THE UNIVERSITY OF ALABAMA® | University Libraries

Ensembles vs. information theory: supporting science under uncertainty

Grey S. Nearing – University of Arizona
Hoshin V. Gupta – University of Arizona

Deposited 10/11/2018

Citation of published version:

Nearing, G., Gupta, H. (2018): Ensembles vs. information theory: supporting science under uncertainty. *Frontier Earth Science*. DOI: 10.1007/s11707-018-0709-9

This work is distributed under the Creative Commons Attribution 4.0 License.

Published by Higher Education Press and Springer-Verlag GmbH Germany, part of Springer Nature

© Author(s) 2018

Front. Earth Sci.
https://doi.org/10.1007/s11707-018-0709-9

REVIEW

Ensembles vs. information theory: supporting science under uncertainty

Grey S. NEARING[1], Hoshin V. GUPTA (✉)[2]

1 University of Alabama, Department of Geological Sciences, Tuscaloosa, AL 35294-1152, USA
2 University of Arizona, Department of Hydrology and Atmospheric Sciences, Tucson, AZ 85721-0158, USA

© Higher Education Press and Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract Multi-model ensembles are one of the most common ways to deal with epistemic uncertainty in hydrology. This is a problem because there is no known way to sample models such that the resulting ensemble admits a measure that has any systematic (i.e., asymptotic, bounded, or consistent) relationship with uncertainty. Multi-model ensembles are effectively sensitivity analyses and cannot – even partially – quantify uncertainty. One consequence of this is that multi-model approaches cannot support a consistent scientific method – in particular, multi-model approaches yield unbounded errors in inference. In contrast, information theory supports a coherent hypothesis test that is robust to (i.e., bounded under) arbitrary epistemic uncertainty. This paper may be understood as advocating a procedure for hypothesis testing that does not require quantifying uncertainty, but is coherent and reliable (i.e., bounded) in the presence of arbitrary (unknown and unknowable) uncertainty. We conclude by offering some suggestions about how this proposed philosophy of science suggests new ways to conceptualize and construct simulation models of complex, dynamical systems.

Keywords information theory, multi-model ensembles, Bayesian methods, uncertainty quantification, hypothesis testing

1 Introduction

A multi-model ensemble does not actually provide us with a reliable way to understand uncertainty, but instead represents a kind of sensitivity analysis, which is not inherently related to model uncertainty. There exists no consistent (i.e., asymptotic or bounded) theoretical relationship between sensitivity and uncertainty, and this distinction has important implications for scientific inference. In particular, if the scientific goal is to build the best hypothesis-driven model possible given (imperfect) experimental data, then any form of hypothesis testing fundamentally requires separating model error from data error.

Further, inferences over a multi-model ensemble only converge if the truth can be sufficiently well approximated as a mixture of the models in the ensemble. In general, however, we cannot know whether a true model is a member or mixture of the models in a particular ensemble. The consequence is that inferences over multi-model ensembles admit unbounded errors under conditions that are not evaluable. Ensemble-based methods are therefore fundamentally unreliable for either uncertainty quantification or hypothesis testing. On the other hand, the argument that we present here is that information theory makes it possible to derive a consistent (i.e., bounded) method for partitioning of data error from model error, and this leads to a reliable hypothesis test.

Throughout this discussion, we will distinguish between two different things:

- Uncertainty decomposition separates the contribution to total uncertainty from different sources of error in a modeling chain (e.g., model inputs, model parameters, model structure, measurement error, etc.).
- Uncertainty propagation provides estimates of uncertainty in forecasts or predictions.

The former is required for hypothesis testing, and the latter for decision-support. Our proposal is that information theory solves the uncertainty decomposition problem, and therefore leads to a reliable hypothesis test. However, information theory does not apparently provide a way to reliably propagate uncertainty into future predictions. To address the latter, we suggest that instead of constructing hydrology models as solutions to sets of partial differential equations representing various conservation laws, we might construct hydrology models as Bayesian networks constrained by conservation symmetries.

Received April 6, 2017; accepted March 4, 2018

E-mail: hoshin@email.arizona.edu

2 Uncertainty

It is impossible to define uncertainty in a meaningful way without reference to an underlying system of logic. This is because uncertainty is an epistemological concept in the sense that any meaningful understanding of uncertainty requires a theory of knowledge. Uncertainty resulting from scientific endeavors can only be understood in the context of a well-defined philosophy of science. For example, under a philosophy of science based in probability theory or probabilistic logic, uncertainty might be understood as the difference between the (unknown) truth-value of some proposition and our state of belief about that truth-value. Uncertainty, in this case, is due wholly to the fact that we must rely on non-deductive reasoning in situations where information is limited.

There have been many attempts to define uncertainty in the Hydrological sciences (e.g., Montanari, 2007; Renard et al., 2010; Beven, 2016). None of these attempts – that we are aware of – achieves logical self-consistency. However even if this minimum standard were met, it is widely understood that no such attempt in any branch of science has yielded a consistent estimator of a measurable quantity (Nearing et al., 2016b). This is a problem because, at least since Popper (1959), it has been understood that hypothesis testing must account for uncertainty. This paper outlines a procedure for hypothesis testing that does not require quantifying uncertainty, but is coherent and reliable (i.e., bounded) in the presence of arbitrary (unknown and unknowable) uncertainty.

In the context of any particular science experiment, there are two fundamental sources of uncertainty: (i) the fact that our hypotheses may not be perfect descriptions of dynamic systems, and (ii) any imprecision and incompleteness inherent in experimental data. The latter is simply the fact that any set of experimental control (input) data will never fully determine any set of experimental response (output) data, even given a hypothetically "perfect" model. While this indeterminacy may be partially due to real, ontic randomness that manifests at scale (Albrecht and Phillips, 2014), it is, more generally, just a matter of limited information in experimental inputs[1)].

We will use the terms *epistemic* and *aleatory* to refer to, respectively, hypothetical (imperfect system description) and experimental (data imprecision and incompleteness) components of uncertainty. The intuition we encourage is that epistemic uncertainty is uncertainty about scientific hypotheses, and aleatory uncertainty is related to incomplete information in experimental input/output data. This differs somewhat from distinguishing between ontic (as aleatory) vs. epistemic randomness, which is not a meaningful distinction given our current inability to solve the Schrödinger equation at watershed scales. Our perspective also differs from distinguishing between aleatory vs. epistemic uncertainty such that the former is variability that can be treated probabilistically while the latter cannot (e.g., Beven, 2016); this is not a valid distinction because probability theory is fundamentally an epistemic (i.e., logical) calculus (Van Horn, 2003).

Given that the job of a scientist is – at least ostensibly – to build models that are as close to perfect as possible, separating these two sources of uncertainty (epistemic vs. aleatory) is fundamental to the project of hypothesis testing, and therefore of science in general. As an example, take the simplest experiment we might imagine (illustrated in Fig. 1), consisting of a model that takes one control variable u to determine one response variable y. The variable z_u represents data about the system control while z_y represents data about system response, and these data may not correspond exactly to the phenomenological[2)] variables u and y. There is, in principle, some "perfect" (but unknown) probability distribution $\mathscr{S}(z_y|z_u)$ that represents the distribution over actual experimental data that is implied by the real dynamic system, including all measurement devices. This distribution represents aleatory uncertainty, as defined above, in the context of the particular experiment, and completely describes everything that it is possible to know about the relationship between measured components of the system inputs and response given available data (one could think of $\mathscr{S}$ as a perfect model).

Our objective is to represent system $\mathscr{S}$ with a hypothesis-driven, usually physically-based, model $\mathscr{S}(z_y|z_u)$ in a way that is, in some sense, explanatory (i.e., reductionist). It is important to remember that because the relationships between the phenomenological variables (u and y) and their corresponding data (z_u and z_y) are determined by measurement devices, which are just physical systems like any other, it is necessary to model measurement devices as part of the whole dynamic system that we are testing. This is a consequence of confirmation holism (Stanford, 2016) – it is impossible to separate a measurement model from the model dynamical system that we want to learn about during any hypothesis test. This means that all hypothetical models $\mathscr{h}$ must be probabilistic to account for aleatory uncertainty, a fact that is well-known and widely discussed in the hydrology literature (e.g., Weijs et al., 2010; Beven, 2016).

The model itself, however, is subject to epistemic uncertainty in all of its component process representations. This epistemic uncertainty about individual process representations, including all measurement devices, man-

1) Even if ontic randomness does exist in macro-scale systems like watersheds, this still manifests as limited information in experimental inputs.
2) Something is said to be phenomenological if it is observable in principle. Theoretical laws are not phenomenological because only their effects on existent objects can be observed, however properties of existent objects – things like porosity, bulk density, precipitation, and streamflow – are phenomenological, even if it is impossible to measure these things exactly in practice.

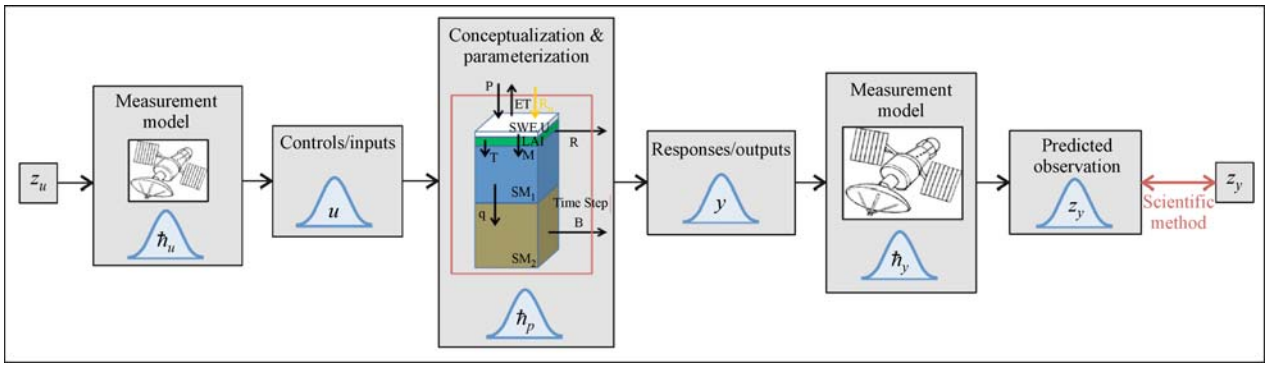

Fig. 1 Diagram of a simple experiment with a single input and single output. Even in the simplest case, we require at least one process hypothesis $\hbar_p$, and at least two measurement models $\hbar_u$ and $\hbar_y$. Aleatory uncertainty is defined as the (unknown) distribution $\mathscr{S}(z_y|z_u)$, which would only be knowable if we had access to both a perfect process model and also perfect measurement models. A full model is the conjunction $\hbar = \{\hbar_p, \hbar_u, \hbar_y\}$.

ifests as uncertainty in all phenomenological components simulated by the model. This means that not only must each individual hypothetical model $\mathscr{S}(z_y|z_u)$ be probabilistic to account for aleatory uncertainty, but also that we must have probability distributions over families of these probabilistic models.

The simple example in Fig. 1 is apparently completely generalized by making z_u and z_y vectors instead of scalars, so that the concepts and philosophy outlined above apply to more detailed modeling problems in the environmental sciences. For example, the list of phenomenological variables might include any types of parameters, boundary conditions, prognostic states, or diagnostic outputs, and we may have experimental data about any of these things.

3 Ensembles can't measure epistemic uncertainty

The most common strategy for treating epistemic uncertainty about the structure and functioning of a dynamical system is to place a probability distribution over some family $\mathcal{H}$ of hypothetical model components: e.g., $\mathcal{H} = \cup_i \hbar^i$ where the superscript i indexes a family of competing or complimentary models. We can estimate a predictive distribution over a finite family of models by Monte Carlo (Metropolis, 1987) integration:

$$\mathcal{M}(z_y|z_u) = \Sigma_{\mathcal{H}} p(z_y|z_u,\hbar^i)p(\hbar^i). \tag{1}$$

Predictions from more general model classes must be integrated differently, of course, but the principle is the same. The notation $p(z_y|z_u,\hbar)$ is equivalent to $\hbar(z_y|z_u)$. Technically, $\mathcal{M}$ is itself a model; however, calling it that might cause confusion with objects like $\hbar$, so we will call $\hbar$ a model, $\mathcal{H}$ an ensemble, and $\mathcal{M}$ the ensemble prediction. $\mathcal{H}$ is exactly the typical multi-model ensemble – we have, thus far, not deviated in any way from standard practice. Sometimes multi-model ensembles are generated simply by choosing different parameter values (e.g., Beven and Freer, 2001) and sometimes by choosing different biogeophysical process representations (e.g., Clark et al., 2015).

We might wish to be more explicitly reductionist in our development of Eq. (1). For example, we might wish to explicitly recognize different components of each model – like different biogeophysical process representations ($\hbar^i_{pj}$ is the ith hypothesis about a jth biogeophysical process), different measurement models (e.g., $\hbar_u$, $\hbar_\theta$, and $\hbar_y$ are models of the measurements of boundary conditions, parameters, and diagnostic outputs respectively), etc. Accordingly, the appropriate expansion of Eq. (1) is over phenomenological variables using a chain rule like:

$$\begin{aligned}&\mathcal{M}(z_y|z_u,z_\theta)\\ &=\Sigma_{\mathcal{H}}\int_{u,\theta,y}\Big(\hbar^{i_y}_y(z_y|y)p(y|\hbar^{i_1}_{p_1},\ldots,\hbar^{i_n}_{p_n},u,\theta)\\ &\quad\hbar^{i_u}_u(u|z_u)\hbar^{i_y}_\theta(\theta|z_\theta)\mu_{u,\theta,y}\Big)\\ &\quad p(\hbar^{i_y}_y,\hbar^{i_1}_{p_1},\ldots,\hbar^{i_n}_{p_n},\hbar^{i_u}_u,\hbar^{i_\theta}_\theta).\end{aligned} \tag{2}$$

The aspects of an ensemble that are actually testable, at least in principle, are its predictions – i.e., the probability distribution $\mathcal{M}$ over experimental data that results from integrations like Eq. (1) or Eq. (2). Notice that these predictions are determined completely by the information that we put into the ensemble – information that comes in the form of hypotheses like the various $\hbar^i_{pj}$s. So, $\mathcal{M}$ does not yet represent anything like real uncertainty, because at this point our ensemble may contain models that conflict with reality, and it certainly does not contain all possible models that do not conflict with what we currently know about reality. Specifically, we know with effective certainty that none of the $\hbar^i$'s are actually correct, and without collecting and conditioning $\mathcal{M}$ on experimental data, we

cannot know the (inexact) relationship between families like $\mathcal{H}$ and reality. We cannot even reliably guess, without empirical testing, whether $\mathcal{H}$ is substantively biased, let alone anything about its predictive accuracy or precision.

Related to prediction and forecasting, it may seem intuitive that if our ensemble $\mathcal{H}$ contains a set of models $\hbar^{i}$ that are plausible representations of system $\mathscr{s}$, then the resulting $\mathcal{M}$ will represent a set of plausible outcomes for data . While this may be true, the problem is that this range of plausible outcomes is not a plausible range of outcomes, and any quantification over $\mathcal{M}$ may be significantly biased related to either true frequencies or true likelihoods of real outcomes. It is, at least in principle, potentially dangerous to confuse an ensemble of plausible models with a plausible representation of uncertainty (Taleb, 2010).

Ensembles like $\mathcal{H}$ also do not help with reliable scientific inference. The problem here is that unless we are willing to believe that at least one of the competing models $\hbar^{i}$ from our family of models $\mathcal{H}$ is actually "true" in a strict sense, then any scientific inference that we make over $\mathcal{H}$ is at least potentially inconsistent. This problem is well-known and well-understood, and is discussed in accessible detail by Gelman and Shalizi (2013). Specifically, the problem is that if the true model $\mathscr{s}$ is not a mixture of the $\hbar^{i}$ models in $\mathcal{H}$, then any application of Bayes' theorem to $\mathcal{H}$ is not guaranteed to yield posteriors that cluster in any neighborhood around the best models in terms of future predictions. Importantly, inference over $\mathcal{H}$ with $\mathcal{M}$ can even result in posteriors that end up predicting with skill worse than the worst model in the class, and sometimes with skill worse than chance (Grünwald and Langford, 2007).

In hydrology, the primary danger is that incorrect phenomenological distributions (or measurement distributions) like $\hbar_{y}$ can result not only in incorrect, but actually in contradictory, inferences over ensembles of process representations like $\hbar_{p}$. Examples of this effect were given by Beven et al. (2008), and discussed by Clark et al. (2011): "*A major unresolved challenge for the ensemble method to work is to ensure that the ensemble includes at least one hypothesis that approximates "reality" within the range of data uncertainty.*" The problem is that we can never know whether such a model is available without first being able to separate aleatory from epistemic uncertainties. However, under Eq. (1), any inferences over any particular component of the modeling chain (e.g., those measurement distributions that together represent aleatory uncertainty) are at least potentially inconsistent under misspecification of the distribution over any other component of the modeling chain (e.g., those biogeophysical processes hypotheses, $\hbar_{p}$, that we actually want to test). Thus we are left with a circular problem in that we cannot reliably test process models unless we have accurate data models, and we cannot test data models unless we have accurate process models. This is exactly a consequence of the aforementioned confirmation holism problem. To reiterate, hypothesis testing necessarily relies on separating epistemic from aleatory uncertainty.

It is worth pointing out that certain practical implications of this problem have been discussed widely in the Hydrology literature. Any application of Bayesian methods for model discrimination requires that a likelihood function be specified a priori, so as to relate model predictions with actual observations. At least in the case of environmental models, this choice is necessarily ad hoc (Beven et al., 2008; Beven, 2016), and Nearing et al. (2016b) point out that this problem of assigning likelihood functions is exactly the problem of separating aleatory from epistemic uncertainty.

So, what we need is a way to reliably (i) separate aleatory from epistemic uncertainty, and (ii) project our epistemic uncertainty onto forecasts and predictions. Information theory provides a solution to the first problem, and we propose that it might offer a solution to the second problem as well.

4 Information theory can be used to decompose uncertainty

Before we continue, let's remember that there are currently two dominant and fundamentally different approaches to scientific inference: Bayesian methods (i.e., ensemble-based methods) and falsification-based methods (i.e., statistical hypothesis testing). Gelman and Shalizi (2013) provide an excellent and accessible treatment of this distinction. For our purpose, the issue is that falsification-based hypothesis testing avoids assigning probabilities to models, and thus avoids the inconsistency problems discussed above, but still requires that the models assign probabilities to predictions, and thus requires specification of a prior.

Gong et al. (2013) laid the groundwork for a falsification-based information-theoretic approach that decomposes aleatory from epistemic uncertainty related to any individual model. Since this approach doesn't rely on an ensemble there is no chance for misspecification, degeneracy, or the resulting inconsistency. More generally, the approach does not rely on assigning any a priori distributions at all – it relies completely on empirical probability distributions, and therefore avoids the problem of requiring a priori distinction between any aspects of the structure of aleatory and/or epistemic uncertainty. To state this another way, by avoiding any direct use of Bayes' theorem it is not necessary to specify either a prior or a likelihood function, both of which cause errors in inference when misspecified.

This approach is as follows. First, we measure the information contained in experimental control data z_u about experimental response data z_y; call this $I(z_y;z_u)$. This measure is equal to 1 if the experimental control data

completely explains the observed variability in experimental response data, and is equal to 0 if the experimental response data is independent of the experimental control data. A model $\hbar$ then makes predictions $\hat{z_y}$ – i.e., $\hat{z_y} \sim \hbar\,(y|z_u)$ – that also contain some fractional information about experimental response data – call this $I\left(z_y;\hat{z_y}\right)$. As long as the measure I used for quantifying information is self-equitable, and the probability distribution over $\hat{z_y}$ is independent of z_y conditional on z_u (i.e., the model $\hbar$ acts only on experimental control data and not on experimental response data), then the data processing inequality (Kinney and Atwal, 2014) guarantees that the information content of the model predictions $\hat{z_y}$ is always bounded above by the actual information content of the data: $I(z_y;z_u) \geqslant I\left(z_y;\hat{z_y}\right)$. This is always true because $\hat{z_y} \sim \hbar\,(y|z_u)$ necessarily results in the Markov relationship: $z_y \rightarrow z_u \rightarrow \hat{z_y}$.

This inequality allows us to ask whether a given model $\hbar$ provides as much information about the relationship between experimental control data and experimental response data as is contained in the data themselves. If the answer to this question is "no" – i.e., $I(z_y;z_u) - I\left(z_y;\hat{z_y}\right) > 0$ – then we know that model $\hbar$ could be improved given only the information that is currently available from experimental data that we actually have in-hand.

More specifically, if measure I is the ratio of Shannon's (1948) mutual information to entropy then we can quantitatively separate epistemic uncertainty from aleatory uncertainty in the context of available experimental data. Details were given by Gong et al. (2013), but the basic idea is that if mutual information is notated as i:

$$i(z_y;z_u) = \iint p(z_y = \varsigma, z_u = v) \ln\left(\frac{p(z_y = \varsigma | z_u = v)}{p(z_y = \varsigma)}\right) \mathrm{d}\varsigma \mathrm{d}v, \tag{3}$$

and entropy as $\hbar$:

$$h(z_y) = \int p(z_y = \varsigma) \ln\left(p(z_y = \varsigma)^{-1}\right) \mathrm{d}\varsigma, \tag{4}$$

then $0 \leqslant I = \dfrac{i}{h} \leqslant 1$ is the fraction of observed entropy in z_y that is explained by observed variability in z_u. The distribution $p(z_y,z_u)$ is estimated empirically. In other words, I quantifies the fraction of variability in experiential response data that is explainable due to variability in experimental input data. A similar procedure is used to calculate $I\left(z_y;\hat{z_y}\right)$.

Ideally $p(z_y,z_u)$ would be an accurate representation of $\mathscr{S}$, but of course this is impossible in practice. The challenge is coming up with an empirical estimator of $I\left(z_y;z_u\right)$ that is as close to the true value as possible, and although we can't estimate $I\left(z_y;z_u\right)$ directly, we can bound it conservatively. Take any mapping $r : z_u \rightarrow z_y$, then by another application of the data processing inequality we have $I(z_y;z_u) \geqslant I\left(z_y;r(z_u)\right)$ such that information missing from model $\hbar$ (notated ε) is bounded by estimator $\hat{\varepsilon}$ as follows:

$$\varepsilon = I(z_y;z_u) - I\left(z_y;\hat{z_y}\right), \tag{5}$$

$$\hat{\varepsilon} = I\left(z_y;r(z_u)\right) - I\left(z_y;\hat{z_y}\right), \tag{6}$$

$$\varepsilon \geqslant \hat{\varepsilon}. \tag{7}$$

If r has known convergence properties over large function classes (e.g., Hornik, 1991), then in principle and as long as metric I itself admits a convergent estimator, we have a bounded and convergent estimate of the information content of experimental data. Regressions like $r(.)$ could be any convergent basis expansion; examples might include neural networks, Gaussian processes (Rasmussen and Williams, 2006), polynomial chaos expansions, and many others. The only strict requirement is that $r(.)$ must be data-driven. It's also essential that the z_u input data to $r(z_u)$ must be the same experimental control data that is used as input into model $\hbar$.

The fact that this approach always underestimates epistemic uncertainty (i.e., Eq. (7)) is essential. If our purpose were uncertainty propagation, then we might prefer over-estimating uncertainty; however, when the purpose is uncertainty decomposition for hypothesis testing, it is essential that epistemic uncertainty is under-estimated. The reason is because of the asymmetry between falsification and verification due to the modus tollens. In this case the proposition we are evaluating for our hypothesis test is $\mathcal{P} = \{\text{there exists no information in experimental data that is not captured by the model}\} \leftrightarrow \{\varepsilon = 0\}$. If there existed a finite probability that $\hat{\varepsilon} \geqslant \varepsilon$, then we should have a finite probability of incorrectly rejecting $\mathcal{P}$.

To summarize, information theory provides a reliable way of bounding epistemic uncertainty in a completely context-free environment, meaning that we are not simply comparing the performance of two competing hypotheses about how the system behaves. We bound the information contained in experimental data using a purely empirical method, so that we minimize conflation of epistemic with aleatory uncertainties, and any conflation that does occur is reliably in one direction – we will always under-estimate epistemic uncertainty. This approach facilitates a bounded hypothesis test whereby we know that any model $\hbar$ for which $I(z_y;z_u) \geqslant I\left(z_y;\hat{z_y}\right)$ has the potential to be improved without collecting any new data. Further, the test is coherent with probability theory, works in the presence of data uncertainty, and does not require subjective choice of a likelihood function. For demonstrations of this theory,

the reader is directed to Nearing and Gupta (2015) which applied this theory and method to conceptual runoff models, and Nearing et al. (2016a) which applied the theory to four distributed land surface models.

5 Information theory for uncertainty propagation

The underlying problem with ensemble-based methods is that we must explicitly formulate each of the hypothetical models in $\mathcal{H} = \cup \hbar^{i}$. Of course, it is impossible to formulate a complete family of plausible models (estimating a range of plausible outcomes is easy, but estimating a plausible range of outcomes is impossible). Instead, we suggest that a better approach would be to develop models that use hypotheses about individual process relationships to constrain model behavior.

It seems very strange to formulate a hypothetical model that we know is wrong, and to then assume that the error in predictions made using that wrong model will remain stationary up to some probability distribution. This goes for models like $\hbar$ and also for models like $\mathcal{M}$. Similarly, it seems strange to require that models admit solution. Currently, if we want to run any existing watershed model, we have to provide values for several unknown (and sometimes unknowable) parameters and boundary conditions. We then have to go back and prescribe uncertainty related to all of these components. If we prescribe that uncertainty using inverse methods, then we are fundamentally using a multi-model approach. If we assign that uncertainty a priori, then we are relying on a priori information.

So, consider instead that we would like to be able to run a hydrology model by starting with no prior information at all regarding the nature of the relationships among relevant variables (except, perhaps, for the fact that the system we are simulating is conservative), and to then impose constraints on its behavior based on available information or assumptions we may wish to make. In such a model, the mathematics would fundamentally act in such a way as to constrain probability distributions, instead of acting to solve sets of differential equations and then imposing uncertainties on top of those solutions. Following this perspective, we might approach hydrologic modeling from a "maximum entropy" perspective in which we first impose a uniform or non-informative distribution over possible outcomes (e.g., a uniform distribution over the real number line representing streamflow), after which the model equations act in such a way that they use hypotheses about process-level system behavior to constrain this distribution.

One approach might be as follows. A model is constructed as a Bayesian network, where each node in the network represents a different simulated variable at a particular location in time and space. A complete simulation of a complex and spatially distributed biogeophysical system is therefore represented by a joint probability distribution over a large Bayesian network. Conservation equations would then be imposed on this type of network as symmetry constraints in either time or space. Any other hypotheses about system behavior (e.g., that soil moisture exerts a threshold control on stomatal resistance) may be implemented by imposing constraints on the full network such that outcomes conflicting with imposed constraints are assigned zero probability.

We do not, at this moment in time, provide an example of such an implementation, in part because we are suggesting a completely different way of thinking about modeling hydrologic systems, and because there are many aspects of how this approach can be implemented that remain to be worked out. At this time, we are effectively suggesting that the community reconsider the first principles of complex models. Instead of solving sets of differential equations that do not admit solution unless they are fully parameterized – often requiring parameter values that are unobtainable in practice – we propose to instead treat conservation laws as symmetry constraints on maximum entropy distributions. This would allow for real quantitative understanding about the implications of our scientific hypotheses on future events, without necessitating any assumptions about statistical stationarity of model error distributions. It would also allow simulations under any available information – in principle if system properties (either relationships or values of physical parameters) are known, then these could be used to impose constraints on model behavior, but if they are not known then the model would not require them to complete a simulation.

6 Summary

It has become popular to employ various types of model ensembles to characterize and quantify uncertainty in model predictions, with multi-model ensembles considered as a way of assessing epistemic uncertainty due to imperfect system representation. However, there exists no consistency theorem, boundedness theorem, or convergence theorem that relates any measure over any feasible model ensemble or ensemble prediction with anything that could plausibly be called "uncertainty".

Because we cannot, in general, ever know whether truth can be sufficiently well approximated as a mixture of the models in a given ensemble, a) Bayesian assessment will almost unavoidably suffer from inconsistency and result in incorrect (even contradictory) inferences, and b) there is no guarantee that the Bayesian posterior will (in any systematic way) be related to anything like real uncertainty.

The solution to assessing epistemic and aleatory components of uncertainty within any modeling framework is, instead, to adopt an information theoretic approach. This makes it possible to quantify information provided by a system hypothesis in a reliable (bounded) way. This approach does not suffer from the typical problems related to assigning likelihood functions, Bayesian degeneracy and inconsistency, nor from the problem of ad hoc statistical rejection criteria.

Of course, we recognize that conceptualizing a model as "uncertainty restricted via constraints" as opposed to the conventional "prediction plus uncertainty" might seem like a difficult shift to make, and the difference in these two approaches does seem to be very fundamental. Whereas the latter (conventional) approach requires two sets of assumptions (model assumptions + error structure assumptions), the former only requires one set of assumptions (model assumptions). In the conventional approach, the mathematics of the model acts to make a prediction, to which we add uncertainty. The point is that we really want the mathematics of the model to act in such a way as to rule out areas of prediction space that are not feasible given our modeling hypotheses, theories, assumptions, and available data. To do this, we need a way of imposing conservation symmetries into mathematical descriptions of complex systems.

More broadly, in a complementary paper (Nearing et al., 2016b) we discuss the fact that uncertainty quantification is itself an incoherent activity and that it would be much easier to address science, engineering, policy, and communication problems from an information-centric perspective. Briefly, the idea is that we should communicate and act on the best available scientific information, and that this information may come in the form of probability distributions. So, while this current paper treats uncertainty as a central object in the scientific endeavor, we do believe that this is not an optimal perspective. Instead, our goal with this current paper is to help bridge the gap between current practice and the more fundamental change in perspective advocated by Nearing et al. (2016b).

The more practical point is that we really can't do model-based experiments in a reliable way without using information theory. What we discuss here is one aspect of that argument – that the standard methods fail to deal reliably with the underlying problem, which is separation of epistemic from aleatory uncertainties. Instead, multi-model assessment – as provided by intercomparison projects (e.g., Eyring et al., 2016) and multi-hypothesis modeling systems (e.g., Clark et al., 2015) – have their value in providing guidance on which process-level theories and equations are better suited as foundations for avenues of exploration about how to better represent different types of systems.

Finally, it is important to understand that there currently does not exist any reliable way to propagate epistemic uncertainty into future predictions. To this end, we encourage the community to focus on methods that use hypotheses and data to constrain the range of possible future outcomes, rather than to propose (necessarily degenerate) families of explicit modeling alternatives. Contrary to Beven (2006), mass and energy balance models expressed as differential equations might not be the "Holy Grail" of hydrologic modeling.

Acknowledgements We thank Professor Steven Fassnacht for his invitation and encouragement to submit this manuscript to the special issue on "Uncertainty in Water Resources", and are grateful to Anneli Guthke, Uwe Ehret, and one other anonymous reviewer for their helpful and constructive comments that helped to clarify points raised herein.

References

Albrecht A, Phillips D (2014). Origin of probabilities and their application to the multiverse. Phys Rev D Part Fields Gravit Cosmol, 90(12): 123514

Beven K, Freer J (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J Hydrol (Amst), 249(1–4): 11–29

Beven K J (2006). Searching for the Holy Grail of scientific hydrology: $Q_t = (SR)A$ as closure. Hydrol Earth Syst Sci, 10(5): 609–618

Beven K J (2016). Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. Hydrol Sci J, 61(9): 1652–1665

Beven K J, Smith P J, Freer J E (2008). So just why would a modeller choose to be incoherent? J Hydrol (Amst), 354(1): 15–32

Clark M P, Kavetski D, Fenicia F (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resour Res, 47(9): https://doi.org/10.1029/2010WR009827

Clark M P, Nijssen B, Lundquist J D, Kavetski D, Rupp D E, Woods R A, Freer J E, Gutmann E D, Wood A W, Brekke L D, Arnold J R, Gochis D J, Rasmussen R M (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. Water Resour Res, 51(4): 2498–2514

Eyring V, Bony S, Meehl G A, Senior C A, Stevens B, Stouffer R J, Taylor K E (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geosci Model Dev, 9(5): 1937–1958

Gelman A, Shalizi C R (2013). Philosophy and the practice of Bayesian statistics. Br J Math Stat Psychol, 66(1): 8–38

Gong W, Gupta H V, Yang D, Sricharan K, Hero A O III (2013). Estimating epistemic and aleatory uncertainties during hydrologic modeling: an information theoretic approach. Water Resour Res, 49 (4): 2253–2273

Grünwald P, Langford J (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. Mach Learn, 66(2–3): 119–149

Hornik K (1991). Approximation capabilities of multilayer feedforward networks. Neural Netw, 4(2): 251–257

Kinney J B, Atwal G S (2014). Equitability, mutual information, and the maximal information coefficient. Proc Natl Acad Sci USA, 111(9): 3354–3359

Metropolis N (1987). The beginning of the Monte Carlo method. Los Alamos Sci, 15(584): 125–130

Montanari A (2007). What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology. Hydrol Processes, 21(6): 841–845

Nearing G S, Gupta H V (2015). The quantity and quality of information in hydrologic models. Water Resour Res, 51(1): 524–538

Nearing G S, Mocko D M, Peters-Lidard C D, Kumar S V, Xia Y (2016a). Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. J Hydrometeorol, 17(3): 745–759

Nearing G S, Tian Y, Gupta H V, Clark M P, Harrison K W, Weijs S V (2016b). A philosophical basis for hydrologic uncertainty. Hydrol Sci J, 61(9): 1666–1678

Popper K R (1959). The Logic of Scientific Discovery. London: Hutchinson & Co.

Rasmussen C, Williams C (2006). Gaussian Processes for Machine Learning. Gaussian Processes for Machine Learning. Cambridge, MA: MIT Press

Renard B, Kavetski D, Kuczera G, Thyer M, Franks S W (2010). Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. Water Resour Res, 46(5): https://doi.org/10.1029/2009WR008328

Shannon C E (1948). A mathematical theory of communication. Bell Syst Tech J, 27(3): 379–423

Stanford K (2016). Underdetermination of Scientific Theory. In: Zalta N, ed. The Stanford Encyclopedia of Philosophy

Taleb N N (2010). The Black Swan: the Impact of the Highly Improbable Fragility. New York: Random House Group

Van Horn K S (2003). Constructing a logic of plausible inference: a guide to Cox's theorem. Int J Approx Reason, 34(1): 3–24

Weijs S V, Schoups G, van de Giesen N (2010). Why hydrological predictions should be evaluated using information theory. Hydrol Earth Syst Sci, 14(12): 2545–2558