Thank you very much for your review. Your detailed comments will be taken into consideration to improve the paper.

Regarding the major comments:
*1. The main aims and goals of the paper are poorly stated. Developing a hydrological model for a particular region as stated as the main goal of the paper is not a 'Cutting edge case study'. Similarly, the key research contributions from the paper are not clearly highlighted within the conclusions. The authors need to think about the novel aspects of the paper and two-three key messages they want the reader to take away.*
*2. The title of the paper is currently misleading – I would remove 'with data uncertainty' as you do not consider uncertainties in stage data and the analysis of precipitation uncertainties is limited.*
Reply to 1-2): All three reviewers pointed out that the paper could benefit from clearer objectives and subsequently a more appropriate title. We agree with the reviewers that the paper needs improvement here. We have submitted our study as a "cutting edge case study". According to HESS "**Cutting-edge case studies** report on case studies that require (a) broadening the knowledge base in hydrology as well as (b) sharing the underlying data and models. These case studies should be cutting edge with respect to the quality and diversity of data provided the soundness of the models employed, and the importance of the study objective."

We present both 1) an important and high quality data set for the data-poor Mara River Basin after detailed analysis of the available rainfall, river stage and discharge measurements and 2) an innovation in rainfall-runoff modelling using river water level time series for model calibration in absence of reliable discharge data which is often encountered in African river basins. In our opinion the latter contributes to the knowledge base of hydrology, in particular rainfall-runoff modelling. In addition, we analysed the influence of rainfall data averaging in semi-arid basins where the rainfall typically has a high spatial and temporal variability.

The main goal was not to merely develop a hydrological model, but to develop a modelling methodology which can help increasing the hydrological understanding in this poorly gauged semi-arid region using water level time series for calibration instead of discharge since the rating curve was of very poor quality. Hence, the challenge was to assess the water availability despite the poor data quality. In the Mara River Basin, there is limited data available, let alone a complete assessment of the data availability and quality. In addition, there are only limited hydrological models of this basin, therefore the understanding of the local hydrological processes is quite limited. Moreover, the absence of good quality discharge time series is not unique to this area, therefore assessing the possibility of calibrating on water levels instead of discharge is very useful for poorly gauged areas and should be explored more detailed in future studies. The advantage of water level time series is the higher availability as it is easier to measure and higher reliability since there is no calculation step in between (using a rating curve). In the future this could be combined with remotely sensed altimetry data.

In short, our key objectives are: 1) present an important data set for the Mara River Basin, 2) illustrate a hydrological modelling methodology where the model is calibrated using river water levels instead of discharge and 3) show the difference between input averaging of the rainfall as typically done and output averaging of the modelled discharge. The latter allows the inclusion of the non-linear behaviour of the rainfall-discharge relation in river basins.

Therefore the key messages for the reader to take away are:

1. In poorly gauged river basins, calibration on water level time series is more reliable than on discharge time series since additional uncertainties arise from fitting rating curves on scarce discharge measurements.

2. In this methodology, the water level-discharge relation is implicitly included in the model; the power exponent of this relation is related to the geometrical data which is observable in the field.
3. The method for dealing with highly spatially distributed rainfall in hydrological modelling is significant to obtain reliable results.

To take this comment into account and highlight these key objectives more clearly, this division into these main topics will be applied throughout the article. In combination with a clearer title, we hope the key messages will be clearer. We suggest changing the title into: **Rainfall-discharge modelling using river stage time series in the absence of reliable discharge information: a case study in the semi-arid Mara River Basin**

*3. There needs to be a broader introduction to data uncertainty in the introduction including rainfall uncertainty as this is considered later in the paper. Furthermore, there should also be a larger section devoted to model calibration and model diagnostics and particularly how to perform robust model evaluation in the face of data uncertainties.*

Thank you for this comment. One of the objectives is indeed on the uncertainty caused by the rainfall heterogeneity, more specific: the difference between averaging of the input precipitation in contrast to averaging the output modelled discharge. Therefore, the introduction should indeed also include rainfall variability. However, in this study, uncertainties in the data for the Mara River Basin were pointed out rather than performing a complete uncertainty analysis to assess the influence of data uncertainty on the modelling results. Therefore, we feel that a section on model uncertainty analysis in the introduction is outside the scope of this article.

*4. A separate section on data would be useful. At the moment, different datasets are introduced at lots of different points throughout section 2 and section 3.*

The reviewer makes a good point here. All data should be introduced in section 2. Those newly mentioned in section 3 (DEM, land cover map, NDVI and remotely sensed precipitation) should have been introduced in section 2 as well. This will be done by subdividing section 2 into multiple sub-sections: Section 2.1 Site description (lines 77-85), Section 2.2 Ground measurements (lines 87-112) and Section 2.3 Remotely sensed data. The latter will be added to introduce remote sensing data that are now newly mentioned in section 3:

Section 2.3 Remotely sensed data

Besides ground measurements, also remotely sensed data were used for the model development. The catchment classification was based on the topography and the land cover. For the topography, a digital elevation map (SRTM) with a resolution of 90 m and vertical accuracy of 16 m was used (U.S. Geological Survey, 2014). The land cover was based on Africover, a land cover database based on ground truth and satellite images (FAO, 1998). Moreover, NDVI maps were used to define parameter constraints.

New information mentioned in section 2.3 will then be excluded from section 3 to avoid repetition.

*5. One of the reasons for calibrating the model to water level is to 'avoid' uncertainties in water discharge. However, by then calibrating the 'c' parameter for the Strickler formula surely you just replace one source of uncertainty with another. As stated in the paper, it is likely that this parameter is also compensating for large sources of uncertainty in your precipitation data so I wonder how robust the results are given all these different sources of uncertainty. This needs to be better discussed in the limitations.*

This indeed is a limitation of this methodology. However, in contrast to the discharge uncertainties, this is a parameter uncertainty that could be quantified more accurately. This is a recommendation for future studies. Therefore, a new section will be added in the discussion to highlight more clearly the short comings of this methodology (e.g. compensation of the slope-roughness parameter *c* for non-closure effects) and recommendations for future studies (e.g. quantification of uncertainties in the parameter *c*, methodologies to constrain or estimate parameter *c*, analysis of the potential of water level based model calibration in well gauged basins to assess the uncertainties more reliably, determination of suitable objective functions for calibrating on water levels instead of flow etc.).

*6. Section 3.3 is really difficult to follow and certain model choices need to be better justified –*
*a. Why was NSE chosen for model evaluation? How appropriate is NSE for calibrating water levels?*
Thank you for this comment. In this case, NSE was chosen for model calibration and validation. However, it was not analysed how appropriate this is for calibrating on water levels. Therefore this is a good recommendation for future studies!
*b. Strickler formula on line 205 needs to be presented as a separate equation – what do 'k' and 'i' denote?*
Thank you for this comment. This will be clarified as such (line 205): […], where R is the hydraulic radius, A the cross-sectional area, k the roughness and i the slope; […]

*7. Results*
*a. Section 4.1. The authors state at a couple of points that 'the observed and modelled water depth were quite similar to each other'. How similar is similar!? It would be better here to state NSE values as a quantitative measure of how similar they are.*
A quantitative measure is indeed useful here. This was done in Table 6.

*b. Section 4.2. How many point discharge measurements were taken? While these can be useful in model calibration and evaluation – I don't think comparing a single point measurement to a whole month of modeled discharge was useful and the fact that the modeled results were 'within an order of magnitude of the point measurement' not a particularly persuasive argument that the model was performing well. I think these could be incorporated much better into the model evaluation framework.*
In total, five point measurements were taken at three locations (see section 2). As there are only a few measurements and a significant time difference between the measurements (2014) and the model (1970s/1980s), it is not possible to use these measurements for model evaluation other than comparing the order of magnitude. If there would have been more measurements, then more accurate comparison methodologies would have been possible.

*8. I was surprised that given the amount of effort that went into defining HRUs and different model structures for the basin based on field observations and interviews, no results or analysis was presented on these different model structures. Was it just data uncertainty that lead to poor model performance or also the definition of model processes? How were model simulations improved by using two different model structures rather than one?*
Thank you for this comment. Analyses on the effect of using different model structures (lumped vs. semi-distributed) were done in an early stage yet the results were indeed not included in the paper. The during the model development, a lumped model structure was compared with a semi-distributed model using two different model structures. This comparison showed applying multiple model structures significantly improved the model, especially during validation (Table 1).

**Table 1: Model comparison: semi-distributed vs lumped for calibration (1988-1991) and validation (1985-1987)**

| | Calibration | | Validation | |
|---|---|---|---|---|
| | $NS_{FDC,log}$ | $NS_{FDC}$ | $NS_{FDC,log}$ | $NS_{FDC}$ |
| **Semi-distributed** | 0.91 | 0.71 | 0.74 | 0.93 |
| **Lumped (SSF model structure)** | 0.87 | 0.42 | 0.00 | 0.41 |
| **Lumped (HOF model structure)** | 0.90 | 0.62 | 0.09 | 0.12 |

**Literature**


FAO: Africover, GLCN, 1998.

Digital Elevation Map: www.earthexplorer.usgs.gov, 2014.