Dear Dr. Ehret,
Thank you for your thoughtful critique of our manuscript. We have addressed each of your comments below (in red text). In particular, your comments have helped to clarify the intent of our manuscript, as well as ensuring that this intent is clearly stated in justification of our approach. These clarifications were focused on the introductory and discussion portions of the text.

Christa Kelleher and Colleagues

Dear Editor, dear Authors,

I have reviewed the aforementioned work (version 2 of the manuscript). My conclusions and comments are as follows:

1. Scope

The article is within the scope of HESS.

2. Summary

The authors present and apply a framework for calibration of distributed hydrological models by applying a hierarchical set of parameter constraints and error metrics to accept or reject randomly generated parameter sets. The constraints and error metrics are distinguished by i) range of applicability (from regional to local), ii) 'softness' (from local observations to heuristically formulated local expert knowledge) and iii) the evaluated characteristic (evaluation of non-dynamical to dynamical aspects).

The framework is presented at the example of the physically based, distributed DHSVM model applied to the 5.5 km² Stringer Creek catchment, whose hydrological behavior is dominated by seasonal snow accumulation and snow melt.

Using 10.000 randomly drawn model parameter sets, 10 signatures and error metrics are applied individually, in groups and in hierarchical combination to identify behavioral parameter sets from the initial set. The resulting subsets are then discussed with respect to the equifinality reduced (i.e. by how much the initial parameter ranges were narrowed).

The authors show that by jointly applying all criteria considerably narrows the behavioral parameter sets (here: to nine). However, these still show large differences, specifically with respect to catchment groundwater table (values and spatial patterns). From the analysis, the authors conclude that i) a multi-criteria approach to identification of behavioral parameter sets is superior to single-criteria approaches, ii) dynamic constraints to be more effective than non-dynamical ones, and iii) that despite substantial narrowing of the parameter space still large differences among the surviving parameter sets remain, especially with respect to spatial patterns of hydrologic states.

3. Overall ranking

The work is ranked 'Major revision'.

4. Evaluation
Major points
I like the work presented in this paper, especially the strong argument towards using multiple, hard and soft sources of information to identify behavioral model parameter sets, and the thorough literature review. However, despite the fact that the authors' focus in this paper is to present the concept, with the choice of the catchment and time series used, they have clearly missed some very good opportunities to make their results more general and interpretable. More specific:

Judging from the presented time series of discharge and snow water equivalent, the catchments' hydrological function is very simple (one major discharge event during snowmelt, rainfall-runoff events are hardly playing a role). Arguably, a very simple conceptual hydrological model could reproduce this behavior at least as well as the applied model with respect to all discharge-related signatures and metrics, but with much less parameters.

- So why choose this very simple-behaving catchment if the goal is demonstrate the usefulness of a targeted constraining approach for a distributed, physically based model? This way, the model stays well below its potential, and this also means a lot of opportunities for more targeted model parameter evaluation are missed.

We selected this catchment for a number of reasons, but especially because it is a place where modelers and experimentalists have collaborated before, where other types of models have been applied. Most importantly, we selected it because it is a place where we have lots of understanding. We sought to incorporate distributed measurements that are more likely to be available in other catchments (e.g., SWE) to frame our application in a general way that could be applied to other sites. Furthermore, as this is a relatively simple system, it provides a first-order test as to whether this type of approach could have merit before we introduce this approach to a more complex set of catchments.

An additional benefit to this system is it has been modeled previously with conceptual models, as highlighted in the discussion section (page 16, section 5.2). We compare our findings to other simple conceptual models and distributed models of this system (across similar periods), and show that we achieve similar levels of fit. To your point above, we have added a discussion of the choice of model complexity, framed by applications within this specific catchment (page 17):

“The three models to which we compare our results demonstrate a range of model frameworks that can be used to evaluate model behaviour: conceptual (Smith et al., 2013), lumped (Ahl et al., 2008), and distributed without physically-based parameters (Nippgen et al., 2015). As is shown in this study, all of these models are able to accurately simulate the hydrograph for this catchment. The primary trade-offs across these models include requirements for inputs and parameters alongside computational requirements, which are inversely related to the complexity of simulated behaviour that can be produced from each of these models. While any of these approaches may be used to simulate streamflow, each will enable researchers to answer different questions related to hypotheses about catchment functioning, the use of field information to inform model parameter constraints, and predictions of spatio-temporal hydrologic processes. Finally, these contrasting models also illustrate the differences between a model like DHSVM that may be applied to many different catchments versus the models introduced by Smith et al. (2013) and Nippgen et al. (2015), in which the model framework and structural equations were developed only for this catchment. In this study, we specifically evaluate the application of physically-based, distributed models to simulate experimental catchments, though we encourage researchers to select the right tool, and therefore the appropriate model, for a given study objective."

- Furthermore, the many degrees of freedom in your model inevitably lead to problems of equifinality, which would not exist in a simpler model appropriate for the simple catchment. So why not choose a more complex catchment, which requires a distributed model?

Selecting a more complex catchment (e.g., in terms of more variable vegetation or soil) would inevitably lead to greater issues with equifinality. Thus, we sought to first constrain this approach using a straightforward framework (e.g., distributed vegetation but undistributed soil types). If the framework were not able to reproduce the hydrograph, this would merit further distribution of inputs. Future work will include evaluation of this type of approach in catchments of differing complexity – e.g., complex inputs, more variable soils, or more variable vegetation. We would suggest any catchment may be modeled using a distributed model if the end goal is to predict distributed hydrologic processes – the simplicity of the catchment does not negate our interest in testing whether observations of streamflow and SWE may constrain the simulation of distributed hydrologic processes.

The use of a distributed model may depend on either the spatial complexity of catchment characteristics, or the desire to simulate spatially-distributed hydrologic behavior. As these types of models are regularly being used to simulate spatially-distributed behavior, regardless of the availability of data to constrain model inputs, parameters, or simulations, we specifically chose a more simplistic catchment where we could maintain a parsimonious number of model parameters.

Along the same lines: Why was a catchment selected with such little available observations as 'hard truth'? Why not choose one with a network of observed groundwater tables, ET, nested discharge observations, spatially distributed information on soil type and soil depth etc.?

We have chosen a site with nested discharge observations, and have presented these nested results in this manuscript. We do not incorporate comparisons to observed well behavior and ET in part because we are interested in the patterns of this information, not just matching these values to a single point. As we discuss on page 8, internal simulations of catchment behavior may still be incorporated into evaluations of distributed model behavior. As internal catchment measurements are often difficult to come by, we have sought to specifically incorporate an evaluation of model predictions in the absence of 'hard truth', to show that this type of evaluation is possible.

In fact, the model elements are all set to the same soil depth and soil type, which makes it much less distributed as it could be.

While we do not distribute soil type or depth, this is in part framed by experimental observations at this site (as outlined on page 5), as well as the strong tradeoff in equifinality that exists as more and more 'types' are added to a given distributed characteristics, contrasted with whether there are enough available observations to truly distribute soil types in space as well as with depth. Thus, we opted to create a parsimonious distributed representation of the system. Vegetation types are distributed based on vegetation height, as we expect this to be an important determinant of system behavior.

Placing the study in a better equipped catchment would offer the opportunity to fill the very nice framework with many more signatures and metrics, especially those evaluating spatial patterns. Furthermore, this would have opened the opportunity to compare the value of constraints formulated as aggregated/heuristic expert knowledge to 'hard' constraints based on observations. This is a clear miss.

From this comment, we have clarified one of our objectives for this study – in the absence of spatially distributed measurements, can point observations, especially streamflow, inform catchment patterns of hydrologic behavior? We have altered the text on page 4 to reflect this point:

"Secondarily, we also explore whether this type of approach, using observations of a subset of hydrologic processes, may inform simulations of other unmeasured spatially-distributed hydrologic processes. Thus, we seek to test whether temporal observations may contain information regarding simulation of hydrologic patterns."

This point is further clarified in the results (4.8) and discussion (5.5) sections.

Along the same lines: All evaluations are done for a single year, and for the calibration period. This way it is impossible to judge

− to which degree the remaining behavioral parameter sets are dependent on the chosen calibration period, and

− whether the behavioral parameter sets found in calibration are still behavioral during a different, validation time. This is a clear miss.

Model predictions are performed for two years – a one year calibration period and a one year validation period (section 2.2.3). We have clarified this point in the text, as it was a source of confusion.

We additionally include all streamflow performance metrics in Table B1 of the Appendix, to enable comparison of results to another period. As can be seen from Table B1, most sets maintain high performance for WY 2007 with respect to streamflow at LSC, with the exception of slightly higher RRE values for sets 3 and 5 and SFDCE for set 4.

While the period considered for model performance is relatively short, it still represents a large computational burden in terms of the size of the catchment being resolved (22.5 km^2, 10 m by 10 m resolution) as well as the sampling procedure employed. We expand on this point, as well as possible ways to address this challenge of computational burden in the future, on page 17, with the following text:

"Ultimately, our ability to resolve issues with equifinality and identify appropriate parameter sets in space and time is challenged, as it was in this study, by the computational demand of complex models. Executing model predictions for the relatively short period of time investigated in this study across 10,000 parameter samples required thousands of computing hours (and even longer periods if the modeller retains or "saves" spatial predictions across the catchment). While distributed, physically-based models like DHSVM may have the ability to resolve predictions of hydrologic processes through space and time, we do not yet have effective, computationally inexpensive approaches for evaluating and representing uncertainties in these types of applications. In order to put these types of models to the test, we need better parameter sampling strategies (e.g., Rakovec et al., 2014; Jefferson et al., 2015) and alternative approaches to those we use for conceptual models, where executing a model many times is not a challenge or limit on analysis. This may come in the form of new methods, or alternatively, approaches that evaluate model adequacy via frameworks for computationally frugal analysis (Hill et al., 2015). While quantifying or limiting equifinality may always be a challenge for physically-based, distributed catchment models, we likely will need to reframe our approaches for evaluating the uncertainties associated with complex model applications. This challenge may be best addressed by encouraging interaction across the conceptual modelling community and the fully, distributed, physically-based modelling community, to address broad issues related to uncertainty and equifinality that, it can be argued, plague all models of any complexity (Hrachowitz and Clark, 2017)."

The authors advocate a multi-criteria approach to identify behavioral parameter sets. However, looking at the criteria in Table 2, two questions arise

Parameterization of ET plays an important role in the model (parameters 22 to 53 in Table 1). However, ET is only used as a very weak constraint (300-650 mm/a). Why not also evaluate the model with respect to ET error, ET timing, ET peaks etc.? From the text, my understanding is that ET estimates from local observations exist. This would be another important and independent criterion.

[Even if the constraints are not very narrow,

We compare our results primarily to streamflow and SWE because we sought to demonstrate results for observations that are likely to be present in other experimental catchments (now clarified on page 10). Additionally, as this is an arid, snowmelt driven system, we chose to compare model simulations to SWE, as we concluded this may be a more important criterion. In a screening sensitivity analysis of model sub-catchments, Kelleher et al. (2015) showed for this system that vegetation parameters (22 to 53 in Table 1) were key to the prediction of metrics assessing SWE. A smaller subset of model parameters directly influenced metrics related to ET. Thus, we conclude that these model parameters are more important determinants of SWE, though are still important to ET, and therefore chose to evaluate model simulations with respect to SWE.

While it may appear that ET is only weakly constrained, we found that, in fact, it is particularly discerning. To assess the impact of ET and other model constraints on the final set of parameters, we calculated the effect of removing a single constraint on the remaining number of parameter sets that meet all other criteria. Out of all possible metric constraints, removing constraint on ET has the biggest impact on the number of final parameter sets. This illustrates that even our weak constraint can have an impact on matching catchment-wide behavior.

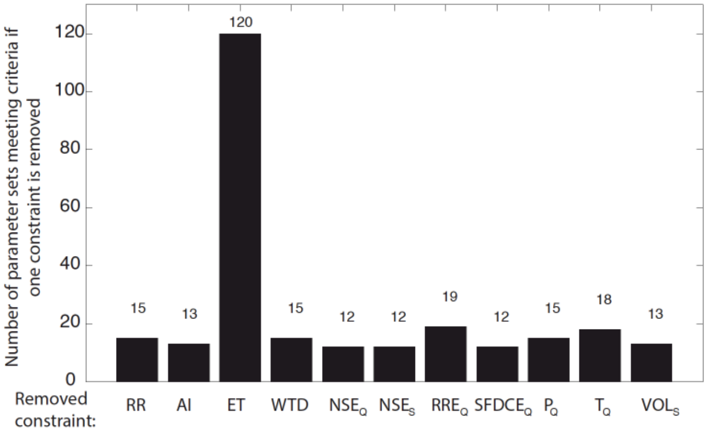

Figure B3: The impact of removing a single constraint on the number of final behavioural parameter sets.

I do not understand the usage of AI: From my understanding of the text, both PE and P are from observations, so AI is independent of the model. So how can this criterion be used as a signature for model evaluation?

PE is calculated using a Penman-Monteith approach, and therefore is also impacted by vegetation parameters. This is stated on Page 9 lines 23-24.

Minor points

From my experience, the main control on parameter equifinality is model structural choice. The authors discuss this important issue briefly in the conclusions. I encourage them to discuss this aspect in more detail, although I am well aware that model structural choice is not the topic of the paper. However, it can offer an avenue of progress to reduce the still-high equifinality of the final behavioral set of parameter sets.

We have centralized this discussion in section 5.6 with the following text:

"In this vein, the choice of model structure may also offer another opportunity to reduce equifinality (Clark et al., 2008; Pokhrel et al., 2008; Samaniego et al., 2010; Rakovec et al., 2016). In particular, the extensive body of literature on parameter regularization may offer a pathway for maintaining spatial complexity and consistency while reducing the number of free model parameters (Hundecha and Bardossy, 2004; Hundecha et al., 2008; Samaniego et al., 2010; Rakovec et al., 2016). Alternatively, there is also a body of work that treats the model framework itself as a form of uncertainty, testing different model structures as hypotheses for how a catchment may function (Clark et al., 2008; Clark et al., 2011; Fenicia et al., 2011;

Hrachowitz et al., 2014). This approach may also provide an alternative to predicting hydrology via a model with fewer parameters than the distributed application shown here, with a model structure that incorporates the level of detail mandated by the complexities of the catchment (e.g., Euser et al., 2015; Zehe et al., 2014). As encouraged by Beven (2002), to best represent catchment behaviour, we may need to not only focus on model parameters, but also the model structure in terms of how this reflects the physical landscape."

P8/l29: RR instead of PET?
PET varies with model parameters in DHSVM as it is calculated following a Penman-Monteith formulation.

P12/L17: Where is Appendix D?
We have corrected this to 'Table B1'.

P13/L3: Why not also compare simulations to well observations? (see also my major comments)
Please see our response above.

P13/L7: Appendix B3 instead of B only?
We have changed this.

P14/L22-23: … suggest that evaluating certain types of internal behavior by point observations…?
We have corrected this sentence to:
"Together, these results broadly suggest that not all observations will reduce equifinality."

P21/L14: Figure B3 instead of Fig 9?
We leave the reference to Figure 9, but include 'Figure B3' in the previous sentence.

Fig 1: − A) The location of the label 'Tenderfoot creek' is misleading − A) is the scale really [km]? − C) add a legend (which gauge is which)
We have corrected these points in Figure 1.

Fig 4, caption: remaining instead of removed?
We have changed this to 'retained'.

Fig 6, caption: Where are the black dotted lines? Where is (a) and (b)?
We have corrected this figure caption.

Fig 7: Please add year indicators (2007, 2008), a legend (which gauge is which) and for clarity add in the caption that these are plots for the final subset of 9 parameter sets
We have made these changes.

Fig 8, caption: For clarity please add in the caption that these are plots for the final subset of 9 parameter sets
We have altered the caption.

Yours sincerely, Uwe Ehret

References:
Ahl, R. S., Woods, S. W., and Zuuring, H. R.: Hydrologic calibration and validation of swat in a snow-dominated rocky mountain watershed, Montana, USA, J. Am. Water Resour. As., 44, 1411-1430, doi: 10.1111/j.1752-1688.2008.00233.x, 2008.

Beven, K. J.: Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system, Hydrol. Proces., 16, 189-206, doi: 10.1002/hyp.343, 2002.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02, doi:10.1029/2007WR006735, 2008.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modelling, Water Resour. Res., 47, W09301, doi:10.1029/2010WR009827, 2011.

Euser, T. Hrachowitz, M., Winsemius, H.C., and Savenije, H. H.: The effect of forcing and landscape distribution on performance and consistency of model structures, Hydrol. Process., 29, 3727-3743, doi: 10.1002/hyp.10445, 2015.

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modelling: 1. Motivation and theoretical development, Water Resour. Res., 47, W11510, doi:10.1029/2010WR010174, 2011.

Hill, M. C., Kavetski, D., Clark, M., Ye, M., Arabi, M., Lu, D., Foglia, L. and Mehl, S.: Practical Use of Computationally Frugal Model Analysis Methods, Groundwater, 54, 159–170, doi:10.1111/gwat.12330, 2016.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G., and C. Gascuel-Odoux: Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, Water Resour. Res., 50, 7445–7469, doi:10.1002/2014WR015484, 2014.

Hrachowitz, M. and Clark, M.: HESS Opinions: The complementary merits of top-down and bottom-up modelling philosophies in hydrology, Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-36, in review, 2017.

Hundecha, Y., and Bardossy, A.: Modeling effect of land use changes on runoff generation of a river basin through parameter regionalization of a watershed model, J. Hydrol., 292, 281–295, doi: 10.1016/j.jhydrol.2004.01.002, 2004.

Hundecha, Y., Ouarda, T. B. M. J., and Bardossy, A.: Regional estimation of parameters of a rainfall-runoff model at ungauged watersheds using the spatial structures of the parameters within a canonical physiographic-climatic space, Water Resour. Res., 44, W01427, doi:10.1029/2006WR005439, 2008.

Jefferson, J. L., Gilbert, J. M., Constantine, P. G., and Maxwell, R. M.: Active subspaces for sensitivity analysis and dimension reduction of an integrated hydrologic model, Comput. Geosci., 90, 78-89, doi:10.1016/j.cageo.2015.11.002, 2016.

Kelleher, C., Wagener, T., and McGlynn, B.: Model-based analysis of the influence of catchment properties on hydrologic partitioning across five mountain headwater subcatchments, Water Resour. Res., 51, 4109–4136, doi:10.1002/2014WR016147, 2015.

Nippgen, F., McGlynn, B. L., and Emanuel, R. E.: The spatial and temporal evolution of contributing areas, Water Resour. Res., 51, 4550– 4573, doi:10.1002/2014WR016719, 2015.

Pokhrel, P., Gupta, H. V., and Wagener, T.: A spatial regularization approach to parameter estimation for a distributed watershed model, Water Resour. Res., 44, W12419, doi:10.1029/2007WR006615, 2008.

Rakovec, O., Hill, M. C., Clark, M. P., Weerts, A. H., Teuling, A. J., and Uijlenhoet, R.: Distributed evaluation of local sensitivity analysis (DELSA), with application to hydrologic models, Water Resour. Res., 50, 409-426, doi: 10.1002/2013WR01463, 2014.

Rakovec, O., Kumar, R., Attinger, S., and Samaniego, L.: Improving the realism of hydrologic model functioning through multivariate parameter estimation, Water Resour. Res., 52, 7779-7792, doi: 10.1002/2016WR019430, 2016.

Samaniego, L., Kumar. R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resour. Res., 46, W05523, doi:10.1029/2008WR007327, 2010.

Smith, T., Marshall, L., McGlynn, B., and Jencso, K.: Using field data to inform and evaluate a new model of catchment hydrologic connectivity, Water Resour. Res., 49, 6834–6846, doi:10.1002/wrcr.20546, 2013.

Zehe, E., Ehret, U., Pfister, L., Blume, T., Schröder, B., Westhoff, M., Jackisch, C., Schymanski, S. J., Weiler, M., Schulz, K., Allroggen, N., Tronicke, J., van Schaik, L., Dietrich, P., Scherer, U., Eccard, J., Wulfmeyer, V., and Kleidon, A.: HESS Opinions: From response units to functional units: a thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments, Hydrol. Earth Syst. Sci., 18, 4635-4655, doi:10.5194/hess-18-4635-2014, 2014.