Hydrol. Earth Syst. Sci. Discuss., 12, C2869–C2884, 2015
www.hydrol-earth-syst-sci-discuss.net/12/C2869/2015/

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Hydrological model parameter dimensionality is a weak measure of prediction uncertainty" *by* S. Pande et al.

S. Pande et al.

s.pande@tudelft.nl

Received and published: 28 July 2015

Response to referee 1:

We thank the referee for constructive comments. We thank the referee for the comment on the innovation that the authors have tried to bring to the subject of hydrological model complexity. We appreciate the concerns that the current version of the manuscript is not sufficiently intelligible. We hope to clarify the concerns that the referee has raised in the following responses as well as in a future version of the work.

Comment: 1. a) Page 3: "Should not this uncertainty in assessing structure deficiency depend on the class of model structures which are used to assess deficiencies?" What is meant here?

Response: What we mean here is that assessment of model structure deficiency may itself be uncertain. Such uncertainty should depend on the model structure under investigation.

Comment 1b) "the characteristics of uncertainty in system representation can then identify the consequences of ill-conditioned model selection problem and hence define ill-conditioned model selection."

Response: An ill-conditioned model selection problem has one or more of the following 3 properties: 1) a solution to the problem does not exist, 2) multiple solutions exist (non-uniqueness) and 3) has unstable solution (i.e. a modeler obtains different model solutions on different data sets). If a model selection problem is ill-posed, the uncertainty in representing the underlying system by a selected model will be characterized by one or more of these properties. For example, if the model selection problem suffers only from non-uniqueness, the uncertainty that follows (in representing the underlying system) will be different from a case when the model selection problem suffers only from instability.

Comment 1c) Page 3: "Meanwhile, instability refers to inconsistency in process representation as more information on the underlying processes is available."

Response: An unstable model selection problem would yield drastically different model solutions on different realizations of the same underlying system. Consider the case of a collection of unstable models. Unstable models have the property that they produce dissimilar outputs when its input forcings are even slightly perturbed. Thus, when such a collection of models are used to identify a model that best suites the observations, dissimilar models may be chosen as different realizations of the underlying system come to fore. Consider the case when these different realizations are put one after the other to form a sequence of realizations of sizes N, 2N, 3N and so on (where N is the sample size for example). In this case, a model that is selected on sample N will be drastically different from the one that is selected on 2N and so. Thus a model that

is selected on data set N is not consistent as it is not selected again as more data is available.

Comment 1d) Page 5: "..stable system representation..." –> what does this mean? How can I assess whether my model is a stable system representation

Response: A system representation is stable if it provides similar performance (for e.g. in streamflow simulations) on independent sets of observations. Such performance need not be the best possible but it should be consistent. Kindly see Arkesteijn and Pande (2013) as well as Pande et al. (2009) where they have assessed stabilities of system representation when model complexity is controlled for in a model selection exercise and when it is not. For example, a stable model representation that has NSE of 0.4 on one set of observations, should have NSE of around 0.4 on another independent set of observations. An unstable representation may have NSE of 0.9 on one and NSE of 0.4 on another.

Comment 1e) Page 5: "Often models with low parameter dimensionality (i.e. less number of parameters) are considered less complex and hence are associated with low prediction uncertainty." I disagree with the latter part of this statement. Models that are simple might not fit the data very well and hence have significant uncertainty - the uncertainty is not made up of parameter uncertainty but in large part of structural uncertainty. More complex models might better fit the data - yet their parameter uncertainty might be larger (structural uncertainty - that is model error smaller). From a statistical point of view their total uncertainty is the same - if the goal is to describe statistically the data.

Response: Many thanks! Kindly note that we define prediction uncertainty as instability in performance of a model over different realizations of data. The model of two equally deficient models that has lower prediction uncertainty (or lower instability in model performance) will have lower possibility of worse performance over future unseen data. This model will also be a robust choice. Thus robust model selection is a complex-

ity regularized model selection problem that trades off model deficiency with model complexity. However, any treatment of prediction uncertainty would remain incomplete unless the role of model complexity in prediction uncertainty and the role of parameter dimensionality within the notion of model complexity is satisfactorily described. We use an upper bound on prediction uncertainty that is a function of model complexity, which in turn measures the sensitivity of model simulations to perturbations in input forcings. Our paper argues that parameter dimensionality only partly describes model complexity and hence prediction uncertainty.

Comment 1 f) Page 5: "...unstable model representation..." –> Antonym of what is used previously- but again this wording is new to me. What does stable/unstable refer to? A model that is numerically unstable (I know this is not what the authors are referring to) - a model that does not describe the data very well? Or?

Response: Kindly see our response to comments 1c) and 1d). We however will minimize the use of word unstable to minimize such confusion.

Comment 1g) Figure 1: Caption not repeated here. o_1 and p_1 are observed and predicted value - how can this be of size "N" (one would a vector then not just a scalar o_1 and p_1). What is mode output space? We will get "N" distance between the observed and simulated data. We can write this in vector notation and plot this? How does "B" (vector notation) then define the overall distance? I find this presentation to be very confusing - and because this is the underlying theory the authors are developing it is hard to assess what is going on. I would strongly recommend to keep the presentation simple - and clarify the figures.

Response: Please note that o_1 and p_1 are indeed described as 2 dimensional vectors {o_1,1, o_1,2} and { p_1,1, p_1,2} for N =2. In general o_1 = { o_1,1, o_1,2,.., o_1,N}. Model output space is a collection of all possible predictions that the model can make corresponding to all possible input forcing realizations. Please note that the figure illustrates the case when two realizations are observed and simulated (with the

model being forced by corresponding input forcings). The N distances between observed and simulated for the two realizations is given by vectors A and C. These vectors plot the distances. The vector B represents the distance between two simulations of the model (forced by input forcings corresponding to the two realizations). Finally, the magnitude of the distance (i.e. the magnitude of the vectors) would depend on the metric being used. Examples include mean absolute error, root mean square error. These will then impose l1 and l2 norms respectively when estimating the magnitude of the vectors.

Comment 1 h) Figure 2: Same issues - "also model structure output space". Why not just "model output space". This Figure generalizes Figure 1 to multiple models. How can the model structure be defined as a union of two models? Each model has its own structure and output space - the union of two models would constitute a part of the feasible model space?

Response: We agree that this figure generalizes Figure 1 to multiple models. We define a model structure as a collection of models and thus distinguish between a model and a model structure. For example, a linear reservoir model structure is a set of linear reservoir models corresponding to recession parameter values in (0,1]. Kindly note in section 2.2.1 we describe how a model structure can be defined as a union of two models (through abstract parameterization).

Comment 1 i) Figure 3: Unclear as well. So model 2 can be considered to be part of the output space of model 1. OK. Thus model 2 has a large variation in the output space for given input data - OK - why would model 2 then have a higher instability in system representation? I would think this is model 1 given that this model exhibits a larger uncertainty in the output space?

Response: We agree, this is a typo. It should be model structure 1 and not model structure 2.

Comment 1 j) Figure 5: Unclear (inset on right side too small. Now model 1 has highest

instability? I find the wording instability very confusing. Has a negative connotation - and as I am trying to follow the logic here I am confused about the logic

Response: Please see our response to comment 1i). Model structure 1 has indeed higher instability. Instability implies dissimilar models are chosen when confronted with different realizations.

Comment 1 k) Page 7: "We define instability of a given model by the variability in the differences between its outputs over two different realizations of data." This is confusing – two different realizations? In practice we only have one realization of the data. Or do you mean we have two or more observations of the same type but at different locations in space or time?

Response: We agree, in practice we have only one realization. We use the idea of two or realizations to define instability. A more constructive example would be splitting a realization of size M*N into M realizations of sizes N. Then variability of model performance on these M realizations will give us an idea of its instability in representing the underlying system. Kindly see Arkesteijn and Pande (2013) where such an idea of instability has been implemented.

Comment 1 l.1) Page 7: "A model then is more unstable if it tends to have larger differences between model simulations for any given pair of data realizations. Such a definition is sufficient to encapsulate the notion of inconsistency in process representation by a model." I fail to understand this logic. I am just missing pieces here to understand the reasoning of the authors. If a model exhibits a large uncertainty in the output space, that is given some input data and prior parameter space, the model simulates a large variability in the output space - is this then equivalent to inconsistency in process representation?

Response: Kindly consider an unstable model 1 that performs really well on one realization X and poorly on another Y. If the class of models contains unstable models, there may as well be another unstable model 2 that performs poorly on X and better

than model 1 on Y. We might end up choosing model 1 on realization X and model 2 on realization Y. This is what we mean by inconsistent system representation.

Comment 1 l.2) Another issues that emerges here is that the output space of a model does not say much about process presentation. a model resolves many different processes (different model components – equations if you will) - the collection of which produces a model output. The focus on model output as vehicle for analysis complicates things further - because what is analyzed is a summary term of different processes. I feel that more progress can be made if authors focus on outputs that relate directly to individual processes in the model. I feel that this is a more realistic assessment of the strengths/stability/consistency of the model - rather than evaluating the range of the outputs.

Response: We agree. The model output space is determined by the choice of output variable. Consideration of multiple output variables that correspond to individual processes is therefore desirable. Yet, the notion of (in)stability applies individually for an output variable and embedded in the output space corresponding to the output variable. Kindly note that there is more to estimation of complexity than evaluating the range of the outputs. Please see our response to your following comment.

Comment 1 l.3) The authors continue on Page 7 with "This is because it is quite likely that a highly unstable model that appears to be a suitable representation of the underlying system on one piece of information may not be a suitable representation on another or more pieces of information." Exactly - this is a common problem - a model might be considered appropriate when evaluated against one type of data - but completely useless when asked to simulate/predict another variable. I guess my problem is with the unstable formulation the authors. I do not view a large range in simulated value as unstable. I might refer to it as uncertain. I think authors can be much clearer in their reasoning if they adopt a more logical jargon.

Response: Kindly see our response to Comment 1 l.2). We agree that we should

adopt a more logical jargon, especially since what we want to convey by the use of term instability is more variability vs less rather than really useful vs completely useless. In this way, instability in system representation contributes to uncertainty. Finally, consider the case when multiple model output variables are considered. This would lead to multiple measurements of instability, one per model output variable, and corresponding measures of complexity. However it is not yet clear how all these measures of complexity can be jointly used (or summarized) to assess and constrain a model selection exercise. Kindly note that complexity measure (and the corresponding notion of instability) that we are proposing is not about large range of simulated values for a given input forcing but about how diverse these simulations are on different realization of input forcings of size N.

Comment 1.m) From Page 7 on - the material is excessively difficult to follow. Perhaps the earlier part is still intelligible - the later part (Page 8 middle forward) is hard to follow

Response: We hope to clarify this further in a future version of the work.

Comment 1.n) Page 9: structure output space?

Response: Kindly see our response to comment 1.h.

Comment 1.o) Page 10:"Figure 4a also demonstrates that deviation in performance of system representations from model structure 2 is often larger than structure 1, to the extent that Pr(..) is larger for nearly all >0." This paragraph is trying to say that models with more parameters (more complex) generally have a larger uncertainty in the simulated output - and hence exhibit a large simulation uncertainty?

Response: Yes. In the light of this referee's comment on model structure error in 1.e) please note that there is no structural deficiency in this synthetic example. Both the model structures contain the 'truth'. Both the structures contain the linear reservoir model that simulates the synthetic streamflow. Thus, only model complexity (which equates to number of parameters here since the 2 reservoir model structure subsumes

a single reservoir model structure) affects prediction uncertainty. The conclusion that we attempt to draw here is that prediction uncertainty can be controlled by controlling model complexity. Arkesteijn and Pande (2013) present real world case studies to demonstrate the same.

Comment 1.p.1) Page 11: "Figs. 4 and 6 suggest that controlling for the complexity in a model selection exercise may stabilize the representation of underlying processes. This is akin to “correcting” the ill-posedness (Vapnik, 1982) of model selection problem by constraining the complexity of the model structures used. This is equivalent to regularized model 20 selection problem" OK. But this analysis is based on synthetic data? What about using real-world data - would one arrive at a similar conclusion?

Response: Kindly note that the proof of concept of complexity controlled model selection results in stabilized (or robust) representation of the underlying processes has been provided by the first two authors in Arkesteijn and Pande (2013). Real world data and complex hydrological models were used in the study. Kindly also see Pande et al. (2009) and Pande et al. (2014) for the same for other studies of similar kind. The use of synthetic data here is solely to illustrate the concept. The over-arching aim here is to suggest that number of parameters does not completely describe model complexity and hence does not completely describe the effect of model complexity on prediction uncertainty.

Comment 1.p.2) But why would one need to constrain the complexity for model selection? The marginal likelihood (Bayesian perspective) will pick the simplest model that still explains the data - so if complexity is inappropriate then this model will not get selected - or if the model is too complex then the integral of prior and likelihood will provide values for p(D) that are smaller than those derived for a simpler model.

Response: We think that maximizing marginal likelihood (integral of prior and likelihood) implicitly controls for model complexity. Consider its approximation such as KIC that trades off the likelihood function with its Hessian (second order derivative) with re-

spect to its parameters, both evaluated at the optimum. The Hessian of the likelihood can be considered as a measure of complexity in Bayesian perspective. This is similar to the notion of complexity being presented. This is because the Hessian measures the curvature of the likelihood function around the optimum and controls how the marginal likelihood function behaves around its maximum. The Hessian thus measures the stability of a model solution that corresponds to the maximum of marginal likelihood. As the referee suggests, if the complexity is inappropriate the model will not get selected, thus complexity ends up playing a role in Bayesian perspective.

Comment 1.p.3) Another emerging issue here: If a model is indeed very complex - has many parameters but the parameters their prior uncertainty appears relatively small. Then the model might be better constrained (more stable in wording of authors) then a model with far fewer parameters but that exhibit a much larger prior uncertainty. All this is taken care of in Bayesian model selection - if looked beyond simple criteria such as the AIC or BIC. So why not compare the arguments made here against full (numerical) integration of the prior and likelihood? This might make arguments more compelling. Because one can view p(D) as a measure of complexity as well. One that integrates quality of fit with uncertainty.

Response: We think that prior specification is user defined but we agree that the specification of a prior can be used to constrain (and partly stabilize) Bayesian model selection problem. We agree that one of the ways to compare complexity regularized method proposed here with Bayesian method can be its comparison with the numerical integration of the prior and the likelihood. Nonetheless, please note that numerical integration of prior and the likelihood is a research challenge in itself. Further, the authors have already presented a comparison of an approximation of the integral with complexity regularized model selection proposed here in Arkesteijn and Pande (2013). The authors there have further provided proof of concept that complexity regularized model selection works. Finally, the intent of this study is to use the complexity estimation method to argue that there is more model complexity than number of parameters.

Comment 1.q) Page 12: dimensions [1/T]ËĘ3 - rather awkward -> each recession parameter has unit 1/T -> not [1/T]ËĘ3

Response: Kindly note that it refers to the dimension of the set of the 3 recession parameters. Hence the dimension of the set should be [1/T]ËĘ3.

Comment 1.r) Page 18: "First we note that E[kBk] is the expected dierence in a model's simulations for two realizations of observations." –> unclear. What is meant by two realization of observations? Which observations? Forcing data. Unnecessary difficult to follow. Again, I only highlight a few of these places - many other sentences can be found that are confusing at best.

Response: Please note that by observations we meant input forcings. Two realizations of input forcings then mean a pair of input forcings of certain length that are sampled. By expected difference in model simulations we then mean the average of the 'distances' between model simulations over many different pairs of realizations. Here the 'distances' are measured by the norm used (for example l1-norm). Comment 1.s) Equation 4 - 11: Here things become confusing. Equations are provided but their relevance remains unclear - again first we need to understand what the expectation of B refers to? Two different realizations of input data? Precipitation data? Or all forcings combined? Dedfinition 1: Am I right that this is the difference between any simulated data point and the mean of these data points for a parameterization alpha? Why not word this - I highly recommend to explain each of the equations this way - and also to illustrate their calculation graphically in a plot. Just plot some data - calculate the mean of the data simulated by a model and then introduce Definition 1. Much easier to follow. Do so with each of the equations/definitions. Then it is much easier to follow for a reader.

Response: We will do so as suggested. The referee is right in pointing out that two different realizations refer to that of input data, of all relevant forcings combined. Further, the referee correctly points out that expectation of B is similar to the difference between

any simulated data point (in the output space) and the mean of these data points for a model parameterized by alpha.

Comment 1.t) Definition 2 is unclear. a model parameterized by alpha by gamma tilde? What is gamma and what is gamma tilde? Two different parameterizations? This is where things become rather unclear. Either use graphics - ideally combined with simple explanation in words. Do not hide behind equations - this will make things unnecessarily complicated.

Response: Kindly note that the model is parameterized by alpha while the expectation is denoted by gamma tilde. The two variables gamma and gamma tilde are two variables and their roles are explained in equation 4. Gamma is an arbitrary positive variable while gamma_bar is the expectation presented in Definition 1. We however will use graphics to explain this further.

Comment 1.u) How does the Markov Lemma come into play here? I miss the connection. What is X >= 0?

Response: X >= 0 refers to a variable X that takes up either 0 or positive values. Markov Lemma allows us to put an upper bound on P_N,gamma. Here P_N,gamma is defined in Definition 3.

Comment 1.v) Page 24: "By doing so we test whether the ordering in terms of its complexity of various model structure set-ups changes with different data sets. Insensitivity of the ordering of structure complexities to the data sets used for input forcings is crucial for any robust statement about the role of parameter magnitudes in determining model complexity" Difficult to follow. Many readers will have lost your arguments here – nor understand the underlying theory. Many elements need to be clarified before one can judge competence, relevance and importance.

Response: We will remove this statement.

Comment 2. How would you evaluate the complexity of an artificial neural network? If

you add such model to the analysis - would the ANN then come out as most complex? I need to see more the results of more than two models to evaluate the findings.

Response: Kindly note that we make a reference to ANN to highlight that "parameter magnitudes and number of parameters both influence model complexity" has been shown to hold for ANNs . We can provide a comparison with the complexity of ANN, however kindly note that a detailed assessment of the method of different hydrological models has already been provided elsewhere (Arkesteijn and Pande, 2013).

Comment 3: The present methodology requires an ensemble of forcing data - to evaluate the range of simulated output for a given parameterization. Is sampling of the prior parameter space not sufficient? Because fundamentally the approach that is presented herein differs from Bayesian model selection in that multiple inputs are considered.

Response: We agree, the referee has correctly pointed out that the present methodology requires an ensemble of forcing data. A comparison, however, of the frequentist method presented here with Bayesian approach may not be straight forward. We think model complexity in Bayesian approach is embedded within the marginal likelihood function (integral of prior and the likelihood). This the referee has also suggested previously. Sampling of prior parameter space may not be sufficient (even in the Bayesian approach), if the referee is implying that this sampling can be used to constrain the model selection problem. One might further need to investigate marginal likelihood function to select a model of appropriate complexity. The approach presented in this paper estimates model complexity prior to (independently of) the model selection step and computation of model complexity is independent of the observations of the response variable. However this is not the case within the Bayesian approach. For example Hessian of the likelihood function requires observations of both input and output variables, if the Hessian can be considered as a measure of Bayesian model complexity.

Comment 4. Going back to my earlier comment. The authors evaluate complexity

by looking at the output space of the model. This is one measure of complexity - the number of parameters used can be another measure of complexity - depending on their ranges as well. Thus one can define different measures of complexity - nevertheless - I believe the authors should benchmark their findings against common complexity criteria – ideally numerical integration of the posterior distribution (marginal likelihood).

Response: Kindly note that the model output space is specific to the hydrologic variable under consideration, for e.g. stream flow, evaporation or soil moisture. The way distances are measured in the model output space depend on the metric used, for example mean square error or similarity based on any signature that is devised. Thus, different estimations of model complexity can be obtained for different combinations of hydrological variable and metric used. However all these estimations will be based on model output spaces corresponding to the choice of hydrologic variable and the metric used. Further, any such estimate will also include the effect of the number of parameters. For example, Arkesteijn and Pande (2013) demonstrated that the complexity of a linear regression model is equal to the sum of squares of its parameter magnitudes, thereby measuring both the magnitude of individual parameter and the total number of parameters. Kindly note that the method has been benchmarked against a Bayesian measure on real world data and using a different set of models in Arkesteijn and Pande (2013). There the authors also presented a proof of concept that complexity regularized model selection results in stabler system representation (than when complexity is ignored).

Comment 5. The definitions the authors provide use a L1 norm for the distance between the simulated and mean data. How does their analysis hold if a different norm was used? Why use a L1 norm? Why not generalize this to any norm? L1-L2-L3-Linf.

Response: The arguments presented would not change as long as distance between two vectors is measured by a valid metric. We agree with the referee that it can be generalized to any norm.

Comment 6. I fear that focus on the model output space gives a rather limited view of complexity. I believe that author should focus on individual components of their model - and better recognize that if the goal is process understanding and analysis - that model output is not the way to go. Much better is to investigate specific metrics that are sensitive only to given components of the model. Such implementation would enhance significantly the impact of this paper. One can define summary metrics of the data and then use those to quantify complexity.

Response: We agree with the referee that complexity should be measured with respect to individual components of the model. However, complexities corresponding to these individual components can be estimated based on model output space since there is a model output space corresponding to a hydrologic variable and a choice of a metric. One may then obtain a summary of these complexities. Why our use of model output space is relevant is because it is the only measure of complexity at present that case summarize the complexity effect of the number of parameters and the magnitude of parameters. Arkesteijn and Pande (2013) have shown that model selection that is regularized with respect to this measure of complexity leads to stabler system representation (as the theory claims) based on real world datasets and a different class of hydrological models. It has also been benchmarked against a Bayesian measure. The authors have also demonstrated that the measure yields results that are similar to what has been demonstrated for other class of models.

Comment 7: What is the unit of complexity the authors are proposing? Something that should be clarified. Also is this metric relative or absolute?

Response: We thank the referee for raising this point. Since it measures the span of model output space, its units are based on the units of the output variable that is used to define the model output space. These units are then transformed by the norm ||.||. For example, if the output variable is streamflow in mm/day and the metric is mean absolute deviation then the units of complexity measure is mm/day. The complexity measure can also be made relative by for e.g. normalizing the output variable being

used (by subtracting the mean and dividing by the standard deviation).

Comment 8. Appendix B: How does this inequality hold for N data points (N>2)?

Response: Kindly note that the ||.|| is defined in N-dimensional space. Let it be a l1 norm and let the components of the vector A be defined as {a1,a2,..,ai,..,aN}. Then ||A|| - ||C||<= ||B|| + ||D|| can be restated as: the sum of the differences between the absolute values of the components of A and C are bounded from above by sum of the absolute values of the components of B and D. Kindly note that this inequality has been derived from two triangle inequalities. And we know that triangle inequalities hold in any dimension as long as valid metrices (see for example 3rd condition in Appendix A) are used to measure distance between the vectors.

Comment 9. Paper has many typos.

Response: We will address the typos in a future version of the study.

References: Pande, S., McKee, M., and Bastidas, L. A. (2009). Complexity-based robust hydrologic prediction, Water Resour. Res., 45, W10406, doi: 10.1029/2008WR007524.

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.

Pande, S. Arkesteijn, L and Bastidas, L. (2014). Complexity regularized hydrological model selection,In: Ames, DP, Quinn, NWT, Rizzoli, AE (Eds.), 2014. Proceedings of the 7th International Congress on Environmental Modelling and Software (iEMSs), June 15-19, San Diego, California, USA. ISBN: 978-88-9035-744-2.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 12, 3945, 2015.