Answer to the comments by anonymous reviewer #1

I reviewed the original submission of the two companion papers, and raised a major concern of seemingly overlapping between these two. Now I believe that this concern has been adequately addressed therefore recommend acceptance as is.

We would like to thank you for your positive comments. Your (past) comments helped us significantly to improve this manuscript.

Answer to the comments by anonymous reviewer #2

The revised version of the manuscript is clearer. The main change the authors made was to extend the test period from 2 to 4 years. I found interesting the main conclusion of the manuscript that the more complex constrained but uncalibrated model structure can provide as good results as the calibrated lumped version.

That said, I felt quite disappointed by the way my comments (Reviewer #2) were accounted for by the authors. Several comments were judged "out of the scope" of the article although I still think they are not. At least a few lines of discussion could have been added on these issues (e.g. previous comments #6, 9, 15). Others (e.g. comments #3, 9) were said to be already discussed in the two recent papers published by the authors (Gao et al., 2013 and Hrachowitz et al., 2014) but this is still not mentioned nor discussed in the revised article (though I noticed these two papers are now cited in the companion paper). For a few other minor comments, the authors said they agree but I did not see the corresponding changes in the revised text (e.g. #1, 11). Last, it seems that some parts of the discussion of results in the revised manuscript were not updated following the introduction of the new results, with comments that are now not supported by results shown (see detailed comments).

We would like to thank you for your comments. We appreciate your disappointment and we regret our failure to take into account your previously mentioned comments properly. In this revision we followed your suggestions more closely, and in the process, the paper has gained clarity and we feel the paper has much improved as a result.
We understand your concern about the use of runoff data in our manuscript (your previous comments #11 and #6 regarding the use of runoff data as constraint on the runoff coefficient). in this study we had runoff data available and we used this information by introducing very loose constraints on yearly and the seasonal runoff coefficients. In the absence of availability of runoff data one can instead use other sources of information which provide an estimate of the runoff coefficient such as the Budyko curve or the runoff coefficient from adjacent catchments. We believe that the explanation given in section 3-3-3-2 clearly reflects this idea by saying "In case of absence of suitable runoff data, the mean annual runoff coefficient can be estimated from the regional Budyko curve using…"
About comments # 9 regarding models without constraints; we introduced the corresponding benchmark model. The benchmark models are simply un-constrained versions of calibrated or un-calibrated models. We compared every section where results are presented (4-1, 4-2, 4-3 and 4-4) with the benchmark models.
About comments # 15 regarding the reliability or narrower uncertainty interval; we agree with the reviewer that a narrower uncertainty interval does not mean more reliability. The reliability of the uncertainty interval illustrated in figures 3 and 4 are directly related to the reliability of the imposed constraints. One might argue some constraints are more reliable than others and therefore it is crucial to have an idea of how important each constraints is; this is what we are currently working on.
About comment # 3 regarding large sample hydrology; we agree with the reviewer that any suggested framework should be tested on a larger number of catchments. But the constraints for each individual catchment would possibly be different. To overcome this challenge an automated strategy could be developed to take specific input characters of each catchment into account, so as to generate specific constraints for each catchment. However at the current moment we are far away from being able to establish such an automated strategy. As you

mentioned we will elaborate other similar work that we did within our group in the discussion part (Gao et al., 2014 and Hrachowitz et al., 2014).
About comment # 1 regarding the definition of complexity; we regret that we didn't make the change we promised. This was missed between the corrections of the manuscript between the authors. To our knowledge there is no formal way of measuring complexity of a model, however, here by complexity we mean more parameter and more processes involved. We changed the first paragraph of the manuscript accordingly.

Overall, my main concern expressed about the robustness of the conclusions still remains. Although the proposed approach seems sensible, the authors could have done a bit more to make their article not "just a simple test" as mentioned in their reply.

Since the authors do not seem willing to deeply modify their manuscript, I just advise the few minor revisions below.

Detailed comments
1. Page 1, Line 48: "are provided"

We believe this should remain the same as models are the subject of the sentence.

2. Page 1, Line 48: "Grayson"

The text is changed accordingly.

3. Page 3, Line 34: "and thereby limit predictive uncertainty" (?)

The text is changed accordingly.

4. Page 4, Line 28: "Drogue"

The text has been changed.

5. Page 7, Line 8: "reservoirs" (?)

The manuscript is modified.

6. Page 15, Line 36: "constraints"

The text is changed.

7. Page 17, Lines 22-27: Although this comment was true in the first version of the manuscript, this is unclear now with the new results. The performance of the FLEXA model only degrades on the ENS criterion. Performance even improves for ENS,log and ENS,FDC between calibration and validation for the three models. Of course, this may happen when the validation period is easier to model than the calibration one. But then the notion of "validation/calibration performance ratio" becomes difficult to interpret. The interpretation would have been easier if the authors had chosen to apply the full split sample test as already suggested in my previous review comments, since it would ease the comparison of performance between calibration and validation (even though I agree with the authors reply that in the case of uncalibrated models, calibration and

validation performance would be the same). Note that the slight change the authors made in their test period changed the way the results should be interpreted, which reinforce my previous feeling that conclusions given in the article could be more robust with enhanced testing scheme.

We agree with the reviewer that the result and conclusion of this paper is case-specific and is only valid for the study catchment (the Wark). We tried to reframe the conclusion in a way that the general conclusions are emphasized rather than specific conclusions. Moreover we carried out a split sample test as the reviewer mentioned. The result of split sample test is presented in the supplementary material.

8. Page 17, Lines 28-33: Again, this statement is less clear now with the new results (FLEXB is better than FLEXC for two criteria) and this conclusion should be revised.
9. Page 19, Lines 8-9: According to results shown in Table 4, this statement is false since the median calibration performance is not better for FLEXA than FLEXB and the uncertainty is the largest for FLEXA.
10. Page 19, Lines 13-15: This statement is not supported by the results shown since FLEXB appears better performing than FLEXC.

We regret that the manuscript did not correspond to the figures and tables. We carefully changed the manuscript to avoid any confusion as such.

11. Figure 5: The caption could indicate which distribution percentiles are shown by the box-plot.

We mentioned the explanation of box plot, percentiles, Whiskers and outliers in the caption of the figure 5.

12. Page 20, line 17: "reaction tends to"

Text has been modified accordingly.

13. Figure 6: Should this figure better show the results in validation, which are more representative of actual model performance than calibration results?

We created figure 6 for validation period. However the general conclusion and the manuscript did not change.

Answer to the comments by anonymous reviewer #3

This article is very interesting and well-written. The authors provide a systematic and potentially useful way to incorporate the expert or soft knowledge in parameter identification. I advise publication after minor revision.

We would like to thank you for your constructive comments.

Specific comments

1.Section 4-1, the model FLEXC having the best performance with the un-calibrated constrained parameters does not necessarily prove that "the imposed relational constraints force the model and its parameters towards a more realistic behavior". The author should do the comparison among these different models with calibrated parameters (using conventional calibration methods, avoiding the effects of these proposed constraints). In my opinion, it is the additional landscape or HRU that improves the simulation performance, far more than the additional constraints.

Based on what the reviewers asked, we implemented further tests. This test was to compare the constrained to un-constrained parameter sets. So for each "constrained but uncalibrated" and "constrained and calibrated" test, we made benchmarks without imposing any parameter and process constraints (the benchmarks are "unconstrained but uncalibrated" "unconstrained and calibrated"). This way we can test how important the model structure is without any constraints imposed on the model structure.

2.Section 4-3, the second paragraph, "the expectation that increasingly complex models will have increasingly poor validation/calibration performance ratios", many similar expectation results have been mentioned in this article, so it is necessary to provide at least one figure to show these expectation results are reasonable. For example, compare the validation results and uncertainty between these three models with calibrated parameters using a conventional method.

We think by conventional method, the reviewer means calibrating a model without any constraints. If so, and as we explained in response to the earlier comment, we introduced benchmark models without any constraints. We elaborated the result and discuss it as a comparison in the manuscript.

3.Section 4-3, the third paragraph, in Figure 5 (Ens), the difference between a calibrated lumped model FLEXA and a more complex constrained but uncalibrated model FLEXC in validation period is small, but when we compare table3 and table 4, the difference between the results of Ens(log) is large (0.75 vs 0.63), and the difference of uncertainty is also large. So the following implication is not reasonable. Please add more analysis and discussion based on the other two metrics.

We thank the reviewer for pointing this out. The finding is based on E_{NS}, indeed it will be different using different measure of performance. We clarify this issue in the manuscript and we also toned down our previously stated conclusion.

4.Although this article focuses on the constraint-based method, I want to know how big is the gap between the constraint-based method and the conventional calibration method (a benchmark), and thus can estimate the potential capacity of the proposed method.

By introducing the benchmark models without any constraints the effect of the model structural differences and added constraints can be evaluated. Comparing the constrained but uncalibrated parameter sets with unconstrained and uncalibrated gives us a clue about how importance the constraints are. The other comparison can be between un-calibrated and un-constrained (benchmark) parameter sets for difference models. This comparison presents us information on the importance of different models structures and inclusion of landscape units into the modelling practice.