

Response to Anonymous Referee #4

Received and published: 19 July 2016

Anonymous Referee (AR): I have to admit that I struggled for a couple of days to understand the message of this paper, and I am disappointed to say I failed to do so. My understanding is this paper is an amalgam of sensitivity analysis, parameter estimation and catchment clustering, however it is not well described how these approaches are linked together. Each of these elements, if performed elaborately, can be a separate paper and mixing them only confuses readers. Neither the abstract nor the introduction sections support the goals of the study. Well to be fair, goals are not clear either! Also I should mention that manuscript is not fluent at times. In my comments, I only focus on major flaws of the manuscript and skip my minor comments:

Simon Höllering (SH): We thank the referee for his critical assessment of our manuscript.

AR: I noticed there are several unsupported claims in the manuscript, one of them is "parameter estimation". I can't find how the parameter estimation is performed. It is not close to sufficient to consider the 91 parameter combinations from the sensitivity analysis and use a selection criteria based on some fingerprints of the catchment to select parameters as behavioral from this limited set. This would lose many behavioral parameter combinations, and doesn't provide any information about the posterior distribution.

SH: Yes, as stated in the reply to referee 1, we are aware that a number of 91 runs is certainly too small to find parameter sets which work well for most of the fingerprints. Nevertheless, we found several to work fine for specific fingerprints and some that are not far from the results (in terms of the NSE) we obtained by automatic DDS calibration (0.75 vs. 0.81). In accordance with a statement by referee Björn Guse we will mention in the revised manuscript that this lower number of model runs is acceptable here to show the goals of this study, which is different from automated model calibration.

AR: I might be wrong, but my understanding is that the sensitivity analysis of model parameters depends on a residual based objective function. At least it is highly dependent on simulation of the system response at individual time steps! This is in contrast with the purpose of using fingerprints of catchments, which are originally defined to constrain model parameters to represent an aggregate behavior of the catchment.

SH: Sensitivity analysis using FAST rests upon the idea that model simulation output can be transferred into a Fourier series. Within feasible, predefined ranges, parameters are varied with specific, independent frequencies. The Fourier coefficients finally allow an estimation of the partial variance, hence model parameter sensitivity, used to identify dominant parameters. In this process of sensitivity analysis an ensemble of model runs (based on different parameter combinations) is required (91 in our case of 6 parameters) which we also used for the assessment of model performance and consistency check in terms of the reproduction of response fingerprints. Given the implementation of FAST we do not need an objective function here (such as NSE or MSE). Similarly, as responded to referee 3 and what we will clarify in the revised manuscript, a structurally adequate model with behavioral parameters sets should consistently allow for acceptable stream flow simulations and reproduce fingerprints derived from stream flow. An inconsistent behavior in the sense that parameters which work well during stream flow simulations, but do not work well with respect to reproduce fingerprints provides evidence for model structural error.

AR: It is not clear how the selection criteria is adopted to delineate the behavioral parameter distribution. There are times that authors discuss one fingerprint is used, whereas in other instances they used a couple of fingerprints jointly!

SH: **Sorry for being imprecise, we will clarify the selection of behavioural parameter sets in the revised version of the manuscript and better discuss the use and value of both one and several fingerprints as constraining measures in separate sections.**

AR: In the original application of fingerprints that authors referred to (Vrugt and Sadegh, 2013), 4 fingerprints were used that are necessary to meet the acceptance criteria jointly. It is not clear if authors have performed their analysis on single sites (headwaters), or they have modeled the entire system altogether.

SH: **The simulation was run and sensitivity analysis was performed on each single headwater and, as we showed in several figures, fingerprints were separately derived. Five most behavioural sets were in the end chosen jointly from seven headwaters (where the overall model performance was acceptable) and further assessed.**

AR: Page 2, line 32: My experience shows that, at least for US catchments, parameters of certain models are more correlated with climatic variables rather than soil characteristics. It is worth mentioning here, although the sentence is correct in how it describes the findings.

SH: **We agree with this statement. This section might be removed, thus catchment classification will not be relevant for the revised manuscript.**

AR: In section 3.3.2, authors talk about the study area before introducing it!

SH: **Yes, we will correct this point.**

AR: Page 8, Line 3: I don't understand the sentence: "relate physiographic and climatic characteristics to sensitivity-confined hydrodynamic response fingerprints"

SH: **We apologize for the unprecise formulation. In fact we try to relate the connection of regional differences in hydrologically relevant landscape properties and climatic factors (which have shown here to be heterogeneous even in nearby locations) and the hydrologic response expressed by fingerprints which can be further related to model parameters which might be sensitive to a fingerprint. This insight can be e.g. used to diagnose model structural deficits.**

AR: Page 10, line 7: Water is withdrawn from the system, it is not lost!

SH: **Yes, fixed this text passage.**

AR: Section 5.3: authors talk about consistency of behavioral parameter sets. In what sense have you analyzed the consistency of parameter sets? For definition of hydrologic consistency refer to: Martinez, G. F., and H. V. Gupta (2011), Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States, *Water Resources Research*, 47 (12).

SH: **Martinez, G. F., and H. V. Gupta (2011) point out that 'the term consistency can mean a great many different things, including what might be called (1) internal consistency with regard to the conceptual representation of a model (...) and (2) external consistency in terms of the ability to reproduce hydrological behaviors and properties seen in the data.' In our case, the evaluation of behavioural parameter sets was done with respect to their consistency in reproducing one single or two combined fingerprints of hydrodynamic response. This might be most similar to what is meant by external sensitivity.**

AR: Page 15, line 25: In the entire manuscript authors are talking about 6 model parameters, and all of a sudden they switch to 52 global mHM parameters! It confuses me which one is the correct number of model parameters.

SH: We changed this part to better highlight that we used 6 parameters out of the whole set of 52.

AR: Page 16, line 26: Authors suddenly talk about temporal sensitivity of parameters! This is completely different from what reader expect from a joint sensitivity-parameter estimation analysis. The latter works with the entire data set, whereas the former is concerned about individual time steps!

SH: In accordance to the response to referee comment 3 we will clarify objectives e.g. in the way that we extend the TEDPAS (temporal dependence of parameter sensitivity) concept to fingerprints (FDCs, HFD, etc.) to depict changes in their parameter sensitivity for changes of the independent variable INDPAS and optionally assess regional differences.

AR: Page 19 lines 6-16: Categorizing catchments based on model parameter assumes that the model is sufficiently describing the system. This assumption is not well justified nor supported by the results.

SH: We agree with the reviewer that any model based categorization of catchments is based on the assumption that the model is structurally adequate. Here we do not categorize the catchments on parameters, but we compared differences TEDPAS and the partial sensitivities among catchments. Differences in the sensitivities underpin the differences in parameter identifiability in these catchments. As parameter ranges of the FAST are the same within all these catchments. These differences may be attributed to differences in the hydro-climatic input. We will better explain this in the revised manuscript.

AR: Page 20, lines 21-22: This is again unjustified claim to say this paper does: “(1) investigate hydrologically relevant structural and functional attributes in terms of consistency and feasibility in classifying similar catchments, (2) assess the value of functional constraints for the parameter spaces of distributed hydrologic models”. I am not convinced that the results of this study support these claims.

SH: We will remove these passages from the revised manuscript, as catchment classification is not within the scope of this study.

AR: Figure 3 is not well explained in the text.

SH: Yes, we think about removing this figure in the revised version.

AR: Figure 10 & 11: How is it possible to differentiate between model simulations of the two gauges?

SH: The simulations here are only shown for gauge Wenholthausen (WEN) but observations for both (with the reference to parameters of highest sensitivity). To be clearer here we will add the names of gauges to the curves.

AR: Figure 12: Why four of the parameters and not all 6?

SH: Good, we also think about showing SDCs of all the six parameters.