We thank A. Bàrdossy for his very precise commentary on the paper. Please find in the following our point by point response.

Comment: The methodology uses an inequality type approach…

Response: We agree with the referee that our analysis is based on inequality type approach. Pande et al (2012) first used Chebysev type inequality for a class of linear reservoir models and found that they are relatively weak. Hence they suggested the use of a tighter inequality called Markov inequality. Arkesteijn and Pande (2013) employed this inequality to estimate model complexity of several model structures including SAC-SMA and SIXPAR.

Arkesteijn and Pande (2013) also proved that complexity estimated based on Markov inequality, as done here, accurately estimates the extent of model output space and hence is accurate in measuring model complexity within the presented framework of prediction uncertainty.

Comment: Further I have a problem of the transfer of the upper limit to the risk. The argumentation is somewhat as if x < a, y < b and a < b would imply x < y which of course is not true.

Response: The recommended use of the upper limit is to select a model that is better than another model with certain minimum probability. Consider two models 1 and 2 with expected risks $E\xi_1$and $E\xi_2$ and empirical risks ξ_1 and ξ_2. Let use define an event A_1 that is $|E\xi_1 - \xi_1| > t$. Similarly let A_2 be $|E\xi_2 - \xi_2| > t$. Let us have two probability inequalities:
$P(A_1) \leq \delta_1(t)$ and $P(A_2) \leq \delta_2(t)$. The upper bounds in both the inequalities depend on corresponding model complexities. Since we know that $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$, we can say that $|E\xi_1 - \xi_1| > t$ and/or $|E\xi_2 - \xi_2| > t$ with a probability that is less than or equal to $\delta_1(t) + \delta_2(t)$. In other words one can say that $|E\xi_1 - \xi_1| + |E\xi_1 - \xi_1| \leq t$ with a probability of atleast 1-$[\delta_1(t) + \delta_2(t)]$. Finally we know $E\xi_1 - \xi_1 - (E\xi_2 - \xi_2) \leq |E\xi_1 - \xi_1| + |E\xi_1 - \xi_1|$.
From the above sequence of valid arguments, we note that the following holds with a probability of atleast 1-$[\delta_1(t) + \delta_2(t)]$ that
$E\xi_1 - E\xi_2 \leq (\xi_1 - \xi_2) + t$.
We can now obtain a valid answer to the question of which model is better in the sense of its (expected) risk. What is the probability that model1 is better than model 2 in expected sense, i.e. what is the probability that $E\xi_1 - E\xi_2 \leq 0$?
Answer: with a probability of atleast 1-$[\delta_1(\xi_2 - \xi_1) + \delta_2(\xi_2 - \xi_1)]$. We obtain this by letting $(\xi_1 - \xi_2) + t = 0$. Thus better models can be truthfully revealed with a confidence level that is resolved by the upper limit on risk.
It can be shown that the above selection procedure is equivalent to minimizing a complexity regularized cost function when model structures that are nested in their complexities are considered.
Nonetheless tight upper bounds are desirable for tight 'least' probabilities of one model being better than the other.
We will add the above discussion in the revised version of our manuscript.

Comment: Algorithm 1 is a pure resampler. The goal of hydrological modelling is not to repeat modelling under the same stationary conditions, but to transfer it to different conditions. It is under the different conditions where one faces risks.

Response: We agree. Algorithm 1 is a resampler. We however wonder if the referee is suggesting that different conditions can only appear under non-stationarity. One can think of different conditions from a stationary distribution as well. Different sequences of input forcing with similar long run statistics are quite possible. Nonetheless, we agree with the referee that Algorithm 1 can be adapted to handle non-stationarity. This is one of the strengths of the proposed algorithm. However it may require assumptions about how the underlying distribution is changing over time as would any other resampler.
We consider regularized model selection problem (wherein a model is selected such that it has an optimal tradeoff between its prediction performance on finite data and complexity) as a problem of transferring models in time under stationarity assumption. This is akin to the concept of transferring models in space. Further, just as in this paper, the current state of the art in transferring models in space (instead of time as done in this paper), for eg based on regionalization of parameters, is an exercise under the assumption of stationarity in our opinion.

Comment: It is under the changed conditions where complex models are sometimes considered as more plausible. (There is a constant debate on the use of very complex so called physically based models for climate change assessment. The argument is often that the risk is lower as the model is based on more founded principles.)

Response: We agree. As we have also stated in the paper, the methodology does not exclude the case when a more complex model is warranted. The methodology sees choice of a model with optimal complexity as the one that has best tradeoff between empirical risk and model complexity. Please note that Algorithm 1 can be adapted to resample from a non-stationary distribution. The methodology of the paper would then still apply and a more complex model may still be chosen if the underlying changing distribution warrants so. However, since the framework of predictive uncertainty presented here is probabilistic (in the sense that it suggests that one model is better than another with certain minimum probability), a model with non-optimal complexity may always be selected with certain probability. However the error in probability of choosing the wrong model is not 0 even in the case when the exact probabilities are available (which is nearly impossible).

Comment: A reasonable weather generator might resolve the problem of stationarity and could give some insight to the complexity vs model transfer problem.

Response: We agree.

Comment: The strong restrictions of Algorithm 1 are contrasted by Algorithm 2. Here the choice of the parameters _ is not completely clear. The use of Latin Hypercube Sampling gives me the

impression that the authors do not restrict their model parameters to such which are useful for the case study example. A large number of uncontrolled parameter combinations may lead to nonsense models for which the risk is irrelevant for any application.

Response: Please note that we use Algorithm 2 only to quantify complexity. Algorithm 1 resamples input forcings. We donot estimate empirical risk or contrast the performance of sampled parameters with observed data. We use algorithm 2 to quantify complexity that may then be used to select robust models. Even if we sample nonsense models and estimate their upper bounds on (expected) risk, they would be swiftly rejected. Further, given clear differences in complexity distributions corresponding to various parameter ranges, sampling more points will not alter the conclusions. Nonetheless, research is underway to reduce the computational complexity of the introduced algorithms and introduce smarter search sub-algorithms within Algorithm 2.
We will add the above discussion in our revised version of the manuscript.

References:

Pande, S., Bastidas, L. A., Bhulai, S., McKee, M. (2012). Parameter dependent convergence bounds and complexity measure for a class of conceptual hydrological models, Journal of Hydroinformatics, Vol 14, No 2 pp 443–463.

Arkesteijn, L., and Pande S. (2013), On hydrological model complexity, its geometrical interpretations and prediction uncertainty, Water Resour. Res., 49, 7048–7063, doi:10.1002/wrcr.20529.