Water Resources Research®

RESEARCH ARTICLE
10.1029/2021WR031164

Special Section:
Advancing process representation in hydrologic models: Integrating new concepts, knowledge, and data

Source Relationships and Model Structures Determine Information Flow Paths in Ecohydrologic Models

Allison E. Goodwell[1] and **Maoya Bassiouni**[2,3]

[1]Department of Civil Engineering, University of Colorado Denver, Denver, CO, USA, [2]Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden, [3]Department Environmental Sciences, Policy, and Management, University of California, Berkeley, Berkeley, CA, USA

Key Points:

- Source dependency and model structure together dictate a range of possible information flow pathways
- Linear and nonlinear source dependencies can mask or reveal incorrect process representations
- Comparing source dependencies with model functional performance can help develop more robust model structures

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

A. E. Goodwell,
allison.goodwell@ucdenver.edu

Citation:

Goodwell, A. E., & Bassiouni, M. (2022). Source relationships and model structures determine information flow paths in ecohydrologic models. *Water Resources Research*, *58*, e2021WR031164. https://doi.org/10.1029/2021WR031164

Received 6 SEP 2021
Accepted 23 JUN 2022

Abstract

In a complex ecohydrologic system, vegetation and soil variables combine to dictate heat fluxes, and these fluxes may vary depending on the extent to which drivers are linearly or nonlinearly interrelated. From a modeling and causality perspective, uncertainty, sensitivity, and performance measures all relate to how information from different sources "flows" through a model to produce a target, or output. We address how model structure, broadly defined as a mapping from inputs to an output, combines with source dependencies to produce a range of information flow pathways from sources to a target. We apply information decomposition, which partitions reductions in uncertainty into synergistic, redundant, and unique information types, to a range of model cases. Toy models show that model structure and source dependencies both restrict the types of interactions that can arise between sources and targets. Regressions based on weather data illustrate how different model structures vary in their sensitivity to source dependencies, thus affecting predictive and functional performance. Finally, we compare the Surface Flux Equilibrium theory, a land-surface model, and neural networks in estimating the Bowen ratio and find that models trade off information types particularly when sources have the highest and lowest dependencies. Overall, this study extends an information theory-based model evaluation framework to incorporate the influence of source dependency on information pathways. This could be applied to explore behavioral ranges for both machine learning and process-based models, and guide model development by highlighting model deficiencies based on information flow pathways that would not be apparent based on existing measures.

Plain Language Summary

Two sources walk into a model together
Who may or may not have previously met
They dictate an output, one with the other
But maybe those outputs are somewhat set.

By "set" we mean that the sources' relation
Really determines the possible states,
Or ways that the sources can give *information*
To the output that they create.

We consider some cases, just for practice
From binary, math, and ecohydrology samples
We find patterns along the source relationship axis
With these simple to complex examples.

If you have a model and think that it's hot
This source dependency framework could show
How your model listens (or maybe does not)
To sources it could possibly know

So models and sources are jointly to blame
For missing observed information flow,
But this little poem is admittedly lame,
And you should just read the paper below.

1. Introduction

The detection of causal interactions is relevant to both understand and model complex ecohydrologic systems. From an observational standpoint, forcing and feedback interactions can explain different ecosystem reactions to perturbations (Goodwell et al., 2018), characterize pairwise dependencies of different strengths and timescales (Ruddell & Kumar, 2009; Sendrowski et al., 2018), or untangle dynamics of large-scale complex systems (Runge et al., 2019). The land-atmosphere interface, watersheds, or the canopy-root-soil continuum are generally non-interventional systems, in that it is impossible or difficult to perturb these types of systems and isolate a response. Instead, the notion of causality must be derived from statistical methods rather than experiments (van Leeuwen et al., 2021). From a modeling perspective, interventions to explore causal interactions can involve manipulating model structures, parameters, or inputs and observing model behavior. Beyond exploring the impact of these modeled interventions on predictions of the target variable, they must also be explored in the context of dependencies between inputs or source variables in complex environmental systems. On one hand, we can leverage relationships that emerge at larger spatial or temporal scales to develop parsimonious ecohydrological models. On the other hand, we need to diagnose the impact of dependencies between variables on model behavior to ensure that our understanding and predictions are robust in unseen conditions.

Measuring casual associations between model inputs, outputs, and observed data can be used to assess the "functional performance" of a model (Bassiouni & Vico, 2021; Ruddell et al., 2019), or the model's representation of interactions relative to observations. This aspect of model evaluation addresses the question of getting the "right answers for the right reasons" (Kirchner, 2006), and can complement or contrast with "predictive performance" measures that compare observations directly to model outputs. For example, an ecohydrologic model that overestimates soil water evaporation but underestimates canopy transpiration may have high predictive performance regarding the total evapotranspiration, but we would find poor predictive performance when we consider canopy transpiration as the target variable. This mismatch ultimately limits our ability to improve understanding of water-vegetation interactions. Misrepresentation of processes within the model would also become clear with a functional performance measure. A model with high predictive performance but low functional performance may be more likely to break down under previously unseen input or forcing conditions. Meanwhile, a model with both high predictive and functional performance is a more robust tool for scientific inquiry.

In both modeling and observational frameworks, causal interactions can occur between multiple sets of sources and targets, rather than only pairwise between a single source and target pair. A simple example is a model in which the target, or output, is the sum of two die rolls. While the knowledge of an individual die provides partial information about the sum, the knowledge of both sources together is needed to completely reduce the uncertainty about the sum. This paper focuses on these multivariate source-target causal interactions in the context of how models use information from source variables that may have varying interdependencies. These source dependencies could be linear correlations, or nonlinear, threshold-based, system state-dependent, or other types of dependencies. This addresses an aspect of the combination of process parameterization equations (Gharari et al., 2021) with input certainty hypotheses. When forcing is applied to a model with a given process parameterization, the full relevant, or physically possible, functional range of that model may not be accessed with the provided forcing data, leading to an incomplete view of how the model can use available information.

We use information theory-based measures to characterize "information flow" from multiple source variables to a single target variable. Information theory is based on Shannon Entropy (Shannon, 1948), a measure of uncertainty of a random variable. Mutual information is the reduction in uncertainty of one variable given another, and can characterize information flows between sources and targets. A suite of information theory techniques based on mutual information provides pathways to causal analysis at different levels (Goodwell et al., 2020). For example, transfer entropy (Schreiber, 2000) is a measure of conditional mutual information and has been used to characterize Granger, or non-interventional, causality in observations and models. In hydrologic modeling research, it has been used to validate and diagnose missing process connections in a delta model (Sendrowski et al., 2018), evaluate a multi-hypothesis ecohydrological modeling framework (Bennett et al., 2019), select time aggregations and lags toward machine learning applications (Tennant et al., 2020), and finally characterize the functional performance of a multilayer canopy model (Ruddell et al., 2019). Transfer entropy has also been applied to study ecohydrological and climate systems as process networks of pairwise connections between observed variables (Ruddell & Kumar, 2009; Ruddell et al., 2016; Sendrowski & Passalacqua, 2017). However, a transfer entropy-based analysis only highlights pairwise causal connections and does not address the feature of joint or simultaneous forcing from multiple sources.

Meanwhile, information decomposition (Williams & Beer, 2010) quantifies causal interactions in which two sources jointly provide information to a target variable, which could be an observation or a model output. In this framework, the total information provided to a target by the two sources is partitioned into four components: redundancy, or overlapping information that both sources provide individually, synergy, which is joint information that is only obtained from the knowledge of both sources together, and two unique components, which is information only provided by a particular source. These information components reveal different aspects of the source-target relationship in observations or model frameworks. In viewing the system in terms of these four information components, we can look more deeply into the nature of model functional behaviors (Bassiouni & Vico, 2021). Understanding the structure of model errors in terms of how information flow is partitioned can provide important insights into the extent to which and why certain model structures are more or less desirable or robust in terms of performance. For example, different arithmetic operations and the parameters that weigh the influence of input variables in commonly used ecohydrological equations can produce varying information types (redundant, synergistic, and unique). A more comprehensive quantification of these effects and how they may vary with source dependencies can further improve our understanding of model functional behavior.

We consider model structure as a mapping from input variables to an output, which is dictated by the parameters and formulas within the model that are relevant to a particular input-output combination. Here, we use the term “source” to refer to any variable that may influence a “target”, or a variable of interest. Sources and targets tend to correspond to model inputs and outputs but could more generally involve variables that are not considered in a particular model. The source dependency, or the connectivity between two sources, is characterized as shared information between the sources. For example, highly correlated sources have a high source dependency, and completely independent sources have a source dependency that approaches zero. Both model structure and source dependencies can influence how information from inputs flows through the model to produce an output. As a simple illustration, consider an observation Z that is forced by a single source X within some range of system states. Within a different range of system states, Z is forced by both X and another source, Y, together. If a simplified model simulates Z as only a function of X, the amount of “missing” information about Z somewhat depends the relationship between the sources, X and Y. If these are perfectly correlated copies, the model will accurately predict Z even though it does not properly account for Y as a source. However, if X and Y are independent, the model cannot incorporate the additional information that Y could have provided.

In some cases, source dependencies could reflect multicollinearities that are known to increase errors in parameter estimates, especially in linear regression models, and fitting robust model coefficients requires selecting more adequate model structures (Bassiouni et al., 2016). Source dependencies in terms of their mutual information, however, go beyond multicollinearities and incorporate other, particularly nonlinear, ways in which sources can provide information jointly to a modeled or observed target. We also consider sources as generically applicable to a wide range of model frameworks, beyond linear regressions, such as logical operations, process-based models, and machine learning models, which may or may not typically account for the issue of multicollinearity. In a process-based model where the sources are physically meaningful time-series variables, it is not typical to evaluate effects of multicollinearity, or any type of source dependency, during model development. On the contrary, dependencies between variables are often leveraged to represent unobserved processes empirically. For example, remote sensing algorithms for evapotranspiration leverage the relation between atmospheric relative humidity and surface conductance to water vapor to down-regulate potential evaporation in water-limited conditions (Baldocchi et al., 2022; Fisher et al., 2020). Equilibrium assumptions under which surface and atmospheric states are coupled at daily timescales are also used to successfully estimate evapotranspiration from standard weather station data (McColl et al., 2019; Rigden & Salvucci, 2015). Evaluating these assumptions along the axis of source dependencies is therefore critical to understand the conditions for which these generalized relations can be employed and select among alternative parameterizations.

Here, we use information decomposition to study how both model structure and source dependencies inherently restrict the ways in which a set of sources can provide information about a target, either relative to a different model or the natural system dynamics. In other words, we explore the influence of both model structure and source dependency on the causal interactions, or joint dependencies, between sources and targets—a dimension of model functional performance that has yet remained unexplored by previous contributions (Bassiouni & Vico, 2021; Ruddell et al., 2019). Specifically, we hypothesize that source dependency, which can be linear or nonlinear, provides a relevant basis with which to compare between different model types and between different forcing scenarios for a single model. With a series of simple synthetic models and ecohydrologic and weather examples, we address the following questions:

- *How does the influence of source dependency on information flows change based on a given model structure?* In other words, we explore the implications of model functional form on the ways in which a pair of sources can possibly inform a target. This addresses the "functional range" in terms of the possible causal associations that can be observed given source variables with different dependencies between them.
- *To what extent are models able to reproduce different types of interactions that we observe from data, given a certain source dependency range*? In other words, we explore whether the modeled information flows and the influence on source dependency on these flows match observations. This addresses the functional performance of a model relative to the range of observed types of causal interactions that occur under different forcing conditions.

This paper is organized as follows: We present the information theory-based methods and case study setups in Section 2 (Figure 1). In Section 3, we discuss information partitioning results for synthetic binary models. In

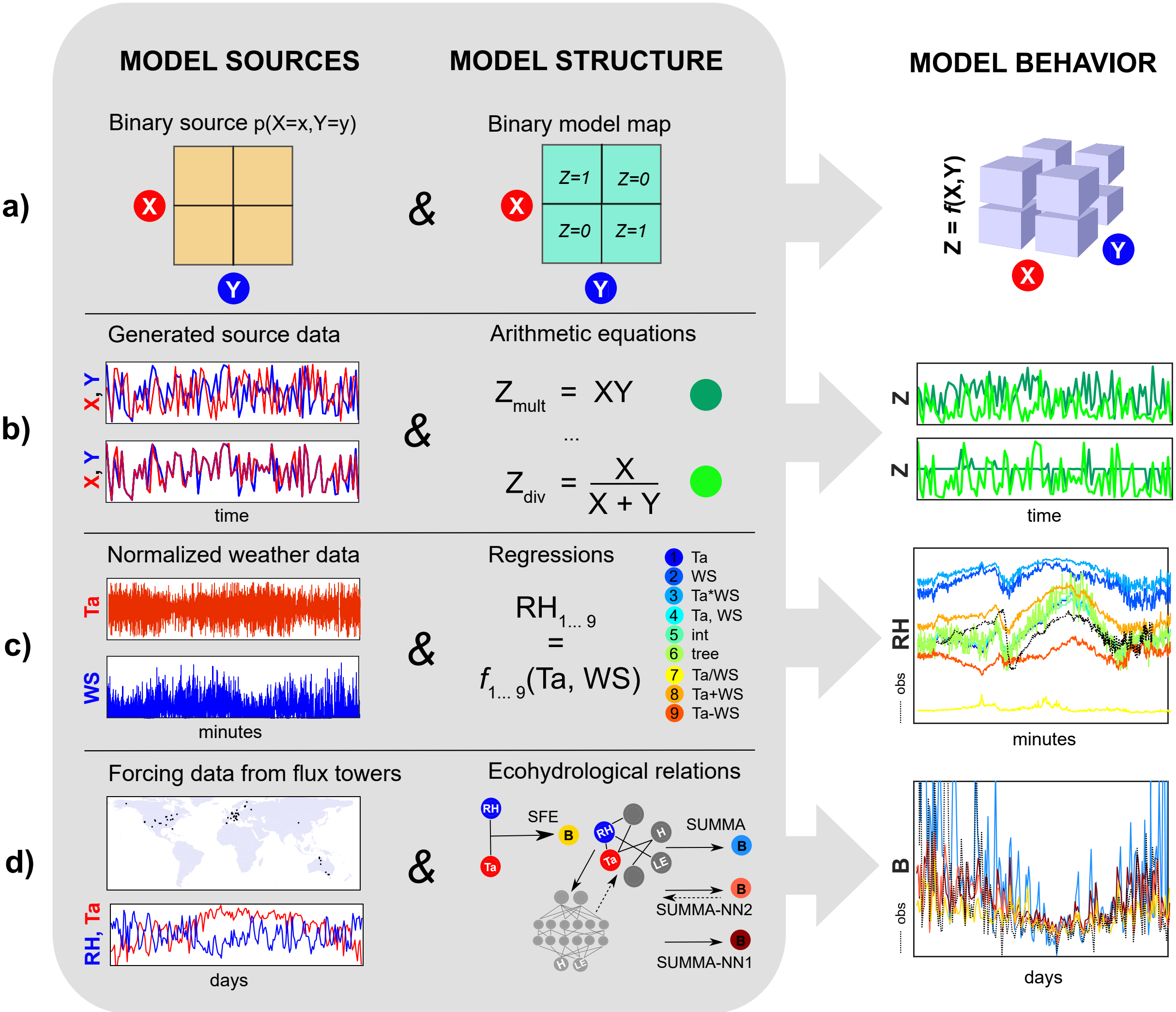

Figure 1. Illustration of model cases, where we show how different source distributions along with model structures dictate model behavior or information flows within a model. (a) Synthetic models based on source distributions and model mappings, (b) time series generated with simple mathematical equations, (c) regressions on air temperature (*Ta*) and wind speed (*WS*) to estimate relative humidity (*RH*), and (d) a comparison between simple (SFE) and more complex models (SUMMA variants) to estimate the Bowen ratio (b) from flux tower data.

Section 4, we discuss information partitioning for several simple two-source models based on arithmetic operations forced with generated time series data. In Section 5, we discuss a suite of models trained on a weather station data set to predict relative humidity given air temperature and wind speed at fast timescales. In Section 6, we compare flux tower observations of the Bowen ratio (B) to simple versus more complex ecohydrologic models. Section 7 provides a discussion, and Section 8 is a conclusion.

2. Methods

2.1. Source Dependency and Information Partitioning Measures

Information theory is based on Shannon Entropy, $H(X) = -\sum_i p(x_i)\log_2 p(x_i)$ (*bits*), where the summation is overall possible states x_i. $H(X)$ is a measure of uncertainty of a discrete random variable X, or the missing information that would lead to its full predictability. In this paper, we focus on reductions in uncertainty, or gains in information, in the form of mutual information between two or more variables.

We define source dependency between two sources, X and Y, as the mutual information between them, as follows:

$$I(X;Y) = \sum p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right), \tag{1}$$

where the summation is overall possible states of $X = x$ and $Y = y$. This has units of *bits* and a maximum bound of $\log_2(N)$, where N is the number of bins used to estimate the *pdf*s. It is upper bounded by the minimum of the source entropies ($H(X)$ and $H(Y)$). Here, we use $I(X;Y)$ to categorize different observed or synthetic data samples into different ranges of source dependency.

To determine information flow pathways in observations and models, we are concerned with the total information that two sources, X and Y, provide to a target variable Z:

$$I(X,Y;Z) = \sum p(x,y,z)\log_2\left(\frac{p(x,y,z)}{p(x,y)p(z)}\right). \tag{2}$$

In an information decomposition (Williams & Beer, 2010), this quantity can be partitioned into four components as follows:

$$I(X,Y;Z) = S_{X,Y} + R_{X,Y} + U_{X|Y} + U_{Y|X}. \tag{3}$$

In the above equation, $R_{X,Y}$ is redundant information, or information that both sources provide in overlap, and $S_{X,Y}$ is synergistic information, or information that is provided only when both sources are known together. $U_{X|Y}$ and $U_{Y|X}$ terms indicate unique information, which is provided individually by one source when considered along with a particular second source. This partitioning of information can be placed in the context of other information theory measures. For example, mutual information is composed of one unique component and one redundant component of information, that is, $I(X;Z) = U_{X|Y} + R_{X,Y}$. Conditional mutual information, which includes transfer entropy as a special case, contains a unique and a synergistic component, that is, $I(X;Z|Y) = U_{X|Y} + S_{X,Y}$. Finally, the interaction information, which is symmetric between all three variables, is equivalent to $S_{X,Y} - R_{X,Y}$, such that a positive or negative interaction information indicates whether synergy or redundancy is dominant. To simplify notation hereafter, we omit subscripts such that $S_{X,Y} = S$ and $R_{X,Y} = R$ given a particular definition of sources and target. We similarly simplify unique information components to $U_{X|Y} = U_X$ and $U_{Y|X} = U_Y$.

While information decomposition is an intuitively useful concept, information theory does not provide an equation to directly perform this partitioning into the four information components. Several methods have been proposed to compute redundancy, synergy, or uniqueness, and one proposed redundancy measure (Barrett, 2015; Williams & Beer, 2010) defines redundancy as the mutual information that the weakest source provides to the target, forcing one unique component to equal zero. However, this is actually a maximum bound for redundancy, and we apply a "rescaled" redundancy measure (Goodwell & Kumar, 2017) in which redundancy is scaled between minimum and maximum bounds that are defined by information theory. The maximum bound is the minimum mutual information that either source provides to the target, $R_{\max} = \min[I(X;Z), I(Y;Z)]$. The minimum bound is zero for cases where the interaction information is positive, or $S - R > 0$, that is, $I(X,Y;Z) > I(X;Z) + I(Y;Z)$. Otherwise, if $S - R < 0$, the minimum bound for redundancy is the negative interaction information, in order for synergy to be non-negative. This leads to a definition of the minimum R as $R_{\min} = \max[0, I(X;Z) + I(Y;Z) - I(X,Y;Z)]$. We then scale redundancy between these bounds based on the normalized mutual information between the source variables, $I_s = \frac{I(X;Y)}{\min[H(X),H(Y)]}$, as follows:

$$R = R_{\min} + I_s\,(R_{\max} - R_{\min})\,. \tag{4}$$

In a linear context, this definition causes highly correlated sources to be maximally redundant with each other, while independent sources are minimally redundant. It has been noted that this scaling assumption may not always hold in cases where a threshold in behavior exists (Weijs et al., 2018), but we maintain this formulation for R in the absence of a redundancy measure with universally desirable properties. A definition for redundancy enables the direct computation of the other information decomposition components, $S_{x,y}$, U_x, and U_y. We normalize components by dividing each by the total mutual information $I(X, Y; Z)$. With this normalization, all information components add up to 1, and a given component indicates the fraction of reduced uncertainty in Z that can be attributed to that information type. Several illustrative case studies of this information partitioning method can be found in (Goodwell & Kumar, 2017), including the addition of two die rolls, addition of two Gaussian random variables, and dependencies between weather variables. Here we expand upon these examples to explore the information decomposition of a wider range of model structures and specifically explore interactions given different source dependencies.

2.2. Model Performance Measures

We build upon previous metrics for model predictive and functional performance based on information theory (Bassiouni & Vico, 2021; Ruddell et al., 2019) by investigating the variability of these metrics along a source dependency axis (Equation 1). We define the predictive performance for a given model as follows, where Z is the observed target variable, Z_{mod} is the modeled output, and $I(Z; Z_{\text{mod}})$ is their mutual information:

$$A_p = 1 - \frac{I\,(Z; Z_{mod})}{H(Z)} \tag{5}$$

A_p represents the information about the observed target variable that is missing from the model output and ranges from 0, for a perfectly accurate model, to 1, if Z_{mod} does not reduce any uncertainty in Z. In other words, A_p is actually a measure of "predictive mismatch", and a low score is desirable. While mutual information is not an ideal measure of predictive performance since it excludes an aspect of predictive "reliability" (Weijs et al., 2010), here we use it as a simple metric that is comparable to measures of source dependency and functional performance measures as described further. Other information theory-based measures, such as the Kullback-Leibler divergence, could be used to more fully evaluate predictive performance, but have the drawback that model predictions need to be in probabilistic form, which we do not consider here.

Functional performance is separated into different components related to information partitioning and total information measures. We consider the difference in an information flow measure for modeled versus observed data as an indicator of functional performance. We first consider the breakdown of information into S, R, and U components, where performance can be assessed as the relative difference between observed and modeled information flows as $A_{f,S}$, A_{f,U_X}, A_{f,U_Y}, and $A_{f,R}$. For example:

$$A_{f,S} = S_{mod} - S_{obs}. \tag{6}$$

The same applies for R, U_X, and U_Y information components, and a positive value indicates that the model overestimates a particular component. As with A_p, all A_f measures represent a degree of "functional mismatch" in that values close to zero indicate similarities between models and observations. We define the overall "partitioning" functional performance as the sum of the absolute values of the four individual functional performance measures as follows:

$$A_{f,Ipart} = |A_{f,S}| + |A_{f,U_X}| + |A_{f,U_Y}| + |A_{f,R}| \tag{7}$$

A low $A_{f,Ipart}$ indicates better model performance in terms of information partitioning components. This measure ranges from 0, for a model that exactly reproduces the information components as observed, to two, for a model that completely trades one type of information for another, or a combination of other information types.

For example, if the observed system shows that $U_X = 1$ (all information is unique to X), but a model system estimates $S = 1$ (that all information is synergistic between the two sources), this leads to $A_{f,S} = 1$, $A_{f,U_x} = -1$, and $A_{f,I\text{part}} = 2$. We compare these functional and predictive performance metrics for the different regression and physically based models and for source dependencies between that range from weak to strong.

We also consider the total mutual information that the two sources, X and Y, provide to the measured versus modeled target:

$$A_{f,Itot} = \frac{I\left(X,Y;Z_{mod}\right)}{H\left(Z_{mod}\right)} - \frac{I(X,Y;Z)}{H(Z)} \tag{8}$$

If $A_{f,I\text{tot}} > 0$, this indicates that the model overestimates the strength of the information flow. This "overly deterministic" case is typical, as found in (Ruddell et al., 2019) and (Bassiouni & Vico, 2021), and most models are expected to follow this pattern. Particularly for a simple model with only a few inputs, we expect those inputs to determine the modeled output relatively more strongly than those same variables influence an observation. However, we find that $A_{f,I\text{tot}}$ is not a reliable performance measure, as a more complex model that better replicates the relationships between sources and targets in terms of all other IT-based measures may have a relatively high $A_{f,I\text{tot}}$. In other words, a model may be very overly deterministic but still capture relevant interaction types better than a more "stochastic" model. With this, it is possible to have an accurate $A_{f,I\text{tot}}$ at the expense of inaccurate partitioning of the individual information components (Bassiouni & Vico, 2021). Additionally, for a two-source model with no other noise, and assuming perfect estimation of the 3D *pdf*, the term $\frac{I\left(X,Y;Z_{mod}\right)}{H\left(Z_{mod}\right)}$ is equal to 1, since X and Y are the only sources of information to Z_{mod}. In this case, $A_{f,I\text{tot}}$ only reflects properties of the observed system.

2.3. Model Cases

We apply information decomposition to several models of increasing complexity to explore the implications of model structure and source dependency on modeled interactions. All synthetic model cases involve two sources, X and Y, and the knowledge of the two sources completely reduces uncertainty of the model target variable, Z (Figures 1a and 1b). In other words, we know that $I(X, Y; Z) = H(Z)$ because only one or both of the two source variables are used as model inputs. The other two cases are based on observations from weather stations and eddy covariance flux towers (Figures 1c and 1d) and do not have this property because no two sources fully inform an observed target. In each case, we explore how different models forced by sources with different dependencies capture information components. Here we provide a more detailed description of these four categories of case studies.

2.3.1. Synthetic Binary Model Maps and Source Distributions

We first generate very simple models with binary inputs and outputs. These models are defined by creating a probability distribution $p(x, y, z)$ directly. We generate a range of $p(x, y)$ source distributions, and a range of model mappings that assign each (x, y) pair to a value of z, as follows for binary ($N = 2$) cases (Figure 1a):

- We create a range of possible source distributions $p(x, y)$. We iteratively set $p(0, 0)$ to values between 0 and 1, $p(0, 1)$ to values between 0 and $1 - p(0, 0)$, $p(1, 0)$ to values between 0 and $1 - p(0, 0) - p(0, 1)$, and $p(1, 1) = 1 - p(0, 0) - p(0, 1) - p(1, 0)$. This results in a number of source distributions that range from independent (e.g., all four $p(x, y)$ terms are equal to 0.25) to highly dependent.
- We create a range of "model maps", $(x, y) \rightarrow z$, in which each (x, y) combination is assigned to a particular value of z. Each model map corresponds to a logical operation performed on one or both sources (e.g., AND, OR, or XOR) with or without the NOT operation ($\neg$) as a preprocessing step.
- Each source distribution $p(x, y)$ is paired with each model map $(x, y) \rightarrow z$ to obtain a $p(x, y, z)$ for that source and model combination.

These synthetic "models" for which we directly obtain distributions are then analyzed in terms of information decomposition components relative to source dependencies (Figure 1a). For the binary case, we generate over 12,000 $p(x, y)$ source distributions and consider 16 (2^4) possible model maps. However, half of these 16 maps are mirror images of each other and two are trivial cases in which $z = 0$ or $z = 1$ regardless of the input. Based on this, we retain seven model maps and apply all source distributions to each.

2.3.2. Arithmetic Models Based on Generated Time Series

For a second model case, we generate time series data for X and Y, where X is a uniform random or normally distributed variable, and Y is a linear or nonlinear function of X to varying extents. For three cases, we generate time series for dependent sources as follows:

$$Y_t = \epsilon f_i\,(X_t) + (1 - \epsilon) w_t \tag{9}$$

where w_t is a uniform random variable, and ϵ ranges from 0 to 1 and indicates a level of strength of the dependency between sources X and Y. The term $f_i(X_t)$ indicates a function through which the sources are related. For a case that induces a linear correlation, we implement $f_1(X_t) = X_t$ as well as a case in which X and Y are linearly correlated Gaussian variables, with correlations that range from 0 to 1. For cases that induce nonlinear source dependencies, we implement $f_2(X_t) = 4X_t(1 - X_t)$ (a logistic map) and $f_3(X_t) = 0.5\cos(4\pi X_t) + 0.5$ (a function with multiple curves). These four scenarios result in two cases where sources are linearly dependent ($f_1(X_t)$ and Gaussian variables) and two cases where sources are nonlinearly dependent and have near-zero linear correlations ($f_2(X_t)$ and $f_3(X_t)$). We apply these source pairs to several simple models as follows:

$$Z_{add} = X + Y \tag{10}$$

$$Z_{mult} = X * Y \tag{11}$$

$$Z_{diff} = |X - Y| \tag{12}$$

$$Z_{div} = \frac{X}{X + Y} \tag{13}$$

For Z_{div}, we use $\frac{X}{X+Y}$ instead of $\frac{X}{Y}$, because that case leads to very small shared information between sources and targets. In other words, for $Z = \frac{X}{Y}$, there is a large decrease in entropy. Later in the weather station regression models, we see an example of this information loss due to the division of two source variables.

For these models, we use fixed binning to estimate the 3D *pdf* $p(x, y, z)$ from the X and Y time series data and model output, Z, using $n = 20{,}000$ data points and $N = 50$ bins (SI: IT measures) and then compute information theory measures and information partitioning components. In all synthetic or generated model cases, since $H(Z) = I(X, Y; Z)$, we assume that the total information value is statistically significant. We also note there is no "error" or associated performance measures in these cases as there is no observational component. In other words, these cases purely focus on how sources with different dependency structures are filtered by a model to produce different types of information.

2.3.3. Data-Driven Models Based on Weather Data

To compare model performance of many different model structures to observations, we apply information decomposition to observed 1-min weather data, where we consider relative humidity (*RH*) to be an output of air temperature (*Ta*) and wind speed (*WS*) (Figure 1c). Data were collected at a restored prairie site in central Illinois over a 2-year period from May 2014–2016 (Goodwell & Kumar, 2017).

We compare observed information components for nine regression models that use different combinations of the two sources. These constitute linear regression models with between one and three inputs that are combinations of *Ta* and *WS*, in addition to a tree model (Table 1). We train the models with randomly selected 50,000 data points and use 10% hold-out validation. To test the robustness of these regressions, we repeat the model training with different random data samples of varying size up to 100,000 points, and find

Table 1
Weather Station Models Based on 1-Min Time Series Data Over a 2-Year Period

Model	Model type	Sources	Abbreviation	RMSE training
$RH_{\mathrm{mod},1}$	Linear regression	*Ta*	lin *Ta*	0.076
$RH_{\mathrm{mod},2}$	Linear regression	*WS*	lin *WS*	0.097
$RH_{\mathrm{mod},3}$	Linear regression	*Ta* * *WS*	lin *Ta* * *WS*	0.095
$RH_{\mathrm{mod},4}$	Linear regression	*Ta*, *WS*	lin *Ta*, *WS*	0.075
$RH_{\mathrm{mod},5}$	Linear regression	*Ta*, *WS*, *Ta* * *WS*	lin int	0.075
$RH_{\mathrm{mod},6}$	Tree	*Ta*, *WS*	tree	0.068
$RH_{\mathrm{mod},7}$	Linear regression	*Ta*/*WS*	lin *Ta*/*WS*	0.098
$RH_{\mathrm{mod},8}$	Linear regression	*Ta* + *WS*	lin *Ta* + *WS*	0.087
$RH_{\mathrm{mod},9}$	Linear regression	*Ta* − *WS*	lin *Ta* − *WS*	0.095

Note. Several additional machine learning models are described in the Supporting Information S1.

very similar results. In general, we are not concerned with parameter optimization or overfitting and instead focus broadly on how different model structures use the same two source variables to estimate a target. We apply models to the entire 2-year (approximately 1 million data points) time period. We assess the predictive and functional performance of each model based on the information theoretical metrics defined above. Specifically, we assess their ability to match *RH* observations (A_p, Equation 5) in addition to their ability to replicate different types of interactions between *Ta* and *WS* that influence *RH* (A_f measures, Equations 6–8).

We filter the diurnal and seasonal cycles from the observed and modeled data using a Butterworth filter, and remove outliers. We also normalize each variable to lie between 0 and 1 as follows: $X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$. To compare model behavior under different source dependency conditions, we partition the 2 years of time series data into 5-day moving time windows (approximately 7,200 data points per window), and compute source dependencies, model predictive performance, and functional performances for each model, for each time window. In other words, the model structure (coefficients) remains constant for each model, but the source dependencies vary between 5-day windows. We test for statistical significance of total mutual information $I(Ta, WS; RH)$, and conditional mutual information terms, $I(Ta; RH|WS)$ and $I(WS; RH|Ta)$ with a shuffled surrogates method (SI: Statistical Significance).

As in the cases based on generated time series data, we compute probability distributions, $p(Ta, WS, RH)$ for each time window for this data set, and also the modeled *pdfs* $p(Ta, WS, RH_{\text{mod},m})$, for $m = 1\ldots9$. We then compare information partitioning components for observed and model cases. We also train several machine learning models, and results and discussion regarding these are included in the Supplementary Information (Figure S2 in Supporting Information S1). These models performed most similarly to $RH_{\text{mod},4}$, the general linear regression model, with slight differences in information partitioning components. However, we focus here on the range of "worse but interesting" models that lead to more highly varying results in terms of functional range, rather than only the most high-performing models.

2.3.4. Ecohydrologic Models Forced With Weather Data

For a physically based model case, we compare information decomposition along the source dependency axis using a simple equation with no free parameters to a calibrated land-surface model, as well as variants of the land-surface model coupled with a neural network. We assess the performance of these models of varying complexities relative to flux tower observations (Figure 1d) in terms of their ability to estimate the daily relation between two sources, *Ta* and *RH*, to the ratio of sensible heat flux (*H*) to latent heat flux (*LE*) or the Bowen ratio (*B*, Brutsaert, 1982). We focus the analysis on this single functional relation between *Ta*, *RH*, and *B* to provide a direct linkage to the previous model setups (Figures 1a–1c) and aim to demonstrate the value and utility of our proposed model evaluation framework in identifying specific weaknesses among the selected model structures.

For the simple physical equation, we estimate *B* according to the principle of surface flux equilibrium (SFE) (McColl & Rigden, 2020; McColl et al., 2019), which requires no land-surface information and only the two source variables *Ta* and *RH* as follows:

$$B \approx \frac{R_v C_p T a^2}{\lambda^2 q^*(T a) R H}, \tag{14}$$

where $\lambda = 2.5 \times 10^6$ (J kg^{-1}) is the latent heat of vapourization of water, $C_p = 1{,}005$ (J kg^{-1} K^{-1}) is the specific heat capacity of air at constant pressure, $R_v = 461.5$ (J kg^{-1} K^{-1}) is the gas constant for water vapor, and q^* is saturated specific humidity as a function of *Ta*. The SFE model thus estimates *B* with the assumption that only two sources (*Ta* and *RH*) encode the necessary information that drives variability in *H* and *LE*.

For the land-surface model, we derive *B* using existing simulations of *H* and *LE* from the Structure for Unifying Multiple Modeling Alternatives (SUMMA, Clark et al., 2015), a modular process-based hydrologic modeling framework, here, set up to mimic the Noah land surface model (see details in Bennett & Nijssen, 2021a). SUMMA estimates *B* by solving a complex series of energy and water balance equations, requiring numerous model inputs beyond *Ta* and *RH*, including land-surface parameters and additional weather forcing variables. We analyze three SUMMA variants: a standalone model, which is only based on physical equations (SUMMA-SA); a one-way coupled neural network, which estimates *LE* and *H* using SUMMA inputs (SUMMA-NN1); and a two-way coupled neural network, which estimates *H* and *LE* using both SUMMA inputs and modeled soil moisture states (SUMMA-NN2).

We use *Ta*, *RH*, *LE*, and *H* observations from 60 selected FLUXNET2015 sites for which SUMMA is locally calibrated and simulations driven by the flux tower weather data are available (Bennett & Nijssen, 2021b). We then calculate *B* from daily average *Ta* and *RH* using the SFE and from daily average *LE* and *H* using SUMMA outputs for each of the three variants to compare model performance against *B* derived from observations at each site. We select daytime conditions for which incoming shortwave radiation >50 W m^{-2}. We exclude data flagged as gap-filled and only consider days with at least 12 half-hourly flux observations. To best meet assumptions of the SFE model and avoid stable atmospheric boundary layer conditions, we exclude days with negative *LE* or *H*. Sites cover a range of wet to arid biomes, including a diversity of sparsely vegetated, grassland, and forested ecosystems. This curated data set also spans a representative range of source dependencies. To compute *pdfs* for monthly time windows, we bin all data globally, that is, we discretize each variable in 50 equally sized bins spanning 0–6 for *B*, −10 to 40°C for *Ta*, and 0–1 for *RH*, and we set all extreme values above and below the prescribed ranges to the first and last bins. We therefore calculate information measures and associated functional and predictive performance metrics for each site using a histogram bin resolution of 2% of the range of observed data, similarly to previous benchmarking studies (Nearing et al., 2018).

3. Results: Synthetic Binary Models Show Functional Ranges

For each of the seven binary model cases based on a range of source dependencies, we compute the total information from sources to the target, $I(X, Y; Z)$, and the associated information decomposition (Figure 2). Although the knowledge of X and Y together completely reduces the uncertainty of Z, the value of $H(Z)$ depends on the source distributions in addition to the model mapping, $(x, y) \rightarrow z$, such that $H(Z)$ can range from 0 to 1. Specifically, $H(Z) = 1$ for cases where the model map and the source *pdf* $p(x, y)$ result in a uniform distribution over Z, and $H(Z)$ approaches zero for cases where the outcome of Z is nearly certain. For example, in the AND(X,Y) model, $(1, 1) \rightarrow 1$ and all other source combinations lead to $Z = 0$, so if $p(1, 1)$ is extremely small, Z is nearly certain to be 0 and has correspondingly low entropy.

We find that several model "types" can be distinguished based on the ranges of U_x, U_y, S, and R information components for different source dependencies. Particularly, the models AND(X,Y), AND(¬X, Y), AND(X, ¬Y), and OR(X,Y) (equivalent to ¬AND[¬X, ¬Y]) have the same range of possibilities for all information components (distribution of gray points in Figure 2). These AND relationships correspond to model maps in which $Z = 1$ or $Z = 0$ for exactly one (x, y) combination (Figure 2a). Meanwhile, the models $Z = Y$ and $Z = X$ are similar to each other, since they are models for which X or Y completely determines Z such that one of the sources does not provide unique information nor is there any synergistic information regardless of source dependency. However, there can still be redundant information in these cases, which arises from the dependency between the sources. Finally, the binary XOR(X,Y) model is the only model that results in high synergistic information and low redundancy for a range of source dependencies. This model structure leads to relatively little overlap in the information that X and Y provide, and instead, they tend to provide synergistic and unique information.

This binary example of how information components vary depending on source dependency and model structure reveals several interesting features. First, for a given model and source dependency, there is typically a restricted range that a given information component can assume. For example, there are no cases for any model where a high $I(X; Y)$ source combination leads to high U_x or U_y. This decrease in unique information with increasing source dependency is expected, since $I(X; Y)$ is a term in the rescaled redundancy equation. More interestingly, S and R often show a branching structure, where for a given source dependency, one or the other is dominant. This implies a trade-off between synergy and redundancy that becomes more extreme as source dependency increases. In other words, for a certain type of model, high source dependency could either lead to high R or high S. We next address the differences between these scenarios.

Differences between models of a given general "category" become apparent when we look at other characteristics of the source variables. For example, we can partially distinguish between the four models with "AND" structures if we look at high versus low correlations between X and Y. For example, high positive covariance between sources corresponds to high R in the AND(X, Y) model, and high S in the AND(¬X, Y) model. For the AND(X, Y) case, a positive source correlation, or high probability for $X = Y$, allows us to distinguish between $Z = 0$ and $Z = 1$ outcomes given only one source variable. Due to this feature, the information that the other source provides is highly redundant. In contrast, for the AND(¬X, Y) model, $(0, 1) \rightarrow 1$, while all other source combinations lead

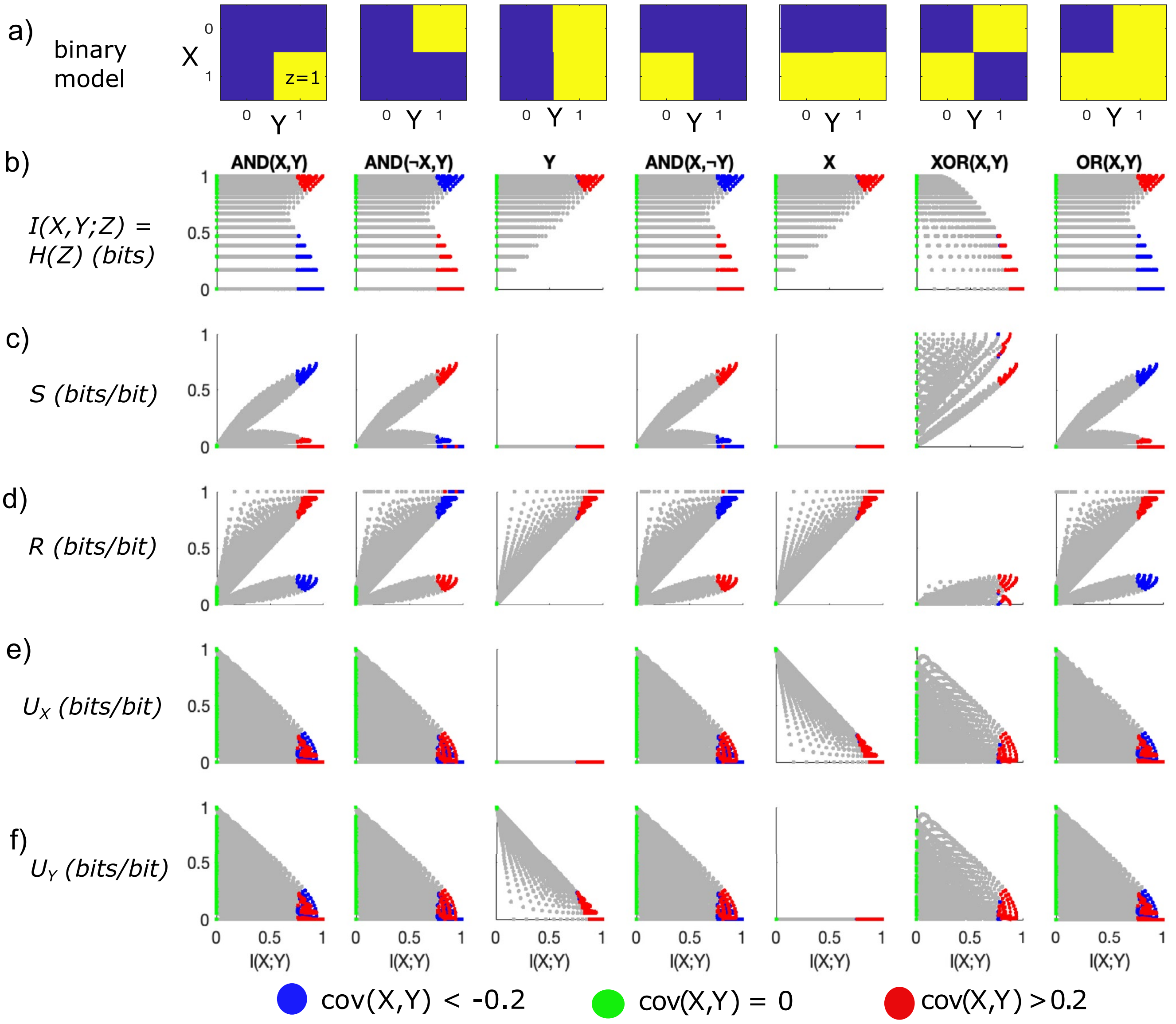

Figure 2. Synthetic binary model case study. (a) Model mappings for 7 different model structures. For example, model AND(X, Y) indicates that $z = 1$ when both $x = 1$ and $y = 1$. The ¬ symbol indicates the logical "not." Information types as a function of source dependency, $I(X; Y)$ for each model, including (b) total information from sources to the target (equivalent to $H(Z)$ for these models) and (c) synergistic (S), (d) redundant (R), (e) unique from X (U_X), and (f) unique from Y (U_Y) information components normalized by total information. Gray dots indicate all generated cases. Blue dots highlight cases with negative covariance between the sources, green dots highlight independent sources (cov(X, Y) = 0 and $I(X; Y) = 0$), and red dots are cases with high source covariance.

to $Z = 0$. Here, positively correlated sources do not distinguish between $Z = 1$ and $Z = 0$, since the knowledge of a single source variable would always lead to a most likely outcome of $Z = 0$. This is exhibited by the low $H(Z)$ for highly correlated sources in the AND(¬X, Y) model (red dots in Figure 2). For this model, since a source combination that falls in the negatively correlated range distinguishes between the possible Z outcomes (blue dots in Figure 2), the knowledge of both sources jointly, or synergistically, informs the outcome of Z.

The key insight gained through these binary model mappings is how both the dependency between source variables and model structure influence the range of synergistic, unique, and redundant information partitionings. Specifically, relatively dependent sources provide avenues for both redundancy and synergy, and relatively independent sources can provide more unique information to the target and lead to a wider range of possible behaviors (green dots in Figure 2). Given input data with the same joint distribution, different model structures lead to different information flow pathways. However, models with similar structures lead to a similar range of possibilities. While this example is useful to explore the simplest possible models in detail, more complex model setups enable us to better understand the functional ranges of more realistic models.

4. Results: Synthetic Time Series Models Show Effect of Functional Forms

Next, we compare four two-source model structures (Equations 10–13) with four induced source dependency types that range from linear to nonlinear. Across all model structures and source dependency ranges, we see broad patterns in information components (Figure 3). In general, S decreases as sources become more dependent, R increases, and unique components U_X and U_Y tend to increase up to a certain source dependency and then decrease. Synergistic information is the dominant component when sources are independent (Figures 3e–3h). As sources become more dependent, S drops relatively sharply as unique information increases. At a certain threshold source dependency, we see that redundant information begins to increase sharply and is the dominant component for completely dependent sources. This $I(X; Y)$ threshold at which R increases corresponds to the point at which $R_{\min}$ in the rescaled redundancy term goes from 0 to a positive value (Equation 4). Below that threshold, $S > R$ and the minimum R value is zero. In other words, there is a source dependency at which the interaction information goes from positive to negative, and R must be scaled based on the interaction information. We see that this threshold exists for all dependency types but tends to be slightly higher for nonlinearly dependent sources (blue colors in Figure 3).

Besides general patterns, there are differences between model structures (columns in Figure 3) and dependency types (colors in Figure 3). For example, multiplication and addition models, $X * Y$ and $X + Y$, show very similar behaviors regardless of whether sources are linearly or nonlinearly dependent. However, this finding would be much different if we considered source covariance on the horizontal axis, rather than mutual information, as a measure of source dependency. Particularly, the nonlinearly dependent sources have near-zero covariances, such that these patterns cannot be discerned and all model cases would collapse to the vertical axis (as shown in Figure S1 in Supporting Information S1). This highlights the relevance of using an information theory-based measure of source dependency rather than, or in addition to, linear correlation. With mutual information, we see that source dependencies have a predictable effect on information partitioning components given a model structure.

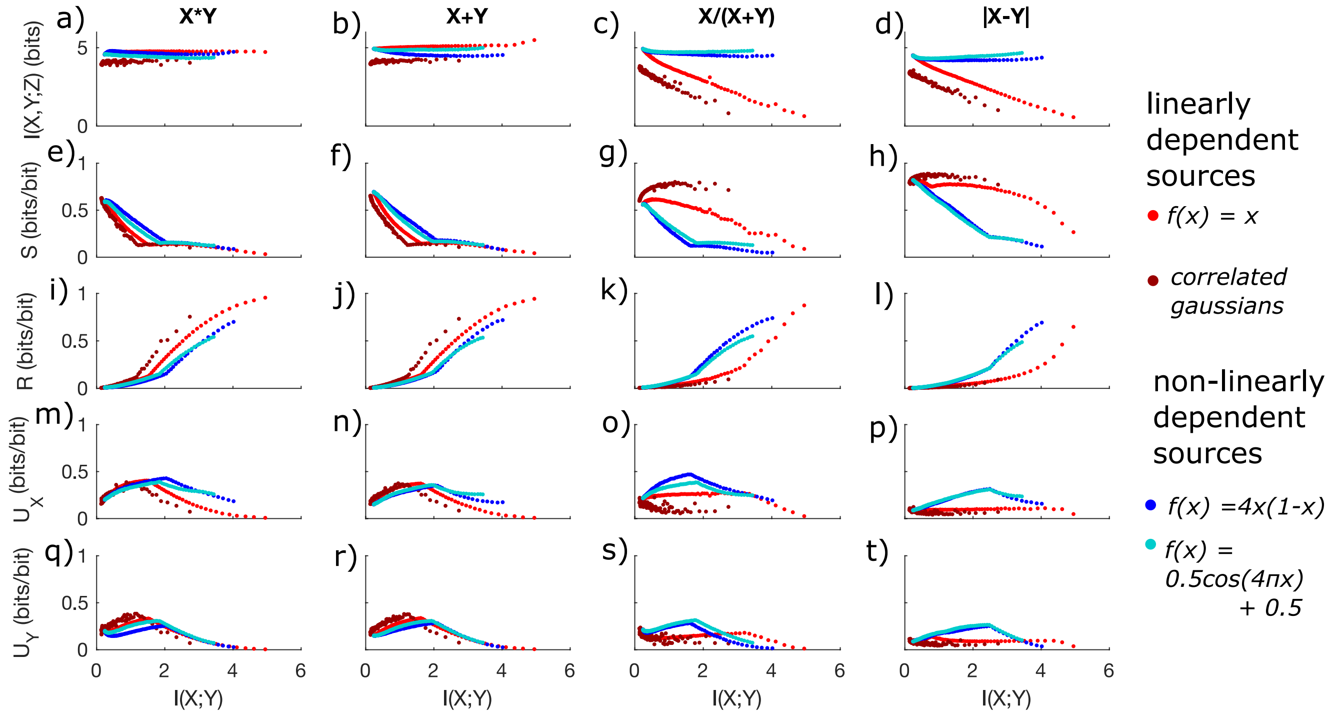

Figure 3. Synthetic time series model case study, based on simple arithmetic operations corresponding to Equations 10–13 and for four different induced source dependency types indicated by colors (two linear and two nonlinear). Information types as a function of source dependency, $I(X; Y)$, for each model, including (a–d) total information, and (e–h) synergistic (S), (i–l) redundant (R), (m–p) unique from X (U_x), and (q–t) unique from Y (U_y) information components for each model normalized by total information.

Meanwhile, the division and difference models (third and fourth columns of Figure 3) show more variability between linearly and nonlinearly dependent sources. Mainly, when sources are linearly dependent, S remains high for a larger range of source dependencies, U_X and U_Y are lower, and R increases more gradually with source dependency. This is also related to a decrease in target entropy as sources become linearly dependent (red lines in Figures 3c and 3d). In other words, as sources become correlated, dividing them or finding their difference goes toward a constant value, and target entropy is zero regardless of the variability of the sources. In contrast, this is not the case for nonlinearly dependent sources. Of all models, the difference, $|X - Y|$, shows the highest S at the expense of the lowest U_X and U_Y.

A key insight gained from these synthetic models is that different mathematical operations, in tandem with certain types of source dependencies, determine the influence of two sources in terms of relative strengths and whether provided information can be unique, redundant, or synergistic. Moreover, we see that different ways in which sources are combined can cause a "loss" of information content from one or both sources, such as in the proportion example (Figure 3c) for which Z_{div} approaches a constant value as sources become linearly correlated. While the target is always fully predictable given both sources, the variability of those sources is not always reflected in the target. This example also highlights the relevance of mutual information as a basis on which to compare model cases, since it captures more general aspects of source dependencies other than multicollinearity.

5. Results: Predictive Performance of Weather Station Regressions Tied to Functional Performance

From the normalized and filtered observed weather station data, we find that on average, the knowledge of Ta and WS reduces about 31% of the uncertainty of RH, or that $\frac{I(Ta,WS;RH)}{H(RH)} = 0.31$. We note again that we have filtered the dominant diurnal cycle from Ta and RH, such that the variables are composed of higher frequency variations. Without this filtering, the shared information between Ta and RH would be higher. From the observed data, we see similar but more muted trends in S and R relative to the arithmetic operations, in which S tends to decrease and R tends to increase as sources become more dependent (dotted lines in Figures 4a, 4c and 4e, g). From observations, we see that the dominant information component is unique information from Ta (U_{Ta}, Figure 4e), but models vary widely in terms of information component strengths. Particularly, models underestimate S and overestimate R (Figures 4a–4d). The main exceptions to this are the addition and subtraction models, $RH_{\mathrm{mod},8}$ and $RH_{\mathrm{mod},9}$, where the single model input is the addition or subtraction of the two original sources. Meanwhile, models are split between over- and underestimating unique information, but any model that overestimates one unique component underestimates the other (Figures 4e–4h).

When RH is estimated as a linear function of either Ta ($RH_{\mathrm{mod},1}$) or WS ($RH_{\mathrm{mod},2}$), the unique contribution from the included source is overestimated, since the synergistic and non-included unique components are inherently zero (Figures 4a–4h). In these two cases, R exists due to dependency between sources. In other words, the non-included source only provides information through its dependency with the included source. In contrast, a source that is not included as a model input can never provide synergistic or unique information. The remaining models account for both sources in different ways. Models $RH_{\mathrm{mod},3}$, $RH_{\mathrm{mod},7}$, $RH_{\mathrm{mod},8}$, and $RH_{\mathrm{mod},9}$ perform a simple math operation on Ta and WS that is then used as a single input in a linear regression model. As noted in the previous synthetic examples, this operation results in some loss of information that the sources could provide to the modeled target. For example, multiplying two variables results in a single source that retains some information about each original source, but the original Ta and WS cannot be completely reconstructed given only the result of the multiplication. Here, this feature leads to relatively poor predictive and functional partitioning performances (high mismatch in terms of A_p and $A_{f,I\mathrm{part}}$, Figures 4i and 4j) for these model types, especially for $RH_{\mathrm{mod},3}$ (multiplication) and $RH_{\mathrm{mod},7}$ (division). Meanwhile, we see that $RH_{\mathrm{mod},8}$ (addition) and $RH_{\mathrm{mod},9}$ (subtraction) models have the highest functional partitioning performance (lowest mismatch in terms of $A_{f,I\mathrm{part}}$) of all models, indicating that the addition or subtraction retains more relevant information about the sources, particularly Ta, which is the stronger driver of RH. Finally, models $RH_{\mathrm{mod},4}$, $RH_{\mathrm{mod},5}$, and $RH_{\mathrm{mod},6}$ all utilize both sources individually and/or jointly. $RH_{\mathrm{mod},4}$ and $RH_{\mathrm{mod},5}$, which are linear regressions with two or more sources, tend to have high performance but always overestimate unique information from Ta and underestimate unique information from WS as the weaker source, and severely underestimate S. This indicates that the simple math operations force sources to be weighted more equally, and eliminate some of the information they could have provided jointly.

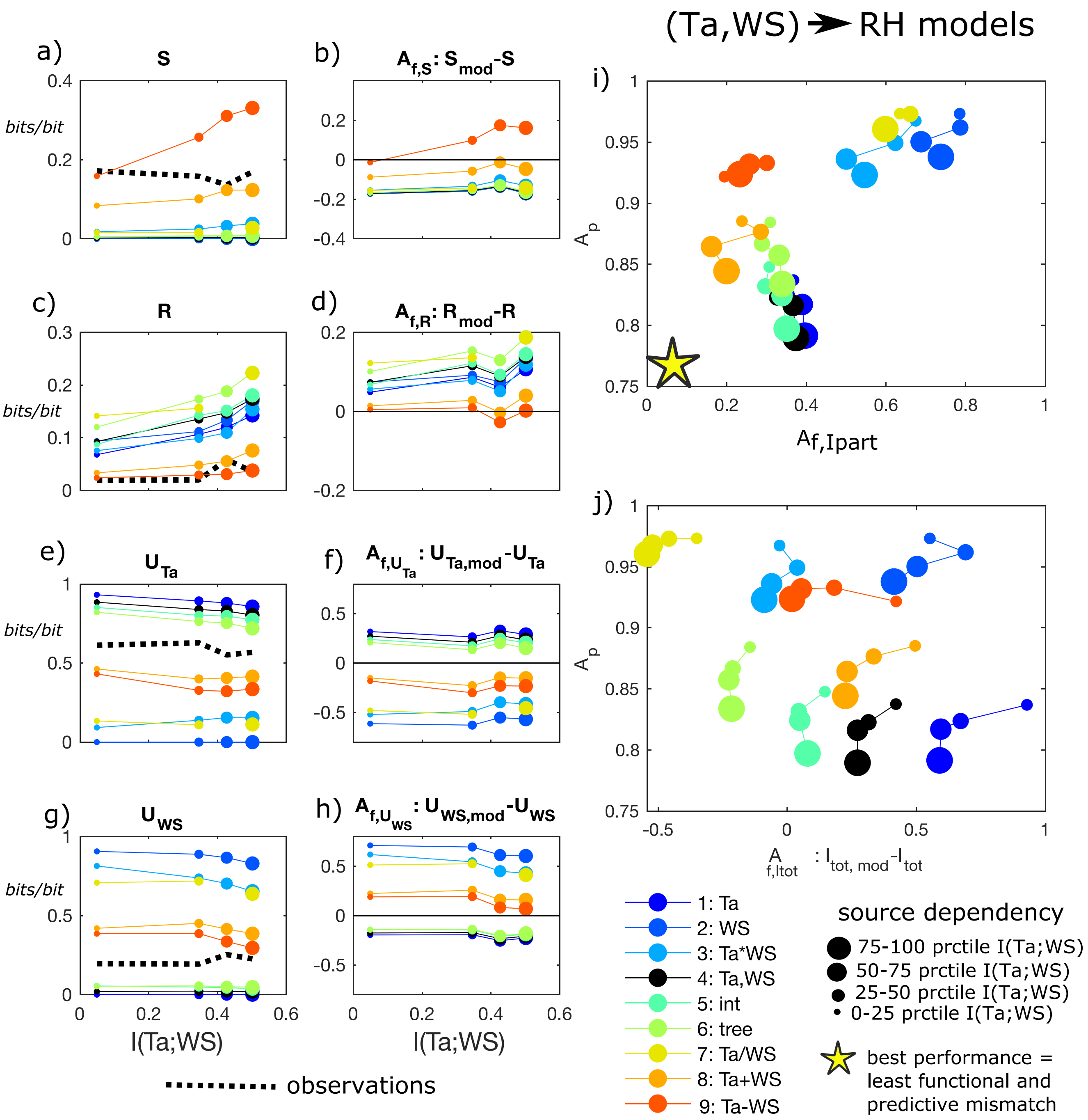

Figure 4. Weather station model case study for different relative humidity (*RH*) regressions as a function of air temperature (*Ta*) and wind speed (*WS*). Comparison of individual information types between different models (colors) and observations (dashed black line) over the range of source dependencies, *I*(*WS*; *Ta*) (circle sizes) and functional performances, including (a–b) synergistic information, *S* and $A_{f,S}$, (c–d) redundant information, *R* and $A_{f,R}$, (e–f) unique information from *Ta*, U_{Ta} and $A_{f,U_{Ta}}$, and (g–h) unique information from *WS*, U_{WS} and $A_{f,U_{WS}}$. (i) Functional partitioning performance $A_{f,Ipart}$ relative to predictive performance A_p. (j) Functional total information performance $A_{f,Itot}$ relative to A_p, for different source dependency ranges.

Most models have better predictive performance for time windows with higher source dependency. The only exceptions to this are $RH_{mod,9}$ (subtraction) and $RH_{mod,7}$ (division), for which there is little or no change in A_p with source dependency (Figures 4i and 4j). This feature of improved predictive performance with source dependency, especially for the "best" performing models, indicates that these models take advantage of source dependencies to compensate for structural errors. When we consider the overall functional performance of these models, we see that $A_{f,I\text{part}}$ improves along with predictive performance for the worst-performing models (upper right corner of Figure 4i) but shows varying behaviors for the addition and subtraction models and fewer changes with source dependency for the "best" performing models.

An analysis of the total information that sources provide to the target, as $A_{f,I\text{tot}}$, reveals a different aspect of these models (Figure 4j). Particularly, the several models with best predictive performance are among the most "overly deterministic." Meanwhile, the addition, subtraction, and tree models more closely approach the observed total information, and the division model is the most "overly stochastic" model. As noted previously, $A_{f,I\text{tot}}$ is not a reliable measure of performance since models can have $A_{f,I\text{tot}}$ arbitrarily close to zero.

When redundancy is overestimated for highly dependent sources, models use information from a source that is actually redundant as unique or synergistic instead. This leads us to infer that when sources are more dependent, these models achieve better predictive performance by using the overlap in information the sources provide without "realizing" that it is overlapping. In this case study, *Ta* is a dominant source and *WS* is a weaker source. When *Ta* and *WS* are dependent and provide more redundant information, a model that does not take *Ta* into account, or generally has some structural error in terms of how it represents the two sources, will perform better than when *Ta* and *WS* are independent. We confirm from the underestimation of redundancy that the model is behaving this way, or attributing unique information to the source for which it more adequately accounts.

A key insight of this case study is that source dependencies can enable models to compensate for missing or misinterpreted sources. This is in agreement with known aspects of multicollinearity, in that model fit tends to improve with more correlated sources. However, our approach further highlights the particular mechanisms behind these behaviors in the form of types of information flows relative to observations.

6. Results: Functional Performance of Bowen Ratio Models Tied to Source Dependency and Complexity

Next, we explore the relationship between source dependency and model performance for simple to more complex ecohydrological model structures in the estimation of daily land-surface sensible and latent heat fluxes (Figure 5). We compare the Bowen ratio (*B*) based on the SFE, which depends only on the two source variables *Ta* and *RH*, with three versions of SUMMA. SUMMA-SA is based on physical equations, while SUMMA-NN1 and SUMMA-NN2 implement neural networks to estimate *LE* and *H* fluxes.

For observed data, $U_{RH} > U_{Ta}$ for all source dependencies, indicating that *RH* informs observed *B* more strongly than *Ta* (Figure 5c, black dotted lines). With increasing source dependency, *Ta* and *RH* contribute an increasing amount of information about observed *B*. Meanwhile, the relative amount of *S* decreases, *R* increases, and both U_{Ta} and U_{RH} show a step-wise behavior (Figure 5a). For source dependencies up to about $I(Ta, RH) = 2$ *bits*, *S* dominates *R* and U_{Ta} and U_{RH} are relatively constant. For higher source dependencies, *R* dominates *S* and U_{Ta} and U_{RH} decrease with source dependency. These patterns are similar to the previous multiplication and addition model examples (Figure 3). All four models follow these general trends but to different extents, as shown from individual functional performance measures (Figure 5b), and thus lead to different predictive (A_p, Figure 5c) and overall partitioning functional ($A_{f,I\text{part}}$, Figure 5d) performances.

The simplest SFE model is overall less accurate in terms of estimating *B* (higher mismatch in terms of A_p) and in representing the causal relation between *Ta*, *RH,* and *B* (higher mismatch in terms of $A_{f,I\text{part}}$) compared to the SUMMA variants (Figures 5c and 5d). The SFE is particularly less accurate when sources are independent because *Ta* and *RH*, the only SFE inputs, are less informative of *B* as $I(Ta, RH)$ decreases. Within the more complex SUMMA models, SUMMA-NN1 has overall lower predictive and functional performance compared to SUMMA-SA and SUMMA-NN2, indicating that the missing information about soil moisture states and therefore the meteorological history in SUMMA-NN1 is necessary to accurately represent the interactions between *Ta* and *RH* and *B*. Additionally, this information is most important in process states in which *Ta* and *RH* are less coupled. Predictive and functional performance for all model variants improves with higher source dependencies. In general, these results reflect the findings of the weather station (*Ta*, *WS*, *RH*) regression models, in that all models take advantage of source dependencies for better predictive performance.

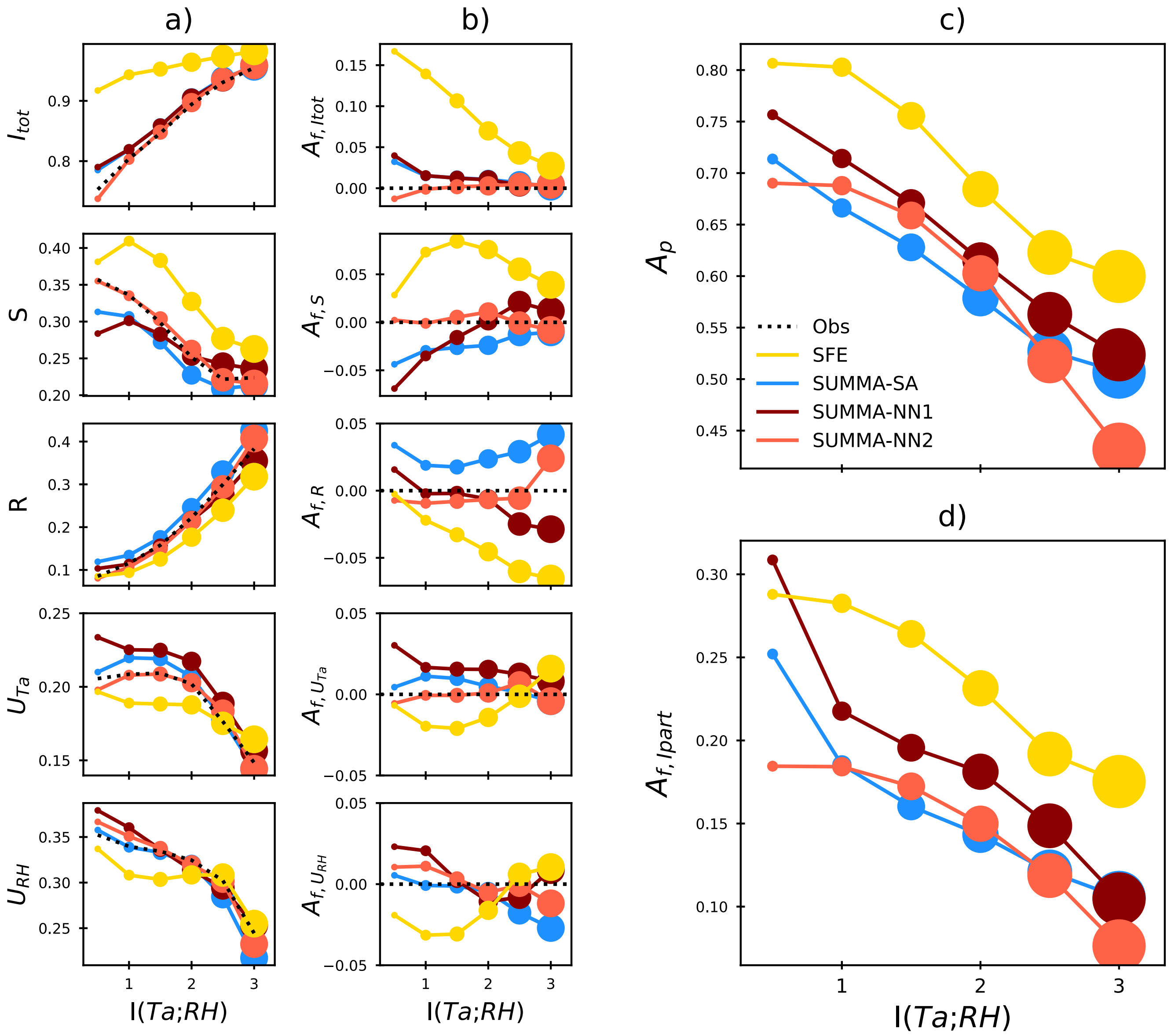

Figure 5. Ecohydrological case study for simple theoretical (SFE) to more complex process-based (SUMMA-SA), data-driven (SUMMA-NN1), and hybrid (SUMMA-NN2) models for the Bowen ratio (b). Comparison of information components between models (colors) and observations (dashed black line) over the range of air temperature (*Ta*) and relative humidity (*RH*) source dependencies, $I(Ta; RH)$ (circle sizes). (a) Total mutual information normalized by Shannon's entropy of *B* (I_{tot}, *bits/bit*) decomposed into synergistic (*S*), redundant (*R*) and unique (U_{Ta}, U_{RH}) information components. (b) Functional performance for each information type ($A_{f,I\mathrm{tot}}$, $A_{f,S}$, $A_{f,R}$, $A_{f,U_{Ta}}$, $A_{f,U_{RH}}$, *bits/bit*). (c) Model predictive performance (A_p, *bits/bit*); (d) overall model functional partitioning performance ($A_{f,I\mathrm{part}}$, *bits/bit*). Markers represent the average individual month-site metrics for six evenly spaced bins along the $I(Ta; RH)$ axis *in bits*.

When we focus on performance for individual information components, we find diverging information flow relationships along source dependency ranges (Figures 5c and 5d), particularly for the SFE model. SFE mostly overestimates the total information contribution from *Ta* and *RH* (higher mismatch in terms of $A_{f,I\mathrm{tot}}$) because these are the only inputs. The SUMMA variants utilize more inputs, and *Ta* and *RH* contribute about the same amount of total information to *B* as in observations. SUMMA-NN2 is the most accurate overall in terms of $A_{f,I\mathrm{tot}}$, especially for low source dependencies. The SFE model underestimates U_{Ta} and U_{RH} for low source dependencies, overestimates *S* for mid to high source dependencies, and underestimates *R* for the range of source dependencies. Specifically, as source dependency increases, unique information performance improves at the expense of greatly underestimating redundancy. This matches with the weather station regressions, in that the model is "using redundant information uniquely" or taking advantage of dependencies between *Ta* and *RH* to make better predictions of B. SUMMA-NN2 and SUMMA-NN1 result in higher S, lower R and higher U_{RH} than SUMMA-SA. This points to the neural network's ability to use information from sources more synergistically, while the physically based equations of SUMMA-SA cause sources to provide more R at the expense of less U_{RH}. We also note that the sensitivity of individual functional performance metrics to source dependency is different for each model variant. For example, $A_{f,S}$ improves with source dependency for SUMMA-SA, while $A_{f,S}$ is relatively constant for SUMMA-NN2. These results show that all models trade off types of information particularly when sources have the highest and lowest dependencies. This effect is greatest for SFE and SUMMA-SA, while individual functional performance metrics for SUMMA-NN2 tend to be more constant along the source dependency axis.

We note that using information theoretical performance metrics to compare SUMMA variants at the 30 min timescale and for LE and H separately, we find that SUMMA-NN1 and SUMMA-NN2 outperform SUMMA-SA in terms of predictive performance as was shown in (Bennett & Nijssen, 2021a). Meanwhile, the pattern in functional performance, especially for H, is more mixed (Figures S3–S4 in Supporting Information S1). This likely led to the more similar performance results between SUMMA-NN2 and SUMMA-SA in terms of daily B and the clearer underperformance of SUMMA-NN1. The daily B results thus highlight the advantage of SUMMA-SA, which enforces energy conservation, in estimating both LE and H together from physical equations. These findings were not previously apparent from more traditional performance metrics (Bennett & Nijssen, 2021a) and when only considering predictive performance of individual fluxes at sub-daily timescales (Figures S3–S5 in Supporting Information S1).

While the SFE is by far the most parsimonious model to estimate B and does have lower performance relative to SUMMA models, we find that it generally follows the observed information flow patterns. Additionally, we find that information flow paths of SUMMA variants that implement neural networks can match observations just as closely as the standalone land-surface model. This example provides insights into how evaluating model performance based on source dependencies enables comparisons between models of very different complexities and structures. All model types use source dependencies to improve their performances, but through different mechanisms, highlighted by information component-level analyses. Specific differences in functional performance between model variants demonstrates how model structure affects information flows and how they match observed dependencies between variables. Additionally, these differences explain why a given variant has the highest performance in different source dependency ranges.

7. Discussion

The scaffolding of increasingly complex case studies presented here highlight how dependencies within model inputs can affect functional and predictive model performance, and that this feature varies depending on model structure. Based on these experiments, we briefly review several key insights and further discuss the advantages and limitations of this model evaluation framework.

From synthetic models, we find that structure and source distribution both restrict the functional range of a model. We see this from the binary examples, in that only one model type (binary XOR) leads to dominantly synergistic information flows. This is also evident in the models based on generated data and arithmetic operations, in that different equation types lead to different relationships between source dependency and information measures. For example, it is possible to develop a model, and likely there are models in current use, that cannot possibly replicate certain information flow pathways that are observed in nature. Our synthetic cases also illustrate the value of a nonlinear measure of source dependency over linear correlations.

From the models tied to observations for which we can evaluate their performance, we find that models use dependent sources to achieve improved performance during those time periods. On one hand, this matches with intuition regarding multicollinearities between sources which is a long-standing area of study in regression modeling. However, while all models tended to have improved performance measures with highly dependent sources, the mechanisms behind these improvements vary as shown by patterns in redundant, synergistic, and unique information flows. For some models, such as the weather station regression models and the SFE, the model structure has some error that causes it to improperly account for one of the sources. For these models, dependent sources allow them to compensate with redundant information that the other source provides. This is revealed as an underestimation of redundant information relative to observations. For other more complex models, such as the SUMMA variations, source dependencies instead enable the model to better represent synergistic information that sources provide together. In other words, sometimes a model "gets the right answers" due to source dependencies, where it is able to utilize information from a source that it does not adequately account for, if that source is somewhat related to another model input. This also points to the fact that sometimes our model assumptions or simplifications are not correct in all process states and some types of information are traded for others depending on source dependencies. For example, we can say the SUMMA stand-alone and NN1 models are less reliable in a hypothetical future case when sources are more independent of each other, and the NN2 is more likely to provide a better result.

Understanding this trade-off of information types and functional performance measures along a source dependency range contributes to the unsolved problem of "disentangling" different types of model uncertainties in hydrological prediction (Blöschl et al., 2019). Uncertainty in models can be caused by errors in forcing data, model parameters, and the structure of a model. Model benchmarking is useful to separate model uncertainties into these different components (Nearing et al., 2016), and it has been found that models often use only a small amount of the information available to them. It has also recently been suggested that machine learning models could be integrated with process-based models and achieve better performance or better understanding of processes than either model would alone (Bennett & Nijssen, 2021a). These studies and others indicate significant potential in terms of model improvements and potential for machine learning techniques to reveal process understanding in addition to improved predictive performance, through combining the strengths of different modeling frameworks (Reichstein et al., 2019). Findings that machine learning models often have more predictive power than physically based models indicate that these models are able to learn complex dependencies from their training data. It is important to diagnose the causal interactions learned in data-driven models that allow them to map input to output with high accuracy. If the data-driven models capture causal interactions better, this could be used to improve process-based models by encoding a better understanding of how information should flow through the model. Meanwhile for physically based models, it has been shown that a trade-off between functional and predictive performance, in which a model may be tuned to higher accuracy at the cost of its process representations, indicates a structural error in the model (Ruddell et al., 2019).

Models of any type implement our knowledge and assumptions about causal interactions. Due to this, performance metrics that can test these assumptions without relying on interpreted system states are useful. For example, we could have evaluated the weather station, SFE, and SUMMA models based on aridity conditions within each time window, and find lower model performance in a certain aridity range. In some cases, this could point to a related process equation within the model, but particularly for a data-driven model, it may not reveal a certain type of error that could be corrected or improved upon. Meanwhile, an information theory-based source dependency can be considered a "system state" (Ruddell & Kumar, 2009) that is agnostic to the magnitudes of the sources or environmental conditions that are assumed to be influential to a certain process. As shown in several examples, this aspect of different behaviors along a source dependency axis can potentially be related to structural components of a model, such as in the addition, multiplication, and division model examples.

While we restrict the focus to two sources that can jointly inform a target, other frameworks could be applied to consider modeled and observed interactions in increasingly multivariate contexts (Jiang & Kumar, 2018, 2020; Runge et al., 2019; van Leeuwen et al., 2021). For example, it could be that there are many "gateway" types of interactions (van Leeuwen et al., 2021), where variables are more or less influential in different ranges, or one variable sets the level of influence of another variable. Meanwhile, physically based models may not take these types of dependencies into account. A more multivariate approach could also be taken to study information flows through time in models with different structures, relative to lagged or instantaneous source dependencies. Particularly, models that account for lagged inputs may show different time-dependent behaviors under different source dependency scenarios, relative to the most simple models.

8. Conclusion

This study fills a gap in model evaluation to answer emerging questions that involve multiple modeling approaches. Metrics that can diagnose more detailed aspects of model behavior and therefore robustly quantify multiple dimensions of model performance are necessary to accompany the proliferation of ecohydrologic models of varying complexities and levels of physical process representations. Such metrics are particularly timely given the debate about process-based, data-driven, and hybrid modeling approaches and the utility of scale-emergent relations that take advantage of source dependencies to estimate ecohydrological fluxes from satellite data. In practice, the framework introduced here is relevant for several aspects of model evaluation and development, summarized below:

- Uncertainty quantification: An information theory-based approach directly quantifies uncertainty and reductions in uncertainty, such that model diagnostics always relate to a fraction of reduced or remaining uncertainty. Particularly, we categorize reduced uncertainty into different types of information and can compare these information uses between models and observations.
- Feature selection: An extended version of this framework could be compared with feature selection or dimension reduction techniques, such as Principal Components Analysis (PCA). While PCA leads to dimensions that are linearly uncorrelated, nonlinear aspects of source dependencies may remain and could be significant. In general, considering source dependencies provides a pathway to determine levels of redundancy or overlap between multiple variables.
- Extrapolation of forcing conditions: The focus on source dependency and the "functional range" of a model highlights the range of model behaviors over the range of source dependencies. With this, we can predict which models are most likely to have the most or least accurate predictions under source dependencies that have not been observed, regardless of the magnitudes of those sources. Additionally, by considering functional performance along with predictive performance, we assume that a functionally more accurate model should have better ability to extrapolate beyond known observations.
- Model development or improvement: This framework introduces a nuanced set of model diagnostics that can be used to identify how sensitive models are to linear and nonlinear source dependencies. This could be used to compare behaviors of very different modeling frameworks, or to compare different parameterizations of a model and choose parameters that optimize functional and predictive performances across a range of source dependencies.

By taking the view that "a hydrological model is not a tool, but a hypothesis" (Savenije, 2009), this paper contributes a framework to test more detailed questions about how models behave, in terms of the flow of unique, synergistic, and redundant information types, and under which conditions, in terms of source dependencies, model assumptions robustly reproduce ecohydrological relationships. Despite our incomplete information, we can gain deeper insights about how our understanding of ecohydrological systems, encoded in models, compares to observations.

Data Availability Statement

The authors thank Andrew Bennett for SUMMA model output available at https://doi.org/10.5281/zenodo.4735592, which utilizes FLUXNET2015 data products available at https://fluxnet.org/data/fluxnet2015-dataset/. We would also like to thank three reviewers who provided valuable feedback on our manuscript. Matlab and Python codes for analyses presented here are available at https://github.com/allisongoodwell/SourceDependency2021 (DOI: https://doi.org/10.5281/zenodo.6600721).

Acknowledgments

AEG is supported by NSF Grant EAR #2012850 for the Critical Interface Network for Intensively Managed Landscapes (*CINet*) and NASA New Investigator Grant 80NSSC21K0934. MB thanks the European Commission and the Swedish Research Council for Sustainable Development FORMAS (Grant No. 2018-02787) for funding in the frame of the collaborative international consortium iAqueduct financed under the 2018 Joint call of the WaterWorks 2017 ERA-NET Cofund. This ERA-NET is an integral part of the activities developed by the Water JPI.

References

Baldocchi, D. D., Keeney, N., Rey-Sanchez, C., & Fisher, J. B. (2022). Atmospheric humidity deficits tell us how soil moisture deficits down-regulate ecosystem evaporation. *Advances in Water Resources*, *159*, 104100. https://doi.org/10.1016/j.advwatres.2021.104100

Barrett, A. B. (2015). Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E*, *91*(5), 052802. https://doi.org/10.1103/PhysRevE.91.052802

Bassiouni, M., & Vico, G. (2021). Parsimony versus predictive and functional performance of three stomatal optimization principles in a big-leaf framework. *New Phytologist*, *231*, 586–600. https://doi.org/10.1111/nph.17392

Bassiouni, M., Vogel, R. M., & Archfield, S. A. (2016). Panel regressions to estimate low-flow response to rainfall variability in ungaged basins. *Water Resources Research*, *52*(12), 9470–9494. https://doi.org/10.1002/2016WR018718

Bennett, A., & Nijssen, B. (2021a). Deep learned process parameterizations provide better representations of turbulent heat fluxes in hydrologic models. *Water Resources Research*, *57*(5), 1–14. https://doi.org/10.1029/2020WR029328

Bennett, A., & Nijssen, B. (2021b). Output data for "deep learned process parameterizations provide better representations of turbulent heat fluxes in hydrologic models". *Zenodo*. https://doi.org/10.5281/zenodo.4735592

Bennett, A., Nijssen, B., Ou, G., Clark, M., & Nearing, G. (2019). Quantifying process connectivity with transfer entropy in hydrologic models. *Water Resources Research*, *55*(6), 4613–4629. https://doi.org/10.1029/2018WR024555

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., et al. (2019). Twenty-three unsolved problems in hydrology (UPH)—A community perspective. *Hydrological Sciences Journal*, *64*(10), 1141–1158. https://doi.org/10.1080/02626667.2019.1620507

Brutsaert, W. (1982). *Evaporation into the atmosphere: Theory, history, and applications*. Springer. https://doi.org/10.1007/978-94-017-1497-6

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, *51*(4), 2498–2514. https://doi.org/10.1002/2015WR017198

Fisher, J. B., Lee, B., Purdy, A. J., Halverson, G. H., Dohlen, M. B., Cawse-Nicholson, K., et al. (2020). ECOSTRESS: NASA's next generation mission to measure evapotranspiration from the international space station. *Water Resources Research*, *56*(4), e2019WR026058. https://doi.org/10.1029/2019WR026058

Gharari, S., Gupta, H. V., Clark, M. P., Hrachowitz, M., Fenicia, F., Matgen, P., & Savenije, H. H. G. (2021). Understanding the information content in the hierarchy of model development decisions: Learning from data. *Water Resources Research*, *57*(6), e2020WR027948. https://doi.org/10.1029/2020WR027948

Goodwell, A. E., Jiang, P., Ruddell, B. L., & Kumar, P. (2020). Debates—Does information theory provide a new paradigm for Earth science? Causality, interaction, and feedback. *Water Resources Research*, *56*(2), 1–12. https://doi.org/10.1029/2019WR024940

Goodwell, A. E., & Kumar, P. (2017). Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables. *Water Resources Research*, *53*(7), 5920–5942. https://doi.org/10.1002/2016WR020218

Goodwell, A. E., Kumar, P., Fellows, A. W., & Flerchinger, G. N. (2018). Dynamic process connectivity explains ecohydrologic responses to rainfall pulses and drought. *Proceedings of the National Academy of Sciences*, *115*(37), E8604–E8613. https://doi.org/10.1073/pnas.1800236115

Jiang, P., & Kumar, P. (2018). Interactions of information transfer along separable causal paths. *Physical Review E*, *97*(4), 1–21. https://doi.org/10.1103/PhysRevE.97.042310

Jiang, P., & Kumar, P. (2020). Bundled causal history interaction. *Entropy*, *22*(3), 1–12. https://doi.org/10.3390/e22030360

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3), 1–5. https://doi.org/10.1029/2005WR004362

McColl, K. A., & Rigden, A. J. (2020). Emergent simplicity of continental evapotranspiration. *Geophysical Research Letters*, *47*(6), 1–11. https://doi.org/10.1029/2020GL087101

McColl, K. A., Salvucci, G. D., & Gentine, P. (2019). Surface flux equilibrium theory explains an empirical estimate of water-limited daily evapotranspiration. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2036–2049. https://doi.org/10.1029/2019MS001685

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

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

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rigden, A. J., & Salvucci, G. D. (2015). Evapotranspiration based on equilibrated relative humidity (ETRHEQ): Evaluation over the continental US. *Water Resources Research*, *51*(4), 2951–2973. https://doi.org/10.1002/2014WR016072

Ruddell, B. L., Drewry, D. T., & Nearing, G. S. (2019). Information theory for model diagnostics: Structural error is indicated by trade-off between functional and predictive performance. *Water Resources Research*, *55*(8), 6534–6554. https://doi.org/10.1029/2018WR023692

Ruddell, B. L., & Kumar, P. (2009). Ecohydrologic process networks: 1. Identification. *Water Resources Research*, *45*(3), 1–22. https://doi.org/10.1029/2008WR007279

Ruddell, B. L., Yu, R., Kang, M., & Childers, D. L. (2016). Seasonally varied controls of climate and phenophase on terrestrial carbon dynamics: Modeling eco-climate system state using dynamical process networks. *Landscape Ecology*, *31*(1), 165–180. https://doi.org/10.1007/s10980-015-0253-x

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., et al. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*(1), 1–13. https://doi.org/10.1038/s41467-019-10105-3

Savenije, H. H. G. (2009). Hess opinions "the art of hydrology". *Hydrology and Earth System Sciences*, *13*(2), 157–161. https://doi.org/10.5194/hess-13-157-2009

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, *85*(2), 461–464. https://doi.org/10.1103/PhysRevLett.85.461

Sendrowski, A., & Passalacqua, P. (2017). Process connectivity in a naturally prograding river delta. *Water Resources Research*, *53*(3), 1841–1863. https://doi.org/10.1002/2016WR019768

Sendrowski, A., Sadid, K., Meselhe, E., Wagner, W., Mohrig, D., & Passalacqua, P. (2018). Transfer entropy as a tool for hydrodynamic model validation. *Entropy*, *20*(1), 58. https://doi.org/10.3390/e20010058

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *196*(4), 519–520. https://doi.org/10.1016/S0016-0032(23)90506-5

Tennant, C., Larsen, L., Bellugi, D., Moges, E., Zhang, L., & Ma, H. (2020). The utility of information flow in formulating discharge forecast models: A case study from an arid snow-dominated catchment. *Water Resources Research*, *56*(8), 1–21. https://doi.org/10.1029/2019wr024908

van Leeuwen, P. J., DeCaria, M., Chakraborty, N., & Pulido, M. (2021). A framework for causal discovery in non-intervenable systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *31*(12), 123128. https://doi.org/10.1063/5.0054228

Weijs, S. V., Foroozand, H., & Kumar, A. (2018). Dependency and redundancy: How information theory untangles three variable interactions in environmental data. *Water Resources Research*, *54*(10), 7143–7148. https://doi.org/10.1029/2018WR022649

Weijs, S. V., van Nooijen, R., & van de Giesen, N. (2010). Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Monthly Weather Review*, *138*(9), 3387–3399. https://doi.org/10.1175/2010MWR3229.1

Williams, P. L., & Beer, R. D. (2010). Nonnegative decomposition of multivariate information. Retrieved from http://arxiv.org/abs/1004.2515