The reviewer's comments are in italics and the responses in regular text.

It is always interesting to read one of John's papers, and this one is no exception.

Thank you.

Given the current penchant for throwing data into the black boxes of machine/deep learning algorithms and declaring success without producing much in the way of understanding, the more thoughtful approach presented here, in a highly original way, is of value.

The work with such "black boxes" is interesting.

I did find the paper not easy to read in places, with some sweeping generalisations and somewhat limited reference to previous work (see below). And the end result is limited to a really simple non-parametric modelling approach that appears to only reflect the basic minimum of hydrological knowledge, that flows reflect today's rainfall and some index of antecedent conditions (here represented as only by matching the pattern of average rainfalls over the last n days, in a least squares sense without any allowance for autocorrelation in those values).

It sort of works (even better than a calibrated rainfall-runoff model for some of the catchments and mostly not much worse in terms of NSE). It sort of works for transfer from a proxy catchment (but see comment below on this). But there are perhaps some issues of hydrological knowledge that are somewhat glossed over.

(A) When compared against a model like GR4J, KERR does "sort of work". That is a little surprising given it is parameterless and is based on trivial hydrologic knowledge. In fact, it works well enough to be seen as a reasonable RR model, and the experiment essentially shows that a model of a layman can give a reasonable RR model. If KERR was a model of us (i.e. the authors), rather than a layman, its median NSE might be higher as a result of our understanding of many of the technical details mentioned in the review. It is important to note, though, that the aim is not to do the best RR modelling possible, but the best knowledge experiment possible, based on a model of a layman. That might help explain why the work seems to be based on sweeping generalisations and limited references to previous work.

There will be uncertainties in the data. These are mentioned but then ignored. But can be significant – e.g. where event runoff coefficients are highly variable and go greater than 1 (see Beven and Smith, 2015; Beven, 2019).

This implies that it might be useful to allow for uncertainty in the mean prediction – you can after all pattern match to get an ensemble of possible values which could be treated as a first estimate of a pdf reflecting uncertainty.

For convenience, the responses to all the questions about uncertainty are given here.

(B). With hindsight, a brief note on uncertainty should have been included in the Future Work section of the paper (Sect 9.1). The RR records used are a snapshot taken at some time before 2014, when the work started. The records change over time. There is now the CamelsGB set, where a different rainfall product was used and for which at least one of the 38 catchments has a revised

rating curve. Relative to CamelsGB, the data errors in the original set are therefore known and their impact on the knowledge experiment could be calculated. There is a sampling problem in that only 38 catchments were used and a fixed period of time selected. Such things could be addressed in future work.

(C). The terminology used in the paper is that there is "predictive headroom" in that GR4J gives a higher NSE, suggesting there is scope for finding better hydrologic knowledge because there seems to be some unexploited structure within the RR data. "Unpredictability" includes what comes under predictive headroom, plus data errors, plus errors in rating curves, etc. The measure for unpredictability is simply the mean deficit w.r.t. perfection, so is the mean difference between 1 and the NSE values achieved using KERR. That mean is what is in the third row in Table 9, and that row is where the value 2 comes from which is attributed to unpredictability in L18 in the abstract and L425 in the summary. Text to this effect will be added to the results section.

(D) In L236 the number of best days is fixed at 10. In the context of the work there is no uncertainty about this. The value 10 was picked and was found to be entirely successful given that testing shows that there is insensitivity to picking 10. It is possible to calibrate the number of best days or to construct numerical schemes that use ensembles. In the context of the knowledge experiment, though, nothing meaningful would be gained by doing that. The case is the same for fixing the pattern length at 365 days.

The use of daily data is convenient in terms of getting hold of the data but will be subject to discretisation issues in small catchments (where the peak occurs in the dayaffects the volume for that day) and autocorrelation issues in larger catchments (every day is here treated as independent, even on recessions).

We all have to work with the data we can get hold of. In other work, experiments using KERR with sub-daily data worked fine.

Snow is mentioned, but then neglected. It may be relatively unimportant in most UK catchments perhaps, but one of the outcomes from the Iorgulescu and Beven (2004) attempt at a similar non-parametric data-based predictor based on CART also with different rainfall period inputs, was that the classification identified anomalous periods associated with delayed snowmelt. That this might happen is hydrological knowledge is easily stated in English!

An experiment could be run for a layman who takes an interest in snow. Iorgulescu and Beven (2004) will be cited.

For the transferability in space, a brute force approach to finding the best KERR model is taken but checking all the catchments in the data set and picking the best as the donor site. That cannot be used if a catchment was treated as ungauged (when the proxy basin transfer is actually required) and really does not seem to be making too much use of hydrological knowledge (though the difficulty of transferring response characteristics using catchment characteristics or model parameters is, of course, well known). But our model hydrologist could perhaps be expected to know that there is an expectation that catchments of different scales might involve different processes, or catchments in hard rock wet areas of the west might be expected to be different to chalk catchments in the south east. So, in respect if the title of the paper the words very naïve might need to be added before

hydrologist (and indeed the MH is referred to elsewhere as a layman or angler rather than someone with better hydrological knowledge).

The fact that a layman can be modelled gives some hope that less naïve hydrologists can also be modelled. The proxy catchment results show there is a deep-seated similarity at play: given any catchment A, a catchment B can be found such that when the catchments have the same rainfall pattern they have the same specific discharge. This deep-seated similarity might help with the ungauged catchment problem.

The paper does not mention that we are often wanting to simulate the potential effects of future change. If that change is only to the inputs then the proposed strategy might work, perhaps with some degradation if the processes change. If, however, if it is change due to reforestation or NFM measures or other changes, then it could be used as a baseline to compare with future observations, but not as a simulator of a changed future (and indeed the changes might be within the uncertainty of the predictions if that was assessed in some way).

If there is a there is deep-seated similarity, it could help with the usual extrapolation problem (e.g. the proxy catchment might have some periods with higher rainfall intensity than the target catchment). The work gets quite far considering it is for a layman. Change impacts for reforestation etc. are outside the scope of the work. Perhaps, though, there is some potential in adapting the similarity in time method used to create Fig. 3 to the study of trends or anomalous variations in the matching of days.

Which then, of course, raises the question of what might happen if the MH had access to that committee of experienced hydrologists (or even inexperienced hydrologists – see the tale of the hydrological monkeys in the Prophecy paper cited). That experience might lead them to think more in terms of model parameters than direct use of data (Norman Crawford, Sten Bergstrom, and Dave Dawdy are examples from quite different modelling strategies but all were known for their skill in estimating parameters for models, including for ungauged sites …. though there may have been some potential for positive bias in tweaking and reporting results there). And there are instances of committees of experienced hydrologists not doing that well in setting up models and even getting worse results as more data were made available (see the Rae Mackay et al. groundwater example from NIREX days).

The committee would need a chairman, and the work would stand or fall on how the chairman is selected. There is no reason why parameters cannot be used, but a method would be needed to track the consequences through to the conclusion drawn. The concrete example in the paper drives a spike in the ground and gives a place to measure from, for the combination of NSE and hydrologic knowledge. Note that the value of the example comes from the fact that is for a combination, and not simply for the NSE on its own. There is no reason why performance should improve when more is known or is assumed known, unless a monotonic rising relationship is built in by design. Jakeman and Hornberger built such a relationship into their performance measurement (i.e. the performance for the modelling as a whole and not simply for hydrographs).

I would suggest that the authors could make more of the difficulties of going further with more experience and knowledge about catchment characteristics. It is an argument for their KERR

approach – but I would also suggest that the KERR approach also be extended to reflect the uncertainty to be expected as a result of that simplicity.

The Future Work section (Sect. 9.1) will be extended to suggest how the method could be taken further and to discuss uncertainty.

Some specific comments

L37 Best available theory – but there is also the issue of whether that theory is good enough when it differs from the perceptual model of the processes.

Yes.

L56. There seems to be a lot of overlap between what is referred to here as hydrological knowledge and the concept of a qualitative perceptual model. Both need to be simplified to make quantitative predictions (and often do so in ways that conflict with the perceptual model because of what is called here selective ignorance).

A mention of the overlap will be added to the text. No limit is placed on the nature of hydrologic knowledge. It could, for example, be quantitative or be about how performance must be measured.

P121. There were earlier suggestions of this approach, e.g. Buytaert and Beven, 2009, or even the donor catchment approach of the FSR/FEH.

Agreed.

L128. This is analogous to the Condition Tree concept in Beven et al, CIRIA Report C721 (also Beven and Alcock, 2012) that results in an audit trail to be evaluated by others.

Noted.

L132. Performance really ought to take account of uncertainty in the data (see earlier comment and papers cited in Beven, 2019)

See responses B-D above.

L137. But catchments that look very similar can also respond quite differently – even if mapped as the same soils/geology. We have an example from monitoring two small catchments on the Howgills. So issue of requisite knowledge here is when such small scale variability might integrate out (or not) – this was discussed in the 80s as a representative elementary are concept – eg. Wood et al.1988; also papers on when variability in stream chemistry starts to integrate out).

The text at L137 just notes that RR modelling utilises similarity in time and place (without specifying the nature of that similarity). The nature of the similarity, in as far as it is apparent in RR records, is considered in other parts of the paper.

L149. I think the "peasant's model" was suggested by Eamon Nash in modelling the Nile before this.

We will look for this.

L166. But again that upper limit will also definitely depend on the uncertainty and inconsistencies in the observations.

See response C above.

L182 nowhere to hide – exactly the point made for the Condition Tree / audit trail

Noted.

L336. Why not use an ensemble here to add in some uncertainty to the process?

Have assumed that this is L236. See response D above.

L394. But it is only a match to rainfall pattern – is there no additional knowledge that could be used? In the case of expected greater autocorrelation in large catchment for example (or extension to shorter time steps in small catchments) perhaps the last few predictions of flow might be useful (avoiding just doing 1 step ahead forecasting, though such a model with updating does also represent the forecasting model to beat – e.g the Lambert ISO model, see RRM book)

See response A above.

P425 relative importance is 2???? Not clear.

See response C above.

P426. Always upredictable – see the inexact science paper again

See response C above.

P456. Does not require scaling – There is an expectation that processes change with increasing scale, and that specific discharge becomes less variable with increasing scale, and generally less in moving from headwaters to large scales, so does this imply that this is compensated by a decline in mean catchment rainfalls so that the power on the area scaling is low, or simply that the variability is within the uncertainty of the predictions so does not have too great an effect on NSE? There are past studies on scaling with area that might be treated as hydrological knowledge here.

Perhaps this similarity in place boils down to a simple thing related to the evolution of the geometry of flow paths, rather than directly to some form of compensation between scales and processes and dynamics.

Wood, E.F., Sivapalan, M., Beven, K.J. and Band, L. (1988), Effects of spatial variability and scale with implications to hydrologic modelling. J. Hydrology, 102, 29-47.

Buytaert, W and Beven, K J, 2009, Regionalisation as a learning process, Water Resour. Res., 45, W11419, doi:10.1029/2008WR007359.

Beven, K. J. and Alcock, R., 2012, Modelling everything everywhere: a new approach to decision making for water management under uncertainty, Freshwater Biology, 56, 124-132, doi:10.1111/j.1365-2427.2011.02592.x

Beven, K. J., and Smith, P. J., 2015, Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models, ASCE J. Hydrol. Eng., 20(1), p.A4014010, doi: 10.1061/(ASCE)HE.1943-5584.0000991.

Beven, K. J., 2019, Towards a methodology for testing models as hypotheses in the inexact sciences, Proceedings Royal Society A, 475 (2224), doi: 10.1098/rspa.2018.0862

Iorgulescu, I and Beven, K J, 2004, Non-parametric direct mapping of rainfall-runoff relationships: an alternative approach to data analysis and modelling, Water Resources Research, 40 (8), W08403, 10.1029/2004WR003094