Response to F. Sarrazin (Referee)

Received and published: 18 July 2016

Fanny Sarrazin (FS): In this manuscript, the authors present a parameter estimation and sensitivity analysis scheme to learn about differences in catchment functioning and to classify catchments. However, many points need clarification, in particular the implications and conclusions of the work. I recommend major revisions of the manuscript. Parts of the manuscript appear to be quite long (specifically the introduction and results sections) and the reader tends to get lost. A general recommendation to the authors is to select in each section the key elements that contribute to their argumentation and to clearly connect them. The authors should also try to keep their sentences short to improve readability. Furthermore, critical information on the implementation of sensitivity analysis is missing (output considered, how the sample size was chosen), which makes it difficult to interpret the sensitivity analysis results. I provide below more detailed comments for the different sections and some minor comments at the end.

Simon Höllering (SH): We sincerely thank Fanny Sarrazin for the time and efforts she invested into the review of our manuscript. We agree that the presentation of our study can be considerably improved and will do this in the revised manuscript as specified in the forthcoming reply to the reviewer comments.

SECTION 1

Introduction

FS: 1) I think the introduction section is unnecessarily long and not sufficiently linked to the objective section (section 2). Specifically, how do the work of Bardóssy (2006) (p2 L11) Wagener et al. (2007) (p2 L18), Castiglioni et al. (2010) (p4 L14), Yadav et al. (2007) (p4 L18) relate to the work presented in the manuscript? In which way the literature review on catchment classification (section 1.2) is useful to better understand the work presented in the manuscript? I suggest the authors select key elements in the literature review and clearly link them to their study, so to clarify what the contributions of the manuscript are.

SH: We will streamline the introduction and focus on studies dealing with regional parameter sensitivity, FAST and their optional combination. As also recommended by the first reviewer we will better focus on fingerprints used as constraints for the parameter space and model diagnostics.

FS 2): There is some confusion in Section 1.3 and it needs clarification. A distinction has to be made here between a model and the underlying system it represents. The term equifinality commonly refers to the model and not the system it represents. In fact, Beven (2006) (p21, paper cited by the authors) characterizes the term equifinality as the 'rejection of the assumption that a single correct representation of the system can be found given the normal limitations of characterisation data.' Therefore, equifinality is not an 'inherent' property of a model equations as stated by the author p3 L31, but it is due to the fact that the information content in the data is not sufficient to identify a unique model representation. I am also asking the question: Is the study of physical and hydrological similarities based on observation data or on model simulation outcomes?

SH: We thank for this comment, which shows the need for clarification. We agree that the system is not equifinal, hence the average residence time of water in a catchment is well defined (might be state dependent). What we meant with equifinality is inherent to our equations is equifinality due to parameter interactions. For instance n and k determine the average "travel time" through a Nash Cascade. Bardossy (2007) showed nicely that an infinite number of parameter yield acceptable results – a doubled n in combination with a 50% reduce k yields the same results. The same holds true for

Darcies law, as system with a twice as large gradient and 50% reduced permeability produces the same flux. We will better explain this in the revised manuscript

In the revised manuscript we will also explain that we use observed streamflow signatures, which have been shown to vary among our target catchments, for regional model evaluation. On one hand we test whether parameter sets which reproduce one or multiple fingerprints, yield acceptable stream flow predictions (or the other way around). A structurally adequate model should be consistent with respect to both. We furthermore use FAST to estimate the sensitivity of the fingerprint of the parameter along their independent variable (for instance for different exceedance probabilities when using the flow duration curve. Differences among different catchments might be a hint for different interplay of runoff generation processes. This information might be furthermore useful for calibrating models directly based on fingerprints.

FS: 3) p4, L20: the expression 'to raise the information considered in the parameter transfer' is quite fuzzy. Please clarify.

SH: We apologize for being unprecise. We will find a more comprehensible formulation for the reference to these studies in the revised version of this section.

SECTION 2 Objective and driving research questions

FS: 1) As mentioned above, the objectives need to be clearly linked to the literature review, Interactive to show the contributions of the manuscript.

SH: Yes, we will link the review more precisely to what we want to achieve with our study. Clearer objectives could be:

- **Explore regional differences in TEDPAS (Temporal Dependence of PArameter Sensitivity) and discuss their implications for model diagnostics.**
- **Extend the TEDPAS concept to fingerprints to explore and depict changes in their parameter sensitivity and along independent variables (e.g. exceedance probability for FDCs).**
- **Employ constraints by fingerprints on the feasible parameter space of a distributed hydrologic model and assess hydrological consistency in terms of model performance and model structural aspects.**

FS: 2) p4 L30-31: This objective needs clarification. First, 'consistent manner' is vague and it is required to better explain this expression. Second, the motivation needs to be clarified. Is the objective to learn about the model ability to reproduce the data, to learn about the value of observation data etc.?

SH: Sorry for being not precise here. We think that 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. In this sense we tested here parameter sets which performed well with respect to one or several stream flow indices with respect to their performance during a stream flow simulation and vice versa.

FS: 3) p5 L8-9: the expression ‘within the range of their independent variable’ needs clarification.

SH: Sorry for being imprecise. We use the Fourier Amplitude Sensitivity Test to assess temporal dependence of parameter sensitivity. We hence vary 6 parameters with preselected independent frequencies within 91 model runs. The idea of FAST is then to perform a Fourier analysis of the 91 model runs for each time step. The amplitudes in the power spectrum provide information which of the parameter contributes most, second most and so on to the spread of the model runs at a given time step. Temporal changes in their weights indicate changes in dominant sensitivities. We do the same for stream flow signatures such, in case they are in the form of a bivariate plot such as the flow duration curve or a double mass curve of a catchment, which is accumulated discharge plotted against accumulated precipitation. The fast ensemble yields 91 flow duration curves or 91 double mass curves. Again we can perform a Fourier analysis of the 91 model runs for a given value of the abscissa (the independent variable). The amplitudes in the power spectrum provide information which of the parameter contributes most to the spread of the model runs. Changes in the weights of the different parameters indicate changes in dominant parameter sensitivities of the fingerprint either for different exceedance properties or for different accumulated rainfall inputs during a hydrological year.

SECTION 3 Concept and methodological approach

FS: 1) I have reservations regarding the implementation of the sensitivity analysis (section

3.3), specifically:

- p7 L12: How was the sample size chosen? It is necessary to check the convergence of the results, that is to say to assess to what extent the results would change if using a new sample. When the sample size is too small, the sensitivity analysis results can be unreliable (e.g. Sarrazin et al., 2016). A way to assess convergence is to derive confidence intervals on the sensitivity indices using bootstrapping. However, this technique cannot be easily applied when using FAST, given the structure of the sample. Therefore, I would suggest to simply look at the stability of the results when using smaller/larger sample sizes.

SH: As far as we understand the method, and as explained in Reusser et al. 2011, the frequencies and the minimum number of model runs are pre-defined by the theory of the FAST Fourier transform method. The sampling size depends on the number of parameters and independent frequencies. We hence do not expect different results when increasing the number of model runs. Results might however definitely change when the sensitivity of the model is heteroskedastic, for instance in the presence of threshold processes, for instance Hortonian overland flow. When choosing a specific range for the related parameter one should assure that are such that the process gets activated. The ranges of the parameters and the parameters themselves were chosen based on a previous local sensitivity analysis.

FS: - It is required to specify the scalar output used for sensitivity analysis, otherwise the results are meaningless. Sensitivity analysis results can strongly vary when considering different model outputs (e.g. van Werkhoven et al., 2008).

SH: We agree that parameter sensitivities depend on the model output/ simulated process. Naturally one expects evaporation simulations to be sensitive to different parameters as streamflow simulations, particularly as these processes depend also in nature on different catchment characteristics. We will present a more accurate characterization of the 91 streamflow time series (as scalar model output) and input for FAST.

FS: - Why analysing main effects (FAST) only and not total effects (eFAST)? A parameter could have an effect through interactions only and therefore would not be detected when applying FAST.

SH: We agree with your statement and are aware of the shortcomings of FAST that solely covers the main effects of parameter changes on model output. Nevertheless, we considered the computationally efficient FAST method with a small number of factors (parameters) and interaction terms, thus smaller 'likelihood of nonnegligible higher-order terms' (Saltelli et al., 1999), as sufficient to illustrate the goals of the study. This reasoning is additionally supported following the main findings of Reusser et al. (2011), where the comparison of 3 different methods for Sensitivity analysis (including FAST and eFAST) yielded similar results in terms of sensitivities to model output. However we will better explain our choice in the revised manuscript.

2) In Section 3.4.1 several points need clarification:

- p8 L3: 'sensitivity confined' is unclear.

SH: Yes, sensitivity confined should somehow express that we relate or connect parameter sensitivity to the dynamic response of catchments either expressed by fingerprints, FDCs or hydrographs. We will clarify that statement by using a different expression.

- p8 L4: this sentence is vague

SH: We will clarify that one, too, or remove it from the manuscript.

- p8 L5: this sentence have to be revised in relation to section 1.3.

SH: We will better explain this part in connection to section 1.3.

3) p9 L10-25: This is quite a long paragraph and I am not sure in which way it connects to the work presented in the manuscript. Again, I suggest to select the relevant information that helps to better understand the present study.

SH: In the revision we will focus more precisely on regional parameter sensitivity linked to signatures (fingerprints) such as FDCs and their value to constrain the parameter space defined by FAST and their consistency in model performance and as control possibility of model adequacy. We will better illustrate the connections.

SECTION 4 Study area

1) The authors should present the data used for the simulations and precise the simulation time horizon chosen for the analyses. Also, why presenting some of the data used before presenting the study area? (section 3.3.2, p7 L15-21)

SH: The presentation of the data is indeed a point which has to be reworked and better structured.

SECTION 5 Results

1) A general comments for this section is that some figures are redundant and could be removed for the sake of brevity (e.g. Figure 6, 9 and 10). Likewise, the text could be more concise.

SH: Yes, as also stated in the answer to the 1st referee, we will reduce the number of figures to the minimum necessary amount and better specify captions.

2) p11 L13: Why is it a ‘qualitative constraint’?

SH: Qualitative might be better removed here or at least replaced.

3) p11 L29-30: The behavioural subset is the lower branch of the ‘arrow’ or the points that are within the circles?

SH: We changed ‘behavioural’ to ‘...model runs with acceptable performance.’ to avoid misunderstanding, since we only assign model runs that are within the circles to the group of behavioural parameter sets.

4) p12 L2-3: What are the implications of the statements ‘which portends [.. .] specifically.’ and p12 L10-11 ‘Circles [. . .] varying strength’?

SH: We will be more precise here and discuss the implications. The finding here is that local differences emerge between different catchments (though closely located) if the same constraint is used for the same parameter space in the vicinity of locations where behavioural parameter sets can be well selected while nearby places hardly produce any behavioural model run for the selected fingerprints.

5) How does section 5.3 relates to section 5.1 which is also about constraint on parameter space?

SH: Sorry for being unprecise here. Section 5.3 is the assessment of the performance of the behavioural parameter sets selected by constraints from observed fingerprints. We will better connect these two sections which build upon each other.

6) In section 5.3, I think it may be more appropriate to identify for each fingerprint and pair of fingerprints not a unique most behavioural parameter set but an ensemble of best performing parameter sets. In fact, observation data are affected by uncertainties and it may not be relevant to make distinctions among the top performing parameter sets.

SH: This is also a good point that we will further discuss and try to consider for the revision.

7) Why keeping correlated fingerprints in the analysis of section 5.3? For instance, CV or HPC could be removed from the analysis since the two fingerprints have the same information content.

SH: Yes, nevertheless this example might show how important it is to choose appropriate, non-redundant signatures. We will find a solution here.

8) p14 L12: Please clarify how the five most behavioural parameter sets were determined.

SH: We selected these 5 most behavioural parameter sets, by counting the number of times they occurred in the top five in reproducing the four combinations of fingerprints shown in fig. 5 and fig. 7c and d (referred to the seven shown headwaters). We will further clarify this selection.

9) p15 L24-26: This statement is not correct. When referring to Sobol’ variance decomposition (Sobol’, 1990), the sum of main effects is equal to 1 when main effects only contribute to the total variance (no interactions). A sum of main effects for the six parameters analysed equal to 0.43 means that a significant fraction of the total variance (1-0.43=0.57) is due to interactions between parameters. If more parameters were added to the analysis, this would not necessarily result in an increase in the sum of the main effects, since the output variance would also change. The same applies for the statement p21 L11-12.

SH: Thank you for this statement. We will check these parts and correct it.

10) Section 5.4.1: I have reservations regarding the interpretation of the sensitivity analysis results: - p16 L4-10: I do not see any clear summer/winter pattern for the sensitivity of Aspect-corrPET and Recharge Coeff for WEN but more 'ups and downs' all year round.

SH: Yes, the summer/winter pattern of AspectcorrPET and RechargeCoeff could be more pronounced but we can see a shift of the first rank sensitivity in the transition winter-summer from RechargeCoeff to AspectcorrPET and its generally higher average level of sensitivity in summer whereas RechargeCoeff doesn't change the average level of sensitivity significantly. We will try to clarify the interpretation.

- p16 L26-35 - p17 L1-11: I am not sure about the significance of Figure 11. The authors consider the most influential parameter only. However, the difference in sensitivity among the most sensitive parameters may be small and even not statistically significant given the approximation error in the sensitivity index values. This is all the more concerning, since all the sensitivity indices take quite small values from Figure 9 (below 0.12).

SH: We will find a way to present the relation of sensitivity to the independent variables (e.g. time or exceedance probability) in a more comprehensive way.

11) p17 L10-11: The sentence 'Nevertheless [. . .] on stream flow generation.' is very fuzzy. It is required to clarify. I am also still wondering why the authors chose to study the main effects only and not the total effects.

SH: We will clarify this sentence. Please also refer to our answer on eFAST above.

SECTION 6 Discussions and Conclusions

1) I suggest splitting this section in two, with a discussion section and a concise conclusion section. This would help to highlight the contributions and implications of the work.

SH: This is a good idea. Referee Björn Guse also recommended to split this section into two sub-chapters. We will incorporate this in the revised version of our manuscript.

2) p18 L33: Either an indicator is normalised or it is not normalised but it cannot be 'less normalised'.

SH: Correct. We changed it to '…more tangible and less biased (Krause et al., 2005) as it is e.g. in the case of NSE.'

3) p19 L10-14 What are the implications of the statements 'The composition [. . .] in the eastern headwaters.'?

SH: We agree to further clarify this point in terms of the implication for the performance analysis and parameter identification.

4) p19 L18-19 'by model [. . .] of meteorological forcing data' is quite fuzzy. Please clarify.

SH: Yes, meteorological forcing especially precipitation might have an effect on local differences in temporal parameter sensitivities. We will discuss this in more detail.

5) p20 L22: The authors do not actually present any result on catchment classification

SH: We apologize for setting the scope of the paper too wide. We will set a better focus on our main objectives.

MINOR COMMENTS

- title of section 1.3: replace 'inconconsistency' by 'inconsistency'

SH: We fixed this mistake.

- there is an error in the reference to Beven (1990). Are the authors referring to Beven, K: Changing ideas in hydrology - The case of physically based models, Journal of Hydrology, 105, 157-172, doi: 10.1016/0022-1694(89)90101-7, 1989.

Or

Loague, K.: Changing ideas in hydrology - The case of physically based models - Journal of Hydrology, 120, 405–407, doi:10.1016/0022-1694(90)90161-P, http://linkinghub.elsevier.com/retrieve/pii/002216949090161P, 1990.

SH: Fixed. The first reference is the one we intended to refer to.

-p5 L8: replace 'of the selected fingerprints' by 'to the selected fingerprints'

SH: Yes, fixed.

-p5 L12 and L23: what do the authors mean by 'dependent'?

SH: Dependent in the sense of the dependent variable streamflow in the case of FDCs and hydrographs. We will better explain these expressions.

- p5 L13: replace 'of simulated stream flow and of the related fingerprints' by 'to simulated stream flow and to the related fingerprints'.

SH: Yes, fixed.

- p6 L21-23: Why introducing eFAST here if it is not used? I think the sentence can be removed to keep the manuscript concise, unless eFAST is actually used.

SH: We agree with your recommendation

-p8 L19: aren't there five dynamic fingerprints?

SH: We incorporated the baseflow index as intermediate fingerprint in the dynamic ones to characterize low flow conditions.

-p10 L14-15: it is required to reformulate this sentence.

SH: We will reformulate this to make this statement clearer.

- p14 L31: replace 'to small' by 'too small'

SH: Yes, fixed.

- p16 L25: I don't think the term 'interacting inversely' is correct since only the main effects (and not parameter interactions) are analysed here. Please reformulate.

SH: We will reformulate this, given that main effects are analysed here.

REFERENCES

Sarrazin, F., F. Pianosi, and T. Wagener. 2016. "Global Sensitivity Analysis of Environmental Models: Convergence and Validation." Environmental Modelling & Software 79: 135–52. doi:10.1016/j.envsoft.2016.02.005.

Sobol', I.M. 1990. "Sensitivity Estimates for Nonlinear Mathematical Models." Matematicheskoe Modelirovanie 2, 112-118 (in Russian), Translated in English (1993). In: Mathematical Modelling and Computational Experiments 1: 407–14.

Van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang. 2008. "Characterization of Watershed Model Behavior across a Hydroclimatic Gradient." Water Resources Research 44 (1): 1–16. doi:10.1029/2007WR006271.

SH: Thank you for this reference list. We will try to incorporate citations to these references in our revised manuscript.