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

Regarding the major comments:
Section 1. The paper states that the "The goal is to develop a reliable hydrological model for the semi-arid and poorly gauged Mara". In my opinion, this is not the kind of objectives that warrants a publication. I am convinced that the authors can identify a set of more appealing objectives for their work.
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.**

Section 1. Can the authors clarify why using water levels for model calibration avoids the effect of discharge uncertainties? This is presented as a fact, with no references to previous literature, and no explanations. I do not find the explanation obvious. Do they imply that rating curves are constant, and that the whole procedure of updating rating curves, as commonly done, is flawed and useless?
Thank you for this comment, this indeed should be explained more explicitly.
It is important to make a distinction between well and poorly gauged river basin. In well gauged basins, sufficient discharge measurements can be available for fitting a rating curve more reliably and updating it regularly. In that case, discharge time series are indeed reliable and useful for model calibration. However, in poorly gauged areas, discharge measurements are generally very scarce. As a result, rating curves are fitted to scarce data and not updated regularly resulting in high uncertainties especially when extrapolating. As a result, there are significant uncertainties in discharge time series. Water level time series however are direct measurements which are therefore more reliable.
For this specific case of the Mara River, data analysis indicated that there are indeed high uncertainties in the discharge data (section 2). Therefore, here water level time series were more reliable than the discharge.
In short, using water levels for model calibration instead of discharge is only an improvement if the rating curve is indeed of poor quality, as often the case in poorly gauged areas.

132: to further delimit HRUs. Which HRUs? Even reading the paragraph further, it is unclear how many HRUs are used. You say 4, but then mention "are mainly cropland and forest, whereas further south the land use is dominated by grassland", which are not in the 4 HRUs.
The HRUs were defined in Line 134: "This resulted in four HRUs in the sub-basin of the Mara River Basin: forested hill slopes, shrubs on hill slopes, agriculture and grassland".

Section 3.3. This section, which is key to explain what was done in the paper, is very convoluted, and impossible to understand. The first sentence states "Parameters and process constraints have been applied to eliminate unrealistic model results". Which model results? "For example, the maximum storage" – why for example? I want to know exactly what was done and how it was done. Instead, there are just a few sentences of how the methodology was carried out, relegating the essential details to even more unclear supplementary materials. "The model was calibrated and evaluated" how was this done? What is the difference between calibration and evaluation? "For the evaluation of this calibration", why this? Is there another calibration? In general every paragraph contains a lot of information in a very convoluted way. It is necessary to describe the methodology in a much more streamlined way.
As the reviewer pointed out, the section on model methodology is indeed quite concise and could benefit from more elaboration. Therefore more details will be added in this section and in the supplements. A table of all the constraints was already included in the supplement (Table S1 and S2). The formulation of "unrealistic model results" is indeed confusing, "unrealistic parameter sets" is more accurate as constraints were applied to eliminate unrealistic parameter sets rather than unrealistic model results; for instance forest interception should be greater than cropland interception. Furthermore, the model evaluation step consisted of several elements: first the model

was evaluated by means of validation (which is what is meant in this section), later on by analysing the discharge on sub-catchment level, analysing the rating curves and the influence of the rainfall (in the discussion).

205. Needs to be expanded and clarified.
1) The procedure for calibration using h and the procedure for evaluation using Q needs to be clearly distinguished.
2) You write that you use d for model calibration and flow duration curves for model evaluation. Flow means Q, but all the plots show d duration curves. Where is the flow used?
3) The value d_{mod} is not present in the Strickler formula (there is A and R). What is the relation to d? It should be written explicitly.
4) What is the relation between the Strickler formula and Q = a â´L°U (h ⍰ h0)b?
5) What is the value of b?
6) If I understand well, the observed and modelled water discharge are obtained using the same formula with the same parameters. Why is then Qrec needed? Trying to explain 3 essential things (model calibration, evaluation and evaluation of rating curves) in the same paragraph does not work.
Reply to 1) There are indeed multiple steps in the use of water level and discharge for calibration and validation that could confuse the reader and should therefore be explained more clearly. First, the model was calibrated on water level (line 202), then the modelled discharge ($Q_{Strickler}$ and Q_{mod}) were compared to the recorded discharge (line 211) for model evaluation.
Reply to 2) The reviewer is right, this will be corrected. Instead of flow duration curve, the duration curves of the water depths were used for calibration. The flow was not used for the calibration.
Reply to 3) Thank you for this comment; this indeed is not written explicitly and should be included:
The cross-sections were simplified as a trapezium with a river width B and two different river bank slopes i_1 and i_2; these coefficients (Table 1) were estimated based on available cross-section data (Supplement S2). In addition, the water depth *d* was calculated from the water level *h* and reference level h_0.

$$A = B * d + \frac{1}{2} * d * (i_1 + i_2) * d$$

$$R = \frac{A}{B + d * \left((1 + i_1^2)^{\frac{1}{2}} + (1 + i_2^2)^{\frac{1}{2}}\right)}$$

$$d = h - h_0$$

Table 1: Coefficients used for the simplification of the river cross-section

	River width B [m]	River bank slope i_1 [-]	River bank slope i_2 [-]	Reference level h_0 [m]
Amala	10.0	3.50	1.83	0
Nyangores	19.05	2.65	5.56	0
Mines	43.81	3.53	3.66	10

Reply to 4) Both equations estimate the discharge using water level data. In the rating curve (Q = a * (h-h0)^b), parameter *a* includes information on the cross-section, roughness and slope; parameter *b* information on the cross-section. This information is more direct in the Strickler formula.
Reply to 5) The value *b* varies for each cross-section. In line 206/7, this information could be included such as:
Note that by using the Strickler formula the exponent of the rating curve is fixed; $Q = a * (h - h_0)^b$, with b = 1.71 at Amala, b = 1.71 at Nyangores and b = 1.70 at Mines using the same water level time series as for the calibration and validation.

Reply to 6) Thank you for this comment, this indeed needs to be explained more clearly. In contrast to what the reviewer stated, the modelled and observed discharge were obtained using different formulas. The modelled discharge Q_{mod} was obtained through the FLEX-Topo model. Q_{Str} was calculated using the Strcikler formula, a calibrated roughness/slope parameter *c* and water level time series. Q_{rec} was obtained from the water department and was calculated locally probably by using a rating curve and the water level time series.
This discharge Q_{Str} was compared to Q_{rec} to compare the recorded and modelled rating curves with each other.

215. Does the model provide simultaneously the output at the 3 stations? Was it calibrated simultaneously to the 3 gauging stations? Or was it calibrated individually to each station? If it was calibrated individually to each station, shouldn't the parameters of the same HRU in different catchment be the same? How was this ensured?
The reviewer has a good point here. The model was calibrated for all three stations individually using the same parameter ranges and constraints. As a result, the parameters were similar, yet slightly different for each station. After calibration, the "best" parameter sets were used for cross-validation. The model performed well when validating at Nyangores using the parameter set based on Mines ($NS_{FDC, log}$ = 0.94 and NS_{FDC} = 0.83) whereas vice versa resulted in poor performance ($NS_{FDC, log}$ = 0.29 and NS_{FDC} = 0.00). This is not surprising as all HRUs were represented when calibrating at Mines and only two HRUs when calibrating at Nyangores, namely forest and agriculture.

230. It appears that the model was calibrated using FDCs. But the objective is to simulate streamflow. Are FDCs sufficient to represent streamflow time series? E.g. I can imagine that information about seasonality as well as timing of peaks is lost when calibrating to FDCs. How were these problems addressed?
This is a good question. By calibrating on FDCs, the focus is on the flow statistics (e.g. how often high flows occur). This information is also in the streamflow, only the exact timings are not included when calibrating on FDCs. However, in this case, the timings were off anyway due to the limited number of rainfall stations available which was insufficient to capture the spatial heterogeneity well. Therefore, in this case the FDCs were good for model calibration.

230. Was it multi objective calibration leading to a Pareto-front? Needs to be clarified.
Thank you for this comment. Instead of analysing a Pareto-front, the values for the objective functions were ordered and the ones with the highest values were considered as "good" parameter sets.