Response to Referee #1

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*.*

In their paper, Hohmann et al. studied the runoff sensitivity to spatial rainfall variability, namely they investigated the impact of station density and interpolation schemes on runoff simulations. I found the novelty of the study very limited. The effects of the density of the network of rain gauges and different interpolation methods on hydrological models have been analyzed many times in the past – as the authors mention themselves in the introduction to the paper. The research questions asked in the paper ("how many stations do we need to reliably model hydrological runoff under heavy rainfall events?" and "how large is the influence of interpolation scheme on runoff results under different station network densities?") are simply not new.

Thank you for your comment. While we agree that the research questions are in general not new, we summarize here three main points of novelty we see and wanted to publish in this 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 combined rainfall event with hydrological modeling, something that was not possible in other regions than 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.

I found the results to be very limited following the decision to explore only two interpolation methods (TP and IDW), which seems to be motivated mainly by the fact that these are the interpolation methods that are built-into the WaSiM model that was used in this study while ignoring other common interpolation methods (e.g. Kriging).

In WaSiM it is in principle feasible to include various external grids for meteorological input. Therefore, we confirm, Kriging was not excluded because it is not supported. We even 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.

And 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 more clear. 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 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).*

But again, having said the above, 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 as a possible alternative of IDW3 in our revised study setup. This will help to have a third interpolation method (Thiessen, IDW2, and Kriging), but still keep the computational time of the model feasible (currently around 4-5 days per simulation run per server at our computational servers). The computational time otherwise would "explode" with too many interpolation-method cases on top of significantly enlarged rainfall event and station subnetwork ensembles.

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.

Are six rainfall events enough for this type of study? Considering the authors discuss the effects of rainfall variability in space, I would expect many more events to be interpolated and simulated to demonstrate the robustness of the results.

Thank for your 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.

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 to 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.

I cannot recommend the manuscript for publication in HESS in its current form. I would suggest extensive revisions of the study that include extending the number of extreme rainfall events in the analysis, exploring additional interpolation methods and, as the authors suggested in the abstract, conducting further investigation to study if a rainfall ensemble can be used to decompose the effects of rainfall records and interpolation uncertainties on modelled runoff – this might be qualified as a novel aspect that can contribute to the originality of the study, which is now missing.

Thank you for this valuable list of suggestions how to improve the manuscript. In the following, we summarize the extensions and advancements we intend to implement in order to improve the study design and the manuscript along these suggestions.

- *As one important point, already alluded to in the answers above, we will extend the number of extreme rainfall events significantly. We will use an ensemble of 20 to 30 short-duration* *heavy rainfall events and 20 to 30 long-duration heavy rainfall events, that is 40 to 60 events in total. For more details see the notes above. This event-ensemble approach, compared to the handful of events of our "initial study approach", will help to statistically and in more detail analyze the effect of different station subnetwork densities. This clearly will give substantially more robustness to our results.*

- *As to a more systematic study of the co-influence of interpolation method choices, at each subnetwork density level, we want to test the "IDW2 plus IDW1-vs-IDW2 and IDW3-vs-IDW2" as a baseline. We, however, would keep the Thiessen as well and want to test also ordinary Kriging in our revised study setup (such ordinary Kriging assumes that the constant mean is unknown, which is reasonable in our case). This makes (at least) five interpolation-method options (at the limit of computational feasibility together with the other demands). We intend to decide what we really formulate into the revised manuscript based on the preliminary results. We are in particular interested to this end what the reviewer's opinion is as to whether the "IDW2 plus IDW1-vs-IDW2 and IDW3-vs-IDW2" alone wouldn't be sufficient to learn the essential effects of how the degree of spatial spreading of station rainfall data would impact the runoff results.*
- *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).*
- *In addition, regarding these ten subnetworks sizes, 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) in order to also 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.*

With all these improvements, that are the best-possible tradeoff also regarding limits of computational load and efforts, we hope that the manuscript could potentially be recommended for publication since in this case we can more robustly and valuably help to better understand the sensitivity of high-flow runoff to highly variable rainfall records and to interpolation choices.

We are interested in the reviewer's opinion as to whether he/she considers that these are adequate improvements and so encourages us to pursue a revision along these lines.

Some specific comments:

Line 150. What is the reason to calibrate the model with an objective function that includes both NSE and KGE? Why not simply using one of the two?

The objective function NSE is more considering the peaks and the KGE more the whole water balance. Both of these aspects are important, especially if we want to analyze different types of rainfall runoff events (short-duration/long-duration). While calibrating the model, we also realized a different behavior of the objective function e.g. one parameter set lead to an improvement of the NSE, but a decrease of the KGE, or the other way around.

Lines 156-157. The fact that the model is calibrated using IDW2 is not affecting the results when compared later to input using IDW3 and TP? I would expect the model to be calibrated to each input separately to reduce the errors emerging from the model parameterization (as your goal is to discuss the errors emerging from the model input).

Thank you for the comment, we had the same thoughts, but comparing the model efficiency for the calibration period (Mai to September 2009) and validation period (Mai to September 2010) for all cases, the NSE value and KGE are almost the same with very little deviation to the calibration case (IDW2 158 Stations). → We will make this clearer in the manuscript and add the table to the appendix.

L 155 – 158: Because of the necessity of such manual recalibration, the model was calibrated with the IDW2 interpolation and 158 precipitation stations and not recalibrated with all different precipitation inputs and interpolation schemes. This setup is assumed to capture the spatial variation of precipitation in our study area. When comparing the NSE and KGE values for all cases, the deviation to the calibration run was found very small, with deviations for the calibration/ validation period exhibiting a maximum deviation of 0.02/ 0.07 in NSE and of 0.05/ 0.08 in KGE, respectively.

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		Validation Period	
	Mai - Sep. 2009		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

Section 3.2.2. I suggest also comparing the rainfall events using some statistics. For example, how similar how the short rainfall events in term of their spatial variability? You can, for example, plot a graph comparing the spatial autocorrelation of the storms.

Thank you for this suggestion. We intend to compute temporal and spatial variations using the standard deviation of rainfall amount (both normalized to the total rainfall amount) and spatial variation at the peak hour (normalized to the peak hourly rainfall amount). These numbers will help to compare short-duration versus long-duration events, but do not come into troubles with robust auto-correlations across events with varying duration.

Section3.2.4. Why using only a single metric to analyze the runoff? Consider adding other metrics, pointing on the timing of the peak runoff, total runoff volume, etc.

Thank you also for this suggestion. We had a short look to the timing of the runoff peak, but just saw little deviations for this specific event. 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. We will also have a look to the total runoff volume and further investigate such different metrics to strengthen the robustness of the results.

Lines 405-406. One option to overcome the model default is to compute the IDW2 outside the WaSiM model and a second option can be changing the code of WaSiM to compute the IDW the way you choose (if I recall the model as an open code policy).

Yes, we also see this point and checked again the possibilities. Now we found an option to separate the areas for the interpolation, including a core area/zone that corresponds to the WEGN region and the area/zone around it, which is not covered be the dense network.

As a further improvement of the study setup, 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.