Thank you Prof. Beven for the quick response and taking the time to write such detailed comments. One could only hope to get useful critical comments. We agree with some of the concerns that you have raised and not so much with the others. Nothing a nice discussion won't solve. Some points were not conveyed properly in the first submission, these will be rectified. Following are our responses. We address the comments published in HESSD first and then the ones on the manuscript. In **black** are your comments while the ones in **red** are our responses to them.

1. *I am always a bit wary of papers that compare models with models in order to constrain the uncertainties involved in the modelling process. I fully understand why they do so – it is so difficult to get any grip on the real uncertainties in model inputs and "observed" stream discharges and other sources of epistemic uncertainty. But the question is whether what is learned is more than what is lost by not addressing the real problems of model uncertainty directly.*

   We don't use just any model, this model has precipitation inverted in order to match the observed flows. Assuming by model uncertainty you refer to all the uncertainties encountered while modeling rainfall-runoff, we believe that this cannot be assessed. You said it yourself, epistemic uncertainty is always there. Simultaneously adjusting model parameters and input data is an ill-posed problem. Any ideas are welcome that address these issues and would greatly benefit the entire hydrological modeling community. Assuming you are referring to model uncertainty strictly, this paper brings attention to the other part i.e. input data uncertainty. Some hydrologists (Kavetski, for example) have tried to bring this to our attention but the rainfall-runoff modeling community seems to be obsessed with model parameter uncertainty. Running models with uncertain data along with uncertain parameters is certainly not the norm. A day doesn't go by that we see studies dealing with model parameter sensitivity while completely ignoring the other part. We are not disputing the problem of uncertain model parameters. We just want others to see that it is not the only problem that needs our attention.

2. *Here the reference model is the SHETRAN model driven by a relatively dense network of rain gauges. How snow accumulation and melt is handled in the observations is not made clear. SHETRAN is described as a physically-based model, but its physical basis is wrong, particularly at the 1km grid scale used (Richards equation with no account of sub-grid variability, effects of assuming effective parameters at the grid scale). It was already described as a lumped conceptual model at the grid scale nearly 35 years ago (Beven, JH1989). It is not made clear how the daily time step is implemented for the reference model. It would have to have a smaller internal time step, so are the inputs simply averaged over the day? And any surface runoff production surely occurs at much smaller scales than 1km? Does that not already suggest and expectation that the reference will already underestimate peaks in unrealistic ways (unless compensations exist, for example in the routing)? As a virtual reality this does not affect the current study (there was no need to see whether this was a valid model of the catchments) – but as an example of the hydrological expectations we might have BEFORE making such a study, it is surely relevant.*

   The German Weather Service (DWD) reports all sorts of precipitation as water equivalent in mm. A flag is also provided for the type of precipitation e.g. snow, hail or liquid precipitation. It is not so bad to assume that anything falling under zero degrees Celsius is very likely to be snow. Just to be sure, we checked the conditions before the selected peak flows for snow melt for HBV. Except for one, all the other events are due to liquid precipitation i.e. temperatures significantly above zero. The one that does have snow melt is a mixed event (the third figure). Figures of these are shown at the end (Fig. 1, 2 and 3). We considered two weeks before the event and one week after it. Thus, the snowmelt processes are practically irrelevant for our research. We are well aware of the shortcomings of the model. SHETRAN has come a long way since then (the core developers can chime in at any moment they please). Coming to its physical basis being wrong, we are still waiting for the one whose physical basis is right even if we get that one, how would it account for the consistent underestimation of precipitation volume? One could argue that a model with wrong assumptions would produce wrong outputs and wrong conclusions would follow but then, all models are wrong. The choice of a model would not make a difference to the final conclusion of this paper. One can take any model. We agree that surface runoff production happens at a scale less than daily and 1km resolution. Personally, we would like input data that is observed at the microsecond scale and observations with a nanometer resolution and an accompanying computer that can handle all the processing. Till then, we have to deal with what we have. Coming to the point that averaging will underestimate the peaks. Yes, it has a smoothing effect but that was not the point. The averaging will happen for the reconstructed reference precipitation. And this is the better model for this catchment, because in our previous studies we inverted the precipitation to match the peaks. There, we performed the inversion using SHETRAN and HBV. To assess whether the inversion is realistic, we switched the inputs for both models to see whether the performance dropped. It didn't, in any case, and always improved the performance, leading us to believe that inversion was not an overfit or model specific. The point here is that as we use less stations, and then interpolate, we get systematic underestimation of precipitation and thus of peak flows. In some instances it could also overestimate, of course, but the bias is much more towards the underestimation than overestimation. Furthermore, here at the institute we have had, and are still having, studies that compare model performances with varying temporal and spatial resolutions. And as expected the higher resolution models outperform the coarser ones. But not by enough for us to conclude that peak flows underestimation is explained by the higher resolution. We believe that is because all of them have one thing in common. The interpolated input data. For a finer temporal resolution input data are even more uncertain and the systematic underestimation for sparse networks is even worse.

3. *And talking of expectations, it is stated that the subcatchments of the Neckar used were large enough to allow for daily time step simulations. The hydrographs shown in Figures 9, 10 and 13 really do not support this. Again, I understand the use of a daily time step so as to maximise the precipitation stations available but the discretisation effects would suggest that this is surely not properly reflecting the hydrology of these catchments. Equally HBV is run at the 1km scale, but it appears as if the outputs from each grid square are simply added, with no account taken of distance from the outlet (and the carry-over from day to day this might produce).*

   From the statement about being large enough, we meant that peak flows show up the next day after a high precipitation event takes place. We consider this to be good enough for practical proposes. Also, we do not use a post smoothing filter on model runoff to make it look nicer. Ideally, we should have modeled them at the 15 minute resolution but the data are not reliable at that scale, even if we could somehow circumvent the remaining problems. For the HBV, yes the output is just added for all the cells, we are aware of the distributed HBVs with routing. This has been tested and the results (as expected) did not change enough for the people who tried this. Furthermore, we wanted to give HBV the full freedom to adjust its parameters to match the output, as was stated in the manuscript. Imposing cell to cell routing will only constrain it more. It would be more correct to do so, this would have been the next step, had the one without routing perform well enough to match the reference discharge. But it did not. One could argue that cells get saturated more and more as we move downstream, but previous studies do not show any significant difference. Again, we agree cell-to-cell routing is more correct.

4. *So we are left with the conclusion that a model might get better (in representing an incomplete virtual reality) as the number of gauges to define the inputs increases. A model using spatially distributed inputs will generally do better than one using an average input (all other things being spatially equal, which of course, in general they are not). But we knew that already, so have we really learned anything new about WHY our models "keep underestimating peak flows"? Except that, as shown in Figure 13, they do not – the global recalibration of HBV, given a set of inputs, can overcompensate for some storms.*

   The reality used here is incomplete but it was good enough to reproduce peaks that were otherwise underestimated by interpolated precipitation. Coming to the point of if we really learned anything, yes we did indeed. Areal precipitation is systematically underestimated for low density networks which leads to an underestimation of the peak flows. The underestimation is not always the case but is much more frequent than the overestimation. The underestimation of precipitation does not occur for most *normal* events. Figure 13, is there to show that even for extremes this is not always the case, we put it there deliberately to support the *other* figures (that have been not commented on). Again, looking at the violin plots we see the direction of the bias, which is clearly towards underestimation. And yes, some storms are over compensated, but in the same time series most are not. Speaking of this catchment (Enz) specifically, the maximum discharge recorded in this time period is about 400 $m^3.s^{-1}$. The one shown here is 150, it is still among the highest but not the highest. The recalibration probably tried to get close to the biggest one and in the process increased flows for the smaller events.

5. *We know there are subtle interactions between different types of model calibration, time steps, routing methods, parameter sets and their complex interactions, and sources of epistemic uncertainty, including both inputs and the rating curve (this study eliminates the discharge uncertainties by design but these can be important in practice). So if we already have an expectation that, in general, catchment averaged or poorly sampled inputs will lead to a tendency to underprediction of peaks, then the question that the paper should be addressing is what to do about that in real applications (where the discharge uncertainties also come into play). We cannot invent input data, so we will normally use as much as is available (which would be even more than the 150 gauges in this study). We know that the calibration will depend on, and compensate for, the limitations of the observed data that are available – but the smaller the sample then the greater the resulting uncertainties in the predicted discharges might be (see figure 13 again). This paper does not address those uncertainties (except in comparing the 5 samples at each density). It does not even make use of the uncertainties associated with the kriging interpolation (which would seem to be a good reason for using the kriging interpolator, despite the necessary assumptions and requirement to already have many gauges to estimate variograms).*

This study does not eliminate discharge uncertainties by design. We intended to drive the attention to the underestimation of the largest events due to a systematic underestimation of the corresponding precipitation input. We believe that for correcting errors the first step is to recognize them. A simple correction by a multiplicator seems unrealistic as *normal* events are well represented and not underestimated.

Originally, this paper was supposed to address the issue of model calibration too. But then we found that it is better to handle input problems first and then deal with the compensation (if it still exists) and the other problem was that the paper was becoming too long when the compensation part was added. Furthermore, we also found that models calibrated on different networks for the same catchments have systematic problems as well. These will be dealt in much more detail in future papers. If systematic input problems are compensated by model parameters via calibration then the users face serious problems. The models calibrated to compensate systematic precipitation errors are specific to the observation network and will fail in case the network changes (new stations added and/or others removed). These models should not be used with meteorological forecasts unless their bias is adjusted to the calibration network. The same problem occurs when using the model for estimating climate change effects.

Coming to not addressing uncertainties, the paper is about peak flows, so we took the biggest five. For low flows, we have different problems such as treated waste water. A different can of worms. The scatter plots show the whole story i.e. underestimation of precipitation. To be clear, the paper does not deal with all sorts of uncertainties (we stated that clearly in the beginning), just the peak flows. Coming to Kriging, the first author has been researching this since the start of his career. The uncertainty of the areal precipitation can be quantified using kriging, but is unrealistically small. Further then problem is not the uncertainty of the areal precipitation but its bias (frequent underestimation). We chose two different approaches - the observation data in which case we do not know the true precipitation and a spatially simulated case. For the first case we demonstrated that while using the same interpolation method the lower density networks frequently estimated less precipitation than the denser network. In the simulated case we were free to use different networks and could compare the results with the virtual reality. Both cases lead to the same results and the underestimation was exaggerated by the hydrological model. The results are not specific to ordinary kriging, same goes for external drift kriging, co-kriging, splines to name a few. All of them suffer from the same issues. In an unpublished study the first author investigated the same problem in South-England with the same type of results, frequent underestimation of the extremes and a few severe overestimation. Ideally, we would like to have an interpolation scheme that results in a zero mean bias for the precipitation corresponding to the peak flows or anywhere else, regardless of the gauging density.

6. *So I do not think the paper can be published in this form. As far as I can see we cannot use any of the results to improve practical applications (other than a general exhortation to use as many input gauges as possible and try to take account of spatial patterns – but even then the common problem of not having gauges in higher elevations has not been mentioned and there is a lack of information about how snow is handled). I would suggest it needs to be more complete and more ambitious and address the question of IF we only have a certain density of gauges available (remembering that in practice we cannot resample from a larger set), then how should the modelling workflow compensate to get better estimates of the (uncertain) peak flows. Does calibration provide sufficient compensation? Do the acceptable*

parameter sets change with the input scenarios? Can the authors suggest "uncertainty multipliers" as the density of inputs decreases (but that requires estimation of addressing the simulation uncertainties more directly)? Does this vary between models (though I understand the problem of calibrating SHETRAN if it was used in the comparison)?

But this paper has results that we can use, i.e. we are systematically underestimating precipitation and our peaks and yes, we need to do something to make up for the missing volume and the answer does not lie in adjusting model parameters only, which is what is normally done. We can ask the same question from everyone, if we know that model results are bad due to few data points then why is the emphasis on adjusting model parameters and not on improvement of model inputs? Another important conclusion is that places with good gauging densities have less to worry about systematic underestimation than the ones with a lower density as shown in all comparison figures. For them, the better density ones, the model outputs came very close to what they originally were, even though the points were not that many. In our opinion, this is no trivial conclusion. Looking at the location of stations in the study area, we think the problem of few gauges in higher elevations does not exist for our case, but yes, having fewer gauges in the higher elevations is a known problem that causes mass balance mismatches. Just to sure, we checked the elevation distribution of the gauges against the simulation grid. Fig. 4 shows that we are safe. For the simulations, the sampling of points were uniform (but still random) over the grid. Also, as we showed in IAHS this year, for our case, calibration compensates but not enough as is again evidenced by the violin plots. If a model is mass conservative, there is no way to make up for the missing volume and consequently match the peaks. While using NSE, the HBV parameters go all over the place, but we haven't encountered any bad ones. Uncertainty multipliers would be a start, but that is like putting a band-aid on an open wound. We think that many minds need to come together to solve this issue. The ultimate solution is more observations which is something only governments can afford and maintain. We can use any model, it is not a coincidence that both models produced similar biases. We knew from start that missing precipitation volumes would cause problems at the end. This paper serves as a concrete example and a reference that will be referred to in our upcoming studies. The future research questions listed at the end of the manuscript are what we are working on right now. Each one will be addressed, if possible, slowly. Surprisingly the measurement errors - if they are unbiased do not lead to a further systematic deterioration of the estimations. Thus effort using opportunistic sensors for precipitation estimation may help us quite a lot. In a recent study (in a German scientific journal, HyWa) of the 2021 floods on the Ahr in Germany we have seen that the much denser private weather station network lead to a higher and more realistic areal precipitation estimation than the official weather service network.

Now coming to the comments in the manuscript.

7. **Line 63:** Also HSJ 2016 and RSPA 2019 Inexact Science papers.

   We took the latest paper that was available. Normally, we prefer to keep citations to a minimum. We are against the trend to have a list of references rivalling the length of the paper. Besides, HESS charges per page and we have a funding limit per paper. One can always google and find something, or at least have a starting point. It is interesting that you mention "Inexact Science", we will come back to that at the end.

8. **Line 64:** Though there are much earlier works concerned directly with patterns of precipitation on model outputs and peak predictions that are relevant - e.g. Beven, K.J., Hornberger, G.M. (1982), 'Assessing the effect of spatial pattern of precipitation in modelling streamflow hydrographs, Water Resources Bulletin, 18(5), 823-829.

   We were unaware of these. Thank you for bringing this aspect to our attention. We will add the findings of this and similar relevant studies to the revised manuscript.

9. **Line 72:** Handling of snow in precipitation records? No account of correlation of precip with elevation - except as recorded at specific sites?

   As we said before, choice of interpolation scheme does not make a decisive difference. It is only slightly better at best. If something like External Drift Kriging would have helped, we would have taken that. Besides, we are using reconstructed daily precipitation. It has very little correlation with topography on a daily scale. It shows up on monthly and annual scales only. Out of the peaks considered only one was partially related to snow, thus it does not influence our results.

10. **Line 100:** Stable and variogram with small numbers of points is a bit of an oxymoron esepcially on a time step by time step basis (though it is not stated yet as to whether that was how it was applied). - this is surely a source of epistemic uncertainty in itself?

    Please read the method in the reference, it does not use a few points. Otherwise, we wouldn't have called it stable. And yes, it is one source of problems.

11. **Line 108:** But that is points - and you are creating a grid, so do you allow for reduced variance due to change of scale and loss of any nugget variance in block kriging?

    For simulating precipitation realizations, we had an automatic correlation function fitting (described in Random Mixing). It may or may not have a nugget. For interpolation/simulation, we assume that an observation station represents the whole cell but given that we have 19000 cells, accounting for the nugget won't make much of difference. Further we compare results obtained using the same geostatistical model. Thus, the dense networks are treated the same way as the sparse ones.

12. **Line 123:** 10 for each N is not many? Not totally clear?

    10 multiplied by 5 is 50. We can do more if you like. Again, for the interpolation with stations in Sect. 5.2, we have sampled 100 times. It is not a coincidence that both cases (sampling from reference and sampling from all observation stations) showed underestimation of high precipitation values.

13. **Line 126:** Not really, surely. One advantage of OK (or co-K if you including some elevation or other effect) is that it allows for the variance to be estimated at each point or grid (as Block Kriging). That would then allow for the generation of consistent random fields with higher values (even if our expectations might be that they might not be extreme enough given the scale of rainfall cells and the smoothing effect of kriging on variances).

    Again, we have no information about the very short distance variation of precipitation so we assume that point to cell transition does not make a big difference. The mean is not influenced, and the problem is the mean. Further the kriging variances are usually not a good representation of interpolation uncertainty. This has been discussed extensively in the late 80-s in the geostatistical community. Again, the effect of topography alone have practically zero effects on the daily precipitation estimates. This has been addressed/discussed by the first author in previous studies.

14. **Line 127:** Statistically that is surely correct because you are integrating (comment about block kriging earlier). But it is not the only possible set of assumptions if you have an expectation that the sample underestimates the true variance or that there is a (complex) elevation correlation in the headwaters etc. Why would you choose to reduce extremes if you think those extremes are important?

    Yes, we can assume differently as we know that the high values are underestimated. We thought of a something where we could take the bias due to a sample into account to find a better estimate of the distribution for the whole area. Random Mixing does this implicitly where it can easily simulate value beyond that of the given constraining points. The main problem is not that. It is that we use an interpolator.

15. **Line 147:** Yes ,but the results will depend on the time step that is used (clearly not daily) especially for any surface runoff production (if any), and how inputs are distributed for any subdaily time steps.

    Yes, they do depend on the used time step, which in this case would have helped us with better peaks. A: we do not have better data. B: All the models do the same here. If we change the reference model then we change all. The point here is that we get a consistent underestimation of peak flows. As is shown in the violin plots.

16. **Line 154:** No routing at all???

    As discussed above, this does not bring much or answer the question of the missing precipitation volume. If anything, it will constrain the simulation even more. But we agree, routing is the way to go.

17. **Line 218:** what happened to 10 as the coarsest?

    That was for the observation stations' case where only precipitation underestimation was shown. Here, we sampled 25 points from the reference precipitation.

18. **Line 219:** an odd phrase - do you mean infiltrated into the soil? But would not all the precip infiltrate in this region given your "natural" soil properties (i.e not urban or other impermeable surfaces) anyway?

    Yes, we meant infiltrated into the soil that did not make it in time to the peak flow.

19. **Line 256:** But that is equivalent to a nugget variance at the point that should theoretically be integrated out in block kriging to the 1km sacale?

    Yes, this will create a nugget. We did not check how much. Variograms are recomputed each time a change is made.

20. **Line 263:** why is this a given? Exposure of standard raingauges has an effect on what is measured...

    Here *given* means *if*. We are currently looking at how this error is distributed. If it is not Gaussian, then we have additional problems to address before we can use the personal weather station networks, like you said with wind and all.

21. **Line 282:** what is correct - you have used 1km scale but saturated areas are at much smaller scales - while Richards equation does not actually apply at such scales?

    Correct depends on context. Here, we meant to say enough gauge density. As for the problem that 1km scale where Richards equations does not apply, you did not mention what you would prefer and why? For catchments so large, it does not really matter if we model at 1km resolution. The entire cell gets the same precipitation and has the same elevation. Either the whole cell is saturated or its not. We would like to see a model and the accompanying data that resolves the soil at the molecular level, because one can always demand more complexity. The point here was to show how using fewer gauges would cause systematic biases, regardless of what model is used. We think this point was proved conclusively.

22. **Line 285:** Not only is this not given, but why is it necessary? That is a problem as to what uncertain information such gauges bring but you do not have to rely on standard statistical assumptions when other issues might be important (wind speed effects on catch and snow drift have not had a mention here?)

    Again, given means if.

23. **Line 288:** Again a consequence of the statistical assumptions you have made - but others could be made to more reflect the epistemic nature of the uncertainties. Given that you KNOW you will get biased predictions given this set of assumptions should not you explore other sets of assumptions even if (for good epistemic reasons) they might be more difficult to explain and justify?????

    Yes, we know that we need something more to address the issue. This paper is a starting point. We are not stopping here. This one will serve as basis for advocating for better/denser observation networks because for some unknown reason the gauging densities are not what they used to be here.

24. **Line 303:** The references are a mess - put into standard format for HESS

    We are using the HESS template. This is what came out.

25. **Concluding remarks from the authors:** Coming back to the rainfall-runoff modeling being an inexact science. Most of the comments made by you demand theoretical and mathematical pureness. Pureness which doesn't help much in practice. We just use it as a starting point. For example, take Kriging. It makes assumptions of covariance being isotropic. In reality, when has it ever happened that a precipitation field at any given time step fulfills such conditions? How about other interpolators and their assumptions? Maybe temperature under special circumstances fulfils some of the assumptions. Same goes for pretty much all variables. Same goes for HBV with its separate soil and routing modules. This didn't stop it from being one of the most widely used models. We use a mass conservative version here. Hence,

the state of its internal variables (soil moisture and reservoir water depth) do point to the occurrence of peak flows when using forecasts as inputs, but still underestimate. There is a reason why conceptual models exist and are used in practice. This is why we play loose with the rules because we know that if we stick to them, outputs of our models are guaranteed to be wrong and less useful. The second author was negatively surprised upon finding out that the physically-based models are not exactly what their name suggests. They are incorrect representations but we can tweak them to be useful enough. Models run by governments to monitor and forecast floods use similar resolution here (they have failed on occasion though). That does not make 1km and daily simulation theoretically right either but they are useful and that's what matters. But you already knew that and we think that the majority of the hydrologists would also agree.

26. **A quick and temporary solution from the authors:**

During these experiments we realised that while using the nearest neighbor (NN) as an interpolator. Sometimes, it does produce discharges that are higher than the observed, as can be seen in the violin plots and the figure 13 that has been referred to multiple times (kriging does it too but no so often). If possible, we could use random subsamples of stations and then interpolate. By doing this many times, we get many inputs series. By using these as inputs, we are very likely to get peak flows that are higher than the observed. This is again playing loose with the rules, but it works. We can see this as a limited input data uncertainty analysis. Maybe someone tried this already. And this won't work in areas with a small density of gauges. Another slightly worse solution could be to transform the distribution of modeled discharges to the observed ones. The transform can be updated as new observed data comes in. There will be a problem with the minimum and maximum values as the transformation may under- or over-predict when going out of bounds. This has to be tested though, as peak under- and over-estimation is a function of where precipitation took place, among other things, as you showed a long time ago. Like we said before, a separate paper will deal with this, given you agree with our responses. Combining all these into one is just too much.

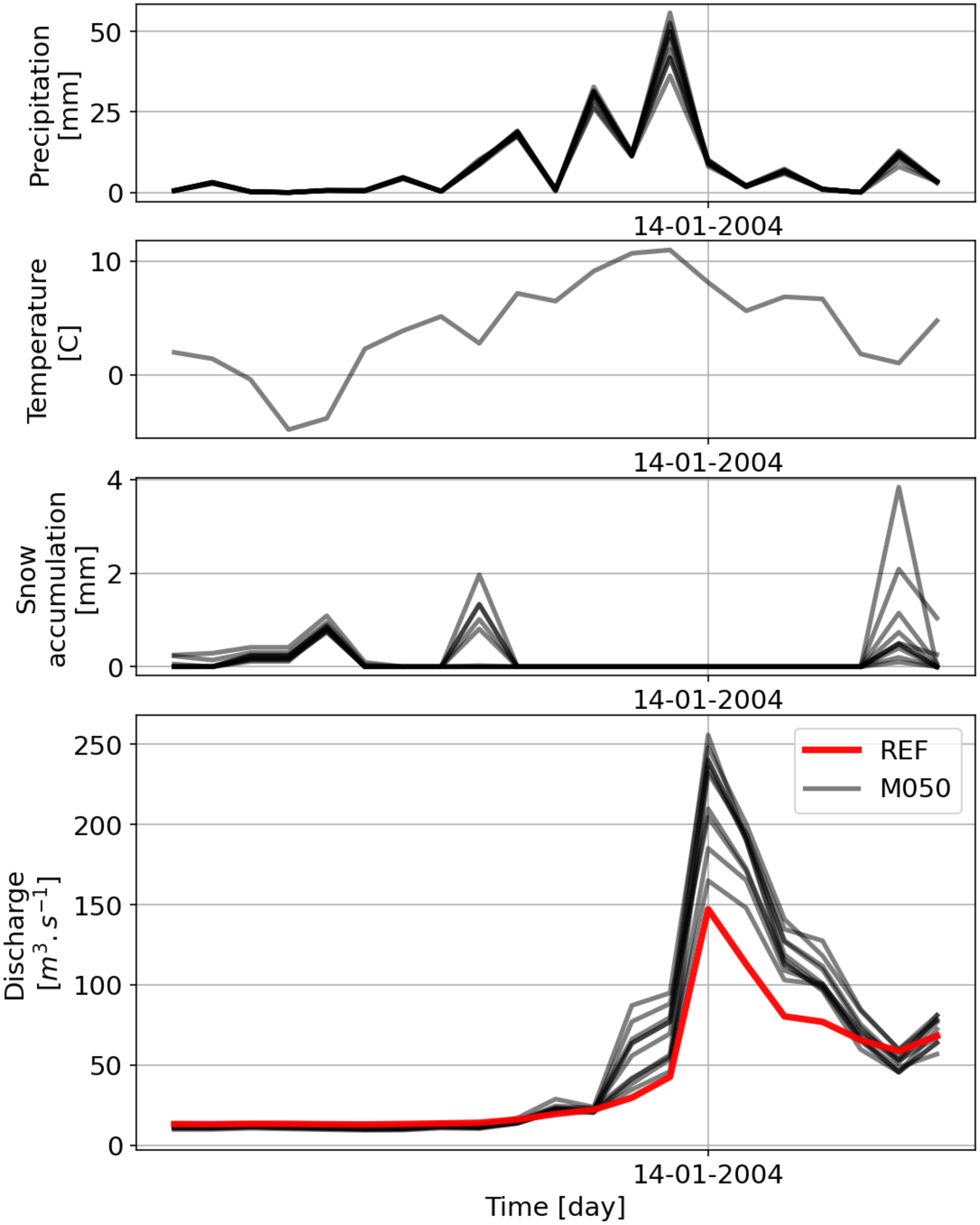

Figure 1. Event hydrograph comparison for interpolation with 50 points from reference using HBV.

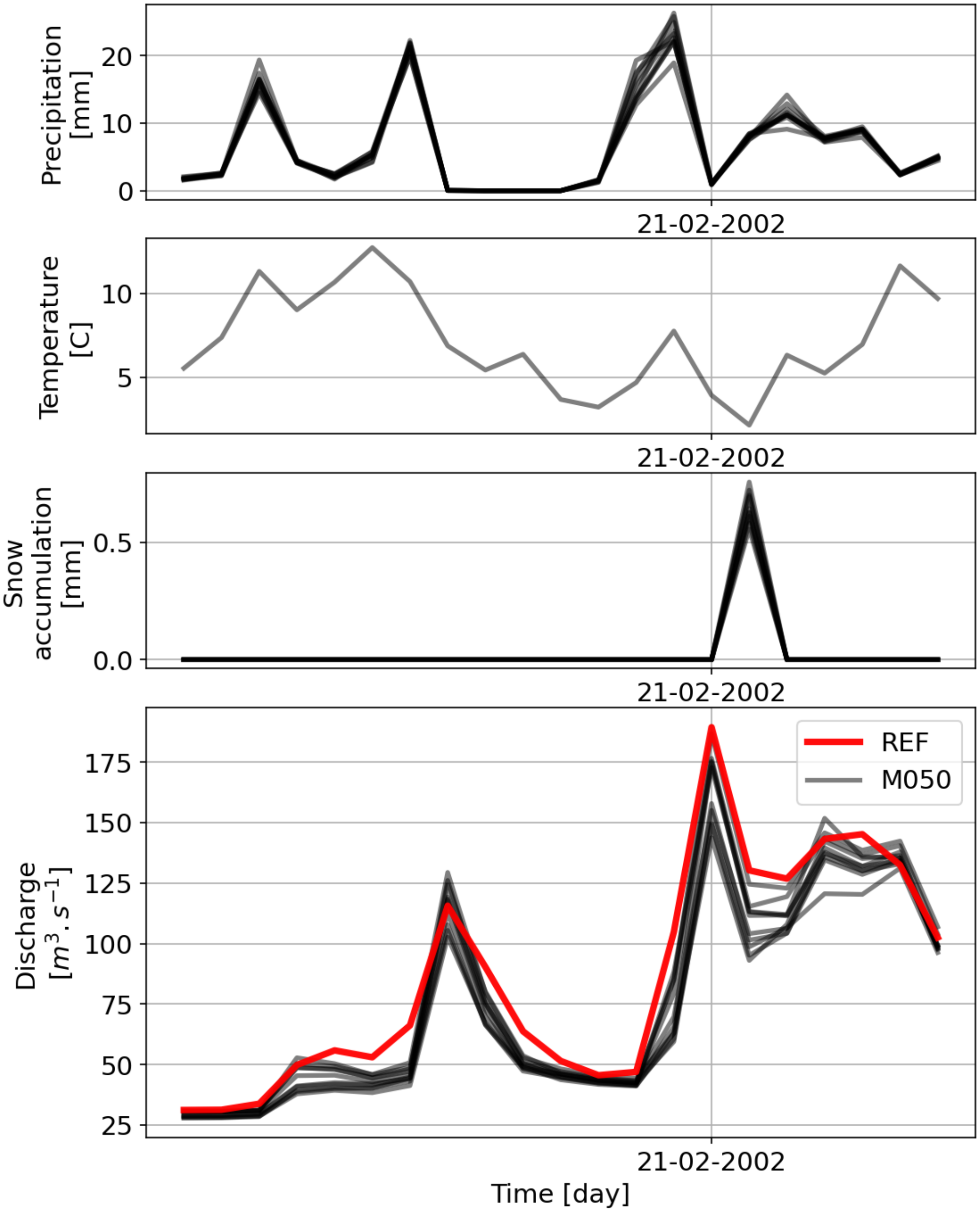

Figure 2. Event hydrograph comparison for interpolation with 50 points from reference using HBV.

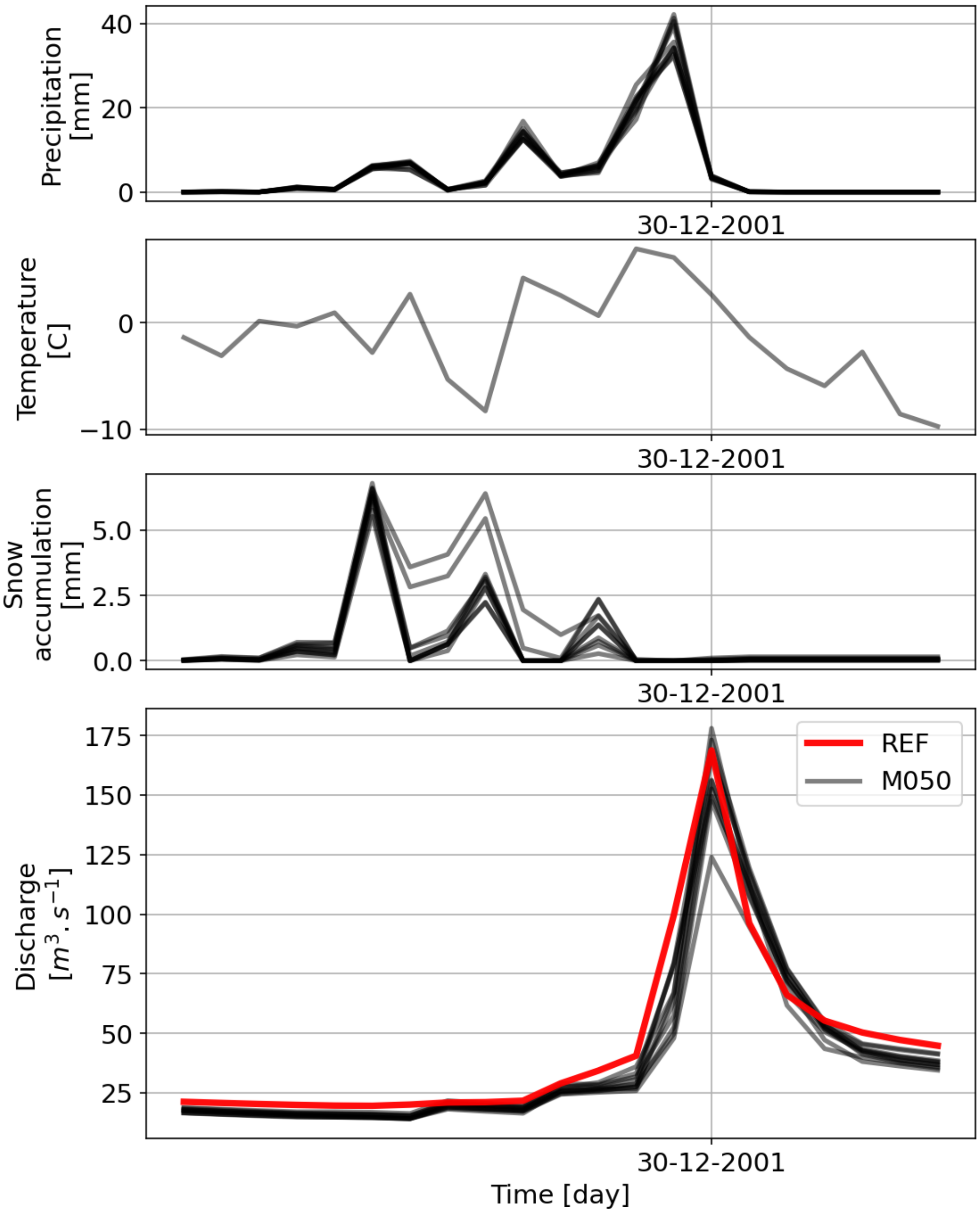

Figure 3. Event hydrograph comparison for interpolation with 50 points from reference using HBV.

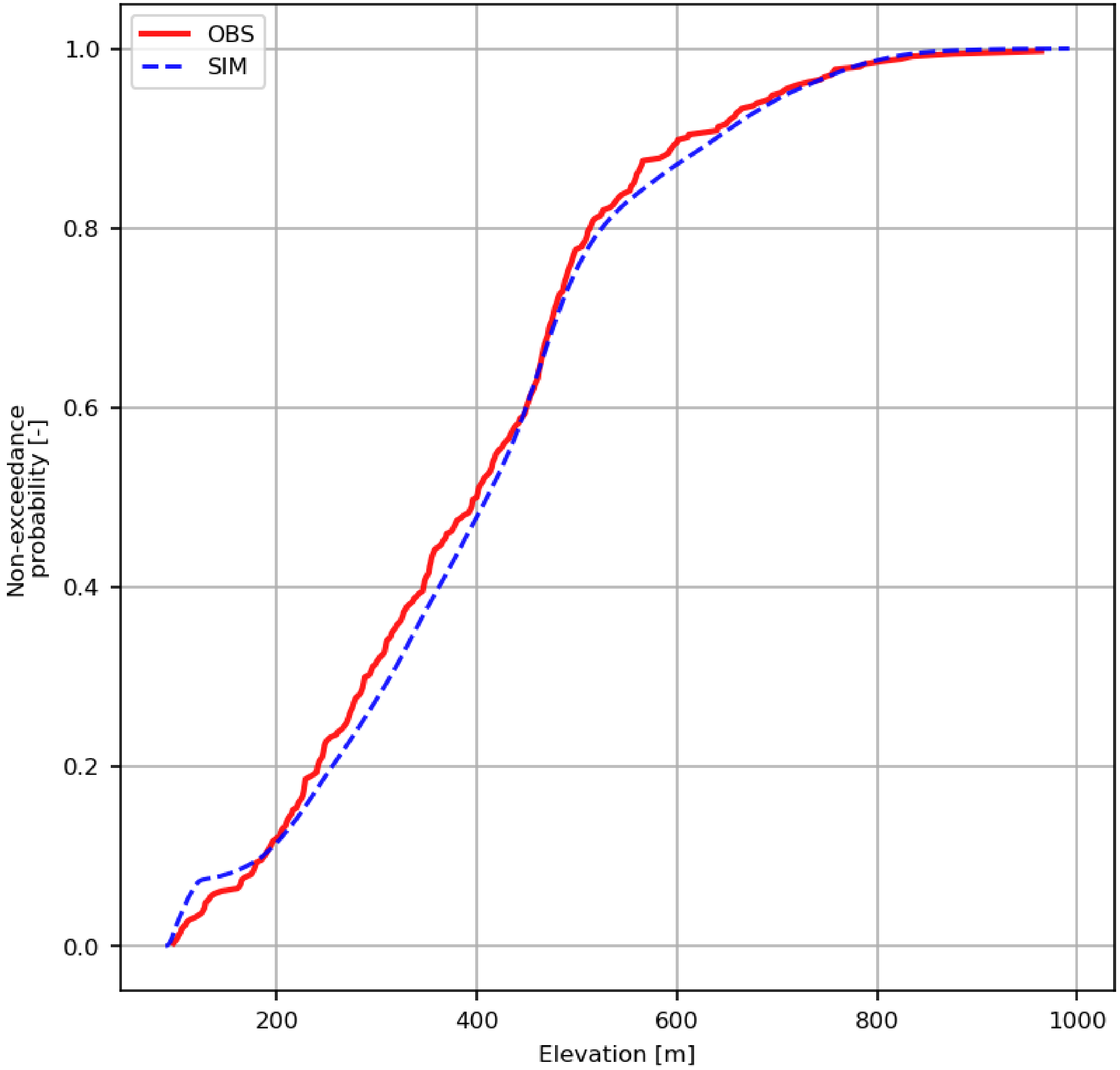

Figure 4. Comparison of elevation distributions of observation locations (red) and the whole simulation grid (blue) using the SRTM 90m grid for the study area.