Hydrol. Earth Syst. Sci. Discuss., 12, C2968–C2996, 2015
www.hydrol-earth-syst-sci-discuss.net/12/C2968/2015/

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: 4 August 2015

Response to the referee.

General response: We thank the referee again for his critical comments and respect the opinion that the central thesis of the paper is "not only demonstrably incorrect, it is antithetical to progress in the field". Our following responses (with no disrespect intended) hopefully persuades the referee otherwise. We also appreciate efforts undertaken by the referee to critically evaluate the ideas presented. But before we embark on our response to the referee, we will like to highlight that the central thesis of the paper is based on generalization theory and solving ill-posed problems that has been in existence for more than 40 years, successfully contributing to our understanding

of uncertainty (for example Vapnik and Chervonenkis, 1971; Tikhonov, 1963). The understanding hydrological model uncertainty has also benefited from frequentist perspective, see for e.g.(some earlier work include Montanari and Brath, 2003; Schoups et al, 2008; Montanari et al., 2009 and related references within).

Comment 1.1: "What the authors discuss in this paper has nothing to do with model complexity. The authors mention Occham's principle. Occham's principle says that we should choose the model with the fewest assumptions. Solomonoff used the principle correctly in the context of model inference, whereas the current authors use it incorrectly.

Response: We disagree with the referee on our interpretation of Occam's principle or of Occam's principle itself. Our interpretation of Occam' razor is based on Vapnik Chervonenkis generalization theory (Vapnik, 1995). It says to choose a model with lowest complexity from a set of best performing models (thus provide similar performance) on a given realization of observations.

We agree that the proposed measure of complexity is not the same as counting the number of assumptions as the referee suggests, except if it happens so in the special case of 'nested' model structures in the model output space. Model complexity 'may' then be monotonic in number of assumptions.

As already discussed in the previous response, our notion of model complexity is inspired by stability of model selection problems, which requires a 'stabilizer' (Vapnik, 1982; Cucker and Smale, 2001; Meir, 2000) to control variation in model simulations over different realizations of N observations. In this context, the measure of complexity should be such that when it is penalized, the model selection problem is stabilized. We therefore proposed a measure that is based on measuring the model output space.

Finally, we respect the opinion of the referee that 'Solomonoff used the principle correctly in the context of model inference, whereas the current authors use it incorrectly.'

Comment 1.2. The current debate about the Everettian interpretation of QM highlights this (somewhat common) mistake: (1; First Interlude, Page 2). "When you have two competing theories that make exactly the same predictions, the simpler one is the better." So, in the authors' example, we would use this principle to (at least a priori) prefer model $\wedge 1$ over model $\wedge 2$ because $\wedge 1$ has fewer ontological commitments (fewer processes). Instead the authors argue that we should prefer $\wedge 1$ (over $\wedge 2$) because of some characteristic of their predictions (namely, that predictions from $\wedge 1$ are are less dynamic in the sense that they respond less to inputs).

Response: We agree with the referee on this statement "When you have two competing theories that make exactly the same predictions, the simpler one is the better."

Where we disagree with the referee with the referee's interpretation of what is simpler. A qualitative statement/interpretation a model with less number of assumptions or less number of processes does not do justice to what is deemed 'simple'. The referee is also incomplete on the referee's specification of 'similarity'. What such an interpretation misses is context. It misses the metric of 'similarity' (of predictions), which is important because it is this metric which determines what is 'simple'.

Our notion of complexity and Occam's principle does exactly that (see for example Vapnik, 1995). A relationship between prediction uncertainty and the notion of 'simplicity' (or complexity) is identified in the model output space (see for example, Cucker and Smale, 2002).

While the referee does not give a justification why model $\wedge 1$ with fewer ontological commitments (fewer processes) should be preferred over model $\wedge 2$ (since the metric space in which the referee is measuring the similarity of predictions is not clear), we base our arguments on geometric interpretation and model complexity in (valid) metric spaces, where similarity in predictions between any two theories can be 'measured' and what is simple can be 'quantified' (or measured) in a consistent manner (by using triangle inequalities). The use of triangle inequality, to which we will refer to later again,

tells us that a theory A will make predictions that are 'closer' (as measured by the metric) to each other on different realizations of observations (each realization being the same for both A and B) if its output space is 'smaller' (measured by the metric) than another theory B. Thus, if theories A and B give similar predictions on one realization that is close to the observed, there is a higher chance that theory B gives a prediction that is way off from the truth than theory A.

We however do not rule out the possibility that model $\wedge 1$ with fewer processes is selected over model $\wedge 2$ but only in the case when: 1) Model $\wedge 1$ is nested within model $\wedge 2$ in the model output space and 2) Models $\wedge 1$ and $\wedge 2$ perform the same on a given realization of data.

We refrain from the use of terms such as 'dynamic ranges' to avoid unnecessary confusion.

Comment 1.3. I hate to argue semantics, But there is already a word for the idea of measuring distances in model output space due to differences in model inputs (what the authors notate here as ||B|| and refer to as "complexity"). These distances are measures of "sensitivity". The manuscript explicitly makes the argument that model sensitivity is the same thing as model complexity.

Response: We agree, the distances measure "sensitivity". Please see step 7 of page 3966 on what our measure of complexity exactly measures. It is the expected deviation of a model simulation from its expected value. The idea of measuring distances in model output space is behind the idea of model complexity proposed by others (Vapnik, 1982; Cucker and Smale, 2001; Meir, 2000).

Comment 2.1: More dangerously, this manuscript argues that we should a priori prefer insensitive models during induction. The stated motivation for this preference is that the dynamic range of a model bounds variation in prediction error, and that large variation in prediction error can lead to multi-modal inference posteriors (i.e., instability and equifinality).

Response: We disagree, we have never argued that. This is a mis-representation of our arguments. Please see equation 3 where it is clearly stated that model selection is a tradeoff between performance on a given realization of data and model complexity. The latter is, as described in the paper, measures the extent of the model output space. Formal, well founded, arguments are provided in the preceding section 2.2.2.

To emphasize again, complexity regularized model selection is not performed by minimizing the upper bound that the referee mentions above. Instead, it is a tradeoff between model performance on a given realization of observations and model complexity as quantified by Algorithms 1 and 2. The proof of the concept that such a model selection does indeed lead to stable/robust model selection, has been provided in Arkesteijn and Pande (2013).

Some intuitive arguments: prediction uncertainty describes the variation of model performance over different realization of observations. This is a function of model complexity. Thus, model performance on one realization of data and some function of prediction uncertainty (or model complexity) bounds the model's performance on other future unseen observations. Thus a tradeoff between model performance on one realization of data and some function of prediction uncertainty (or model complexity) allows us to select models that are less likely to poorly perform on future unseen observations.

Finally, we never argued that "The stated motivation for this preference is that the dynamic range of a model bounds variation in prediction error, and that large variation in prediction error can lead to multi-modal inference posteriors (i.e., instability and equifinality)" since our methods does not deal with posteriors! – such a suggestion is an extrapolation of our arguments in the terminology used by the referee.

Comment 2.2. The preference for unimodal posteriors is justified here based on an argument about uncertainty that is strictly and factually incorrect. Uncertainty is (incorrectly) defined in this paper as related to our ability to select between a set of candidate models.

Response: We fail to see how the referee has arrived to a conclusion that we prefer unimodal posteriors. We reiterate that our approach has no role for posteriors. Our approach is not Bayesian.

Nonetheless, we do suggest that uncertainty does influence our ability to select between a set of candidate models. Please see response to comment 2.1.

Comment 2.3. No a priori set of candidate models will ever contain a true model, so even if we integrated perfectly over the inference posterior, we would still have some unmeasured (and un-measurable) uncertainty related to the fact that the true model is not assigned a finite probability.

Response: We agree. We think the referee here is reiterating our response to his comments in the first round on the need to have full specification. However, even in the case of assigning a finite probability to the true model, it is not sufficient that a true model can be identified. Please see Pande (2013) and other references within such as Feldman (1991).

Nonetheless, marginal likelihood in our opinion is the best way to holistically incorporate model complexity in Bayesian model selection, since all other Bayesian treatment of model complexity are approximations of marginal likelihood. Further convergence of model inference based on criteria such as BIC does not require any assumption on the truth being in the set of models (Ye et al., 2008; Cavanaugh and Neath, 1999)

Comment 2.4: When there is no true model in the inference class there is no reason to prefer a unimodal posterior (i.e.,one that avoids either instability or equifinality). Instead we should prefer a posterior that assigns probabilities to each candidate model in proportion to their agreement with observations

Response: There is no reason to believe that we prefer unimodal posteriors since a) we donot deal with posteriors, b) our approach does not promote minimization of prediction uncertainty (instead a tradeoff between model performance on one realization

of observations and prediction uncertainty). Please see response to comment 2.1. and c) it is not clear how instability in our context leads to multi-modal posteriors – we appear to comparing apples with oranges

We do not see how assigning probabilities to each candidate model in proportion to their agreement with observations would lead to any better convergence to the truth, if the Bayesian problem of model selection is mis-specified. Assigning weights to each candidate models is similar to prior specification, which does not help the 'convergence to the truth' if none of the corresponding likelihoods support the 'truth'.

Nonetheless, we would like to highlight that BMA is one of the best techniques for model selection currently available, given that its convergence properties can be tracked (Ye et al., 2008; Cavanaugh and Neath, 1999).

Comment 2.5: The stated goal of the proposed strategy for regularizing inference is to constrain inference posteriors over classes of candidate models in such a way as to encourage those posteriors to be uni-modal –to converge uniformly to a single model as more data becomes available.

Response: We here hesitate to explore what the referee wishes to express. We donot see how he has arrived at a conclusion that we prefer unimodal posteriors.

We agree that our proposed frequentist approach provides only one solution. It provides a model that is most robust in representing the underlying system. However, multiple models can be obtained by changing the metric used. Please see our response to the comments of referee 1.

If prediction is the issue at hand and not system identification, non-uniqueness (equifinality) does not pose a problem. Hence multi-modal posteriors due to non-uniqueness are acceptable.

The link between instability and multi-modal posteriors is not clear.

Even if we explore an equivalence between our (frequentist) approach and Bayesian

approach: it possibly would be exploring a tradeoff between likelihood at the maximum and the curvature of the likelihood at the maximum (something similar to KIC) across models: the latter becoming a measure of complexity. This, as already stated in our first response, is an approximation of marginal likelihood. With full specification this is anyhow the best that can be constructively done within Bayesian model selection literature.

Comment 2.6. By artificially constraining the inference posterior, we articfically constrain our ability to *measure* uncertainty – we do not actually constrain uncertainty.

Response: We find this interpretation confusing. Complexity regularized model selection problem provides a robust system representation based on an upper bound on prediction uncertainty (please note that this does not mean we minimize this upper bound to obtain the model, please see our earlier responses). The model selection is based on constraining uncertainty since it is based on a tradeoff between model performance on a given realization of data and a function of prediction uncertainty (Arkesteijn and Pande, 2013), where the latter is penalized.

The Bayesian equivalence (KIC) will be approximation of marginal likelihood, which thus would maximize the likelihood that the data is generated by the selected model. For a given specification of Bayesian model selection, this is the best that can nonetheless be done for measure uncertainty, especially in the light of convergence arguments of Cavanaugh and Neath (1999).

Comment 2.7. In fact if the inference posterior is artificially constrained, then we actually *increase* that portion of uncertainty that is impossible to measure.

Response: Our approach does not have a posterior and so we do not artificially constrain it. Nonetheless, the statement "we actually *increase* that portion of uncertainty that is impossible to measure" requires full specification, i.e the knowledge of all uncertainty. One can increase a portion of uncertainty only if one can measure that portion of uncertainty.

Comment 2.8. The method proposed here is justified via an argument to improved uncertainty management (actually using an argument that it will reduce uncertainty). In actuality, it only hides uncertainty by reducing our ability to measure uncertainty, and increasing the portion of uncertainty that is fundamentally not measurable. The effect of what is proposed here is to decrease our ability to manage uncertainty.

Response: Please see our responses above.

Comment 3.1 The justification for this paper is not the triangle inequality –that is the tool used to implement the idea. The justification for the proposed regularization scheme is that it will help attenuate three problems: i)non-uniqueness, (ii) equifinality, and (iii) non-identifiability (actually, these are really only two distinct issues).

Response: We appreciate the insights of the referee, but the triangle inequality is the key to demonstrate that the regularization scheme will 'stabilize' model selection problem. At best, what we would like to demonstrate that the regularized model selection problem will deal with i) non-existence of a solution (not an issue), ii) non-unique solution and iii) instability of a model selection problem. A formal proof can be provided but see Vapnik (1982), proofs of theorem A.1 on page 24 that will the basis. Please see our response to referee 1.

Comment 3.2: A convergence proof is necessary for two reasons: (1) to show that the proposed regularization scheme actually does mitigate these problems, and (2) to show how this regularization scheme affects convergence rates. That is, in the presence of only finite inference data, what is the potential for this regularization scheme to cause us to make a mistake, and prefer a non-optimal model?

Response: What the referee means by convergence proof is now clear. The convergence proof will require additional conditions on lambda, the penalty factor (see equation 3). Evidence to support the convergence of such a regularization scheme has been provided in Arkesteijn and Pande (2013). This was in response to similar comments made by a referee.

It shows how regularized model selection stabilizes model selection problem on small data sets and how solution of regularized model selection converges to solution of un-regularized model selection as the sample size increases.

Comment 3.3. A proof of convergence is needed. This is non-negotiable. If you propose a regularization scheme for any type of induction, you must show how this regularization affects rates of convergence. The only alternative is to remove all claims about regularizing inference procedures in the manuscript (essentially, the entire introduction), in which case there is no motivation for calculating what is here (incorrectly) called "model complexity"

Response: We can provide a proof of convergence. Please also see our response to comment 3.3.

Comment 4: Citing your own work is not terribly convincing. In fact, you made the same errors there, but apparently no reviewer caught them

Response: We respect the opinion of the referee. What surprises us is that almost all the arguments of the referee till now has been extensively discussed in the literature cited in responses to comments 3.2

Please see our response to comment 3.2 above and elsewhere when citing our own work. Many of the concerns raised by the referee were addressed already in this paper.

Comment 5.1: Yes it is. By Cox' theorem (2) all induction is Bayesian(3) unless you violate one of the three Aristotelian axioms (4) (e.g., (5)). You propose to use your "complexity measure" to regularize the problem of model inference (i.e., the discussion about non-identifiability, equifinality, etc).

Response: We respect the opinion of the referee.

As earlier said, we would like to refrain from the debate of whether all induction exercises are Bayesian given that arguments exist that argue otherwise. For example please see Vapnik (1995), page 116-118.

Comment 5.2: In particular, any regularization scheme imposed on an inference procedure is equivalent to specification of a Bayesian prior.

Response: We agree that prior specification can be shown as being equivalent to regularization model selection problem 'under full specification' (otherwise) but we fail to see how this makes model selection approach presented here to be the same as Bayesian model selection approach.

We fail to see how any regularized model selection, for e.g. the one presented here, can be 'constructively' converted into Bayesian model selection exercise with a particular specification of a prior.

Such a step has implications for the specification of likelihood function, which is not as straight forward as the referee has suggested.

Comment 5.3 All I have done is to use the language of Bayesian inference to discuss what is proposed in this manuscript, and since the topic of this manuscript is induction over hypothetical models, this is an appropriate language.

Response: We respect the opinion of the referee. We respectfully disagree.

The language has only added confusion. Please see our response to earlier comments. Since we do not obtain a posterior (this can be shown), how can arguments based on the type of posterior be appropriate when discussing the approach that we have presented.

Comment 5.4. If you were to drop the discussion about equifinality, identifiabiltiy, and about how this complexity measure might help with these problems, then you can claim that your approach is not Bayesian, because then you would not be proposing any aspect of an inference procedure. But then I see no motivation for calculating model complexity in the first place

Response: The above argument of the referee is circular.

The issues of non-uniqueness (equifinality) and instability can be discussed without invoking posteriors or Bayes law (Vapnik, 1982; page 308). As also stated in the previous response to the referees, unstable model selection problems are ill-conditioned. The measure of complexity as proposed here 'stabilizes' such problems.

Comment 6.1: Of course you don't. Such a thing is impossible. This is my point.

Response: Naturally, we agree. Never in our manuscript we mention identification of the 'truth'.

Comment 6.2: Your definition of uncertainty ("We characterize uncertainty in hydrological system representation as composed of non-uniqueness and instability in system representation") only works if the true model is in the class that you are performing inference over. It is precisely because you don't (and can't) propose to test a true model that your definition of uncertainty fails.

Response: What we fail to understand is why the referee thinks we ignore the issue of model deficiency. If instability in model performance (uncertainty in system representation) is zero and let non-uniqueness not be present, system representation is certain but may still not be accurate, i.e. model performance may not be perfect. A posterior can be sharp and uni-model but can still be biased (see for example the synthetic example in Pande, 2013).

Comment 6.3. To be clear, instability and non-uniqueness refer to the fact that inference with different finite data sets results in preference for different models from the inference class (different posterior modes), and equifinality refers to the fact that, given a single finite set of inference data, there are sometimes many local modes in the posterior.

Reponse: We do not agree, instability refer to the fact that inference with different finite data sets results in preference for different models from the inference class and non-uniqueness (equifinality) refers to the fact that, given a single finite set of inference data, multiple solution to a model selection exercise exist.

Comment 6.4: To restate what I said in my original comment, Defining uncertainty related to these issues is insufficient because uncertainty also must consider the fact that, inevitably, whatever set of candidate models that we propose will not contain a "true" model.

Response: We agree that model deficiency should be considered in a model selection exercise. Our definition of uncertainty does allow for model selection and hence allows for the case when the set of candidate model do not contain the true model. We are well aware of this. Please see response to comment 6.2.

We fail to see why the referee thinks that our approach does not consider this. We have emphasized this in our response to the last round of comments. Please see Arkesteijn and Pande (2013) (again with apologies). Equation (4)-(7) clearly demonstrate how one goes from prediction uncertainty to a model selection problem that minimizes a tradeoff between model performance on one realization of data and a measure of complexity (exactly the same measure of complexity that we deal with in this paper). This acts as an upper bound on model performance when very large amount of observations are available. No assumptions are made that model deficiency is 0, since it is never assumed that model performance on very large amount of observations is 0.

The prediction uncertainty defined by variability in model (parameterized by α) performance, measure by a risk function $(\xi_N(\mathbf{y^0}, \mathbf{x}; \alpha)$ (higher risk $\Leftrightarrow$ poorer performance), over different realizations can be defined as (additional steps can be provided if needed):

$$\Pr\left(|\xi_N(\mathbf{y^0}, \mathbf{x}; \alpha) - E[\xi_N(\mathbf{y^0}, \mathbf{x}; \alpha)]| > \gamma\right) \tag{1}$$

.

Here $E[\xi_N(\mathbf{y^0}, \mathbf{x}; \alpha)]$ represents the expected risk of the model (average over all possible sample realizations and $\gamma \geqslant 0$. This is one measure of model deficiency. Also, $(\mathbf{y^0}, \mathbf{x})$ represents a set of observations of an outcome variable of interest $(\mathbf{y^0})$ and

input forcings $(\mathbf{x})$ of size N.

It can be shown that this is bounded by a function of complexity, the same measure that we discussed in the paper, where $\eta > 0$ is some constant.

$$\Pr\left(|\xi_N(\mathbf{y^0}, \mathbf{x}; \alpha) - E[\xi_N(\mathbf{y^0}, \mathbf{x}; \alpha)]| > \eta\gamma\right) \leqslant \frac{f(h,N)}{N^2\gamma^2} = \frac{F(h,N)}{\gamma^2} \tag{2}$$

.

Using the above, it can then be shown that for some $c_N \geq 0$ the following holds:

$$E[\xi_N(\mathbf{y^0}, \mathbf{x}; \alpha)] \leqslant \xi_N(\mathbf{y^0}, \mathbf{x}; \alpha) + c_N\sqrt{F(h,N)} \tag{3}$$

What is minimized in the regularized model selection problem is the right hand side of the above. It is never assumed that model deficiency is 0. This is because if we did what the referee says we did, we will be minimizing only the one addendum of the right hand side, i.e. $\sqrt{F(h,N)}$ which essentially will minimize our notion of prediction uncertainty.

Comment 6.5 Thus if we aim, as claimed in this paper, to reduce instability and equifinality (which are characteristics of the inference posterior), then we do NOT necessarily reduce uncertainty, we simply reduce that portion of uncertainty that is quantifiable by integrating the posterior. We artificially narrow or otherwise constrain the posterior.

Response: We disagree, it not the aim of the paper to 'just' reduce instability and non-uniqueness. Also, we also donot do what the referee claims we do, i.e. that we artificially narrow or constrain the posterior. There is no use of a posterior here.

The closest equivalent of our complexity regularized model selection approach will be maximizing an approximation of marginal likelihood (or numerical approximation of marginal likelihood itself). If referee means that maximizing marginal likelihood is the same as uncertainty that is quantifiable by integrating the posterior, then we disagree

with his statement that “then we do NOT necessarily reduce uncertainty, we simply reduce that portion of uncertainty [..corresponding to instability and non-uniqueness] that is quantifiable by integrating the posterior.” We approximately reduce overall uncertainty as quantified by the marginal likelihood (or integrating the posterior in this sense).

Comment 6.6. That is, if the proposed regularization scheme does what the paper claims it does (we don’t know because no proof is provided) then the result is toartificially reduce aleatory uncertainty at the expense of increasing non-aleatory epistemic uncertainty(see (6) for definitions of these terms). That is, we reduce the portion of uncertainty that is represented by the inference posterior at the expense of increasing the portion of uncertainty that is not represented by the posterior. This is precisely what we do *not* want to do.

Response: This paper is not about the regularization scheme. Arkesteijn and Pande (2013) is about regularization scheme, upon which this paper builds. Empirical support that regularization scheme works is provided in that paper. However, given the comments 6.1-6.5 and our response, we disagree with the referee interpretation of our approach presented here.

Nonetheless, as we have also responded earlier to previous similar comments, we fail to see how one can make a statement on increasing the portion of uncertainty that is not measurable. Full specification is not a requirement if the objective is to select the best available approximation of the truth and not to describe the truth (see for e.g. Cavanaugh and Neath (1999) on regularity conditions for convergence).

Comment 7. That is not what I said. Any inference procedure should consider these two things PLUS the reality that any model we test is wrong. As Beven routinely points out (e.g., (6)) it is wholly insufficient to consider only these two issues (which both contribute to aleatory uncertainty, as both non-uniquness and instability are properties of the posterior and thus quantifiable), and to ignore the issue that I discussed (which

is a non-aleatory aspect of epistemic uncertainty).

Response: Apologies for the misunderstanding. Please see our response to comments 6.1-6.6. We never ignored the issue of model deficiency.

Comment 8.1: Yes, you do focus solely on this aspect of uncertainty. To the detriment of the real issues related to uncertainty (again, as Beven routinely accuses the community of doing). Again, instability qua instability is NOT an issue. Instability and equifinality are *good* things if they stem from a real representation of the fact that none of the models that we test are perfect, and thus our finite data cannot distinguish uniformly between candidate models.

Response: Please note our comment referred to stabilizing a model selection problem and not taking care of non-uniqueness (equifinality). Also, as we have responded previously, we do not ignore model deficiency in our proposed approach.

The referee then is implying that unstable model selection problems are ‘good’, that there is no need for regularization on finite data.

We disagree. Complexity regularized model selection will select a model that reveals itself to be a bad choice on future unseen data with a no larger probability than any other model. Why we use ‘no larger’ is because there can be cases where system representation does not require complexity regularization due to the set of model used for inference. We here emphasize again that the approach does not ignore model deficiency.

Regarding the need to obtain a collection of models, the proposed approach depends on the choice of response variable and the choice of metric (please see our response to the comments of referee 1). Thus a collection of models may be chosen that represent different processes or different quantiles of the same response variable.

Comment 8.2. This is real uncertainty and should be represented in our inference posteriors. Forcing our representations of this real facet of uncertainty causes us to

underestimate real uncertainty. If the posterior is multimodal (perhaps even in the limit), then we Simply recognize uncertainty in the model selection process related to the fact that our candidate model set is incomplete. Of course, this does not fully measure epistemic uncertainty, but at least we are not artificially reducing our ability to quantify epistemic uncertainty.

Response: The regularization factor that determines the tradeoff between model performance on a single realization of data and model complexity is not exogenous. It depends on the problem and in practice it is first estimated on oservations itself before it is applied for regularizing a model selection problem. A 0 estimate for the regularization constant c_N is possible. Please see Arkesteijn and Pande (2013) on how a regularization constant can be estimated.

Comment 8.3. The problem here is that the a priori objective of this paper is to reduce that part of aleatory uncertainty that is related to finite inference. The stated aim is to under-estimate uncertainty! This is not a consequence of the method, it is the stated objective! Setting an a priori goal of reducing instability is not appropriate, and is actually dangerous.

Response: We have emphasized again and again (as we had also done in the previous round) that we donot ignore model deficiency, that our method does not depend on the assumption that truth is in the model space. Please see our responses to comments 6.1-6.6.

Yes, we seek to reduce uncertainty in system representation by stabilizing model selection problem. This will ensure that we choose a model that reveals itself to be a bad choice on future unseen data with a no larger probability than any other model. We fail to see why it is so antithetical or dangerous.

Please see our response to 8.1 and 8.2 as well.

Comment 9: Yes, this suggestion is why the manuscript crosses the line from simply

wrong to dangerously wrong. I want to reiterate that this is not a matter of rejecting a new idea. This is a matter of rejecting an idea that makes a very well understood mistake – a mistake that has been discussed in our literature for almost two decades. I have rarely seen such a clear example of this mistake. You are doing here precisely the wrong thing – trying to force the inference procedure to prefer a single (incorrect) model. Instead we should be working for our best estimates of the (possibly multimodal) inference posteriors, so as to appropriately characterize within-class uncertainty related to model selection, and to understand how this in-class variability relates to real uncertainty (if it does, at all).

Response: Through our method we suggest to select a model that reveals itself to be a bad choice on future unseen data with a no larger probability than any other model. We agree that we suggest only a single model, and (as also provided in the response to referee 1) a distribution of models can be selected based on various combinations of response variable and metric used (as shown elsewhere that one can represent uncertainty by using asymmetric loss function as a metric).

To move from a single model choice to multi-model choice is not an issue. However, we disagree with the referee's implication that since we agree that a distribution of models is a wiser choice than a single model, we should go for any multi-model distribution. We agree that the need to have more than 1 model is known for more than 2 decades but there can be more than one way to respond to this need.

Comment 10.1. Of course, it is true that inference balances the prior with the likelihood –i.e., in this case balances the prior that favors a lack of model sensitivity – not complexity –with the whatever error function is used to measure predictive power (this error function necessarily implies a likelihood). This is why Occham's razor must be applied *only to distinguish between models that make the same predictions*. Because otherwise you use this metaphysical principle to a priori prefer simpler models that make worse predictions.

Response: Our interpretation of model complexity is not new (Meir, 2000; Cucker and Smale, 2002) though we respect that this referee chooses to call it model sensitivity. Model complexity has a role to play in ill-conditioned model selection problem and the definition of model complexity is consistent with that.

We further disagree with the suggestion of the referee that "this error function necessarily implies a likelihood". It is nonetheless well known that minimization of an error function may imply maximization of a likelihood. This is yet another reason why we disagree with the referee's extrapolation of our approach to desiring a uni-modal posteriors.

Finally on "Because otherwise you use this metaphysical principle to a priori prefer simpler models that make worse predictions.": we find this statement incomplete. We agree that we use the principle to prefer less complex (simpler) models to make worse predictions 'on a given sample of data'. The proposed approach does not do it arbitrarily. The method ensures that the chosen (less complex) model on one set of observations will reveal itself to be a bad choice on future unseen data with a no larger probability than any other model (that the chosen model is a robust choice).

Comment 10.2. Things like AIC, BIC, KIC all make this same compromise, however, the priors in each of these metric (yes, each have a term representing a Bayesian prior), are all justified via some reasonable a priori principle

Response: As we have been emphasized earlier, there is more to them than prior specification (our approach accommodates a general treatment). We agree that AIC and BIC differ only in the specification of prior but BIC and KIC do not. BIC is KIC for large sample size and flat prior. KIC can accommodate any prior specification. Finally, there are two more steps (or element to it than just prior specification), that is of the conditions under which Information matrix equality holds and Laplace method for approximating an integral to go from marginal likelihood to KIC.

All 3 are approximations of a marginal likelihood but AIC and BIC can be seen as

special cases of KIC for large N. Thus they all do not make the same compromise: AIC and KIC makes more compromise by specifying a prior than KIC – there is more to regularization than just specification of priors.

This is exactly the issue here, there is more model complexity than prior specification. As we have also emphasized earlier (contrary to the referee's suggestion), prior specification can lead to certain regularization but any regularization implies a prior specification may be incorrect. Representation of our method to a Bayesian equivalent of prior specification is incorrect. As we have mentioned before, our method is perhaps equivalent to a KIC version with Hessian instead of Fisher information matrix.

Comment 10.3. Moreover, all are associated with some convergence description (a convergence proof).

Response: We agree, and it does not require full specification. Further the proof is needed only for KIC and is more based on the notion of consistency (see for example regularity conditions of Cavanaugh and Neath, 1999). But given that all are approximation of a marginal likelihood, a numerical integration of such a method is most desirable to compare models.

Comment 10.4: All regularization schemes are Bayesian priors, and must therefore be based on some principle that is justifiable in an a priori sense.

Response: We disagree in case of non-Bayesian approaches, please see our response to comment 10.3. There is more to it than specification of priors.

Comment 10.5: The a priori justification for the proposed regularization scheme is literally that it reduces our ability to quantify uncertainty (not that it actually reduces uncertainty).

Response: Please see our response to comments 10.3 and 10.4. We donot agree with the conclusion that the referee is drawing. As we have shown above, the impression that all regularization is prior specification is incomplete and drawing conclusions based

on such an impression can be misleading.

Comment 11: Yes, and this is why you need a convergence proof. Is this tradeoff handled correctly? Does this regularization increase or decrease the rate of convergence to the true model?

Response: Please see our response to comment 3.2.

Comment 12: Yup, but now any predictions made from the resultant posterior distribution over models is more deficient than it was before we implemented the proposed regularization because it artificially narrows the posterior probability distribution.

Response: The reasoning here of the referee is incomplete. We donot deal with posteriors. Further regularization is not just about prior specification. It is misleading to suggest that our regularization is the same as looking for a unimodal posterior since our approach is not just about minimizing prediction uncertainty. If there is an equivalent to our approach in Bayesian induction, it is with maximizing a marginal likelihood.

Comment 13: Not really. Aleatory uncertainty is uncertainty that can be quantified via probability distributions, and epistemic uncertainty is uncertainty related to the fact that our models are incorrect (6). Notice that portions of epistemic uncertainty can be quantified, and are thus also aleatory –the two are not a strict duality. To be specific, if we have a distribution over potential models, then this distribution is both aleatory and epistemic since it is both due to our lack of complete information about the specification of the system, but is also quantifiable. But since this distribution necessarily does not consider all possible models (the set discussed, for example, by (7, 8)), then there is an additional component of epistemic uncertainty that is not quantifiable, and therefore not aleatory. What the proposed method does is to decrease the quantifiable component of uncertainty (by encouraging unimodal posteriors),at the potential of increasing the non-quantifiable component (i.e., by getting the posterior wrong).

Response: Please see our response to comment 12. The referee argument that we

"decrease the quantifiable component of uncertainty (by encouraging unimodal posteriors)," is misleading.

Comment 14: So what exactly does this paper offer over Arkesteijn and Pande (2013)?

Response: The mentioned paper provided evidence of convergence of the proposed method (complexity regularized model selection, with complexity defined as in this paper). It showed how regularized model selection stabilizes model selection problem on small data sets and how solution of regularized model selection converges to solution of un-regularized model selection as the sample size increases. It also compared this method (complexity regularized model selection) with a Bayesian criteria.

This paper is not Arkesteijn and Pande (2013). This paper formally presents the algorithms that quantify model complexity. It further builds on this measure of complexity to suggest that there is more to model complexity that just counting the number of parameters of a model.

Comment 15: In what way is Kolmogorov complexity not "constructive"? See, for example, (9), who use it constructively in the context of hydrological model selection. Given that they use a comprehensive definition of uncertainty (you do not), that they use an appropriate definition of complexity (you do not), and that they appropriately apply Occham's principle (you do not), I ask again: what does your method offer over existing methods?

Response: We respect the opinion of the referee, in particular his opinion that 'that they use an appropriate definition of complexity (you do not), and that they appropriately apply Occham's principle (you do not)'. Please see our response to comments 1.1 and 1.2.

However the comment of the referee is not constructive. We fail to see in this comment how Kolmogorov complexity has been constructively used in model selection, how the model selection problem was specified and how model complexity enters in

this problem of model selection.

Finally we would be surprised that the reference cited has used the mentioned complexity for hydrological model selection. This is because finding a smart quantization is an extremely hard problem for a given set of functions, such as hydrological models. We are sure that authors are on it, given that they mentioned in the paper that this task is for future work.

Comment 16.1: Absolutely this is false. In fact all that the paper shows is that the *difference* in residuals (not the actual magnitudes of the residuals) is bounded by the dynamic range of the model (what the authors call “complexity”) PLUS the dynamic range of the observations. Not only that, but uncertainty is *not* strictly related to model residuals.

Response: We respect the opinion of the referee. We agree that this is what the proposed method presented in the paper is based on. This is consistent with our definition of prediction uncertainty, and bounds it.

From this definition of prediction uncertainty, one can construct a bound on the expected value of a performance measure (a function of residuals). Please see response to comment 6.4.

Different performance measures (combination of response variable and metric used) can be used to informally measure other definitions of uncertainty.

Comment 16.2: There is absolutely no sense in which the proposed sensitivity measure measures uncertainty.

Response: Please see our response above. Our proposed measure of complexity is instrumental in bounding prediction uncertainty from above. This does not mean we ignore the issue of model deficiency. If instability in model performance (uncertainty in system representation) is zero and let non-uniqueness not be present, system representation is certain but may still not be accurate, i.e. model performance may not

be perfect. A posterior can be sharp and uni-model but can still be biased (see for example the synthetic example in Pande (2013).

Comment 16.3. This is really important. The authors argue that ||B|| should be low because this will favor unimodal posteriors. But .. if ||D|| is high, then it is entirely possible for the model to make very bad predictions. Simply minimizing the dynamic range minimizes the *differences* between model residuals, but it does not minimize model residuals at all –if the observations exhibit dynamical response but the model doesn't, then ||D|| is large, ||A|| is large, ||C|| is large, but ||B|| is small. ||B|| has nothing to do with measuring uncertainty.

Response: Once again, this is a misunderstanding that we favor unimodal distributions.

This mis-understanding is clearly evident here. We do not propose model selection based solely on minimization of ||B||. Our approach to model selection proposes to minimize a tradeoff between a function of ||B|| and performance on a given sample of data, say ||A||. Please see our response to comment 16.1. and comment 6.4.

Comment 17.1: Occham's razor does not argue to prefer models that make simple predictions, it argues that we should prefer simple models that make accurate predictions.

Response: We agree. Our method as well does not propose to pick models that make simple predictions (we donot solely minimize ||B||).

Comment 17.2: Parsimony is favored in the ontological requirements, not the ontological consequences of the model. What is proposed here is *not* Occham's razor. It is also not in any way related to model complexity.

Response: We believe metrices are important here. A metric that examines similarities and differences between various concepts and relationships should be consistent with the metric that examines similarities and differences between the consequences of these concepts and relationships. This is a must for a consistent interpretation of Occam's razor. Counting the number of parameters of a model may not sufficiently

quantify model complexity. Please see our response to comments 1.1 and 1.2.

Comment 18: Yet none of that literature is cited in the manuscript. I cited it in my review, but the authors did not. If the authors actually took the time to formally compare what they propose against existing methods, then I expect they would not have submitted the idea - as it is a very bad idea.

Response: We respect the opinion of the referee.

We indeed cited relevant work in our paper. For example, please see lines 16-21 on page 3974. Assessment of complexity regularized model selection has been provided in Arkesteijn and Pande (2013). The intention of this paper is to reinforce evidence provided elsewhere, for example see references in lines 16-21 on page 3974, that model complexity is a function of number of parameters and the magnitude of parameters. The authors believe that the underlying theory is very strong (see for example Cucker and Smale, 2002).

Comment 19: There is a prior implied by your regularization scheme (like all regularization schemes). Just because you have not taken the time to derive the implied probability distribution (like (10)and others did), does not mean that it does not exist –it simply means that you have neglected a formal statement of your idea.

Response: We appreciate referee's Bayesian take on our approach. As said, our arguments are geometric and such a view on uncertainty assessment is formal. Nonetheless, again we do not find this comment constructive.

Why, because as the author (of the reference cited) himself states "The author feels that all problems in inductive inference, whether they involve continuous or discrete data, or both, can be expressed in the form of the extrapolation of a long sequence of symbols. . ., the known sequence of symbols is very long, and contains all of the information that is to be used in the extrapolation.

By a formal solution is meant a mathematical equation that in some sense expresses

the probability desired as a function of the sequences involved. It will not, in general, be practical to evaluate the probability directly from this equation. In most cases, there is some question as to whether it is even possible in theory to perform the indicated evaluation."

Further, without fully specifying the model selection problem (i.e. our incorrect set of models in tandem with correct specification for residuals), this Turing machine will not be able to tell us which model is most appropriate since after all it will take in sequence of residuals (conditional on the set of models used) estimated on an infinite set of observations. Even if one is able to construct such construct, we will not be surprised if the Turing machine tells us that there is more to model complexity than number of parameters- that parameter magnitudes matter as well.

Comment 20: Doesn't change the fact that you are performing inference, and therefore your regularization scheme can be interpreted as a Bayesian prior.

Response: We agree that we are performing inference but we fail to see how a regularization scheme can be interpreted as a Bayesian prior (prior specification as regularization is valid in Bayesian selection problem, based on Bayes law). This is an Bayesian extrapolation that we fail to associate with. Please see our response to comment 10.2 as well.

Comment 21.1. Yes, every inference procedure requires full specification of the probability distribution from which observations are generated. Often, we use simplified error models, but every error model implies a probability distribution(11, 12).Even if we ignore observation uncertainty, we have simply used a Dirac distribution.

Response: We agree that Bayesian inference procedure requires full specification of the probability distribution from which observations are generated, as we have discussed this elsewhere. Except that we donot associate with referee's comment on Dirac distribution. Please consider the case of sampling uncertainty (even when measurement uncertainty is absent) in tandem with model deficiency.

Comment 21.2. It is strictly impossible to perform inference without full specification of the prior and likelihood, and nothing that the authors have proposed changes this reality. Any inference procedure that they might apply their regularization scheme to will also necessarily include full specification of an error distribution.

Response: This is clear evidence that the author has misunderstood our approach. Our approach is based on weak apriori information, it does not require full specification of any distribution (Vapnik, 1995). Our approach is distribution free.

This is the reason why this approach works on upper bounds of 'true' error (that is expected value of model performance with respect to the underlying but unknown distribution). Had we made the assumption of full specification, we would have worked with the (assumed) 'true' error.

Nonetheless, as Cavanaugh and Neath (1999) have shown, the use of KIC without full specification is justified, given certain regularity conditions hold.

Comment 22: This is false. There is no need for a non-parameteric pdf in Solomonoff induction

Response: There must be a reason why it is called a 'universal' probability distribution (In: Hutter, M., Universal Algorithmic Intelligence. 2003, Technical Report IDSIA-01-03). Please see our response to comment 19.

Comment 23: Anyway, I could go on, but this should be sufficient. The problems with this article are fundamental –the idea is flawed and is actually counter-productive to meaningful progress in uncertainty estimation. This really shouldn't be a matter of debate, as the errors made here are of a type that is common, widely-recognized, and well-documented in hydrology and other literature.

Response: We respect the opinion of the referee.

We believe most of the comments of the referee regarding the accuracy of our approach are either a consequence of misunderstanding our approach or misrepresen-

tation of our approach.

Finally, we believe that a meaningful progress in uncertainty estimation cannot be made by institutionalizing ideas that have been in a field for two or more decades (list of references is Bayesian but other perspectives on uncertainty exist such as in our list of references for example) but instead enriching these with fresher perspectives on uncertainty.

References: Cavanaugh, J. E., and A. A. Neath (1999), Generalizing the derivation of the Schwarz information criterion, Commun. Stat. Theory Methods, 28,49–66.

Cucker, F., and S. Smale (2001), On the mathematical foundations of learning, Bulletin of the American Mathematical Society, 39(1), 1-49.

Feldman, M. (1991), On the generic non-convergence of Bayesian actions and beliefs, Econ. Theory,1, 301–321.

Meir, R. (2000), Nonparametric time series prediction through adaptive model selection, Machine Learning, 39, 5-34.

Montanari, A. and Brath, A. (2004) A stochastic approach for assessing the uncertainty of rainfall–runoff simulations, Water Resour. Res., 40, W01106, doi:10.1029/2003WR002540.

Montanari, A., C. A. Shoemaker, and N. van de Giesen (2009), Introduction to special section on Uncertainty Assessment in Surface and Subsurface Hydrology: An overview of issues and challenges, Water Resour. Res., 45, W00B00, doi:10.1029/2009WR008471.

Pande, S. (2013), Quantile hydrologic model selection and model structure deficiency assessment: 1. Theory, Water Resour. Res., 49, 5631–5657, doi:10.1002/wrcr.20411.

Schoups, G., van de Giesen, N. C., and Savenije, H. H. G.(2008). Model complexity control for hydrologic prediction, Water Resour. Res., 44, W00B03,

doi:10.1029/2008WR006836.

Vapnik, V. (1995), The Nature of Statistical Learning Theory, Springer, New York.

Vapnik, V., and A. Chervonenkis (1971), On uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl., 16(2), 264–280, doi:10.1137/1116025.

Vapnik, V., E. Levin, and Y. LeCun (1994), Measuring the VC dimension of a learning machine, Neural Comput., 6(5), 851–876, doi:10.1162/neco.1994.6.5.851.

Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, Water Resour. Res. ,44, W03428, doi:10.1029/2008WR006803.

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