Response to referee #1

We thank the anonymous referee for his or her comments and thoughts on our paper.
Below we reply to each of the comments. The original comments are quoted in *italics*.

Major comments

1.The interpolation approaches

In Section 2.2.3, the description of interpolation of station data using downscaled reanalysis data is quite confusing. Personally, I think it would be more appropriate to describe the approach as "Adjusting downscaled precipitation data with station observations", rather than "interpolation of station data using spatial fields from downscaled reanalysis data". This is because the scaling factor is applied to the daily downscaled reanalysis precipitation data to derive the final products (WRFadj-all and WRFadj-ind), while the station data is only used to derive the scaling factor. Also, there exists a technical issue with the approach: why not using the relationship between monthly station data and monthly downscaled reanalysis, since it is more reliable/stable than the relationship between daily station data and monthly reanalysis?

Reply: We now added equations to the description of the interpolation approaches in order to make the approaches more comprehensible. As the performance of the downscaled reanalysis data is much better on the monthly time step, only monthly values of the downscaled reanalysis data are used in this interpolation method. For the daily precipitation amounts the method relies on the observed precipitation data. We use a relation between daily station data and monthly downscaled reanalysis data instead of the relation between monthly station data and monthly downscaled reanalysis data, because we are interested in generating daily time series. By calculating the factor $F_i = \frac{P_i}{M_i}$ at the gauge locations

(with F_i : factor at gauge *i*; P_i : observed daily precipitation at gauge *i*; M_i : monthly values of the downscaled reanalysis data at gauge *i*),
interpolating this factor using IDW and multiplication with the monthly values of the downscaled reanalysis data, the downscaled reanalysis data only provide the spatial field, while the daily values are given by the gauge observations. Therefore we think "interpolation of station data using spatial fields from downscaled reanalysis data" describes the method in a more appropriate way than "adjusting downscaled precipitation data with station observations".

In addition, for the IDW approach, although it is commonly used, some brief description and references should be provided (e.g., how is the distance calculated? Is it based on x, y, z or x and y only?) This applies to the IDW approach mentioned in Sections 2.2.4 and 2.2.5.

Reply: A brief explanation of the IDW method and relevant references were added. As the IDW technique is also referred to in the sections "Interpolation of station data using spatial fields from downscaled reanalysis data" and "Interpolation of station data using monthly fields derived by multi-

linear regression" we decided to move the description of the IDW method to before the other two methods.

In Section 2.2.4, some explanations for the stepwise backward vs. forward MLR approaches (as well as relevant references) need to be provided.

Reply: We added more explanation for the stepwise regression approach and for the decision of forward vs. backward MLR. A reference was also inserted.

Personally, I think it would be more effective if equations are used to summarize the approaches presented 2.2.3, 2.2.4, and 2.2.5. 2.

Reply: We agree and inserted equations for the description of the interpolation methods.

2. The calibration experiments

My major concern on the experimental setup is that the different precipitation datasets are evaluated for four different calibration periods without any validation. As the paper also points out, calibration tends to have the parameters adjusted toward compensating for the errors from other different sources. As a result, a precipitation dataset that performs best in a calibration period may not be the best for an independent validation period. In other words, the approach by elevating the performance within calibration periods is essentially flawed. Hence, I would recommend that the authors use two of the four time periods for calibration (e.g., 1st and 3rd) and the other two for independent validation (e.g., 2nd and 4th), and evaluate the precipitation datasets based on their performance in the validation.

Reply: In this study, the purpose of the model calibration is to evaluate different precipitation estimates. This is somewhat different from the case where one tests a hydrological model. When evaluating a hydrological model, the model is after calibration typically also applied in a validation period in order to test the transferability of the model and its parameters to different time periods. For the evaluation of precipitation data it is however not necessarily required to assess the precipitation data in a validation instead of the calibration period.

For our study, there are two reasons why we decided to evaluate the precipitation data in the same period used for calibration. The precipitation estimates are evaluated using the value of a calibrated precipitation bias factor and the objective function value. The precipitation factor naturally always relates to the calibration period. Therefore it is useful to also evaluate the objective function value in the calibration period so that both criteria are evaluated over the same period. Second, evaluating the objective function values in a validation period would have the disadvantage that the objective function value would also include the effects of a possibly different bias of the precipitation estimate in the validation period. Thus one would not be able to differentiate easily between the performance with respect to the temporal dynamics and with respect to the overall over- or underestimation of precipitation. In contrast, using our approach, a change in the bias of the precipitation estimate is directly indicated by different values of the calibrated precipitation bias factors for the two periods and the performance with respect to the temporal dynamics can directly be inferred from the objective function value.

In order to demonstrate the ability of the WASA model to simulate runoff also in periods different from the calibration period, we exemplarily show some model validation results here (Fig. 1). The results refer to the model calibrated in the first period (1961-66) and validated in the second period (1967-72). The model calibrated in the first period performs also well in the second period, but the model performance is worse than for the model calibrated to the second period. The decline in model performance is larger for those cases where there was a strong change of the precipitation bias factor from the first to the second period (as for example for the precipitation data set WRF in most catchments), which results into a large streamflow bias in the second period. However, as explained above, for the purpose of this study to evaluate different precipitation data sets, their different performance can be seen from the differences in the bias factor for the two calibration periods. Thus, to focus on the main aspects, we decided to skip presenting the validation results in this paper.

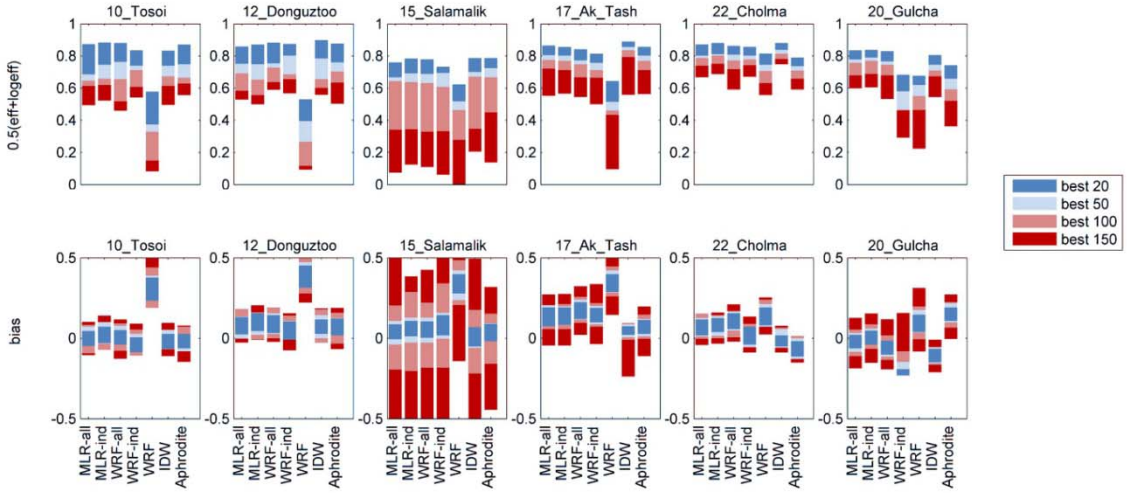

Fig. 1 Validation results for the model calibrated in the period 1961-66 and validated in the period 1967-72. Top row: 0.5*(efficiency + log-efficiency), bottom row: bias.

The results presented indicate that the precipitation bias factor has a dominant influence on the calibration process, making it less effective on constraining the other parameters. My recommendation would be to conduct a two-step calibration experiment by first calibrating all the relevant/important parameters (as it has already been done by the authors), and then follow up with a second calibration with the precipitation bias factor fixed at the value(s) from the first calibration, in order to more effectively constrain the other parameters.

Reply: Fixing the bias factor to the values from the first calibration, i.e. narrowing the parameter bounds of the bias factor in a second calibration step, would in our case probably only have a small effect on constraining the other model parameters. In contrast to a Monte-Carlo framework, where simulations with an unsuitable precipitation factor would be discarded, DDS runs which start with an unsuitable precipitation factor would in most cases evolve the precipitation factor to a suitable value. Fixing the bias factor to one particular value (instead of fixing it to the range from the first calibration) would not serve the objectives of this study, as we are interested in the uncertainty range of the precipitation bias factor. It would also probably not improve constraining the other model parameters very much, as there are no strong correlations of the precipitation bias factor to other parameters (as described in section 4.2.1). A larger effect would be expected if those

parameters which show strong correlations to other parameters and which can therefore hardly be constrained could be fixed to a particular value. However, this might have an effect on the simulation results under changed conditions. As the objectives of this study are not impeded by the fact that several model parameters are not well constrained, no further steps to possibly better constrain the model parameters were taken.

The dominance of the precipitation bias factor also makes it less meaningful to examine the sensitivity of the bias factor to inputs and parameters while it is being calibrated. A more sensible approach would be to assess the sensitivity of various inputs and parameters prior to calibration and identify the most sensitive/important parameters and inputs to be included in the calibration process. Including non-sensitive parameters in a calibration experiment may interfere with the constraining of other parameters, rendering the calibration process ineffective.

Reply: We see that the motivation behind our sensitivity analysis was not explained very well and therefore revised the description of the sensitivity analysis (see section 3.3.5). The objective was to investigate in what order of magnitude uncertainties in inputs would have an influence on the precipitation bias factor. The focus was particularly on inputs to the evapotranspiration module, as it is expected to have a strong influence on the water balance and thus also the precipitation bias factor. The sensitivity analysis therefore differs from the type of sensitivity analyses which are performed prior to model calibration and aim at identifying the most sensitive model parameters for model calibration. The variation factors introduced for this purpose are not intended as possible calibration factors. It would not be possible to identify a correction factor for radiation, wind speed, plant height etc. through model calibration, as these factors can partly compensate each other. Therefore the available data (from literature or output from the regional climate model) were applied as best estimate, but through the sensitivity analysis it was investigated what influence a variation of these inputs would have on the precipitation bias factor.

The paper only discusses the parameter distributions from one calibration case (3.2.1). Please be specific about the sub-catchment, precipitation dataset and calibration period for this case. What about the rest calibration cases? Are the parameter distribution patterns in those cases significantly different from (or very similar to) the presented case? And why?

Reply: The subcatchment, precipitation dataset and calibration period of the shown example is now also named in the text. Generally the parameter distributions of the other calibration cases show similar characteristics. The parameter ranges of the precipitation bias factor can be seen for all calibration cases in Fig. 10. Similar to the example shown in Fig. 8 the precipitation bias factor is well constrained in all calibration cases. If one considers the best 150 instead of the best 50 simulations the range the precipitation bias factor is constrained to gets much larger for all cases in the catchment Salamalik. However, as can be seen from Fig. 9, this would then also include simulations with low objective function values so that it would make little sense to consider these simulations for the evaluation of the parameter ranges. The behaviour of the other parameters in the shown example in Fig. 9 can also be seen as typical for the other calibration cases. For example the parameters k_sat factor, kf_corr factor and sat_area_var are never well constrained, while the groundwater and snowmelt parameters are usually better defined.

Finally, more explanation/clarification of the optimization algorithm is needed (Page 10735, last paragraph of Section 2.4.4). What does DDS-AU stand for? How are the short DDS runs used to assist the long run?

Reply: We added more explanation on the DDS-AU algorithm and particularly pointed to the fact that all DDS runs are independent from each other. The short optimisation runs with 33 to 77 model evaluations help to approximate the uncertainty bounds, while the long run with 3000 model evaluations is meant to come very close to the global optimum.

3. The presentation and overall structure

The overall presentation and structure of the paper need to be improved. For example, the paper loosely wrap up too many pieces of information into Section 2, including study domain, datasets, interpolation approaches, the evaluation approaches, the hydrologic model, the calibration algorithm, and the calibration runs etc. These need to be more tightly re-organized into smaller, more distinctive sections.

Reply: We see that due to the content of the paper the methods section is relatively extensive and therefore acknowledge the suggestion of the referee aiming at making the section "Methods and data" easier to understand. Re-organising the various subsections into own distinctive sections would however also have disadvantages. In order to allow for a quick orientation of the reader it was in our opinion preferable to keep the headings "Introduction", "Methods and data", "Results and discussion" and "Conclusions" on the first level. At the same level, we added the description of the "Study Area" which has been moved from the methods section in the revised version. The section "Methods and data" is now structured into three subsections, which we think is an acceptable number, and we also think that these remaining subsections group individual points in a comprehensible way.

The paper first discusses the point based evaluation in Section 2.3 and then presents the results in 3.1.1. For better organization, one could give an overall summary in an earlier section on the evaluation strategy and then discuss the evaluation details and results together in a later section. For example, Section 2.3 can be combined into 3.1.1, to make things easier to follow.

Reply: Structuring the manuscript in thematic paragraphs (e.g. "Point based evaluation of the precipitation data" and "Evaluation of the precipitation data based on simulated discharge") which would each have a method and a results/discussion section, or structuring the manuscript by first describing all methodological approaches and then describing and discussing the results both has advantages and disadvantages. In the first approach the results directly follow the description of the method, which is then still fresh in the readers mind. On the other hand, such a structure is less clear, as some parts of the methods are relevant for different thematic paragraphs so that they would require an additional paragraph. In our opinion, this would make it more difficult for the reader to quickly grasp the structure of the manuscript, particularly as most readers are very familiar with a structure which first describes all methods, followed by "results" and "discussion" or a combined "results and discussion" section.

Section 3.1.3 (Comparison to global gridded datasets) seems to be out of place and not making much contribution to the study. More importantly, the downscaled reanalysis precipitation datasets (WRFadj-all and WRFadj-ind), the focus of this study, are not included in the comparison. Hence, this section can be safely removed from the paper.

Reply: Global gridded data are often used for validation or bias correction of climate scenario data and could also be used for hydrological studies. However, from the estimated bias of the precipitation data set MLR-all and the comparison of the precipitation data set MLR-all to the global gridded data sets, we could infer that the global gridded data sets considerably underestimate precipitation for the Karadarya basin. This makes it an interesting additional point, which also underlines the necessity of studies like this one. MLR-all was used as a reference, as it had the lowest bias in all six subcatchments, which simplifies this comparison. Additionally, we now also refer to the precipitation fields from the downscaled reanalysis data and other interpolation methods applied in this study.

Minor comments:

P17036, L7: shouldn't it be Table 4?

Reply: Thanks, yes, Table 4 is correct.

P10738, L21: what does "this method" refer to? Please be more specific.

Reply: The text now refers to the abbreviations of the methods; this has also been changed in the following sentence.

Please consider increasing the font size of texts in the plots for better readability in Figures 5, 6, 7, 10, 11, 12.

Reply: Has been done.

Figure 6: what is 'WRFdir0?

Reply: Meant was the WRF simulation without adjusting to station data. This has been changed to 'WRF' as in the other figures.