**Response to reviewers' comments**

"Flood and drought hydrologic monitoring: the role of model parameter uncertainty" by N. W. Chaney, J. D. Herman, P. M. Reed, and E. F. Wood.

We thank the reviewers for their time and helpful comments. We have addressed each point below. Reviewer comments are shown in *blue italics*, while author responses are shown in unformatted text.

**Reviewer #1**: *This is an interesting study looking at parameter uncertainty and its impact on extreme hydrologic event modeling by applying annual, monthly, and daily scale constraints to ensemble simulations corresponding to 10,000 Latin hypercube sample sets*

*Page 1698, line 19: To me "accurate" means unbiased. I think the priors are better to be accurate (unbiased), precise (reduced uncertainty, narrow distribution), but also appropriately represented (e.g., derived with minimum-relative-entropy or maximum-entropy concepts). The shapes of the prior pdfs might significantly affect the sensitivity analysis results, especially for a problem with a high-dimensional parameter space.*

Here we refer to improvement in general rather than a precise statistical meaning. The reviewer raises a good point and we have revised this sentence accordingly. This sentence aims to reflect discussion section 5.2. Here it is recognized that using the same prior distribution at every grid cell throughout the globe without accounting for local characteristics is a missed opportunity. Ongoing research by the the authors is looking into using available high resolution data and available observation networks to provide improved prior distributions at each grid cell.

*Page 1699, line 27: Yes I agree that it is possible that an optimization get the right answer (e.g., good fits) for wrong reasons, for example, when model structural uncertainty or data uncertainty is large. The ensemble framework would make it possible to separate the parameter uncertainty from the data/model structural uncertainty.*

Yes, the ensemble approach allows an exploration of both parameter uncertainty and model structural uncertainty. This study focuses on the issue of parameter identifiability, which proves a difficult task even for a single model structure given observations of global runoff. One recent study to quantify model structural uncertainty is Gong et al. (2013), who take an information theoretic approach to dividing the effects of aleatory and epistemic uncertainty. Although not the primary focus, the impact of model structural uncertainty is apparent in Figure 1 (now Figure 2) for the regions in which no parameter sets meet the error criteria, primarily in arid regions.

*Page 1701, line 1: the use of 10000 sample sets is arbitrary. Please justify. It is unclear whether this is adequate without a convergence test (e.g., evaluating the SA results vs the number of LH samples). The required numbers of samples depends on choices of*

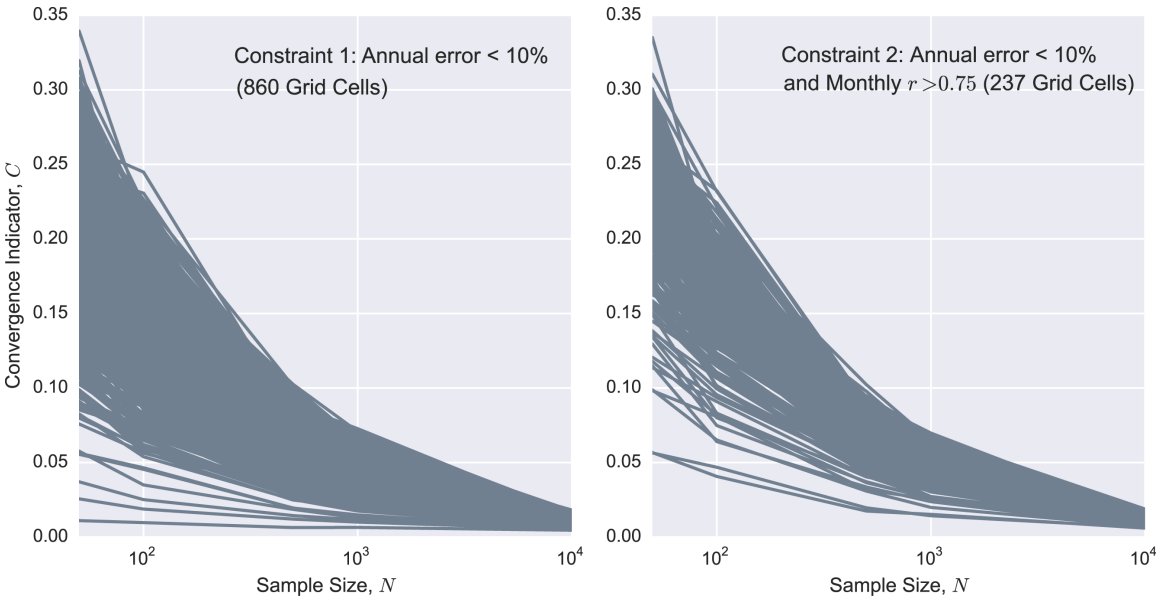
*response variables/metrics. BTW, one advantage of LHS is that you can add augmented samples to the existing ones if necessary.*

The ensemble size for this study was chosen to balance the need for a suitably sized behavioral ensemble with computational requirements. As Figure 1 (now Figure 2) shows, many grid cells contain a very small percentage of the original 10,000 parameter sets in the behavioral set; those containing zero could be considered an indication of structural error. The ensemble size of 10,000 represents the limits of current computational power, and to the authors' knowledge is the largest Monte Carlo study of parameter uncertainty in the field to date.

The variable for which convergence should be monitored is the number of behavioral parameter sets—for example, the ratio of the standard error of this estimate to its mean. We suggest the following convergence indicator:

$$C = \frac{\sigma_{\overline{N}_b}}{\overline{N}_b}$$

Where $\overline{N}_b$ is the number of behavioral parameter sets at a particular sample size. Using a bootstrapping scheme with 100 resamples, we observe the convergence of this indicator to approximately zero as a function of increasing sample size:




At an ensemble size of 10,000, the standard error of the number of behavioral parameter sets is approximately 1-2% of the estimates themselves, suggesting that this is an appropriate size. The convergence of the CDF distance metric follows from the convergence of the number of behavioral parameter sets.

*Page 1702, line 21: that is, assume that model structural uncertainty and data uncertainty are negligible.*

To build the observation dataset, we assume that the model structure is adequate to capture the spatial heterogeneity of runoff at a 1.0 degree spatial resolution. Understanding that model structural uncertainty will exist and the observations are uncertain leads us to use a parameter screening criteria (relative error< 10% and linear correlation > 0.75) that is able to account for existing uncertainties in the derived observation runoff fields.

*Page 1703, line 9: what is "temperature" climate group? It should be "temperate" or "mesothermal". Why not spell out the 5 veg groups and the 5 precipitation groups as well?*

Thank you for this observation. The typo has been corrected. It is now defined as the temperate climate group.

*Page 1704, line 5: the range for the parameter Ksat seems is too narrow. And is it sampled in log10 space?*

Yes, the parameter $K_{sat}$ is sampled in $\log_{10}$ space. As noted in the caption of Table 1, all parameters spanning two or more orders of magnitude are sampled in $\log_{10}$ space. The chosen minimum and maximum values of $K_{sat}$ are selected based on the look-up table used for the VIC land surface model. This table which can be seen at http://www.hydro.washington.edu/Lettenmaier/Models/VIC/Documentation/Info/soiltext.shtml has a $K_{sat}$ minimum value of 297.5 and a maximum value of 9602.5. The range selected in this study (100-10,000) encompasses these values.

*Page 1704, line 11-25: I am fine with the parameter set screening criteria (e.g., relative error > 10%, and correlation < 0.75), but I am not sure it is the best we can do by assuming the behavioral parameter values to have the same weights in the posterior distributions. The procedure is similar to rejection sampling, but without replacement and is not dependent on previously accepted sample values. Two simple practices might yield better estimate of the posterior distributions: 1) the samples are accepted with a probability as a function of the corresponding misfits; 2) the samples are assigned weights as a function of misfits (which assumed to be normally distributed). Again, the point is that the behavioral sample values are not equally probable.*

We treat the behavioral samples with equal probability when developing the posterior CDF due to the potential for errors in the observed data and model structure. Although we do not explicitly estimate these errors, they may significantly alter the goodness-of-fit measures (annual relative error and monthly correlation). Therefore, we use the behavioral criteria as a filter and do not attempt to distinguish the likelihood of individual samples beyond that.

*Page 1705, line 21: an alternative metric to CDF distance could be relative entropy (or Kullback-Leibler distance), which measures the relative change in information/uncertainty.*

The two metrics share some similarities. The CDF distance used in this study between two cumulative distributions $F(x)$ and $G(x)$ can be written as:

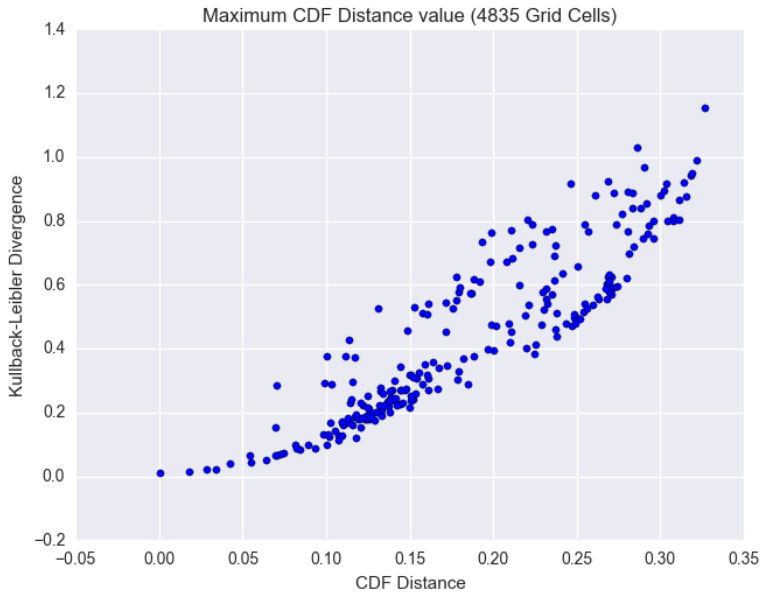
$$D_{CDF} = \int_{x_l}^{x_u} |F(x) - G(x)|\, dx$$

Where $x_l$ and $x_u$ are the lower and upper bounds of the parameter in question, which we have normalized to [0,1] to improve interpretability of the result. This equation has been added Section 3.2.1.

By contrast, the Kullback-Leibler divergence (or expected log-likelihood ratio) between two probability distributions $f(x)$ and $g(x)$ is given by:

$$D_{KL} = \int_{x_l}^{x_u} f(x) \ln \frac{f(x)}{g(x)}\, dx$$

For the parameter in each grid cell with the maximum CDF distance value, the two measures are strongly related:




It should be noted that the estimation of PDFs to compute $D_{KL}$ requires an additional step. In this case, simple histogram binning was used, but this could be improved with kernel density estimation.

*Page 1707, line 10: how many cores/cpus are involved? Did you run the simulations or part of them in parallel?*

The model was parallelized to run one grid cell on each processor, for a total of 15,836 processors. The scale of the experiment was made possible by the Blue Water supercomputer at the University of Illinois.

*Page 1708, line 22: "a limited number of behavioral. . ." I would view the issue as existence of significant model structural errors. The screening criteria for "being behavioral" might need to be relaxed for these regions.*

Figure 1 (now Figure 2) shows the global distribution of behavioral parameter sets for progressively more difficult error criteria. Even with annual error below 20% as the sole criterion, several regions contain zero behavioral sets. It is reasonable to attribute these cases to model structural uncertainty. This issue has been clarified in the manuscript. We wish to also note that for operational flood and drought monitoring we would relax the screening criteria since we wish to have a simulation in all areas. However, the purpose of this study is to show the skill of the model using a set of predefined acceptance criteria.

*Page 1711, line 1-4: the statement is not clear to me.*

The sentence has been revised: "While predictions in tropical climates are not well constrained with this approach, the results are encouraging for monitoring the hydrologic cycle with properly-constrained land surface models in continental and polar climates".

*Page 1713, line 18: do you meant "local" temporally, or spatially, or both? Regarding the prior distributions, the shape should be considered carefully in addition to refining its range.*

This refers to the use of spatial information to inform prior distributions. (Ideally, in the absence of structural error, the optimal parameter values would not change in time, though this is rarely the case in practice).

This section has been revised:

> Given the need to rely on monthly and annual observations to constrain the model parameter uncertainty, local prior distributions should be informed by spatial land surface characteristics to constrain the initial ensemble spread and the flow duration curves. Spatially distributed information could also be used to refine the distribution family and shape of the priors in addition to their ranges.

*Page 1714, line 9: add refs for "random forests".*

The reference to Liaw and Wiener (2002) has been added.

*Page 1715, line 12: I agree that adding process models lead to higher parameter dimensionality and more parameter uncertainty. Such additions have the potential to*

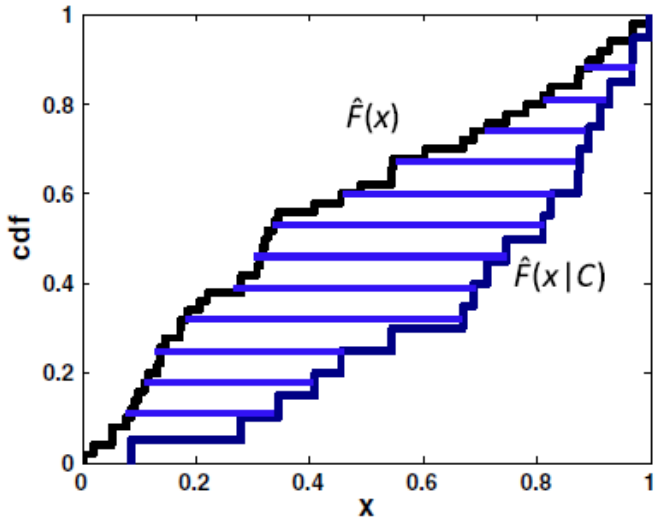
*reduce model structural uncertainty; meanwhile the increased parametric uncertainty can be reduced through inversion, hopefully in a physically more plausible way.*

We agree with the reviewer's comment, and would add that a reliable approach to inversion for more highly parameterized models will depend on accurate observations of fluxes at higher spatial and temporal resolutions.

**Reviewer #2**: *The authors setup a quite impressing experiment for global scale simulation of hydrological extremes. The paper is nice, the approach is sound and the results allow learning very much about uncertainty of model parameters*

*The use of the CDF-distance is interesting. A figure on the concepts would be useful to present it.*

Fenwick et al. (2014) provide an illustration of the CDF distance concept:




The area between the prior and posterior CDFs is used as a simple distance measure. The method proposed by Fenwick et al. (2014) can be generalized to any number of classes conditioned on a performance metric, but in this study we use only two classes (behavioral and non-behavioral). In the manuscript, we have added the equation to compute the CDF distance in Section 3.2.1 to clarify this.

*I do not catch why the authors used the whole 1948-2010 period to select their parameter ensemble and do not include any validation of their ensemble estimations of discharge. I argue that if they would have split their period into two or three slices, they could have isolated an even more narrow parameter ensemble.*

We agree that dividing the 1948 to 2010 period into different slices and then using some slices to constrain the ensemble and others for validation is preferable.

However, the observations used in this study make this approach unfeasible. As discussed in section 2.3, the observations used are the GRDC monthly climatology (12 data points per grid cell) of runoff at a 1.0 degree spatial resolution. To define the climatology, we use the longest extent possible with the available meteorological forcings (i.e., 1948-2010).

*1700 – 12-14: Please provide some references here.*

Thank for pointing this out. After a more comprehensive literature review on the subject we have decided that "Many" is not an appropriate term for the current state of accounting for meteorological uncertainty in these hydrologic monitoring systems. We have included a reference and have changed the sentence to "A growing number of hydrologic monitoring systems already include the impact of uncertainty in meteorological forcing, which should be extended to include model parameter uncertainty".

*1701 – 15-20: I understand that the Sheffield et al. (2006) is a well-cited source for having a description of the meteorological forcing. I wonder if it anyway possible to elaborate here on the impacts of using TRMM (available since 2002) on the homogeneity and accuracy of precipitation estimates. It can be here or in the discussion.*

The TRMM data is not used directly in the Princeton Global Forcing (PGF) dataset. Instead, it is used to spatially and temporally downscale coarse observation datasets that cover the entire period (e.g., CRU and the NCEP-NCAR reanalysis). For the temporal downscaling, the precipitation is downscaled from the daily NCEP-NCAR reanalysis product to a 3-hourly product by using a probabilistic approach based on sampling from TRMM. The TRMM data is also used to spatially downscale the 2.0 arcdegree meteorological data down to a 1.0 spatial resolution. This is done via a probabilistic approach based on relationships between precipitation intensity and grid cell fractional precipitation coverage. In that sense, the TRMM data is mostly used to determine the temporal and spatial properties of rainfall; the data is not actually merged therefore it does not create artificial inhomogeneities. For further details, see [*Sheffield et al.*, 2006].

*- You sample your parameter space from table 1 10'000 times (this is well declared)*
*- You apply each set to all 1 grid cells of your domain. Eg. 10'000 simulations with no spatial distribution of the parameters (this is hinted at lines 1704:7-8, but I'm not sure this is what you mean there.*
*- You evaluate separately for each cell the behavioural parameters (well declared) - You make the analysis*

This is correct. The text has been updated to clarify the spatial distribution issue:

> Each parameter is drawn from a uniform distribution; parameters that cover 2 or more orders of magnitude are sampled in $\log_{10}$ space. For each LHS

parameter set, the model is run at a 3 h time step between January 1948 and December 2010 with a 10 year spin up period. Parameter values are assumed to be uncorrelated in space. The 10 000 ensembles are run for all 1.0 degree land grid cells over the globe excluding Greenland and Antarctica (15 836 grid cells in total).

*Another question here: Why using the whole period 1948-2010 to discriminate the parameters and not dividing into let's say 20 to 30 years slices and have for instance a periods for training and evaluation?*

(See response to duplicate comment above)

*1708 – 4-9: I think that when the number of "behavioral parameter sets" rapidly declines when introducing additional constraints, then many of the realization fulfilling the first constraints were right for the wrong reason. Nice way to show this here!*

We thank the reviewer for highlighting this contribution. It is possible that the inclusion of more observed data (daily runoff, for example) would invalidate an additional subset of the ensemble members.

*1708 – 20-26: True statements. Does this in any way deviate from your expectations?*

Over the past years, the authors have been involved in the development and implementation of the African Flood and Drought Monitor [*Sheffield et al.*, 2014]. This experience brought to light the strengths and weaknesses of these types of systems due to model uncertainty. The primary goal behind this study was to quantify and characterize this uncertainty. Although not surprised by the results, it has made it clear that structural uncertainties play an appreciable role in these types of monitoring systems.

*1709: Section 4.2.1 reads well, but I ask myself what can a non-VIC user learn from this model specific sensitivity analysis. Some general recommendations for readers willing to explore such approaches would be welcome.*

Even though each land surface model (LSM) has its own characteristic model structure, many times they share parameterizations; this leads to, at times, strong similarities between LSMs. For example, the Jarvis model of canopy resistance is widely used in many LSMs. This scheme relies on the minimum stomatal resistance parameter. Given the substantial role that this parameter has in this study, it is straightforward to hypothesize that minimum stomatal resistance will also play an important role in other LSMs. Indeed, this appears to be the case as noted by a sensitivity analysis of the Noah land surface model [*Rosero et al.*, 2010]. Furthermore, although other models do not use VIC's characteristic variable infiltration curve or its baseflow parameterization, they do tend to use similar simple parameterizations. They share the weakness with VIC in that they rarely

account for local characteristics to either define the parameter prior distribution or to define the model structure of runoff generating processes. This is discussed in sections 5.2 and 5.3. The revised manuscript ensures that the connection between the results in section 4.2.1 and these discussion sections is more apparent.

*1711: Again, here you elaborate on parameters and hydrological extremes in a very VIC-focussed perspective. Does this somewhat deviate from the perceptual model implemented in the VIC algorithms? Do you learn something here that you could later implement to reduce uncertainty? In this respect: Currently a paper by Pechlivanidis and Arheimer (2015, HESSD) is being also discussed for publication in HESS. They also look at large scale hydrological modelling and try to implement the PUB recommendations. Could your approach also being adapted to implement the PUB recommendations?*

This section aims to provide insight into the drivers of remaining uncertainty after applying the monthly and annual runoff constraints. The spearman correlation helps us understand the parameters that are driving the spread in the flow duration curves in Figure 5 (now Figure 6). The spatial differences in these parameters are driven by the unique characteristics of each region. For example, when there is a distinct seasonal cycle to precipitation, the partitioning between baseflow and surface runoff plays an important role – this does not seem to be the case in areas without that distinct seasonal cycle. These results coincide with our understanding of the model structure of the VIC model. The challenge here is that not using local environmental characteristics to provide improved local prior parameter distributions and a lack of daily observations leads to a large spread in the daily flow duration curves. As discussed in section 5.2 and 5.3, providing improved prior distributions and defining a model structure that can use the available data provides a path to begin to resolve these challenges The revised manuscript ensures the connection between the discussion section and these results is more apparent.

A key focus of the PUB initiative is parameter regionalization, which Pechlivanidis and Arheimer (2015) achieve by clustering sub-basins on the basis of land surface and climate characteristics. The present study suggests the difficulty of properly constraining model parameters even for locations in which runoff observations are available, so parameter regionalization will require some additional effort. However, the PUB recommendations for large scale modeling proposed by Pechlivanidis and Arheimer (2015)—in particular, the classification of sub-basins using spatial characteristics—are a focus of the first author's ongoing work.

*1714-1715: How long do you think the FAO Map would be still the state of the art?*

Although the FAO Soil Map of the World is still widely used in global land surface modeling, there are upcoming alternatives. One especially exciting dataset is the GlobalSoilMap data product that will provide soil properties over the globe at a 100 meter spatial resolution [*Arrouays et al.*, 2014]. This data will provide a

breakthrough in the representation of soil properties in global land surface and hydrologic modeling.

*Artwork: I know that it is difficult to create adequate visualization of global data, but my eyes are really struggling when inspecting Figure 1,3 and 5. I would welcome a supplement with high-resolution versions of these Figures.*

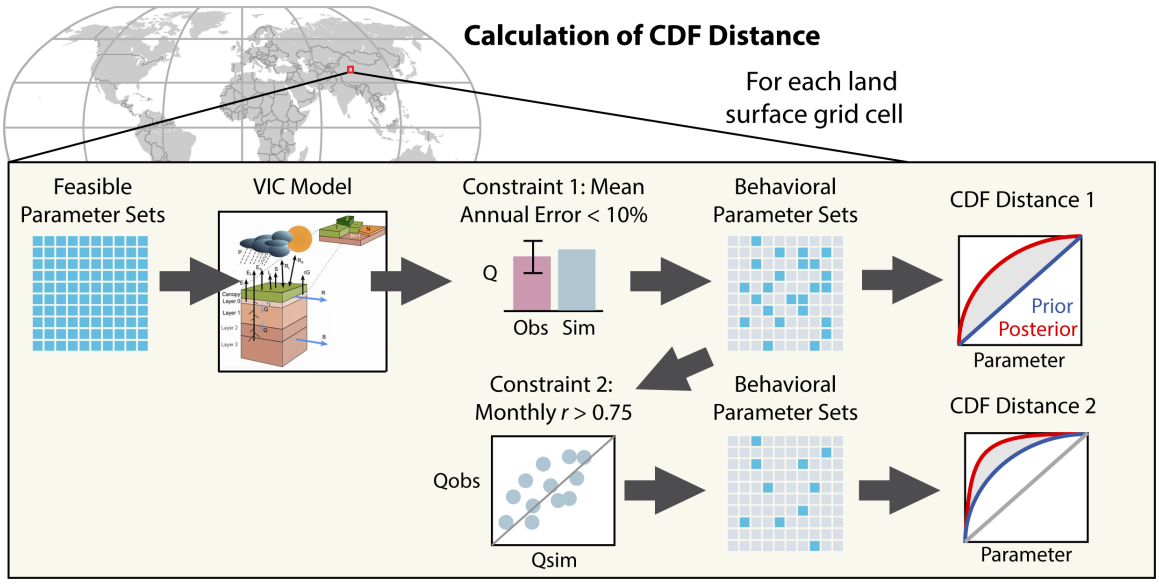
Agreed, we will request for these to appear as full-page figures in vector format (zoomable) during proofing.

*1703 – 1-9: Concerning the Köppen-Geiger climate you might find some ideas in Teuling (2011). Just something that came to my mind reading the paragraph, no need to implement anything to reply here.*

We thank the reviewer for this recommendation. The shift from a discrete climate classification to a continuous one would be an interesting way to more properly understand how the change in behavioral parameter sets changes in a spatially continuous way. However, this work aims to allow the reader to quickly distinguish the differences in model parameter and structure uncertainty across discrete climate types. By using a simple discrete classification, we ensure that the abrupt changes are clearly identified. As one of the co-authors is exploring in another project using these types of continuous data will be helpful as we attempt regionalize the behavioral parameter sets.

*1705: I find the described approach quite interesting. I wonder if I am the only person that would welcome here a graphic rendering of the steps involved here (flowchart with boxes, arrows and references).*

We thank the reviewer for this suggestion. The following graphic has been added in the updated manuscript (now Figure 1) to visualize the different steps involved in creating and constraining the ensemble.




*1707 – 22-23: "However, the most prominent feature is the lack of runoff observations (grey areas)". Well, this finding could have been made also a priori and so further reduce the number of grid cells to be computed (or increase resolution to 0.5 degrees).*

We thank the reviewer for this important insight. For the purposes of this study, we agree that solely running the model on the grid cells that have observations would have been sufficient. However, another goal of this project was to make the 10,000 member ensemble simulation available to the greater scientific community for future additional analysis with other observation datasets that might cover some of the regions without observations in the GRDC runoff climatology database. For this reason, we wished to run the model over all land grid cells.

*1708- 1-9: You describe Figure 2 (top panel) as follows: "Tropical and dry climates see the largest decrease in behavioral parameter sets while continental, polar, and temperate regions experience the least." I understand what you mean, but I find the formulation could be improved. When the number of "behavioral parameter sets" decreases, than what is the maximum number of "behavioral parameter sets"? I think here you should use a more straight formulation and just say: "According to the evaluated criteria Tropical and dry climates see smallest portion of behavioral parameter sets while continental, polar, and temperate regions experience the highest number." Please also consider to switch the two sentences "In this case, the number of acceptable parameter sets over arid regions is significantly smaller than other climates." and "This is especially true over the North American mountain west, the Sahel, and most of Australia."*

Thank you for the suggestion. We have revised this paragraph as follows:

> Figure 2 further summarizes these results as a function of climate classification. Although most of the regions with observations meet the annual constraints (10 and 20 percent relative error), there are distinct differences between

climates. Tropical and dry climates see the smallest proportion of behavioral parameter sets while continental, polar, and temperate regions experience the largest. The number of behavioral parameter sets decreases even further for all climate types when applying the monthly constraint (Pearson correlation between the simulated and observed normalized monthly climatology). In the case of arid regions, the number of acceptable parameter sets is significantly smaller, especially for the North American mountain west, the Sahel, and most of Australia.

*This is sound, well written manuscript. Very compact and with the right balance of pictures and tables. I experienced all along the manuscript that the authors assume that all readers are perfectly familiar with VIC. I can recommend publication after minor revisions.*

We would again like to thank both reviewers for their time and helpful comments.

## References


Arrouays, D. et al. (2014), Chapter Three - GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties, *Adv. Agron.*, *125*, 93–134.

Baroni, G., and S. Tarantola. 2014. "A General Probabilistic Framework for Uncertainty and Global Sensitivity Analysis of Deterministic Models: A Hydrological Case Study." *Environmental Modelling & Software* 51 (January): 26–34. doi:10.1016/j.envsoft.2013.09.022.

Fenwick, D., C. Scheidt, and J. Caers. 2014. "Quantifying Asymmetric Parameter Interactions in Sensitivity Analysis: Application to Reservoir Modeling." *Mathematical Geosciences* 46 (4): 493–511. doi:10.1007/s11004-014-9530-5.

Gong, W., H.V. Gupta, D. Yang, K. Sricharan, and A.O. Hero. 2013. "Estimating Epistemic and Aleatory Uncertainties during Hydrologic Modeling: An Information Theoretic Approach." *Water Resources Research* 49 (4): 2253–73. doi:10.1002/wrcr.20161.

Herman, J. D., P. M. Reed, and T. Wagener. 2013. "Time-Varying Sensitivity Analysis Clarifies the Effects of Watershed Model Formulation on Model Behavior." *Water Resources Research* 49 (3): 1400–1414. doi:10.1002/wrcr.20124.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

Pechlivanidis, I. G., and B. Arheimer. 2015. "Large-Scale Hydrological Modelling by Using Modified PUB Recommendations: The India-HYPE Case." *Hydrol. Earth Syst. Sci. Discuss.* 12 (3): 2885–2944. doi:10.5194/hessd-12-2885-2015.

Troy, T.J., E.F. Wood, and J. Sheffield. 2008. "An Efficient Calibration Method for Continental-Scale Land Surface Modeling." *Water Resources Research* 44 (9): W09411. doi:10.1029/2007WR006513.

Rosero, E., Z. Yang, T. Wagener, L. E. Gulden, S. Yatheendradas, and G. Niu (2010), Quantifying parameter sensitivity, interaction, and transferability in hydrologically enhanced versions of the Noah land surface model over transition zones during the warm season, *J. Geophys. Res. Atmos.*, *115*(D3), D03106.

Sheffield, J., G. Goteti, and E. F. Wood (2006), Development of a 50-Year High-Resolution Global Dataset of Meteorological Forcings for Land Surface Modeling, *J. Clim.*, *19*, 3088–3111.

Sheffield, J. et al. (2014), A Drought Monitoring and Forecasting System for Sub-Saharn African Water Resources and Food Security, *Bull. Amer. Meteor. Soc.*, *95*(6), 861–882.