General Response

Thank Prof. Sadegh very much for your careful review and kindly comments. We take your comments into account seriously and will do the revision for our manuscript accordingly which include: First, we will add some reference papers recommend in your comments. Second, we will redesign our calibration experiment using multi measures of agreement (RMSE, RMSEln, NSEln and NSE) instead of the single value of NSE (see the response to comments of Prof. Zappa). Last but not least, we will check the writing carefully to correct some writing mistakes. Replies to each detail comment are as below:

1. **I would see this study as a step-wise calibration rather than diagnostic calibration. In diagnostic calibration (diagnostic model evaluation, I would prefer to use), as introduced by Gupta et al. (2008) signatures of the system (data) are used instead of an ad hoc residual based likelihood (model evaluation) function. In this study a NSE measure was used for step-wise calibration of each parameter which doesn't correspond to the original term of "diagnostic model evaluation". Also when the term diagnostic is used, reader would expect to see it points out some kind of model/data error, while this study doesn't pin point which part of the model needs correction/modification.**

Thanks for this comment. The calibration method proposed in this study is a step-wise calibration. However, each calibration step is based on a signature extracted from hydrograph. Through analyzing the spatial-temporal dynamic of temperature, precipitation data and snow/ice coverage, we extracted hydrological meaningful information from the available data and developed four signatures: groundwater baseflow hydrograph partition, snowmelt hydrograph partition, glacier melt hydrograph partition and rainfall direct runoff hydrograph partition. Each signature was related to independent model components. Parameters were estimated on the difference between the observed and simulated partitions. To quantify the difference, the measure of agreement was used as function (a similar procedure can be found by Hingray et al.(2010)). Our procedure works within the framework of diagnostic problem proposed by Gupta et al. (2008): Diagnostic evaluation consists of noting the behavioral (signature) similarities and differences between the system data D^{obs} *and the model simulations* D^{sim}*, and correction procedures by relating these to relevant model* *components. The proposed calibration method aims at the diagnostic evaluation problem, i.e. to determine those components of the model, which, when assumed to be functioning properly, will explain the discrepancy between the computed and observed system behavior (Reiter, 1987;Gupta et al., 2008). The model components in this study consists of groundwater baseflow, snow/glacier melt and rainfall direct runoff. To evaluate the model, the effects of each component on simulation runoff were separated by hydrograph partitioning. The degree of a realistic representation of each component achieved by the model was evaluated on each calibration step. The proposed method can used to diagnose model structure and this can be left for further study, but not included in the current study.*

2. **Introduction doesn't connect to the body of paper. In the introduction section authors present a literature review of diagnostic model evaluation studies using several indices (signatures) of the watersheds and in the current study they just use NSE!**

In this study, we extracted hydrological information from available data and partitioned the hydrograph pertaining to water source for runoff generation. Partitions were developed as signatures for model calibration and can be used for model component diagnostic (see the above reply to comment 1). And we will redesign the calibration experiment using different objective function to quantify the difference between the observed and simulated partitions in each step here. The proposed method aims at the model diagnostic problem in an alpine area, so a literature review of diagnostic model evaluation studies using several indices of watersheds was presented in the introduction section.

3. **Recently a formal statistical framework for diagnostic model evaluation is introduced in the literature. Authors can include the following papers (amongst all) to give readers a better overview of diagnostic model evaluation literature: Olden, J. D. and Poff, N. L. (2003), Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. River Res. Applic., 19: 101–121. doi: 10.1002/rra.700 Vrugt, J. A., and M. Sadegh (2013), Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, Water Resour. Res.,**

49, 4335–4345, doi:10.1002/wrcr.20354. Sadegh, M. and Vrugt, J. A.: Bridging the gap between GLUE and formal statistical approaches: approximate Bayesian computation, Hydrol. Earth Syst. Sci., 17, 4831-4850, doi:10.5194/hess-17-4831-2013, 2013. Several step-wise CRR model calibration papers also exist in the literature than should be referred to in the paper.

Thanks. We will add some of these references directly relevant in the Sect. 1.1.

4. **Fig. 2: In some months like April and June, temperature estimated from equation 1 (based on temperature lapse rate) nicely follow the fluctuations of observed temperature while in others like February and November it fails to simulate the temperature dynamics. How do you explain this phenomenon?**

The temperature lapse rate was estimated based on temperature series in THS station and two automatic weather stations (XT AWS and TG AWS) in the Tailan basin. While the validation of the lapse rates shown in Fig.2 was carried out on the BT AWS in Kumalak basin. The monthly relationship between temperature and elevation can be slight different in different basins in Northwest China according to Zhang et al. (2012). This can explain the different performance of the lapse rate in different months. Two points can be derived from Fig.2.: the trend of temperature varies with elevation can be captured using the estimated lapse rates, especially in the hot months (April to August) when snow/ice melt mainly occur; the monthly lapse rates performed better than the annual constant lapse rate.

5. **What explains the significant temperature laps rate difference in different months (-0.36 to -0.86)? In the most basic form, this lapse rate is a constant number for the whole year.**

The difference between monthly temperature lapse rates should be attributed to the seasonal variation of air flow and prevailing wind direction in the mountain areas, while it is still not very clear. However, many studies have pointed out the importance of considering the monthly varied temperature lapse rate in melt simulation in the alpine areas (Aizen et al., 2000; Zhang et al., 2012; Ji et al., 2013). According to Zhang et al. (2012), the mean monthly temperature lapse rate in northwest China can vary from -0.29 to -0.7 ℃/100m. The lapse *rates used in this study was estimated according to temperature data series gauged in weather stations in the basin. Fig.2. shows that monthly lapse rate perform better than a constant number for the whole year in capturing temperature spatial distribution.*

6. **Your objective function for estimating the lapse rate needs an "abs" function (absolute value), otherwise negative and positive residuals will cancel out. This might explain why we don't see a good fit to measured temperature in some months?**

Yes, thank you very much. There is a mistake of the objective function in Eqn. (2) which should be corrected as below:

$$Z \;=\; Min[\sum\left(T_i - (T_{oi} + T_p \cdot (H - h))\right)^2]$$

7. **Suggestion: Fig 5a and 5b can be presented as subplots in one plot.**

Thanks, we agree this, and the Figure will be modified (Figure1).

8. **Months names in Fig 5a legend are not in order! Is it just a typo?**

The month names will be reordered, thanks.

9. **I expected to see all the curves in Fig 5a-b continue to a common elevation (_5000), although might be horizontal at the end. Your study area does not change with month, just the melt area changes which can be represented by a horizontal line at higher elevations.**

Yes, the curves will be extended to higher than 5000m, where melt areas are constant.

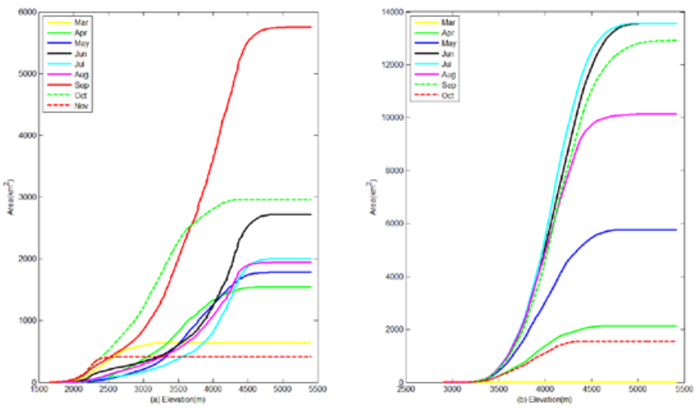

Figure1. (a) Cumulative monthly snowmelt area distribution by elevation for 2003 to 2012. (b) Cumulative monthly glacier melt area distribution by elevation for 2003 to 2012. The

snowmelt areas in December, January and February and the glacier melt areas in November, December, January and February are zero and are not shown in this figure.

10. You should evaluate your model as a complete package for the evaluation period and don't partition the hydrograph into several constituents. Eventually, your model parts should work as a whole. Also in principle you don't know what type of process generates your surface runoff in the evaluation period, so it doesn't make sense to partition your hydrograph.

We evaluated the model for the whole simulation period eventually. As parameters were identified based on different hydrograph partitions, we evaluated the simulation of each partition firstly. And finally, the complete simulation in the evaluation period was composed of different partition simulation. As described in Sect.2.1, the runoff concentration time in TRB is less than 1 day, we can then separate the runoff components that generate surface runoff in a step-wise way: in the SM period, both glacier melt and rainfall runoff don't occur, so surface runoff in this period is generated by snowmelt alone; in the SM+GM period, rainfall runoff doesn't occur, surface runoff is generated by snow and glacier melt; in the SM+GM+R period, surface runoff is generated by both snow/glacier melt and rainfall. We partitioned the hydrograph according to the runoff components, and estimated parameters through evaluating the simulation of each partition.

11. Page 1273, line 18-21: It is mentioned that results of this study is comparable to an automatic calibration method. If so, why do we need to partition the hydrograph? And what has been diagnosed in this study?

The automatic calibration method used here is a benchmark method which uses the single whole hydrograph as a measurement to evaluate all of the model parameters. The comparison between the proposed and automatic method is to demonstrate the difference calibration efficiency by using partitioned hydrograph and the whole hydrograph respectively. Although the results between the two methods are similar, the proposed calibration method has some good features: one is to reduce the parameter uncertainty during the calibration procedure. Another is to diagnose model components through parameter calibration. In the *proposed calibration, we firstly extract hydrological information from available data to separate runoff components in the hydrograph. Then we relate model parameter group to each hydrograph partition. Each parameter group was calibrated upon the corresponding runoff component separately, uncertainty from equifinality can be reduced in this way, while the automatic calibration method does not have this function. During the calibration procedure, we can diagnostically identify the components/processes which should be improved to account for the under/over estimation. In this way, the efficiency of the calibration procedure can be improved.*

12. Table 5: for the evaluation period, we see a better performance for automatic calibration rather than step-wise calibration! How do you explain this? And why would a researcher leave automatic calibration for step-wise calibration?

Results in Table 5 show that the partitioned calibration method has a better performance than the automatic method. In the revised manuscript, we will do a new comparison between the proposed and the automatic calibration method using a benchmark model and seasonal runoff simulations (see also reply to comments of Prof. Schaefli and Prof. Zappa). Thanks.

13. Page 1274, lines 1&2: "number of criteria handled by an automatic calibration procedure should be lower than 5 : 1"! Number of evaluation criteria is not important, the amount of information that they extract from data is more important.

Thanks. Information extracted from data is also one kind of criteria. To calibrate model parameters, hydrological meaning information can be used as criteria for measure of agreement as done in the proposed calibration method in this study. The most significant difference between the proposed and the automatic calibration method used here is that the amount of information that they extract from data is different. In the automatic method, we just used the single whole hydrograph to evaluate parameter groups, and no more information had been extracted. It is may weak at constraining more additional criteria such as simulation error of snow and glacier melt. Parameter uncertainty can be reduced by constraining multi-criteria, while, the automatic calibration method can usually not handle *more than five criteria. The comparison between automatic and the proposed method shows that the proposed method can expand the amount of information extracted from data and use to constrain parameters. The number of model parameter and evaluation criteria can be matched well in the proposed method, which is weak in the automatic method. When number of calibrated parameter is increased, the proposed method should perform much better than automatic method, which is however left for future study.*

14. Page 1274, lines 9-12: It is mentioned that automatic calibration methods are sensitive to the calibration data period while step-wise calibration is not. Different calibration periods provide different events and might affect step-wise calibration in the same way it affects automatic calibration. Actually it does affect step-wise calibration as well. In the cross validation step (same page lines 19-21) it is shown that the value of parameter B changes from 0.2 to 0.8 due to different calibration events.

The proposed calibration method should be sensitive to calibration data in some degree, but should be less than the automatic method. The cross validation results show that only the value of parameter B varies significantly due to much higher peak flow in the late evaluation period (2008-2012). Peak flows are mainly generated by storm-rainfall. Rainfall data series is one of the main factors that influence the simulation of peak flows. The significant variation of parameter B should be attributed to the error of rainfall input data. In the proposed step-wise calibration method, data time series are not used for calibration directly. Information of hydrological processes is extracted firstly and used to partition hydrograph. The calibration data in the proposed method is not simple discharge series, but hydrograph partitions which relates to the hydrological process physically. The relationship between parameter and corresponding hydrological process is distinguished, and each parameter is determined by the hydrology process it controls separately. The role of parameter on discharge simulation is separated in the proposed method, calibration data in this method is more hydrological meaningful than simple data time series usually used in the automatic methods.

15. Page 1275, line 11-12: "the low performance of the model for extreme summer storm events indicated the inadequacy of rainfall measurement". Cross validation shows that storm-runoff parameter (B) which controls the highflow to a high extent varies if the calibration period changes (0.2 to 0.8), so you can't simply attribute the poor model performance to the lack of rainfall measurement for the extreme summer events!

Yes, many factors should be responsible for the low performance for the high flow, including model input (precipitation is the most important), model parameter and model structure. Model structure could be the uncertainty source, which is, however, difficult to quantify. Also, as the THREW model has been applied to a dozen of watersheds with varied climatic/geographic characteristics (e.g., Tian et al., 2012), we have confidence to say that the corresponding model structure reflects the-state-of-the-art modeling approach. Model parameter sensitivity results show that the most sensitive parameter to high flow is B and WM (see Table 1. In the reply to comments of Prof. Schaefli), and the two parameters are calibrated on the high flow separately according to an optimized NSE value. The low performance can be contributed to the model input, i.e. insufficient of rainfall measurement.

16. Some mistakes in writing should be taken care of before publication including but not limited to: a. Page 1262, line 12: a similar procedure for temperature: : :!a similar procedure as temperature: : : b. Page 1262, line 26: downloaded from the website: : : ! downloaded from the NASA website c. Page 1263, line 10: was combined ! were combined d. Page 1265, line 8: annual mean ! inter-annual mean e. Page 1266, line 12: indexes ! indices f. Page 1271, line 15: years have ! years with g. Page 1274, line 9: calibration data ! calibration period

Thanks, we will correct these and some other mistakes.

Reference

Aizen, V., Aizen, E., Glazirin, G. and Loaiciga, H. A.: Simulation of daily runoff in Central Asian alpine watersheds, J. Hydrol., 238, 15-34, 2000.

Hingray, B., Schaefli, B., Mezghani, A. and Hamdi, Y.: Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments, Hydrolog. Sci. J., 55 (6): 1002-1016, 2010.

Ji, X. and Luo, Y.: The influence of precipitation and temperature input schemes on hydrological simulations of a snow and glacier melt dominated basin in Northwest China, Hydrology and Earth System Sciences Discussions, 10, 807-853, 2013.

Gupta, H. V., Wagener, T. and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrol. Process., 22, 3802-3813, 2008.

Reiter, R.: A theory of diagnosis from first principles. Artificial Intelligence ,32, 57–95, 1987.

Tian, F. Q., Li, H. Y., Sivapalan, M.: Model diagnostic analysis of seasonal switching of runoff generation mechanisms in the Blue River basin, Oklahoma. J. Hydrol, 418, 136-149, 2012.

Zhang, S., Gao, X., Ye, B., Zhang, X. and Hagemann, S.: A modified monthly degree-day model for evaluating glacier runoff changes in China. Part I: model development, Hydrol. Process.,26, 1686-1696, 2012.