Benchmarking and Process Diagnostics of Land Models

GREY S. NEARING
Department of Geological Sciences, University of Alabama, Tuscaloosa, Alabama

BENJAMIN L. RUTTDELL
School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, Arizona

MARTYN P. CLARK
Research Applications Laboratory, NCAR, Boulder, Colorado

BART NIJSSEN
Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington

CHRISTA PETERS-LIDARD
Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland

(Manuscript received 6 November 2017, in final form 29 June 2018)

ABSTRACT

We propose a conceptual and theoretical foundation for information-based model benchmarking and process diagnostics that provides diagnostic insight into model performance and model realism. We benchmark against a bounded estimate of the information contained in model inputs to obtain a bounded estimate of information lost due to model error, and we perform process-level diagnostics by taking differences between modeled versus observed transfer entropy networks. We use this methodology to reanalyze the recent Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER) land model intercomparison project that includes the following models: CABLE, CH-TESSEL, COLA-SSiB, ISBA-SURFEX, JULES, Mosaic, Noah, and ORCHIDEE. We report that these models (i) use only roughly half of the information available from meteorological inputs about observed surface energy fluxes, (ii) do not use all information from meteorological inputs about long-term Budyko-type water balances, (iii) do not capture spatial heterogeneities in surface processes, and (iv) all suffer from similar patterns of process-level structural error. Because the PLUMBER intercomparison project did not report model parameter values, it is impossible to know whether process-level error patterns are due to model structural error or parameter error, although our proposed information-theoretic methodology could distinguish between these two issues if parameter values were reported. We conclude that there is room for significant improvement to the current generation of land models and their parameters. We also suggest two simple guidelines to make future community-wide model evaluation and intercomparison experiments more informative.

1. Introduction

The Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER) benchmarking experiments by Best et al. (2015) showed that some of the world’s most sophisticated operational land models (CABLE, CH-TESSEL, COLA-SSiB, ISBA-SURFEX, JULES, Mosaic, Noah, ORCHIDEE) were outperformed in their ability to simulate short-term surface energy fluxes by simple regressions. Specifically, the PLUMBER experiments used piecewise-linear regressions trained on data from 20 globally distributed FLUXNET meteorological and eddy covariance towers to predict half-hourly latent and sensible heat fluxes given inputs of half-hourly near-surface air temperature, relative humidity, and shortwave radiation.
The regressions were trained and evaluated using a leave-one-out approach, so that training data came from different FLUXNET sites than where the regression was evaluated. These out-of-sample, instantaneous regressions, extrapolated over globally distributed FLUXNET sites, generally yielded improved performance compared to predictions made by the land surface models according to several different metrics (mean bias error, normalized mean error, standard deviation, correlation coefficient), but the land models generally yielded better performance according to metrics that considered full probability distributions (5th and 95th percentiles, third and fourth statistical moments).

This type of result is not unique. Abramowitz et al. (2008) reached similar conclusions after benchmarking CLM and ORCHIDEE against linear and nonlinear regressions (again with no state memory), and Nearing et al. (2016) used benchmarking to show that Noah, SAC-SMA, Mosaic, and VIC each use generally less than half of the information available to them from parameters and boundary conditions. Results like this indicate that there is substantial potential to improve even our most sophisticated land models.

One result lacking in these benchmarking studies was any indication about where exactly information is lost in the different hydrometeorological processes simulated by the land models. There certainly exists some intuition that model benchmarking should help us understand how to improve models, however, to our knowledge, there has not been any methodology proposed that allows us to do this in a systematic way. For example, Luo et al. (2012) outlined a set of desiderata to bridge the gap between benchmarking and diagnostics, but did not offer any methods to implement those desiderata.

The primary purpose of this paper is to suggest how model intercomparison studies might be designed to facilitate more diagnostic results. We do this by first formalizing a quantitative theory of model benchmarking based in information theory and then extending the conceptual underpinnings of that theory to the problem of model diagnostics. We demonstrate the proposed theory and methodology using data from the aforementioned PLUMBER study, and use this demonstration to suggest some specific improvements that might be adopted in the experimental design of future benchmarking and intercomparison studies. In particular, the PLUMBER study did not distinguish between parameter error versus model structural error. This could be remedied by simply reporting the parameter values used for each model and then applying an extension of the information-theoretic method proposed here to disaggregate these two types of error (e.g., Nearing et al. 2016).

We want to be clear that the purpose of this paper is not to argue for using statistical, data-driven, or regression models in place of physically based or process-based models for operational forecasting of terrestrial hydrological systems. We do not want to do this because of the potential for nonstationarity—some type of mechanistic understanding of the system is necessary to predict under changing conditions (Milly et al. 2008). That being said, we cannot ignore the fact that regression models regularly outperform our best biogeoophysical models (e.g., the benchmarking references above). The pertinent question we address is how benchmarking can guide the improvement of complex dynamical systems models that contain explicit representations of interacting biogeoophysical processes.

We also want to be clear that we are not advocating rejecting land models if they do not pass a particular benchmark(s). Just like the PLUMBER land models should not be falsified or rejected because they were outperformed by any of the linear benchmarks, those same land models should not be rejected because they do not outperform a benchmark here. Simplistic model rejection or falsification by either traditional statistical hypothesis testing or other methods (e.g., Beven 2016; Liu et al. 2009) does not seem particularly helpful when dealing with large and complicated models. Instead, we view benchmarking as a tool to help guide a model’s conceptual development and performance improvement, and our goal in this paper is to advance that objective.

The paper proceeds as follows. Section 2 briefly revisits the PLUMBER experiments, which will serve as an example for the rest of this paper. Section 3 proposes a more formal framework for model benchmarking using empirical models such that the benchmarking strategy used by Best et al. (2015) is a special case. This section also outlines a theory for process-level diagnostics that extrapolates that formalization. Section 4 applies several approximations of both the benchmarking and diagnostics theories to the same models and field data used by Best et al. (2015), first to clarify and extend those previous results, and then to look at specific process-level deficiencies in the models. The concluding discussion in section 5 speaks to how we see model benchmarking practice developing in the future, and lists opportunities to improve both models and model intercomparison projects.

2. The PLUMBER experiments

The PLUMBER observational dataset consists of half-hourly measurements of surface boundary conditions and eddy covariance surface energy fluxes from 20 globally distributed FLUXNET towers. In particular, model inputs consisted of half-hourly longwave $R_{lw}$ and shortwave $R_{sw}$ radiation, near-surface air temperature $T_a$, wind speed $W_s$, relative humidity $H_r$, and precipitation $P$. Model outputs evaluated were half-hourly sensible $Q_h$ and
latent heat fluxes $Q_l$. These FLUXNET data were originally collated and used for land model analysis by PALS (Abramowitz 2005, 2012; Abramowitz et al. 2007, 2008; Luo et al. 2012). The reader is referred to Table 1 and Fig. 2 in Best et al. (2015) for a description of these sites.

The PLUMBER performance metrics were (i) bias, (ii) standard deviation ratio (model over observations), (iii) linear correlation coefficient, and (iv) the normalized mean squared error. These were all calculated on time-step (half-hourly) sensible and latent heat fluxes ($Q_h$ and $Q_l$, respectively), as measured by FLUXNET.

The PLUMBER reference values for these metrics were determined by two simple conceptual models and three statistical models. The conceptual benchmarks were (i) Penman–Monteith (i.e., that evapotranspiration $E$ is always equal to potential evapotranspiration $E_P$) and (ii) Manabe Bucket (Manabe 1969), which was the first land model. The three empirical benchmarks were regressions with the following regressor sets: (i) $R_{sw}$; (ii) $R_{sw}$ and $T_a$; and (iii) $R_{sw}$, $T_a$, and $H_r$. The one- and two-variable regressions were linear, and the three-variable regression was piecewise linear over 27 $k$-means clusters. All of the regressions were calibrated using a leave-one-out approach on the 20 FLUXNET sites so that the regressions predicting observations at each site were not calibrated using data from that site. The error metrics from each model and all five benchmarks were ranked and the rank average was reported.

Best et al. (2015) applied the above benchmarks to test the land models that contribute to most of the major weather and climate forecasting centers in the United States, Europe, and Australia. For a complete list and description of these models, the reader is referred to Table 2 in Best et al. (2015). They found that, on average across the 20 sites, the half-hourly $Q_h$ predictions from all of the institutional land models were outperformed by all three regressions, in bias, error, and correlation metrics and that the $Q_l$ predictions from all land model predictions were outperformed by the three-variable piecewise regression in the same variables. Their conclusion was that, since it is possible to develop regression models that outperform biogeophysical land models, the land models do not make full use of the information content of the atmospheric input data—some information that is present in the forcing data goes missing in the model simulations, due to model error.

3. Theory
a. Benchmarking

1) WHAT IS BENCHMARKING?

Best et al. (2015) recognized three approaches that the land modeling community has generally used to understand model consistency, accuracy, and precision. They defined model evaluation as a process in which model outputs are compared against observations, model intercomparison as a process in which multiple models are applied to specific test cases, and model benchmarking as a process in which model outputs are compared against a priori expectations of model performance. Our first purpose here is to formalize the concept of model benchmarking.

A benchmark consists of three distinct elements: (i) a specific observational dataset, (ii) a particular performance metric defined over that dataset, and (iii) a specific criterion or reference value for that performance metric. The point of benchmarking is to define a priori objectives and/or adequacy criteria for model performance (Gupta et al. 2012).

The main challenge when evaluating a complex systems model is to understand (and quantify) how well a model can be expected to perform given that all experiments contain some amount of randomness and uncertainty. That is, even a hypothetically perfect model will not exactly simulate any given set of observation data, simply because of uncertainty in both the measured input data used to drive the model and in the measured response data (Beven 2016). Our goal is therefore to develop a strategy for testing models that leads reliably toward minimizing the component of predictive error associated with the model and to do this in the presence of predictive error due to uncertainty in experimental input and response data. Ideally, the benchmark criteria must be somehow related to data uncertainty, so that we can separate data uncertainty from model uncertainty. Our approach is to have the benchmark reference value estimate the information content of the (imperfect) model input data relative to the (imperfect) measured response data. Gong et al. (2013) laid out a general theory for dealing with this problem that we will adapt here.

2) BENCHMARK CRITERIA

We must choose all three components of a benchmark (observation data, performance metric, and reference value), and again, the major challenge is to quantify model error in the presence of data uncertainty. The idea is that if we knew how accurate a “perfect” model would be over a given set of experimental data, then we could measure how far away any given hypothesis-driven model is from this ideal. We want to quantify information loss due to inevitable model error.

We will rephrase the question slightly: instead of asking how far away we are from a perfect model, we will ask whether there is any measurable information in the experimental input/response data about the relationship between inputs and outputs that is not captured by the
model. This is the question that the PLUMBER experiments addressed.

As Nearing and Gupta (2015) discussed, this approach requires that our benchmark criteria come from a purely data-driven model, simply because our goal is to isolate data uncertainty from model uncertainty. We always have many options for data-driven models, and this choice will affect benchmarking results. For example, the standard Student’s t test can be viewed as benchmarking against a parametric (Gaussian) distribution with a sample variance parameter. Along with the chosen α value, the null hypothesis represents a benchmark criterion. Persistence and climatology are common benchmark criteria used in numerical weather forecasting. Best et al. (2015) used a parametric (piecewise linear) regression as their “null hypothesis” to set benchmark reference values relative to several standard evaluation metrics. Similarly, Gong et al. (2013) used independent component analysis and information-theory metrics to separate model input uncertainty versus model structural and parameter uncertainty, and Nearing et al. (2016) extended that method to set benchmark reference values that separate parameter uncertainty from model structural uncertainty. Given that the desideratum we proposed is a benchmark reference value that somehow quantifies predictability in the presence of data error, the general approach—as in all of the above studies—will be to set the reference values using data-driven models.

One general strategy for developing data-driven models is to rely on superposition theorems. For example, Hornik (1991) showed that an appropriate superposition of sigmoidal bases covers any smooth function. This means that, in principle, a typical single-layer, feed-forward neural network will converge to extracting all information from experimental observations that are about smooth functional relationships between model inputs and outputs. Convergence is difficult to demonstrate in practice, and while this is a very large class of potential relationships, it does not include all possible (e.g., discontinuous or nondifferentiable) relationships. However, it is a very good start to approaching our nonparametric objective, and more sophisticated versions of, perhaps layered, superpositions (e.g., deep learning) can be expected to improve on the first-order approach outlined here.

Importantly the question we want to ask is how much information about the systematic (nonrandom) relationships between controls and responses as we can extract directly from our experimental data.

3) BENCHMARK METRICS

So far, we’ve outlined a general strategy for setting benchmark reference values when the goal is to separate data uncertainty from model uncertainty, which we claim is at the core of the challenge of testing models. Benchmarking against nonparametric regressions allows us to test a hypothesis-driven model in absence of any a priori assumptions about the functional dynamics or process behavior of the system, and to do this in the presence of (arbitrary) experimental randomness.

The next question is about how to choose a benchmark metric. Our objective in this regard is to quantify the ability of the hypothesis-driven model to emulate the nonrandom, or systematic, portion of the relationship between input and response data under the assumption that a nonrandom relationship between experimental input and response data is due to physical processes that we want to capture in our model. Of course, we may use metrics that quantify any particular aspect of relationship in data (e.g., mean additive bias, linear correlation, or Euclidean distance; Tian et al. 2016); however, it is possible to formulate a family of metrics that captures the consistent portion of probabilistic relationships directly. These metrics are called divergences and take the form

\[
D = \int_{x \in \mathcal{X}, y \in \mathcal{Y}} f \left[ \frac{p(x, y)}{p(x)p(y)} \right] dx \, dy = \int_{x \in \mathcal{X}, y \in \mathcal{Y}} f \left[ \frac{p(x|y)}{p(x)} \right] dx \, dy.
\]

In Eq. (1) \(x\) and \(y\) are random variables defined on domains \(\mathcal{X}\) and \(\mathcal{Y}\) respectively, \(p(.)\) is a probability density function, and \(f(.)\) is an unspecified function that we will call an integrating function. The class of metrics \(D\) measure the probabilistic dependence between random variables \(x\) and \(y\) by measuring the divergence from their joint distribution \(p(x, y)\) to a product of their marginals \(p(x)p(y)\). This divergence is always equal to the divergence from a conditional distribution over either of the two random variables—\(p(x|y)\) or \(p(Y|x)\)—to the respective marginal distribution, that is, either \(p(x)\) or \(p(y)\). It is the latter interpretation that is most intuitive for our purpose as we continue.

Nearing and Gupta (2015) showed that certain standard evaluation measures (additive bias, linear correlation, and mean squared error) are special cases of metric \(D\) given particular choices for the integrating function and particular parametric approximations of the various probability distributions.
There is one class of divergence metrics that is especially relevant to benchmarking. This is the class of divergences used by Shannon (1948). In particular, Shannon outlines a set of very basic requirements for a theory of information, and then demonstrates that \( f(v) = -\log(v) \) is the only integrating function that satisfies those conditions. It has been argued that his information theory provides the most appropriate set of metrics to evaluate hydrological models (Weijs et al. 2010), and although certain theoretical arguments for the uniqueness of this choice also exist (e.g., Knuth 2005), we can, in principle choose any integrating function so that \( D \) represents some particular aspect of the systematic (i.e., probabilistically dependent) relationship between model predictions and experimental response data. It is also worth noting that the random variables \( X \) and \( Y \) can be any transform of modeled and/or measured data—for example, we might want to predict Box–Cox transformed streamflow from a hydrology model (e.g., Bulygina and Gupta 2011), or we might want to predict extreme events. In cases like these, we would select relevant random variables.

In our case, we will choose Shannon’s integrating function based on arguments by Nearing and Gupta (2018), and this gives us a useable family of benchmark metrics and benchmark criteria. It is helpful to outline these explicitly. The variable that we want to predict is notated \( X \) (again, the theory does not change depending on how we transform or select the variables), and the predictive distribution from our hypothesis-driven model \( m \) is \( p_m(X|U) \), where \( U \) are the experimental forcing data or boundary conditions. The probability distribution produced by the data-driven model used to set benchmark criteria is \( p_r(X|U) \). The benchmark metric and reference value are therefore of the form:

\[
D_m = \int_{x \in X, u \in U} f \left[ \frac{p_m(x|u)}{p(x)} \right] dx \, du, \quad (2a)
\]

\[
D_r = \int_{x \in X, u \in U} f \left[ \frac{p_r(x|u)}{p(x)} \right] dx \, du. \quad (2b)
\]

Parameter \( p_m(X|U) \) is the modeled probability distribution of the observation data, whereas \( p(X) \) is the empirical distribution of the observed variable \( X \). Typically for a deterministic model, \( p_m(X|U) \) is an empirical distribution of the observed \( X \) data conditional on the model-predicted \( X \) data, so we could think about this as a distribution \( p[X|X = m(U)] \), where \( X \) is the observed data and \( X \) is simulated prediction of \( X \) by model \( m \)—we notate this as \( p_m(X|U) = p[X|m(U)] \). Parameter \( p_r(X|U) = p[X|r(U)] \) is the same thing for the data-driven benchmark model. The probability distribution \( p(X) \) is the marginal distribution over the predicted variable that is derived directly from experimental response data without conditioning on any ancillary information like experimental controls \( U \) or any hypothesis-driven models like \( m \)—this is the same \( p(X) \) distribution used in Eq. (1).

We obtain this simply as an empirical distribution derived from data, and it represents the variability in the observations that our model will (partially) account for. A hypothesis-driven model “passes” the benchmark if \( D_m \leq D_r \), and this means that we have not (at least by this particular test) discovered any information in our experimental data that could be used to improve the overall model performance. It is important to understand that if we were to know \( p_r(X|U) \) exactly, then it is impossible for \( D_m \leq D_r \). However, we cannot ever estimate \( p_r(X|U) \) perfectly from data—the experiment-driven benchmark is defined by our ability to estimate this quantity purely from data, and therefore it is possible for the hypothesis-driven model to exceed this data-driven estimate.

Using Shannon’s integrating function, Eqs. (2a) and (2b) become

\[
I[X; m(U)] = E_{x,u} \left\{ \ln \left[ \frac{p_m(X|U)}{p(X)} \right] \right\}, \quad (3a)
\]

\[
I[X; r(U)] = E_{x,u} \left\{ \ln \left[ \frac{p_r(X|U)}{p(X)} \right] \right\}. \quad (3b)
\]

Parameter \( E_{x,u} \) indicates the expectation over both \( X \) and \( U \), and \( r \) is the data-driven regression model acting on input data \( U \). Parameter \( I[X; m(U)] \) is called the mutual information between measured and modeled data. This differs from \( D_m \) only in that we have specified that \( f(u) = -\log(u) \), and similarly for \( I[X; r(U)] \) and \( D_r \).

In all cases in this paper, we will report information values normalized by the total entropy of \( X \), so that \( I_n \) and \( I_r \) are the following ratios, which for discrete random variables take values between zero and one:

\[
I_n = \frac{E_{x,u} \left\{ \ln \left[ \frac{p_m(X|U)}{p(X)} \right] \right\}}{E_x \left( -\ln[p(X)] \right)}, \quad (3c)
\]

\[
I_r = \frac{E_{x,u} \left\{ \ln \left[ \frac{p_r(X|U)}{p(X)} \right] \right\}}{E_x \left( -\ln[p(X)] \right)}. \quad (3d)
\]

The difference \( I_r - I_m \) is an estimate of the (normalized) information missing from predictions by model \( m \). This difference is bounded in an important way due to the data-processing inequality (Kinney and Atwal 2014). Specifically, if our regression model is not overfitted (we will discuss overfitting more in appendix A), then \( I_r \) cannot overestimate the amount of information available in model input data because
Parameter $I(X; U)$ is the (unknown) “actual” amount of information contained in $U$ about $X$ that we would measure if we had access to the perfect model of the probabilistic relationship between $U$ and $X$—this is explained in somewhat more detail by Nearing and Gupta (2018). The consequence is that $I_r$ always underestimates the amount of information in model input data $U$ so that there is no possibility of a model passing the benchmark while at the same time extracting all information from model input data (this is similar to type I error in hypothesis testing).

Our primary benchmarking question is about whether this difference between the model-derived and data-derived estimates of information in $U$ about $X$ is positive or negative:

$$I_r - I_m < ? > 0.$$ (4)

If $I_m \approx I_r$ then we have not been able to identify any information in the experimental data that is not captured by the hypothesis-driven model $m$. Otherwise, to the extent that $I_r - I_m > 0$, we have discovered that the hypothesis-driven model loses at least some of the information from boundary conditions, and therefore could be improved.

To reiterate, the predicted variable $X$ can be any aspect or transformation of the actual measured response data. For example, we might be concerned with extreme events, and so $X$ might be data points that correspond to some definition of “extreme.” Or we might be concerned with some property of a time series, in which case $X$ might be quantitative results from a frequency analysis. Or perhaps, we might be concerned with long-term water or energy balances, in which case $X$ might be integrations or summations of short-term fluxes. The important thing is that the random variable $X$ must represent the same conceptual quantity in the calculation of both $I_m$ and $I_r$.

The above discussion assumes that our objectives for evaluating model $m$ are related purely to scientific inference. We might instead want to set benchmark criteria relative to some concept of engineering adequacy (Gupta et al. 2012), perhaps related to a specific decision problem that we hope to use a particular model to inform. These sorts of objectives are in the domain of decision theory, and outside the scope of this discussion. Our objective here is to outline a theory of benchmarking that is related to scientific learning and hypothesis testing under arbitrary data uncertainties. We expect that many of the same principles will apply to any arbitration-driven evaluation effort, but we have no explicit argument to this effect.

### 4) BENCHMARK DATA

It is, of course, necessary to choose benchmark data that are as representative as possible of the various situations where we expect our model to be applicable. Again, with reference to the nonstationarity argument above, the purpose of a model is to produce emergent behavior from process descriptions that are themselves stationary, and there may be both differences and similarities in the functional dynamics of a particular set of biogeophysical processes at different locations (different climates and biomes). This means that we may ask two distinct benchmarking questions. The first is about the extent to which a particular hypothesis-driven model is able to represent the behavior of any particular dynamical system, and the second is about the extent to which a particular model is able to represent those aspects of behavior that are consistent across many different dynamical systems. Thus, we advocate a data collection strategy that emphasizes both breadth and depth (Gupta et al. 2014).

In the above discussion, we said that we ultimately want models that are able to reproduce all of the information contained in the experimental data about patterns between observed control (input) and response (output) variables. The way we advocated approaching this is to create $p_r(X|U)$ using a regression trained directly on whatever experimental data are available, and noted that it is essential to ensure that our benchmark regressions are not overfitted.

The PLUMBER benchmarking experiments used a split record approach. They took data from 20 FLUXNET sites and used a leave-one-out approach on the site locations to train their regressions. By comparing their models against benchmarks trained on data from different sites, they effectively asked whether their models could reproduce those portions of the patterns in the FLUXNET input/output data (input data were meteorological forcings, and output data were surface energy fluxes) that were consistent between sites. This means that, effectively, they assessed the ability of their models to reproduce the globally consistent components of surface energy balance, and they found that land models often could not reproduce the linearized component of this globally stationary portion of the signal. This is only one of the two benchmarking questions mentioned in the first paragraph of this subsection. We might also ask about the site-specific aspects of the surface energy balance (as measured by FLUXNET). To do the latter we would train separate data-driven models at each site. We will apply and compare both of these benchmarking approaches in section 4a.
b. Process diagnostics

Our next objective is to obtain some type of insight about any deficiencies in our model(s). To state this another way, we not only want to understand predictive accuracy (or information content of model predictions), but we also want to make sure that our models “get the right answers for the right reasons” (Kirchner 2006). Here we suggest doing this by looking at the dynamics of information transfer within the model, in addition to looking at predictive skill. We outline a general conceptual framework for quantifying interactions and feedbacks in complex systems.

To do this, we will treat dynamical systems and their models as information flow dynamical process networks (DPNs or iDPNs). DPNs have been applied to infer process structure, feedback, and nonstationarity in ecohydrological (Kumar and Ruddell 2010) and ecoclimatological (Ruddell et al. 2016) systems based on field observations. An example illustration of a DPN (for a FLUXNET tower) is given in Fig. 1. Each node in a DPN represents a particular modeled or observed variable, and each edge represents directed influence between variables. Even if all equations in the model are deterministic, each variable is still probabilistic conditional on only a subset of the other variables that determine its value at any point during the simulation.

Our diagnostic approach will be to quantify the influence that each variable has on all others in the observed system and then to see whether our hypothesis-driven model reliably simulates these partially informative relationships. That is, we will measure information transfers within the model and within the ecohydrological system to identify particular relationships in the LSM that are either overconstrained or underconstrained. If a model simulates a smaller-than-observed transfer of information from one variable to another, then that modeled pathway or modeled relationship is underconstrained, and vice versa.

To quantify the influence that one variable, say $y_1$, has on another variable, say $y_2$, in a dynamic (time evolving) Markovian system, we integrate over the expected effect of probabilistically conditioning $y_2$ at time $t + \tau$ on the value of $y_1$ at time $t$ given all of the variables in the model other than $y_1$ (notate these as $y_{-1}$) at time $t$. It is important to understand that even in deterministic models each variable is probabilistic conditional on only a subset of other variables. Schreiber (2000) proposed a computationally feasible Markovian approximation of this metric called transfer entropy:

$$I(y_2^{t+\tau}; y_1^t | y_2^t) = \int p(y_2^t) \int p(y_2^{t+\tau}, y_1^t | y_2^t) \ln \left( \frac{p(y_2^{t+\tau} | y_1^t, y_2^t)}{p(y_2^{t+\tau} | y_2^t)} \right) dy_2^{t+\tau} dy_1^t dy_2^t.$$  \hspace{1cm} (6)

The probability distributions are derived empirically from either modeled or observed data, and this metric can be applied at any spatiotemporal scale. If sufficient observation data are available, this metric can be directly calculated over observations at or across scales (e.g., Paluš 2014), and directly over model runs at similar spatiotemporal scales. In that case, any differences in the individual transfer metrics represent
nonisomorphism in the process-level behavior of the model.

One major challenge with any type of process-level diagnostics is that there is typically insufficient observation data about each individual simulated variable. It is possible that this could be dealt with by using data assimilation (Bulygina and Gupta 2011), but this is out of the scope of the current paper. Our PLUMBER reanalysis in section 4 follows Ruddell and Kumar (2009) in assuming the system is identified by readily available FLUXNET data variables rather than trying to identify the correct system boundary and resolution in this paper.

To calculate information flow diagnostic benchmarks, a custom computer code was written in MATLAB. This code is available at https://github.com/greyNearing/plumber_diagnostics.git. The code was validated by comparison with a community code, ProcessNetwork version 1.5 (Ruddell 2016), and found to be in good agreement.

4. Reanalysis of the PLUMBER experiments

We will now return to the PLUMBER experiments and apply the theory outlined in section 3. Before presenting the details of these experiments in sections 4a–c, we will briefly outline the primary results.

Figure 2 reports out-of-sample benchmarks that measure the overall ability of the PLUMBER land models to utilize information in meteorological inputs for estimating half-hourly surface energy balance components. Figure 3 shows results from a highly simplified version of the theory outlined in section 3 to benchmark long-term water balances, rather than the short-term energy balances shown in Fig. 2. The Budyko-type analysis in Fig. 3 is from a simplified version of benchmarking that is useful when insufficient data are available to implement the full information-theoretic procedure. Figure 4 summarizes the primary process-diagnostics results, and compares these against the benchmarking results. Figure 5 looks specifically at process-level differences between different land models for partitioning net radiation in particular. Figure 6 shows all of the process-level diagnostic results for all of the land models.

a. Local versus global PLUMBER benchmarks

The goal of benchmarking, as outlined in section 3a(1), is to set a priori expectations on model performance, given error in experimental input and response data, that are free of any conceptual understanding about the modeled system. So we first trained a set of nonparametric regressions that estimate the information content of the FLUXNET meteorological input data about the FLUXNET response data, which here are half-hourly surface latent and sensible heat fluxes. Then we asked whether the hypothesis-driven land models (CABLE, CH-TESSEL, COLA-SSiB, ISBA-SURFEX, JULES, Mosaic, Noah, ORCHIDEE) each provide more or less information about this relationship than we were able to extract directly from observation data [Eq. (5)].

The exact method that we used for constructing our benchmarks is given in appendix A. It is, of course, an...
approximation of the philosophy outlined in section 3, but it is an approach that is coherent in the presence of arbitrary uncertainties, and which relies on explicit and well-understood approximation. Each sigmoidal regression used lagged meteorological data to account for the conservation-related memory of the land surface, and we used an empirical analysis to ensure that our benchmark regressions were not overfitted.

The first set of benchmarks that we considered were analogous to the PLUMBER benchmarks in that they were trained at out-of-sample FLUXNET sites. We trained single-layer, feed-forward artificial neural networks on each permutation of 19 out of 20 FLUXNET sites using time-lagged input data, and then used each of these 20 out-of-sample regressions to make surface energy flux predictions at the single excluded site. These benchmarks tell us something about the ability of the land models to represent only the globally stationary components of the forcing-response relationships at the land surface.

We also want process-based land models to represent site-specific components of the forcing-response relationships. In practice, spatial heterogeneity is represented in these models by parameter maps that reflect characteristics of soils, vegetation, topography, etc. PLUMBER did not report the parameter values used by each model, so we cannot directly assess information loss due specifically to model parameters, as was done by Nearing et al. (2016). However, we can assess the ability of the parameterized model (Gupta and Nearing 2014) to capture spatial heterogeneity in system processes by performing site-specific benchmarking. In this case, benchmark regressions are trained at each site individually using a split-record approach. Here we used a $k$-fold with $k = 5$ validation approach, where five regressions were trained at each site using four-fifths of the available training data without resampling, so that we obtained out-of-sample regression predictions over the whole data record at each site.

We will refer to the first set of benchmarks (trained using a leave-one-out procedure over the 20 FLUXNET sites) as global benchmarks, and the second set (trained using $k$-fold validation at each site) as local benchmarks. In the hypothetical situation where our regression models were perfect at extracting information from the observation data, the global benchmarks could be beaten by even an imperfect parameterized hypothesis-driven model, whereas the hypothetically perfect local benchmarks could not be beaten by even a hypothetically perfect hypothesis-driven model due to the data processing inequality. Appendix B gives the technical details of how the normalized information metrics [Eqs. (3c) and (3d)] were calculated, and Fig. 2 compares the information content of the PLUMBER models against these benchmarks.
To understand the primary message of the results in Fig. 2, it is important to understand two things. First, the metrics we report are the ratio of discrete information in the model or benchmark to the total discrete entropy of the FLUXNET surface flux data (either $Q_e$ or $Q_h$). These metrics therefore exist on the range of zero to one, with a value of zero indicating that there is no information about the observations in the benchmark or model data, and a value of one indicating that there is a one-to-one relationship between the benchmark or modeled data and the observations. Second, the information and entropy metrics were calculated over all sites at once—they were not calculated at each site and then averaged. That is, regressions were trained out of sample, and then used to predict $Q_e$ or $Q_h$ over the excluded data (either leave-one-out site data or $k$-fold validation data), but then all of these out-of-sample regression predictions were used to calculate a single (global or local) value of $I_r$ from Eq. (3d). Similarly, for each land model, all $Q_e$ or $Q_h$ data from all 20 FLUXNET sites were used to calculate $I_m$ from Eq. (3c). This requires that we use the same number of data points from each FLUXNET site, so that no site is overweighted compared to another in the calculated metrics. The reason for using data from all sites to calculate a single metric is that this allows systematic model biases at individual sites (that differ across sites) to be reflected in the calculation of the benchmarking metrics—taking site-specific information metrics would effectively ignore site-specific systematic biases in either the data or the models.

Figure 2 shows that the land models evaluated by the PLUMBER experiments provide less than about half of the total information in FLUXNET observation data about the relationship between near-surface atmospheric forcings and surface energy fluxes. This benchmark fully accounts for all observation error, unless there is systematic measurement error that is stationary across all sites. These results also show that the land models generally produce about as much information as the global benchmarks. This means that the models are not capturing spatial process heterogeneity that exists between different ecoclimates. Again, it is impossible to tell whether this is because of parameter error or model structural error because the PLUMBER parameter values were not reported, but these two effects could be segregated using information-theoretic benchmarking by training a third set of regressions that act directly on model parameters (Nearing et al. 2016).
We can reframe the results in Fig. 2 in terms of an uncertainty decomposition (Gong et al. 2013). By subtracting the information content of the model simulations from the information content of the benchmark regressions, we get a (bounded) estimate of the amount of information lost due to model error $U_m$. Similarly, by subtracting the information from the regressions from total entropy of the observations, we get a bounded estimate of the uncertainty due to errors and incompleteness in input data $U_d$:

$$U_m \geq \frac{I[X; r(U)] - I(X; U)}{H(X)}, \quad (7a)$$

$$U_d \leq 1 - \frac{I[X; r(U)]}{H(X)}. \quad (7b)$$

Parameter $H(X) = E_x(-\ln[p(X)])$ is the total entropy of the benchmark evaluation data [$Q_e$, $Q_h$, or net ecosystem exchange (NEE)]—this is the same quantity as the denominators in Eqs. (3c) and (3d). Nearing et al. (2016) reported these as fractions of total missing information, which is obtained here by normalizing both $U_m$ and $U_d$ to their sum: $[H(X) - I(X; U)]/H(X) = U_m + U_d$.

We did the same here, so that $U_m$ and $U_d$ always sum to one and found that uncertainty due to missing information in the input data $U_d$ was in the range of 77%–83% for $Q_e$ and $Q_h$, and $U_d$ was in the range of 85%–95% for NEE, depending on the model. Likewise, the fraction of information lost due to model error $U_m$ was in the range of 17%–23% for $Q_e$ and $Q_h$, and in the range of 5%–15% for NEE, depending on the model. Nearing et al. (2016, their Table 2) reported similar fractions for a different set of land models (including the Noah and Mosaic models used here) and forcing data—they found a $U_d$ fraction for $Q_e$ of about 70%.

Finally, it is worth directly comparing the local versus global benchmarks. This difference provides us some intuition about the observed systems themselves. In particular, from Fig. 2, we see that the about half of the information in FLUXNET data comes from globally stationary patterns between forcing and response data, while about half comes from site-specific relationships. The objective of a science-informed model is (generally) to capture both types of information. The difference between these two quantities (scaled by the local benchmark) is something we will call heterogeneity-induced information loss fraction (HILF):

$$\text{HILF} = \frac{I[X; r_{\text{local}}(U)] - I(X; r_{\text{global}}(U))}{I[X; r_{\text{local}}(U)]}. \quad (8)$$

For $Q_e$, $Q_h$, and NEE, the HILF was about 44%, 39%, and 28%, respectively, at these FLUXNET sites.

b. Benchmarking the PLUMBER water balance

In addition to simulating the surface energy balance, we are interested in understanding the ability of land models to simulate long-term water balances. We will use this benchmarking objective to motivate an example that demonstrates how to approximate the theory outlined in section 3a in cases where we may only have limited data.

One purpose of land models is to predict spatial variations in the partitioning of precipitation into evapotranspiration and runoff. Long-term relationships between water and energy controls in a watershed are, to a first-order approximation, described by a Budyko-type relationship. For example, the Turc–Pike model (Gerrits et al. 2009) relates the dryness index, which is long-term total potential evapotranspiration over long-term cumulative precipitation $E_p/P$ with the long-term evaporative fraction, which is total evaporation over precipitation $E/P$:

$$E/P = \left[1 + \left(\frac{E_p}{P}\right)^{-\nu}\right]^{-1/\nu}. \quad (9)$$

The parameter that we used for this study was $\nu = 2$. The Turc–Pike curve suggests that $E$ should be close to $E_p$ at sites with low dryness indices (i.e., energy limited sites, where precipitation is much greater than $E_p$) and that $E$ should be close to precipitation at high values of the dryness index (i.e., water limited sites, where $E_p$ is much greater than precipitation).
Given that hydrological theory can generally explain at least certain departures from the Turc–Pike curve by the dominant controls on the water balance, including intraseasonal-to-interannual variability in precipitation, and the controls of vegetation, topography, and soils on runoff generation and transpiration (Eagleson 1978; Milly 1994; Zhang et al. 1999), it is reasonable to expect that sophisticated simulation models should also be able to predict departures from the Turc–Pike curve.

To assess this, we trained a regression model to act on model inputs, which here were taken to be long-term averages of net radiation (sum of short- and longwave radiation), air temperature, and cumulative precipitation, to predict departures from climatological evaporative fraction calculated according to Eq. (9). In this case we only have observation data from 20 FLUXNET sites—each of which yields a single value for the dryness index and a single value for the evaporative fraction. Similarly, calculating the evaporative fraction from each set of land model output represents a transform of the simulated data into a new random variable $X$, as discussed in section 3a(3). Nineteen data points are not enough to train robust nonparametric regressions, so we had to use a simpler benchmark reference value—in this case, a linear regression.

Figure 3 illustrates the mean absolute error (MAE) of evaporative fraction across FLUXNET sites for the various models and benchmarks. The spreads in this figure were calculated by calculating error statistics over 1000 separate random samples drawn with replacement from the 20 FLUXNET sites. Both the Turc–Pike curve itself and the regression on departures from this curve outperform all of the physically based land models. There are not enough data to develop reliable empirical probability distributions for Eqs. (3), and so we cannot estimate the fraction of information missing from the models relative to this benchmark criterion. Instead, we calculated the MAE over all models and benchmarks relative to the observed evaporative fractions (Fig. 3) and used a nonparametric bootstrap hypothesis test on the MAEs across the 20 sites. The out-of-sample regression benchmark here had lower MAE on the evaporative fraction ($\mu = 0.140$) than all of the physically based land models (all $\mu > 0.147$), but the differences were only significant ($\alpha = 0.10$) against Noah 2.7.1.
Ruddell and Kumar (2009) identified step of the FLUXNET observation data. Second, time step of the model—here we are limited by the time state and not prior lagged states. So, in the context of perform time-step integration using only the current hydrology models are Markovian in the sense that they step of the reported FLUXNET data and second, most variables at the FLUXNET station they analyzed.

The important takeaway from these plots is that there is no identifiable model-specific pattern in the behavior of differences between measured versus modeled transfer entropies—each individual PLUMBER model has different error structures at different FLUXNET sites. However, there are readily identifiable patterns in the transfer entropy differences of all models at each individual FLUXNET site. This behavior indicates that models are not correctly capturing the intersite process-level variability that is ultimately measured by the HILF [Eq. (8)].

To formalize this, the strength of site-related groupings versus model-related groupings was quantified by calculating the fractional reduction in mean distance to center of cluster due to clustering the results in each subplot in Fig. 4 by model versus clustering by site. Cluster centers for each of the model-specific and site-specific clusters illustrated in Fig. 4 were calculated using a Euclidean distance in the $x$-$y$ space defined by (i) the difference between modeled and measured transfer entropies, and (ii) the missing information of each model for each prognostic variable ($Q_c$, $Q_h$, and NEE). Then, average distance to center of cluster was calculated for each subplot in Fig. 4. Clustering these results by model reduced dispersion by about 6% for $Q_c$ and $Q_h$ and about 20% for NEE, whereas clustering by site reduced dispersion by between 42% and 45% for $Q_c$ and $Q_h$ and about 30% for NEE—this is shown in Fig. 5. This preference for site-specific clustering indicates that all of the PLUMBER land models are generally wrong for the same reasons—all models share similar patterns of process-level deficiencies at each individual site, although those patterns of process-level deficiencies (that are similar between all models) change from site to site. Conditional on the assumption that there is no (or minimal) systematic error in the observations at each site, this collection of models would not constitute a useful ensemble to represent epistemic uncertainty, because these models are process-biased in similar ways.

The $R_{sw} \rightarrow Q_c$, $R_{sw} \rightarrow Q_h$, and $R_{sw} \rightarrow$ NEE transfer entropy pathways showed the most model-specific reduction in clustering dispersion (16%, 10%, and 31% for $Q_c$, $Q_h$, and NEE, respectively). The model-specific results for the two energy partitioning pathways are explored further in Fig. 6. In this case, there is clear delineation between model groups. Specifically, the Noah and JULES models show consistent bias; they underestimate the role of net radiation on determining both $Q_c$ and $Q_h$ at all sites. The other models—CABLE, COLA-SSiB, ISBA-SURFEX, etc.—do not have this consistent bias.

Understanding the causes of this model-specific bias will require more data than was collected as part of the PLUMBER study. In particular, it will be necessary in the future for each of the modeling groups that contribute to the model intercomparison project to report the parameter values. It is impossible from the data provided to the PLUMBER study to know whether the energy partitioning biases in Noah and JULES is
due to poor parameter values or to specific problems with model physics. All of these models are based on Penman-style evapotranspiration equations; both Noah and Mosaic used climatological vegetation. Stomatal conductance parameterizations in the Noah, Mosaic, and CH-TESSEL models use a Jarvis-type stomatal resistance, while most other models including JULES, CABLE, and ISBA-SURFEX use a Ball–Berry–Leuning model for stomatal conductance. Given the similarity in parameterization schemes between process-biased versus process-unbiased models, and the dissimilarity between Noah and JULES, it is at least possible that the biases in the Noah and JULES models are due to specific parameter values, rather than to model physics.

d. Summary of results

The results outlined above related to PLUMBER models and data can be summarized as follows:

- **Figure 2** shows that the PLUMBER land models use less than half of the information about half-hourly surface energy fluxes that is available to them from the meteorological forcing data as measured by FLUXNET. This figure also shows that about one-half of the information in FLUXNET forcing-response data is due to locally specific patterns of behavior that are not shared across sites.

- **Figure 3** shows that the PLUMBER models underutilize information in meteorological forcing data about long-term (multiyear) water balances. However, due to a lack of data at this time scale, we can only estimate missing information using a parametric (RMSE) statistic.

- **Figures 4 and 5** show that all of the PLUMBER models have quantitatively similar process-level information transfer biases, but that these biases differ between different ecoclimates.

- **Figure 6** shows that some of the PLUMBER models (especially the Noah models) have consistent bias across all sites in the shortwave energy partitioning relationships, while the other models do not.

5. Conclusions and outlook

Benchmarking is often used to assess model performance against baseline criteria—for example, Best et al. (2015) used linearized benchmark criteria under the perspective that complex land simulation models should capture at least some of the nonlinearity of hydrometeorological systems. Similarly, benchmarks related to persistence and climatology are often used to assess numerical weather prediction models and forecasts, because these are simple criteria that a competent model should be able to beat.

We take a different perspective here, one that is developed around a more general philosophy. Our proposal is that model evaluation—and hypothesis testing in general—should generally consider model performance for a given set of experimental data in the context of the inherent information content of that data. Even a perfect model cannot predict more accurately than is allowed using noisy input and response data. We are also not interested in using benchmarking to reject complex systems models, but rather to guide their improvement.

Our benchmarking approach is therefore designed to do two things: 1) help assess model performance in the context of experimental observations with unknown and arbitrary data uncertainty and 2) connect model evaluation and benchmarking with process-level model diagnostics. Under the proposed theory, benchmarking allows us to partition uncertainty between data error versus model error, and separating these two things allows us to judge model performance independent of any error in the observation data. This is probably the intuition behind the benchmarking approach used by Best et al. (2015) and Abramowitz (2005), and our contribution is to take steps toward formalizing this intuition. We further propose that an information-theoretic perspective on model benchmarking suggests at least an informal link between benchmarking and process diagnostics, and we suspect that it will likely be possible to formalize this relationship by deriving an aggregating relationship to relate process-specific transfer entropy metrics within the model directly to the holistic mutual information metrics used in section 3.

Empirical results from the application of our benchmarking theory to the PLUMBER models and data support the primary conclusions by Best et al. (2015) that modern land models do not take full advantage of the information content of input data. Our specific experimental conclusions about the PLUMBER models are summarized in section 4d. We agree with the Best et al. (2015) conclusions and expand on their findings by showing how results like theirs can be exploited to guide model development.

Our methods and results also provide some insight into how we might improve the design of future model intercomparison experiments. The Best et al. (2015) PLUMBER experiments did not collect sufficient data about the various models and simulations to enable highly detailed process-level model diagnostics. The first suggestion from our analysis would be that modelers should report their parameter values, so that benchmarking can be used to segregate the effects of model parameters from model structure (e.g., Nearing et al. 2016). However, if
the objective is to use model benchmarking or intercomparison projects to inform continued model development, then we would ideally want full posterior parameter distributions over model parameters, so that parameter effects could be marginalized out in the integrations used to calculate information metrics. This would allow us to make quantitative statements directly about model processes in the presence of parameter error, and would require that future benchmarking studies adopt (for example) a Markov chain Monte Carlo approach to dealing with model parameters.

A second suggestion for future model intercomparison studies is that it would be very helpful if there were a succinct description of the process-level differences between different land surface models. This would be an invaluable resource for qualitatively tracing diagnostic signatures to differences between common process parameterizations used in modern land models. One way to approach this might be to host a collaborative workshop where developers of the various land models worked together to produce a systematized report of the process-level differences between land models. Tools like the Structure for Unifying Multiple Modeling Alternatives (SUMMA; Clark et al. 2015) are processor-flexible in the sense that they have the ability to implement several of the more common process-specific parameterizations used commonly in land models, and it might be worthwhile to focus a concerted effort on collecting functional equivalents of the major process descriptions used in the various PLUMBER models in SUMMA, so that individual processes can be tested in a common computational framework. Both of these efforts would require significant buy-in from the developers of at least several of the modern land models. To derive significant scientific value from model intercomparison experiments, it will be necessary to start by outlining a formal theory and methodology of model benchmarking and diagnostics and then to design an intercomparison protocol around that theory.

Acknowledgments. This work was partially supported by the National Science Foundation under Grant EF-1241960 through Northern Arizona University, and by the NASA Earth Science Technology Office Advanced Information Systems Technology program through NCAR, NASA, and the University of Washington. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies. All code and data used for this project are publicly available at https://github.com/greyNearing/plumber_diagnostics.git.

APPENDIX A

Constructing PLUMBER Benchmark Criteria

Our PLUMBER benchmark criteria were set by single-layer, feed-forward neural networks that projected time-lagged meteorological inputs onto FLUXNET eddy covariance estimates of half-hourly surface latent and sensible heat fluxes. The regression inputs were hour of day, wind speed $W$, 2-m air temperature $T_a$, incident short- and longwave radiation $R_{sw}$ and $R_{lw}$, relative humidity $H_r$, and time-step cumulative precipitation $P$. Nearing et al. (2016) looked at the number of time lags necessary to capture the majority of the signal in the lagged time series of similar model inputs, and we used enough lagged data to capture the period that they recommended—here, four half-hourly lagged data of each type, two daily-aggregated (mean or sum, as appropriate) lagged data of each type, and an additional two weekly- and two monthly-aggregated lagged data of each type. This resulted in a total of $D = 1 + (6 \times 10) = 61$ inputs to each neural network, and we used $(D/2) + 2$ hidden nodes with sigmoidal activation functions, trained with back-propagation against a mean squared error objective function. Using a mean squared error objective function means that the neural network represents the mean of a Gaussian prediction with variance that is stationary over the input domain. This regression model could be improved by using a kernel density estimator, as was done by Nearing et al. (2016); however, Nearing and Gupta (2018) point out that Eq. (4) is bounded against type I error regardless of the choice of regression model. We know exactly the logical environment from which our results are derived.

To ensure that the regressions were not overfitted, we pulled 600000 data points—30000 from each of the 20 FLUXNET sites—and then randomly sampled a fraction of those data points to train various neural networks. Performance statistics were calculated over the training data and also over the remaining portion of the 600000 samples that were not used for training. We repeated this procedure 10 times for a number of different training sample sizes, and Fig. A1 reports the mean and two standard deviations of the normalized mutual information [Eq. (3d)] over these resampling tests as a function of the number of training data. The important takeaway from this figure is that the performance of these regressions over training and test data converge to within a few percent of the same information content at around 11000 samples. This does not guarantee that such regressions account for all of the information about functional relationships contained in these data, but it
Figure A1. Convergence of neural network training and test performance statistics [normalized mutual information from Eq. (3d)] as a function of the number of training samples. Training sets were chosen randomly from a total of 600,000 data points sampled randomly from 20 FLUXNET sites, and test sample statistics were calculated on the remainder that were not used for training. Error bars show two standard deviations calculated from 10 random samples of training data at each sample size. Both the mutual information ratio and the more traditional correlation coefficient are stable at around 10,000–15,000 training samples, indicating that this number of training data is likely sufficient to avoid overfitting the benchmark regressions.

APPENDIX B

Calculating Information Metrics

The information ratios that we report are the ratio of maximum likelihood estimators (Paninski 2003) of discrete mutual information between the model predictions and the evaluation data over the discrete entropy of the evaluation data. We use discrete (i.e., histogram) estimators to ensure that these ratios are bounded both below (by zero) and above (by one). This means that we have to discretize the model outputs and the evaluation data, and the precision of this discretization will affect the mutual information and entropy statistics, as well as their ratios. For the remainder of this essay we will report information ratios calculated using discretizations with precision equal to 1% of the total observed range of each type of observation data ($Q_e$, $Q_h$, and NEE) over all 20 FLUXNET sites. We use the same discretization for model outputs as for observation data.

The steps to calculating the information metrics are as follows. First, we define a bin width as a fraction of the range of the observed data (either $Q_e$, $Q_h$, or NEE). We then discretize the observation data and the model or benchmark outputs at this histogram bin resolution. We then form an empirical joint histogram from the record of model/benchmark and observation data pairs. We first calculated the individual and joint entropies by summing over the histogram bins according to

$$H_o = \frac{1}{N_o} \sum_{i=1}^{N_o} \ln\left(\sum_{j=1}^{N_m} K_{ij}\right),$$

$$H_m = \frac{1}{N_m} \sum_{j=1}^{N_m} \ln\left(\sum_{i=1}^{N_o} K_{ij}\right),$$

$$H_{o,m} = \frac{1}{N_o N_m} \sum_{i=1}^{N_o} \sum_{j=1}^{N_m} K_{ij} \ln(K_{ij}).$$

Here $K_{ij}$ is an individual histogram bin and $N_m$ and $N_o$ are the number of bins in the model/benchmark and observation space. We used the convention that

does mean that our models are not overestimating information in data due to overfitting.
\[ 0 \ln(0) = 0. \] The mutual information statistics were calculated according to

\[ I_m = H_o + H_m - H_{o,m}, \]

and these were normalized by the total entropy of the observations so that the statistics we reported were of the form \( 0 \leq I_m / H_o \leq 1. \) Mutual information statistics for the benchmark were calculated analogously as above, and we reported \( I_o / H_o. \)

REFERENCES


