We thank reviewer 2 for his/her very valuable comments regarding the novelties and presentation issues of this study.

Page and line numbers indicated in this document refer to the latest version of the manuscript submitted, including improvements based on suggestions by both reviewers. Reviewer comments are formatted in *italics* and our responses in normal formatting. Sentences added to the manuscript are indicated in blue colour.

(1)
I wonder however, to what extent and what is the significantly new scientific contribution (for international scientific audience). This needs to be formulated in a more clear way. In its current form, the manuscript reads more as a combination of a case study/technical report assessing the potential climate impact on hydrology in selected basin and a commentary discussing the value of bias correction for improving future hydrological projections. In both cases it is not fully clear, what is the novel finding.

As formulated in the introduction [page 3, line 26], the objectives of our study are indeed twofold. Hence, as reviewer 2 is arguing, it is in fact a combination of a case study addressing climate change impacts in the Blue Nile River basin that additionally investigates bias-correction practices and their possible impacts on the result of such an impact study.

On the one hand a comprehensive impact study, as conducted here, is valuable solely by contributing to the portfolio of climate impacts knowledge in a specific region, particularly if the focus region is subject to rather high uncertainties with regard to climate change projections. What is the novelty in this respect? As explained in the Introduction, many studies use very simple approaches to assess the likely impacts of climate change on the hydrology in the Upper Blue Nile River basin. These approaches are simple either because they apply very simple modelling approaches and/or they use only one climate model. Not even a handful of studies can claim to be as comprehensive as our study in terms of the number of models and scenarios used.

On the other hand we shade some light on the issue of using uncorrected and bias-corrected climate input data and their impacts on the model results. We propose a method to evaluate climate model performance for regional hydrological impact studies using hydrological indicators (e.g. Taylor diagrams to easily identify outliers [page 12, line 27] or flow duration curves to characterise extreme and non-extreme discharges [Section 4.2.3]). Using hydrological indicators in this connection may not be a novelty as such (it has been applied for instance by Elshamy et al. 2013) but has, however, not yet been practised in many studies. The method consists of criteria and their thresholds used to select only those climate models that provide reasonable input. The proposed model selection method does also address the usability of climate simulations for different purposes, because it matters whether climate projection data are used to investigate qualitative or quantitative impacts/changes [page 3, line 20; page 9, .line 23]. Especially with regard to land and water management studies (e.g. reservoirs and irrigation), it is important that climate projections are quantitatively accurate to a certain degree. Studies addressing projected changes of floods and droughts would require a good representation of respective extremes. The selection method developed in this study was designed to support the decision of which models may be chosen to be members of the model ensemble finally used in the impact study to be conducted. Thereby, we evaluate model performance not only based on their performance in the reference period but we also consider the behaviour in future periods.

(2.1)
I would suggest to narrow the focus (main objective of the paper) to some novel contribution. For example, I found interesting the question to what extent can bias-correction alter the magnitude of change signals in hydrological simulations, however the results are given just in the supplement…
(2.2) ...and not discussed and upscaled/generalized to other regions.
(2.3) Generally starting presentation of results with figures/tables in supplement is formally not very attractive (it looks that such results are only supplementary to the paper objectives).

Concerning (2.1):
We agree that the question to what extent can bias-correction alter the magnitude of change signals in hydrological simulations deserves more attention. However, in the manuscript, we started the related discussion actually already in Section 4.2, where the impact of bias-correction on precipitation and discharge performance is intensively discussed on three pages. The impact on discharge projections is discussed in Section 4.4 and 4.6. Here, the reviewer is probably right and the Figures S6 and S7 should be part of the main paper instead of being placed in the supplement. Hence, we included these figures as Figure 10 and 11.
Moreover, we added following sentences to Section 4.4:
Looking at the change of average peak magnitudes between UC and BC ESM simulations in the reference and future period, the change signals are in a similar order, except for simulations based on IPSL. They are also in the order of average peaks simulated with WFD input, see Figure 3.
The projected changes of peak discharge magnitudes between UC and BC RCMs is significantly higher in BC simulations in 50% of the models. This is not surprising, because bias-correction of RCMs led to significant overestimation of high flows already in the reference period, as was discussed in Section 4.2.3. This behaviour is exaggerated in future periods.

However, as previously explained, the objectives of the paper are twofold. On the one hand we provide a climate impact study of the Blue Nile River basin and on the other hand the aim was to add information on how these results were achieved by investigating the uncertainties of using uncorrected and bias-corrected climate inputs, considering different climate model ensembles (GCMs and RCMs). Narrowing the focus to one novel contribution only is from our perspective not really possible without losing important details required for the messages we intend to give, especially with regard to impacts on the hydrology in the Blue Nile.

Moreover, by answering the last question (3) of reviewer 2 at the end of this document, we discuss, among other bias-correction-related issues, how we addressed the basic question in the paper that has been raised here (...altering the magnitude of change signals).

Concerning (2.2):
As we understand the criticism of reviewer 2, there is a lack of investigating the transferability of the method to other regions. Scientific methods are usually required to be generalizable, which is of course valuable in many contexts. However, in studies focusing on a specific region with its unique characteristics, the application of generalised methods that are not adapted to region-specific conditions, may result in loss of information. Hence, we provide a framework that assists the reader/user in choosing criteria to evaluate model performance for specific purposes, such as for qualitative or quantitative impact assessment studies.
In the "Discussion and conclusions" chapter we state that: "The authors of this study conclude that a purpose-driven selection of a climate model subset is a reasonable approach, **particularly in a regional context.** To identify ``good performing" models, the **selection process should include** an analysis of climate inputs, seasonal discharge patterns, volumetric deviations, daily characteristics (FDCs for extremes) and an assessment of the magnitude of projected future changes. It is also worth mentioning that the **thresholds defined to evaluate model performance have a subjective component** and are based on statistical parameters, graphical data interpretation, and modelling expertise. **In another river basin with different characteristics**, e.g. with a nival regime or a bimodal rainfall regime, the performance criteria and their thresholds may have been defined differently."
Following sentence was added to the Discussion section:
Hence, the model selection method can be applied to other river basins but it is always necessary to consider region-specific characteristics that may require the introduction of new criteria adapted to the situation at hand.

In Section 4.4, we added the following sentence to show that it is important to not only consider model performance in the historical period but also to account for model behaviour in projection periods.
Aich et al. (2014) applied the same five BC ESMs in four large African River basins and found that also in the Niger basin (comparable climate zone as the Blue Nile River) one of the five models projects extreme and unexplainable changes although it performed adequately in the historical period. In the case of the Niger River basin, it was the MIROC model that behaved awkward in the projection period whereas the IPSL behaved in the range of the other models. As in the study at hand, in the study of Aich et al. (2014), the IPSL model showed the same behaviour in the UBN.

Concerning (2.3):
The subjects investigated in this manuscript are very comprehensive using 30 climate model runs (15 uncorrected and 15 bias-corrected), 2 RCPs and 2 future time slices. To provide a meaningful analysis of climate change, bias-correction of climate models, and their impacts on the hydrology, it was necessary to condense the produced information to an understandable level without losing too many details. Unfortunately, it is not entirely clear to which figures and tables reviewer 2 is referring to. In case he/she meant the Figures S6 and S7, this issue would have been solved by introducing those in the new version of the main manuscript (see answer to comment 2.1 above).
In case the reviewer was referring to Figures S2 to S4 our answer would be the following: In the analysis of climate model performance, the main focus in this study is basically on hydrological performance indicators. That is why we start the description of precipitation characteristics and performance in Section 4.2.1 with figures and tables provided in the supplement. However, including these figures into the main part of the manuscript would mean to have 15 more figures, which is unreasonable.

(3)
In the debate, I would expect some more discussion whether the application of bias correction in climate change impact studies is generally a scientifically sound approach (per se). Is it meaningful to apply/analyse future projections, if the climate simulations have bias already in the reference/historical period? I did not find a solid/clear answer to this question in the manuscript.

In the three pages Discussion and conclusions we discuss on approximately one page the impacts of climate change and their impacts on discharge projections in the Upper Blue Nile River basin. The other two pages are devoted to discuss the issues of bias-correction and its consequences. From our perspective, the questions raised in the introduction are sufficiently answered. However, for some reasons this did not satisfy the expectations of reviewer 2. In the following we list here the questions that were central to the bias-correction topic again and give examples how these were tackled in the paper. Sentences added to the new manuscript version are indicated in blue colour.

To what extent can bias-correction alter the magnitudes of change signals in hydrological simulations in the study area?
The basic message that is given in the paper is that the bias-correction methods applied in this study improved the behaviour of average daily and monthly precipitation (and discharge) data considerably. However, bias-correction did also increase the variability of precipitation and discharge amounts resulting in under and overestimation of extremes. This has consequences for the applicability, particularly if changes of extreme values are the subject of investigation. Another finding was that the multi-model mean of BC simulations project always higher increases of discharges than the UC model ensemble.
The sentences and two figures in Section 4.4 included in the new version of the manuscript are explained in the section addressing comment 2.1 in this document. Following examples show how the question was addressed in the paper:

- Although bias-correction improved the performance of average climate conditions, **the range of monthly precipitation amounts increased critically in several models, producing some extreme outliers** in both ensembles. The same is true for daily precipitation characteristics.
- Average daily precipitation and the number of rainy days were considerably improved but **13 out of 15 BC models overestimate daily precipitation maxima and many of them significantly**.

- All **BC RCMs overestimate maximum daily precipitation**, many of them significantly.
- Almost all **BC simulations show higher SD than the UC simulations**
- ...the **bias-correction methods applied** to ESMs and RCMs in this study could be considered as **only partly successful.**
- Moreover, the multi-model means of **BC simulations** (both RCPs and periods) **always project higher increases in average annual discharges than the UC multi-model means.** Example: In the near future (2030-2059) in both RCPs, the range of UC models is between 7.4% and 19%, the range of BC models between 11.3% and 27.7%. In the far future (2070-2099) considering both RCPs, the range of UC models is between 7.5% and 21.6%, the range of BC models between 20.3% and 56.7%
- The HadGEM2 model is **the only model where bias-correction changed the sign of the discharge signal.** The simulation with UC climate input projects a decrease of average annual discharges of -2.9% and the BC simulation an increase of +2.2%.
- However, a hydrological impact study in the Danube River basin showed that relative changes of average monthly discharges projected by using UC and BC climate models are overall comparable (Stagl2015).
- ...**maximal discharge peaks** simulated with RCM climate input is **often much higher than average peaks simulated with WFD** (6000m^3/s).
- Looking at projected peaks in the period 2030-2059 (RCP8.5) shows that **nine out of ten BC RCM-driven** and **five UC RCM** simulations simulate **peaks that are higher than 7000m^3/s.**
- Another way to deal with low performance in the simulation of extremes in impact studiesis to analyse changes in return periods of extreme events only (Hattermann et al., 2016).

In how far can we trust simulations that required a strong correction?

- **bias-correction increased the range of monthly precipitation sums critically** in several models in both ensembles...**particularly if the deviation of monthly medians between UC simulation and WFD is rather large**
- ...the bias-correction **methods applied** to ESMs and RCMs **in this study** could be considered as **only partly successful**.
- …**bias-correction may help to overcome some quality issues but** it was also found in this study that **improving climate simulations in the reference period does not guarantee higher quality or reliability in simulating future periods.**
- **On the contrary, the greater the necessity to correct a particular model, the higher the risk that BC simulations will show unexpected behaviour in future periods**, where exceptions confirm the rule:
    - Interesting are the results of the NorESM1 model. The UC model simulates a bimodal rainfall and runoff system with a dry period during the rainy season in July to September. Although the model was forced by bias-correction into a completely different system, by pushing the dry season into a rainy season, the projections seem not anywhere near as disrupted as the IPSL simulation. Hence, the **NorESM1 results do not support the assumption that strong bias-correction necessarily results in unexpected behaviour in future periods**.
- It should be emphasised that the analysis of climate model performance in this study is only valid for the region of the UBN. It does not imply that a model which performed poorly in this study area is generally performing ``poorly'' in other regions, too.
- It is also worth mentioning that the thresholds defined to evaluate model performance have a subjective component and are based on statistical parameters, graphical data interpretation, and modelling expertise. **If the thresholds would have been set more critically in this study, almost no climate model would have passed the evaluation process successfully.** The rather weak thresholds were a compromise and reveal the fact that the **performance of many climate models is still far beyond being adequate for applied quantitative impact studies**. This statement includes bias-corrected simulations and implies that the ability of bias-correction can, depending on the approach, be rather limited and is thus not per se necessarily improving the reliability.
- **This study demonstrated that neither the trend-preserving method** applied to the five ESMs **nor the harmonic-based method** used to bias-correct the ten RCMs **was able to generate fully satisfactory climate inputs for a regional hydrological impact study with high demands in terms of accuracy.** Hence, further research is required to improve regional climate simulations

and/or to investigate alternative correction methods or approaches to make available climate simulations meaningful for application-oriented regional studies.

Are we using the right fuel?

- The value of using uncorrected climate simulations to answer those questions is, due to the lack of spatio-temporal accuracy and the lack of statistically representing observed weather characteristics, usually rather limited. **Bias-correction of climate simulations is an attempt to overcome at least some of these deficiencies.**
- **While achieving significant improvement in terms of average daily, monthly, and annual precipitation characteristics, increasing variability of precipitation amounts and therefore under and overestimation of extremes was the result in many simulations.** This phenomenon is problematic for impact studies and the application of hydrological models. Particularly if changes of extreme values are the subject of investigation.
- Unsurprisingly, discharge simulations show similar deficiencies as precipitation simulations.
- ...**with few exceptions, the performance of high and low flows was not improved, in fact has worsened in most of the simulations.** Many BC discharge simulations tend to exaggerate high (overestimation) and low flows (underestimation).
- Comparing **peak discharges** using UC and BC climate input, for instance, **showed a tremendous increase in some BC simulations** although average monthly precipitation patterns of BC models achieved a much better fit than their UC counterparts.
- **Large overestimation of precipitation on some days or in some months for instance, which are balanced by dry months in the long term, can lead to large amounts of excess water that may be simulated almost entirely as surface runoff by the hydrological model. Therefore, it is reasonable to use hydrological performance indicators to evaluate the suitability of climate simulations**, particularly for quantitative impact studies, and to create a subset of models for the impact assessment.
- Knowing these limitations, **one should carefully consider the model's suitability and the purpose it is being used for.**
- **An impact study focusing on relative changes of future water availability may have lower requirements in terms of model accuracy than a study with the aim to investigate future extremes,** such as floods and droughts or a study addressing land and water management issues including irrigation and/or reservoir operations.
- Whenever complex water management is involved, bias-correction is often unavoidable because reservoir volumes, specific thresholds, and irrigation capacities demand for more accurate hydrological input. However, to simply trust in climate input only because it was bias-corrected would be naive.