Answers to reviewer 01 comments

We would like to thank reviewer 1 for the valuable and constructive comments that helped to improve the manuscript. We accepted most of the minor grammatical suggestions and are addressing (red) all reviewer comments (black) here.

Comments to the RC1 supplement.

p. 3, line 7

"... as mentioned for other studies below, Elshamy et al. (2009) used a distribution mapping approach to downscale and bias correct simultaneously..."

The sentence was modified accordingly: Elshamy et al. (2009) used a distribution mapping approach to simultaneously downscale and bias-correct 17 CMIP3 GCMs (SRES A1B) and applied the corrected climate data to run the Nile Forecasting System in the UBN.

p. 4, line 5

It seems there are discrepancies for the area as Elshamy et al (2009) mention 185,000 km2 for the same point while Mengistu and Sorteberg (2012) used 174,000 km2, other sources may have other different figures. Please comment on the impact of such uncertainty on the results.

We are not able to quantify the impact of different assumptions of the UBN catchment areas but included following sentence: "Elshamy et al (2009) estimates a catchment area of 185,000 km^2 and Mengistu and Sorteberg (2012) an area of 174,000 km^2 for the UBN. These discrepancies are certainly based on different digital elevation models and GIS algorithms used to delineate the catchment area and thus may add to the uncertainties of such studies, which are not easily quantifiable."

Moreover, the catchment area we used in this study was actually 172,000 km^2 (almost similar to Mengistu and Sorteberg (2012)) not 166,000 km^2 as stated in the manuscript. The numbers were changed accordingly.

p.4, line 24

I do not see any discussion of how the model simulates evapotranspiration which typically represents 70-80% of the balance.

The simulated average annual evapotranspiration corresponds to about 73% of the balance.

Following sentence was included in the model description section: "*Actual evapotranspiration is determined by simulated soil evaporation and transpiration from the vegetation cover.*" Due to the focus of the manuscript, we are not providing a detailed description of the SWIM model. For this, the user is referred to the publications cited in this section.

Moreover, to better address the evapotranspiration issue, a figure was included in Section 4.3 showing ET projections similar to those figures showing precipitation and temperature projections of the selected ensemble. Following paragraph was added here:

"Although surface air temperature increases already until 2050 in both scenarios by up to 2.2 K, actual evapotranspiration remains rather stable on the level of the reference period. Only in the second half of the 21st century the projected values increase by up to 50mm per annum. Hence, it can be concluded that actual evapotranspiration is already at its maximum and can only increase if water availability increases, too, as is the case after 2050."

p. 5, line 4

does is this mean the water balance is not closed. I understand there could be no further interaction between deep groundwater and shallow groundwater or river discharge but lost from the system means the balance has been closed.

It is correct that there is no further interaction between deep and shallow ground water once the water percolates from the shallow into the deep ground water aquifer. If these amounts would not be considered in the calculation of the water balance, the water balance would not be closed. However, by considering these losses, it is closed. Average percolation into the deep aquifer is 7% in our simulations. To account for this missing information, following sentences were included into Section 4.1 Model calibration and validation:
"The simulated amount of water percolating into the deep aquifer is about 7% in average. Without this recharge component, it was not possible to achieve good simulations during the dry period."

p. 5, line 6
This statement needs rephrasing noting that the weir was constructed in 1996.
Done.

p. 5, line 10
was that done for all simulations even where radiation was available? and if so, how did the derived values compare to those available?
Yes, for the sake of consistency, radiation was calculated in all simulations. Following sentence was included in Section 3.2: "*The simulated radiation data were calibrated to fit average annual observed radiation data of about 1800 kWh/m².*"
The data source is: http://solargis.info

p. 5, line 13
which variables were downscaled? and if statistical methods were used, please indicate the method briefly
We used only precipitation and temperature data as climate input. The downscaling and bias-correction methods applied are described in the respective sections 3.5.1 and 3.5.2.

p. 6, line 27
what is the impact having two different bias correction methods for ESMs and RCMs on the uncertainty of results?
First of all, the reason why two different bias-correction methods were used is that corrected data were readily available for ESMs (from ISI-MIP) but not for RCMs. We knew already from previous studies that the correction method applied to ESMs can have some deficiencies in different regions. Hence, we intended to apply an improved correction method to the RCMs. However, as shown in this study, also this method has its deficiencies, particularly in the extremes.
Concerning the impact on the results by using two different bias-correction methods, one should distinguish the different kinds of results here. Quantifying the impact of different correction methods on the climate change impacts in terms of a changed hydrology is not easy because the number of models in both ensembles is different (theoretically more outliers possible in the larger RCM ensemble) and the models show different performance in representing present-day climatology. However, the impacts or characteristics of the two bias-correction methods on simulated discharges are discussed in Section 4.3
Another definition of "results" that are provided in this study are the uncertainties regarding different climate inputs used in impact studies. We actually see a benefit in showing results of two slightly different correction methods because not only one but two examples are provided where lessons can be learn from, also with regard to uncertainties related. One may also conclude from this exercise that it might be useful or worth testing if different models would require different correction methods, depending on their characteristics of biases they show in the historical period.

p. 8, line 9
It is worth checking the work of Elshamy et al, (2013) in this regard. They used a GLUE-like methodology to exclude and weigh the models for assessment of impacts

Following sentence was included:
“A similar approach was used by Elshamy et al. (2013) who used a GLUE-like methodology to exclude and weigh climate model performance.”

p. 8, line 22
This metric is not commonly used - it is neither the coefficient of determination not the coefficient of efficiency (Nash-Sutcliffe). I wonder why the others did not use either instead of introducing a new metric whose performance is not well documented. The numerator measure the distance of the simulation and the mean of reference not the mean of simulation. At least some discussion on why this metric is used and how it is defined is necessary. Naming it R squared causes confusion as well.
Sorry, for the confusion. Indeed we used the coefficient of determination but a wrong equation was printed in the manuscript. We removed the equation, because R^2 is really commonly used, and we included an explanation where R^2 is mentioned the first time.

p. 9, line 10
I would think 30% is too large for the PBIAS especially for 30-year means - but after reaching the conclusions, I see the reasoning. Perhaps, it is worth hinting at this here. same comment applies to deviations in quantiles especially in the NED range. 30% deviations for extreme quantiles are fine.
We hint at the selection of threshold values with following sentences that were included in the subsection: Evaluating the suitability of climate simulations.
“Note that the definition of threshold values is somewhat subjective and was influenced by the simulation results of the model ensemble. However, if the thresholds would have been set more critically, almost no climate model would have passed the evaluation process successfully.”

p.9, line 21
so you really think these future changes are implausible? What about the far future?
We agree here in general that assuming “projected future changes of a certain magnitude in the near or far future being implausible or unlikely” is difficult. However, this is not what we intended to say, we are rather arguing that it was found that: Page 9, lines 20-21: “*...several simulations project enormous increase in annual river discharge already in the period 2030–2059. This was particularly the case in simulations where bias-correction resulted in stupendous increase of extreme daily rainfall and therefore extraordinary high peak discharges.*” The reason why the selection criterion (rate of change) was introduced, in addition to those analysing model performance based on the historical period, was to exclude those models from the selected ensemble that are difficult to be used in studies where a certain quantitative precision is required. This applies particularly to studies where future water availability for management, for instance irrigation and reservoir operations, are to be analysed.
However, in the discussion section we emphasize that by communicating climate change impacts, it is always necessary to show the full extend of model results, not only the result of models selected for a certain reason. Page 18, lines 13-14: “*However, model selection for regional impact studies is only a reasonable, justifiable, and recommended approach if the uncertainties of the selected ensemble are communicated within the context of the whole model ensemble.*”

p.9, line 27
would you please give more details of the calibration process: is it automatic, or manual, incremental or using the 3 gauges simultaneously? You mentioned NSE and PBIAS as your criteria, were these used in a multi-objective way, using weights, or seperately? Please clarify the calibration and validation periods. PBIAS for Lake Tana is large. Why did you calibrate to monthly flows and not daily flows?
SWIM was calibrated to the Blue Nile catchment using a semi-automated approach.
The model was calibrated to monthly discharge data because daily discharge data were not available. The periods for calibration and validation were on the one hand chosen according to data

availability and on the other hand to cover periods of wet and dry years.
The first step was a sensitivity analysis used to estimate reasonable parameter ranges. These ranges were used to set the boundary conditions (ranges and initial parameter values) for the automatic calibration algorithm PEST.
PBIAS for Lake Tana is rather large (23%) because observed discharges in the years 1973-1975 are not reliable, as is explained the last paragraph of Section 4.1.

Following sentences were added to Section 4.1:
"Due to limited data availability, the model was calibrated to the monthly time step using a semi-automated approach. The calibration (1981--1986) and validation (1987--1992) periods for gauge El Diem were on the one hand chosen according to data availability and on the other hand to cover periods of wet and dry years. Data availability for the gauges Lake Tana and Kessie were limited to the years 1969--1975 and 1976--1979, respectively. The gauges were successively calibrated where a parameter sensitivity analysis was performed in a first step to assess reasonable parameter ranges as boundary conditions for the automatic calibration algorithm PEST (Model-independent parameter estimation \& uncertainty analysis software). The objective functions to measure model performance are the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) and PBIAS, where NSE was the primary criterion."

p. 10, line 33
please add the same comment as notes under the tables for them to be self-complete
Done

p. 11, line 24
I guess using NSE would have been better for such reason, although the R^2 used is not the same as the coefficient of determination.
NSE is commonly used to measure the performance of monthly or daily time series but normally not for averaged time series (annual cycle) as are used here. The R^2 (coef. of determination) is more adequate for averaged daily values and was therefore used in this application. Moreover, R^2 was used in combination with PBIAS where R^2 determines the performance of seasonality and PBIAS the volumetric performance. The title of the respective subsection was changed from "Performance of daily discharge using UC and BC climate input" to "Performance of **average** daily discharge using UC and BC climate input".

p. 11, line 33
Visual inspection is recommended any way, but NSE would have been more indicative. and for low flows, NSE of log transformed discharge gives a good performance measure.
Same as previous comment. We think that NSE or NSE log criteria should not be applied to averaged time series.

p.12, line 2
The standard deviations in Taylor diagrams are normalised by the standard deviations in such a way that the max. normalised SD equals 1.

p. 14, line 13
I would say this should be 4.3, i.e. discussing temperature and precipitation before discharges. Additionally I do not see any analysis of the impact on evapotranspiration (actual or potential)
According to the suggestion, we restructured the document so that previous Section 4.5 (Temperature, precipitation, and evapotranspiration projections) becomes Section 4.3.
Moreover, Figure 9 (development of evapotranspiration) was added and described in the respective section.

p. 15, line 18
Therefore, this is a very critical assumption and should have additional explanation as why such changes are considered non plausible? Are there any physical phenomena to support this?
We agree that this is a critical assumption and the 30% threshold is based on a subjective estimate rather than on possible physical phenomena. The selection criterion was used to limit the number of climate models in the subset that represents climate projections to be used in quantitative and application-oriented water management studies. Reservoir management studies, for instance, would certainly have crucial difficulties to deal with extreme projections, although they might be reasonable. However, later on we discuss that serious impact studies should always take the full range of model ensembles into account.

p. 17, line 10
better be specific and state those specific purposes
specific purposes was replaced by “, particularly for quantitative impact studies,”

p. 17, line 11
I do not see this as a fact; there are methods to map different years of climate models to years of observations (e.g. based on proximity of precipitation totals or some other measures) so that discharge simulations correspond. Additionally, delta change methods allow such comparisons. It is common, however to analyze average hydrographs in climate impact studies.
It is true that alternative methods to those being applied in our study exist. However, our point is that it is not possible on a real-time daily or monthly basis as can be performed with observations where real-time discharge events correspond to real-time precipitation events. Therefore, we see the options to evaluate performance of climate model driven discharge simulations as limited compared to real periods of observations.
The methods applied in this study cover a large range of evaluation options considering various flow conditions (flow duration curves) and the analysis of daily hydrographs averaged over the 30-years baseline period.

p. 17, line 18
in fact, it has worsened.
This statement has been added to the respective sentence.

p. 17, line 19
I think this is repetition of what has been just mentioned.
The sentence has been deleted as it was indeed redundant.

p. 18, line 10
It is important to hint at this in section 3.7
The corresponding hint is given at the end of section 3.6.

p. 28
normalized by the Sdref?
Yes. Description added to Figure 5.