Thank you again Prof. Beven for the very quick response. We have considered your comments and added some more clarifications to the main script. Regarding the comments, we agree with what you propose to some extent. Same as before, in **black** are you comments and in **red** are ours.

1. *Just a couple of points arising from your comments. You say that: We don't use just any model, this model has precipitation inverted in order to match the observed flows. Inverted rainfall is not mentioned at all in the paper. Only that the gridded rainfall is produced so as to match the observed point rainfalls. Nor does it seem to be necessary as the "observed" flows are produced by SHETRAN (with its deficiencies and need for some clarifications I mentioned).*

   We guess there was a confusion arising from what we mean by inverted precipitation. Just to be clear, it is the one that is iteratively simulated to produce model discharges that closely match the observed ones. This does make a difference compared to using interpolated preciptiation, even if we use SHETRAN as reference as it helps with producing dicharge that is closer to reality. We have expanded the reference precipitation section. One way of putting it is that we produced gridded rainfall that produced runoff that matched the observed flow. We don't know how to run the model backwards.

2. *Secondly, I am not sure I mentioned parameter uncertainty anywhere in my comments (except to ask if the acceptable (or in your case calibrated) parameters might vary systematically with the input resolution). I am not sure that the hydrological community is obsessed with parameter uncertainty to the neglect of other sources, but in practical applications (where we would normally use ALL the raingauges available) it is a simple way of generating potentially behavioural models. Different model structures can easily be incorporated as well in a GLUE or Bayesian weighting framework. Input realisations are somewhat more problematic in that it is usually not obvious how to construct them (as you point out Kriging and its variance estimates depend on some rather strong assumptions, and gauges can be spaced more widely than raincells).*

   We explicitly said the rainfall-runoff modelling community. It is our experience, that almost all the time, we see papers that work with, say, model parameter sensitivity analysis or something similar. Even during calibration, we look for the optimum parameters in general and not consider the other sources of problem that may be responsible for less than desirable results. Apart from you and may be a handful of modelers, nobody deals with the other types. That being said, we are guilty of this too (to some extent). It is good that you mention that for practical applications, we take ALL the available gauges. We just show that how gauging density affects the final precipitation field. Here, we would like to bring the attention to the point that the interpolated precipitation is likely to underestimate large events, something which cannot be corrected by models. The problem is that we cannot tell in advance how the under- or over-estimation will be. Only that on average, we will underestimate. Kriging is problematic but there exist other algorithms (e.g., Random Mixing and Phase Annealing.) that do alleviate the problem but they produce, theoretically, infinitely many fields.

3. *Your quick and temporary solution suggests a number of ideas but I am not sure they help much towards what we should do about the problem in practice. You ask for ideas – my analysis of the problem is that we know that interpolated rainfalls might be biased towards underestimation (especially where upland gauges are missing) and that traditionally model calibration (alone) or including rainfall multiplier parameters in the calibration (with a history back to the Stanford Watershed Model) has been a simple way of trying to compensate implicitly for that. That expectation of compensation extends to any errors there might be in the discharges observations (particularly at the highest and lowest flows), conditional on the model or models used. As noted above, allowing that an ensemble of behavioural models might result from that compensation is a further extension that can be used to produce some range of predicted outcomes that (again implicitly) reflect both sources of uncertainty. The analysis of runoff coefficients in the Inexact Science paper is another reflection of the joint uncertainties (there are a couple of follow-up papers in review that make use of that approach).*

   You are right. We do not know as of yet what to do either. We have tried non-linear monotonic transformations for bias correction using various assumptions. But the bias is not the only issue here. Each event behaves differently. Many are underestimated but then some get overestimated. There is no way of knowing this in advance at short temporal scales, say hourly or daily, how the precipitation field pattern will look like (no, not from radar) and its effect on the final field that is interpolated. We are investigating this right now. The problem of upland gauges is indirectly covered by our study

as well. Although, not having gauges at high elevations is more of a systematic problem. Also, we prefer not to comment on event multipliers. Regarding the analysis of runoff coefficients, we would again like to draw the attention to that case when all of the gauges on a given day significantly miss the high precipitation region. Interpolating it (IDW, NN or OK) will result in a much smaller volume, leading us to believe that runoff coefficient is wrong (when it is more than one). If you think about it, how can it be greater than one? Is there a siphoning effect of the outflow? We are not sure. For us, it is more likely due to a small gauging density.

4. *But, as you demonstrate, the compensation of calibration does not always produce an underestimation of flows. BATEA etc have revealed how this is (as you note) an ill-posed problem, with resulting huge (and unrealistic?) variations in event to event rainfall multipliers when the compensation starts to allow for model structural deficiencies and consequent antecedent conditions from event to event. These are certainly all forms of epistemic uncertainties – or put it another way, even if we use all the raingauges available we cannot know what effect the interpolation will have on individual events. We can only make assumptions about what that effect might be. So to pose a hypothesis relevant to your paper: are there any systematic biases that can be detected in the reduced rainfall networks (or their effect on the discharges) that might be used to inform the model simulations (or more generally constrain their uncertainties).*

   We agree that input and parameter uncertainty should be somehow considered at the same time. But this would definitely result in a mess. To answer the question about the systematic biases. Yes, this is what the end results show partly. For an existing network, we won't know how much is due to the model and how much due to the data. Only that the on average, we should expect reduced interpolated precipitation because that is model independent.

5. *This is a more challenging, but also more important, problem. I think I would approach it somewhat differently. I would eliminate the SHETRAN virtual reality – yes it is mass-conserving but that is not really relevant for practical applications for which we are trying to reproduce the observed discharge. I would try to assess the uncertainty in the observed discharges and allow for that in the model evaluations in some way. I would generate different rainfall realisations based on the samples of raingauges and compare both the rainfall biases (ie. no compensation by calibration) and the predicted discharge biases (ie. with implicit compensation by calibration) with results using the full network. For the kriging case the realisations could reflect the grid estimation variances (though there is still the issue of what minimum network numbers you need to determine a variogram).*

   The reduction in the variability of both inputs and simulated flows as the sample of gauges increases would still be revealed, surely. It might be considered as representative of what to expect in areas with a similar distribution of gauges with elevation (as demonstrated in the hypsometric curves in your reply). I am not sure it would apply in our Cumbrian catchments where we are much more deficient in higher elevation gauges (again there is a paper in review on this looking at different interpolation methods – though this was not extended to the type of sampling study you have done). So that might provide a range of potential outcomes for other applications with similar sampling densities but less gauges (the only problem being that we would not know where in that distribution that actual particular sample lies... though perhaps looking at the effect on the simulated outputs might help there even if after the compensatory effect of calibration – that is something that you could look at).

   We agree with your ideas and would follow them in the future, if possible. As we said before, this paper is too long already and is meant to bring out attention to the problem of interpolated precipitation. It is true, we don't need SHETRAN necessarily it could be any other model. The problem of assessing precipitation uncertainty is not trivial. During this time we tried different ways to see find various causes or variables that were related to increased uncertainty but haven't found anything that consistently explains the underestimation for all the considered catchments here. Again, the problem of not having gauges at higher elevations is what this study also points to, it is similar to missing the high precipitation locations consistently.

   Regarding considering the uncertainty of discharge; in our experience, we consistently underestimate the values in the upper tails when we look at the discharge of a calibrated model against observed, even when using Nash-Sutcliffe (it concentrates on peak a lot). You must have observed this as well. This creates a problem when dealing with discharge uncertainties i.e. we will always go for values in the lower confidence bounds in the upper tails of the observed. Statistically, speaking this should not be the case.

6. *This might be a quick and more enduring way of going beyond demonstrating the problem towards what we might do about it. I think we would certainly agree that there is no better solution than getting more raingauge (and discharge) observations but, certainly for historical data, we are often constrained to limited networks so we need some practical solutions.*

   KB

   Yes, this is what we also conclude, there is no way around a denser network, to get rid of the input uncertainty. Then, we can deal with other types.