Hydrol. Earth Syst. Sci. Discuss., 12, C2885–C2899, 2015
www.hydrol-earth-syst-sci-discuss.net/12/C2885/2015/

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 2:

We thank the referee for such an in-depth criticism of the manuscript. Following is our response.

General response: We are unable to see where a proof of convergence is needed (we miss the context). The justification of our idea comes from the triangle inequality that we have illustrated in the first few figures. It suggests that variability in a model's response to different realizations of input forcings is bounded from above by complexity, where complexity is a function of model output space. The former is linked to prediction uncertainty or uncertainty in system representation. Such an idea is not new and has

been used as a basis to regularize model selection problem. We agree that the inequality does not mean that prediction uncertainty is tightly linked to model complexity. However, this inequality can be used (and has been used in literature on complexity regularized model selection, see for e.g. a discussion on this in Arkesteijn and Pande (2013) and references within) to control prediction uncertainty by controlling for model complexity. We agree that not all definitions of the terms used may have provided beforehand. We will improve upon this aspect in a future version.

Comment: "..some desiderata for finite inductive inference",

Response: Our approach is not about inferring a model from a posterior. Our approach is not Bayesian. Also, we never claim that we would like to test a 'true' model. However, we agree with the referee that any model selection should deal both with the issue of non-uniqueness and instability. We focus on the issue of instability in our paper, i.e. we propose an approach to stabilize a model selection problem through complexity regularized model selection. What we are suggesting is that such a regularization will reduce uncertainty in system representation due to increased stability. Comment: "it is important to point out that putting a constraint on E||B|| effectively favors models with a smaller dynamic range. E||B|| is minimized when the function is a constant and generally favors models with no sensitivity to their inputs. My intuition is that this type of regularization will *increase* uncertainty (defined somewhat more comprehensively)...,":

Response: We agree that putting a constraint favor model with smaller model output space (and hence a model with lower complexity), but complexity regularized model selection does not necessarily pick a model with lowest complexity. A complexity regularized model selection will trade off model performance on one available set of observations with model complexity. It will thus tradeoff epistemic uncertainty with aleatory uncertainty. The extent to which it will reduce epistemic uncertainty upon uniform convergence (as the referee puts it) will depend on the class of models that were used. If the class of model contains the "truth", model that is selected will converge to this

truth. If the class of models does not contain the truth, complexity regularized model selection will give a model solution that is still 'deficient' in explaining the truth.

Comment: "A single empirical analysis is presented, however it is trivial to show that this analysis is a special case. The authors make two false generalizations about nested models (see below), however even without these errors there is no way to know whether the empirical results here can be generalized. The example given represents a particularly simple type of nested structure, and so there is no chance for a different model structure to contribute different information during different periods (see, for example, [1], [2] for senses in which we may want different models for different data)."

Response: The use of empirical analysis was solely to illustrate how the inequality works. Arkesteijn and Pande (2013) provide a more exhaustive analysis of the complexity regularized model selection approach presented here (with different model structures and real world data). We agree that at present there is no chance to for a different model structure to contribute to different information during different periods.

Comment: "Another thing that should be mentioned is the relationship between this idea and existing related theories. The objective of the paper is to define the concept of model complexity in a way that is measurable in model output space. The purpose is to set up a complexity-based regularization constraint on inductive inference. Of course, there already exists a definition of model complexity[3] that is used to regularize the type of inductive inference[4] known as universal induction[5]. This theory is used in hydrology[6]. Similarly, the ||B|| value, which is used here as the basis for a measure of "complexity", is actually a primitive measure of sensitivity[7], not complexity. The authors are indeed correct that there is a direct relationship between the sensitivity of a function to its inputs and whether or not inverse problems related to that function are well-conditioned[8]."

Response: We agree that definitions of complexity exist. We here have been motivated by the need to find a 'constructive' definition of 'hydrological' model complexity that can

implemented in complexity regularized model selection. We are glad that the referee sees the connection between stability (as measured by what we propose) and well posed model selection problems. However, we are unware of any constructive definition that can be used to measure hydrological model complexity (including those that the referee has cited). We ourselves are curious to know if the references provided by referee do so. Kindly note that the approach that we have presented is not Bayesian. The intent of this paper is to measure model complexity, perhaps primitively.

Comment: No criticism is offered either (a) of the theory that the paper purports to replace (complexity -> universal induction), nor (b) of the theory that it actually attempts to replace (sensitivity -> condition numbers)."

Response: Regarding the criticism that the referee wants us to offer, we do not claim any of what the referee alludes to, i.e. neither a) nor b).

Comment: Does anyone argue that parameter dimensionality is a measure of prediction uncertainty? I'm not sure how this argument would proceed – the idea is prima facie absurd. Of course the two can be related, and the former can even contribute to the latter (via non-monotonicity), but to assert that one "measures" the other is untrue. The authors claim that this is a common assertion in line 20 of page 3949,but no examples or references are given. I'm genuinely curious what this argument looks like.

Response: Our argument is simple. Often models with higher parameter dimensionality are considered more complex and associated with higher possibility to overfit. If prediction uncertainty (see also comments of referee 1) is attributed to both model deficiency and model complexity, the latter measures a portion of prediction uncertainty. For example, consider a set of competing models that contain the 'truth.' Then complexity will measure prediction uncertainty. However number of parameter will be a poor measure of prediction uncertainty since there is more to model complexity than just the number of parameters. The same holds in the case if prediction uncertainty is defined such that an inaccurate model can predict with certainty and an accurate model can

be uncertain. In this case, prediction uncertainty is measured by complexity. Relevant references can be provided.

Comment: “This paper shows that instability of hydrological system representation in response to different pieces of information and associated prediction uncertainty is a function of model complexity” No, it doesn’t show that. It shows that instability in a performance metric is bounded by a sum that includes what the authors call “a measure of complexity” (i.e., ||B||). At best, the argument is that so-called “complexity” contributes (functionally) to a bound on the *potential for* instability. Never is it argued that instability is “a function of” ||B||, E||B||,Or . It is trivial to find counter-examples – for example a model of a projectile built on Newton’s laws will have less variability in presumably any performance measure than, say, a linear regression, even though the latter may have very small variability in its output (i.e., small regression coefficients and small sigma).

Response: We agree with the referee that we provide an upper bound on instability in a performance metric. This bound is a function (sum) of instability in model simulations and another quantity. The latter measure of instability (or sensitivity) is a measure complexity for the following reason. For two models of similar model deficiencies, the model with lower instability in model simulations (our definition of model complexity) will have lower possibility of worse performance over future unseen data, i.e. robust model selection (Occam’s razor). What we are proposing is something standard in regularized model selection literature. The model selection problem should trade off deficiency with model complexity. So, if a projectile model built on Newton’s laws is less deficient than a linear regression to the extent that its overpowers the effect of model complexity (or instability in model simulations) on future (over unseen data) model performance then the former should be the model of choice. We therefore donot see the projectile example as a counter-example.

Comment: To extrapolate on this, I wonder why we should perform inference over a prior that favors the linear regression. The objective of inductive inference is to use available information to condition our current knowledge. So what is the a priori knowl-

edge that favors a model with smaller dynamic range? Solomonoff bases his a priori preference against complexity on Occham’s principle, and then proves that convergence is guaranteed[10]. Here the concept of “instability” is used as the justification, and the rational given for this a priori preference is that instability contributes to “uncertainty”. Convergence criteria are not offered, let alone any demonstration thereof.

Response: These are Bayesian arguments to what is a frequentist approach to model selection. There is no prior in our argument. Our arguments are geometric. The arguments that the referee is attempting to invoke using [10] requires full specification of the probability distribution from which the observations of input and output are being generated, i.e. the set of models and the description of what remains unknown together can be fully specified by a distribution. In absence of full specification, convergence is impossible. To further explore referee’s suggestion of [10], one would need a ‘universal probability distribution’ that has all possible probability distributions in its support. The unknown distribution from which the observations have been sampled should also be from a ‘computable probability distribution’. And it is only then that the problem of model selection ‘may’ (in the sense of converging to the unknown distribution from which observations have been generated) be solved if ‘infinite computational resources’ are available (additional conditions may be required since for convergence full specification is not sufficient). Pande (2013b) has provided similar arguments.

Comment: The problem with this argument is trivial – instability (actually equifinality, which is the symptom of instability in the context of induction) contributes to uncertainty about *choosing among an a priori set of possibilities*. Really, the goal should be to find the best representation of the system (not just the best among a particular class). Since this is generally impossible in practice, we must recognize that there is always uncertainty about the extent to which the best model from a particular class represents the system. So simply increasing our ability to pick a single choice from some arbitrary a priori set does not seem to be a worthwhile objective, and can easily lead to increased uncertainty.

Response: We disagree with the referee on equating instability with equifinality. Non-uniqueness, one of the 3 conditions that define ill-posed problems, is the same as equifinality. Please see our response to comment 1b of referee 1. We agree that there will always be model deficiency, even when chosen within a Bayesian approach that only has finite computational resources and cannot avail of the universal probability distribution, since one would 'atleast' need to have a fully specified model selection problem. Please see our response to the previous comment. We totally agree with the referee that "we must recognize that there is always uncertainty about the extent to which the best model from a particular class represents the system." But given that a particular class representation is a specification (though not complete) that we often come up as our set of hypotheses, we must make an effort to identify a subset of those with low probability of performing worse on future unseen data. This is what complexity regularized model selection aims to do.

Comment: What is an "unstable model representation"? Is a model represented, or does the model represent the system? If the latter (which is what I assume is meant), how can a model's representation of a system be "unstable"? As written, this sentence means that there is connection between the location of poles in the complex plane and something called "model complexity", which is not defined. This is obviously not what is intended.

Response: We meant the latter. Kindly see our response to comments 1.c, 1.d and 1.l.1 of referee 1. The proposed measure of model complexity measures its instability in representing the underlying system.

Comment: "complementary pieces of information to select a better constrained model" What is a "better constrained model"? I assume you mean that the multiple pieces of information are used to better constrain the inference procedure, however it is possible that you really mean that you will select a constrained probabilistic model. Can you elaborate?

Response: We mean "that the multiple pieces of information are used to better constrain the inference procedure."- to suggest such an inference procedure is similar to complexity constrained (regularized) model selection.

Comment: "But is the issue of ill-conditionness limited to the discourse of the number of measures used?" The answer to this question is well known. Obviously, the extent to which an inverse problem is well conditioned depends on several things including the form of the prior (or at least the sensitivity and injectivity of the function to be inverted, if we prefer to conceptualize the problem that way), and the form of the likelihood (or optimization objectives). More important is the fact that we don't have a very clear idea here what you mean by "conditionedness". In the classic sense of the word, the problem is formulated around some function that must be inverted. In the more general context of inductive inference, we may want to consider including (as is done here) inference in function space. It would be good to put the content of the second paragraph before the first, as it introduces the content necessary to understand the discussion of "multiple measures", etc. More to the point, you state that "a definition of ill-conditionedness [is] needed" but you don't actually offer one. Define first, discuss second.

Response: Many thanks! We agree that the conditions that the referee has mentioned are for whether or not an inverse problem is well posed (since injectivity is akin to the condition of non-uniqueness while sensitivity is akin to the condition of stability). Meanwhile conditioness is only linked to the sensitivity or stability condition. In both the cases, the problem of inference (finding an inverse) is in a function space. Since ill-posed inverse problems are ill-conditioned and that non-uniqueness in absence of instability does not contribute to instability in a performance metric, we choose instability (in model simulations or in system representation) to define model complexity. This is because this definition of model complexity then has all the effects that a model can introduce in prediction uncertainty (instability in a performance measure).

Comment: "what happens when an ill-conditioned model is selected to represent the

underlying hydrological system" How can a model be ill-conditioned? Previously you discussed that "reconciling models with observations" is ill conditioned, now we are assigning that property to the model itself.

Response: We understand the confusion. By ill conditioned model, we mean that the inverse of the model equation y = f(x;beta) for given parameters beta is ill-conditioned, where y is sampled from some distribution. Meanwhile for ill-conditioned model selection we mean that the inverse of model equation y = f(x;beta) in terms of beta where data (y,x) is sampled from some distribution.

Comment: "Since it fails to exploit interesting information in the data, there is uncertainty in system representation." This sentence appears to be gibberish. I'm not even going to guess. First, we don't know what these words refer to (see previous comments). Second, even if we assume that an "ill conditioned model" is a model that results in equifinal parameter distributions during calibration, then this model in no way "fails to exploit" interesting information in the data, it simply means that the information in the data does not yield a unimodal inference posterior. If we instead are talking about a model with variable performance metrics, then it is also not necessarily true that it "fails to exploit" any information in the data because we have no idea what information is in the data to begin with. Perhaps the input data simply is insufficient to predict the observed response.

Response: We will delete the sentence.

Comment: "Should not this uncertainty in assessing structure deficiency depend on the class of model structures which are used to assess deficiencies?" First, what is a "class of model structures"? Second, how is a class of model structures used to assess deficiencies? Third, deficiencies in what, exactly? We have not assessed any structure deficiency up to this point, we have only discussed whether an inverse problem is ill-conditioned.

Response: We will delete this sentence.

Comment: "The characteristics of uncertainty in system representation can then idetify the consequences of ill-conditioned model selection problem and hence define ill-conditioned model selection." So now we are back to the inference problem (over models) being ill-conditioned, whereas earlier in this paragraph it was the model. Second, you have made no argument that "characteristics of uncertainty" can identify anything. I'm sorry, but this whole paragraph (and the one preceding) is gibberish. There is no discernable content here; it is just a bunch of buzzwords strung together to make semi-coherent sentences.

Response: "we are back to the inference problem (over models) being ill-conditioned, whereas earlier in this paragraph it was the model" – because these two are connected. The uncertainty in system representation due to non-uniqueness in model selection is different in nature from uncertainty in system representation due to instability in model selection (or ill-conditioned model selection). We will rephrase the paragraph based on previous comments of the referee.

Comment: "We characterize uncertainty in hydrologic system representation as composed of non-uniqueness and instability in system representation." Then you would be wrong. Uncertainty includes many things, one of which is the fact that any inference posterior will have finite entropy after a finite number of experiments. This fact results both in what you call equifinality and also inconsistency. In particular, both of these things result if the posterior is generally multi-modal. There are, however, other sources of uncertainty. One of which is that we may not test an accurate model, so even if we have low-entropy unimodal posteriors we may still not know whether we have an accurate system representation. Even if we had an infinite number of experiments. Also, even if we did happen to test a "true" model, we could never know that we did due to Hume's problem. Anyway, this "characterization" is false.

Response: Kindly note that our approach does not deal with posteriors and multi-modal distribution. We appreciate how the referee has pointed out how equifinality (non-uniqueness) and inconsistency (instability) leads to multi-model distributions. However,

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 (2013a)).

Comment: "different measures of closeness, which when orthogonal, provide complementary pieces of information to select a better constrained model (Sivapalan et al., I wonder what orthogonal means here? Unless I missed something Sivapalan 2003 says nothing about orthogonal measures of closeness. Both of these references actually talk about using multiple types of data for inference, which is a particular kind of "more information", and is not the same thing as "multiple measures".

Response: Kindly see our response to comments 1 l.2, 1 l.3, 4 and 6 of referee 1.

Comment: "Instability can then be rephrased as a changing set of good system representations (models) when different data sets are used." What is a "good" system representation? Why are we talking about value judgments("good") when we could be talking about inference posteriors? How is this definition different than before? More importantly, why is this undesirable, and why will it help anything for us to force this not to happen?

Response: We will rephrase.

Comment: "This paper demonstrates that instability of a given model over realizations of data can be understood and controlled by what we term as model output space."No, it doesn't. First of all, you argue that instability can be bounded by a measure on model output space. Second, you argue that inference on models that demonstrate instability can be "controlled" using measures of model output space. But you don't actually show that this is true at all, you merely assert that it is so. The paper is a set of unjustified (and I imagine unjustifiable) assertions with minimal supporting evidence.

Response: The synthetic case study indeed 'demonstrates' the above said. Kindly also see Pande et al. (2009) and Arkesteijn and Pande (2013), where we have demonstrated the same on real world data sets and different types of models.

Comment: "We call the extent of model output space as the measure of complexity since its regularization would lead to a stabler representation of underlying processes" Sure, if a model has no variability in its response to inputs, the limiting magnitude of the difference in values of a performance metric during different applications is minimal (bounded only by ||D||, which does not depend on the model). Yes, a constant is a stable system representation; it is not, however, one that offers any insight into the system response. That is, a model that returns a constant value for all inputs is favored according to this philosophy because the bound on the "variability in the errors" is minimized (at the variability in the observations). I fail to see a priori why we should favor models that lack a dynamic response or any sensitivity to inputs.

Reponse: The proposed approach does not suggest selection of a model of lowest complexity. It proposes a tradeoff between model performance on one realization of data and model complexity. Kindly see our response to previous comments on the same. Also see Pande et al. (2009) and Arkesteijn and Pande (2013).

Comment: "The deviation in performance of a model over two different information sources is bounded by ||B|| that measures how large is the model output space." Strictly read, this sentence states ||A||-||C||< ||B||.

Response: We agree.

Comment: "Since ËĘ2 is nested within ËĘ1, if p_1ËĘ1 is not the same as p_1ËĘ2 then p_1ËĘ1 is closer to the observation o_1." This is true only in special cases: when the additional processes that differentiate ËĘ1 and ËĘ2 can be parameterized so that they have no effect. If this is not possible then the stated conclusion is not guaranteed. It is very easy to imagine a situation where this is not the case. For example, if we have a Nash cascade with N and N+1 reservoirs respectively then this will not hold. Similarly,

if we have a cascade of independently parameterized linear reservoirs treated with an explicit solver that takes the form (e.g., HyMod): x_(i,t)=(x_(i,t-1)-k_(i-1) x_(i-1,t))/(1+k_i), we must allow k_(N+1) to approach infinity for pËĘ1=pËĘ2 (i.e., there is no value of k_(N+1) for which this is true, and at best ËĘ2 can only approach the accuracy of the less complex ËĘ1 when ËĘ1 is the true model). The point is that the concept of nested structures is not very illuminative.

Response: The statement made by the authors holds. Since structure 2 (structure = a collection of models) is nested within structure 1, best model chosen from structure 2 can never be closer to the observed than the best model chosen from structure 1. This is because all possible models that can be chosen from structure 2 can also be chosen from structure 1. Nash cascade structure with N reservoirs is not nested within Nash model structure with N+1 reservoirs since there should be a subset of models from N+1 Nash structure that should form N reservoir Nash model structure (as a nested structure will require). It can never happen since all reservoirs in a Nash model have the same storage constants. The only model parameterization when the two model structures (and models) overlap is a trivial one with all storage constants are 0. Independently parameterized cascade of linear reservoirs models should form nested structures.

Comment: We now note that the model output of any hydrological model is continu-ous in its parameters."What? No. Of infinite-capacity linear bucket models, yes. Of anything with a threshold or a classification (e.g., vegetation, soil), no. It is simply false to claim that (y(alpha)) is continuous in alpha.

Response: Our assertion that hydrologic models are continuous in parameters is correct. It may not be differentiable, as the referee's example of thresholds suggests.

Comment: Moreover, if we are here using the y(alpha) definition of a model, then this is an absurd statement, since changing parameters changes the structure, and this change may be discrete (e.g., number of linear reservoirs, choice of infiltration equa-

tion, etc.). Also, the authors are correct that a model necessarily includes the distributions over parameters (i.e.., a determined parameter .. is actually a Dirac distribution at x= ..). The problem with this discussion is that it only considers uniform distributions. I am not sure that this actually affects anything, but the presentation is careless.

Response: We agree that perhaps the presentation is careless but the argument on continuity stands. Jumps from one infiltration equation to another will require two abstract parameters (in addition to the set of parameters corresponding to both the equations), one going from 0 to 1 and the other from 1 to 0 to represent a switch from one equation to another. These switches from 0 to 1 and vice versa can be smoothed by sigmoid type distribution functions (which are not uniform distributions)

Comment: "a definitive statement on structure complexity based on parameter ranges or parameter dimensionality, i.e. without knowing their complexity in advance, can only be made if the corresponding structures are nested." I can't tell what the point is. If we use the definition of complexity offered here, then I can make a definitive statement about complexity on any model using the proposed algorithm. Yes, if we have a (very) special case of nested structures, then we can – perhaps – say that one is more complex than the other without running the algorithm, but so what? I'm not sure what Gupta et al (2008) says about a "continuum of model structures". There is no Renard et al (2010) in the reference list.

Response: The point is fundamental to any structural risk minimization problem (see for e.g. Vapnik, 1982) and any hydrological model selection problem that seeks to increase complexity in a step wise manner (step wise manner essentially builds a nested set of model structures). Kindly note again that what we are proposing is not that model complexity is used in isolation, but that it is used in a tradeoff with model performance on a given realization of data. It justifies why one can ignore the issue of complexity and only focus on model performance on one given realization of data (since we know that model complexity surely increases as one builds one complicates the model structure in a step wise manner). It justifies increase model complexity in a step wise manner as

long as model performance does not degrade. Once it starts to become difficult to find a better model (or one starts to find a worse model) based on a performance metric with increasing model complexity, it means additional complexity is not warranted. This is also the basis for top-down modeling approach or the downward approach (Sivapalan et al., 2003).

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. Pande, S. (2013a). Quantile hydrologic model selection and model structure deficiency assessment 1: Theory, Water Resources Research, 49, 5631–5657, doi:10.1002/wrcr.20411. Pande, S. (2013b). Quantile hydrologic model selection and model structure deficiency assessment 2: Applications, Water Resources Research, 49, 5658–5673, doi:10.1002/wrcr.20422. Vapnik, V. (1982). Estimation of Dependencies Based on Empirical Data, Springer Verlag, New York. Sivapalan, M., Blöschl, G., Zhang, L., and Vertessy, R.: Downward approach to hydrological prediction, Hydrol. Process., 17, 2101–2111, doi: 10.1002/hyp.1425.

.

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