Dear Dr. Hrachowitz,
We thank you for your suggested revisions. In line with your comments, we have revised and checked all figure captions, references, and numbering of figures, tables, and appendices. Our introduction and discussion sections now include references to examples of regularization in watershed modeling, an oversight on our part. We have also addressed many of your additional minor comments, which were helpful for clarifying the direction of the manuscript. We have responded to each of your comments below (in red text), and noted changes we have made in response.

Many thanks,
Christa Kelleher and Colleagues

The manuscript "Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding" by Kelleher et al. addresses a topic that is of critical importance for hydrological modelling applications: the respective value of multiple, different data sources and model evaluation metrics to identify meaningful parameters sets within limited uncertainty. The experiment is well-designed and described. I would be glad to see this contribution eventually be published and there are only two, relatively minor concerns I would like to encourage the authors to address to strengthen the manuscript:

(1) The manuscript would strongly benefit from being proof-read with a bit more care. The present version comes across as a bit sloppy with e.g. wrong figure numberings, wrong or missing references to figures, wrong references to or missing appendices and relatively imprecise figure captions.

In line with these suggestions, we have revised captions and checked mentions of references, figures, and appendices.

(2) Being a highly important and ubiquitous topic in environmental systems models, I was quite surprised that the link to related techniques dealing with ill-posed inverse problems commonly applied in other fields is completely missing. This includes for example regularization which is quite a standard technique to reduce parameter equifinality in e.g. ground water models. It is just that it has not yet found, for whatever reason, its way into mainstream surface hydrology. Linked to that is the complete lack of this manuscript to refer to distributed model frameworks that make use of regularization (although they often refer to it as regionalization, which is the same thing, really), e.g. the work of Luis Samaniego's group on the mhM model (e.g. Samaniego et al., 2010; Kumar et al., 2013)

We appreciate the recommendation to include this subset of literature in our revised manuscript. We have made changes to the introduction and discussion to address this oversight.

We have added the following to the introduction (page 3): "One noted exception is the work that has been done to parameterize models via regularization (Hundecha and Bardossy, 2004; Hundecha et al., 2008; Samaniego et al., 2010; Rakovec et al., 2016). Regularization creates global functional relationships describing transfer functions that link model parameters and catchment characteristics (e.g., Samaniego et al., 2010; Kumar et al., 2013). Regularization has the ability to limit parameter equifinality in part because the number of parameters used to describe global functional relationships is far fewer than the number of possible total model parameters using cell-by-cell parameterization."

Detailed comments:
(3) P.1,l.11-12: distributed models do not necessarily require more parameters, if only the input is distributed (e.g. Ajami et al., 2004; Das et al., 2008; Kling and Gupta, 2009; Fenicia et al., 2008; Euser et al., 2015). I would suggest to make this distinction clear somewhere in the introduction section.

To address this point, we have amended the introduction to include a paragraph describing different types of distributed models. In particular, we include the statement:
"Alongside variations in the structural equations, models may take a number of different forms, including those that use only distributed inputs (e.g., Ajami et al., 2004; Das et al., 2008; Kling and Gupta, 2009; Fenicia et al., 2008; Euser et al., 2015), those that may lump together portions of a watershed with similar characteristics in a semi-distributed fashion (e.g., Leavesley et al., 1983; Flugel, 1995; Ajami et al., 2004; Das et al., 2008; Zehe et al., 2014), those that may upscale sub-grid variability to parameterize models (e.g., Samaniego et al., 2010; Kumar et al., 2013), and those that use physically-based parameter values (e.g., Wigmosta et al, 1994; Qu and Duffy, 2007) versus conceptual values or transfer functions (e.g., Samaniego et al., 2010; Kumar et al., 2013).

(4) P.1,l.18: I find the term "certainty" quite problematic. Given all the different sources of uncertainty, how can we ever think of a parameter as being "certain". What would that mean in that context? "Certain" with respect to what? Reality? Uncertain observations?

We have removed the term certainty from the abstract, and all other mentions of this terminology from the manuscript.

(5) P.1,l.27ff: it is not quite clear what the authors define as distributed model in this study. Note that both, physical models but also conceptual models (for example done with mhM) can be applied in distributed ways. Please provide a more precise definition.

We see this as a very important point. We have clarified this point by adding a paragraph to the introduction (page 2): "Alongside variations in the structural equations, models may take a number of different forms, including those that use only distributed inputs (e.g., Ajami et al., 2004; Das et al., 2008; Kling and Gupta, 2009; Fenicia et al., 2008; Euser et al., 2015), those that may lump together portions of a watershed with similar characteristics in a semi-distributed fashion (e.g., Leavesley et al., 1983; Flugel, 1995; Zehe et al., 2014), those that may upscale sub-grid variability to parameterize models (e.g., Samaniego et al., 2010; Kumar et al., 2013), and those that use physically-based parameter values (e.g., Wigmosta et al, 1994; Qu and Duffy, 2007) versus conceptual values or transfer functions (e.g., Samaniego et al., 2010; Kumar et al., 2013). These differences can be separated into two primary categories (1) the nature of parameter values (physically-based or conceptual) and (2) whether and how parameter values are distributed (cell-by-cell, grouped or lumped in some meaningful way, or undistributed). It is also worth noting that an important feedback on these decisions is the scale of the application alongside the scale at which inputs and parameters are distributed. In this study, we focus specifically on physically-based, fully distributed (cell-by-cell basis) models applied at the small-scale watershed scale (<25 km^2)."

(6) P.2,l.2: maybe rephrase to "Distributed models should represent the. . .."

We have made this change.

(7) P.2,l.10: why would distributed models *require* a single parameter set? This statement goes against much we know about the uncertainties involved in environmental systems models

While we agree with you (in fact, addressing approaches that use a single parameter set is the point of this paper), the reality is this is still how many researchers use fully distributed, physically-based models. We have revised this statement to:

"Model simulations and predictions require specification of parameter set(s) or ranges; selecting these set(s) and appropriate ranges is especially challenging given that distributed models require a larger number of model parameters (~50-100 or more) and longer model run times than conceptual or lumped models."

(8) P.2,l.14-15: not only valid for distributed models but for *any* hydrological model

We agree, and have altered the text to reflect this.

(9) P.2,l.19-21: not untrue, but the contribution of the developers of the mhM framework, who went to great lengths to exploit the value of regionalization/regularization, should not go unnoticed here (see above)

We have included the following text on page 3: One noted exception is the work that has been done to parameterize models via regularization (Hundecha and Bardossy, 2004; Hundecha et al., 2008; Pohkrel et al., 2008; Samaniego et al., 2010; Rakovec et al., 2016). Regularization creates global functional relationships describing transfer functions that link model parameters and catchment characteristics (e.g., Samaniego et al., 2010; Kumar et al., 2013). As the number of parameters used to describe global functional relationships is far fewer than the number of possible total model parameters using cell-by-cell parameterization, regularization is able to limit parameter equifinality."

(10) P.2,l.27: I would suggest to formulate this in a more general way by emphasizing the degree of information that is used on the prior distributions

We have changed this to:
"Decisions regarding how to constrain ranges and distributions for priors on model parameters often depend on…"

(11) P.3,l.3: should read as ". . .to justify the selection. . ."

Noted and changed.

(12) P.4,l.18,figure1: where do all the tall trees hide?

There are a small number of tall trees, and a very large matrix of values (our simulated grid is 560 by 760 cells). We have increased the contrast of this figure to highlight tall trees, but note that the dominating pattern here are the differences between understory and canopy vegetation.

(13) P.5,l.2: what does "tandem" mean?

By 'In tandem' we meant 'at the same time' – we have changed this to 'concurrently'.

(14) P.5,l.7,table 1: add units in table 1; what is referred to as "ranges"? is it the uninformed prior parameter distributions? Please clarify.

We have added units to Table 1, and clarified our description of parameter ranges as "the minimum and maximum bounds on uninformed (uniform) prior parameter distributions".

(15) P.5,l.24: please be more specific here – what does shading include? Does it combine shading due to aspect and shading due to topographic features (e.g. mountain ridge) obstructing sunlight to reach certain locations?

We have clarified this point on page 6 with the following text: "These shading maps incorporate both the effects of topographic shading, diel variability, and monthly variability in sun position and angle."

(16) P.5,l.26: please be more specific here – what are "allowable" ranges?

This became a problem with spatial boundary conditions, where humidity and vapor pressure deficit on the coldest days (< -40 deg F) at the far edges of the model reached unrealistic conditions. We have amended the word 'allowable' to 'unrealistic'.

(17) P.5,l.30ff: not entirely clear – which period was the model calibrated for? The entire 01/10/2006-01/10/2008 period? If so, I was wondering if it was not more instructive to retain some months for post-calibration evaluation (i.e. "validation") to see the effect of signatures etc. under these conditions as well.

As stated in the text (Page 6 lines 27 – 29), we use the 2008 water year for calibration, and retain the 2007 water year for validation. We use April 1, 2006 through September 30, 2006 to initialize the model. We have added dates for clarity, and have moved this text to earlier in the paragraph. We exclude a period of initialization (Apr 1, 2006 – Sept 30, 2006) from any model analysis.

(18) P.8,l.27,table 2: which runoff ratio was used? Over the entire period? Annual runoff ratios? Or seasonal runoff ratios? They carry different information content, and from earlier work we

found that in particular the seasonal runoff ratios carry considerable information. This may be worth looking into.

We use an annual runoff ratio for the 2008 water year. To target seasonality, we use the slope of the flow duration curve for 10th through 30th percentile flows (mainly to fit to the rising and falling limbs of the hydrograph), though a seasonal runoff ratio is an interesting suggestion for future analysis to ensure model simulations match seasonal behavior.

(19) P.8,l.28ff,table 2: PET is constrained with regional data, but how were these regional PETs estimated? Are they more reliable than the penmen-monteith derived estimates here? if so, why? If not, what is the point of using them. I think this point warrants some discussion.

The estimate of aridity used to constrain the model is an average value based on daily precipitation and daily maximum and minimum air temperature data for the period of 1950 to 2000, with potential evapotranspiration (PET) modeled via the Hargreaves method (Hargreaves et al., 1985). Thus, this constraint is meant as a check on long-term average conditions, to ensure that the model stays within regionally realistic estimates. We have modified the text to reflect this point:

"Constraints on AI from this dataset represent long-term average conditions for the region, and are used to broadly eliminate simulations outside of the realm of regionally realistic average estimates."

To corroborate the validity of this 'space-for-time' approach, we calculated PET following a combination method described by Oudin et al. (2005) and calculated AI from estimated PET and annual precipitation from the Stringer Creek SNOTEL location. Values for AI ranged from 0.39 to 0.64, well within the constraints extracted from the regional estimates.

Table 1. Estimates for aridity index (AI) based on annual precipitation (Stringer Creek SNOTEL) and potential evapotranspiration calculated following Oudin et al. (2005).

Year	PET (m)	P (m)	AI (-)
2006	0.47	0.80	0.59
2007	0.48	0.75	0.64
2008	0.41	0.73	0.56
2009	0.43	0.74	0.57
2010	0.40	1.04	0.39
2011	0.41	0.86	0.47

DHSVM uses a cascading approach (from overstory to the understory) to estimate potential evapotranspiration, with differences largely based on vegetation height parametrizing aerodynamic resistance. As indicated by Figure B2, AI does appear to have information content, but it is largely with respect to very poor model runs likely produced as the combination of the extreme vegetation height parameter combinations.

To test the effect of constraining AI on the final set of parameters, we recalculated the number of parameter sets that would meet all constraints excluding the constraint on AI. We found that this number would increase from 9 to 13 parameter sets, demonstrating that this constraint may only serve to remove a small number of sets that do not meet other constraints. We have added a figure demonstrating how the number of final parameter sets changes if only one constraint is removed to Appendix B, Figure B3, as this addresses points raised by both reviewers.

(20) P.9,l.15,table 2: does this refer to the maximum flow of each year? In many cases in particular the extreme events are subject to unproportionally high observation errors. Thus, do you not run into the risk of forcing the model to reproduce a value that is relatively likely to be considerably off from reality (cf. epistemic errors!!)? I also think that this point needs some more reflection.

This metric does refer to the maximum flow of each year. We found that without this constraint, the model consistently underestimated peak flows. We do agree with your point regarding potential for fitting to epistemic errors. This is in part why we left windows on acceptable "errors" wide (e.g., +/- 35%), so as to ensure that we do not overly constrain these single observation points in time. While we agree that overly constraining this value could lead to problems, in previous iterations of this manuscript, criticisms of our modeling approach primarily focused on whether or not our simulations emulated this peak value. Thus, we aimed to add a target error for this value, but one that was wide.

To reflect this, we have added the following text to page 10:
"Additionally, as peak flow is likely to be most influenced by epistemic errors (e.g., Westerberg and McMillan, 2015), we left bounds on peak flow to be reasonably wide, so as to avoid incorrectly constraining a potentially uncertain measurement."

(21) P.9,l.16,table 2: the equation for the error in peak timing (again: does this apply for the timing in each individual year?) seems to involve some magic – I could not figure out how a ratio of days over days would result in days. In addition, I am not sure if this ratio makes sense, as we talk about the day of the year, at least as I understand it. Would be probably more meaningful to just minimize or constrain the absolute errors, i.e. abs(Tq,s-Tq,o)

This was an error in Table 2, and has been corrected. The constraints were correctly identified.

(22) P.10,l.8ff,figure 4: this is not entirely clear. Do the grey and coloured bars show the parameter sets that are retained or those that are discarded? The text and figure captions seem to somehow contradict each other. Please check.

We note the confusion, and have corrected the captions to 'retained'.

(23) P.10,l.10-11: what about the effect of observational uncertainty in precipitation (in particular snow accumulation!!) and peak runoff - see comment (20)??

Again, we note that we are not fitting exactly to this value, but including any model simulations that are within +/- 35% of this value. We agree that there are likely to be uncertainties in these measurements, hence our choice to leave bounds on these values to be wide. As many sets match SWE behavior, we do not think we are overly constraining peak flow behavior on the basis of observational uncertainty in snow accumulation.

(24) P.11,l.11,figure 6: I am struggling with this figure. Caption seems to describe an entirely different figure. Where is the black dotted line (the one I see merely seems to separate the two hydrological years)? Where are subplots (a) and (b)?? what is the light blue line (observation?)? please clarify.

This figure caption has been corrected.

(25) P.12, l.3, figure 8(?): figure number seems to be wrong. Is this not figure 7? if not, where is figure 7? Please indicate years on x-axes. What are the grey shaded areas? Are these model ensembles or uncertainty ranges? If the latter, how were they constructed and, in particular, where do the gaps come from??

We have corrected the text to state this is Figure 7. We have added years to the x-axis on Figure 7. The grey lines each correspond to a simulation from a different parameter set. We have clarified the figure caption to reflect this point.

(26) P.12,l.4: what is meant by "relative timescales"?

We have amended this to "match periods" for clarity.

(27) P.12,l.17: I cannot find Appendix D (and C not either, for that matter).

This has been changed to 'Table B1'.

(28) P.12,l.23: please reconsider the use of the term "parameter certainty" (see above)

We have changed this to 'Parameter Uncertainty'.

(29) P.12,l.30: what is meant by "original" distribution? Is it the uninformed prior distribution (i.e. table 1)? If so, please call it also like that.

This has been changed.

In addition, although I see and understand the intention here, I am wondering how much this also depends on the choice of the parameter range in the prior distribution. If a narrow prior range was chosen, the normalized values shown here may be less constrained than if a wider prior range was chosen.

We agree that this depends on the width and magnitude of the uninformed priors. While this likely has an effect, we do note that parameters where ranges appear to be more constrained do have a variety of magnitudes and ranges. For instance, lateral conductivity (parameter 5), varies across three orders of magnitude, exponential decrease in conductivity with depth (parameter 6), varies across two orders of magnitude, and understory maximum stomatal resistance (parameter 26) varies across one order of magnitude. The meaning of "narrow" and "wide" ranges also has some implications for our assessment – we intentionally left all ranges relatively wide in terms of values and ranges reported in the literature. Our point with this figure is mainly to highlight that there are a few parameters where ranges are narrowed by this type of analysis, which is interesting given the number of interactions between parameters in models of this type.

In addition, are the individual parameter sets likelihood weighted to give more weight to better solutions? if not, why not? Any thoughts on this?

We did not assign likelihood weights to parameter sets, in part because the parameter sets shown in this figure meet all criteria for what we have identified to be 'behavioral' solutions. We are mixing both signatures and error metrics, as well as a number of different signatures and error metrics. Thus, while some may have better values for one objective function, they may have poorer values for another. As there was no clear or easy way to "weight" these parameter sets across these many 'behavioral' criteria, we opted to leave these values unweighted.

(30) P.13,l.6: where are these predictions included? Figure b3? Please specify.

This has been changed.

(31) P.13, l.13, figure 9: the figure and captions are a bit confusing. What is meant by set 1/5/9? And how is set 1 drier than set 9? Please clarify.

This labeling originally referred to Figure B3. We have replaced the labeling with sets 1, 2, and 3, to indicate that these predictions come from three different parameter sets, and now reference Figure B3 in the caption. We have also replaced the words wetter and drier with "lower" and "higher" water table.

(32) P.18,l.6: I think the wrong reference is provided in the reference list. The paper you are referring to is Euser et al. (2015, HP), but in the reference list Euser et al. (2013, HESS) is provided.

We have corrected this.

(33) P.18, l.9: some references to mhM would fit in nicely here.

We have included the following text on page 19 (section 5.6):
"In this vein, the choice of model structure may also offer another opportunity to reduce equifinality. 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., 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., Zehe et al., 2014; Euser et al., 2015)."

Best regards, Markus Hrachowitz

References:

Ajami, N. K., Gupta, H., Wagener. T., and Sorooshian, S.: Calibration of a semi-distributed hydrological model for streamflow estimation along a river system, J Hydrol., 298, 112-135, doi: 10.1016/j.jhydrol/2004.03.033, 2004.

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.

Das., T., Bardossy, A., Zehe., E, and He., Y.: Comparison of conceptual model performance using different representations of spatial variability, J. Hydrol., 356, 106-118, doi:10.1016/j.jhydrol.2008.04.008, 2008.

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., Savenije, H. H., Matgen. P., and Pfister, L.: Understanding catchment behaviour through stepwise model concept improvement, Water Resour. Res., 44, W01402, doi: 10.1029/2006WR005563, 2008.

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.

Flugel, W.-A.: Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Brol, Germany, Hydrol. Process., 9, 423-436, doi: 10.1002/hyp.2260090313, 1995.

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.

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.

Kling, H., and Gupta, H.: On the development of regionalization relationships for lumped watershed models: The impact of ignoring sub-basin scale variability, J Hydrol., 373(3), 337-351, doi: 10.1016/j.jhydrol.2009.04.031, 2009.

Kumar, R., Livneh, B., and Samaniego, L.: Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme, Water Resour. Res., 49, doi:10.1002/wrcr.20431, 2013.

Leavesley, G. H., Lichty, R. W., Troutman, B. M., and Saindon, L. G.: Precipitation-Runoff Modeling System: User's manual, Water Resources Investigations Report 83-4238, United States Department of the Interior, Denver, Colorado, USA, 1983.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andreassian, V., Anctil, F., Loumagne, C.:Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modeling, J. Hydrol., 303, 290–306, 2005.

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.

Qu, Y., and Duffy, C. J.: A semidiscrete finite volume formulation for multiprocess catchment simulation. Water Resour. Res., 43, W08419, doi:10.1029/2006WR005752, 2007.

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.

Westerberg, I. K. and McMillan, H. K.: Uncertainty in hydrological signatures, Hydrol. Earth Syst. Sci., 19, 3951-3968, doi:10.5194/hess-19-3951-2015, 2015.

Wigmosta, M.S., Vail, L., and Lettenmaier, D. P.: A distributed hydrology-vegetation model for complex terrain, Water Resour. Res., 30, 1665-1679, doi:10.1029/94WR00436, 1994.

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.