The Quantity and Quality of Information in Hydrologic Models

Grey S. Nearing – NASA Goddard Space Flight Center
Hoshin V. Gupta – University of Arizona

Citation of published version:

Nearing, G., Gupta, H. (2015): The Quantity and Quality of Information in Hydrologic Models. *Water Resources Research*, 51. DOI: 10.1002/2014WR015895

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR015895

Key Points:
- Models provide information of variable quality
- Information theory must be adapted to measure model info
- Dynamic systems models store information via induction

Correspondence to:
G. S. Nearing,
grey.s.nearing@nasa.gov

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

Received 27 MAY 2014
Accepted 25 NOV 2014
Accepted article online 5 DEC 2014
Published online 26 JAN 2015

The quantity and quality of information in hydrologic models

Grey S. Nearing[1] and Hoshin V. Gupta[2]

[1]NASA Goddard Space Flight Center, Hydrologic Sciences Lab, Greenbelt, Maryland, USA, [2]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA

Abstract The role of models in science is to facilitate predictions from hypotheses. Although the idea that models provide information is widely reported and has been used as the basis for model evaluation, benchmarking, and updating strategies, this intuition has not been formally developed and current benchmarking strategies remain ad hoc at a fundamental level. Here we interpret what it means to say that a model provides information in the context of the formal inductive philosophy of science. We show how information theory can be used to measure the amount of information supplied by a model, and derive standard model benchmarking and evaluation activities in this context. We further demonstrate that, via a process of induction, dynamical models store information from hypotheses and observations about the systems that they represent, and that this stored information can be directly measured.

1. Introduction

To explain a phenomenon is to find a model that fits it into the basic framework of the theory and that thus allows us to derive analogues for the messy and complicated phenomenological laws which are true of it.
Cartwright *[1983]*

The purpose of this paper is to provide a foundation for quantifying the role of models in science—both their ability to make informative predictions and their role in the process of induction. In doing so, we adopt *Cartwright*'s [1983] definition of a model as a tool that translates a set of hypotheses and/or theories into testable predictions: *"theories are schemata that need to be concretized and filled with the details of a specific situation, which is a task that is accomplished by a model"* [*Frigg and Hartmann*, 2009]. This definition agrees with that used in Hydrology (sometimes implicitly, sometimes explicitly) [e.g., *Beven*, 2000, 2001; *Clark and Kavetski*, 2010; *Fenicia et al*., 2011; *Gupta et al*., 2008, 2012].

Induction is the process of refining our beliefs about the truth value of a given hypotheses using experimental observations [*Howson and Urbach*, 1989, pp. 3–11]. However, it is rarely the case that hypotheses can be tested directly, precisely, because a model is necessary to translate any hypothesis into testable predictions. In particular, hydrological models require a hierarchy of hypotheses about appropriate representations of the hydrological system including about (i) what processes are important constraints on system behavior (the conceptual or process model), (ii) how we represent those processes mathematically in space and time (the mathematical model or system parameterization), and (iii) what assumptions are appropriate to make in the solution of the differential equations (the computational model) [*Gupta and Nearing*, 2014]. This is especially a problem in hydrology where the systems of interest are of relatively high complexity but intermediate randomness, which makes it difficult to develop general physical theories that both significantly constrain our uncertainty about the behavior of hydrological systems and also hold true across diverse watersheds [*Dooge*, 1986]. This means that hydrologists are typically restricted to using and testing models that are associated with either significant uncertainty (imprecision) or significant error (inaccuracy).

Although hydrologists have discussed extensively the philosophy and techniques behind developing and testing models, this question has recently been posed in the context of the idea that models provide information. For example, model benchmarking has been proposed as a way to differentiate between information provided by the model and information provided by boundary conditions [*Abramowitz*, 2005, 2012; *Luo et al*., 2012], and in this context, *van den Hurk et al*. [2011] posed as an open question whether *"model physics actually [add] information to the prediction system."* Similarly, *Gong et al*. [2013] attempted to measure

the amount of information lost due to model errors, and *Weijs et al*. [2010] argued that a particular information-theoretic measure provides more insight than traditional model evaluation metrics.

This paper unifies these perspectives by deriving the theory of model evaluation from first principles. It will come as no surprise to most scientists that all model evaluation is necessarily conducted in a benchmarking framework where the proposed model is compared to some baseline alternatives (e.g., the null hypothesis). That this is always the case follows directly from the inductive philosophy of science [*Jaynes*, 2003, p. 310] and allows us to reconcile *Gong et al.'s* idea that model error causes information loss with the intuition that even imperfect models may potentially provide information to forecasts. This reconciliation provides support for *Weijs et al.'s* argument that model evaluation is fundamentally an information theoretic endeavor.

In particular, we do two things: we define and measure information provided by a model in a completely general sense in the context of induction (section 2), and we show that it is possible to measure the amount of information stored in a model through induction (section 3). We provide an application example (section 4) that demonstrates how to use this theory to measure the amount of information obtained both from hypotheses that are encoded into models and from processes of conditioning models on observations.

2. Information Provided by Models

2.1. Uncertainty and Information

To understand models in terms of their information is motivated by the fact that induction is derived from Boolean logic [*Jaynes*, 2003, chapter 1] and therefore allows us to assign measures of belief to statements that can take on a Boolean truth value (T or F). Since we know that no meaningful hydrological model is correct in an absolute sense—because *"no substantial part of the universe is so simple that it can be grasped and controlled without abstraction"* [*Rosenblueth and Wiener*, 1945]—it is unreasonable to search for a "true" model. Moreover, practical application of the philosophy of induction allows us only to test models relative to a definite set of specified alternatives [*Jaynes*, 2003, p. 310]—i.e., Universal Induction is impossible in practice [*Rathmanner and Hutter*, 2011]. Therefore, it is necessary to adopt the perspective that *"all models are wrong but some are useful"* [*Box and Draper*, 1987, p. 424], and specifically to measure the information provided by a hydrological model. In this case, we can use induction on Boolean statements like "model $\mathscr{M}$ provides more and better information than the proposed alternatives." We argue that this question is, in fact, the one that is generally used in practice when developing and testing models.

Because all models are wrong, each should be associated with some amount of uncertainty. Since *Cox*'s [1946] theorem showed that *"the belief system of any rational agent must obey the standard axioms of probability"* [*Rathmanner and Hutter*, 2011, p. 5], every model is fundamentally probabilistic. Strictly deterministic models are degenerate probability distributions; however, even most deterministic models are implicitly treated as having greater than infinitesimal support when evaluated against data using standard statistical methods [*Weijs et al*., 2010].

In the context of induction, and since rational beliefs are measured probabilistically, information is a property (of either data or models) that changes our probability distributions [*Jaynes*, 2003]. *Shannon* [1948] offered what is perhaps the most famous measure of information in the context of *Cox*'s [1946] axioms (equivalently by proof, *Kolmogorov*'s [1956] axioms): that the expected amount of information in one random variable φ about the value of another random variable ξ is measured as the expected Kullback-Leibler (KL) [*Kullback and Leibler*, 1951] divergence to the marginal distribution over ξ from the distribution of ξ conditional on φ (this is called the mutual information between ξ and φ):

$$I(\xi;\varphi)=E_{\varphi}\left[D_{KL}\left(p_{\xi|\varphi}\|p_{\xi}\right)\right]. \tag{1}$$

These generic random variables ξ and φ may represent any phenomena about which we have some uncertainty, and in our context are typically models, model inputs, model predictions, or experimental observations. The KL-divergence is the integrated difference between negative log-transforms of two probability distributions over the same random variable:

$$D_{KL}(p'_{\xi}\|p_{\xi})=E_{p'_{\xi}}\left[\ln\left(p'_{\xi}(\xi)\right)-\ln\left(p_{\xi}(\xi)\right)\right]. \tag{2}$$

2.2. Model Information

Shannon's measure leads to results that apparently conflict with intuition in the context of estimating information provided by models. Whereas scientists believe that models at least potentially provide information

about the modeled system [*Doherty and Simmons*, 2013; *Reichle et al*., 2008; *van den Hurk et al*., 2011], *Gong et al*. [2013] point out that the data processing inequality [*Cover and Thomas*, 1991, p. 34] states that the amount of information about a predictand $\boldsymbol{Y} \in \mathbb{R}^{d_y}$ (terms like d_* refer to the dimension of the variable in the subscript) provided by the model $\mathscr{M}$ is no greater than the amount of information contained in model inputs (e.g., boundary conditions) $\boldsymbol{U} \in \mathbb{R}^{d_u}$ about the predictand:

$$I(\boldsymbol{Y};\boldsymbol{U}) \geq I(\boldsymbol{Y};\mathscr{M}(\boldsymbol{U})). \tag{3}$$

This seems to imply that the answer to *van den Hurk et al.'s* question about whether modeled physics add information to predictions is always "no." This is because equation (1) requires that we have perfect knowledge of the joint distribution between $\boldsymbol{U}$ and $\boldsymbol{Y}$: $p_{u,y}=p_{y|u}p_u$. We assert—contrary to what the data processing inequality may seem to imply—that $\boldsymbol{U}$ tells us nothing about $\boldsymbol{Y}$ without a mapping like $p_{y|u}$, which is the role of the model. What equation (3) actually states is that any model we may hypothesize will underperform compared to the true underlying (physical) relationship between inputs and predictands. Given a vector of inputs $\boldsymbol{u}$, a prototypical model $\mathscr{M}$ results in a prediction of $\boldsymbol{Y}$ (capital letters represent random variables and lowercase letters represent realizations of random variables):

$$p_{\mathscr{M}}(\boldsymbol{Y}|\boldsymbol{U}=\boldsymbol{u}) \leftarrow \mathscr{M}(\boldsymbol{u}). \tag{4}$$

$\mathscr{M}$ is a discriminative [*Ng and Jordan*, 2001] Bayes approximator in the sense that it directly facilitates an estimate of the distribution of $\boldsymbol{Y}$ conditional on $\boldsymbol{U}=\boldsymbol{u}$ (as opposed to a generative Bayes approximator, which would derive the $\boldsymbol{Y}|\boldsymbol{u}$ conditional from a joint distribution). $p_{\mathscr{M}}$ can be notated as a distribution of $\boldsymbol{Y}$ conditional on $\mathscr{M}$ by treating the model itself as a random variable [e.g., *Draper*, 1995; *Liu and Gupta*, 2007]:

$$p_{\mathscr{M}}(\boldsymbol{Y}|\boldsymbol{u})=p(\boldsymbol{Y}|\boldsymbol{u},\mathscr{M}). \tag{5}$$

To apply equation (1), we also need an estimate of the prior distribution of $\boldsymbol{Y}$ conditional on $\boldsymbol{u}$. Any relationship between $\boldsymbol{Y}$ and $\boldsymbol{u}$ (including independence) constitutes a model, and so the question of whether a model adds information about any predictand of interest is really a question of comparing two models. Equation (1) is similar to a likelihood ratio test except that we simply measure a (log-transformed) ratio of two predictive probability distributions rather than a ratio of the likelihoods of two models given observation data.

Following *Jaynes* [2003], we therefore adopt the following definition of information from a model—*any model provides information about* $\boldsymbol{Y}$ *when it changes the probability distribution over* $\boldsymbol{Y}$ *relative to a prior model* $p_p(\boldsymbol{Y}|\boldsymbol{u})$. For example:

$$I(\boldsymbol{Y};\mathscr{M}|\boldsymbol{u})=D_{KL}\left(p_{\mathscr{M}}\|p_{\not\mathscr{M}}\right), \tag{6}$$

where there is no expectation over the divergence since we are interested in the information from a single model. This definition is completely general under the constraints that beliefs (and uncertainties) are quantified using probability measures in accordance with Cox' theorem (really this means that we perform induction in the context of Boolean logic [*Jaynes*, 2003, chapter 1]).

2.3. The Prior

A prior is essential in any learning context [*Rathmanner and Hutter*, 2011]. For example, *Abramowitz* [2005, 2012] recognized the need to discriminate between the information provided by a model and *"information provided to a model."* Whereas *Gong et al*.'s [2013] use of a prior was formal while *Abramowitz'* was informal, both used empirical or semiempirical models to determine the amount of information in model inputs $\boldsymbol{u}$. This works because empirical models can give us some idea about the predictand given model inputs *before* adding specific (e.g., physics-based) hypotheses. Of the many techniques that we might use to define empirical information priors, examples include (i) neural networks with bootstrapping [e.g., *Schaap et al*., 2001], (ii) Gaussian process regressions [*Rasmussen and Williams*, 2006], and (iii) empirical density functions with estimates of asymptotic variance [*Guillou and Merlevède*, 2001]. It is not strictly necessary to use an empirical prior—any model can be used as the baseline for assessing the information contribution of any

other model. Empirical information priors simply allow us to test specific sets of hypotheses in a relatively context-free setting.

2.4. Information Quality

Because information is here defined simply as a change in beliefs (uncertainty), and because models inevitably contain errors, it is important to consider the *quality* of information provided by a model [*Nearing et al.*, 2013a]. Measuring information quality is an exercise in model evaluation in that it requires some concept of truth—generally a target estimate of the distribution over the predictand $\boldsymbol{Y}$, which we call the evaluation distribution and notate p_e. Classically, this distribution is obtained by extrapolating a set of repeated observations (the frequentist approach), or by an a priori estimate of the uncertainty on a single observation. The latter, used in data assimilation and many instances of likelihood-based parameter estimation [*Evensen and van Leeuwen*, 2000; *Reichle et al.*, 2008; *Vrugt et al.*, 2006], is necessary when observations are not individually repeatable. Alternatively, if the model is to provide information in lieu of an observing system, we would want observations to add as little information as possible conditional on the model estimate. In that case, an N-bin histogram-type (discrete) evaluation distribution can be obtained by minimizing the cost function:

$$\sum_{i=1}^{N} \frac{e_i \; p(\boldsymbol{z}|\boldsymbol{y}_i)}{\sum_{j=1}^{N} e_i \; p(\boldsymbol{z}|\boldsymbol{y}_i)} \ln\left(\frac{\sum_{j=1}^{N} e_i \; p(\boldsymbol{z}|\boldsymbol{y}_i)}{p(\boldsymbol{z}|\boldsymbol{y}_i)} \right), \tag{7}$$

with respect to the N e_i values, where $e_i = p_e(\boldsymbol{Y} = \boldsymbol{y}_i)$, and where $\boldsymbol{z} \in \mathbb{R}^{d_y}$ are observations.

In general, the concept that information has quality simply means that models do not always increase prediction accuracy, or even increase the probability assigned to the true prediction. As an example, *Nearing et al.* [2013a] defined *good* information for a scalar Y as accumulating when, for a given value of y, the value of $p_{\mathscr{H}}(y|\boldsymbol{u})$ is strictly greater than or strictly less than both $p_e(y)$ and the approximate Bayesian posterior (for model benchmarking this is $p_{\mathscr{M}}(y|\boldsymbol{u})$), and *bad* information as accumulating when $p_{\mathscr{H}}(y|\boldsymbol{u})$ is between $p_e(y)$ and $p_{\mathscr{M}}(y|\boldsymbol{u})$ (illustrated in Figure 1):

$$I_{good}(Y; \mathscr{M}|\boldsymbol{u}) = \int \left(1_{p_{\mathscr{H}}(y|\boldsymbol{u}) > p_e(y) > p_{\mathscr{M}}(y|\boldsymbol{u})} + 1_{p_{\mathscr{H}}(y|\boldsymbol{u}) > p_{\mathscr{M}}(y|\boldsymbol{u}) > p_e(y)} + 1_{p_{\mathscr{M}}(y|\boldsymbol{u}) > p_e(y) > p_{\mathscr{H}}(y|\boldsymbol{u})} + 1_{p_e(y) > p_{\mathscr{M}}(y|\boldsymbol{u}) > p_{\mathscr{H}}(y|\boldsymbol{u})} \right) \left(\ln\left(p_{\mathscr{M}}(y|\boldsymbol{u})\right) - \ln\left(p_{\mathscr{H}}(y|\boldsymbol{u})\right) \right) p_{\mathscr{M}}(y|\boldsymbol{u}) dy, \tag{8.1}$$

$$I_{bad}(Y; \mathscr{M}|\boldsymbol{u}) = \int \left(1_{p_e(y) > p_{\mathscr{H}}(y|\boldsymbol{u}) > p_{\mathscr{M}}(y|\boldsymbol{u})} + 1_{p_{\mathscr{M}}(y|\boldsymbol{u}) > p_{\mathscr{H}}(y|\boldsymbol{u}) > p_e(y)} \right) \left(\ln p_{\mathscr{M}}(y|\boldsymbol{u}) \right) - \ln\left(p_{\mathscr{H}}(y|\boldsymbol{u})\right) p_{\mathscr{M}}(y|\boldsymbol{u}) dy \tag{8.2}$$

Intuitively, good information accumulates at a particular value of y when the model moves the probability associated with $Y = y$ toward the evaluation distribution relative to the prior, and bad information accumulates when the model moves the probability of $Y = y$ away from the evaluation distribution. Two disjoint integrations over the differences between (log-transformed) $p_{\mathscr{H}}$ and $p_{\mathscr{M}}$ are used to measure each type of information, and these integrations sum to the divergence so that good and bad information sum to total information provided by the model.

Using the KL-divergence, good and bad information are unbounded (either can be negative), and values close to zero do not necessarily indicate that no good (bad) information has accumulated anywhere in the domain of the predictand. The method is more intuitive when using other types of divergences, for example, certain f-divergences [*Csiszár*, 1972], of the form:

$$D_f(p_y \| p_y') = E_{p_y'} \left[f\left(\frac{p_y}{p_y'} \right) \right], \tag{9}$$

where f is concave (to obtain the KL-divergence set $f(\omega) = -ln(\omega)$). However, no divergence is generally preferable to all others [*Ullah*, 1996], and we do not include a particular divergence or difference measure in our definition of information because the modeler should be free to select a preferred divergence function for any given application (equations (6) and (8) are examples that use the KL-divergence). Notice that given a deterministic observation (i.e., a delta evaluation distribution at $\boldsymbol{z}$), the log of the likelihood ratio of the model to the prior is proportional to the amount of good information measured using the

KL-divergence. Similarly, most common model evaluation statistics are actually information measures—some examples are given in Appendix A.

3. Information Stored in Models

Models encode hypotheses about the structure of the system, and we may employ some process of induction to refine these hypotheses. It therefore seems intuitive that models at least potentially contain information about the system collected from hypotheses and data. When new hypotheses or observations result in changes to model structure, the new information can be measured as a divergence from the posterior to prior model.

This concept is meaningful in the context of dynamical systems models, for example:

$$d\boldsymbol{x}_t = m(\boldsymbol{u}_t, \boldsymbol{x}_t)dt + v(\boldsymbol{u}_t, \boldsymbol{x}_t)dW_t, \tag{10}$$

where $\boldsymbol{x}_t \in \mathbb{R}^{d_x}$ is the time-dependent dynamic state. The Wiener process, W_t, with variance $v(\cdot)^2$ is a common approximation [*Evensen and van Leeuwen*, 2000; *Mitchell and Houtekamer*, 2000; *Reichle et al*., 2008] of epistemic uncertainty (i.e., effects due to processes not accounted for, or misrepresented by, the drift function $m(\cdot)$). Dynamic systems models are typically solved by Monte Carlo integration [*Metropolis*, 1987]—for example, equation (10) is usually solved by sampling a discrete-time Euler-Maruyama integration:

$$\boldsymbol{x}_t = m(\boldsymbol{u}_t, \boldsymbol{x}_{t-1})\Delta t + \boldsymbol{v}_t, \tag{11}$$

where $\boldsymbol{v}_t \in \mathbb{R}^{d_x}$ is white noise to approximate epistemic Brownian motion.

Information stored in a model such as equation (11) is measured as a divergence from the modeled state transition distribution $p_m(\boldsymbol{x}_t|\boldsymbol{u}_t, \boldsymbol{x}_{t-1})$ to a prior state transition distribution. Notice that the epistemic uncertainty need not be Gaussian; we chose this example simply for convenience in notation and because its application is nearly ubiquitous in hydrology and the atmospheric sciences. It is, however, necessary that both the prior and posterior model structures be distributions over states drawn from the same sample space. One nonparametric technique that adapts well to Monte Carlo integration is to emulate state transition models such as equation (11) using kernel density estimators—this is illustrated in section 4.

Since it is impossible to know a "true" model structure, it is impossible to measure the *quality* of information stored in a model. While this is a fundamental limitation, quality can be approximated given a "best" estimate of model structure that may be achieved as the end result of an inductive process. In that case, quantities of good and bad information measure agreement and compensation between hypotheses and observations included in the inductive process.

4. Illustrative Example

To make these ideas concrete, we illustrate the development of a model to estimate daily streamflow over a 25 day evaluation period during and after a storm event. We test the information added to the model via five distinct steps, each of which results in a new model that provides a distribution over a time series of daily streamflow from a time series of daily precipitation and potential evaporation. The first three steps involve building a model from one or more hypotheses about appropriate descriptions of the watershed. These hypotheses are cumulative so that inclusion of each refines the model that results from the preceding set. The final two steps involve the incorporation of observational evidence via parameter estimation and system identification. The five steps are (further detail in Appendix B):

1. Continuity Hypothesis (m_{cont}): the watershed preserves mass. The scalar state represents total water stored in the watershed.
2. Infiltration and Basic Routing Hypotheses (m_{rout}): the watershed stores mass in an unsaturated zone x^{sm} (mm) and a routing process x^{rt} (mm) ($d_x = 2$). This introduces two parameters: soil moisture capacity and a runoff coefficient.
3. Saturated and Overland Flow Hypotheses (m_{HyMod}): the routing process is segregated into flow within a saturated zone and overland flow. This model has seven parameters (Table 1) and $d_x = 5$. The model that results from this final set of hypotheses is the HyMod simulator [*Boyle*, 2000].

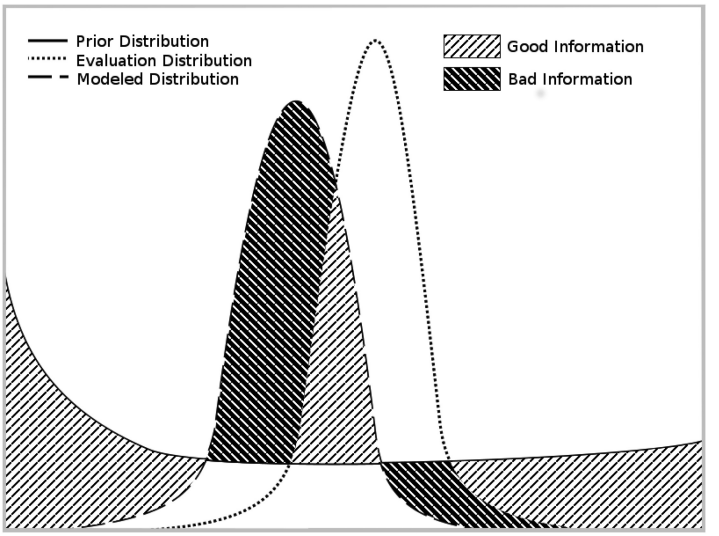

Figure 1. Areas between the (negative-log transformed) prior, modeled, and evaluation distributions that contribute to good and bad information.

4. Parameter Estimation ($\mathscr{m}_{\theta est}$): HyMod parameters and initial states are calibrated using data from a 1000 day calibration period (calibrated values are listed in Table 1); $d_x = 5$.
5. System Identification ($\mathscr{m}_{SysID}$): the mathematical structure of the calibrated HyMod is updated by conditioning on calibration period observations using five iterations of expectation-maximization system identification (EM-ID) [*Nearing*, 2013, Appendix D]; $d_x = 5$.

An empirical prior distribution over each daily streamflow value $p_{\not\emptyset}(y_t|\boldsymbol{u}_{1:t})$ was estimated from calibration period data by sparse Gaussian process regression (SGPR) [*Snelson and Ghahramani*, 2006] acting on 16 lagged input values. The effect of varying the number of lagged inputs on the divergence from the empirical prior to the evaluation distribution is illustrated in Figure 2 (a *nat* is a unit of information where f in equation (9) includes a natural logarithm). The prior and each of the five models were used to simulate streamflow at each day during the evaluation period: $p(y_t|\boldsymbol{u}_{1:t}, \mathscr{m}_{HyMod})$; these predictions are presented in Figure 3.

Information contributed to predictions by each model was estimated as the divergence from the model conditional distribution to the empirical prior:

$$I(y_t; \mathscr{m}) = D\big(p(y_t|\boldsymbol{u}_{1:t}, \mathscr{m}) \| p_{\not\emptyset}(y_t|\boldsymbol{u}_{1:t})\big). \tag{12}$$

We used *Jeffreys'* [1946] divergence with $f(\alpha) = (\alpha - 1)ln(\alpha)$ to ensure that good and bad information values are finite and that values close to zero indicate that little good or bad information was accumulated anywhere in the domain of y_t. Equation (12) was estimated at each evaluation time step using the method of sieves [*Paninski*, 2003] with bin width according to Scott's rule [*Scott*, 2004].

Catchment models are usually calibrated and evaluated using a time series of individual (rather than repeated) observations of streamflow. Absent specific knowledge about measurement error observational uncertainty is often assumed zero-mean Gaussian [*Vrugt et al.*, 2006]; we used a truncated (at zero) Gaussian uncertainty distribution centered on the observation value with standard deviation 0.5 (mm/d).

Figure 4 shows the time-averaged values of total, good and bad information about predictions contributed by each model relative to the empirical prior, as well as the time-averaged ratio of good information. Each model adds between one and two nats of good information; however, all models except $\mathscr{m}_{SysID}$ add more than four nats of bad information. In this case, conditioning on observations (especially via EM-ID) largely serves to reduce bad information about predictions.

The state transition distribution $p(\boldsymbol{x}_t|\boldsymbol{u}_t, \boldsymbol{x}_{t-1}, \mathscr{m})$ of each model was emulated by d_x conditionally independent SGPR—that is, a separate SGPR mapping was constructed for each state dimension: $p(x_t^i|\boldsymbol{u}_t, \boldsymbol{x}_{t-1}, \mathscr{m})$. Information stored in model $\mathscr{m}(I_{\mathscr{M}})$ was measured as the sum of KL-divergences from the d_x conditionally independent posterior Gaussian processes (GP) to standard normal GP priors:

$$I_{\mathscr{M}} = \sum_{i=1}^{d_x} D_{KL}(\mathscr{G}_i \parallel \mathscr{G}_0) \tag{13.1}$$

$$D_{KL}(\mathscr{G}_i \parallel \mathscr{G}_0) = \frac{1}{2}\left[trace(\boldsymbol{K}_i) + \boldsymbol{\mu}_i^{\mathsf{T}}\boldsymbol{\mu}_i - q + \ln|\boldsymbol{K}_i|\right]. \tag{13.2}$$

$\boldsymbol{K}_i \in \mathbb{R}^{q,q}$ is the covariance matrix of $\mathscr{G}_i$, and $\boldsymbol{\mu}_1 \in \mathbb{R}^{q,1}$ and $\boldsymbol{\mu}_0 \in \mathbb{R}^{q,1}$ are means of the posterior and prior GPs, respectively. q is the number of times each GP was sampled; we used $q = 10^5$.

Table 1. HyMod Parameters and Initial States, Their Uncertainty Ranges and Calibrated Values

Parameter or Initial State	Range	Calibrated Value	Units
ϕ	100–600	380.27	mm
b	0.05–1.95	0.15	
ϱ	0.5–0.95	0.90	
k_{sf}	0.001–0.1	0.07	
k_{qf1}	0.3–0.95	0.79	
k_{qf2}	0.3–0.95	0.77	
k_{qf3}	0.3–0.95	0.91	
x^{sm}	0–600	328.65	mm
x^{sf}	0–300	76.08	mm
x^{qf1}	0–100	95.06	mm
x^{qf2}	0–50	39.77	mm
x^{qf3}	0–50	2.59	mm

Figure 5 compares the information stored in each model structure relative to the standard normal GP prior. It is intuitive that increasing the state dimension increases the information in the model. We also see that choosing a particular set of parameter values removes information as compared to using uniform distributions over parameters. This simply means that relative to a standard normal GP prior we lose some amount of information when we move from a uniform distribution to a delta distribution over parameters. There is no indication as to whether this lost information due to calibration is good or bad because in this case we did not consider an evaluation model. Notice that we could measure the divergence from the calibrated HyMod state transition distribution to the uncalibrated HyMod state transition distribution; in that case, parameter estimation would necessarily add net information relative to the prior defined by HyMod with uniform parameter distributions. Finally, we notice that conditioning the mathematical structure of the model on observations (via EM-ID) adds information to the model structure—we have no way of directly assessing the quality of this information other than indirectly against observations of model predictions.

The conclusion from Figures 4 and 5 is that system identification (structure updating) is, in this case, much better at extracting good information about predictions from observations than is variational parameter estimation. There is more information in the observations than can be extracted via parameter estimation—namely, information about errors in the model structure, which the model translates into information about predictions. This does not mean that system identification can fix fundamentally broken models—if there is an important process represented in the model that does not contribute substantially to predictions of available past observations but may instigate a regime change in the future, this cannot be detected by system identification.

5. Concluding Discussion

In this paper, we discuss how to quantify the role of models during induction, and this discussion provides a basis to formalize the process of model benchmarking. Although *Cover and Thomas* [1991, p. 13] argue that "*the concept of information is too broad to be captured completely by a single definition*," we argue that it is straightforward to define information in a standard way when applied to the problem of scientific induction under the formalism [*Schement and Ruben*, 1993, pp. 6–9] that information results from communication rather than being the communicated signal itself. This is because all learning problems can be quantified

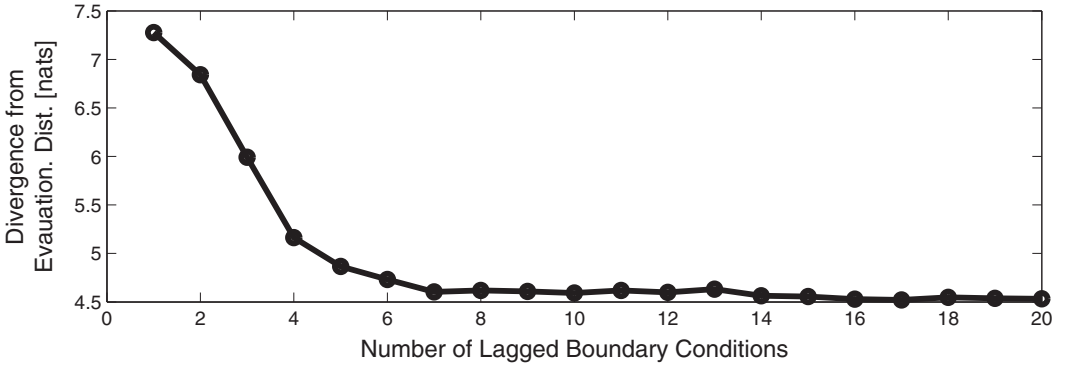

Figure 2. Sensitivity of the diagnostic empirical SGPR prior to the number of lagged input vectors (boundary conditions precipitation and potential evaporation) estimated by the time-averaged *Jeffreys'* divergence from the SGPR distribution to the evaluation distribution. We chose to use 16 lagged inputs vectors.

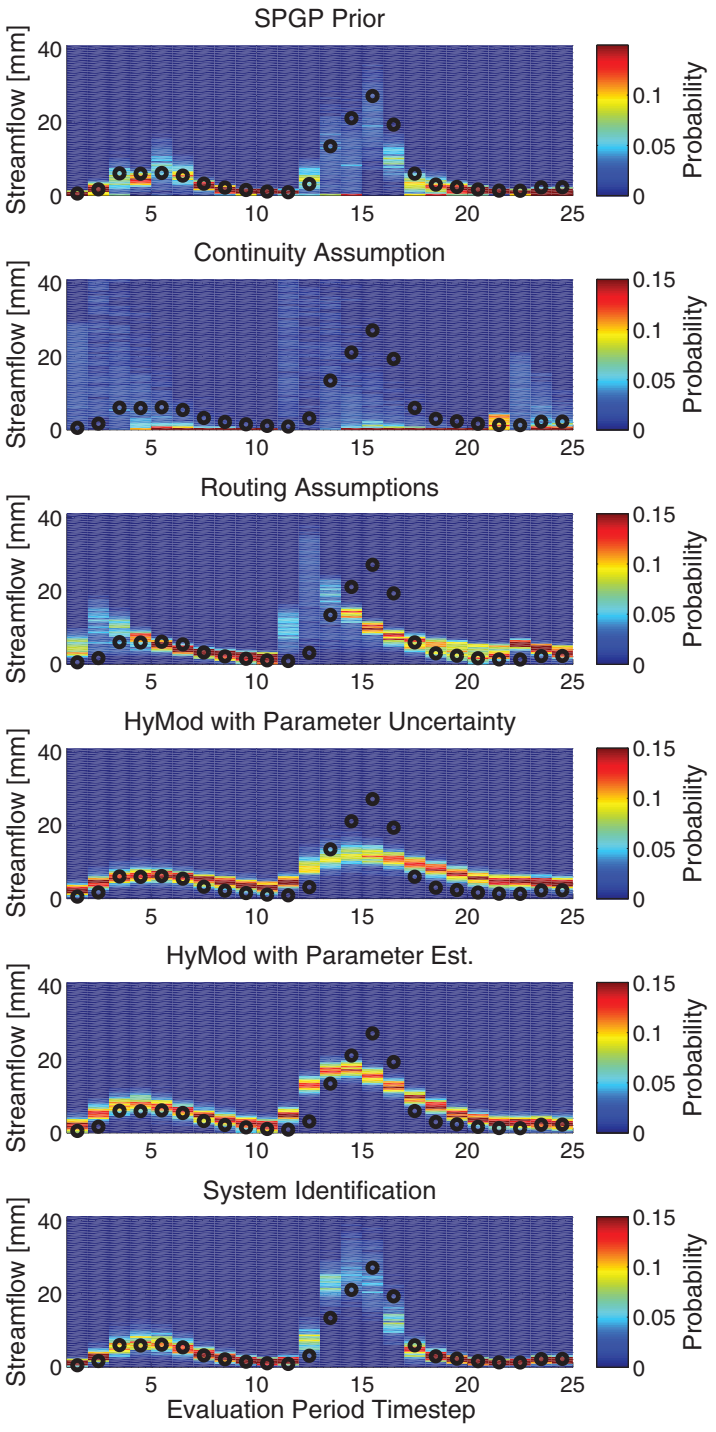

Figure 3. Evaluation period predictions of daily streamflow made by the empirical SGPR prior and each of the five inductive models. Black circles mark daily streamflow observations.

using the calculus of probabilities, and information can be defined generally in this context as a change in probability distributions.

Even with a universal definition and concept of information, it is not possible to define a unique *measure* of information that suits all purposes. The typical measures and theorems (i.e., Shannon's) that comprise information theory only apply in the case when we can apply Bayes' law exactly [*Nearing et al.*, 2013b], and are therefore not particularly meaningful when evaluating models because models are discriminative approximations of Bayes' theorem that by their very nature contain errors. So the question that this paper addresses is how to interpret and measure the idea that models provide information even though we know that no model is correct. It turns out that comprehensive understanding of model information requires a concept of information quality, and also an application-oriented divergence measure. Again we point out that most common model performance metrics (e.g., bias, mean squared error, correlation coefficient) are measures of information under our definition (see Appendix A).

It is also necessary to designate a prior distribution that codifies the learning *context*. This has been done in hydrology using empirical models [*Abramowitz*, 2005; *Gong et al.*, 2013], which allow us to test physics-based models in an isolated but meaningful context by asking questions like "how much information does this physics hypothesis add to predictions?" We should be careful to point out that we are not necessarily looking for a hypothetical *"true"* prior but rather a prior that is meaningful in context of evaluating the particular models we are interested in. In other words, it is not the purpose of the prior to codify everything we know before building the model, only to provide context for testing various models.

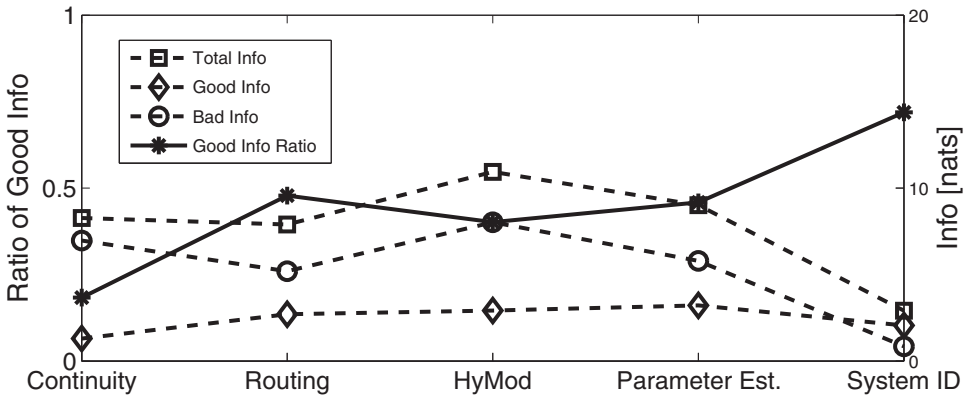

Figure 4. Time-averaged total, good, and bad information (right-axis) and time-averaged fraction of good information (left axis) about evaluation period streamflow contributed by each model. Each plotted point represents an average over the 25 (divergence, partial divergence, or ratio) values from the evaluation period, and each divergence is measured from the modeled distribution to the empirical SGPR prior.

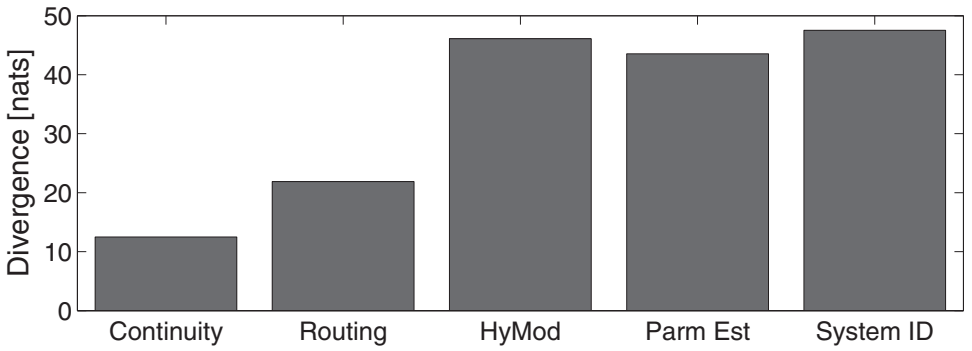

Figure 5. Measurements of information stored in progressive model structures. Each information divergence is from the model-conditional state transition mapping $p(\boldsymbol{x}_t|\boldsymbol{u}_t, \boldsymbol{x}_{t-1}, \mathscr{M})$ to the d_x-dimensional independent standard normal GP prior.

Most importantly, we have demonstrated precisely what it means to say that a model provides information and how this perspective generalizes and provides a formal connection between the philosophy of induction [*Jaynes*, 2003] and the need of applied scientists to evaluate and benchmark their models.

Appendix A: Derivation of Common Statistics as Measures of Information

1. Bias: The bias is obtained for any model and evaluation distribution using the Csiszár function $f(u)=y[u-1]$:

$$E_{p_{\mathscr{M}}}\left[y\left(\frac{p_e(y)}{p_{\mathscr{M}}(y)}-1\right)\right]=\int y\frac{p_e(y)}{p_{\mathscr{M}}(y)}p_{\mathscr{M}}(y)dy-\int yp_{\mathscr{M}}(y)dy. \tag{A1.1}$$

$$=\int yp_e(y)dy-\int yp_{\mathscr{M}}(y)dy=E_{p_e}[y]-E_{p_{\mathscr{M}}}[y]. \tag{A1.2}$$

2. Mean Squared Error: The mean squared error (MSE) assumes either a Gaussian model and a deterministic evaluation distribution or a Gaussian evaluation distribution and a deterministic model. Following *Weijs et al.*'s [2010] argument that deterministic models are treated as probabilistic during evaluation, the example is that some deterministic simulator predicts the outcome of n experiments as $\{y_i\}_{i=1,\ldots,n}$ which are collectively taken to be the mean of an n-dimensional joint Gaussian with unit variance and independent marginals: $p_{\mathscr{M}}(Y_i|\boldsymbol{u})=\mathscr{N}_{y_i,1}(y_i)$. The evaluation distribution is taken to be a delta function at the value of the observation:

$$p_e(\boldsymbol{Y})=\prod_{i=1}^{n}\delta_{Y_i,z_i};\qquad \delta_{Y_i,z_i}=\begin{cases}1; & Y_i=z_i\\ 0; & Y_i\neq z_i\end{cases}. \tag{A2}$$

Using the Csiszár function $f(u)=u(2n^{-1}\ln(u)-\ln(2\pi))$ in the divergence from the evaluation to the model, we obtain:

$$\int \frac{\prod_{i=1}^{n}\delta_{y_i,z_i}}{\prod_{i=1}^{n}\mathscr{N}_{y_i,1}(y_i)}\left(2n^{-1}\ln\left(\frac{\prod_{i=1}^{n}\delta_{y_i,z_i}}{\prod_{i=1}^{n}\mathscr{N}_{y_i,1}(y_i)}\right)-\ln(2\pi)\right)\prod_{i=1}^{n}\mathscr{N}_{y_i,1}(y_i)d\boldsymbol{y}, \tag{A3.1}$$

$$=\int\left(-2n^{-1}\ln\left(\prod_{i=1}^{n}\mathscr{N}_{y_i,1}(y_i)\right)-\ln(2\pi)\right)1_{\boldsymbol{y}=\boldsymbol{z}}d\boldsymbol{y}, \tag{A3.2}$$

$$=\int n^{-1}\sum_{i=1}^{n}(y_i-y_i)^2 1_{\boldsymbol{y}=\boldsymbol{z}}d\boldsymbol{y}, \tag{A3.3}$$

$$=n^{-1}\sum_{i=1}^{n}(z_i-y_i)^2, \tag{A3.4}$$

where the first equality uses the convention that $0\ln(0) = 0$. The MSE therefore measures the information about the observations that is missing from a Gaussian model centered at a deterministic prediction (*Nearing et al.* [2013a] note that this is actually a combination of bad and missing information). Equivalently, the MSE can be interpreted as the divergence from a Gaussian observation to a deterministic model with $f(\omega)=-(2n^{-1}\ln(\omega)+\ln(2\pi))$.

3. Coefficient of Determination: The coefficient of determination ρ^2 for a linear model measures the information provided by a Gaussian linear model as the divergence to a Gaussian prior with mean and variance given by the observation sample mean and variance $\bar{z}$ and σ_z^2 from a discriminative joint Gaussian conditional with mean given by the α, β model coefficients and $\sigma_{\mathscr{M}}^2$ variance:

$$p_{\mathscr{M}}(\boldsymbol{Y}|\boldsymbol{u})=\sqrt{2\pi}(\sigma_{\mathscr{M}})^{-n}\exp\left\{-\frac{1}{2}\sigma_{\mathscr{M}}^{-2}\sum_{i=1}^{n}(\alpha u_i+\beta-Y_i)^2\right\}. \tag{A4}$$

Using the Csiszár function $f(\omega)=1-\omega^{\frac{2}{n}}$, the information in the linear model is:

$$I(\boldsymbol{Y};\mathscr{M}|\boldsymbol{u})=E_{\mathscr{M}}\left[1-\left(\frac{(\sigma_z)^{-n}\exp\left\{-\frac{1}{2}\sigma_z^{-2}\sum_{i=1}^{n}(z_i-Y_i)^2\right\}}{(\sigma_{\mathscr{M}})^{-n}\exp\left\{-\frac{1}{2}\sigma_{\mathscr{M}}^{-2}\sum_{i=1}^{n}(\alpha u_i+\beta-Y_i)^2\right\}}\right)^{\frac{2}{n}}\right] \tag{A5.1}$$

$$=E_{\mathscr{M}}\left[1-\left(\left(\frac{\sigma_{\mathscr{M}}}{\sigma_z}\right)^{n}\frac{\exp\left\{-\frac{1}{2}\sigma_z^{-2}\left(\sigma_z^2(n-1)\right)\right\}}{\exp\left\{-\frac{1}{2}\sigma_{\mathscr{M}}^{-2}\left(\sigma_{\mathscr{M}}^2(n-1)\right)\right\}}\right)^{\frac{2}{n}}\right] \tag{A5.2}$$

$$=E_{\mathscr{M}}\left[1-\frac{\sigma_{\mathscr{M}}^2}{\sigma_z^2}\right]=1-\frac{\sigma_{\mathscr{M}}^2}{\sigma_z^2}=\rho^2. \tag{A5.3}$$

The second equality results from the fact that the variance of the linear model is by definition the sample variance of the residuals. The correlation coefficient is the square root of the coefficient of determination when α, β, and σ_z^2 are estimated in the usual least squares manner, and the skill score (Nash-Sutcliffe efficiency) is the coefficient of determination when $\alpha=1$ and $\beta=0$.

Appendix B: Description of Example Models

The streamflow demonstration uses observations of daily precipitation u_t^{pp} (mm), potential evapotranspiration u_t^{pe} (mm), and daily streamflow y_t (mm) from 1951 to 1954 from the Leaf River catchment, Mississippi, USA. Thirty days were used for model spin-up and the next 1000 days (calibration period) to perform parameter estimation and system identification. The evaluation period was 14 March 1954 to 8 April 1954 (25 days); the evaluation period consists of a single storm event and was kept short to facilitate a simple demonstration.

1. Continuity Hypothesis: the continuity hypothesis asserts that mass flow out of the watershed cannot exceed the sum of past inflows and outflows. We encode this as a discrete-time hidden Markov model:

$$x_t=x_{t-1}+u_t^{pp}-\mu_t^{ae}-y_{t-1}; \tag{B1.1}$$

$$y_t \sim U[0, x_t], \tag{B1.2}$$

$$\mu_t^{ae} \sim U[0, \max(0, \min(u_t^{pe}, x_{t-1} + u_t^{pp} - y_{t-1}))], \tag{B1.3}$$

where x_t (mm) is water storage, and $U[\psi|a, b]$ is the probability measure at ψ of the uniform distribution with support on (a, b). Actual evapotranspiration μ_t^{ae} (mm) is sampled from a uniform distribution between zero and the minimum of u_t^{pe} or available water. The initial state is drawn from a uniform distribution $x_0 \sim U[0, 1100]$. This model was sampled $N_s = 1000$ times to estimate $p(\boldsymbol{x}_t|\boldsymbol{u}_t, \boldsymbol{x}_{t-1}, \mathscr{m}_{cont})$ and $p(y_t|\boldsymbol{u}_{1:t}, \mathscr{m}_{cont})$.

2. Infiltration and Basic Routing Hypotheses: routing assumptions partition the storage state into an unsaturated zone (x^{sm} (mm)), which controls infiltration, and the routing process (x^{rt} (mm)). Soil moisture capacity is determined by parameter $\phi \sim U[100, 600]$ (mm); infiltration is:

$$INF_t = \int_{x_{t-1}^{sm}}^{\min(\phi, u_t^{pp} + x_{t-1}^{sm})} \left(\frac{\phi - \xi}{\phi}\right)^b d\xi. \tag{B2.1}$$

$b \sim U[0.05, 1.95]$ represents spatial variability of infiltration capacity. Initial soil moisture is sampled from $x_0^{sm} \sim U[0, x_0]$ where x_0 is the total storage state, and the initial routing state is: $x_0^{rt} = x_0 - x_0^{sm}$. Distributions over parameters and initial states are model assumptions. Soil moisture is updated by infiltrating to capacity and subtracting potential evapotranspiration scaled by soil moisture:

$$x_t^{sm} = \frac{x_{t-1}^{sm} + INF_t}{1 + \frac{u_t^{pe}}{\phi}}. \tag{B2.2}$$

Rainfall exceeding infiltration capacity is routed, and streamflow is a linear function of the lagged routing state according to:

$$x_t^{rt} = \frac{x_{t-1}^{rt} + INF_t}{1 + k_{rt}}, \tag{B2.3}$$

$$y_t = k_{rt} x_{t-1}^{rt}. \tag{B2.4}$$

where $k_{rt} \sim U[0.01, 0.95]$. This model was sampled $N_s = 1000$ times to estimate $p(\boldsymbol{x}_t|\boldsymbol{u}_t, \boldsymbol{x}_{t-1}, \mathscr{m}_{rout})$ and $p(y_t|\boldsymbol{u}_{1:t}, \mathscr{m}_{rout})$.

3. Saturated and Overland Flow Hypotheses: HyMod [*Boyle*, 2000] partitions routing into overland and subsurface flow. Overland flow is represented by storage states x^{qf1}, x^{qf2}, x^{qf3} with linear coefficients k_{qf1}, k_{qf2}, and k_{qf3}, and saturated flow is represented by a single storage state x^{sf} with coefficient k_{sf}. Coefficient ϱ determines the fraction of excess rainfall that becomes overland flow. The state vector is (x^{sm}, x^{sf}, $x^{qf1}, x^{qf2}, x^{qf3}$), equation (B2.3) is replaced by:

$$x_t^{qf1} = \frac{x_{t-1}^{qf1} + \varrho INF_t}{1 + k_{qf1}}, \tag{B3.1}$$

$$x_t^{qf2} = \frac{x_{t-1}^{qf2} + k_{qf1} x_t^{qf1}}{1 + k_{qf2}}, \tag{B3.2}$$

$$x_t^{qf3} = \frac{x_{t-1}^{qf3} + k_{qf2} x_t^{qf2}}{1 + k_{qf3}}. \tag{B3.3}$$

$$x_t^{sf} = \frac{x_{t-1}^{sf} + (1 - \varrho) INF_t}{1 + k_{sf}}. \tag{B3.4}$$

and equation (B2.4) is replaced by:

$$y_t = k_{sf} x_t^{sf} + k_{qf3} x_t^{qf3}. \tag{B3.5}$$

HyMod parameters and initial states are listed in Table 1. We use $\boldsymbol{v}_t$ (state transition uncertainty in equation (10)) with diagonal covariance matrix $diag(2, 2, 1, 1, 1)$, which represents an additional model

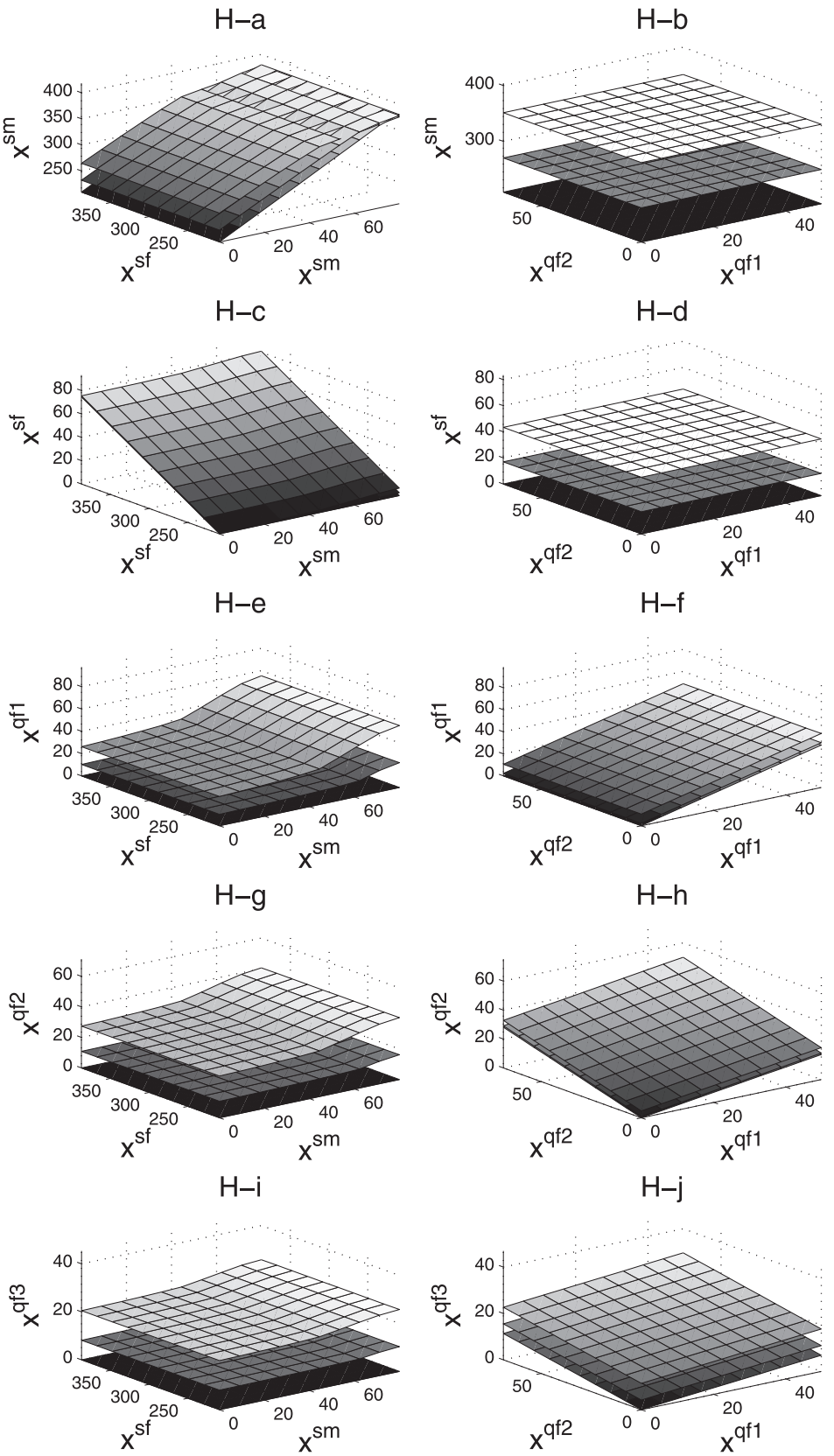

Figure 6. Partial response surfaces of the scaled drift functions of the calibrated HyMod (H-a through H-j) and the SGPR model identified by EM-ID (M-a through M-j). The H and M plots differ due to information introduced from observations during system identification.

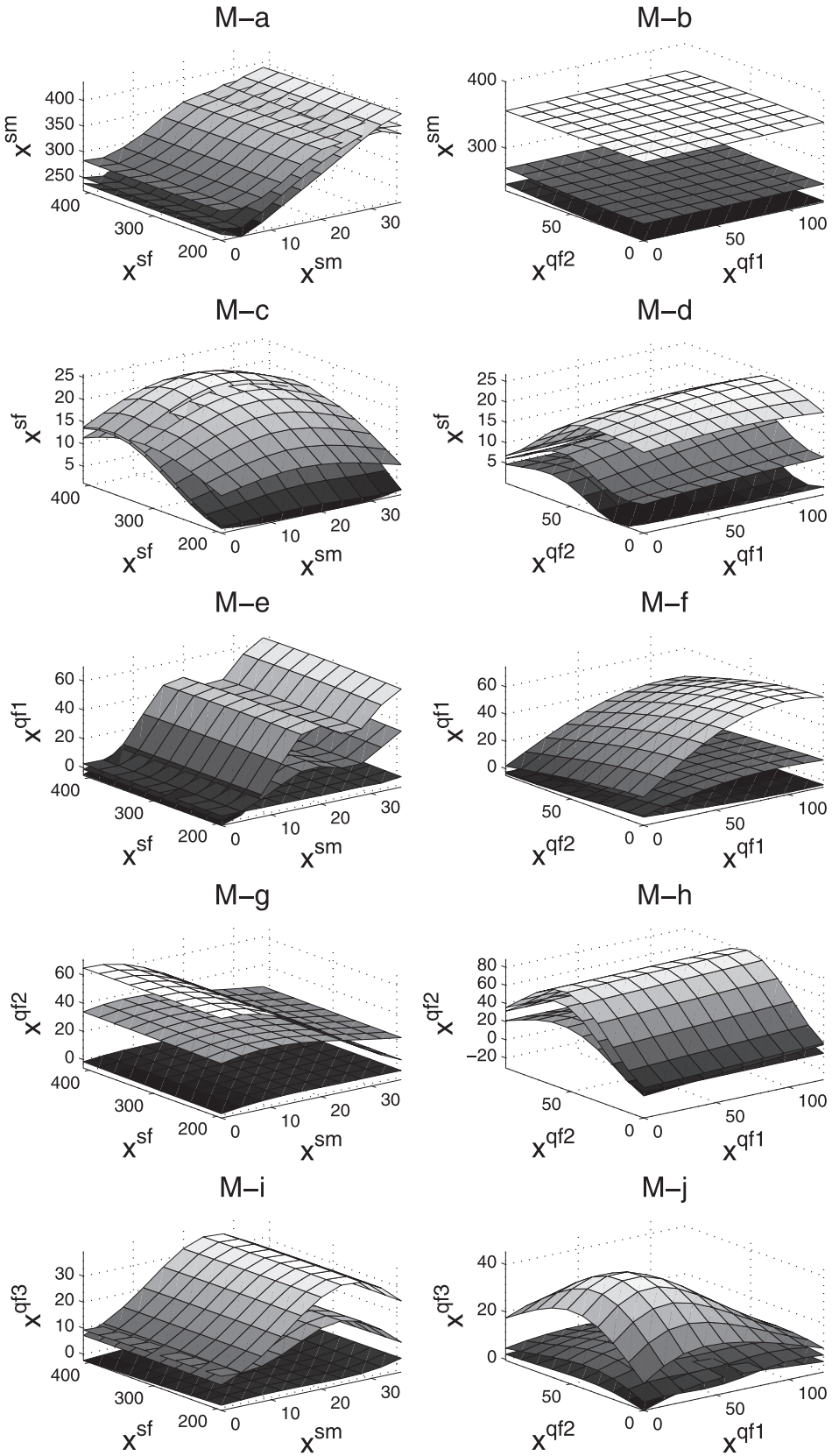

Figure 6. (continued)

assumption. This model was applied to estimate $p(\boldsymbol{x}_t|\boldsymbol{u}_t, \boldsymbol{x}_{t-1}, \mathscr{m}_{HyMod})$ and $p(y_t|\boldsymbol{u}_{1:t}, \mathscr{m}_{HyMod})$ using a *Monte Carlo* ($N_s = 1000$) uniform sample of the parameter space.

4. Parameter Estimation: shuffled complex evolution [*Duan et al.*, 1993] was used to minimize the mean squared error between calibration period streamflow observations and the expected value of HyMod streamflow estimates by optimizing the seven model parameters and five model states over ranges listed in Table 1.
5. System Identification: we use the method of expectation-maximization system identification [*Nearing*, 2013] based on original work by Ghahramani and Roweis [*Ghahramani and Roweis*, 1999]. The calibrated HyMod was emulated by a set of SGPRs [*Snelson and Ghahramani*, 2006], such that every state dimension at each time step was independent of all other state dimensions *conditional on the model inputs* $\boldsymbol{x}_{t-1}$ *and* $\boldsymbol{u}_t$. Therefore, each SGPR emulator consisted of $d_x = 5$ independent GPs, each emulating one $p(x_t^i|\boldsymbol{u}_t, \boldsymbol{x}_{t-1}, \mathscr{m}_{spgp})$ for $1 \leq i \leq d_x$. Observations of streamflow during the calibration period were assimilated using ensemble Kalman smoothing [*Evensen and van Leeuwen*, 2000], and a new SGPR emulator was trained using the updated states. This process was repeated five times to obtain $\mathscr{m}_{SysID}$. The identified model was sampled $N_s = 1000$ times to produce estimates of the state transition and predictand density functions: $p(\boldsymbol{x}_t|\boldsymbol{u}_t, \boldsymbol{x}_{t-1}, \mathscr{m}_{SysID})$ and $p(y_t|\boldsymbol{u}_{1:t}, \mathscr{m}_{SysID})$. A comparison between partial response surfaces of the calibrated HyMod model and the model identified through system identification is given in Figure 6.

Acknowledgments

Data, models, and code may be obtained from the first author on request. The first author acknowledges support from the NASA ROSES Terrestrial Hydrology Program (NNH10ZDA001N-THP). The second author acknowledges support by the Australian Centre of Excellence for Climate System Science (CE110001028). Thank you to Patrick Reed, Tobias Krueger, Steven Weijs, and an anonymous reviewer for their very insightful comments.

References

Abramowitz, G. (2005), Towards a benchmark for land surface models, *Geophys. Res. Lett.*, *32*, L22702, doi:10.1029/2005GL024419.

Abramowitz, G. (2012), Towards a public, standardized, diagnostic benchmarking system for land surface models, *Geosci. Model Dev. Discuss.*, *5*(1), 549–570.

Beven, K. J. (2000), Uniqueness of place and process representations in hydrological modelling, *Hydrol. Earth Syst. Sci. Discuss.*, *4*(2), 203–213.

Beven, K. J. (2001), *Rainfall-Runoff Modelling: The Primer*, John Wiley, Chichester, U. K.

Box, G. E. P., and N. R. Draper (1987), *Empirical Model-Building and Response Surfaces*, John Wiley, Oxford, U. K.

Boyle, D. P. (2000), *Multicriteria Calibration of Hydrologic Models*, Univ. of Ariz., Dep. of Hydrol. and Water Resour., Tucson.

Cartwright, N. (1983), *How the Laws of Physics Lie*, Cambridge Univ. Press, N. Y.

Clark, M. P., and D. Kavetski (2010), Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes, *Water Resour. Res.*, *46*, W10510, doi:10.1029/2009WR008894.

Cover, T. M., and J. A. Thomas (1991), *Elements of Information Theory*, Wiley-Interscience, N. Y.

Cox, R. T. (1946), Probability, frequency and reasonable expectation, *Am. J. Phys.*, *14*, 1–13.

Csiszár, I. (1972), A class of measures of informativity of observation channels, *Period. Math. Hung.*, *2*(1), 191–213.

Doherty, J., and C. T. Simmons (2013), Groundwater modelling in decision support: Reflections on a unified conceptual framework, *Hydrogeol. J.*, *21*(7), 1531–1537.

Dooge, J. C. I. (1986), Looking for hydrologic laws, *Water Resour. Res.*, *22*(9S), 46S–58S.

Draper, D. (1995), Assessment and propagation of model uncertainty, *J. R. Stat. Soc., Ser. B*, *57*(1), 45–97.

Duan, Q. Y., V. K. Gupta, and S. Sorooshian (1993), Shuffled complex evolution approach for effective and efficient global minimization, *J. Optim. Theory Appl.*, *76*(3), 501–521.

Evensen, G., and P. J. van Leeuwen (2000), An ensemble Kalman smoother for nonlinear dynamics, *Mon. Weather Rev.*, *128*(6), 1852–1867.

Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, *Water Resour. Res.*, *47*, W11510, doi:10.1029/2010WR010174.

Frigg, R., and S. Hartmann (2009), Models in science, in *Stanford Encyclopedia of Philosophy* (Fall 2012 Edition), edited by E. N. Zalta, Stanford Univ., Stanford, Calif. [Available at http://plato.stanford.edu/archives/fall2012/entries/models-science/.]

Ghahramani, Z., and S. T. Roweis (1999), Learning nonlinear dynamical systems using an EM algorithm, *Adv. Neural Inf. Processing Syst.*, *11*, 431–437.

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

Guillou, A., and F. Merlevède (2001), Estimation of the asymptotic variance of kernel density estimators for continuous time processes, *J. Multivariate Anal.*, *79*(1), 114–137.

Gupta, H. V., and G. S. Nearing (2014), Debates—the future of hydrological sciences: A (common) path forward? Using models and data to learn: A systems theoretic perspective on the future of hydrological science, *Water Resour. Res.*, *50*, 5351–5359, doi:10.1002/2013WR015096.

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*(18), 3802–3813.

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

Howson, C., and P. Urbach (1989), *Scientific Reasoning: The Bayesian Approach*, Open Court Publ., Chicago, Ill.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, edited by G. L. Bretthorst, Cambridge Univ. Press, N. Y.

Jeffreys, H. (1946), An invariant form for the prior probability in estimation problems, *Proc. R. Soc. London, Ser. A*, *186*(1007), 453–461.

Kolmogorov, A. N. (1956), *Foundations of the Theory of Probability*, Chelsea, N. Y.

Kullback, S., and R. A. Leibler (1951), On information and sufficiency, *Ann. Math. Stat.*, *22*(1), 79–86.

Liu, Y. Q., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, *43*, W07401, doi:10.1029/2006WR005756.

Luo, Y. Q., J. Randerson, G. Abramowitz, C. Bacour, E. Blyth, N. Carvalhais, P. Ciais, D. Dalmonech, J. Fisher, and R. Fisher (2012), A framework of benchmarking land models, *Biogeosci. Discuss.*, *9*(2), 1899–1944.
Metropolis, N. (1987), The beginning of the Monte Carlo method, *Los Alamos Sci.*, *15*(584), 125–130.
Mitchell, H. L., and P. L. Houtekamer (2000), An adaptive ensemble Kalman filter, *Mon. Weather Rev.*, *128*(2), 416–433.
Nearing, G. S. (2013), *Diagnostics and Generalizations for Parametric State Updating*, 210 p., Univ. of Ariz., Tucson.
Nearing, G. S., H. V. Gupta, and W. T. Crow (2013a), Information loss in approximately Bayesian estimation techniques: A comparison of generative and discriminative approaches to estimating agricultural productivity, *J. Hydrol.*, *507*, 163–173.
Nearing, G. S., H. V. Gupta, W. T. Crow, and W. Gong (2013b), An approach to quantifying the efficiency of a Bayesian filter, *Water Resour. Res.*, *49*, 2164–2173, doi:10.1002/wrcr.20177.
Ng, A. Y., and M. I. Jordan (2001), On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, *Adv. Neural Inf. Processing Syst.*, *14*, 605–610.
Paninski, L. (2003), Estimation of entropy and mutual information, *Neural Comput.*, *15*(6), 1191–1253.
Rasmussen, C., and C. Williams (2006), *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Mass.
Rathmanner, S., and M. Hutter (2011), A philosophical treatise of universal induction, *Entropy*, *13*(6), 1076–1136.
Reichle, R. H., W. T. Crow, and C. L. Keppenne (2008), An adaptive ensemble Kalman filter for soil moisture data assimilation, *Water Resour. Res.*, *44*, W03423, doi:10.1029/2007WR006357.
Rosenblueth, A., and N. Wiener (1945), The role of models in science, *Philos. Sci.*, *12*(4), 316–321.
Schaap, M. G., F. J. Leij, and M. T. van Genuchten (2001), ROSETTA: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions, *J. Hydrol.*, *251*(3), 163–176.
Schement, J. R., and B. D. Ruben (1993), *Between Communication and Information*, Trans. Books, New Brunswick, N. J.
Scott, D. W. (2004), Multivariate density estimation and visualization, in *Handbook of Computational Statistics: Concepts and Methods*, edited by J. E. Gentle, W. Haerdle, and Y. Mori, pp. 517–538, Springer, N. Y.
Shannon, C. E. (1948), A mathematical theory of communication, *Bell Syst. Tech. J.*, *27*(3), 379–423.
Snelson, E., and Z. Ghahramani (2006), Sparse Gaussian processes using pseudo-inputs, *Adv. Neural Inf. Processing Syst.*, *18*, 1257–1264.
Ullah, A. (1996), Entropy, divergence and distance measures with econometric applications, *J. Stat. Plann. Inference*, *49*(1), 137–162.
van den Hurk, B., M. Best, P. Dirmeyer, A. Pitman, J. Polcher, and J. Santanello (2011), Acceleration of land surface model development over a decade of GLASS, *Bull. Am. Meteorol. Soc.*, *92*(12), 1593–1600.
Vrugt, J. A., H. V. Gupta, and B. O. Nuallain (2006), Real-time data assimilation for operational ensemble streamflow forecasting, *J. Hydrometeorol.*, *7*(3), 548–565.
Weijs, S. V., G. Schoups, and N. Giesen (2010), Why hydrological predictions should be evaluated using information theory, *Hydrol. Earth Syst. Sci.*, *14*(12), 2545–2558.