Editor's comment

Dear Authors, your original paper has received some contrasting comments and a few criticisms requiring an additional round of referees' evaluations. Please, try to improve as much as possible your revised version, allowing for the good suggestions received during the discussion phase.

Reply:

We thank the Editor for his overall positive assessment of our manuscript. We revised our manuscript addressing the main requests of the reviewers by adding a new figure with the scheme of the main processes involved in the hysteresis and by adding a sensitivity analysis of the hysteresis index to the model parameters uncertainty. These 2 points are the main changes made, but we also integrated all the comments from the review process in order to improve as much as possible our manuscript as detailed below in the point-by-point responses to each review.

Reply to Anonymous Referee #1

"This paper is very well written and surely of interest for the hydrology community. I agree with the Authors in that the purpose of hydrology is not to maximise performance measures but to correctly understand/reproduce what happens (in this case, what are the catchment internal dynamics). This is valid for practical purposes too, since models that can correctly capture the processes going on are expected to be more reliable in predicting the catchment response in conditions non observed in the past. I am definitively supporting for the publication of the paper in HESS. I have some specific comments below, but since they mostly involve additional discussion, from my side the resulting revision should be minor."

We thank the reviewer for the positive assessment of our article and his/her comments and suggestions which helped us to make our manuscript clearer for the reader and to extend the discussion. Below we reply to each of the specific comments.

Specific comments:

1. *"The analysis is done on only one (very small) catchment, while from the title I would have expected more examples"*

   We agree on the ambiguity in the title using plural form of "catchments" therefore we propose to reformulate as "Hydrological hysteresis and its value for assessing consistency in catchment conceptual models".

2. p. 5669, section 2.3. *"is the normalisation of the storage/saturation values using the minimum and maximum observed values a robust choice? How much does it depend on the record length? How sensitive are the hysteresis indices to this choice? The Authors should add one sentence here to justify that this choice is robust and/or that it has no effect on the results of the study."*

Reply: The hysteresis index is defined as a difference between recharge and recession storage values which correspond to either a normalized groundwater level or normalized soil moisture. As highlighted in the following comment (3) the catchment is characterized by a strong annual cycle with clear recharge and discharge periods. At the inter-annual scale, while stream flow varies in a quite large range due to rainfall variability, the groundwater levels and the soil moisture values are varying within a narrower range of values from one year to the other. The normalization of storage values aims at making the comparison between measurement points (upslope/downslope for the piezometers and depth for the soil moisture sensors) more readable. Using the maximum and the minimum values is a simple solution for normalizing because it is difficult to estimate the actual storage capacity of both unsaturated and saturated zones. In the Hysteresis Index as it is defined here, the normalization is equivalent to dividing the difference between recharge and recession non-normalized values by the maximal amplitude over the records:

Denoting Z the groundwater depth (always negative) and θ the soil moisture (always positive) HI could be written as

$$HI = \frac{Z(t_R) - Z(t_r)}{\text{Min}(Z) - \text{Max}(Z)}$$

Or

$$HI = \frac{\theta(t_R) - \theta(t_r)}{\text{Max}(\theta) - \text{Min}(\theta)}$$

According to the respective ranges of variation of Z and θ, the denominator is always a positive real. **Therefore the normalization does not affect the sign of HI.** The value of the denominator increases with the amplitude of groundwater level variations or soil moisture variations in the record, thus HI values are likely to decrease when the amplitude of variations increases. However it **does not affect our results because**:

- The normalization would tend to increase the absolute values of HI computed from the dowslope piezometers where groundwater fluctuations are lower than in the upland piezometers but we still observed that HI absolute value tends to decrease from upslope to downslope areas so the normalization does not erase this trend.
- When HI is used to compare the model to the observations the normalization is done on the same period.

As suggested by the reviewer we propose to add a sentence in order to explain this point at the end of section 2.4:

"The normalization of the observed variables related to the storages (here either groundwater level or soil moisture) has no effect on the sign of HI, the HI values are being only divided by the maximal amplitude observed in the storage during the whole period. Therefore, as long as the normalization is applied for the whole period (for all years and for both measurements and simulations), it does not affect the interpretation related to absolute values of HI. "

3. p.5670, Eq. (1). *"This definition for the hysteresis index is used by the Authors at the annual scale. This makes sense in this work because the storage dynamics have an annual period (see Fig. 2). Do the Author expect this to be the case in general? I would think that in other catchments there could be more cycles in one year or even a non-periodic behaviour (in arid climates). However Eq. (1) would still be valid but at the event (rather than annual) scale. If so, a sentence could be added here as a guidance for researchers willing to use the same index in different hydrological settings."*

Reply: We agree with the reviewer that the HI index as defined here is useful for characterizing periodic behaviours. Here, the index led to annual values because of the strong seasonal cycle occuring in the studied catchment. The same index could be used similarly for flood events if they are more or less uni-modal. For multiple-seasonality, e.g. if there are 2 recharge periods in the year, the hysteresis is likely to exhibit a double loop and 2 indices may be relevant to describe each of them. In particular one can imagine that 2 successive loops may have different directions (so different signs of HI) due to the successive activation of different flow paths and the fact that hillslope storage is likely to be less empty at the beginning of the second recharge period than at the beginning of the first one. In snow-melt driven catchments, the hysteresis relative to the snow cover storage should be taken into account too (as a third storage). In arid catchment where the groundwater recession can occur during several years (see e.g. Ruiz et al., 2010), it would be more relevant to compare these relationships among the identified pluri-annual cycles composed by at least both a recharge and a recession rather than at the annual scale. In order to help the identification of the limits of our HI, we propose to add the following precision before explaining how HI is computed:

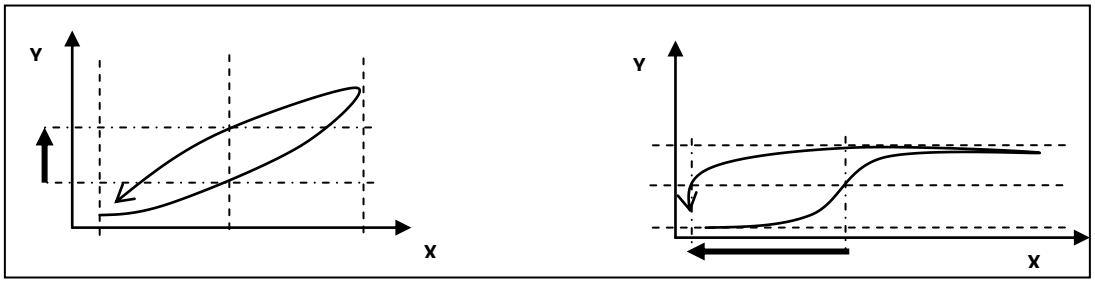
"In this paper, **as the hydrological variables exhibit a strong annual uni-modal cycle,** we calculated the hysteresis index (HI) each year as the difference between water storages at the dates of mid-point discharge in the two phases of the hydrological year"

4. p. 5671 Eq. (2). *"Related to the previous comment, is the choice of Q_{mid} robust? This is because I expect Q_{max} to be very variable from year to year and maybe related to short term rainfall response (flood event)."*

Reply: Regarding the observed hysteresis over the 10 years, the choice of Q_{mid} succeeded in catching the difference of the saturated storage states in recharge and recession more or less in the middle of these periods in the study site (cf. Figure 4). Even if stream flow is varying considerably among years, these variations are not so abrupt. Moreover forgot to mention that we have used smoothed time series using 7-day moving average. To clarify this important point we added the following explanation in section2.4:

"In order to reduce the impact of the quick variations of discharge or groundwater level due to individual storm events, we smoothed the time series using 7-day moving averages. The strong seasonal discharge cycle led to identify two occurences of Q_{mid} per year only: during the recharge period (t_R) and during the recession period (t_r) , while high and low stream flow values are taken several times per year as explained by Lawler et al. (2006)."

However we agree that the relevance of Q_{mid} will depend on the shape of the loop. As cited in the corresponding section 2.4, some authors prefer to describe the loop width using the extreme values of the Y variable (X variable is always stream flow, Y is storage in our case but can be either a concentration or the turbidity in hysteresis studies). If the stream flow maximal values flatten the loop, a better metric of the hysteresis width could be the difference of stream flow values between recharge and recession for the annual mid-point storage value (cf. schematic representations below).



5. p.5671, lines 15-17. *"What do clockwise and anticlockwise hysteresis loops mean from a process point of view?"*

Reply: at this stage of the manuscript (Material and method section) the processes are not further developed as the interpretation of the underlying processes is discussed in the results and discussion section, but the direction of the hysteresis indicates immediately which variable (storage or flow) is reacting first to rainfall. We propose the following additional explanations related to HI information:

"HI is a proxy for the importance of lag time response between variations in catchment storages (unsaturated and saturated) and stream discharge**, its sign indicates if storage reacts before or after the stream flow**."

6. p.5671, line 26. *"how does this work differ from Hrachowitz et al. (2014)? That paper is under review in WRR and has a title which could be the title of this manuscript, although more general. A sentence should be added in the introduction (and maybe also here) to clarify what are the different contributions of the two papers."*

Reply: The contribution of the previous work from Hrachowtiz et al., used as a basis of the present work, is detailed in sections 2.5 and 2.6 of the Material and Method section as it is considered as previous results/knowledge. To clarify this, as suggested by the reviewers, we propose:

i) to add a mention to this previous work at the end of the introduction :

" (...) to which degree a suite of conceptual rainfall-runoff models with increasing complexity, **which were calibrated and evaluated for this catchment in a previous work, using a flexible modelling framework (Hrachowitz et al., 2014)**, can reproduce the observed storage-discharge hysteresis (...)"

ii) to explain this choice to the reader at the beginning of the corresponding section

"In a previous work, a range of 11 conceptual models were calibrated and evaluated for the Kerrien catchment in a stepwise development using a flexible modelling framework (see Hrachowitz et al., 2014). **This section aims at summarizing the results of this previous study that are used as a basis for the present work.**"

(ii) and to provide additional information about this work in a new table (cf. New Table 4 at the end of this reply) for clarifying the objective functions used for calibrating the models, the hydrological signatures used for assessing them; and Figure 3 has been revised to provide the hydrographs on both calibration and evaluation periods (provided at the end of this reply).

It has to be noticed that the **manuscript submitted to WRR is now accepted** and available for further details on the previous work (doi: 10.1002/2014WR015484).

7. p. 5674, line 18. *"in Hrachowitz et al. (2014) more model structures were considered while here just four of them are analysed. What is the rationale for the choice of these four?"*

Reply: The selection of only 4 of the 11 models from the previous work of Hrachowitz et al. (2014) has been motivated by the fact they correspond to the main different model architectures. The other models are rather constrained differently. We agree that this rationale should be explained and we propose to add this precision at the beginning of the section 2.5:

"Four of these 11 models (noted M1 to M4) were selected for the present work, as they correspond to the sequence of model architectures that provide the most significant performance improvements."

8. p. 5677, section 3.1.3. *"I like this section a lot. Just a suggestion: a figure/schematic that illustrates the mechanisms leading to opposite directions of the hysteresis loops in the hillslope and riparian zone (hypothesis 3) would be very useful (here or later)."*

Reply: We thank the reviewer for this suggestion: we propose to add this new figure for illustrating our interpretations in terms of mechanisms (see new Figure 9 at the end of this reply).

9. p. 5679, sections 3.1.4, 3.1.5. *"Maybe also the sensitivity to Q_{mid} could be explored. Do the results change if the second annual peak is chosen as Q_{max}?"*

Reply: This would be interesting but as explained in reply to comment 4 we are working on smoothed data in order to eliminate the highest Q values due to rapid storm. However, we agree that its sensitivity analysis would be needed if the index was to be used e.g. to calibrate the models. For this purpose, we suggest that a range of indices rather than a unique one should be used to fully describe the hysteresis. Therefore, the sensitivity analysis would be worth when all this range of indices will be defined. What we propose to add from the revision process is an investigation of the sensitivity of HI values to the parameter uncertainties (see below and also comments of the other reviewers)

10. p. 5683, lines 25-27. *"The Authors state that "...a model able to represent the internal catchment behaviour will generate a wrong discharge value but consistent with the storage value and will be rejected in traditional calibration procedures". This is a very valid point. If the Authors could show that this actually happens in the study they made, that would be great. The model parametrisations chosen for the analysis are optimal in maximising the performance measure Eq. (5) (page 5674, lines 23-26). It would have been very interesting to find out whether non-optimal parametrisations result in better modelling of the hysteresis."*

Reply: We agree with the reviewer and therefore we included a sensitivity analysis of HI on basis of the parameter sets retained as feasible (see also replies to the other reviewers). (see end of section 3.2.1 and Figure 11b). It seems that the best parametrisation provides also the best hysteresis modelling.

"The hysteresis index sensitivity to parameter uncertainty increases with the number of parameters from M1 to M2 and then stays in the same range from M2 to M4 (Figure 11b). This analyse confirms the importance of considering the Hysteresis Indices both between saturated and unsaturated storage (HSS and HUS) to avoid accepting an inadequate model architecture. For example, considering only the performance on the HSS(Q) relationship could lead to accept model M1 while its performance on HUS is lower and it is not able to reproduce the Riparian compartment hysteresis. For readability purposes, Figure 11b illustrates this sensitivity for the different HI in the year of 2011-2012 only but similar behaviour is observed every year. It showed that best behavioural parameters sets (bbp) lead to modelled HI values closer to the observed values than average modelled HI values. Using an additional calibration criterion related to the hysteresis could reduce the sensitivity of HI to parameter uncertainty and lead to narrow range of feasible parameter sets."

11. p. 5685, line 25. *"please recollect what are the four periods mentioned here."*

Reply: We proposed to summarize the four periods in the conclusion as following:

"Four periods have been identified along the hydrological year: **1) first, at the end of the dry period, rainfall starts to refill unsaturated storages; 2) in the wetting period, riparian unsaturated storage is filled and the saturated storage starts to supply the stream while hillslope unsaturated storage is still being replenished; 3) during the wet period, unsaturated storage in the hillslope is also filled and the saturated hillslope storage also feeds the stream. Finally when rainfall declines, flow from the riparian groundwater recedes and during the recession period, the stream discharge is sustained only by hillslope groundwater**."

12. Figures 7 and 8: *"just a suggestion: the years could be associated to the points in the graphs (e.g., "03-04", "10-11") so that the relationship with the other figures can be seen explicitly."*

Reply: We agree. A new version of Figures 7 and 8 is proposed at the end of this reply.

13. Figure 9: "*please indicate the direction of the loops"*.

Reply: The direction of the loops has been added on Figure 10 (previously Figure 9) at the end of this reply.

Reply to Anonymous Referee #2

"Most parts of the manuscript are well written and structured. Its scientific contribution will fit well into Hydrology and Earth System Sciences after some revision have been performed. Part from some more elaborations about the Hysteresis Index and some necessary shortening of subsection 3.1 I have two major comments".

We appreciate the reviewer's positive assessment of the manuscript. In the following, we provide further precisions about the discussion points highlighted by the reviewer and suggest practical modifications to integrate these comments.

Major comments:

- *"Already in the methodology the authors refer to another study (Hrachowicz et al., 2014, in revision at WRR) that is not available for the reader. In particular the reference to hydrological signatures that are not explained in the text or shown in the figures through the entire text made some of the interpretations and conclusions hardly understandable."*

Reply: We agree that the definition of the signatures was missing so we propose to provide additional information about this work in a new table (cf. New Table 4 at the end of this reply) for clarifying the objective functions used for calibrating the models, the hydrological signatures used for assessing them; and as new version of Figure 3 (provided at the end of this reply). It has to be noticed that the **manuscript submitted to WRR is now accepted** and available for further details on the previous work (doi: 10.1002/2014WR015484).

- *"In the description of the models and their parameters (which is partly referring to the above-mentioned study) the authors choose one final parameter set for each of the four models based on a weighted performance measure that only uses discharge observations. However, many preceding studies showed that models with more than 4-6 parameters face problems of over-parameterization when they only use discharge for calibration (Jakeman and Hornberger, 1993; Wheater et al., 1986). The low spread of weighted efficiencies/Euclidean distances in Fig3 in the manuscript might disprove that but the distributions shown there are re-shaped (with an exponent of 10) and might appear much more uniform in their original distribution. Since the model simulations are a substantial part of the interpretations and second part of the manuscript the authors need to provide some more information about the reliability of their models and the chosen parameters"*

Reply: We fully agree with the reviewer that, if insufficiently constrained, models with elevated numbers of parameters are subject to increased parameter uncertainty and associated predictive uncertainty, as many parameter combinations will merely provide a mathematically suitable fit while essentially misrepresenting the internal dynamics of the system as pointed out by many previous studies (e.g. Beven, 2006; Kirchner, 2006; Gupta et al., 2008; Andreassian et al., 2012). To reduce that problem while in the same time allowing for higher process complexity, we chose a double strategy for model calibration/selection in the Hrachowitz et al. (2014) manuscript: 1) multiple objective calibration based on 4 calibration objectives, which in

the past has been shown to be in itself already a valuable tool for identifying parameter sets that would otherwise be kept as feasible if only 1 calibration objective (e.g. Nash-Sutcliffe Efficiency) was used (e.g. Gupta et al., 1998; Seibert and McDonnell, 2002; Vrugt et al., 2003; Fenicia et al., 2007; see also a recent review paper on the topic by Efstradiadis and Koutsoyiannis, 2010), highlighting the different information content of objective functions based on different catchment signatures (Euser et al., 2013). 2) to further increase the confidence that the selected parameter sets actually reproduce the observed system dynamics to a certain extent, we complemented the multi-objective calibration strategy with the use of expert-knowledge and data driven parameter and process constraints to ensure that the selected parameter sets are in themselves consistent (e.g. the unsaturated storage capacity in wetlands needs to be lower in wetlands than on hillslopes) and that they reproduce system dynamics that do not contradict what we know about the system (e.g. unreasonably high/low long-term average base flow contributions or actual evaporation as estimated from the Budyko relationship) within certain limits of acceptability (e.g. Winsemius et al., 2009; Gharari et al., 2013; Gao et al., 2014). Applying these constraints a wide range of mathematically feasible parameter sets, violating these constraints, can be discarded significantly reducing the ill-posed nature of the problem. This is also reflected in our Figure 3: not only does the calibration performance and its spread improve with the progression of M1-M4, more importantly, the performance and its spread during VALIDATION also improves, indicating improved predictive power of the model, which in turn points toward potentially improved process representation.

We will clarify this in the revised manuscript. Please also note, that the performance measurements used, i.e. the Euclidean distances, are not weighted, rather the Euclidean distances themselves (together with an exponent of 10) were used as informal likelihood measurements (following the concept of GLUE) to construct uncertainty bounds around the modelled variables in the Hrachowitz et al. (2014) manuscript. Thus, the weights do not affect the actual performance assessment for the manuscript under review. We will therefore remove any reference to it in the revised version.

Detailed comments:

1. p. 5565, line 20*. "Please elaborate the link between the scale problem of lumped and (semi) distributed models and the storage behavior in a bit more detail".*

Reply: In order to clarify our statement we suggest adding the following precisions:

"A time-series of groundwater table level from a single piezometer is not representative of the behaviour of the groundwater, even at the hillslope scale; therefore it is difficult to link it with either a reservoir volume simulated by a lumped model or an average water table level of a grid point simulated by a fully distributed model."

2. p. 5667, line 7. *"Please also mention studies that used water quality data for model assessment, e.g A Hartmann, T Wagener, A Rimmer, J Lange, H Brielmann, M Weiler Water Resources Research 49 (6), 3345-3358 or Hartmann, A., Weiler, M., Wagener, T., Lange, J.,*

Kralik, M., Humer, F., Mizyed, N., Rimmer, A., Barberá, J. A., Andreo, B., Butscher, C. and Huggenberger, Hydrol. Earth Syst. Sci., 17(8), 3305–3321" .

Reply: The use of water quality data is indeed another example of multi-calibration studies, in the early version of the manuscript we cited only quantitative examples but we suggest adding the following references as examples of the use of tracer data in multi-data modelling approaches:

"chloride concentrations (Hrachowitz et al., 2011), atmospheric tracers (Molenat et al., 2013) or nitrates and sulfate concentrations (Hartmann et al, 2013 a), and water isotope as $\delta^{18}O$ (Hartmann et al., 2013 b)".

3. p. 5668, line 7. *"PET + drainage are smaller than precipitation - please elaborate. In addition, please mention also that there is a strong seasonal behavior. Otherwise the definition of the HI would be hard to understand (see comments below)."*

Reply: There is indeed a high water deficit in the annual budget of the catchment almost each year (PET+Q<R). It is true for the neighbour catchments and nested catchments too (not presented in the paper but also part of the ORE AgrHys). This deficit is likely to be due to underflows below the outlet, as it was suggested by previous studies (e.g. Ruiz et al. 2002). We propose to add the following sentences in the study site section to notify these points to the reader more explicitly:

"Mean annual rainfall over the period 1992-2012 is 1113 mm (+/-20%) and mean annual Penman potential evapotranspiration (PET) is 700 mm (+/- 4%). Mean annual drainage is 360 mm (+/- 60%) at the outlet. **There is a high water deficit in the annual budget almost each year due to underflows below the outlet (Ruiz et al., 2002)**. The catchment is laying under granite called leucogranodiorite of Plomelin, which upper part is weathered on 1 to more than 20 m deep. Soils are mainly sandy loam with an upper horizon rich in organic matter, depths are comprised between 40 and 90 cm. Soils are well drained except in the bottomlands which represent 7% of the total area. Agriculture dominates the land use with 86% of the total area. The base flow index is about 80 to 90%, thus the hillslope aquifer is the main contributor to stream (Molenat et al., 2008; Ruiz et al., 2002). **Both stream flow and shallow groundwater tables exhibit a strong annual seasonality in this catchment (Fig. 2)**".

And to illustrate the strong annual seasonality we suggest adding the hydrographs in new version of Figure 3.

4. p. 5669, line 9. *"I think you mean 'groundwater storage dynamics' and not groundwater storage', which would be related to volumes rather than dynamics"*

Reply: Yes the word "dynamic" has to be added: as explained in the introduction (and related to comment 1) the groundwater level and soil moisture measurements are more representative of the dynamics of the storages rather than of the volumes themselves.

5. p. 5669, l. 23-24. "*Rephrase*".

Reply: We propose the following reformulation of the sentence: "The two profiles are located on the upslope and downslope parts of the hillslope respectively. Thus, we assumed that

averaging their normalized values will allow us to build a proxy for the dynamics of the unsaturated zone storage on the whole hillslope".

6. p. 5670, lines 8-12, *"This is not clear. Please provide some more detail why an index has to be used."*

Reply: To make it clearer we propose to reformulate the sentence as

"For storage-discharge hysteresis at the annual scale, this approach is not sufficient as the same type of hysteretic loop is likely to happen for almost all the years **because despite of stream flow inter annual variations, the seasonality (with a high flow period during winter) is the same for all years**. **Moreover a preliminary cross correlation analysis revealed that storage and stream flow are strongly correlated**"

7. p. 5670-5671 Eq. (1): *"How often can you calculate this difference within one year? Is HI their mean? (by reading through the proceeding chapters it appears that there is a strong seasonality in discharge and these values might only be passed once a year, but this is not clear at this stage of the manuscript)", "define Qmid before eq. 1 and mention how often it occurs within one hydrological year.*"

Reply: Indeed, the strong seasonal cycle observed on the studied catchment allowed us to compute a HI based on a Q_{mid} value which is taken only 2 times per year : during the recharge period and during the recession period. Actually, high and low stream flow values are more likely to occur several times a year in this catchment than medium values. Moreover, we smoothed the data using a 7-day moving average to remove highest Q values due to rapid storms. We suggest explaining and presenting the presence of this seasonal cycle in the case study section (cf. Reply to comment 3 and illustration in new Figure 3), and adding some precision when defining Q_{mid} in this section:

"In this paper, **as the hydrological variables exhibit a strong annual uni-modal cycle**, we calculated the hysteresis index (HI) each year as the difference between water storages at the dates of mid-point discharge [...] In order to reduce the impact of the quick variations of discharge or groundwater level due to individual storm events, we smoothed the time series using 7-day moving averages. The strong seasonal discharge cycle led to identify two occurrences of Q_{mid} per year only: during the recharge period (t_R) and during the recession period (t_r) , while high and low stream flow values are taken several times per year as explained by Lawler et al. (2006).

8. P. 5671, lines 2-4*. "Parts of subsection 2.4 appear like a literature review that could also be part of the introduction*".

Reply: We would like to keep the literature related to the hysteresis descriptions in this section as it is really specific to this methodological point and we do not see how it could fit in the general introduction

9. P. 5671, lines 15. "*The authors should also add some more detail about their reasons to exactly choose this hysteresis index. Considering only 2 points instead of shape/rotation/etc. might omit some convolutions but it also gives the impression that a lot of information is ignored and misinterpretations might be possible, too*"

Reply: We agree that HI does not integrate a full description of the hysteresis and does not pretend to do so. It is only an index which gives already 2 types of information as explained in the paragraph, and allows a quantitative comparison between simulations and observations. Classification methods for storm events are indeed based on the rotational pattern (clockwise/anticlockwise), curvature (shape), and trend or rotational angle. Rotation pattern is given by the sign of HI. Curvature is defined from concave to convex, the trend is generally used to identify on concentrations-discharge hysteresis if the solutes are diluted (negative trend) or concentrated (positive trend) during a storm. Our hystereses are always concave with a positive trend. This pattern similar among years tends to support the hysteresis as a signature of internal catchment behaviour. Note also that all our interpretations are driven by the full observations and not only on the HI values.

But we agree that if the objective was to calibrate models, more analysis would be required to identify the best index (or a combination of several ones) to lose as little information as possible.

10. p. 5672 l. 7, "*Just considering water balance it appears obvious that M1 will not work (see my comment at the study site description)*".

Reply: We fully agree with the reviewer, yet, although many catchment worldwide exhibit similar water balance deficits due to deep infiltration losses/underflow (e.g. Le Moine et al., 2007), only a small minority of models actually caters for this process. M1 has been used in the Hrachowitz et al. (2014) manuscript as a starting point and benchmark model that resembles many frequently used models (e.g. HBV, Sacramento, FLEX, etc.) The use of such an overly simplistic model structure allowed us not only to illustrate that it cannot sufficiently well reproduce the hysteretic behaviour, but it also helped in model diagnosis to see where and how the model fails to reproduce the catchment behaviour. Thus, M1 was included merely for instructive purposes to demonstrate how wrong modelling can go with frequently used, yet unsuitable model architectures.

11. p. 5673, line 1, "*This is only three objective functions*".

Reply: There are four objective functions: the Nash Sutcliffe Efficiency Criterion applied to the stream flow (1), to the logarithm of stream flow (2), and to the flow duration curve (3), and finally the Volumetric Efficiency criterion applied to stream flow (4). In order to clarify this point we suggest presenting the four criteria with the hydrological signatures in the additional table (cf. New Table 4 at the end of this reply).

12. p. 5673, "*this is really high did Freer et al also use the Same value ? Is it really necessary to introduce p at all?*"

Reply: It is indeed a high value and Freer et al. (1996) even tested the effect of using an exponent of 30. To further answer the question: no it is not absolutely necessary to introduce an exponent for the informal likelihood measure. However, an exponent, in particular a high one, can serve two purposes: (1) it reduces the width of the uncertainty interval by giving relatively

little weights to poor model realizations, thereby addressing the frequently raised criticism of GLUE that it overestimates uncertainty (e.g. Mantovan and Todini, 2006; Stedinger et al., 2008) and (2) it reduces the sensitivity of uncertainty interval to the subjective choice of behavioural models, which is a second point frequently criticised in GLUE (e.g. Montanari, 2005).

Please note that the GLUE and the informal likelihood weights (+exponent) do not affect the actual performance assessment for the manuscript under review. We will therefore remove any reference to it in the revised version.

13. p. 5674, line 1. , p. 5680, line 23 and p. 5681, line 25 , p. 5682, line 26, *"Signatures are not defined anywhere and not shown".*

Reply: We suggest adding the signatures used in the modelling work presented in Hrachowitz et al., (2014) in an additional table (see new Table 4) and the performances of each model on these signatures.

14. p.5674 lines 3-16. "*This whole paragraph presents results that should be moved to the results section. Furthermore the authors explain differences among the model by state variables and fluxes that are not shown in Figure 3. It is not clear, which part of these results was done by Hrachowitz et al. 2014. "*

Reply: These results are a contribution from a previous work (Hrachowitz et al., 2014) therefore they are presented here in this material and method section as they are considered as previous results/knowledge for the present study. We agree with the reviewer that this has to be clarified (see also reply to reviewer 1). To make this point clearer we suggest an explicit reformulation in the introduction of this paragraph.

iii) to add a mention to this previous work at the end of the introduction :

" (...) ii) to which degree a suite of conceptual rainfall-runoff models with increasing complexity, **which were calibrated and evaluated for this catchment in a previous work, using a flexible modelling framework (Hrachowitz et al., 2014)**, can reproduce the observed storage-discharge hysteresis (...)"

iv) to explain this choice to the reader at the beginning of the corresponding section

"In a previous work, a range of 11 conceptual models were calibrated and evaluated for the Kerrien catchment in a stepwise development using a flexible modelling framework (see Hrachowitz et al., 2014). **This section aims at summarizing the results of this previous study that are used as a basis for the present work.**"

We also suggest providing additional information about this work as, especially the results of model calibrations in a new version of Figure 3.

Moreover the **manuscript submitted to WRR is now accepted** and available for further details on the previous work (doi: 10.1002/2014WR015484).

15. p.5674, line 26., p. 5683 line 20, *"It is very confusing to refer to the results of another study. It is also critical to go on only with the "best" parameter set, which might be very similar to other parameter sets in the sample if p was not applied on the original likelihoods. Is there any proof that the selected parameter set were sufficiently identifiable, ie. that there is no equifinality ? "*

Reply: The study of Hrachowitz et al. (2014) aimed at proposing a stepwise modelling approach where increasing model complexity (and increasing model number of parameters) was always associated with an increase of model constraints (parameter or architecture constraints) and always motivated by the need of reducing the predictive uncertainties, and the difference between calibration and evaluation period uncertainties (so called model consistency) rather than increasing model performance in the calibration. This approach limits the equifinality which may appear when increasing the model complexity (see also reply to major comment #2). To provide the reader more information about this point, we suggest adding some details from the previous work in the new version of Figure 3 with illustrations of model performances on the objective functions, for both calibration and independent evaluation period.

16. p.5676, line 12. *"It would be very helpful to show schematic figures /conceptual models that visualise these interpretations ".*

Reply: We thank the reviewer for this suggestion, an additional Figure with a scheme of the interpreted mechanisms is proposed as a new Figure 9 (see at the end of this reply).

17. p. 5676, *"the subsection above is also about observations in hill slope - please choose other header".*

Reply: Indeed the previous section is already dealing with hillslope observations so we suggest modified the 2 titles as "3.1.1 Observations in hillslope and riparian zones: saturated storage vs. Flow" and "3.1.2 Observations in hillslope: saturated and unsaturated storages vs. Flow".

18. p.5677, section 3.1.3. "*There is quite a lot of interpretation subsections 3.1.1 to 3.1.3. which overloads the manuscript combined with the proceeding modeling. I recommend to shorten this part (Subsections 3.1.1 to 3.1.3) significantly and providing conceptual drawings of the interpreted system behavior instead*".

Reply: We agree that the addition of a conceptual scheme would be useful. The proposed new Figure 9 would provide the required conceptual drawing. However, we propose to maintain the text, also as it has been appreciated by reviewer 1.

19. P.5680, line 24, *"As mentioned before, this is already obvious only by considering water balance..."*

Reply: we agree with the reviewer on the fact that model M1 can be expected as non consistent regarding our knowledge of the catchment (see reply to comment 10). However it is interesting to see that the hysteresis comparison shows clearly and immediately what is inconsistent and how the model behaved to compensate the error: "model M1, the overestimation of the hillslope saturated storage was partially compensated by the underestimation of the hillslope unsaturated storage".

1. p. 5681 line 28, *"this can only be answered when the model parameters are evaluated for their sensitivity and parameter interactions. With the present information equifinality in calibration could also be a very probable explanation"*

Reply: We included a sensitivity analysis of HI on the basis of the parameter sets retained as feasible (see below and also replies to the other reviewers). These results have been integrated at the end of section 3.2.1 and in Figure 11b. About the equifinality see also reply to comment 15.

"The hysteresis index sensitivity to parameter uncertainty increases with the number of parameters from M1 to M2 and then stays in the same range from M2 to M4 (Figure 11b). This analyse confirms the importance of considering the Hysteresis Indices both between saturated and unsaturated storage (HSS and HUS) to avoid accepting an inadequate model architecture. For example, considering only the performance on the HSS(Q) relationship could lead to accept model M1 while its performance on HUS is lower and it is not able to reproduce the Riparian compartment hysteresis. For readability purposes, Figure 11b illustrates this sensitivity for the different HI in the year of 2011-2012 only but similar behaviour is observed every year. It showed that best behavioural parameters sets (bbp) lead to modelled HI values closer to the observed values than average modelled HI values. Using an additional calibration criterion related to the hysteresis could reduce the sensitivity of HI to parameter uncertainty and lead to narrow range of feasible parameter sets."

2. p. 5686 line 4-8. *"Large part of this paragraph is referring to Hrachowitz et al. 2014, which was not part of this study".*

Reply: The previous work of Hrachowitz et al. (2014), analyzed the model ability to reproduce the classical hydrological signatures but the comparison with the performance on the hysteretic signature is actually a result of the present work. We propose to reformulate in order to avoid any ambiguity:

"The tested models were characterized by an increasing degree of complexity and also an increasing consistency, as shown in a previous study using classical hydrologic signatures. In this study, we showed that if all of them simulated a hysteretic relationship between storages and discharge, their ability to reproduce hysteresis index also increased with model complexity. In addition, we suggest that if classical hydrological signatures help to assess model consistency, the hysteretic signatures help also to identify quickly when the models give "right answer for wrong reasons" and can be used as a descriptor of the internal catchment functioning."

Reply to Anonymous Referee #3

"The manuscript is a relevant and important step towards diagnostic analysis of hydrologic models, with specific contribution through testing internal hydrologic process representation via comparing observed and simulated hysteretic patterns that exist between storage and streamflow discharge in the selected watershed. I think the topic is of interest to the HESS readership and the manuscript is well written, well-structured and the use of language is generally good. However, part of the analysis is described at an abstract level with frequent citation to a recent unpublished manuscript (Hrachowitz et al., 2014) by the authors. Therefore a clear distinction between the contributions by these two manuscripts should be made in the manuscript.

Reply: We thank the reviewer for the positive assessment of our article.
As highlighted also by the other reviewers the distinction with the results from the previous study of Hrachowitz et al. should be made clearer, this can be done by the addition of explicit sentences at the end of the introduction and when the results of this previous work are presented, i.e. in the "materials and methods" section, as they are previous knowledge for the present work.
Also we suggest to provide additional information about these previous results in a new table (cf. New Table 4 at the end of this reply) for clarifying the objective functions used for calibrating the models, the hydrological signatures used for assessing them; and in a new version of Figure 3 for illustrating the results obtained from this previous study (provided at the end of this reply).

It has to be noticed that the **manuscript submitted to WRR is now accepted** and available for further details on the previous work (doi: 10.1002/2014WR015484).

"I also suggest below a few cases where explicit discussion of the analysis should be provided. Overall, my assessment for the manuscript is minor revisions. The manuscript could be published after the authors address the comments listed in "Main Comments" and "Minor Comments" sections listed below."

Reply: We also thank the reviewer for his/her comments and suggestions, our corresponding discussions and suggested modifications to take these comments into account are detailed below.

Main comments:

1. *"The authors state that 86% of the study area is dominated by agriculture (Section 2.1). A discussion on the source of irrigation water (groundwater?) and how the agricultural use could affect the hysteretic patterns should be provided in the manuscript. Similar discussion related to percentage of snow and its possible impact on the results should be provided."*

Reply: Due to the temperate oceanic climate, there is no snow cover on the study catchment which is located only 10 km from the mouth of the Odet River in the Atlantic Ocean. This climate is indeed characterized by rainy and mild winters (with minimal temperature around 5.9 Celsius deg) and relatively wet and cool summers. The average total rainfall in the summer is more than 200 mm, and the numbers of day with rain per year is around 150 days in average. Due partly to

this abundance of regular rainfall, agriculture in Brittany represents only 4 % of water uptakes. Irrigation is used in only 2% of the agricultural area and mainly for vegetable cropping. Irrigation is absent in the study catchment. We added in section 2.1:

"Agriculture dominates the land use with 86% of the total area **covered by grassland, maize and wheat, none of them irrigated."**

2. *"Hysteresis Indices (Section 2.4): The description of the mid-point discharge is not clear. In addition the manuscript lacks a hydrograph, which many hydrologists would very much like to see in a manuscript related to hydrologic processes and models. Therefore the authors should provide a representative hydrograph of the watershed with clear description of the mid-point discharge values on the hydrograph. Below the hydrograph a time series of water levels/moisture levels should be provided again indicating the selected points used in calculating the Hysteresis index. A figure as described above is very important for understanding the HI concept used in the study. My main concern is that the HI concept followed in the manuscript is only specific to the selected watershed. I also think that mid-point discharge could occur multiple times during recharge and recession periods, therefore which time to select should be clearly described in the manuscript".*

Reply: (See also reply to reviewer 1 and 2) In order to clarify the mi-point values we proposed to add a new Figure (Figure 9 the end this reply) that helps to identify that within the seasonal pattern observed on the studied catchment, the mid-point discharge value is taken only twice per year during the recharge and the recession periods respectively; and to add the observed and simulated hydrographs in new version of Figure 3. Regarding the observed hysteresis over the 10 years, the choice of Q_{mid} succeeded in catching the difference of the saturated storage states in recharge and recession more or less in the middle of these periods. This behaviour is not specific to the studied catchment, and is also supported by previous studies e.g. Lawler et al., (2006) who "argue that computing HI by using mid-point discharge usually allows avoiding the small convolutions which are frequently observed at both ends of the hysteretic loop." (p.5670, lines 17-19).

We agree that HI does not integrate a full description of the hysteresis and does not pretend to do so. However it gives already 2 types of information as explained in the paragraph and allows a quantitative comparison between simulations and observations. It is relevant for studying annual pattern with strong seasonal cycle. It could be used similarly for flood event if it is uni-modal. For multiple-seasonality, e.g. if there are 2 recharge periods in the year, the hysteresis is likely to exhibit a double loop and 2 indices may be relevant to describe each of them. In particular one can imagine that 2 successive loops may have different directions (so different signs of HI) due to the successive activation of different flow paths and the fact that storages are likely to be less empty at the beginning of the second recharge than at the beginning of the first one. In snow-melt driven catchment, the hysteresis relative to the snow cover storage should be taken into account too (as a third storage). In arid catchment where the groundwater recession can occur during several years (see e.g. Ruiz et al., 2010), it would be more relevant to compare these relationships among the identified pluri-annual cycles composed by at least both a recharge and a recession rather than at the annual scale. As cited in the corresponding section 2.4, some authors prefer to describe the loop width using the extreme values of the Y variable (X

variable is always stream flow, Y is storage in our case but often either a concentration or the turbidity in hysteresis studies).

In order to help the identification of the limits of our HI we propose to add the following precisions:

"In this paper, **as the hydrological variables exhibit a strong annual uni-modal cycle,** we calculated the hysteresis index (HI) each year as the difference between water storages at the dates of mid-point discharge in the two phases of the hydrological year [...]"In order to reduce the impact of the quick variations of discharge or groundwater level due to individual storm events, we smoothed the time series using 7-day moving averages. The strong seasonal discharge cycle led to identify two occurences of Q_{mid} per year only: during the recharge period (t_R) and during the recession period (t_r), while high and low stream flow values are taken several times per year as explained by Lawler et al. (2006)."

3. *"Model calibration and Evaluation (Section 2.6): It seems that the whole section is taken from Hrachowitz et al. (2014). This should be stated right at the beginning of the section and specific contributions should be clarified. Overall, model calibration and evaluation steps need further explanations and clarifications in the current manuscript. First the selected likelihood measure is not mathematically correct and needs further discussion on the validity and specifically selection of parameter p=10. The authors should include a figure showing only a selected single 2-D representation of the 4-D objective function space to show the projected pareto surface together with the uncertainty intervals. The readers will then be able to understand the calibration and uncertainty analysis procedures with above information. Also, authors state that 13 signatures were used for evaluation, however there is no description of these signatures (perhaps only four is given at an abstract level; Page 5673-Line 22). The signatures used should be explicitly stated".*

Reply: Results from the previous work in Hrachowtiz et al. (2014) are used as a basis of the present work. They are presented in sections 2.5 and 2.6 of the Material and Method section as they are considered as previous results/knowledge. To clarify this, as suggested by the reviewers, we propose:

v) to add a mention to this previous work at the end of the introduction :

" (...) ii) to which degree a suite of conceptual rainfall-runoff models with increasing complexity, **which were calibrated and evaluated for this catchment in a previous work, using a flexible modelling framework (Hrachowitz et al., 2014)**, can reproduce the observed storage-discharge hysteresis (...)"

vi) to explain this choice to the reader at the beginning of the corresponding section

"In a previous work, a range of 11 conceptual models were calibrated and evaluated for the Kerrien catchment in a stepwise development using a flexible modelling framework (see

Hrachowitz et al., 2014). **This section aims at summarizing the results of this previous study that are used as a basis for the present work.**"

(iii) and to provide additional information about this work in a new table (cf. **New Table 4** at the end of this reply) for clarifying the objective functions used for calibrating the models, the hydrological signatures used for assessing them; and in a new version of Figure 3 (provided at the end of this reply).

Moreover, the **manuscript submitted to WRR is now accepted** and available for further details on the previous work (doi: 10.1002/2014WR015484).

(iv) In spite of ongoing discussions of the most suitable technique to assess uncertainty and criticisms of GLUE for using formally incorrect likelihoods (e.g. Beven, 2006, 2008; Mantovan and Todini, 2006; Stedinger et al., 2008; Montanari et al., 2005; Clark et al., 2012; Hrachowitz et al., 2013a), the use of GLUE with its informal likelihood measures, as used here, has proven valuable in many studies in the past. Fully acknowledging the limitations of the approach, we would, however, also argue that other approaches using formally correct likelihoods suffer from other limitations. To further answer the comment, using likelihood measures that are unweighted or that have relatively small weighting exponents result in a relatively high sensitivity of uncertainty intervals to the choice of the threshold. In this study, we used a relatively high exponent p = 10 (which is still lower as p=30, as tested by Freer et al., 1996) to weight the informal likelihood measure (here: DE), so as to give higher weights to models with good performance and penalize models with poor performance, in an attempt to reduce the impact of subjectivity of the choice of feasible parameters in the results. High values of p significantly reduce the sensitivity of the uncertainty intervals of the modelled variables to changes in the subjectively chosen thresholds. In other words, the value of the threshold becomes irrelevant with high p as all performance thresholds will produce effectively identical uncertainty intervals.
However, we believe that a more detailed description of the calibration and uncertainty assessment strategy used here (in particular as it is in detail given in the referenced manuscript) is somewhat out of the scope of this manuscript and will distract the reader from the actual story.

4. *"A sensitivity analysis investigating the sensitivity of the hysteretic pattern simulation to the model parameters will significantly improve the manuscript. Currently it is not clear whether the improvements are solely due to the increase in the number of model parameters, or rightly due to addition of new conceptual component to the model as the complexity is increased".*

Reply: The previous study aimed at proposing a stepwise modelling approach where increasing model complexity (and increasing model number of parameters) is always associated with an increase of model constraints (parameter or architecture constraints) and always motivated by the need of reducing the predictive uncertainties, and the difference between calibration and evaluation period uncertainties (so called model consistency) rather than increasing model performance in the calibration. This approach limits the equifinality which may appear when increasing the model complexity. To provide the reader more information about this point, we suggest adding the relevant results from the previous work in the new version of Figure 3 with

illustrations of model performances on the objective functions and on the signatures, for both calibration and independent evaluation period.

We agree with the reviewer and therefore we included a sensitivity analysis of HI on basis of the parameter sets retained as feasible (see also replies to the other reviewers). According to this analysis, when looking only at one of the hysteretic relationships such as the Hillslope saturated zone-flow, the increase of parameter number does not help to improve the hysteresis modelling, but taking into account the hysteresis between all components simultaneously, the increase of complexity allows for an improved overall simulation.

"The hysteresis index sensitivity to parameter uncertainty increases with the number of parameters from M1 to M2 and then stays in the same range from M2 to M4 (Figure 11b). This analyse confirms the importance of considering the Hysteresis Indices both between saturated and unsaturated storage (HSS and HUS) to avoid accepting an inadequate model architecture. For example, considering only the performance on the HSS(Q) relationship could lead to accept model M1 while its performance on HUS is lower and it is not able to reproduce the Riparian compartment hysteresis. For readability purposes, Figure 11b illustrates this sensitivity for the different HI in the year of 2011-2012 only but similar behaviour is observed every year. It showed that best behavioural parameters sets (bbp) lead to modelled HI values closer to the observed values than average modelled HI values. Using an additional calibration criterion related to the hysteresis could reduce the sensitivity of HI to parameter uncertainty and lead to narrow range of feasible parameter sets."

5. *"Low flow signatures vs. hysteresis patterns: Overall it can be seen that (e.g. Page 5680, Lines 22-23) hysteresis patterns are associated with the low flow signatures as expected. Although authors state briefly, an analysis showing the correlation between the low flow signatures to the hysteretic patterns should be provided. Perhaps a figure could be added showing the low flow signature performance vs. hysteretic index performance of different model structures. Currently low flow signatures are not analyzed independently in the manuscript to investigate their link to the hysteresis signature".*

Reply: at this stage are HI performances are not measured the problem is that we already know that HI is not sufficient to fully describe the hysteresis (as discussed in the 3 reviews in comments related to HI) and therefore was used as an assessment quantitative index but not as a calibration criterion. As discussed in section 3.3, perspectives from this work are a second step where a set of indices would be used to build a process-based objective functions usable for calibration. From our point of view a lot of work is still required to achieve such a set of indices, which would be the focus of a future work.

6. *"Sensitivity of HI to annual rainfall (Section 3.1.5): The annual hysteretic patterns are sensitive to the timing of rainfall however the sensitivity to the annual rainfall is tested. My concern is that the HI will be sensitive to when the rainfall occurred (recharge period, recession period etc.) but less on the total annual rainfall. A discussion is needed".*

Reply: This is an interesting point raised here. Indeed, annual rainfall is only a proxy for annual recharge. In our case, it seems that in fact when the annual amount of rainfall increases it is related to an increase of precipitations mostly in the wet season i.e. the high flow period. Therefore, when the amount of annual rainfall increases, the recharge is almost unchanged

while the recession is delayed. If the beginning of the recession is delayed, groundwater levels (saturated zone storage) will be still high when stream flow reaches the mid-point value.

7. *"The authors presented the degree of hysteretic pattern mismatch between models with different complexities (Figures 9 and 10). A modeler will immediately wonder why the authors did not re-calibrate their models to improve the hysteretic patterns? Calibration using hysteretic patterns could provide additional information on the validity of the model structures and help to understand trade-offs in matching flow based vs storage based signatures. This comment is also linked to Comment 6 which is related to Sensitivity Analysis*"

Reply: As in the previous study, Hrachowitz et al. (2014) did not use the signatures in the calibration procedure but rather use them as a post-calibration diagnostic tool for complementary evaluation of the model performances. We kept here the same approach, assuming that hysteresis pattern would provide other information about the model behavior than classical hydrological criteria. Proper calibration on the hysteretic signatures will require a full set of indices to really constraint the hysteretic pattern (see also reply to comment 5) and will be considered in further work.

8. "Conclusion (Section 4): P5686, L4-7: "They were previously calibrated using classical objective functions and assessed using classical hydrological signatures, and their overall performance at reproducing hysteretic signatures was consistent with their overall performance at reproducing the classical signatures. The analysis of the simulated hysteresis signatures helps to identify why the model may give a right answer for wrong reasons" The above statements by the authors are undermining their work and conflicting. According to the first sentence, one can conclude that the classical hydrologic signatures provided the same information as the hysteretic signatures with respect to the overall performance of the model (both say right model or wrong model). The second sentence is then conflicting since classical signatures are already capable of identifying right model from wrong model (right model for right reason). The authors should provide more in depth discussion about their contribution".

Reply: The first statement is that the general improvement of model performance with increasing complexity observed on the classical hydrological signatures is consistent with a general improvement of the hysteretic signatures too, e.g. from all signatures point of view M4 is more consistent than M1. The second sentence states that the hysteretic signatures might help to identify why a model is wrong when it is assessed wrong by both groups of signatures: e.g. in model M1, performance is decreasing for low flows. It is visible on the low flow signatures and on the hysteresis in dry (low-flows) years. Looking at the HI values, one can see that the model M1 systematically under estimate the unsaturated hillslope storage. In other words, the model calibration tends to put all the water in the saturated storage for reaching the high flow values quickly, therefore in low flow there is not enough water stored in the unsaturated storage and dynamic is wrong. To summarize in other words we could say that classical and hysteretic signatures allow to identify when a model is wrong (or right for wrong reasons) and that additionally, hysteretic signatures allow to identify why the model is wrong (or for which wrong reasons) in a quick and easy way.

We propose to avoid ambiguity the following reformulation:

"The tested models were characterized by an increasing degree of complexity and also a increasing consistency, as shown in a previous study using classical hydrologic signatures. In this study, we showed that if all of them simulated a hysteretic relationship between storages and discharge, their ability to reproduce hysteresis index also increased with model complexity. In addition, we suggest that if classical hydrological signatures help to assess model consistency, the hysteretic signatures help also to identify quickly when and why the models give "right answer for wrong reasons" and can be used as a descriptor of the internal catchment functioning.

9. *"The authors should be more careful about the use of plural nouns, e.g. P5684, L23: soils moisture, P5683, L18: parameters sets etc. Please check throughout the manuscript as many more exist."*

Reply: all the manuscript has been reviewed with a particular attention to this point.

Minor comments

Figure 4: "*An explanation on how the recharge and discharge periods are represented on Figure 4 should be included*"
Reply: Recharge and recession periods are both indicated in Figure 4 (see new version of Figure 4 below).

Figure 8 a. "*The markers overlap and hence some of them are hidden behind. Improve the marker representation*"
Reply: A new version of Figure 8 is proposed at the end of this reply

p. 5679, line 14: *"There is no information related to unsaturated zone in Table 4. Page 5680, Line 24: Figure 3 does not have sub figures a and c"*
Reply: The information related to the unsaturated zone is the initial storage of this zone denoted Initial HUS.

p. 5684, line 6-7: "*Clarify the sentence*"

Reply: We propose to reformulate as

"We argue that rather than increasing the number of constraints or objective functions to satisfy, an alternative could be **to use some objective functions based on a combination of different variables** as stream flow and the groundwater level, soil moisture, or stream concentrations."

p. 5676, line 22. *"Clarify the sentence"*

Reply: We propose to reformulate as

"First, **stream flow was close or equal to zero and was almost exclusively sustained by drainage of the saturated storage**, while the unsaturated zone exhibited a significant storage deficit and only minor fluctuations due to transpiration and small summer rain events (dry period)"

"Page 5682, Line 16: Replace "discharge" with "recharge"."

"Page 5684, Line 11: Replace "though" with "through"."

“Page 5684, Line 2: “realisms constraints” choose one.”

“Page 5683, Line 28: Remove “objectives”.”

“Page 5674, Line 13: Replace “lower that” with “lower than that”.”

“Page 5679, Lines 1-3: Typos with regard to citations.”

All the spelling mistakes and typos regarding to citations raised by the reviewers have been corrected.

Additional references in reply to reviewers

Clark, M. P., D. Kavetski, and F. Fenicia. 2012. Reply to comment by K. Beven et al. On ''Pursuing the method of multiple working hypotheses for hydrological modeling'', Water Resource Research, 48, W11802.

Efstradiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modeling: a review. Hydrological Sciences Journal, 55: 1, 58-78.

Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister, 2007, A comparison of alternative multiobjective calibration strategies for hydrological modelling. Water Resource Research, 43, W03434

Gao, H., Hrachowitz, M., Fenicia, F., Gharari, S. A and Savenije, H.H.G. 2014, Testing the realism of a topography driven model (FLEX-Topo) in the nested catchments of the Upper Heihe, China, Hydrological Earth System Sciences, 18, 1895-1915.

Hartmann, A., Wagener, T. , Rimmer, A. , Lange, J., Brielmann, H., and Weiler, M. 2013 a. Testing the realism of model structures to identify karst system processes using water quality and quantity signatures. Water Resources Research, 49, 3345–3358, doi:10.1002/wrcr.20229.

Hartmann, A., Weiler, M., Wagener, T., Lange, J., Kralik, M., Humer, F., Mizyed, N., Rimmer, A., Barberá, J. A., Andreo, B., Butscher, C., and Huggenberger, P. 2013b. Process-based karst modelling to relate hydrodynamic and hydrochemical characteristics to system properties, Hydrological Earth System Science., 17, 3305-3321, doi:10.5194/hess-17-3305-2013.

Jakeman, AJ. And Hornberger, GM., 1993. How much complexity is warranted in a rainfall-runoff model. Water Resources Research, 29 (8), 2637-2649.

Le Moine, N., Andréassian, V., Perrin, C. and Michel, C. 2007. How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments, Water Resource Research, 43, W06428.

Mantovan, P. and Todini, E., 2006. Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. Journal of Hydrology, 330(1-2), 368-381.

Molenat, J.,Gascuel-Odoux, C., Aquilina, L., Ruiz, L. 2013. Use of gaseous tracers (CFCs and SF6) and transit-time distribution spectrum to validate a shallow groundwater transport model. Journal of Hydrology, 480 , 1–9.

Montanari, A. 2005. Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations, Water Resource Research, 41, W08406.

Ruiz, L., Varma, MRR., Kumar, MSM., Sekhar, M., Marechal, JC., Descloitres, M., Riotte, J., Kumar, S., Kumar, C., Braun, JJ. 2010. Water balance modelling in a tropical watershed under deciduous forest (Mule Hole, India): Regolith matric storage buffers the groundwater recharge process. Journal Of Hydrology, 380(3-4), 460-472, doi:10.1016/j.jhydrol.2009.11.020

Stedinger, JR., Vogel, RM., Lee, SU., and Batchelder, R. 2008. Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. Water Resource Research, 44, W00B06.

Vrugt, J., Gupta, H. V., Bastidas, L. A., Bouten, W., and Sorooshian, S. 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models, Water Resources Research, 39(8), 1214.

Wheater HS., Bishop KH., Beck MB., 1986. The identification of conceptual hydrological models for surface-water acidification. *Hydrological Processes*, 1, 89-109.

Winsemius, H. C., B. Schaefli, A. Montanari, and H. H. G. Savenije, 2009. On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. Water Resource Research. 45, W12422.

Table 4: Model calibration performances of the 4 calibration objectives used and post-calibration evaluation with respect to 13 additional hydrological signatures. The performance metrics include the Nash-Sutcliffe Efficiency (E_{NS}; Nash and Sutcliffe, 1970), the Volume Error (E_V; Criss and Winston, 2008) and the Relative Error (E_R; e.g. Euser et al., 2013). For all variables and signatures, except for Q, Qlow and GW, the long-term averages were used.

	Variable/Signature	Performance metric	Performance							
			M1		M2		M3		M4	
			Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation
Calibration	Time series of flow	$E_{NS,Q}$	0.82 (0.68/0.81)	0.51 (0.25/0.56)	0.84 (0.37/0.82)	0.64 (0.10/0.63)	0.85 (0.19/0.78)	0.59 (0.16/0.58)	0.85 (0.40/0.80)	0.59 (0.08/0.59)
		$E_{NS,log(Q)}$	0.71 (0.45/0.73)	0.66 (0.42/0.67)	0.76 (0.24/0.67)	0.72 (0.27/0.68)	0.75 (0.34/0.68)	0.75 (0.37/0.66)	0.75 (0.40/0.70)	0.74 (0.49/0.72)
		$E_{V,Q}$	0.67 (0.55/0.67)	0.48 (0.36/0.48)	0.74 (0.35/0.69)	0.64 (0.32/0.60)	0.75 (0.35/0.66)	0.62 (0.30/0.58)	0.74 (0.43/0.68)	0.61 (0.36/0.58)
	Flow duration curve	$E_{NS,FDC}$	0.92 (0.63/0.87)	0.85 (0.53/0.85)	0.96 (0.33/0.82)	0.87 (0.26/0.99)	0.96 (0.67/0.99)	0.96 (0.47/0.99)	0.96 (0.71/0.99)	0.96 (0.54/0.99)
	Calibration Euclidean Distance[a)]	$D_{E,cal}$	0.12 (0.12/0.22)	0.20 (0.19/0.31)	0.09 (0.13/0.33)	0.15 (0.16/0.38)	0.09 (0.13/0.31)	0.15 (0.17/0.34)	0.09 (0.12/0.26)	0.15 (0.16/0.32)
Evaluation	Groundwater dynamics[b)]	$E_{NS,GW}$	-0.07 (-0.52/-0.01)	-0.17 (-0.56/-0.06)	0.88 (0.17/0.95)	0.87 (0.46/0.94)	0.84 (-0.30/0.95)	0.84 (0.23/0.95)	0.93 (-0.33/0.93)	0.93 (0.20/0.94)
	Flow duration curve low flow	$E_{NS,FDC,low}$	0.83 (0.14/0.68)	0.75 (0.07/0.69)	0.95 (-0.94/0.97)	0.74 (-0.57/0.99)	0.96 (0.32/0.97)	0.97 (-0.13/0.99)	0.94 (0.35/0.97)	0.96 (0.06/0.99)
	Flow duration curve high flow	$E_{NS,FDC,high}$	0.91 (0.68/0.98)	0.64 (0.42/0.81)	0.93 (0.37/0.95)	0.70 (0.48/0.91)	0.99 (0.04/0.96)	0.91 (0.57/0.91)	0.92 (0.10/0.97)	0.80 (0.58/0.93)
	Groundwater duration curve[b)]	$E_{NS,GDC}$	-0.07 (-0.52/-0.01)	-0.17 (-0.56/-0.06)	0.88 (0.17/0.95)	0.87 (0.46/0.94)	0.84 (-0.30/0.95)	0.84 (0.23/0.95)	0.93 (-0.33/0.93)	0.93 (0.20/0.94)
	Peak distribution	$E_{NS,PD}$	0.23 (0.29/0.94)	0.72 (0.37/0.95)	-0.36 (-3.45/0.97)	0.62 (-1.04/0.98)	0.43 (0.33/0.99)	0.61 (0.45/0.98)	0.34 (0.34/0.99)	0.60 (0.49/0.98)
	Peak distribution low flow	$E_{NS,PD,low}$	-2.60 (-1.89/0.94)	0.34 (-0.42/0.94)	-3.81 (-16.7/0.92)	0.26 (-3.55/0.92)	-1.29 (-1.55/0.96)	0.19 (-0.14/0.96)	-1.85 (-1.57/0.97)	0.19 (-0.05/0.96)
	Rising limb density	$E_{R,RLD}$	0.75 (-0.05/0.86)	0.83 (0.27/0.89)	0.90 (0.84/0.99)	0.93 (0.83/0.98)	0.98 (0.90/0.99)	0.95 (0.82/0.91)	0.99 (0.90/0.99)	0.89 (0.82/0.92)
	Declining limb density	$E_{R,DLD}$	0.28 (-0.86/0.42)	0.45 (-0.07/0.63)	0.47 (0.80/0.98)	0.60 (0.77/0.96)	0.63 (0.74/0.97)	0.75 (0.78/0.97)	0.73 (0.72/0.99)	0.90 (0.80/0.98)
	Auto-correlation function of flow[c)]	$E_{NS,AC}$	0.98 (0.91/0.99)	0.26 (0.10/0.87)	0.99 (0.04/0.99)	0.36 (0.48/0.95)	0.94 (-0.03/0.97)	0.40 (0.62/0.97)	0.96 (0.40/0.96)	0.32 (0.61/0.97)
	Lag-1 auto-correlation of high flow	$E_{R,AC1,Q10}$	0.24 (0.23/0.28)	0.80 (0.78/0.86)	0.25 (0.36/0.91)	0.79 (0.59/0.98)	0.26 (0.37/0.91)	0.81 (0.66/0.91)	0.30 (0.36/0.78)	0.85 (0.78/0.98)
	Lag-1 auto-correlation of low flow	$E_{R,AC1,low}$	0.48 (0.48/0.49)	0.90 (0.89/0.91)	0.48 (0.50/0.95)	0.91 (0.57/0.96)	0.52 (0.56/0.96)	0.92 (0.77/0.99)	0.53 (0.57/0.97)	0.94 (0.79/0.99)
	Runoff coefficient[d)]	$E_{R,RC}$	0.84 (0.73/0.92)	0.65 (0.60/0.67)	0.93 (0.75/0.97)	0.88 (0.67/0.94)	0.93 (0.76/0.98)	0.86 (0.70/0.90)	0.93 (0.79/0.99)	0.85 (0.73/0.96)
	Evaluation Euclidean Distance[e)]	D_E	0.13 (0.17/0.27)	0.17 (0.18/0.27)	0.08 (0.09/0.29)	0.08 (0.08/0.22)	0.07 (0.07/0.19)	0.06 (0.06/0.13)	0.07 (0.07/0.18)	0.07 (0.06/0.11)

[a)] Euclidean Distance to perfect model with respect to the 4 calibration objectives

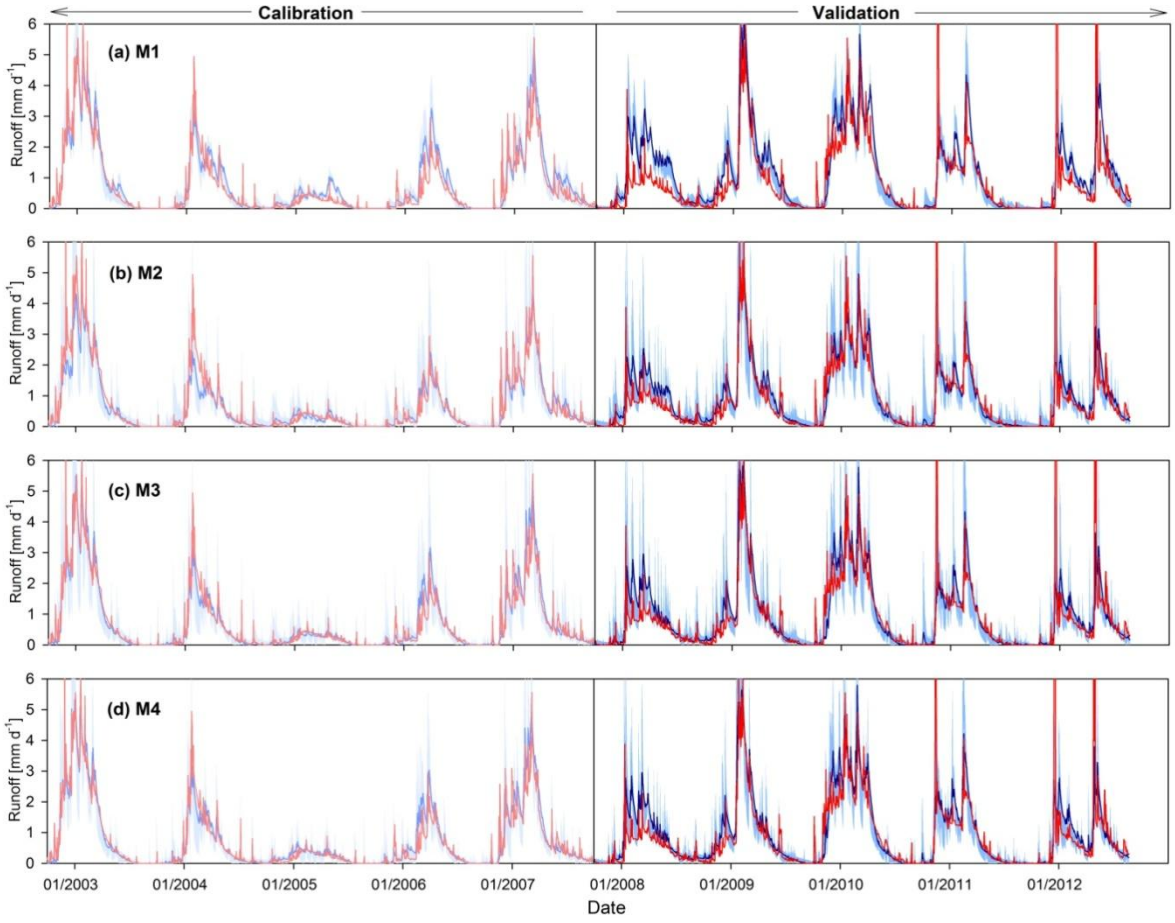
[b)] Averaged and normalized time series data of the five piezometer were compared to normalized fluctuations in model state variable SS

[c)] Describing the spectral properties of a signal and thus the memory of the system, the observed and modelled auto-correlation functions with lags from 1-100d where compared

[d)] Note that in catchments without long-term storage-changes and inter-catchment groundwater flow, long-term average RC equals the long-term average 1-EA

[e)] Euclidean Distance to perfect model with respect to all above given performance metrics

Figure 3. a. Observed (red line) and modelled runoff for model set-ups (a) M1, (b) M2, (c) M3 and (d) M4 in calibration and independent evaluation (validation) periods. Modelled runoff shown as most balanced solution (dark blue line) and the 5/95th uncertainty bounds (light blue shaded area). Adapted from Hrachowitz et al. (2014).



b. Overall model performance for all model set-ups (M1-M4) expressed as Euclidean Distance from the "perfect model" computed from all calibration objectives and signatures with respect to calibration and validation periods. Triangles represent the optimal solution, i.e. the solution obtained from the parameter set with the lowest Euclidean Distance during calibration. Box plots represent the Euclidean Distance for the complete sets of all feasible solutions (the dots indicate 5/95th percentiles, the whiskers 10/90th percentiles and the horizontal central line the median). Adapted from Hrachowitz et al. (2014).

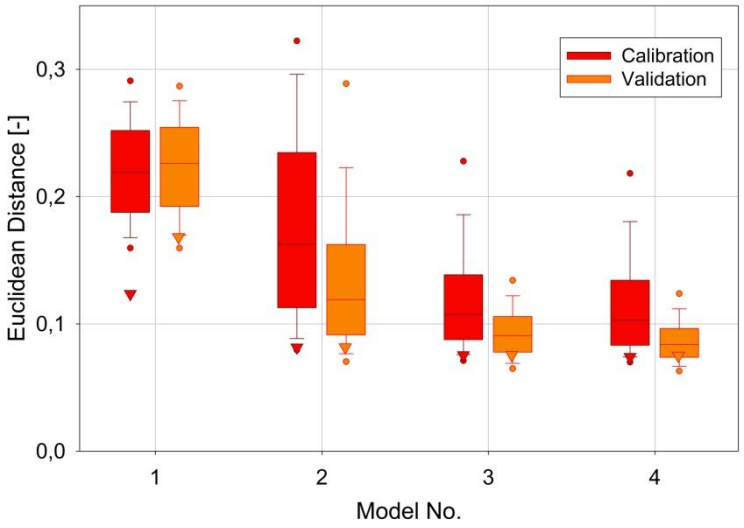


Figure 4. Examples of annual hysteretic loops for saturated zone storage vs. stream flow which are clockwise on the riparian zone (a, b) and anticlockwise on the hillslope (c, d) for the wet year 2003-2004 (a, c) and the dry year 2007-2008 (b, d).

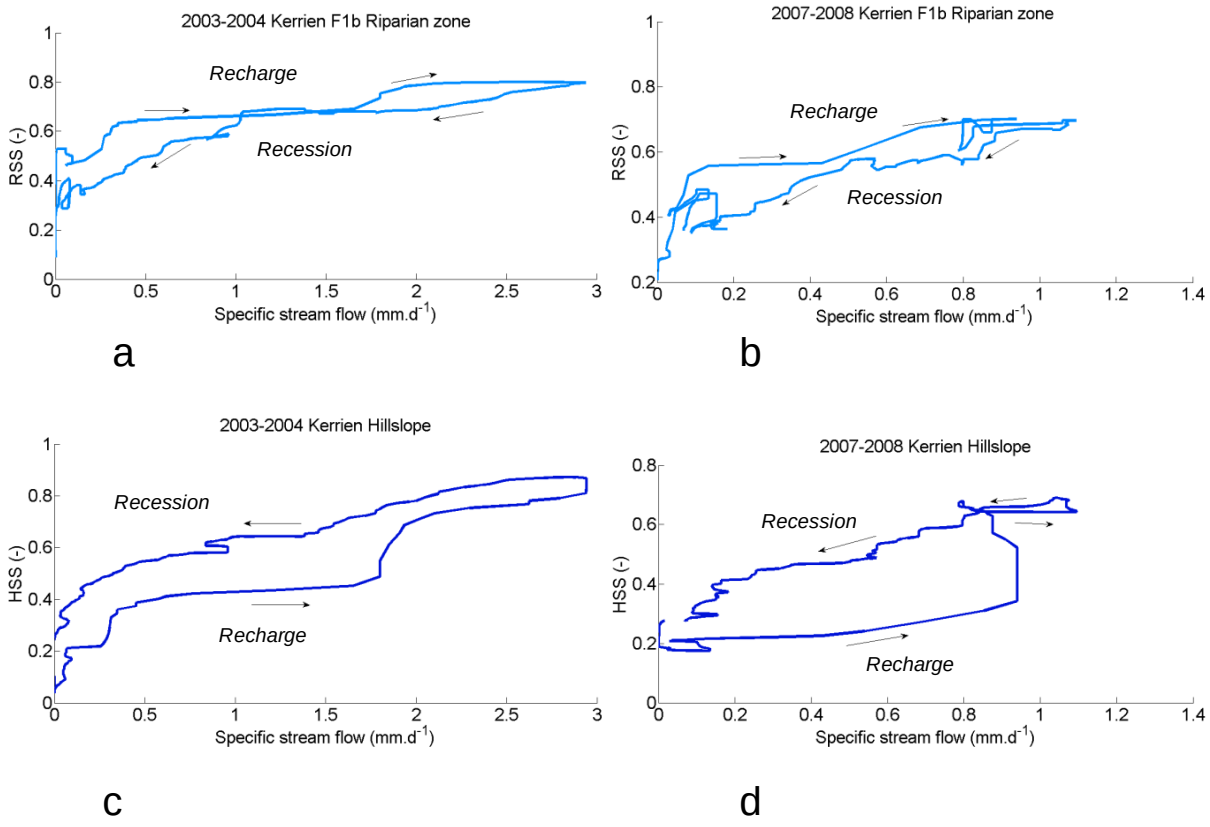


Figure 7. Year to year variations, for the 10 monitoring years, of the hysteresis indices a) HSS-F5b(Q) and HSS-F4(Q) (HI) versus the initial groundwater table level depth in the corresponding hillslope piezometer (F5b or F4) and b) HSS-F1b(Q) versus the initial groundwater table level depth in the piezometer in the riparian area (F1b). Hydrological years are labelled in italic.

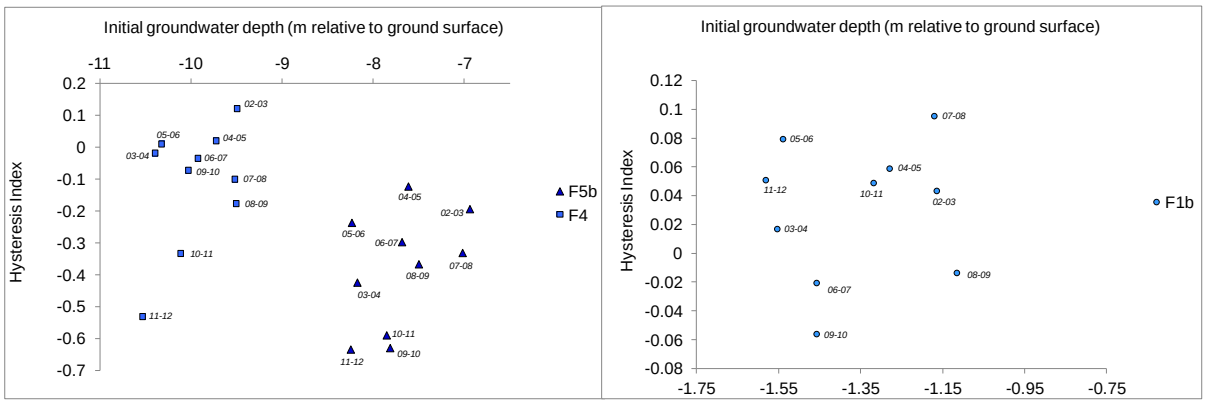
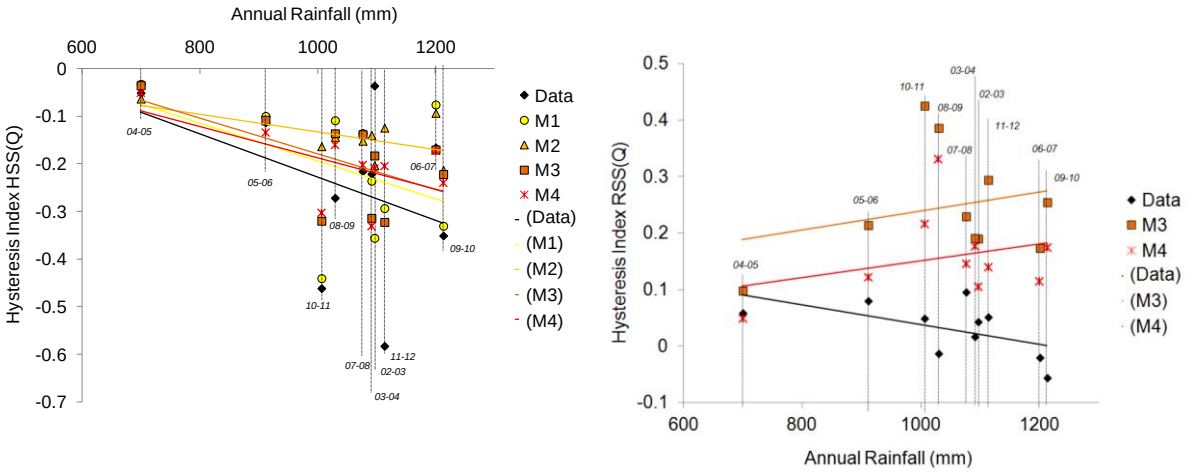


Figure 8. Variations of observed and simulated hysteresis Index versus annual rainfall for the 10 monitored water years for (a) Hillslope Saturated Storage versus discharge HSS(Q), (b) Riparian Saturated Storage vs. discharge RSS(Q). Solid lines indicate the linear regressions. Thin gray lines mark the hydrological years labelled in italic.



a b

Figure 9: Conceptual scheme of successive mechanisms which explain the annual hysteresis between storages and stream flows. HUS: Hillslope unsaturated storage, HSS: hillslope saturated storage, RUS: riparien unsaturated storage, RSS: riparian saturated storage, Q: stream flow, bold characters indicate varying components, grey arrows indicate if the component is increasing or decreasing, black arrows indicate the water flow paths.

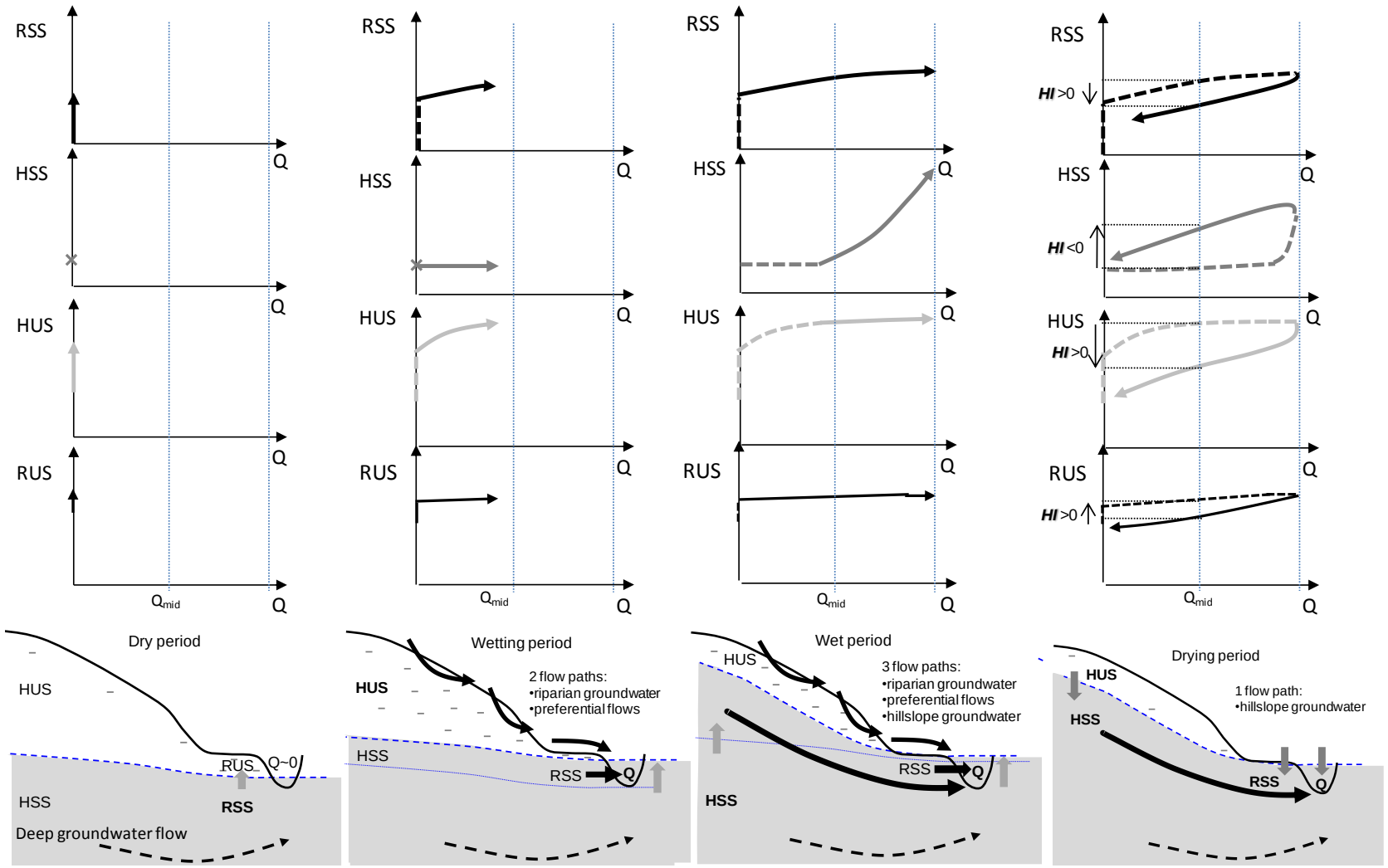


Figure 10. Observed and simulated annual hysteresis between stream flow (Q) and (a, b) Saturated Storage in the hillslope HSS (for observed, HSS is the average of F5b and F4) and (c, d) Saturated Storage in the riparian area RSS (for simulated, only M3 and M4 represent the riparian area), for the water years (a, c) 2003-2004 (wet year, calibration period) and (b, d) 2007-2008 (dry year, validation period).

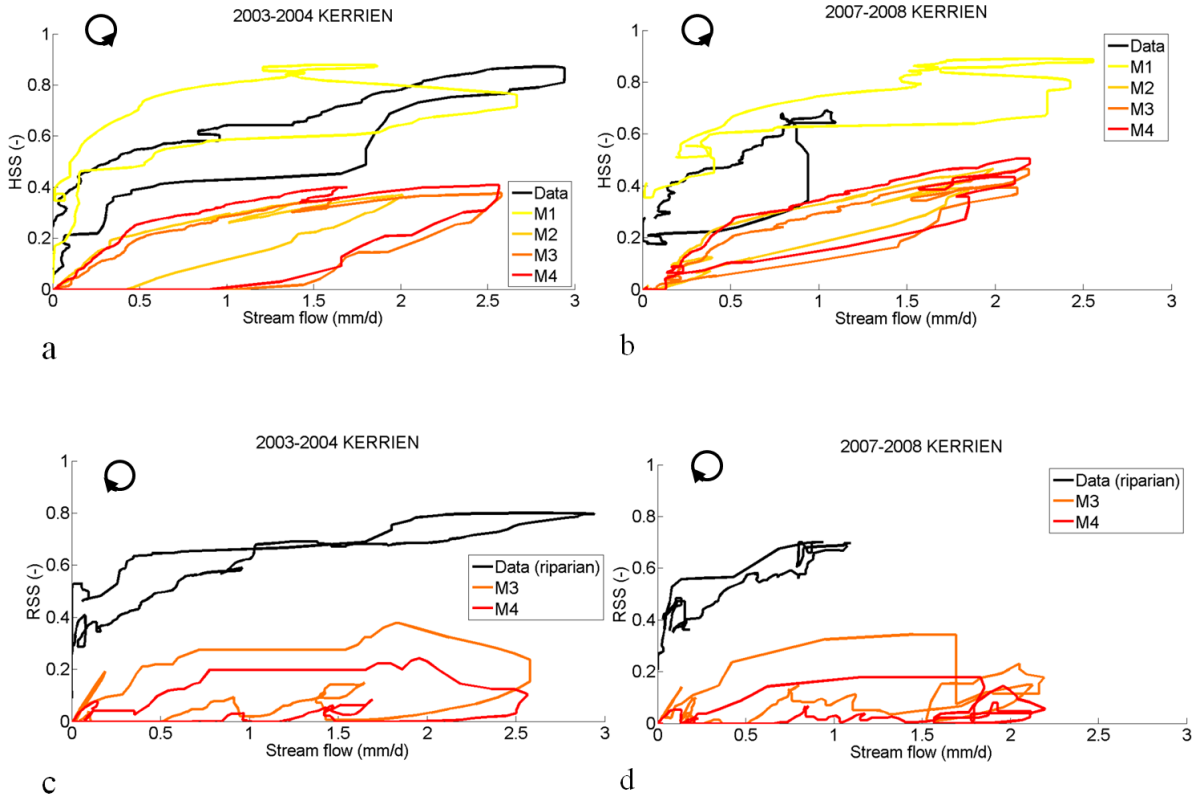
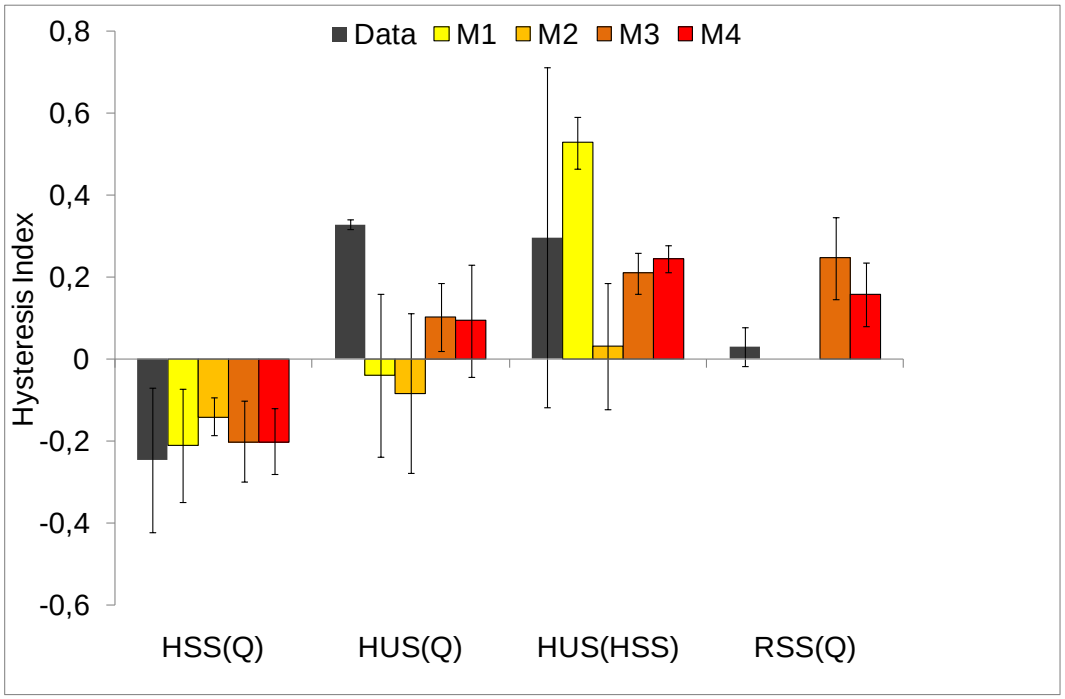


Figure 11. a. Mean annual hysteresis Indices observed and simulated with the 4 models M1 to M4, for Hillslope saturated storage vs. discharge HSS(Q), Hillslope unsaturated storage vs. discharge HUS(Q), Hillslope unsaturated storage vs. Hillslope saturated storage HUS(HSS), and Riparian saturated storage vs. discharge RSS(Q). RSS is simulated only in models M3 and M4. Error bars show the standard deviation for the 10 years for HSS(Q) and RSS(Q), and the values for the two available years for HUS(Q) and HUS(HSS).



b. Sensitivity of Hysteresis Index values to parameter uncertainty for the year 2011-2012. Mx bbp indicates the value for best behavioural parameter sets, the circles, triangles, squares ,and diamonds indicate the mean HI value for the all the behavioural parameter sets, and the corresponding bars its range of variation.