Dear reviewer,

Thank you very much for your constructive comments which help us to clarify the contribution of our study and describe the knowledge gained from this study more precisely. All questions being addressed will be answered in details. If you feel it is still insufficient, please do not hesitate to contact with us. We would like to make further explanations and revisions.

From the authors

Two watersheds is not enough to conclude your study provides general conclusions! There are groups that use thousands of watersheds, look up large sample hydrology, for example: http://meetingorganizer.copernicus.org/EGU2015/session/18271 http://www.hydrol-earth-syst-sci.net/18/463/2014/hess-18-463-2014.html

Responses:

The researches mentioned by the reviewer are expected to get general knowledge in many situations by applying hydrological model in large number of basins. The objective of our study is to demonstrate short periods of observations data have the possibility to be useful for physically-based hydrological model calibration in data-sparse basin. The paper could inspire more researchers to think about using such dataset to calibrate distributed hydrological models in basins lacking of streamflow data. In the past, this approach didn't draw much attention for solving the calibration problem of distributed model in ungauged basins. We focus the potential of this method, rather than the general applicability. Based on the objective of this study, two basins with very different climatic and geophysical conditions were selected to conduct our study.

The contribution of the paper is not clear. Such analysis on the quality of calibration data dates back to 1996 (http://www.sciencedirect.com/science/article/pii/0022169495029184) and republication in HESS is not justified!

Response:

The model used in Yapo *et al.*(1996) is a conceptual rainfall-runoff model. In comparison, our study focuses on *physically-based distributed hydrological models*. This is one major difference between our study and Yapo *et al.* (1996). These two types of models face the same problem: lacking of streamflow data for calibrations in ungauged basins. For conceptual rainfall-runoff models, effectiveness of using short period of streamflow data to calibrate the model has been demonstrated in several papers, such as the one Yapo et al.(1996) mentioned by the reviewer, also Perrin et al.(2006), Beven and Tada (2012). However, based on our knowledge, such approach has not been widely tested for physically-based distributed hydrological models. The contribution of this paper is to demonstrate that, using short period of streamflow data also have the possibility to be useful for calibrating physically-based distributed hydrological models, which are usually preferred, because of their better description of the spatial heterogeneity and details of the water cycle at the basin scale.

There is no figure provided of how calibration of SWAT with limited data translates into model simulation! How do I know 1 month of data is enough for calibration if I don't see how the

model works graphically? NSE is certainly not enough!

Response:

Thank you for this suggestion. For the Jinjiang Basin, a figure showing the simulated hydrograph corresponding to the calibration using data of one month will be added to the revised manuscript to demonstrate the effectiveness of using such short period of data for calibration. The figure is also shown here (Figure R3 below). For the validation period (2008 to 2009), it is shown that the best simulations of ensemble predictions corresponding to the calibrations using three year data (2005 to 2007) and one month data (July 2006) is quite similar visually. As a response to the concern about NES, we computed the Mean Absolute Error (MAE) of best simulations in low flow period (September to next March) and high flow period (April to September). In high flow period, the MAE of the best simulation corresponding to calibration using three-year data and one-month-data is 66.3 m^3/s and 63.3m^3/s, respectively. In low flow period, the MAE of the two best simulations is 36.4m^3/s and 43.1m^3/s, respectively. Generally, similar performance level is achieved by the two best simulations. These results could support our conclusion that in the Jinjiang Basin, it is possible that one month data is informative to calibrate the SWAT model effectively.

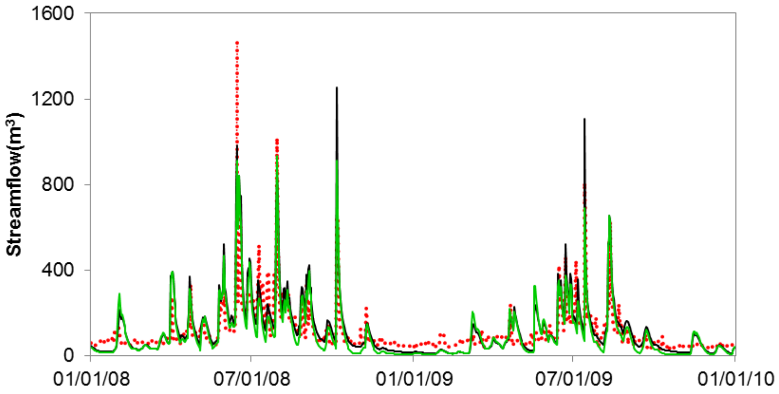

Figure R3 Observed streamflow of 2008 to 2009(dashed red line), best simulations of ensemble prediction for the validation period (2008-2009) corresponding to calibration based on streamflow data of 2005 to 2007 (solid black line) and July 2006(solid green line) in the Jinjiang Basin.

Page 2, lines 1-5: I don't agree with your statement that models like SWAT are able to predict droughts and floods! Droughts and floods respond to climatic forcings and climatic models are used to forecast them, certainly not SWAT!

Response:

As mentioned by the reviewer, hydrological models, like SWAT, use climatic data as forcing data to predict streamflow in the river. When climatic forecasting are available, hydrological models can employ it as input and predict hydrological drought and flood, from the perspective of quantity of streamflow.

Page 2, line 8-9: "Most parameters of hydrological models are conceptual without explicit physical meaning, which makes it necessary to identify parameter values through model calibration based on streamflow data". This refers to conceptual models mostly. Physically

based distributed are supposed to have parameters with clear physical meaning, that can ideally be measured in the field.

Response:

Although values of parameters with explicit physical meaning can be measured, the scale of measurement and model simulation is different, which makes it difficult to apply measured values to hydrological models directly. Also, such measurements require intensive field survey, which are not available in most researches. Therefore, usually model parameters are obtained from model calibration based on streamflow data. Such understanding can be found in published literatures (e.g., Gupta et al. 2005).

"Many recent works have focused on using in situ or remote sensing observations of hydrological processes other than streamflow for model calibration, e.g., soil moisture (e.g., Silvestro et al., 2015; Vrugt 20 et al., 2002), evapotranspiration (Vervoort et al., 2014; Winsemius et al., 2008), groundwater level(e.g., Khu et al.,2008)." My understanding is that since streamflow measurements are not available, one can alternatively use other variables such as soil moisture, ground water table and evapotranspiration as calibration data. This is certainly not the case, since measuring these variables is much more difficult and costly than streamflow. I suggest you phrase your sentences more carefully to avoid such confusions.

Response:

Our intention is not to make new observations of other variables of hydrological cycles, but to make best use of available data of such variables. There are cases that in some basins, streamflow data is unavailable, but measurements of the other variables are available. In such situation, the available data maybe valuable for model calibration. To avoid confusions, these sentences will be revised based on above understanding.

Page 3, line 4: What do you mean by "changing environment"?

Response:

This term of "changing environment" comes from the paper of Montanari et al. (2013), which introduces the IAHS Scientific Decade 2013–2022 " Panta Rhei—Everything Flows". It means the all factors that influence hydrological system, including both changes in nature (e.g., climate changes) and society (e.g., global population growth).

Page 3, lines 4-6: You argue "For hydrological simulations or predictions in changing environments, physically-based distributed hydrological models are usually preferred, because of their better description of the spatial heterogeneity and details of the water cycle at the basin scale (Finger et al, 2012; Wu and Liu, 2012)." I understand that some physical modelers would make such arguments, but it is certainly a debated issue, so I wouldn't make such strong claims. This being said, in a changing climate, even physically based models are not proven to be working properly. The argument made by developers of physically based models is that since they use specific description of the watersheds, their models can handle land-use change (change of physical characteristics of watersheds). This also needs a lot of research still.

Response:

We agree with the reviewer's opinion about the comparison between conceptual and physically based hydrological model. The sentences will be revised in the new version of manuscript to express our understanding more precisely.

Equation 2 is all WRONG! You want to use an objective function of NSE, do, but you can't call it a likelihood function and use it as in the Bayes theorem! There is no scaling in Bayes law! You may call this weight, but not posterior likelihood.

Response:

For GLUE, the term of likelihood is used in a very general sense, as described by developer of GLUE, Keith Beven, in the paper of Beven and Binley (1992): the likelihood function in GLUE works as a fuzzy, belief, or possibilistic measure of how well the model conforms to the observed behavior of the system, and not in the restricted sense of maximum likelihood theory. The likelihood measure quantifies the difference between simulation and observations. The only requirement for a likelihood measure is that it should be assigned as zero for all parameter sets that cannot reproduce the observations and should increase monotonically as the performance rises. Based on the theory of GLUE, in our opinion, using NSE as a likelihood function is proper in our study. Also the Bayes equation is employed in GLUE in a general sense. Equation 2 has been used in many studies related to GLUE, such as Freer and Beven (1996), Beven and Freer (2001).

I have a hard time with equation 3 also! Weights (or as you call them posterior likelihoods) are calculate based on overall performance of the model (t=1:N), but are used at each time step to estimate the cumulative probability of streaamflow. This is not right! You want to estimate the 95% uncertainty range, take the 2.5 and 97.5th percentiles of your streamflow simulation ensemble.

Response:

For application of GLUE, the posterior likelihood is computed based on the overall model performance in the period when observations are available to compare model behavior with observations. Then at each time step, the posterior likelihood is used to estimate cumulative probability of streamflow. This is the way how GLUE estimates simulation uncertainty, and equation 3 can be found in many studies using GLUE. In our opinion, this equation is correct.

Page 6, lines 4-6: I don't understand this sentence, and it is critical to evaluate the methodology proposed in this paper: "As a first trial for the distributed model, we sought to explore the highest possible performance using certain short lengths of records, not the general performance of specific lengths."

Response:

As a first trial for distributed model calibration using limited observations, our objective is to show the possibility of using short period of data for calibration, not to identify the exact minimum requirement for the length of calibration data. Therefore, what we concern is the best performance when using certain short length period of records, not the general performance when using different datasets of certain short length period. This explanation will be added to the revised version of paper.

The entire section 2.4 is poorly written and methodology is poorly described, making it really difficult to assess the paper.

Response:

This section will be revised in the new manuscript, to make it easier to be understood as follows:

For the two basins, firstly we carried out benchmark calibration using three year daily observations for model calibration: For the Jinjiang Basin, the calibration period is 2005-2007 and the validation period is 2008 and 2009. For the Heihe Basin, the calibration period is 2003-2005 and the validation period of is 2006-2008. Secondly, in order to test whether using short period of observations could calibrate the model effectively (i.e., achieve similar performance as benchmark calibration), subsets of the data used for model calibration in the two benchmark calibrations will be selected and used for model calibration. Then, the results of these calibrations (i.e., performance of the calibrated model) will be compared with the benchmark calibration of each basin. If the performance is similar to that of benchmark calibration, it will be conclude that it is possible that data of that specific short period is as informative as three year observations (i.e., data used in benchmark calibration) for parameter calibration and then lead to the conclusion that for calibrating physically based distributed hydrological model in data-sparse basin, resorting to calibration using short period of observations is a possible way.

For all calibrations of each basin, the calibration period is either three-year observations (benchmark calibration) or a subset of the three-year observations (calibrations using short periods of data). But the validation period of all calibrations are made same (2008-2009 for Jinjiang Basin; 2006-2008 for the Heihe Basin), to ease the comparison between benchmark calibration and calibration using short period of data. In such contexts, two issues are extremely important to achieve the goal of this study, the method for assessment of calibrated model performances, and the strategy of selecting short periods of data from streamflow observations used in benchmark calibrations. The details about the two issues are briefly introduced as follows:

The evaluation of each calibration was performed from the aspects of general performance and simulation uncertainty. The general performance was represented by the NSE of the best behavioral parameters set (i.e., the one with the highest likelihood value constrained by the calibration data). The simulation uncertainty is quantified by an index named as “U”, which combine the percentage of observations covered by the uncertainty band and the average width of the uncertainty band. The definition of U can be found in the original manuscript. The NSE and U are computed for the calibration and validation period for all model calibration. For the evaluation, we focus on values of NSE and U for validation period, as information in this period was not used for model calibration. The validation period for all calibrations in one specific basin is made to be same (2006-2008 for the Heihe Basin; 2008-2009 for the Jinjiang Basin.), for conducting the comparison among calibrations in an objective manner.

For extracting subset from the three-year data used in the benchmark calibrations, it was impossible to follow the studies of conceptual models that could conduct calibrations many times, due to the high demand of time for distributed model simulation. We have to perform the calibration in manageable times. In such situation, how to select subset from the three-year dataset is important. Three calibrations using 1-year data record that covered both the rainy and dry seasons, and five calibrations using 6-month data record that covered either a rainy season or a dry season were undertaken. Many studies of conceptual model showed that, generally, when the number of observations for model calibration is same, the data for high flow period are most informative for model calibration. Based on this general understanding, the 3-month, 1-month and 1-week period with highest average streamflow in the best performed 6-month dataset were selected as the representative dataset for the above mentioned temporal scale. Then these subsets were used for model calibration.

By such arrangement, the possibility of using 1-year, 6-month, 3-month, 1-month and 1-week dataset for model calibration were evaluated from the aspects of general performance and simulation uncertainty. Meanwhile, the total time needed for all model calibration is acceptable.

Page 7, line 2: You admit that it is very time consuming to calibrate SWAT using GLUE. Why not using more intelligent calibration approaches like Markov Chain Monte Carlo? It has been shown in the literature that MCMC is orders of magnitude more efficient than GLUE.

Response:

For daily-step model simulation of several years, it is common for lumped conceptual models to finish the simulation within one second. However, for distributed model like SWAT, usually, one such model run may need several minutes. We have no information about prior distribution of model parameters, based on which, parameter sets will be generated randomly. The uniform distribution is used as the prior distribution of each parameter. In such situation, implementing GLUE with Latin hypercube sampling is usually preferred. This strategy has been used in many literatures related to GLUE and SWAT simulation.

Page 7 line 11: "Kim and Kaluarachchi (2009) and Yapo et al. (1996) showed that data from high-flow periods are more informative than data from low-flow periods for model calibration, because most model modules are activated in high-flow periods." I tend to disagree! It depends on your performance metrics, if you use NSE which is sensitive to peak flows, then yes you are right! But if you use metrics such as baseflow index this is not going to hold. Different processes of a model are activated under different forcings, so you can't simply ignore several processes and focus on the one (few) process(es) that are activated at the wet condition.

Response:

We agree with the reviewer that the statement of "because most model modules are activated in high-flow period" is improper and it will be deleted from the manuscript.

Page 7 line 24: Uncertainty bound not band! Correct throughout the manuscript.

Response:

Uncertainty bounds are two lines consisted of the 2.5% and 97.5% quantiles of simulated streamflow in each time step. The uncertainty band is the area bounded by these two lines in hydrograph. Both term of "uncertainty bound" and "uncertainty band" (e.g., Beven and Benley, 1992; Yang et al., 2008) are used when applying GLUE for uncertainty analysis.

Page 8, lines 2-3: "For the 1-year period, all three calibrations performed similarly to the benchmark calibration, and the dataset for 2006 even outperformed the benchmark" It is interesting and concerning that a shorter calibration period provides a higher performance. It requires explanation as to how it happened! You can't just leave it like that which might spuriously suggest smaller calibration period is sometimes even better! Here are my thoughts: 1. You are not using a consistent period to evaluate your model! 2. Your calibration approach did not converge to the right posterior distribution (as might happen with GLUE) 3. Your data includes some misinformation, meaning not only it doesn't provide any good information to constrain model parameters, but also it misguides the model! In the cases of 1 & 2, extra data

can only be redundant and cannot deteriorate the performance of the model!

Response:

The evaluation of each calibration is based on judging model performance in the validation period. The validation period is made to be same for all calibrations in each basin. For the Jinjiang Basin, the validation period is 2008 to 2009. For the Heihe Basin, the validation period is 2006 to 2008. Streamflow data are obtained from the water administrative department in local government and have used in many studies. The data quality is guaranteed. We apologize that the statement of "the dataset for 2006 even outperformed the benchmark" is misleading. It is only based on the NSE of best performed behavioral parameter set. Another important aspect of the evaluation is simulation uncertainty. It is quantified by the index of "U" as defined in the manuscript. As shown in the Table 4 of the original manuscript, it is indicated that the simulation uncertainty of calibration using data of 2006 is a little bit higher than benchmark calibration. So we agree that the statement is improper and it is not our intention to conclude that using shorter period of calibration is better than using long period of observations. The method presented in this study is only expected to be useful for model calibration in data-sparse basins where streamflow data of several years are unavailable. Therefore the statement of "the dataset for 2006 even outperformed the benchmark" will be deleted.

Page 8, lines 14-16: "The calibration using the 1-month dataset still achieved similar performance to benchmark calibration. Thus, it is indicated that in the Jinjiang Basin, it is possible to calibrate the SWAT model effectively using only 1-month's continuous daily observations of streamflow." This claim is rather strange to me! One month is enough to capture all the processes? Some processes might not even be activated in one month! Again, this is because you focused all your attention on NSE, and what is most important in NSE is the high peaks. So if you activate the processes that reproduce the high peaks, you get a good performance. This doesn't mean one month is enough to calibrate a model!

Response:

We apologize that we didn't give enough details for the evaluations of calibration using 1-month data in the manuscript. As our responses to previous comments, from simulated hydrograph and all model performance indexes (NSE, U, MAE for both low and high flow period) in the validation period, it is indicated that the calibrated model corresponding to 1-month calibration data performs similar to the benchmark calibration.

We realize that our expression about the results using 1-month data is too strong and confusing. In the revised manuscript, instead of the original expression, we will conclude that it is possible that 1-month's continuous daily observations can contain much of the information content of 3-year continuous streamflow data for model calibration.

References:

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Process., 6, 279-298, 1992.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249, 11-29, 2001.

Freer, J. and Beven, K., Bayesian estimation of uncertainty in runoff predication and the value of data: An application of the GLUE approach, Water Resour. Res. 32, 2161-2173. 1996.

Gupta, H. V. ,Beven, K. J. and Wagener, T.: Model Calibration and Uncertainty Estimation, In Encyclopedia of hydrological science, Anderson MG (eds), John Wiley & Sons, Ltd,2006.

Montanari, A., Young, G., Savenije, H.H.G., Hughes, D., Wagener, T., Ren, LL., Koutsoyiannis, D., Cudennec, C., Toth, E., Grimaldi, S., Blöschl, G., Sivapalan, M., Beven, K., Gupta, H., Hipsey, M., Schaefli, B., Arheimer, B., Boegh, E., Schymanski, S.J., Baldassarre, G.D., Yu, B., Hubert, P., Huang, Y., Schumann, A., Post, D.A., Srinivasan, V., Harman, C., Thompson, S., Rogger, M., Viglione, A., McMillan, H., Characklis, G., Pang, Z., Belyaev, V.: "PantaRhei—Everything Flows": Change in hydrology and society—The IAHS Scientific Decade 2013-2022. Hydrolog. Sci. J., 58(6),1256-1275, 2013.

Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., and Mathevet, T.: Impact of limited streamflow data on 30 the efficiency and the parameters of rainfall-runoff models, Hydrolog. Sci. J., 52, 131–151, 2007.

Tada, T. and Beven, K. J.: Hydrological model calibration using a short period of observations, Hydrol. Process., 26, 883-892, 2012.

Yang, J., Reichert, P., Abbaspour, K. C., Xia, J. and Yang, H.: Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China, J. Hydrol., 358, 1-23, 2008.

Yapo, P. O. and Gupta, H. V. and Sorooshian, S.: Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, J. Hydrol., 181, 23-48, 1996.