These responses largely serve to substantiate my concerns. I will try to explain the problems succinctly again, and then respond directly to a few of the authors' comments. But ultimately, the central thesis in this paper is not only demonstrably incorrect, it is antithetical to progress in the field.

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. It does *not* prefer hypotheses that make parsimonious predictions, it prefers parsimonious hypotheses. Solomonoff used the principle correctly in the context of model inference, whereas the current authors use it incorrectly. 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 ^1 over model ^2 because ^1 has fewer ontological commitments (fewer processes). Instead the authors argue that we should prefer ^1 (over ^2) because of some characteristic of their predictions (namely, that predictions from ^1 are are less dynamic in the sense that they respond less to inputs).

   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.

2. 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).

   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. 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. 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.

   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. By artificially constraining the inference posterior, we articfically constrain our ability to *measure* uncertainty – we do not actually constrain uncertainty. In fact if the inference posterior is artificially constrained, then we actually *increase* that portion of uncertainty that is impossible to measure.

   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.

Replies to Replies:

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 re-sponse 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 in-equality does not mean that prediction uncertainty is tightly linked to model complexity.

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). 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?

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".

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.

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

Our approach is not Bayesian.

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

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.

Also, we never claim that we would like to test a 'true' model.

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

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.

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. 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.

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 to artificially 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.

However, we agree with the referee that any model selection should deal both with the issue of non-uniqueness and instability.

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).

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.

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. 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.

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.

What we are suggesting is that such a regularization will reduce uncertainty in system representation due to increased stability.

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 multi-modal) 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).

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.

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.

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. Moreover, all are associated with some convergence description (a convergence proof). All regularization schemes are Bayesian priors, and must therefore be based on some principle that is justifiable *in an a priori sense*. 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).

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.

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?

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.

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.

It will thus tradeoff epistemic uncertainty with aleatory uncertainty.

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).

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.

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

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 be uncertain.

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?

In this case, prediction uncertainty is measured by complexity.

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

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**.

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).

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. 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.

What we are proposing is something standard in regularized model selection literature. The model selection problem should trade off deficiency with model complexity.

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.

These are Bayesian arguments to what is a frequentist approach to model selection. There is no prior in our argument.

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.

Our arguments are geometric.

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

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.

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. 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.

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.

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

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.

---

1. D. Wallace, *The emergent multiverse: Quantum theory according to the Everett interpretation*. (Oxford University Press, 2012).
2. K. S. Van Horn, Constructing a logic of plausible inference: a guide to cox's theorem. *International Journal of Approximate Reasoning* **34**, 3-24 (2003).
3. E. T. Jaynes, *Probability Theory: The Logic of Science*. G. L. Bretthorst, Ed., (Cambridge University Press, New York, NY, 2003).
4. B. Russell, *The problems of philosophy*. (Oxford University Press, 2001).

5. B. Kosko, Fuzziness vs. probability. *International Journal of General System* **17**, 211-240 (1990).
6. K. J. Beven, *Facets of Uncertainty: Epistemic error, non-stationarity, likelihood, hypothesis testing, and communication*. *Hydrological Sciences Journal*, (2015).
7. S. Rathmanner, M. Hutter, A philosophical treatise of universal induction. *Entropy* **13**, 1076-1136 (2011).
8. M. Hutter, *Universal Artificial Intelligence - Sequential Decisions Based on Algorithmic Probability*. (Springer, New York, 2005).
9. S. Weijs, N. van de Giesen, M. Parlange, HydroZIP: How Hydrological Knowledge can Be Used to Improve Compression of Hydrological Data. *Entropy* **15**, 1289-1310 (2013).
10. R. J. Solomonoff, A formal theory of inductive inference. Part I. *Information and Control* **7**, 1-22 (1964).
11. G. S. Nearing, H. V. Gupta, The quantity and quality of information in hydrologic models. *Water Resources Research* **51**, 524-538 (2015).
12. S. V. Weijs, G. Schoups, N. Giesen, Why hydrological predictions should be evaluated using information theory. *Hydrology and Earth System Sciences* **14**, 2545-2558 (2010).