AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

COMMENTARY
10.1029/2019WR024918

Key Points:
- We advocate a hypothesis testing paradigm that explicitly accounts for the fact that models are not true and uncertainty is not quantifiable
- Develop a perspective on hypothesis testing and model evaluation based on information theory

Correspondence to:
G. S. Nearing,
gsnearing@ua.edu

Citation:
Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., & Gupta, H. V. (2020). Does information theory provide a new paradigm for earth science? Hypothesis testing. *Water Resources Research*, *56*, e2019WR024918. https://doi.org/10.1029/2019WR024918

Received 4 FEB 2019
Accepted 28 DEC 2019
Accepted article online 4 JAN 2020

Does Information Theory Provide a New Paradigm for Earth Science? Hypothesis Testing

Grey S. Nearing[1], Benjamin L. Ruddell[2], Andrew R. Bennett[3], Cristina Prieto[4], and Hoshin V. Gupta[5]

[1]Department of Geological Sciences, University of Alabama, Tuscaloosa, AL, USA, [2]School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ, USA, [3]Civil & Environmental Engineering, University of Washington, Seattle, WA, USA, [4]Environmental Hydraulics Institute "IHCantabria", Universidad de Cantabria, Santander, Spain, [5]Department of Hydrology & Atmospheric Sciences, University of Arizona, Tucson, AZ, USA

Abstract Model evaluation and hypothesis testing are fundamental to any field of science. We propose here that by changing slightly the way we think and communicate about inference—from being fundamentally a problem of uncertainty quantification to being a problem of information quantification—allows us to avoid certain problems related to testing models as hypotheses. We propose that scientists are typically interested in assessing the information provided by models, not the truth value or likelihood of a model. Information theory allows us to formalize this perspective.

1. Introduction

We propose for debate the following proposition: "Information quantification provides an alternative to uncertainty quantification as a basis for model evaluation and hypothesis testing in the Earth Sciences." This proposition is at least partially motivated by well-known problems related to model evaluation and hypothesis testing discussed by Oreskes et al. (1994), who argued that "it is impossible to demonstrate the truth of any proposition, except in a closed system" and concluded that because of this "we can never verify a scientific hypothesis of any kind." As described by Laudan (1990), this is perhaps an overly strong reading of the underdetermination problem, but it is nevertheless a serious issue when applying quantitative techniques that rely on assigning set memberships to models (including but not limited to likelihoods, probabilities, etc.).

Based on our experience in discussing these ideas within the community and as highlighted in several of the articles in the recent WRR debate series on hypothesis testing (Blöschl, 2017), most hydrologists recognize that no model, measurement, or prediction is true in a strict sense and that the goal of evaluating models or hypotheses is to measure something like utility (e.g., Beven, 2018), adequacy (e.g., Gupta et al., 2012), or as a tool in or compliment to a larger heuristic process (e.g.,Baker, 2017 ; Oreskes et al., 1994). The problem is that when we do hypothesis testing in practice, we use quantitative methods that are based on explicit epistemic axioms or assumptions, and there is—at least often—a fundamental discrepancy between those axioms versus what we actually expect to learn from these exercises. This problem is widely recognized in our discipline. One line of attack has been to abandon coherency (Beven et al., 2008)—by declining to be explicit about the epistemological axioms of our quantitative system (see Nearing, Tian, et al., 2016 for further explanation); the argument appears to be that this provides latitude to ignore the inherent inconsistencies between philosophy and practice.

In contrast, our goal here is to argue for a philosophy and practice for testing models and/or hypotheses that starts with explicit first principles and allows us to derive a quantitative hypothesis test that is coherent within the limits of the underdetermination problem mentioned above. Our argument is that we can do this using information theory, which allows us to ask questions directly about the information content of (strictly false, but still informative) models relative to arbitrarily uncertain data. The test that we end up deriving is organized around the following question: "Is there any information in experimental data that could be used to improve my model?" This is in principle an objective and binary question with a strict yes/no answer. This question does not require, at any point, assigning a truth value, degree of belief, likelihood, or any other type of set membership to any model or hypothesis. If the answer to our question is "yes," then we know that the

model could be improved without further experiment, and if the answer is "no," then we would expect to perform new experiments to improve the model.

2. Background

2.1. Hypothesis Testing Requires Model Evaluation

Before we begin, it's necessary to understand what we mean by "hypothesis testing" in the title of this article and in the preceding discussion. Explaining or predicting a phenomenon requires modeling it (Cartwright, 1983), and a model is necessarily a collection of hypotheses (Duhem, 1954; Hempel & Oppenheim, 1948; Stanford, 2016). Only an entire model, not its individual hypotheses, can be tested directly (Laudan, 1990). This perspective can—of course—be refined in ways that are meaningful for hydrological modeling (e.g., Clark et al., 2011), but for our purpose, it is sufficient to understand that hypothesis testing requires model evaluation. For the remainder of this essay—until the concluding discussion in section 4—we will not recognize a further distinction between model evaluation and hypothesis testing. This is a simplification made for brevity, and the reader is encouraged to consider how the ideas developed here might apply to more refined types of hypothesis testing.

2.2. Hypothesis Testing and Uncertainty Accounting

It is well-known that quantifying uncertainty in hydrology is a hard problem (Beven, 2016; Clark et al., 2011; Montanari, 2007; Renard et al., 2010), which is a polite way of saying that strictly reliable uncertainty accounting is impossible. We cannot quantify what we do not know. This is a problem because most hypothesis testing methods in the natural sciences require quantitative uncertainty characterization. In classical statistics, the central premise is to calculate the likelihood that some experimental data come from a specific model. Under a Bayesian framework, we estimate something like the probability of a model given a particular set of data. In both cases, and in all types of hypothesis testing that the authors are aware of, the test reduces to a comparison between two or more alternative models that are already known and clearly stated. For example, in classical statistics, this is the null versus test hypothesis, and in Bayesian statistics, any model that has finite support in the prior would be compared. Meaningful comparison with a "true" but unknown and unstated model is, and always will be, impossible.

Calculating the relevant quantities for a hypothesis test (e.g., p values, likelihood ratios, model evidences, etc.) requires one or more distributions representing uncertainty in the model and/or experimental data. It is the fundamentally unknowable nature of these uncertainty distributions that causes problems. In a Bayesian framework, this manifests as "misspecification," which occurs when the prior distribution puts zero probability density on the true model. This usually happens because we are unable to state any true model. Misspecification leads to potential for unbounded errors in inference (Gelman & Shalizi, 2013; Grünwald & Langford, 2007). This problem of "inconsistency under misspecification" was demonstrated, for example, in the context of hydrological parameter estimation by Beven et al. (2008), who showed that the wrong likelihood function leads to bias in estimated parameter values. Inconsistency is a serious problem.

2.3. Probability Theory

Before we describe our proposal for information-theoretic hypothesis testing, it is useful to outline certain epistemological foundations. More details on these foundations in the context of hydrological modeling were given by Nearing, Tian, et al. (2016).

The most common theory for reasoning about belief in the natural sciences is probability theory (e.g., Howson & Urbach, 1989). Following Jaynes (2003, Chapter 1,2), probability theory derives from the Aristotelian syllogisms, which take as axiomatic that any well-formed proposition is either true or false but not both. Probabilities represent the degree of belief that we place on the truth value of a model. Non-Aristotelian theories of reasoning do exist; for example, fuzzy logic allows for propositions to be simultaneously true and false to different degrees (Kosko, 1990), but this still requires that we assign set memberships to our models (e.g., models are partially true and partially false). While Cartwright (1983) argued that theories are not truth apt, that is, they do not admit a truth value one way or another, we argue that regardless of whether models admit truth values, the fact that no model is true in a logical sense means we should avoid basing our methods of inference on systems of logic based on truth values.

As an example, the prototypical Bayesian method of differentiating between models is

$$p(M|D) \propto p(D|M)p(M), \tag{1}$$

where M is a random variable related to the choice of model and D is a random variable related to the data from observed experimental outcomes. This can be expanded through various applications of the chain rule to account for different components of a particular type of modelling problem (e.g., Liu & Gupta, 2007; Montanari & Koutsoyiannis, 2012). However, when the purpose is to test a hypothesis about a complex system, what exactly is the set of Aristotelian propositions represented by the random variable M? Certainly, such propositions cannot be of the form "model $M = m$ is true" since we know that all models are false in an Aristotelian sense. In practice, we recognize that these probabilities are measures of relative informativeness of different models with respect to some particular data, yet we still use quantitative objects based on assigning truth values to models. As an example, Oreskes et al. (1994; footnotes 39 and 40) recognized an inherent contradiction in the concept of "weak verification," but did not recognize the inherent contradiction in assigning degrees of belief to models that are known to be false.

2.4. Information Theory

Oreskes et al. (1994) pointed out, correctly in our opinion, that what scientists really do is evaluate consistency between modeled versus observation data as part of a heuristic strategy for gathering evidence. What we actually interpret when we compare models with data is the ability of a model to provide information about observable outcomes. What we interpret is different than what we measure—we measure things like probabilities, likelihoods, degrees of belief, levels of credence, or degrees of truth, but we interpret these as informativeness. It's easy enough to ask our interpretation question directly, as we did above (i.e., "is there a way to improve my model given current experimental data?"). In the spirit of debate, is there any other question a scientist might want to ask? What about an engineer or a policymaker who has a practical, rather than epistemic, motivation?

In the case of the policymaker, who makes decisions using predictions from scientific models, what justification of a model could be more meaningful than a claim like "This model is the best we've been able to build with the currently available information"? Of course, a decision maker might still be interested to consider uncertainty in future predictions, and for that, she will—tautologically—need methods of uncertainty quantification. But for the purposes of hypothesis testing, we need not (and, we argue, must not) go further than quantifying the presence of information. Information measurement is a supportable and affirmative statement of what we know, whereas probabilistic error models and uncertainty are practically useful but fundamentally unsupportable negative statements of what we do not know. We cannot measure what we do not know.

Our strategy for measuring information starts with Shannon (1948), who derived a quantitative theory of communication from probability theory. His three (explicit) desired properties for a measure of information were related fundamentally to how that measure would interact with probability distributions. However, the fact that information theory was derived from probability theory and not the other way around is—at least from the perspective of the theories themselves —an accident of history. That is, both probability theory and information theory can be derived in parallel from a more basic set of axioms (Knuth, 2005). It is perhaps telling about human nature that our most successful quantitative theory of learning grew originally out of an interest in betting games rather than an interest in communication.

We argue that a sufficient logic of hypothesis testing is incomplete without information theory. To see this, consider the following thought experiment: We will roll two dice and observe their outcome. Before observing the two rolls, our doxastic (belief) states about the repeated experiments behave multiplicatively: There are $6 \times 6 = 36$ possible outcomes, and each outcome is associated with a (scalar) measure of belief that behaves according to a product rule. After observing the outcome of the rolls, our epistemic (knowledge) states about the repeated experiments behave additively: We now require $1+1 = 2$ pieces of information (the actual outcomes of each experiment) to capture everything we know. Doxastic states behave according to a product rule before observing experimental outcomes and epistemic states behave according to a sum rule afterwards:

$$\text{Product Rule}: \quad p(e_1, e_2) = p(e_2|e_1)p(e_1), \tag{2.1}$$

$$\text{Sum Rule}: \quad h(e_1, e_2) = h(e_2|e_1) + h(e_1). \tag{2.2}$$

Here e_i are the outcomes of the repeated experiments and p and h are measures of probability and information, respectively.

Probabilities describe how belief states interact before running an experiment, that is, probabilities are associated with possibilities. Conditional probabilities describe how belief states change due to collecting new information, but probabilities don't describe how interactions between belief states change during the process of observation (from multiplicative before to additive after).

3. Hypothesis Testing with Information Theory

3.1. Theoretical Basis

Gong et al. (2013) provided a unique perspective on testing models under information theory. The philosophy that we describe in this debate paper was developed by Nearing and Gupta (2015) and Nearing and Gupta (2018) on top of that foundation.

To establish notation, we will distinguish between two types of experimental data: experimental perturbations u and experimental responses y. Perturbation data are related to what we believe to be "causes" in our experimental apparatus, and response data are related to what we believe to be "effects" (for a discussion about causality in the context of information theory, see the companion article by Goodwell et al., 2019). The hypothetical models that we will discuss take perturbation data as inputs and predict response data, so that u are model inputs (forcings, parameters, initial conditions, etc.) and y are evaluation or test data. We will notate model predictions as $m(u)$.

Independent of our model/hypothesis, and working strictly with the experimentally observed data, there exists some quantity of information about y (measured system responses) contained in u (measured system perturbations). We measure this quantity as the information shared between y and u and notate $I(y; u)$. Similarly, after running the model, there is some quantity of information about y (measured system responses) contained in $m(u)$ (model predictions), which we measure as the mutual information between y and $\hat{y}$ and notate $I(y; m(u))$. In deterministic models, predictions are a function of inputs, $m(u)$, and in probabilistic models, the distribution over predictions is a function of the inputs. In either case, these three variables represent a Markov chain $y \rightarrow u \rightarrow m(u)$. The data processing inequality, which is a theorem in information theory (Kinney & Atwal, 2014), states that $m(u)$ cannot contain more information about y than is contained in u:

$$I(y;u) \geq I(y; m(u)). \tag{3}$$

Accordingly, a perfect model will always yield $I(y; u) = I(y; m(u))$. In more realistic situations, the data processing inequality tells us that the amount of information provided *by* the model will always be a fraction (less than 1) of the total information provided *to* the model. The difference between this fraction and unity measures the effects of model error. This concept is illustrated in Figure 1, which is partially adapted from Gong et al. (2013).

Data error and/or incompleteness is accounted for in the information metric $I(y; u)$. The perturbation data may not contain enough information to fully determine the response data. This insufficiency might be for a variety of reasons including instrument accuracy and/or precision, incomplete coverage of the boundary conditions (e.g., spatial interpolation of variables like rainfall), or any other source of data imperfection. This is not the same thing as uncertainty, because we never need to characterize or estimate the relationship between data and truth to admit that data are imperfect. However, data imperfection accounts for all of the same reasons or causes for why data are typically considered to contain uncertainty.

3.2. Bounded Estimation

The challenge is that, in practice, we cannot precisely measure the information content provided by the perturbation to explain the response data. Estimating information shared between two data sets requires constructing or somehow otherwise integrating over a joint empirical distribution, and this is difficult when u

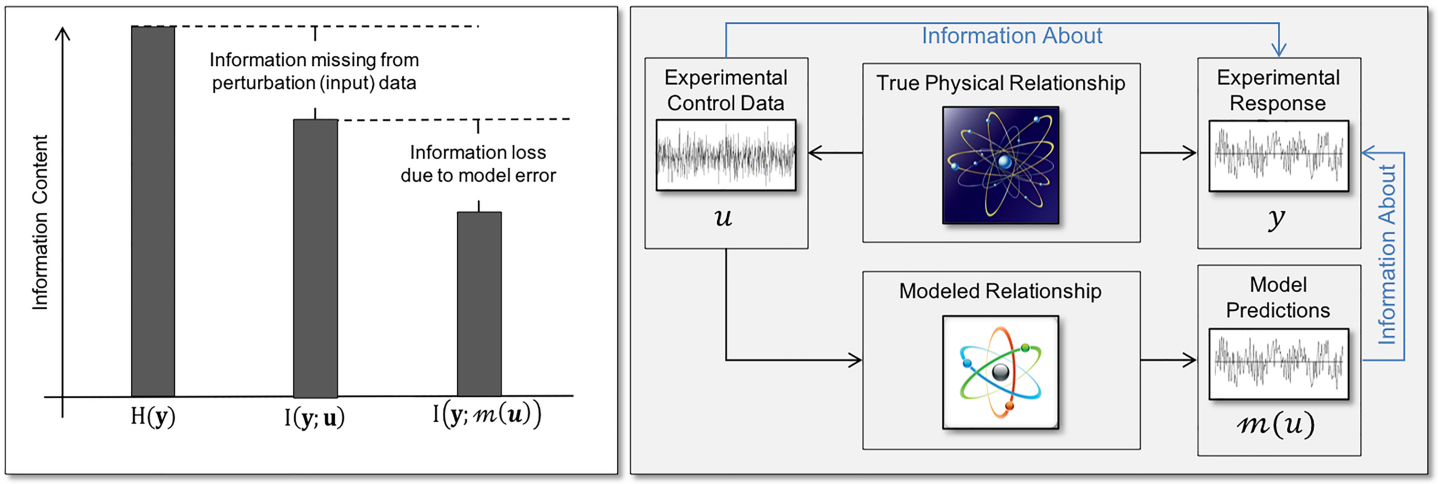

Figure 1. (Left) An information theory perspective on model performance, from Gong et al. (2013). $H(y)$ is the entropy of the observed response data, $I(y; u)$ is the fraction of that total entropy that is explained by perturbation data, and $I(y; m(u))$ is the fraction of $H(y)$ explained by model predictions. (Right) A diagram of a hypothesis testing procedure structured around this type of information partitioning. All of the z$_*$'s represent data sets, and the blue lines represent the information quantities we need to estimate.

is high dimensional. Most Earth systems models require high-dimensional inputs; in particular, any dynamical systems model requires boundary conditions specified through time and sometimes through space. Inputs to most hydrological models are multivariate time series. This dimensionality problem means that any procedure for estimating $I(y; u)$ will be approximate. This is a practical problem, but not a fundamental one.

Our goal, therefore, is to estimate $I(y; u)$ in a way that is consistent in some meaningful sense. In our case—for hypothesis testing—we want this estimate to be conservative in that it should be bounded above by the unknown true value of the information shared between u and y. We will notate this estimate $\widehat{I(y; u)}$. Because this is not the true value of $I(y; u)$, it is possible for the information content of model m predictions to exceed the estimated information content of the model inputs, despite the bounding relationship set by the data processing inequality. However, by another application of the data processing inequality, we have another theoretical statistic that always underestimates information missing due to model error. The reason for this is that we (necessarily) bound $I(y; u)$ by using another model, which we will notate r. This is our analog of the null hypothesis from classical statistics. The null model r takes the same input data u, so that

$$\widehat{I(y;u)} = I(y;r(u)) \leq I(y;u). \tag{4}$$

This yields a bounded quantity, $\widehat{\mathscr{E}}$, on the real quantity of information lost due to model error, $\mathscr{E}$:

$$\mathscr{E} = \mathrm{I(y;u)} - \mathrm{I(y;}m\mathrm{(u))} \qquad \widehat{\mathscr{E}} = \mathrm{I(y;}r\mathrm{(u))} - \mathrm{I(y;}m\mathrm{(u))} \qquad \mathscr{E} \geq \widehat{\mathscr{E}}. \tag{5}$$

We will refer to the mapping function r as a benchmark for model m. Equations (4) and (5) hold for any benchmark function r.

Like all hypothesis tests, equation (5) reduces to a comparison between two known and stated models, neither true: one being our hypothesis m and the other being a benchmark or null hypothesis r. The result of the test is simply a bounded measure of the information about available observation data from the null versus hypothesis model. This test is simple because the only assumption is use of a measure that obeys the data processing inequality. Almost any divergence measure will work (e.g., Csiszár, 1972); however, only Shannon's measures (i.e., mutual information) satisfy the thought experiment discussed in section 2.3. All we are doing is comparing models, but instead of prescribing or estimating uncertainty distributions, we use the empirical distribution between modeled versus measured data. The result is an interpretation of a test that does not violate probability theory, but also does not require models to be truth apt.

We believe that any benchmark model can be used for comparison, but we suggest choosing a high-quality data-driven model to set the benchmark $\mathscr{r}$ for a model $\mathscr{m}$ that we want to test. This is because using a purely data-driven model with minimal parametric assumptions, and certainly no assumptions about the (biogeo)physical properties of the hydrological system itself, means that the benchmark purely measures a quantity of information that can be extracted directly from observation data, rather than being a comparison between two different physical system conceptualizations, which is conflated with the limitations of the observation data. In our previously reported applications of this technique (Nearing, Mocko, et al., 2016; Nearing et al., 2018; Nearing & Gupta, 2015), $\mathscr{r}$ was usually a probabilistic, nonparametric regression that mapped $u \rightarrow y$—for example, a neural network or Gaussian process.

3.3. Discussion

To reiterate, our motivating question for this type of model testing is: "Is there any information in experimental data that could be used to improve my model?" By using a data-driven regression $\mathscr{r}$ as the benchmark, we can ask whether we were able to discover any information in data pairs $\{u,y\}$ that was not captured by model $\mathscr{m}$. The answer to this question could be no; since the benchmark $\mathscr{r}$ is another model, it is possible for the model $\mathscr{m}$ that we want to test to produce more information about experimental response data than the benchmark model $\mathscr{r}$.

It is important that the model error metric, $\widehat{\mathscr{E}}$, be bounded above because of the asymmetry of hypothesis testing. Under the Aristotelian logic, there is no syllogism that allows for scientific verification, only for falsification. In this case, the proposition that we are testing is "there exists no quantity of information in the available data that is not captured by the model." By having a conservative bound on $I(y;u)$, this proposition can be falsified by the *modus tollens*, which says that if proposition P implies proposition Q and the negation of Q (i.e., $\neg Q$) is observed, then proposition P is false: $(P \rightarrow Q) \wedge \neg Q \rightarrow \neg P$. Therefore, this perspective is fundamentally falsificationist and is a logically valid hypothesis testing perspective, but instead of falsifying models as hypotheses, we attempt to falsify propositions about the information content of models relative to experimental data.

It is worth noting that many of the standard model performance metrics used in the Earth Sciences are divergences (Nearing & Gupta, 2015) and preserve ordering according to the data processing inequality. This means that scientists have been intuitively circling the proposed perspective for many years. The Shannon-type information metrics (entropy and mutual information) are additive in the sense that the difference $I(y;\mathscr{r}(u))-I(y;\mathscr{m}(u))$ can be related directly to the fraction of uncertainty (as measured by entropy) about the response data y that could be extracted from input data, u, by an improved model (Nearing, Mocko, et al., 2016).

What we propose does not address the problem of underdetermination in the sense that no single metric can ever account for all aspects of any (finite) data set. However, it is worth noting that y and $\widehat{y}$ can be any transformation of observed or modeled data, respectively. As an example, we might be interested particularly in modeling peak flow. In this case, we would use a transformation of observed flow that emphasizes peak flows, for example, by only looking at 95th percentile flows, or by using an exponential transform of the discharge time series. All of the arguments we have made thus far apply to information about any particular aspect of any particular observed and/or modeled data, and it will typically be necessary to use multiple hypothesis tests on any given model.

4. Summary, Conclusions, and Practical Outlook

To summarize, the basic propositions argued here are the following:

a It would benefit us to change a fundamental assumption near the center of (at least) several philosophical theories about hypothesis-driven science—that is, that models admit (even unknown and unknowable) truth values.
b There exists a certain duality between probability theory and information theory that we might exploit in a way that is directly relevant to empirical science.
c Exploiting these two perspectives allows us to derive a hypothesis test that is bounded under arbitrary uncertainty in both data or models without ever having to directly characterize that uncertainty.

The arguments we have made here are mostly philosophical; we are trying to reconcile the theory and practice of model testing. A fair characterization of our argument is that we see room to improve how hydrologists and other Earth system scientists conceptualize and communicate about learning from models and data. One reviewer of this article suggested that the community already knows that we are not measuring the physical truth value of models as hypotheses, and so, the arguments presented here are redundant. While we fully agree, our point is that hydrologists—perhaps unintentionally—do indeed make these truth claims every time we assign an error distribution to our models or data, or any time we use any of the standard methods for model evaluation.

There is an incongruency between what we say we are doing versus what we actually do in practice. Our objective here is to address this inconsistency in a structured way that is explicit about all of our basic assumptions. We are arguing for a philosophical theory (and corresponding discussion) of model testing that looks more like what we do in practice, or vice versa. Our experience attempting to communicate this information-theoretic perspective has been that the intuition of at least many working scientists is based on an often incongruent mix of theory and practice, and we argue that this is not ideal for making sound and meaningful progress against the hard problems of inference in a science of complex systems. We see the effects of this nonrigorous intuition in our community's—often somewhat heated—discussions about uncertainty (Nearing, Tian, et al., 2016).

In our previous applications of this method (cited above and described below), we compared several different types of conceptual or physically based hydrology and hydrometeorology models with machine learning as the benchmark. The former included sophisticated land surface models currently used by several operational forecasting centers internationally (Nearing et al., 2018). In all cases that we have explored, high-quality machine learning models outperformed traditional hydrology models at estimating states (soil moisture) and fluxes (evapotranspiration, sensible heat flux, streamflow). This set of results is not fundamental to the theory outlined in the previous sections, but it does suggest an opportunity for Earth system scientists to improve the current generation of simulation models given data available from in situ networks and remote sensing.

As an example of the more refined types of hypothesis testing mentioned in section 2, Ruddell et al., 2019 combined the hypothesis testing theory outlined in this article with the process network theory that is described in the companion article in this debate series by Goodwell et al. (2019). The effect was to isolate model structural error from parameter error in specific process components of a complex ecohydrology model. The approach is based on the idea that if the model structure is correct, information loss from a model should be minimal at the same points in the parameter space where the information transfers between variables within the model are similar to the information transfers between those same variables in an observed system. Using a similar approach, Nearing et al. (2018) evaluated the viability of using ensembles of operational land surface models to characterize model structural uncertainty.

One interesting aspect of practical applications in the context of problems that are typically related to uncertainty characterization is that information theory often provides bounds, due to the data processing inequality. As an example of this, outside the context of hypothesis testing, Nearing et al. (2017) derived bounds on metrics of measurement error under arbitrary (and unknown) uncertainty in a triple collocation setting, which typically assumes a linear error model (McColl et al., 2014; Stoffelen, 1998).

To reiterate, the point of what we are trying to do here is to treat seriously the discrepancy between what scientists expect to learn from testing hypotheses or models versus the basic epistemic principles that underlie our formal, quantitative methods for testing hypotheses. We are not satisfied with previous attempts to handwave away these discrepancies by saying that hypothesis testing is only one part of a larger heuristic structure of scientific learning—while this is certainly true, the fact remains that the way we derive, teach, and apply hypothesis tests requires underlying assumptions that are in fundamental disagreement with how we interpret those tests. We argue that information theory provides a way around this problem.

Data Availability Statement

No data or code was generated as part of this project.

References

Baker, V. R. (2017). Debates—Hypothesis testing in hydrology: Pursuing certainty versus pursuing uberty. *Water Resources Research*, *53*, 1770–1778. https://doi.org/10.1002/2016WR020078

Beven, K. J. (2016). Facets of uncertainty: Epistemic error, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal*, *61*(9), 1652–1665. https://doi.org/10.1080/02626667.2015.1031761

Beven, K. J. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. *Wiley Interdisciplinary Reviews Water*, *5*(3), e1278. https://doi.org/10.1002/wat2.1278

Beven, K. J., Smith, P. J., & Freer, J. E. (2008). So just why would a modeller choose to be incoherent? *Journal of Hydrology*, *354*(1-4), 15–32. https://doi.org/10.1016/j.jhydrol.2008.02.007

Blöschl, G. (2017). Debates—Hypothesis testing in hydrology: Introduction. *Water Resources Research*, *53*, 1767–1769. https://doi.org/10.1002/2017WR020584

Cartwright, N. (1983). *How the laws of physics lie*. New York, NY: Cambridge University Press.

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*, W09301. https://doi.org/10.1029/2010WR009827

Csiszár, I. (1972). A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, *2*(1), 191–213.

Duhem, P. M. M. (Ed) (1954). *The aim and structure of physical theory*. Princeton, NJ: Princeton University Press.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Gong, W., Gupta, H. V., Yang, D., Sricharan, K., & Hero, A. O. (2013). Estimating epistemic & aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water Resources Research*, *49*, 2253–2273. https://doi.org/10.1002/wrcr.20161

Grünwald, P., & Langford, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, *66*(2-3), 119–149. https://doi.org/10.1007/s10994-007-0716-7

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*, W08301. https://doi.org/10.1029/2011WR011044

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175. https://doi.org/10.1086/286983

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. Chicago, IL: Open Court Publishing.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. New York, NY: Cambridge University Press.

Kinney, J. B., & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, *111*(9), 3354–3359. https://doi.org/10.1073/pnas.1309933111

Knuth, K. H. (2005). Lattice duality: The origin of probability and entropy. *Neurocomputing*, *67*, 245–274. https://doi.org/10.1016/j.neucom.2004.11.039

Kosko, B. (1990). Fuzziness vs. probability. *International Journal of General Systems*, *17*(2-3), 211–240. https://doi.org/10.1080/03081079008935108

Laudan, L. (1990). Demystifying underdetermination. *Minnesota Studies in the Philosophy of Science*, *14*(1990), 267–297.

Liu, Y. Q., & Gupta, H. V. (2007). Uncertainty in hydrologic modeling: toward an integrated data assimilation framework. *Water Resources Research*, *43*, W07401. https://doi.org/10.1029/2006WR005756

McColl, K. A., Vogelzang, J., Konings, A. G., Entekhabi, D., Piles, M., & Stoffelen, A. (2014). Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. *Geophysical Research Letters*, *41*, 6229–6236. https://doi.org/10.1002/2014GL061322

Montanari, A. (2007). What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology. *Hydrological Processes: An International Journal*, *21*(6), 841–845. https://doi.org/10.1002/hyp.6623

Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, *48*, W09555. https://doi.org/10.1029/2011WR011412

Nearing, G. S., & Gupta, H. V. (2015). The quantity and quality of information in hydrologic models. *Water Resources Research*, *51*, 524–538. https://doi.org/10.1002/2014WR015895

Nearing, G. S., & Gupta, H. V. (2018). Ensembles vs. information theory: Supporting science under uncertainty. *Frontiers of Earth Science*, *12*(4), 653–660. https://doi.org/10.1007/s11707-018-0709-9

Nearing, G. S., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V., & Xia, Y. (2016). Benchmarking NLDAS-2 soil moisture and evapotranspiration to separate uncertainty contributions. *Journal of Hydrometeorology*, *17*(3), 745–759. https://doi.org/10.1175/JHM-D-15-0063.1

Nearing, G. S., Ruddell, B. L., Clark, M. P., Nijssen, B., & Peters-Lidard, C. (2018). Benchmarking and process diagnostics of land models. *Journal of Hydrometeorology*, preprint online, *19*(11), 1835–1852. https://doi.org/10.1175/JHM-D-17-0209.1), null, doi: 10.1175/jhm-d-17-0209.1

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijs, S. V. (2016). A philosophical basis for hydrologic uncertainty. *Hydrological Sciences Journal*, *16*(9), 1666–1678.

Nearing, G. S., Yatheendradas, S., Crow, W. T., Bosch, D. D., Cosh, M. H., Goodrich, D. C., et al. (2017). Nonparametric triple collocation. *Water Resources Research*, *53*, 5516–5530. https://doi.org/10.1002/2017WR020359

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(5147), 641–646. https://doi.org/10.1126/science.263.5147.641

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 Resources Research*, *46*, W05521. https://doi.org/10.1029/2009WR008328

Ruddell, B., Drewry, D. T., & Nearing, G. S. (2019). Information theory for model diagnostics: Structural error is indicated by trade-off between functional and predictive performance. *Water Resources Research*, *55*, 6534–6554. https://doi.org/10.1029/2018WR023692

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Stanford, K. (2016). Underdetermination of scientific theory. In N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=scientific-underdetermination

Stoffelen, A. (1998). Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *Journal of Geophysical Research*, *103*(C4), 7755–7766. https://doi.org/10.1029/97JC03180