Response to Referee #2

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*.*

This study tried to explore the effect of interpolation methods and station density on runoff simulation. However, this work has some severe problems and does not reach the standard of HESS in the current form. My comments are as follows.

Major comments:

1. It can hardly say that this study is novel or contributes much to our knowledge. The authors used two simple and widely used interpolation methods, TP and IDW. Results and conclusions are generally similar to previously published work.

Thank you for your comment. Here we summarize three main points of novelty which we see and wanted to publish based on our original "initial study approach". We also include some points on how we intend to advance the study design and improve the manuscript.

1. *As far as we know, the combination of station density and interpolation scheme in the context of runoff at high resolutions (1 to 10 km scales) has not been studied in this way. In the manuscript we mention in the introduction (Lines 72 – 74) the studies of Gervais et al. (2014), Avila et al. (2015), and Herrera et al. (2018) which combine the impact of station density and interpolation scheme on rainfall data quality, but do not go the next step in the direction of integrating hydrology (i.e., sensitivity of runoff). So, we see new added value especially in the combination between the station density and interpolation scheme in the context of process based hydrological modelling at these high resolutions, with a particular focus to highly variable convective rainfall variability. → We will definitely have a closer look to the manuscript to strengthen this focus target even more. For example, we intend changes such as indicated in the following sentence:*

   L72 – 75 Since previous studies have assessed the impact of the station density and interpolation on precipitation data quality such as mean and extreme rainfall values (Gervais et al., 2014; Avila et al., 2015; Herrera et al., 2018). We go a next step and focus on the impact of such precipitation uncertainty on hydrologic simulations, especially runoff peaks and the combination of station density and interpolation method.

2. *Even though the method of studying sensitivities is obviously not new, we find that our results are. Especially given that we have a station (sub)networks and (sub)catchments design and combine rainfall events with hydrological modeling, something that was not possible in other*

regions than the WegenerNet so far. We admit though that the design of the rainfall event ensemble, the subnetworks, and the interpolation benefits from further significant extension and improvement, which we now suggest below to implement during revision. The other studies who analyzed station densities (e.g., Dong et al., 2005; Bárdossy and Das, 2008; Meselhe et al., 2009; Xu et al., 2013; Zeng et al., 2018; Huang et al., 2019) mention a threshold with no significant increase of model performance with a denser network. In our study we analyzed as one focus short-term convective events which do not show such an expected threshold especially for small catchments. Only the long duration stratiform rainfall events show such a threshold, with no improvement with more stations. The fact that the short convective events do not show this threshold behavior was, in our perception relative to the previous studies, one of the "unexpected" new results. → As an improvement of the study we intend to include a larger rainfall event ensemble, in line with reviewer suggestions, to further investigate and assess the robustness of such "unexpected" behavior. Specifically, we intend to use an ensemble of 20 to 30 short-duration events and 20 to 30 long-duration events. We will also strengthen the manuscript by discussing these events and results in more detail.

3. *The highly dense precipitation station dataset, with 150 gauging station per 300 km² in its core region (complemented by eight operational network stations over 1000 km²), is also a new opportunity to explore our study questions. Another quite dense network was used by Lopez et al. (2015), with 12 station per 1000 km²; and even though they added another 40 hypothetical stations in the Thur basin (~1700 km²), this study is still not as dense as ours. The other studies show even smaller station densities. → With an improved design in combination of a bigger rainfall event ensemble and a more and improved station subnetworks ensemble we can better capitalize on this novel network density for such a study. And we may point to this fact and the station data density in the manuscript in the introduction:*

   L 65-67 The highly dense gridded station network WegenerNet (WEGN) (about 1 station every 2 km² over an area of 300 km²) in the southeastern Alpine forelands of Austria allows us to study these questions related to the Raab catchment and its subcatchments.
   L68-69 Because of the exceptionally dense data availability (Sect. 2.2) it is possible to analyze the influence of precipitation station densities on runoff in detail.

Besides, as a study focusing on the effect of station density and interpolation methods on hydrological modeling, the design is too simple. It is better and not difficult to (1) design station networks with random numbers and locations of stations instead of using pre-defined networks (Table 3), and (2) conduct a sensitivity test of the weighting power instead of using 2 and 3. A negative example caused by (1) is that the authors state "32 stations would be enough . . .", but this resulted from a large jump from 16 to 32 to 64 stations.

Thank you very much for the suggestion. We first explain our thoughts behind these points and then also give our suggestions how to improve.

To point 1) 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 an 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) each for 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 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.

To point 2) Thank you for this suggestion, which we also will have to deal with a due eye to feasibility. 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 would like to study this now in the following way: 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). 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 additionally 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 Raabtal region 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 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. Instead of including Thiessen Polygons and/or additional interpolation schemes like Kriging. Since 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.*

2. A large part of the study area is not covered by 158 rain gauges. Both IDW and TP cannot achieve reasonable estimates outside the WEGN network. As a result, only rainfall in the middle reach is largely affected by different density and interpolation methods, while rainfall in the upper and lower reach could always contain much larger errors. Meanwhile, the hydrological model is built over the whole domain, and runoff at catchment and sub-catchment outlets is compared. The effect of biased estimates in upper and lower reach on the runoff simulation could be very large.

Yes, that is true. To overcome this uncertainty, we choose only subcatchments, which were mostly covered by WEGN. Since the subcatchments are in focus, we should make this clearer. And it is true, that we still should mention this uncertainty of the missing coverage for gauging station Neumarkt/Raab and Feldbach/Raab in the manuscript. We intend to do so.

Another aspect is that the setup with the 5 or 8 stations (around 5 or 8 stations per 1000km²) with well distributed gauges over the total catchment is already more than other studies have to choose from. E.g. Xu et al. (2013) studied the Xiangjiang River (94 660 km²) with around 2 rain gauges per 1000 km² (threshold with 1 station per 1000 km²). Dong et al. (2005) studied the Qingjiang river (12 209 km2) with also with around 2 rain gauges per 1000 km² (threshold around 0.4 gauges per 1000 km²).

To overcome the point that only the middle reach is affected, we intend to only change this WEGN area/zone in the precipitation input. Therefore, we will 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. In this way we also can more adequately adjust the maximum distance of IDW as needed, without inducing undue/unhelpful change in information for the surrounding (low density) stations and areas.

3. Please perform the analysis based on a larger collection of precipitation events. Currently, only three small-scale short-duration and three large-scale long-duration events are selected. Although the six events may be representative, a comprehensive view in a long historical period is useful and necessary to demonstrate the all-aspect effect of interpolation and station density. The current results and analyses are all based on those limited events, making the results more “casual” than “causal”. For example, the interpolation based on 5 stations could be largely affected by the location of storm centers, and thus the results based on 100 events could be different with results based on 6 events.

Thank for your comment. Yes, it is true, they might be affected. 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.

As an improvement, 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 2012 out to 2018 or 2019 or so. 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.*

With this approach, we hope to be more representative for the study area. And we also hope to overcome the problem with the location of single storm cells a bit more.

4. Figure 5: The IDW precipitation map based on 158 stations looks quite unrealistic. An actual precipitation event should be spatially continuous like that in Figure 4. However, the bull eye effect in Figure 5 is too obvious. The authors stated that all stations within a 50 km searching radius would be used in IDW. 50 km is quite large based on the measuring scale in Figure 1. This makes the bull eye effect even weirder.

We will double check this point. But to overcome the problem with just one search radius, we will (as already mentioned in point 2.) only change within the WEGN area the different interpolation schemes and network densities. Now, we will also have the option to change the search radius for every network density respectively. We also hope to overcome the "bull eye effect" with this new setup a bit more, to the degree that area rainfall flux into subcatchments appears clearly plausible.

5. The authors strengthened "spatial rainfall variability" in the title. But the spatial variability is not explicitly analyzed in this study. The title does not reflect what this study actually did, i.e., station density and interpolation methods. Besides, the effect of station density and interpolation methods cannot be simply represented using "spatial variability".

Thank you for pointing this out. We will consider changing the title.

Specific comments:

6. Please adjust the font of units which is often different from that of texts.

Thank you for the hint, it might be because the units are introduced in LaTeX with \unit{XX} e.g. \unit{km}, like the Copernicus LaTeX package said.

" %% Copernicus Publications Manuscript Preparation Template for LaTeX Submissions

%%% PHYSICAL UNITS: Please use \unit{} and apply the exponential notation"

But we will try to solve this problem.

7. Line 25: It is better to state the "measurement uncertainties of rain gauges" because other approaches of rainfall measuring are not mentioned here.

We will change the sentence to:

(Line 24-26) Beside the measurement uncertainties of rain gauges, considerable uncertainty can arise when point-level measurements are spatially interpolated for final gridded products (Goodrich et al., 1995; Mcmillan et al., 2012; Huang et al., 2019; O and Foelsche, 2019).

8. Line 32-34: The pros and cons of radars and satellite sensors are similar in many cases and should be stated together.

We will change the sentences to:

(Line 31-34) On the other side, indirectly estimate precipitation like radar systems and satellites show a higher spatial resolution of the precipitation cells, but do not give specific precipitation amounts (~~e.g., Sun et al., 2000; Tetzlaff and Uhlenbrook, 2005~~). They indirectly estimate precipitation and therefore their data are subject to errors and uncertainties (e.g., Sun et al., 2000; Tetzlaff and Uhlenbrook, 2005; Tian and Peters-Lidard, 2010; Kirstetter et al., 2012; O et al., 2017; Lasser et al., 2019).

9. Figure 1: It will be helpful to add latitude and longitude.

Thanks, yes, we will add latitude and longitude to both maps.

10. Line 124: Please complete the reference for WaSiM.

Sorry for that. We will correct the sentence accordingly:

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.

11. Line 150: Although I understand what you mean by 50% NSE and 50% KGE, please rephrase to be more formal.

Thank you. We attend to rewrite the sentence e.g. like this:

The model performance was assessed with ~~50 %~~ the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) and ~~50 %~~ the Kling-Gupta efficiency (KGE) (Gupta et al., 2009) both weighted half of the total model efficiency.

12. Line 155-162: Given the authors state that manual recalibration is necessary, please add some descriptions on the benefits of manual recalibration. For example, what's the KGE and NSE before manual recalibration? Besides, the NSE decreased from calibration to validation periods but KGE increased. Please add some explanations.

The efficiency measures NSE and KGE were pretty good with the SCE-UA, since they were the goal of the algorithm. But with this parameter set, the runoff components were physically unrealistic, e.g., with no baseflow at all, which is not realistic for our catchment in the Alpine foreland. Therefore, we did a manual recalibration with a focus on the runoff components.

We will make this clearer in the revised manuscript and add some more lines about our calibration strategy.

13. Line 167: Strictly speaking, you used two, not three methods. Two different parameters do not make IDW two different methods.

Yes, this will be rephrased, depending on the new setup of interpolation methods.

14. Figure 4: Please use the shapefile of the catchment to replace the black box, which can help identify whether storm centers are located within or outside the river basin.

Thank you for this hint. We will change the figure, including the catchment appearance.