# Response to Referee #3

*We thank the reviewer for the helpful comments. As can be seen from the detailed responses below, we intend to carefully consider all comments and will aim to adequately address them by carefully selected additional model simulations, analysis, and changes to the manuscript. Clearly one practical constraint that we need to sensibly account for is computational load and related efforts; our suggested ensemble extensions and advancements try to deal with this challenge in a best possible way. In the responses below, our comments are inserted with italicized, black text. The changes in the manuscript text are inserted in green. The original response from the referee is in blue.*

## General comments and recommendation

The manuscript by Hohmann et al. presents an analysis of the effect of different meteorological network densities, as well as different interpolation schemes on the simulated runoff for a meso-scale catchment in southeastern Austria. While indeed many questions about the optimal meteorological network density, about the adequacy and representativeness of stations' distribution and about the suitability of different interpolation schemes for hydrological modelling have not yet been fully answered, I do not see right now in this study an adequate and robust assessment, offering an added value to what we already know. The setup and design of the experiment is not thorough enough to support reliable statements with evidence.

*Thank you very much for seeing the importance of an "optimal meteorological network density" in the context of hydrological modelling. We would like further strengthen the setup and study design to be able to support reliable statements with more evidence. Here we will just name our ideas and further describe these points at the specific comments below.*

- *Include an ensemble of events with 20 to 30 short-duration and 20 to 30 long-duration events*
- *Strengthen the stations networks, with more setups: 5, 8, 12, 17, 25, 36, 56, 75, 109 and 158 stations*
- *Include a second setup of station subnetworks of interim sizes (12 to 109), to further investigate the uncertainty of pre-defined networks with a given station number*
- *Redefine interpolation methods, and make their sensitivity investigation more systematic*

The manuscript has per se a clear structure, the methods are generally described in a comprehensible way or supported by relevant sources, however, some explanations could be clearer and more concise (e.g. the calibration procedure). Even though the discussion provides some good points, generally there are quite a few redundant paragraphs, while more interesting and critical points are not examined closely enough. The manuscript generally features high-quality and interesting figures; some of the tables and their captions should be reorganized in a more meaningful and efficient way. I would also suggest a native speaker to read it and correct it, some sentences definitely need to be rephrased.

*Thank you for mentioning these points. We will revise the calibration chapter in our manuscript to further explain our calibration procedure. For more specific details see the point below under "Calibration of the hydrological model". The discussion part will be adjusted addressing the new study setup, including more critical discussion and avoiding redundant paragraphs. Also, the tables and captions will get a close look. And for the revised manuscript we will also ask a native speaker to read* *through it. Thank you also for mentioning some sentences to be rephrased (see points below in the "Discussion" part).*

Because of these considerations, I think the manuscript requires and extension of the experimental design, a more critical discussion and further work, before it can be possibly recommended for publication.

*As you can see above in our first answer, we intend to improve our experimental design in several major points. We will also give more attention to the discussion part to further investigate the critical points. With all these improvements we hope that the manuscript could be recommended and we can help to more robustly answer open questions about the optimal precipitation network density and the suitability of different interpolation methods for hydrological runoff simulations.*

Please find my specific and technical comments here following.

**Specific comments**

- Experimental design:
    - One important drawback of your setup is the fact that despite the stations' density is high, it is not covering your whole study area, but only its central part.

      *As an explanation, the WEGN was buildup 2007 for meteorological and climate change setups, which led to have the near-rectangular domain, instead of covering a catchment domain.*

      *That the dense station network it is not covering the whole study area it true, but mostly effecting the total catchment, so the gauging stations Feldbach/Raab and Neumerkt/Raab. To overcome this uncertainty, we choose only subcatchments which were mostly covered by WEGN stations. Since the small-scale subcatchments are in focus, we agree that we need to make this clearer in the manuscript. And additionally, we will also mention this uncertainty of the missing coverage for gauging station Neumarkt/Raab and Feldbach/Raab in the manuscript.*

      *Beside these changes in the manuscript and to further overcome the uncertainty of the dense station network, compared to the surrounding stations, we intend to only change the WEGN area/zone in the precipitation input. Therefore, we intend to set the surrounding areas to one baseline setup of precipitation input (like the 8 Stations ZAMG&AHYD case with IDW2 interpolation) and then only change within the WEGN area/zone. We will create the precipitation input maps, which only show changes in this area, where we have the opportunity to well control the density of the setup. Hence, we can have a better focus on the area where we have additional information each time when the subnetwork is densified.*

    - You have "fix" a priori chosen subnetwork configurations, but for a fair evaluation of the effect of stations' subnetwork density you should try different (random?) configurations for the same number stations. Tentatively, you could also consider smaller jumps between one configuration and the next one.

      *We selected the station densities in this way: Our base setup is the full network with 158 stations, we assume to catch the precipitation events in the best possible way. Then, step* *by step we reduced the number of stations by removing them randomly, while ensuring a uniform spatial distribution by considering the area-of-influence of each station (see also supplement of O and Foelsche 2019 https://doi.org/10.5194/hess-23-2863-2019-supplement -> Rain-gauge sub-networks for more details about the method). Again, we assume to catch the precipitation with these well distributed setups as good as possible. We end up with the 8 and 5 stations case, which is the operational stations network (of the meteorological and hydrological services of Austria, ZAMG and AHYD), without any stations from the research network. With this, we tried to get a reasonable setup of stations to cover the area as good as possible. The idea of doubling the station number was based on our expectation as well as based on literature, that the biggest uncertainties/sensitivities are observed for fewer stations, where the "jumps" are smaller.*

*However, we will significantly improve the station subnetwork density sampling, to strengthen the study setup overall. We still need to accept overall computational load limits. → So, we will make the sequence of subnetwork cases denser and more systematic as follows: we still take the 5, and 5+3=8 operational ZAMG&AHYD station cases as the "background network" baseline, and otherwise only increment from one subnetwork to the next within a factor of 1.5 (rather than 2 or more), where the factor 1.45 was found helpful as a guide. This leads to a cascade of overall ten subnetworks with station number as follows: 5, 8, 12, 17, 25, 36, 56, 75, 109, 158 (with the seven number-cases 12, 17, 25, 36, 56, 75, 109 being the core of interim cases between just the operational network of 8 stations and the full 158 stations).*

*We also see the potential utility of a selection of many more station networks. But the option with random picking, and hundreds of runs is really not feasible for our setup, because of computational time with the process-based modelling approach. → To further investigate the uncertainty also related to pre-defined subnetworks of given number, we intend to use two different subnetworks (spatially complementary, using different actual stations from the WEGN) for each of the seven interim number-cases (i.e., from 12 to 109 stations). Thus, we have a sensitivity crosscheck to actual spatial station distribution at a given total number of stations. Hence in total 17 subnetwork cases need to be analyzed within this design, which together with the ensemble expansions in rainfall events and on interpolation choices (see bloew), stretches the computational load and efforts to the feasible limits. We agree that also this more balanced selection of subnetwork cases will substantially contribute to improved robustness of results.*

- You chose 2 interpolation methods (I wouldn't refer to three methods, as you simply changed a parameter of the second method), I think it would be more appropriate to use more interpolation methods, such as kriging, etc.. for a more robust evaluation of the effect given by the interpolation method, and also definitely to use a smaller searching radius (second spurious peaks in your hydrological simulations are not surprising, given how you interpolate precipitation).

*Thank you for your comment. We tested Kriging, but did not see additional value. We more had seen the problem of how to decide which variogram to use, especially for a long-time frame of many years and a half-hourly resolution. But again, we will have a closer look as well to the use of Kriging. And we will find a reasonable way to use the interpolation method of ordinary Kriging. But we still have to deal with the computational time,* *otherwise it would “explode” with too many interpolation-method cases on top of significantly enlarged rainfall event and station subnetwork ensembles.*

*Also for the purpose of exploring the key issue how individual-station rainfalls (point-scale time series) are spread by a certain interpolation method into the space, we have a new idea after receiving all three reviews: The IDW with exponent 2 will be kept as a baseline, and in contrast exponent 1 (quite more spread vs exponent 2) or exponent 3 (quite less spread vs exponent 2), provides the essential insights needed. Based on the evidence we have seen; we believe the key for the area rainwater flux received into the (sub)catchments is certainly how the spatial spreading plays out. Hence in the revision we tend to “simply” make our concept of systematic testing of interpolation influence really clearer. And we definitely will improve our interpolation setup per station subnetwork case, in particular that the overall catchment region around the WegenerNet core region (i.e., the general Raab catchment covered by the eight stations of operational ZAMG+AHYD stations) is kept at baseline settings (also for interpolation). The subnetwork’s densification is properly accompanied at each station density level by adequate interpolation settings (e.g., IDW2, IDW1 vs IDW2, IDW3 vs IDW2).*

*We are interested in the reviewer’s opinion as to whether the systematic assessment of “spatial spread influence” of interpolations is not in his/her view also usefully covered already in the context of this study by the “IDW2 plus IDW1-vs-IDW2 and IDW3-vs-IDW2” approach. Especially now that we focus with this interpolation influence on the WegenerNet core region with its dense stations. Kriging and related issues may induce undue additional work and trials. We do not expect to additionally learn on the effect of how the increase of spatial spreading of point rainfalls impacts on runoff, beyond what we can learn from looking across the IDW1, IDW2, IDW3 cases at each subnetwork density level.*

*To be able to further adjust the search radius of the IDW interpolation we will split up the interpolation maps in the core area/zone of the WEGN and the surrounding area, which is not covered be the dense network. And then the idea is to only change this core WEGN area/zone in the precipitation input. The surrounding areas will be set to a baseline setup of precipitation input (like the 8 Stations ZAMG&AHYD case with IDW2 interpolation), which will not be changed. Therefore, will create precipitation input maps, which only show changes in the WEGN area/zone. With this approach we can adjust the search radius of IDW individually for each density network, without losing necessary information of surrounding precipitation stations.*

The number and sample of events you are analyzing is simply too limited to allow you to make any meaningful assessment, now your analysis might show only very specific and localized effects, and might not hold for a larger ensemble of events. Is there any way to increase the number of events you analyze? You say you selected first heavy precipitation events among the top 10% heaviest rainfall days during summer, and out of these you selected through visual inspection your 6 events. I would suggest you rather select the top 100 events for 1d and 3d accumulation periods, for instance, and further analyze these?

*Thank for this comment. We also checked more events, but then decided to stick to these 6 ones, since they have been among the most extreme rainfall cases in our time frame.*

*We agree with the reviewer that a larger ensemble of events will strengthen the robustness of our analysis. So, we now intend to expand to an ensemble of events with 20 to 30 short-duration, heavy convective rainfall events and 20 to 30 long-duration heavy rainfall events. The selection of the events will still target the 10% heaviest rainfall events (i.e., above 90th percentile in hourly intensity). To be able to include enough "really" heavy rainfall events the time period will be extended as needed from currently to 2012 out to 2018 or 2019. With such change, which comes at significantly increased computational load and effort though, we have the opportunity to extract heavy/extreme events from a much larger pool of rainfall events. So clearly the robustness of the results will substantially increase compared to our "initial study approach" we followed so far.*

- (You are only using discharge gauging stations on the main river trunk, but you analyze also the contributing subcatchments. Sure, this is fine for looking at the effect of the different precipitation inputs, but not enough to disentangle the effects possibly stemming from the parameters of the hydrological model. You implicitly assume the hydrological model is working equally well on much smaller subcatchments, with the same parameters. As the model is process oriented I am fine with this assumption, but you might want to spend a few words on this?)

  *Thank you for pointing this out. Yes, unfortunately we do not have measured runoff data for the subcatchments. That is one reason why we decided to use a process-oriented model. We will also improve the description in the manuscript to this end.*

- **Data**:

- You have a 10 year period to choose from resp. analyze, is there some way to extend it?

  *From the WEGN we have data from 2007 until now. We chose the time frame of 2009 to 2012, because of known especially extreme events in this period. We will check the extreme events in our time frame, and the time period will be extended as needed from currently to 2012 out to 2018 or 2019 or so. Another aspect we have to care about, is the computational time, so we tried to keep each model run as short as necessary.*

- You only report the return period for discharge, but actually it would be interesting – and relevant? - to know the return period associated with your rainfall events too.

  *Thank you, we will look into this issue again, too. We consider that we only have data from 2007 onward within the WEGN. The runoff timeseries of station Neumarkt/Raab is quite longer, with data starting from 1991.*

- **Model setup:**

- P7-L128-130: Lumped doesn't necessarily mean that a model is not process-oriented. I guess you mean conceptual? (that are often lumped, but not always)

  *Yes, exactly that is what we meant. We will change it to:*

  Line 128-130: We focused on a process-oriented model to keep the model uncertainty small, compared to ~~lumped~~ conceptual models, which are often used for similar precipitation runoff studies (e.g., Dong et al., 2005; Zeng et al., 2018; Huang et al., 2019).

- **Calibration of the hydrological model**

  - Please describe better and provide more details on your calibration procedure.

    *Yes, we will provide such a paragraph with more details, and maybe include a section in the appendix to have the space to further describe the calibration without unduly lengthening the manuscript.*

  - Why do you calibrate the model only basing on one summer, and only for one interpolation method?

    *We calibrated the model just for one interpolation method, but we also checked the efficiencies for the other ones. In Table 1 you find the NSE and KGE efficiencies for the other interpolation methods and also different station densities.*

*Table 1: NSE and KGE efficiencies for the calibration period (Mai to September 2009) and validation period (Mai to September 2010) for all station numbers and interpolation method at gauging station Neumarkt/Raab. The model was calibrated with the IDW2 158 stations case at gauging station Neumarkt/Raab.*

| | Calibration Period<br>Mai - Sep. 2009 | | Validation Period<br>Mai - Sep. 2010 | |
|---|---|---|---|---|
| | NSE | KGE | NSE | KGE |
| IDW2 158 Stations | 0.79 | 0.75 | 0.67 | 0.81 |
| TP 158 Stations | 0.81 | 0.76 | 0.65 | 0.81 |
| IDW3 158 Stations | 0.79 | 0.76 | 0.66 | 0.81 |
| IDW2 64 Stations | 0.79 | 0.74 | 0.66 | 0.81 |
| TP 64 Stations | 0.8 | 0.76 | 0.65 | 0.81 |
| IDW3 64 Stations | 0.8 | 0.75 | 0.66 | 0.81 |
| IDW2 32 Stations | 0.8 | 0.75 | 0.68 | 0.81 |
| TP 32 Stations | 0.81 | 0.77 | 0.66 | 0.81 |
| IDW3 32 Stations | 0.8 | 0.76 | 0.67 | 0.81 |
| IDW2 16 Stations | 0.8 | 0.72 | 0.69 | 0.81 |
| TP 16 Stations | 0.8 | 0.73 | 0.69 | 0.82 |
| IDW3 16 Stations | 0.8 | 0.73 | 0.69 | 0.81 |
| IDW2 8 Stations | 0.8 | 0.7 | 0.66 | 0.79 |
| TP 8 Stations | 0.79 | 0.75 | 0.63 | 0.77 |
| IDW3 8 Stations | 0.8 | 0.7 | 0.66 | 0.79 |
| IDW2 5 Stations | 0.8 | 0.76 | 0.64 | 0.76 |
| TP 5 Stations | 0.78 | 0.78 | 0.6 | 0.73 |
| IDW3 5 Stations | 0.8 | 0.77 | 0.62 | 0.75 |

*We calibrated the model for the summer 2009, since in 2009 many extreme events occurred which are in focus of our study. Beside that we also wanted to keep the computational time in a reasonable way, to be able to do the first calibration steps with the SCE-UA. And as we mentioned before the high-resolution process-oriented model setup need quite a lot computational work.*

- Why do you use "only" NSE and KGE to define your objective function, instead of including further goodness-of-fit measures and criteria, perhaps more specifically tailored for floods, their volume and/or their timing?

  *The objective function NSE is more considering the peaks and the KGE more the whole water balance. Both aspects are important, especially if we want to analyze different types of rainfall runoff events (convective/advective). We wanted to focus one these two aspects, but we can of cause also have a look to further goodness-of-fit measures.*

- **Runoff analysis approach:**

- I would recommend you expand your analysis approach by including also the peak timing, as this also is an important aspect in your modelling exercise, e.g. considering the effect of superposition of peaks simulated in the subcatchments.

  *Thank you also for this suggestion. For some events we even checked the peak timing and saw no huge difference, therefore we did not include it. But now, especially when adding many more events, we will analyze the timing of the peak runoff and include it in our results and discussion. Thereby we will have a second source to analyze the runoff behavior of the (sub)catchments.*

- **Results**:

- P16-L271: I wouldn't say that, e.g. Kornbach is rather systematically overestimated.

  *Yes we agree, we may delete the sentence, but we will see how the analysis results of the ensemble of events come up. This will help to further investigate the effect of overestimation or underestimation of specific sub-catchments.*

  Line 270 to 271: ~~In the comparison between the (sub)catchments, no single (sub)catchment is especially noticeable.~~

- P16-L275-277: this is not true, see for example the short-2 event or the long-3 event. –

  *We wanted to say that, if we simulate the runoff with 5 Stations under the TP interpolation scheme and then with 8 Stations and the TP interpolation scheme we will get the same runoff. So, the same runoff deviation compared between the two runs. As an example: For the short-2 event the peak flow deviation will be in the Grazbach -11 for the 5-Stations case and -11 for the 8-Stations case. For the short-3 e.g. at Auersbach the peak flow deviation is 32 for the 5-Stations and 8-Stations case.*

  *We will change the sentence to make this more clear and easier readable, so that no misunderstanding should occur.*

*Line 275-277:* The northern catchments Auersbach, Kornbach and Grazbach do not show differences if we simulate with 5- or 8-Stations subnetworks under the TP interpolation scheme, because of their location in relation to these station locations. *(Will be changed)*

**Discussion**:

- P 20: it is confusing you mentioning first Lopez et al.2015 report no increase in performance after including 24 gauges per 1000 km2, and later on saying they actually used 12 stations per 1000km2. They used gridded products to be able to increase the number of stations with "hypothetical" gauging stations. I think you should either specify this, or not go so much into detail.

  *Yes, while working with the other reviewer comments we also recognized this confusion. We will make sure to clarify this point and adjust the manuscript.*

- P20-L360-363: Rephrase please this last paragraph.

  *Yes, we will rephrase these sentences in the revised manuscript.*

- P21-L380-381: Isn't this possibly case specific? Refer also to Lobligeois et al. 2014: In all regions, natural variability allows for contradictory examples to be found, showing that analyzing a large number of events over varied catchments is warranted.

  *We hope to be able to further investigate to this point, when we analyze the ensemble of events and then adjust the sentence in the manuscript.*

- I think you should expand more on the representativeness of stations' location.

  *Good hint, we will add a paragraph in the discussion part. Besides the option of just one best distributed station network we will include a second setup with also a very good distribution of stations (see in the answers above). This will help to further analyze the stations location. This analysis will also be carefully discussed in the manuscript.*

### Technical corrections

- P6-L124: developed..by who et al.?

  *Sorry for this. We will correct the sentence:*

  We used the hydrological model WaSiM, developed by Schulla et al. (1997), at the ETH Zurich in Switzerland for climate change studies in Alpine catchments.

- P7-L128: BAFU is FOEN in english

  *Thank you, we will change the part to:*

  […] up to operational use (e.g. at ~~BAFU~~ FOEN Switzerland).

- Figure 2: shouldn't you remove the box with snow accumulation/melt, as far as I understood you are not using it?

  *The model is running with the snow module, to be able to simulate continually over time. But the focus time are only the summer month. So, it would be an option to exclude it, because it is not in focus and will not be calibrated.*

- Figure 4: You should be consistent, and either use the min-max range in both cases, or the 5th and 95th percentile range. And why don't you also show the same kind of information for the 3rd source of precipitation data?

  *We intend to update the figure using min-max range for all three datasets.*

- Figure 6: Why don't you show observations in the last column?(i.e. for the Neumarkt/Raab gauging station)

  *We wanted to keep the comparability between the different catchments. Another line in the Neumarkt/Raab figure was more confusing than supporting, from our experience. But we will double check this and may change this.*

- P17-L301 ..in these subcatchment should be in *this* subatchment

  *Yes, we will correct the sentence accordingly:*

  *Line 300-301:* Only the 64-Stations subnetwork under the IDW3 interpolation scheme result in a –25 % peak flow deviation in ~~these~~ this subcatchment.

- P21-L373: ..more event-depended should be event-*dependent*

  *Yes, we will correct the sentence accordingly:*

  *Line 373:* The runoff for the short-duration events is much more event-~~depended~~ dependent