Cover letter to second revision of hess-2020-128

Dear Editor, dear Referees,

Please find our replies to the second round of reviews and the related Editor comment below. We provide a point-by-point reply and indicate changes made to the manuscript. The related page and line numbers refer to the second revised version of the manuscript. The Editor and Referee comments in this document are marked in blue, in the manuscript all changes are marked in red.

Editor note

Comment 1: Dear Authors, as you can see, both Referees recommend major revisions. Referee #1 Elena Toth still casts doubts on the way the comparison among the models is carried out. Indeed, although you argue that the bit-by-bit approach you propose allows comparing models which make use of different input data, I still see Elena's point: if two modelling approaches CAN make use of the same data, comparing them with different inputs by means of an index that accounts (also) the obtained performance can be misleading.

Reply 1: As suggested by Elena Toth, we added a simple ANN with the same input as the autoregressive model (Q(t-1), Q(t-2), Q(t-3)) to the set of models used in the paper, and now provide a more in-depth discussion about how the bit-by-bit method can be used to guide model optimization and model selection in sections 3.1 and 3.2. Please also see our related reply to Comment 1 by Elena Toth.

Comment 2: The other Referee #2 raises more fundamental issues about the way you present your approach. A less narrow review of the wide literature about model parsimony would help the reader to better place your contribution in the context of existing research on the topic.

Reply 2: Agreed. As model parsimony and model complexity are strongly linked (parsimonious models are models of minimally adequate complexity), we have decided to add an overview on related literature to section 1.4 (uses of 'complexity' in the hydrological sciences). Also, we created in section 1.2 (Guidelines for developing parsimonious models) a new paragraph about 'Model selection by applying complexity penalization measures' and added this to Figure 1 (b), as i) these approaches are often used and ii) can be considered a particular class of methods.

Comment 3: And also, to better support your claim about the approach being capable of guiding in model optimization, a clearer training-validation approach to your performance comparison would be probably useful, as suggested by Referee #2.

Reply 3: We agree that the manuscript will benefit from better explaining how the bit-by-bit method can be used to guide model optimization and model selection. We think this can be best done by discussing several typical use cases of model optimization and model selection. We now do so in section 3.2. Please see a detailed description of the use-cases in our reply to Comment 1 by Elena Toth.

Comment 3: Given all this, I definitely agree that major revisions are still required, and I invite you to take seriously into consideration all the raised issues, providing a convincing rebuttal in case you still think that some of them have not to be addressed. I look forward to receiving a new version of your manuscript. Best regards, Roberto Greco

Reply 3: We hope that with our replies and changes made to the manuscript, we could address the Editor's and the Referees' concerns in a satisfactory manner.

RC 1 (Elena Toth)

Comment 1: Dear Authors, I do appreciate that you are now presenting the results of a calibration/validation split-sample test and you improved the manuscript in many points. I am also glad that you have better explained the contribution of the paper and highlighted that "It is important to note that the purpose of the model comparison here is not primarily to identify the best among the different modelling approaches", since actually I keep thinking that for a fair comparison of the performance of the models you cannot use different input. When you write (in the discussion) "It is therefore neither required nor useful to establish models that all rely on the same input data, because the bit by bit approach might be used for exactly the purpose of comparing two models relying on different input data", I think that this holds true for the complexity measure but not for the performance one. In fact, the correct input strongly influence the 'side information' that is needed to reduce the information loss, since the input information is crucial to the information content embedded in the model prediction. As written in Campolo (WRR, 1999), the last observed discharges are generally included as inputs to ANN runoff-prediction models, since they provide information on the state of saturation of the basin, which is a function of the history of the meteorological input in the period preceding the streamflow generation: the capability of the system to respond to rainfall perturbation is represented by the ongoing streamflow measured in real-time at the closing section. And in fact you use just such input for the best performing model (#7). Such model is by far the best performing one since it is the only one that uses such a precious information (and actually also a fair comparison of the performances of the conceptual models would require, see Toth and Brath 2007, a real-time updating procedure making use of such data). For this reason, the new ANN you applied, still fed by rainfall only, still provides a relatively poor performance in comparison to Model 7 (and the complexity seems to be even worse than in the first version). If, instead of using your complex LSTM fed by the rainfall values, you had used a simple multi-layer ANN (again see what I did in Toth and Brath, WRR, 2007) with the same input (past Q values) that you use with the autoregressive model, you would have got performances similar (or probably better) than Model 7, even if the complexity would be, I guess, still much higher than that of Model 7. Such comparison would be not only fairer, but also more significant than what you are presenting now. Such a simple ANN model implementation would take very little energy for you to set up, and, respectfully, I would insist for you to do so, since it would provide sounder results, supporting the scientific relevance of your interesting approach. Elena Toth

Reply 1: We are glad that Elena Toth acknowledges most changes we made to the manuscript and accept her request for providing a fairer model comparison by using the same input for the models that are compared. In fact, we had the same intention when we decided to provide model 8 (the LSTM) with the same input data (precipitation) as the bucket models, to allow for a fair comparison among these. Elena Toth correctly states that this neglects all the information available in past observations of discharge, which are used by model 7 (the autoregressive model), and therefore provides an advantage for the AR-model when comparing it to the LSTM. To resolve this issue, and to also better demonstrate how the bit-by-bit method can be used to guide model optimization and model selection, we

- added a simple ANN (single hidden layer with five neurons) with the same input as the autoregressive model (Q(t-1), Q(t-2), Q(t-3)) to the set of models used in the paper, and kept the LSTM
- we have completely re-written section 3.2 and now discuss several use cases of model optimization and model selection in section 3.2.

We also acknowledge Elena Toth's remark about "the input information being crucial to the information content embedded in the model prediction" by adding a related sentence in the summary and conclusions (P18 L425)

RC 2 (Anonymous)

Comment 1: I have gone through the revised version of the manuscript. Unfortunately the authors either have evaded almost all my comments (except putting the paper in the landscape of select complexity studies) or have decided to respond to comments of their own. They have accentuated their reliance on Weijs and Ruddell without providing either theoretical or empirical support for their claims. Following are some of my comments (only on the added text in red by the authors)

Reply 1: We would like to emphasize that we addressed in a point-by-point manner all comments made by the referee, and we justified each case where we did not follow the referee suggestions.

The referee mentions we responded to comments of our own. As the referee does not provide detail information about where this is the case, we assume that these are cases where we responded to comments made by the other referee.

The referee mentions the strong linkages of the manuscript to Weijs and Ruddell (2020). We are not sure what to do with this comment. Yes, there is a strong connection of this manuscript to Weijs and Ruddell (2020), and throughout the manuscript we explain how they are related to each other (especially in section 1.2), but we do not understand why this should be a problem.

The referee mentions that we do not provide theoretical or empirical support for our claims. As there is no detail information about which claims the referee refers to, we assume it is linked to referee comment 3 and comment 6 of the first round of reviews:

- Comment 3: ("How max parsimony + max performance -> max generalizability? Theoretical rigor behind the claim is missing").
- Comment 6: (" … Author's claim to universality should first provide a rigorous theoretical treatment that has not even been provided in the WRR paper that the authors allude to").

With respect to these comments, we responded that for better readability of the hydrological readership of HESS, we refer to Weijs and Ruddell (2020) and related literature about AIT rather than including it into our manuscript. Following the referee's repeated request, we have now included a brief discussion on the theoretical foundations of our claims in section 1.2 (a detailed list of additions is provided at the end of this reply)

Re comment 3: Adding some form of performance measure and some form of complexity penalization is a widely accepted way to do model selection (which often has the underlying aim of inference or prediction, both of which are related to generality). See for example AIC and BIC information criteria, statistical learning theory, algorithmic information theory, and the minimum description length principle. We have now made this clearer in the paper by adding some discussion on AIC and BIC and how they relate to the concepts we discuss.

Re comment 6: Since proofs related to AIT rely on a wide range of prior knowledge about theoretical computer science, we see little point in repeating those in a paper addressed at an audience of hydrologists. To the long explanation in the previously reply to the reviewer, which we did not include in the paper for the reasons above, we could add the following: Since the space of all self-delimiting programs for a universal Turing machine form a prefix-free code (no valid program is the first part of

another valid program), a program that is one bit longer than another program will be half as likely to be encountered in random noise on the input tape. Therefore, the algorithmic probability of a certain output sequence is proportional to the probability of encountering that output string from a Turing Machine fed with random noise.

We honestly believe adding this to the paper would serve little purpose for the reader. We therefore kept our descriptions at an intuitive level, without too much computer science terminology, but elaborated a bit more on those in the present paper.

Related changes made to the manuscript:

- P4 L68-72: We created in section 1.2 (Guidelines for developing parsimonious models) a new paragraph about 'Model selection by applying complexity penalization measures' and added this to Figure 1 (b). In this paragraph, we discuss AIC and BIC in more detail
- P4 L74-78, P4 L93-96: Added some more information about AIT
- P4 L87-88: Added some more information about the approach of Weijs and Ruddell (2020)

Comment 2: Please also discuss the work of hydrological complexity by Sivapalan, Wagener and colleagues. Where/how does your paper compare and why if not?

Reply 2: We have added related literature and a discussion to section 1.4 (P7 L173-175). The main difference between the work on complexity of Sivapalan, Wagener and colleagues and what we propose in our manuscript is in the definition of 'complexity'. We also added more literature about ways to define and measure complexity (P7 L165-168).

Comment 3: "Weijs and Rudell (2020), we express model performance in terms of information losses" is limited by how p(.) is estimated (it being prone to misspecification) in Equation 2. This in itself is a model, subject to complexity associated challenges. I would therefore suggest caution in suggesting Weijs and Rudell (2020) at the same level of rigour as some other fanstastic work done on Occam's razor in Mathematical Physics and Applied Probability. For example, no underlying arguments for uniform convergence, asymptotic consistency etc are provided other than authors guiding us to literature on information theory.

Reply 3: The point about estimation of p being a model was raised by the referee in comment 11 of the first round of reviews. We agreed and added a discussion in section 2.4.1. In terms of convergence, several estimators for discrete distributions based on limited samples have been proposed that both converge asymptotically towards the true distribution and provide uncertainty bounds as a function of sample size and binning choice. In Darscheid et al. (2018), both a Bayesian approach and a Maximum-Likelihood approach are presented. We added this reference and a short explanation to the manuscript (P11 L246-249).

As for Weijs and Ruddell (2020), in that case information loss is expressed as the minimum description length needed to reproduce the observations. As such, if probability is used, the model to describe p (e.g. the storage of the histogram) is included in the description length and a misspecification or overfitting of the probability model is penalized automatically. One key philosophical advantage of AIT is that the information measures do not refer to an underlying probability distribution, but works directly on the observed data. This gain in generality and rigor is paid for by less practical applicability. In this paper, we do not consider descriptive complexity of the model nor the probability model, and therefore use classic entropy measure in combination with binning approaches recommended in the literature. We agree this is an area worthy of further future research. Since this is again a quite complex and subtle

discussion to explain well in the paper, and not directly within the scope of the central points of the present paper, we decided not to elaborate on this in the paper.

Comment 4: What are the implications of measuring information gains compared to a lower benchmark – the entropy of a uniform distribution, in terms of quantifying model performance. Is it an approximation? Or just some measure unrelated to approximation error (which I think is the case)?

Reply 4: Upper and lower benchmarks are helpful to put a particular model performance into a global perspective. For example, Nash-Sutcliffe efficiency (NSE) is limited to values [-∞ 1]. Models showing NSE-values smaller than zero (the NSE of the mean as a lower benchmark model) are typically rejected, and hydrological models showing NSE-values > 0.8 are typically considered well-performing. So clearly there is some merit in knowing general upper and lower bounds of NSE to put the performance of a particular model into a wider perspective. The same is the case for information gain, where the entropy of a uniform distribution provides a global lower bound. The difference between the two lower bounds is that for NSE it is a function of the data under investigation (as we use them when calculating the mean), and for entropy it is a function of the resolution (number of bins the value range is divided in) for which we investigate the agreement of observation and simulation. We can interpret the entropy as the missing information about which bin (at the chosen precision) the value falls in, when we have a priori knowledge of the range of the data.

Comment 5: "Choice of n is typically guided by the objective to balance resolution and sufficiently populated bins" – exactly and this somehow sounds like Occam's razor again. See my comment above.

Reply 5: Yes, choice of the number of bins can be seen as an optimization problem (see e.g. the binning approach suggested by Knuth (2013). Please also see our replies to comment 3.

Comment 6: Not clear what validation set approach actually doing. I would like to see how model chosen based on the principle presented here performs on the validation set (if it somehow identifies a generally good performing models) in comparison with other 'demoted' models. Otherwise its just a trivial exercise, sure you can have very simple but really bad models. We do not need to understand it bit by bit.

Reply 6: We do not understand what is unclear about the validation set approach as we explain it in section 2.1, and which is standard practice in hydrological modelling. We hope we could clarify this issue by the use cases of model optimization and model selection that we now discuss in section 3.2.

Comment 7: "We continued by describing several paradigms to guide model development:.. Weijs and Ruddell (2020) express both model performance and descriptive complexity in bit, and by adding the two obtain a single measure for what they call 'strong parsimony';.." that ok, but the authors are not the only ones attempting to describe parsimony. Don't even think Weijs and Ruddell (2020) is a paradigm. Please delete or list many others who have dealt with concepts of parsimony.

Reply 7: Following the suggestion by the Referee and the Editor, in section 1.2 (Guidelines for developing parsimonious models), we now give a short overview on approaches to develop parsimonious models in addition to those that are already referred to and used in the paper (Occam's razor, complexity penalization, strong parsimony as suggested by Weijs and Ruddell (2020), validation set, and bit by bit).

Comment 8: "Occam's razor puts an emphasis on descriptive complexity, considers performance as a side condition, but it ignores computational complexity;.." this is incorrect interpretation (on performance being side condition). Please delete in order not to misguide readers.

Reply 8: As described in Weijs and Ruddell (2020), and repeated in our manuscript (P X L X), "Occam's Razor, a bedrock principle of science, argues that the least descriptively complex model is preferable, at a given level of predictive performance that is adequate to the question or application at hand." Predictive performance is used as a threshold to select a set of models, which are then compared in terms of descriptive complexity, and the model with least descriptive complexity is then selected as the best. Predictive performance here serves as a threshold-like filter, which we wanted to express by 'side condition'. As this was apparently misleading, we rephrased the sentence the referee refers to:

"When applying Occam's Razor, the parsimonious among the well-performing models are identified, but comparisons of models of different complexity for model selection are not possible by this principle alone." (P4 L65-67).

"Occam's razor puts an emphasis on descriptive complexity, and is often combined with performance considerations, but it ignores computational complexity;" (P17 L395).

Comment 9: "i) measuring computational complexity by 'Strace' is general in the sense that it can be applied to any model that can be run on a digital computer; ii) 'Strace' is sensitive to all aspects of a model, such as the size of the model itself, the input data it reads, its numerical scheme and time-stepping; iii) the 'bit by bit' approach is general in the sense that it measures two key aspects of a model in the single unit of bit, such that they can be used together to guide model analysis and optimization in a pareto trade-off manner in the general setting of incremental learning. It can be useful.." In general I agree with i). I find their iii) statement misguided – I don't think strace can be used for incremental learning, since computational complexity doesnot make us unlearn something. "Descriptive complexity'" may do so (not based on bit by bit idea however) if interpreted appropriately in how it affects learning. Therefore computational complexity is a redundant concept in context of learning.

Reply 9: We agree that if the computational effort of a model is not an issue, computational complexity is indeed not an important criterion for model optimization and model selection. However, there are many applications in the Earth Sciences where computational effort is critical (e.g. global circulation models or global land surface models), and for which a lot of time and effort is devoted to increase their computational efficiency. In such a setting, computational complexity can be an important criterion for model evaluation. E.g. consider two land surface models, which are identical except for their spatial resolution. If the two models are equal in terms of performance, but the coarser-resolution model is computationally more efficient, we will prefer this model, and we have at the same time learned something about the adequate resolution to represent the natural system in the model. In that sense computational complexity also plays a role in learning. We hope this is also made clearer in the manuscript by adding the use cases in section 3.2.

Comment 10: Over all, I think another revision is at least needed where Weijs and Rudell (2020) is not glamorized (and reference to it kept to a minimum) since all arguments based on it are rather weak.

Reply 10: There is a strong connection of this manuscript to Weijs and Ruddell (2020), and throughout the manuscript we explain how they are related to each other (especially in section 1.2), but we do not understand why this should be a problem (please see also our reply to Comment 1).

Comment 11: The authors should clearly show how computational complexity (or even Weijs and Ruddell description complexity) is affecting learning based on performance (only) of all the models on the validation set. My contention here is that computational complexity in this regard (of generalized learning) is rather of little value and that the paper is merely another demonstration of an interesting concept that does not add much to learning from data.

Reply 11: We hope we addressed this comment appropriately by our replies to Comment 6 and 9, and by adding to the manuscript the discussion of use cases in section 3.2.

Comment 12: I would advise humble response in the next iteration, where limitations and purpose of their manuscript is clearly documented.

Reply 12: With all respect, we would like to mention that while we welcome the scientific debate associated with this review process, and appreciate the time and effort spent by the referee, to our understanding the scientific debate is not to supposed to be humble but rather respectful and professional. Therefore, we aim to convey neither arrogance not humbleness but solely the respectful, professional civility of a scientific debate. That said, we did add an elaborate section on the use cases of the method to clarify its purpose. We think limitations are fairly discussed, and we hope the manuscript will contribute to the scientific discussion.

Yours sincerely,

Uwe Ehret, Elnaz Azmi, Steven Weijs, Benjamin Ruddell, Rui Perdigao

References

Darscheid, P., Guthke, A., and Ehret, U.: A Maximum-Entropy Method to Estimate Discrete Distributions from Samples Ensuring Nonzero Probabilities, Entropy, 20, Article:-601, 10.3390/e20080601, 2018.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrol. Process., 22, 3802-3813, 10.1002/hyp.6989, 2008.

Knuth, K.: Optimal Data-Based Binning for Histograms, arXiv:physics/0605197v2 [physics.data-an] 2013.

Weijs, S. V., and Ruddell, B. L.: Debates: Does Information Theory Provide a New Paradigm for Earth Science? Sharper Predictions Using Occam's Digital Razor, Water Resources Research, 56, e2019WR026471, 10.1029/2019wr026471, 2020.