between runoff production on the seasonal and on the event time scale. On the latter we furthermore differentiate according to the nature of the runoff formation process. Accordingly, we use different normalization schemes.

At seasonal scale: The double mass curves (DMC), i.e. normalized forcing-response plots, which we use at the seasonal scale are in fact pretty similar to common practice in soil physics where tracer breakthrough is plotted against cumulated irrigation (and not against time) to study transport and adsorption. Contrary to soil physics we do however have two forms of water release in catchment hydrology (evaporation and stream flow) and the proposed double mass curves are particularly suited to separate regimes where either the one or the other is dominating. In the current version of the manuscript the DMCs are normalized with total precipitation. This has the advantage that the abscissa is always scaled from 0 to 1 which facilitates the comparison of different sites and years. It has however the weakness that the same ordinate values do not reflect the same mass input. An alternative would be to normalize with storage volume (despite that the uncertainty is high), which expresses mass input in terms of storage volume. A further alternative would be to use a data-driven estimate for the maximum potential evaporation such as net solar radiation divided by the latent heat of vaporization (assuming the entire incoming energy would be consumed for evaporation). This would efficiently separate cold from warm years. We will further elaborate and test the potential of the different normalization schemes for describing the seasonal water balance in the revised manuscript. Furthermore, we will include common-practice signatures and relate insights obtained from the DMCs to results obtained from flow duration curves. The latter are widely used as hydrological signatures in similarity studies.

At event scale: at event scale we distinguish between "capacity" controlled runoff production and "intensity" controlled runoff production. The former relates to mechanisms where the relative amount of water which is stored in the control volume dominates the response. We expect that capacity controlled runoff production, meaning that streamflow monotonically depends on storage amount is the most important mechanism in many, particularly humid environments. The key challenges in this context are a) to characterize different storage compartments, which is done using the different storage estimators and b) to estimate their relative content, which is done by normalizing the storage estimator using a surrogate for the storage capacity in the subsurface. Intensive runoff production refers to the activation of rapid flow paths such as surface runoff or preferential flow. The activation of these flow paths requires an intensive, typically convective rainfall forcing (Beven and Germann, 2013) and is largely independent from the amount of water which is stored in the subsurface. (Relative) storage is hence a poor predictor for intensity controlled runoff production and it is clear that the detection of these mechanisms by means of normalization requires to consider the nature of the process. The use of (storage) capacities, which behave additive, is less appropriate in this case. Intensities, which do not behave additively, actually are the much more relevant properties in that event. Estimating all relevant intensities, or more generally the "intensive state variables" is however not straightforward. It would require structural information on the rapid flow paths, i.e. related conductivities, knowledge on the intensity of the forcing and on that of the response, both in a very high spatio-temporal resolution. Unfortunately, the former are not available and the latter, i.e. discharge, is a convolute of different distributed runoff generation processes, concentration and routing. For this reason we suggest to focus on high intensive rainfall events and to consider associated high intensive runoff responses as evidence for the activation of rapid flow paths. To detect such processes within rainfall-runoff events and based upon data at hourly resolution, which are poorly resolved in this context, we propose to relate the temporal derivatives of rainfall and runoff. This implies to "edge-filter" the time series and to normalize the temporal derivative of discharge using the temporal derivative of precipitation. This emphasizes rapid changes in intensities much better than the observed values.

We recognize that in the first version of our manuscript we did not concisely differentiate between "normalized" and "dimensionless" variables. While many of our normalized quantities are dimensionless, the latter is of course not a prerequisite for the former. We therefore thank the editor for pointing out that both terms must not be mistaken and clearly differentiated. We will clarify this in the revised version of our manuscript.

What is the logic of defining and evaluating the different storage predictors?

The importance of storage on runoff production is beyond question. Characterizing and normalizing storage is however difficult, as elaborated in section 1.3. Basically, we intend to 1) assess the importance of different sub-surface storage compartments on event runoff production (and associated spatial patterns) and 2) to detect similarities in storage capacity controlled runoff production among different sites by means of an inter-comparison study. To characterize storage we selected total active storage dS (Sayama et al. 2011) (Eq. 1) which we associated mainly with deeper storage compartments, and theta (Eq. 2) as a surrogate for the near-surface storage. Pre-event discharge (Eq. 3) was included as it proved to be a valuable estimator for the bulk catchment moisture state (Graeff et al. 2012), although it cannot be attributed clearly to either the near-surface or deeper storage compartments.

To evaluate the impact of the different storage compartments on runoff response we correlate storage estimators and event runoff coefficients. The number of significant correlations gives insights into the overall importance of the different storage surrogates, i.e. the importance of the different sub-surface storage compartments, within our data set. Spatial patterns provide information on regional differences in the importance of the storage compartments. The site-specific normalization is necessary to judge the relative importance of the storage estimators which allows a meaningful comparison of different catchments.

Currently, the evaluation focuses on general aspects of the proposed method. Therefore, we mainly provided results for individual catchments and not for the inter-comparison of different sites like it is done in Fig. 5, bottom right. Splitting the manuscript will allow us to provide more results and to explain them in more detail. Thanks also to RC#2 for pointing out that there is no "best" storage estimator as different predictors represent different properties, yet they are all related to catchment storage. We will clarify this in the revised version of the manuscript. Therein, we will also detail on our rationale for defining the time of integration/summation in the proposed storage measures (Eq. 1-3). It is of course true that the time of integration/summation introduces a high degree of subjectivity and thus, uncertainty. This has already been pointed out by Heggen (2001) and Graeff et al. (2012).

In the revised manuscript we will also provide the units of all quantities which we use in our analyses and show that Eq. 1-3 are dimensionless. We thank the editor for highlighting that this aspect need clarification.

Which data have been used for which purpose and what is the role of the hydrological model?

The fact that observables are generally less uncertain than model outputs motivated us to employ data-driven signatures whenever possible. Data-driven means that our signatures are based upon observables and not upon model outputs. This applies for all signatures, except for those which include ET estimates. ET was calculated based upon Penman-Monteith using a water budget model (LARSIM). Calculating ET required standard hydro-meterological time series, i.e. observed station data of radiation, wind speed, humidity, temperature, etc. We use the same hydrological model for all catchments to ensure that the ET estimates are calculated and interpolated in a consistent way.

In the revised version of the manuscript we will clarify the use of the different data and explain which of them were used for modelling and which of them were directly used to derive the different signatures. Following the suggestions of RC#2 and RC#3 will also comment on the associated uncertainty.

References

Beven, K., Germann, P., 2013. Macropores and water flow in soils revisited. Water Resour. Res. 49, 3071–3092. doi:10.1002/wrcr.20156

BGR, 1995. Bodenübersichtskarte von Deutschland 1:1,000,000 (BÜK1000). Federal Institute for Geosciences and Natural Resources (BGR).

Merz, R., Blöschl, G., Parajka, J., 2006. Spatio-temporal variability of event runoff coefficients. J. Hydrol. 331, 591–604. doi:10.1016/j.jhydrol.2006.06.008

Graeff, T., Zehe, E., Blume, T., Francke, T., Schröder, B., 2012. Predicting event response in a nested catchment with generalized linear models and a distributed watershed model. Hydrol. Process. 26, 3749–3769. doi:10.1002/hyp.8463

Heggen, R.J., 2001. Normalized Antecedent Precipitation Index. J. Hydrol. Eng. 6, 377–381. doi:10.1061/(ASCE)1084-0699(2001)6:5(377)

Sayama, T., McDonnell, J.J., Dhakal, A., Sullivan, K., 2011. How much water can a watershed store? Hydrol. Process. 25, 3899–3908. doi:10.1002/hyp.8288