<u>Comments Reviewer 1</u>

This manuscript deals with the problem of disaggregating daily rainfall records into an hourly resolution and particularly assesses the added value of considering the spatial correlation between neighboring stations. The problem of rainfall disaggregation is particularly relevant for hydrological modelling when only daily data are available but high resolution simulations are still required as for flood forecasting or flood predictions in case of fast reacting catchments (with the concentration time smaller than a day). Thus, the manuscript is certainly of a broader interest. Generally, it is also well written. However there are several issues that need more explanations and thus I recommend a major revision.

The authors thank the reviewer for her comments and suggestions. All replies are indicated with page and lines, based on the track-change document.

Major comments:

1. One of my major comments is related to the hypothesis tested by the authors. The authors assume that differences apparent between differently disaggregated rainfall series (V1, V2, V3) will also be present in runoff simulations with a hydrological model. For testing this hypothesis, they use a bucket-type model (HBV). From their results, the authors did not observe any significant difference between these three different rainfall series when fed into the model. I think this is not surprising because this type of model, due to its rather simple structure, may smooth slight differences between different time series present at instant steps, as it reacts to the cumulative rainfall sums over the event rather than to its small variation in time. Despite that, I would expect that you could still see some differences if you analyze instantaneous flows (e.g. event peaks). Yet, if you look only at cumulative statistics of runoff such as monthly average discharges or flow duration curve, you most likely cannot see any differences because these statistics are derived from averaged runoff values. Consequently, these differences could indeed be minute. The only visible effect could be expected on summer and winter extremes. However, these extremes most likely occur in your catchments due to large (and most likely long lasting) rainfall events, for which an exact rainfall distribution within a day is less important. This may explain why you do not observe any differences in these statistics neither. I think these issues should be at least discussed in the manuscript, and particularly the choice of the runoff statistics for the method evaluation.

We thank reviewer 1 for raising her concerns about the choice of the model type. We try to discuss all concerns in the order of her comment. Reviewer 1 is worried about only slight differences between the different rainfall products V1, V2 and V3. However, as shown in Fig. 7 for areal rainfall of one subcatchment, the extreme values differ strongly (5-12 mm) between the different rainfall products for return periods from 2-50 years. Despite that, more frequent rainfall intensities also differ as indicated by Fig. 5 and Fig. 6. So there are differences between the rainfall products.

Also, the reviewer questions the type of model for the actual investigation. We have added a discussion on that in the manuscript (p13 15-21) and want to point out, that the catchments were divided into subcatchments (see also our reply to comment 4 of reviewer 1).

Nevertheless, we agree with reviewer 1, the runoff statistics FDC and Qmon on a daily basis are not suitable to show differences between the rainfall products. However, this was never the intention. Both runoff statistics have been applied only to achieve an overall plausible runoff behaviour for the continuous simulations. We have added a brief discussion on p15 5-9:

"FDC and Q-mon are used to represent the more frequent discharge values. Q-mon accounts for the temporal dependency on the inter-annual variation of the discharge. As mentioned before, the analyzes of FDC and Q-mon allows no direct validation of the rainfall products, but enables an overall plausible simulation of rainfall-runoff processes."

We thank reviewer 1 for the useful suggestion with the possibility of long-lasting events as main reasons for extreme runoff values. Therefore, we have distinguished between summer and winter extremes, so that e.g. the convective events in the fast responding catchment Pionierbrücke and

long-lasting events in the winter period are taken into account. We mention the season-dependent genesis and its influence on runoff extremes on p3 34-p4 1:
"to take into account e.g. summer and winter floods with their different genesis and resulting runoff behaviour."
2. The authors use throughout the manuscript terms: recording and non-recording stations. It is however never explained what they mean with that and I assume this is not a generally used term and thus should be explained as it is significant for this manuscript. It appears that by recording stations they mean stations with hourly records and by non-recording - stations with daily records only. To my understanding, daily stations could also be assumed as recording. Consider using different terms or provide an explanation to the terms used.
We thank the reviewer for this hint. We have rephrased both terms to hourly (former: recording) and daily (former: non-recording) stations and believe, these terms are more intuitive for the reader.

3. Not all important details regarding the calibration of the hydrological model is given. Particularly, the fact that the model is calibrated independently with three different disaggregated datasets appears only in the discussion. These independent calibrations obviously lead to different parameter sets which are then used for simulating runoff. As these calibrated parameter sets compensate for possible errors in the model structure and in rainfall data, these errors are propagated on the simulated runoff (and computed statistics). This makes a direct comparison of runoff simulated with these three different time series difficult. Although the authors are aware of that, in my opinion, it would make more sense to use the model with the same set up. In this way, you could focus only on the effect of different rainfall time series and minimize the possible effect of parameter and model errors. Indeed, it could be worth a try to use only parameter sets derived from one calibration, e.g., with the V1 data set, and use it for both other disaggregated sets, i.e., V2 and V3 and in this way assess the gained effect of introducing the spatial consistence between stations into the set V1.
Again, we want to reply to all points/suggestions in the order of appearance. We added the information about the separate calibration for each rainfall product in the method section (p14 18). Also, we are aware of the possible compensation of rainfall product differences by the model parameters. This is the reason for investigations "c1) HBV-simulation results without calibration using three rain gauges as input" and "c2) WaSiM- simulation results without calibration using three rain gauges as input". For both, c1 and c2, no calibration was carried out. A neutral parameter set was applied to avoid (dis-) advantages for one of the rainfall products/biased results. Nevertheless, no differences with HBV and only slight differences with WaSiM could be identified.

4. The figure 1 and the Table 1 suggest that each of three studied catchments is divided into smaller sub-catchments. Do you actually use these sub-catchments for hydrological modelling or do you model the catchment as one unit? I expect that the spatial representation of rainfall may play a role when using sub-catchments instead of the entire catchment.
We agree with the argument of the reviewer, that subcatchments are important for the investigation of the spatial consistence. The discretization shown in Fig. 1 in several subcatchments (with approx. 20 km²) is applied for the investigation.

5. The disaggregation scheme (p. 8): how exactly do you decide which time step is considered to be wet and which as dry? Also, is the same disaggregation scheme used for all three catchments (i.e., the same scheme of distributing daily totals into hourly intervals) or is that adapted for each catchment independently? In addition, the authors write in lines 16-17 p. 8 that parameters of the disaggregation (which exactly?) are extracted directly from the observed high resolution data, which data do you mean exactly (the most recent hourly data)?
The brief explanation of the cascade model leads to open questions by reviewer 1. However, in accordance with our reply to reviewer 2 we have even shortened the description in the manuscript and strongly refer the interested reader to the original manuscript by Müller and Haberlandt (2015). However, we want to answer the questions also in our reply. The decision about the wetness-state of each time step is made randomly, based on probabilities estimated from observed time series.

Hence, a second run of the disaggregation leads to a different time series. This is the reason why 10 different realizations have been used as input for the rainfall-runoff simulations. So the scheme in Fig. 3 shows only one possible realization for a single day. The parameters are estimated from the nearest hourly stations with a minimum record length of 7 years. This could be the most recently hourly time series or from a station installed from e.g. 1980-2002 For a description of the cascade model parameters the reviewer is kindly referred to Müller and Haberlandt (2015).

6. The disaggregation V3 uses the station with the highest values per day for deciding on an exact disaggregation way. Can you somehow verify that, i.e. how good it works for other stations? or could you justify this choice?
Reviewer 1 is confused by V3. V3 is not a new disaggregation method, only an alternative to V2 to implement spatial consistence into the disaggregated time series of V1. We have tried to clarify this by adding an additional sentence (12 27): " and is also based on the already disaggregated time series of V1."

Minor comments:
1. P. 5 l. 12-13: change the sentence into: An overview of rain gauges used in this study is given in Fig. 1 while their measuring periods in Tab. 2.
Thanks for the suggestion, we have rephrased the sentence.
2. Table 2; it could be a good idea to add the intervals of rainfall recording.
We thank the reviewer for the hint. We have changed the table and now the temporal resolution of the stations is included.
3. Use "rainfall-runoff model" instead of "rainfall-runoff-model" throughout the manuscript.
Thanks for the hint, we have rephrased it throughout the manuscript.
4. P. 6 l. 12-13: could you give a reference for the finding regarding the nonsensitivity to potential evapotranspiration?
The analyzes was carried out in a pre-study by a student (Herzog, 2013) and is hence not citable in a scientific journal. However, the HBV-IWW model was analyzed regarding its sensitivity to input time series of rainfall, temperature and potential evapotranspiration. The model was not sensitive to the latter one.
5. P. 6. L. 7-9: is temperature data corrected for the elevation and if yes how exactly?
The temperature data has not been corrected. Based on station data, an interpolation was carried out for the study area using External Drift Kriging. The additional information used is elevation.
6. p. 7 l. 9-10: do you mean here "hourly" observed time series? From the table 3 it appears that some records are available from much longer period.
We thank reviewer 1 for pointing out the missing information. Indeed, it should be "observed hourly time series". We have rephrased it.
7. p. 7. L. 13-14: it is not clear which data sources were used to extract the maxima over half of year (hourly, daily, monthly)?
We thank reviewer 1 for pointing this out. We added the data sources to avoid misinterpretations (p8 23-224).
8. Use terms "section" and "subsection" instead of "chapter" and "subchapter".
Thanks for the hint, we have rephrased it throughout the manuscript.
9. The paragraph in lines 18-21 on p. 7 could also be removed.
We thank reviewer 1 for the suggestion, but due to the length of section 3 we think that a brief overview at the beginning enables a better understanding of the investigation.
10. p. 10, l. 15: how many different realizations of the disaggregated time series did you use for these simulations?
We are thankful for pointing out the missing information. We have used 10 realizations for each rainfall product and added the information (p14 18-19).
11. P. 10. L. 26, R1 is not explained before.
We have removed this paragraph.

12. Table 5. The values for the HBV parameters: k1, k1, k2 and kperc are given in days. If the model is run at an hourly time interval, should not these parameters be expressed in hours?
We thank the reviewer for his concern about the units. As it can be identified for k0, not only integer values are possible (minimum for k0=0.25 d = 6 h). The unit [d] was only chosen here for a better understanding of the value, e.g. k2=500 d is more suitable/easier to follow in comparison to k2= 12000 h.
13. Fig, 5 and next: one realization from how many?
10 realizations for each rainfall product, please see our answer to minor comment 10.

References:
Müller, H. und Haberlandt, U. (2015). „Temporal Rainfall Disaggregation with a Cascade Model: From Single-Station Disaggregation to Spatial Rainfall". *Journal of Hydrologic Engineering* 20 (11), p. 04015026.
Herzog, Y. (2013): "Sensitivität eines hydrologischen Modells auf Veränderungen in den Klimavariablen Niederschlag, Temperatur und potentielle Verdunstung". *Student thesis* at the Institute of Water Resources Management, Hydrology and Agricultural Hydraulic Engineering, Faculty of Civil Engineering and Geodesy, Leibniz Universität Hannover (in German).