Ecohydrologic process networks: 1. Identification

Benjamin L. Ruddell[1] and Praveen Kumar[1]

Received 11 July 2008; revised 25 November 2008; accepted 17 December 2008; published 25 March 2009.

[1] Ecohydrological systems may be characterized as nonlinear, complex, open dissipative systems. Such systems consist of many coupled processes, and the couplings change depending on the system state or scale in space and time at which the system is analyzed. The arrangement of couplings in a complex system may be represented as a network of information flow and feedback between variables that measure system processes. The occurrence of feedback on such a network provides sufficient conditions for self-organized and nonlinear behaviors to emerge. We adapt an information-theoretic statistical method called transfer entropy for the purposes of robustly measuring the directionality, relative strength, statistical significance, and time scale of information flow between pairs of ecohydrological variables using time series data. A process network may be delineated where variables are cast as nodes and information flows as weighted directional links between them. The process network captures key couplings and time scales and represents the state of the complex system as a whole, including functional groups of variables (subsystems) and synchronization resulting from feedbacks. It is therefore able to identify interactions which are not detectable using methods which examine the system using one relationship at a time. We assemble an information flow process network using July 2003 FLUXNET data for a Midwestern corn-soybean ecohydrological system in a healthy, peak growing season state and compare the results with those using July 2005 data for the same site during a severe drought. We find that the process network during drought is substantially decoupled, and regional-scale information feedback is reduced during the drought. We conclude that the proposed process network methodology is able to identify the differences between two states of an ecohydrological system on the basis of variations in the pattern of feedback coupling on the network.

Citation: Ruddell, B. L., and P. Kumar (2009), Ecohydrologic process networks: 1. Identification, *Water Resour. Res.*, *45*, W03419, doi:10.1029/2008WR007279.

1. Introduction

[2] Ecohydrology is the study of the dynamic interaction of climate, soils, and vegetation, as they exchange carbon, water, energy, and nutrients at a range of scales in space and time [*Rodriguez-Iturbe*, 2000]. Ecohydrologic processes constitute open dissipative systems [*Jorgensen et al.*, 2007, *Kumar*, 2007] that exhibit emergent patterns that are usually identified through their geometric, dynamical, or statistical characteristics. The emergent patterns are a result of nonlinear feedback interactions between the components of the systems, and/or the system and the environment, where a small change leads to a series of interrelated changes that are not predictable from the knowledge of the behavior of the individual components. The occurrence of feedback in such a network indicates the existence of self-organizing emergent structures [*Capra*, 1996; *Hubler*, 2005]. In such situations, knowledge of a few state variables or coupling relationships is often insufficient to characterize the behavior of the entire system.

[3] The arrangement of couplings in such a complex system may be represented as a network of feedback relationships that transport material, energy, and information across different components [*Reiners and Driese*, 2003]. Each coupling has directionality such that change in one component causes change in another, albeit with a time lag. The hierarchy of these couplings involving the entire water cycle has been characterized as a hypercycle [*Kumar*, 2007]. A process network is defined as a network of feedback loops and the associated time scales that depicts the magnitude and direction of flow of matter, energy, and/or information between the different variables. The process network may be embedded in a physical space. For example, in a stream network transfer of water, nutrients, and biotic material sustains the habitat and food web of aquatic species. On the other hand, the process network may be regarded as an abstraction of a very complex set of interactions without any geographical embedding. For example, the interaction of a multitude of processes that are associated with the uptake of water by fine roots of plants, their passage through the xylem cells, and eventual dis-

[1]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

0043-1397/09/2008WR007279$09.00

charge into the atmosphere through the stomatal openings constitute a process network.

[4] It is proposed that the ideal description of a complex system's state is a process network (that may or may not have a geographical embedding). In other words, the characteristics of a process network describe emergent properties of the system resulting from forcing and feedback couplings. Changes in the feedback strength and rearrangement of coupling may characterize a shift in the dynamic regime [*Kumar*, 2007; *Foley et al.*, 2003; *Folke et al.*, 2004] of the system. This proposed process network approach to ecohydrological analysis has conceptual precedents in other fields of research; for example, *Gather et al.* [2002] represents human medical states as characteristic patterns of connection between patients' various vital signs, *Ma and Bohnert* [2007] represents each Arabidopsis phenotypical state as a network pattern of connection between genes using microarray data, and *Percha et al.* [2005] represents seizures as a pathologically synchronized state of the brain's neural network.

[5] To study ecohydrological systems, a variety of data are being collected from a natural laboratory perspective, that is, uncontrolled experiments in the natural environment, providing an opportunity to uncover hitherto unknown couplings. Since the variables measured have differing dimensionalities (for example, rainfall (volume per unit time), solar radiation (energy per unit area per unit time), net ecosystem exchange (mass per unit area per unit time)) and the coupling strength, and the space-time lag of influence may be different in different directions of the coupling, new methods are needed to statistically identify these dependencies from the data. Given the diversity and incompleteness of observational data available for complex ecohydrological systems, a new approach for the delineation of process networks is needed, which can maximize the use of all available data to describe the system state, without a prior knowledge of how the associated variables are related. That is, an approach is needed that can robustly identify feedback, both strength and direction, from the time series measurements of observable system variables.

[6] There are many methods that can be used to extract properties of coupling between different variables, such as Granger causality, partial correlations [*Opgen-Rhein and Strimmer*, 2007], Bayesian inference [*Knuth*, 2005], directional nonlinear modeling [*Veeramani et al.*, 2004], and Gaussian models [*Markowetz and Spang*, 2007]. Given time series measurements of two variables, traditional correlation-based methods are capable of identifying the time scale of linear couplings between them. However, they are not able to capture nonlinear relationships or unambiguously identify the directionality of the coupling if they are asymmetric, that is if variable X influences variable Y at one time scale while variable Y influences variable X at a different time scale. These limitations prevent the use of correlation-based techniques for the robust analysis of complex systems where feedback is important.

[7] Information entropy–based methods [*Shannon*, 1948] have emerged as alternate powerful tools that overcome these limitations [*Schreiber*, 2000]. They are attractive because of their basis in the theory of statistical information which is becoming increasingly popular because of its deep unity with nonequilibrium statistical thermodynamics [*Nicolis and Prigogine*, 1989], Bayesian inference and predictability [*Knuth*, 2005], and the observer-relative information contained in patterns [*Heylighen*, 2001; *Roederer*, 2005]. Information entropy is also particularly attractive for the analysis of ecohydrological systems, where nonlinearity due to feedback at the subdaily time scale is a well-recognized but poorly understood phenomenon [*Baldocchi et al.*, 2001a; *Katul et al.*, 2007b]. *Baldocchi et al.* [2001a] were able to explain 73% of the variance of carbon dioxide flux using three periodic modes: the annual (21%), the diurnal (43%), and the semiannual (9%). This leaves nearly a third of the total variability to be explained by the subdaily time scale, where feedback between the ecosystem, land surface, and the atmosphere is known to be important. Therefore, the methods must unambiguously resolve this feedback.

[8] Information flow between components of a system provides a unifying way to deal with the different dimensionality of variables in a system. Information flow is the contribution of uncertainty-reducing or predictive knowledge by one variable to another. An information-theoretic approach is presented which delineates process networks by characterizing the flow of information between pairs of coupled variables using measurements of the associated time series. It is essential that feedback be robustly identified, which in turn requires the establishment of (1) the direction and (2) the lag in time and space of each coupling. Feedback occurs when a circle of directional, lagged couplings links a variable back to itself via the network. If directionality and lag are not established, it is impossible to distinguish the case of synchronization due to one-way forcing from the case of synchronization due to feedback [*Rulkov et al.*, 1995]. The basis of the proposed methods is the transfer entropy statistic introduced by *Schreiber* [2000], which was devised as a tool for the analysis of coupled chaotic systems. Using this statistic the directionality, relative strength, statistical significance, and time lag and scale of information flow that couples two time series data sets is determined. The conjugation of all pairwise couplings allows construction of the process network where the variables constitute the nodes in the process network and links between them represent flow of information, and the two together define the organizational structure and state of the complex system as a whole. The process networks resulting from this method feature flows of information, rather than mass or energy. Statistical tests are developed to ensure that the characterization of information flow from pairs of ecohydrological time series is methodologically robust.

[9] To verify the effectiveness of the methods, the information flow process network is delineated for a system of twelve ecohydrological time series variables for a Midwestern corn-soybean ecosystem, at subdaily time scales. The identified couplings and time scales are evaluated against the prevailing understanding of the system to judge whether the process network thus characterized provides a valuable representation.

[10] The paper is organized as follows. In section 2, the basics of the information theory and statistical methods are laid out and applied to derive process networks. In section 3, the data set used to study a corn-soybean ecohydrological system is explained. In section 4, the results of the analysis are presented, including process networks for the healthy

year, July 2003, and the drought-afflicted ecosystem, July 2005. Here a closer analysis of selected couplings between variables is provided, and a study of the numerical robustness of the analysis is conducted. Section 5 outlines the conclusions and implications of the work. Some technical details of the methods are provided in Appendix A. In part 2 [*Ruddell and Kumar*, 2009], network statistics are derived to compare process networks and information flow on the network in the presence of seasonal and climatological variations and identify emergent patterns.

2. Methods

2.1. Mutual Information and Transfer Entropy

[11] Entropy-based statistics, particularly the mutual information I (also known as the transinformation, or in generalized form as the information redundancy), have been used in hydrology and time series analysis and are a mature methodology for establishing codependency between variables [*Fraser and Swinney*, 1986; *Chapman*, 1986; *Vastano and Swinney*, 1988; *Fraser*, 1989; *Sharma et al.*, 2000; *Mays et al.*, 2002; *Molini et al.*, 2005; *Khan et al.*, 2006]. The transfer entropy T [*Schreiber*, 2000; *Kaiser and Schreiber*, 2002] is a useful variation on the conditional mutual information [*Cover and Thomas*, 2006] that is applicable to time series data sets. Transfer entropy has been applied in the study of financial time series [*Marschinski and Kantz*, 2002], communication within the brain's neural network [*Singer et al.*, 2006], fault propagation in chemical plants [*Bauer et al.*, 2004], and deformation and damage in structures [*Nichols*, 2006]. The transfer entropy has not been applied in practice to study directional coupling and feedback in environmental time series data, a problem which poses a number of additional challenges including (1) coupling of dimensionally dissimilar variables, such as carbon fluxes and rainfall; (2) feedback effects in a time series dominated by periodic forcings; (3) coupling of processes at multiple scales; and (4) small data sizes. Here, methods are developed to cope with these challenges, and their methodological limitations are examined as described below.

[12] Let $X_t = \{x_t\}_{t=1,2,\ldots,n}$ be a discrete time series with marginal pdf $p(x_t)$. Two time series are denoted as X_t and Y_t, or equivalently as $X_t^{(i)}$ and $X_t^{(j)}$ where i and j are the indices of variables X_t and Y_t. Assume that X_t takes m discrete values or that continuous values can be binned into m discrete partitions. The Shannon entropy $H(X_t) = -\sum p(x_t)\log p(x_t)$ is bounded as $0 \le H(X_t) \le \log(m)$. While the entropy is a measure of uncertainty, the mutual information measures the reduction in uncertainty due to the knowledge of another variable. It is obtained as

$$I(X_t, Y_t) = \sum_{x_t, y_t} p(x_t, y_t) \log \frac{p(x_t, y_t)}{p(x_t)p(y_t)}, \tag{1}$$

where x_t and y_t are values assumed by time series X_t and Y_t. It is the Kullback-Leibler distance (KLD) [*Cover and Thomas*, 2006], between the joint distribution $p(x_t,y_t)$ and the product $p(x_t)p(y_t)$ of the marginal distributions. A large value of I implies that X_t and Y_t share a lot of information. The mutual information is symmetric with respect to X_t and Y_t and is bounded as $0 \le I(X_t,Y_t) \le \min[H(X_t), H(Y_t)]$. Since it measures a reduction in uncertainty of Y_t due to knowledge of X_t, if no uncertainty in Y_t exists, no mutual information can exist between X_t and Y_t.

[13] *Schreiber* [2000] defined transfer entropy from X_t to Y_t as

$$T^S(X_t > Y_t) = \sum_{y_t, y_t^{[k]}, x_t^{[l]}} p\left(y_t, y_t^{[k]}, x_t^{[l]}\right) \log \frac{p\left(y_t \middle| \left(y_t^{[k]}, x_t^{[l]}\right)\right)}{p\left(y_t \middle| y_t^{[k]}\right)}, \tag{2}$$

where $x_t^{[l]}$ and $y_t^{[k]}$ are the immediate history of X_t and Y_t of "block length" l and k, respectively. T measures the reduction in the uncertainty of the current state of Y_t that is gained from the l length history of X_t that is not present in the k length history of Y_t itself. In other words "it measures a distance from the hypothesis that the dynamics of Y_t can be described entirely by its own past and no information is gained by considering the dynamics of X_t" [*Nichols*, 2006].

[14] If it is assumed that no other variable's information flow to Y_t overlaps with that of X_t to Y_t (pairwise decomposability or independence), and that the history block length used captures all relevant information (completeness), then T^S may be interpreted as measuring a "flow" of information from X_t to Y_t. This flow of information may be interpreted as the presence of a process coupling from X_t to Y_t. The lack of flow of information means either that the two variables are not coupled, or that their states are so closely synchronized that X_t provides no additional information about Y_t.

[15] In practice, the history order is estimated as the first local minimum of the time-lagged mutual information series that compares X_t or Y_t to itself at various lags [*Kantz and Schreiber*, 2000, p. 132]. However, it is difficult to apply this formulation to observed data sets of finite size, because its data requirements grow exponentially with the block length l and k. *Schreiber* [2000] and others [*Marschinski and Kantz*, 2002; *Sabesan et al.*, 2003] have therefore used synthetic time series data sets of length greater than 10,000 examples in their studies. When studying real-world systems, data lengths are usually much shorter. Even if many data points are available, a researcher will want to comparatively study subsets of the data. The issue of estimation is discussed in detail in section A1, where it is shown that 10–20 bins and 500–1000 data points are generally sufficient to obtain a qualitatively robust estimate of the transfer entropy using time series data.

[16] One may consider variants on this basic theme and describe transfer entropy more generally as

$$T^G(X_t > Y_t, \tau, k, l, \omega) = \sum_{y_t, y_{t-\omega\Delta t}^{[k]}, x_t^{[l]}} p\left(y_t, y_{t-\omega\Delta t}^{[k]}, x_{t-\tau\Delta t}^{[l]}\right) \cdot \log \frac{p\left(y_t \middle| \left(y_{t-\omega\Delta t}^{[k]}, x_{t-\tau\Delta t}^{[l]}\right)\right)}{p\left(y_t \middle| y_{t-\omega\Delta t}^{[k]}\right)}. \tag{3}$$

In the above the lag parameter ω accounts for the situation that the k length history of Y_t that provides the most information about Y_t may not be its immediate history at $\omega = 1$ but is located $\omega > 1$ time steps earlier (time steps are in units of Δt or dt). Similarly the time lag τ, allows consideration of l length history of X_t at a distance τ that provides additional information over and above that is contained in Y_t's own history.

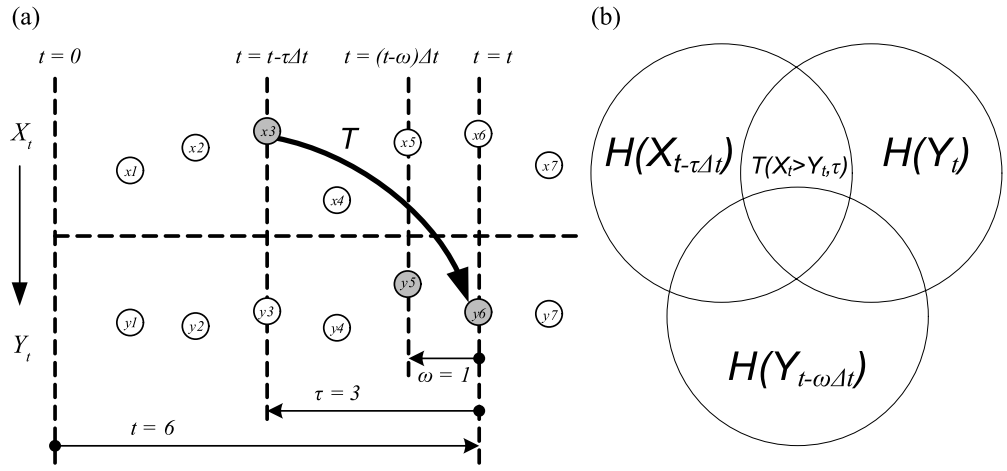

Figure 1. (a) The time lag scheme used to compute the transfer entropy. (b) Venn diagram illustrating that the information flow from time series variable X_t to Y_t, $T(X_t > Y_t,\tau)$ is equivalent to reduced uncertainty in variable Y_t, where uncertainty is measured by the Shannon entropy of the sink variable, $H(Y_t)$. Transfer entropy $T(X_t > Y_t,\tau)$ is the time-lagged mutual information of X_t and Y_t conditioned on the history of variable Y_t so that all information contributed by X_t to Y_t is additional to the information contained in Y_t's own history.

[17] *Marschinski and Kantz* [2002] and *Sabesan et al.* [2003] follow *Schreiber* [2000] in assuming X_t's block length $l = 1$ [*Sabesan et al.*, 2003]. The assumption is conservatively biased (it underestimates transfer entropy) because it neglects information contributed by X_t to Y_t at block lengths $l > 1$. It is also possible to assume that Y_t's block length $k = 1$, using similar arguments. This is a less conservative assumption, because it fails to remove information shared by X_t and Y_t that is contained in Y_t's history at block lengths of $k > 1$. However, this assumption allows minimization of the data requirements and computational demands of the methods, which renders the problem tractable [*Bauer et al.*, 2004; *Nichols*, 2006].

[18] Using $l = k = 1$ makes it simple to treat each time lagged state of X_t as an independent contributor of information to Y_t. The most conservative choice for ω should be the single time lagged history value of Y_t which contributes the most information to the current state of Y_t. In a Markov process $\omega = 1$ because the immediate history always contributes the most information. In this work it is assumed that $l = k = \omega = 1$ and τ as an unknown to be determined (Figure 1a). That is, transfer entropy is estimated at a lag τ at which transfer of information takes place from X_t to Y_t in comparison to the single point immediate history of Y_t,

$$T(X_t > Y_t,\tau) = \sum_{y_t, y_{t-\Delta t}, x_{t-\tau\Delta t}} p\left(y_t, y_{t-\Delta t}, x_{t-\tau\Delta t}\right) \cdot \log \frac{p\left(y_t \middle| \left(y_{t-\Delta t}, x_{t-\tau\Delta t}\right)\right)}{p\left(y_t \middle| y_{t-\Delta t}\right)}. \quad (4)$$

In this paper this transfer entropy is computed from the component Shannon entropies using the form given by *Knuth et al.* [2005] (see Figure 1b),

$$T(X_t > Y_t,\tau) = H(X_{t-\tau\Delta t}, Y_{t-\Delta t}) + H(Y_t, Y_{t-\Delta t}) - H(Y_{t-\Delta t}) - H(X_{t-\tau\Delta t}, Y_t, Y_{t-\Delta t}). \quad (5)$$

It is important to note that transfer entropy is asymmetric, that is $T(X_t > Y_t,\tau) \neq T(Y_t > X_t,\tau)$, and the two transfers may occur at different lags. This property is exploited to characterize both the magnitude and the corresponding lag of bidirectional information transfer between the two time series.

[19] Entropy, mutual information, and transfer entropy can be normalized with respect to the maximum possible entropy H of a distribution where all states are equally probable, i.e., $H = \log(m)$. This normalization eliminates differences in entropy that are caused simply by the number of bins used for discretization or the resolution of the partition, and renders a metric as a fraction of possible entropy or information from zero to one. The resulting normalized metrics are denoted with an apostrophe, as H', I', and T'.

[20] Because these entropy metrics are computed on the basis of estimated marginal and joint probability distributions, the accurate estimation of those distributions is critical to the robustness of the results. In this paper a fixed-interval partition scheme is used for probability density estimation, owing to its simplicity, well-understood limitations, and favorable data requirements. It is necessary to verify that the bin-counting scheme that is sufficiently detailed, and that enough data is available to accurately estimate I and T. The details of density estimation and a validation of this approach are contained in section A1.

[21] Although mathematical methods based on time lags are employed, it is appropriate and accurate to use the language of time scale as a more intuitive substitute for the language of time lag. For transfer entropy estimation, lag and scale manifest themselves in similar ways. This assertion is based on a detailed investigation of the relationship between time lag and time scale which is presented in section A2.

2.2. Significance Test and Periodic Noise

[22] To establish the statistical significance of measured couplings, that is, whether or not a coupling is significantly

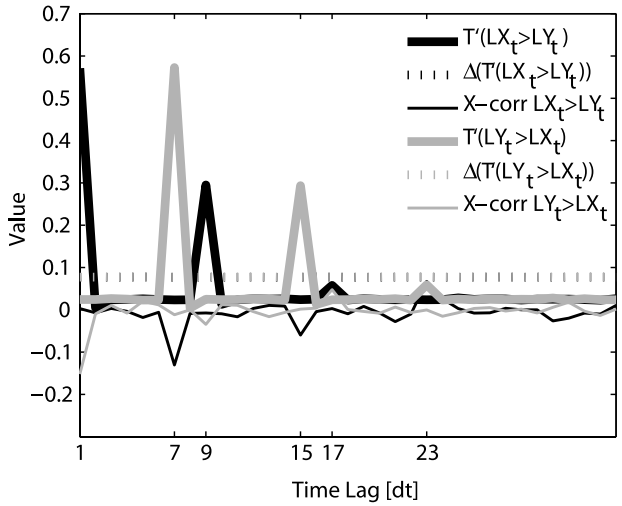

Figure 2. Results for a synthetic system of two logistic maps bidirectionally coupled at a time lag of 1 for the $X_t > Y_t$ coupling and a time lag of 7 for the $Y_t > X_t$ coupling. The linear cross correlation (X-corr) correctly identifies time lags of coupling, but no directionality may be identified. Using T', the lag, directionality, and statistical significance of the coupling is correctly assessed against the significance threshold $\Delta(T')$. Significant coupling time lags are clearly separated from the insignificant. Feedback echoes appear at additive time lags, correctly indicating a repeated cycling of predictive information through the coupled system. The two significance thresholds take the same value in this example.

stronger than that which would occur through random chance between unrelated time series, the method of shuffled surrogates is used, following the example of *Kantz and Schurmann* [1996], *Marschinski and Kantz* [2002], *Sabesan et al.* [2003], and *Nichols* [2006]. To estimate the shuffled-surrogate transfer entropy T_{ss}, the values of time series X_t and Y_t are shuffled randomly in time to destroy time correlations between them, forming new time series' X_{ss} and Y_{ss}. Surrogate transfer entropy $T_{ss}(X_{ss} > Y_{ss})$ is computed for several realizations using Monte Carlo simulations resulting in a Gaussian distribution of surrogates with mean $\mu(T_{ss})$ and standard deviation $\sigma(T_{ss})$. A one-tailed hypothesis test is applied to determine whether $T(X_t > Y_t)$ exceeds $T_{ss}(X_{ss} > Y_{ss})$ at c standard deviations above the mean where c corresponds to $(1 - \alpha) \times 100\%$ level of confidence ($c = 2.36$ for $\alpha = 0.01$, and $c = 1.66$ for $\alpha = 0.05$). The null hypothesis is rejected if $T > (\mu(T_{ss}) + c \cdot \sigma(T_{ss}))$. The critical value that must be exceeded to reject the null hypothesis is termed the "significance threshold" or Δ.

[23] The above method is demonstrated using a coupled logistic map LX_t and LY_t obtained as

$$LX_t = a \cdot LY_{(t-l_{yx})} \cdot \left(1 - LY_{(t-l_{yx})}\right) \tag{6}$$

$$LY_t = a \cdot LX_{(t-l_{xy})} \cdot \left(1 - LX_{(t-l_{xy})}\right). \tag{7}$$

The coupling lags are $l_{xy} = 1$ and $l_{yx} = 7$, and $a = 3.99$ which is within the chaotic range of $3.57 < a < 4$ of the logistic map. T' with 99% confidence thresholds and the corresponding linear cross correlation (X-corr) are plotted for this coupled system of logistic maps for time lags ranging from 0 to 36, using 10000 synthetic data points in Figure 2. Linear cross correlations identify the relevant lags, but cannot identify directionality. Furthermore, the suggestion of a negative correlation is misleading in the presence of this nonlinear coupling. T' succeeds in correctly identifying lags and directions, and in separating lags with significant coupling strength from lags with insignificant coupling strengths. Echoes in the T' time lag plot occur at additive feedback lags, and identify the circular flow of information which repeatedly cycles from one variable to the other.

[24] In environmental data, often very interesting couplings are buried inside dominant periodic drivers such as diurnal, seasonal or annual cycles. To understand the performance of the methods to identify such couplings the coupled logistic time series is corrupted with sinusoidal signal for different signal to noise (SNR) ratio, defined as the ratio of the standard deviation of the logistic time series to the standard deviation of the sinusoidal time series. The results shown in Figure 3 indicate that the method is able to correctly identify the time lags of significant information transfer in the presence of periodic contamination as long as the signal-to-noise ratio is on the order of 1 or greater. Above SNR = 1, the variance of the time series anomaly is more than 50% explained by the signal, as opposed to the noise. The effective SNR can be boosted by filtering out the periodic signal from the data. A periodic anomaly is used to filter out the diurnal cycle in the FLUXNET data as described in section 3.

2.3. Characterizing Process Coupling

[25] A large variety of useful transformations and normalizations have been applied to I and T to produce derivative metrics, such as partial I used by *Sharma et al.* [2000], effective transfer entropy (ETE) and relative explanation added (REA) used by *Marschinski and Kantz* [2002], $H(X)$-relative I used by *Mays et al.* [2002], time lag averaged net T used by *Bauer et al.* [2004], and a normalization $\hat{c}$ used by *Singer et al.* [2006]. It is possible to normalize H, I, and T by each other or any of their component entropies, to render them relative to significance thresholds, and to recombine them in any number of ways. Caution is in order for the use of all transformed metrics for two reasons. First, information values that fall below the significance threshold cannot be considered statistically meaningful for the establishment of a significant coupling regardless of the transformation employed. Second, relative quantities can be deceptive if the normalizing denominator quantity is correlated with the numerator I or T, as is the case when I and T are normalized by some value of H.

[26] A particularly useful combination is the ratio of transfer entropy to the zero-lag mutual information, which will be termed the synchronization ratio Tz, where

$$Tz(X_t > Y_t, \tau) = \frac{T(X_t > Y_t, \tau)}{I(X_t, Y_t)}. \tag{8}$$

Tz measures the transfer of information from $X_{t-\tau}$ ($\tau > 0$) to Y_t, as compared with the shared information at the zero lag. This ratio enables characterization of the nature of coupling

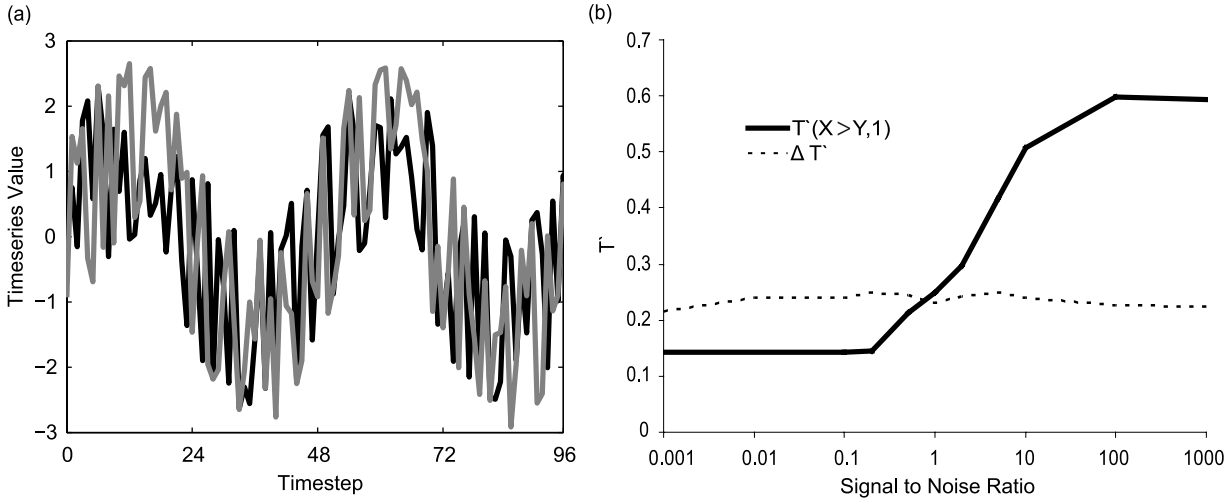

Figure 3. Assessment of the sensitivity of transfer entropy to time-lagged process couplings in the presence of confounding periodic noise. (a) Example of periodic noise added to a pair of coupled logistic maps for a specific signal-to-noise ratio (SNR); in this case, SNR = 1, and the two lines represent each of the sine-corrupted coupled logistic maps. (b) T' at the process coupling time lag of 1 (the correct lag) for the coupled logistic map, and a 99% confidence threshold, plotted against a range of SNRs. Transfer entropy is able to distinguish significant time-lagged process couplings in the presence of confounding periodic noise for SNR $\geq$ 1.

between the dynamics identified through the time series. Four types of couplings between X_t and Y_t can be identified using the ratio Tz (Figure 4 and Table 1). In the list below, the language of synchronization is used as an intuitive approximation for the concept of mutual information, and the language of forcing to approximate the concept of transfer entropy. These couplings occur in pairs between each pair of variables, such that the coupling in one direction takes one type and time scale, and the coupling in the other direction takes an independent type or time scale.

[27] Type 1, synchronization-dominated coupling: If X_t and Y_t are highly synchronized with each other, that is,

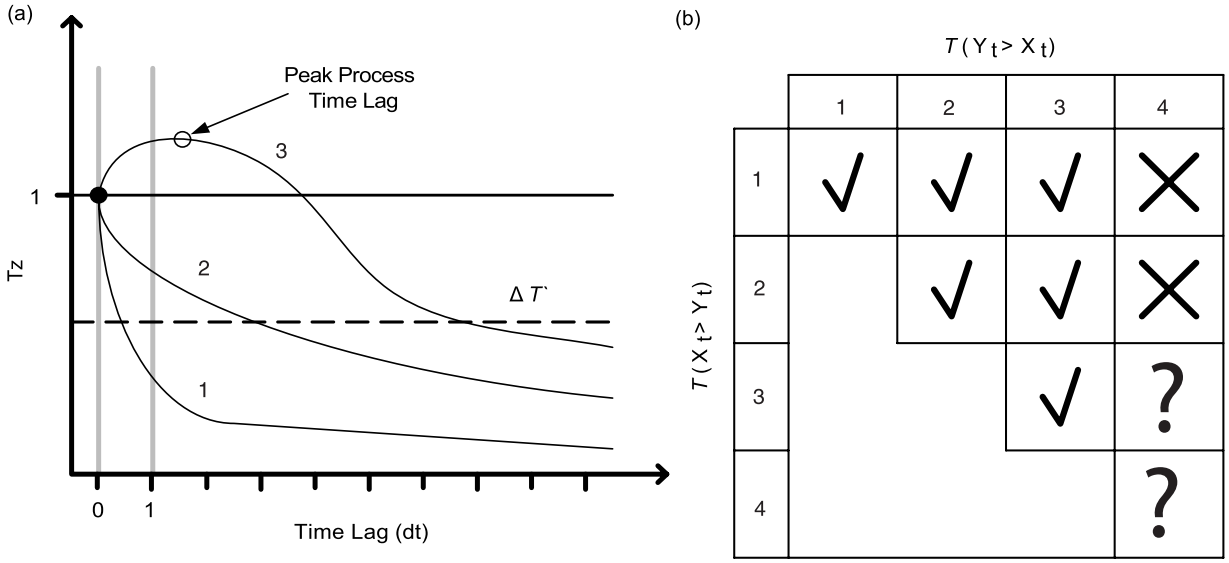

Figure 4. (a) Conceptual sketch giving examples of three types of coupling relationships from one variable to another. Plotted is the line relating Tz to time lag for three hypothetical coupling relationships. Tz is the ratio of T to the zero-lag I, in combination with a statistical significance threshold of information transfer (Δ). $T = I$ when the time lag is zero, so $Tz = 1$. After lag zero, information flow T either increases or decreases relative to I and eventually drops below $Tz = 1$ and $Tz = \Delta$. The peak process time lag is that lag at which the most information is transferred. (b) Conceptual matrix illustrating the six pairwise combinations of canonical couplings identified in the process network for July 2003 and July 2005 (check marks), two possible canonical coupling pairs not observed in these networks (question marks), and two canonical coupling pairs that are not possible (crosses). Matrix is symmetric.

Table 1. Logical Criterion for Coupling Type Classification[a]

	$T > \Delta(T)$	$I > \Delta(I)$	$T > I$	Description
Type 1, synchronization dominated	F	T	F	significant shared information, no significant information flow
Type 2, feedback dominated	T	T	F	significant information flow greater than significant shared information
Type 3, forcing dominated	T	–	T	significant shared information greater than significant information flow
Type 4, uncoupled	F	F	–	no significant information flow or shared information; decoupled

[a]T is true; F is false.

interacting strongly with no time lag, but no significant information flow occurs from X_t to Y_t, it is expected that $T(X_t > Y_t, \tau) < \Delta(T)$, $I(X_t, Y_t) > \Delta(I)$, and consequently $Tz(X_t > Y_t, \tau) < 1$. This is a synchronization-dominated coupling.

[28] Type 2, feedback-dominated coupling: If there is a significant information flow from X_t's history to the current value of Y_t, but this flow is smaller than the shared information at zero time lag, it is expected that $Tz(X_t > Y_t, \tau) < 1$, $T(X_t > Y_t, \tau) > \Delta(T)$, and $I(X_t, Y_t) > \Delta(I)$. X_t and Y_t are closely related, but some information flow still occurs. Type 2 is a sort of middle ground between the synchronization-dominated type 1 and the forcing-dominated type 2, where substantial forcing and synchronization both exist. If type 2 couplings exist in both directions between a pair of variables, this type of information feedback may indicate self-organizing behavior.

[29] Type 3, forcing-dominated coupling: If information flow from X_t to Y_t is significant and larger than the information they share at zero time lag, it is expected that $Tz(X_t > Y_t, \tau) > 1$ and $T(X_t > Y_t, \tau) > \Delta(T)$. This is a forcing-dominated coupling.

[30] Type 4, coupling (uncoupled): if there is no significant information flow or shared information, X_t and Y_t are decoupled and it is expected that $T(X_t > Y_t, \tau) < \Delta(T)$, and $I(X_t, Y_t) < \Delta(I)$.

2.4. Constructing a Process Network

[31] Given a number of variables, it is possible to construct a process network by casting each variable as a node in the network and computing the information flow between each pair of nodes. A network adjacency matrix **A** is constructed where each cell indicates whether there is a directional coupling between two nodes. **A** is a matrix of dimensions $[n_V \times n_V \times n_\tau]$ such that n_V is the number of variables and $n\tau$ is the number of time lags studied. The direction of this information flow coupling always runs from the source to the sink node; in our notation for $T(X > Y)$, X is the index of the source node and Y is the index of the sink node. There are three types of adjacency matrices: "weighted," where each cell contains the coupling strength weight measured by some entropy metric (such as T', I', Tz), "unweighted," where each cell contains a one or zero to indicate the presence (above Δ) or absence (below Δ) of a coupling, and "weighted cut," where cells with no coupling contain a zero but cells with significant coupling contain the value of the weight [*Wilhelm and Hollunder*, 2007].

[32] An adjacency matrix may be constructed for every time lag. To reduce the dimensionality of the process network adjacency matrix from three dimensions $[n_V \times n_V \times n_\tau]$ to a more manageable two dimensions $[n_V \times n_V]$, a rule must be applied to identify one time scale as the "characteristic" time lag, τ', of each coupling. At least two simple approaches exist: the first significant time lag may be chosen, or the statistically significant time lag τ_{max} where T' is strongest may be chosen. It should be noted that either approach will fail if multiple distinct coupling time lags exist. In this paper the first significant local peak time lag is chosen as the characteristic time lag. The two-dimensional weighted-cut adjacency matrix **ATz**(i,j) used to construct the process network in this paper is computed using Tz at the peak time lag of T', $\tau' = \tau_{max}$, as

$$ATz(i,j) = \begin{cases} Tz\left(X_t^{(i)} > X_t^{(j)}, \tau\prime\right) & if \quad T > \Delta T \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where $X_t^{(i)}$ and $X_t^{(j)}$ are the time series corresponding to the ith and jth variable. The corresponding time lag τ' is captured in the matrix $\mathbf{\Gamma}(i,j)$ and the mutual information percentage, $100 \times I'(X_t^{(i)}, X_t^{(j)})$ and the relative mutual information percentage, $100 \times I(X_t^{(i)}, X_t^{(j)})/H(X_t^{(j)})$ are captured in the weighted-cut matrices **AI**(i,j) and **AIr**(i,j), respectively. The adjacency matrices **ATz** and time-lag matrix $\mathbf{\Gamma}$ will allow understanding of the structural organization of the coupling between the variables. This framework is applied to the results in section 4.

3. Study Site and Data Description

[33] The methods described in the previous section are applied to develop a process network of interactions between variables measured at an eddy flux tower. Data from a FLUXNET site [*Baldocchi et al.*, 2001b] located in Bondville, Illinois, USA [*Hollinger et al.*, 2005; *Meyers*, 2008] is used. The Bondville site is located in the central Corn Belt ecoregion which is dominated by more than 90% corn or soybean land cover. The Bondville site comprises a 10 m eddy covariance flux tower and monitors no-till corn (peak canopy height 3 m) in odd years and no-till soybeans (peak canopy height 0.9 m) in even years since August 1996, with a relatively continuous high-quality record from 1998 to the present.

[34] The radiative, meteorological, soil, and eddy flux data are sampled at a resolution of 1 to 10 Hz, averaged at a data set resolution r = 30 min resolution, processed for quality, and then transformed into derivative variables for hundreds of sites across the globe [*Baldocchi et al.*, 2001b] (data available at the Oak Ridge National Lab Carbon Dioxide Information Analysis Center (CDIAC), http://cdiac.ornl.gov, and ORNL-DAAC, http://daac.ornl.gov). The best available gap-filled FLUXNET data is used [*Falge et al.*, 2001a, *Falge et al.*, 2001b], which is the level 4 (L4) product. This study uses the L4 marginal distribution sampling filled products rather than neural network filled products [*Reichstein et al.*, 2005].

Table 2. List of Variables

Symbol	Description	Units
R_g	total incoming shortwave radiation	W m^{-2}
Θ_a	air temperature	deg C
VPD	vapor pressure deficit	KPa
Θ_s	soil temperature (surface layer)	deg C
P	precipitation	mm
θ	soil water content (surface layer)	m^3 m^{-3}
γ_H	sensible heat flux	W m^{-2}
γ_{LE}	latent heat flux	W m^{-2}
GER	estimated gross ecosystem respiration	μmol CO_2 m^{-2} s^{-1}
NEE	net ecosystem exchange	μmol CO_2 m^{-2} s^{-1}
GEP	estimated gross ecosystem production	μmol CO_2 m^{-2} s^{-1}
C_F	cloud fraction between 12,000 feet and surface	fraction

[35] The Bondville site is located 2.5 miles south of the KCMI aviation weather station at the Willard Airport in Champaign, Illinois. METAR (acronym roughly translates from French as Aviation Routine Weather Report) data is observed hourly at U.S. airports (more often when weather is changing), and includes a useful measurement of cloud coverage below 12,000 feet (above surface elevation) which can be interpreted as an index for the cloud fraction (data is available from NOAA at http://weather.noaa.gov/weather/metar.shtml). To create a uniform data set, this METAR data is resampled to the same temporal resolution as the L4 flux tower data, by filling each 30 min value as the most recently observed previous cloud cover value in the METAR record. There are seven possible codes for cloud cover, which are mapped to integers from one to seven in the unified data set. In this way the METAR data is combined with the L4 flux tower data to create a twelve variable set (see Table 2) for this study. These variables cover the most important meteorological, hydrological, radiative, and carbon exchange processes that take place in the system at the subdaily time scale, and represent the full set of variables made available in the L4 data. It is possible to examine a large number of additional variables which contribute to the system's dynamics, but the scope of this paper is limited to a minimal adequate set for this initial study of process networks.

[36] Results for July 2003 and 2005 are presented. The year 2003 is a good growing year that set record crop yields for Illinois (USDA National Agricultural Statistics Service, http://www.nass.usda.gov/), due to the absence of drought, flood, or other adverse conditions during the growing season. The month of July is the peak of the annual growing season, when both the corn and soybean plants show maximum carbon and nutrient assimilation [*Hanway*, 1966; *Hanway and Thompson*, 1967]. It is expected that coupling strength (measured as information flow and feedback) between variables should be the strongest when the ecosystem is actively growing and the magnitude of its interactions with the climate are strongest. Living systems function through the production, utilization, and communication of information, and ecosystems and organisms of higher-order produce the most information [*Roederer*, 2005; *Jorgensen et al.*, 2007, pp. 107, 155]. This is true both at the biochemical level, and also at the ecosystem level where the complexity of the ecological network and the total quantity of information flowing on the network is a measure of its "ascendancy" or developmental maturity. For example, *Jorgensen et al.* [2007, p. 158] explain how ecosystems follow a developmental curve on both the annual and life cycle time scales, in which the information flow and connectedness in the system increase to a peak when the system reaches developmental maturity, and then fall off again as the system declines. If the analogy is valid between this systems approach to ecology and the information flow process network approach, we should expect the most feedback and information production by the ecosystem when the ecosystem is at the peak of its developmental cycle (in midsummer). We will see that this expectation is verified for the Bondville system. The derived process network should therefore feature feedback between vegetation and climate in this healthy July 2003 ecohydrologic system. The process network thus derived will describe the "normal" state of the system under very healthy peak growing conditions. This healthy system is analyzed in close detail in this paper, and then used as a baseline example for a more extensive multiyear study in part 2 of the paper.

[37] The year 2005 was afflicted by an unusually severe drought, which caused a small but economically damaging reduction in crop yields in Illinois [*Gu et al.*, 2006; *Kunkel et al.*, 2006; *Zhang et al.*, 2006]. NOAA's 6-month standardized precipitation index (negative numbers indicate lower than normal precipitation, http://www.ncdc.noaa.gov/oa/climate/research/prelim/drought/spi.html) for July 2005 is -1.2, compared with 0.48 for July 2003. It is expected that the process network can clearly characterize the differences between these two states in terms of the presence or absence of key couplings and feedback.

[38] To understand the subtle patterns of coupling between variables, the diurnal cycle is removed and the anomaly signal thus obtained is analyzed. The anomaly signal is obtained by taking the difference between the values of a variable at a specific time of the day from the average value over the following 5 days of the same variable at the same time. Examples of the transformed anomaly data are given in Figure 5, for latent heat (γ_{LE}) and net ecosystem exchange (NEE). The practical effect of this transformation is that events which produce a variation are emphasized over patterns which recur. These anomaly patterns will generally tend to emphasize daytime events over nighttime events, to the extent that the magnitude of changes in variables during the day is greater than the magnitude of those at night. However, any significant event (such as a weather front) which impacts the system at night is also resolved. Only subdaily time scales of interaction are studied in this paper.

[39] The 1488 half-hour data points for July are reduced by $48 \times 4 = 192$ data points when the 5 day periodic anomaly is taken, and by additional 36 data points to account for time lags of 18 h. The L4 FLUXNET data and METAR cloud data do have some gaps, but where gaps do exist they are discarded. If any of the twelve variables at a given time step is discarded, the values of all twelve variables at that time step are discarded. After preprocessing, the number of data points used to generate a typical month's statistics varies from roughly 800 to 1250 data points, depending on the month, year, and time lag used. July 2003 and 2005 have over 1200 data points. The

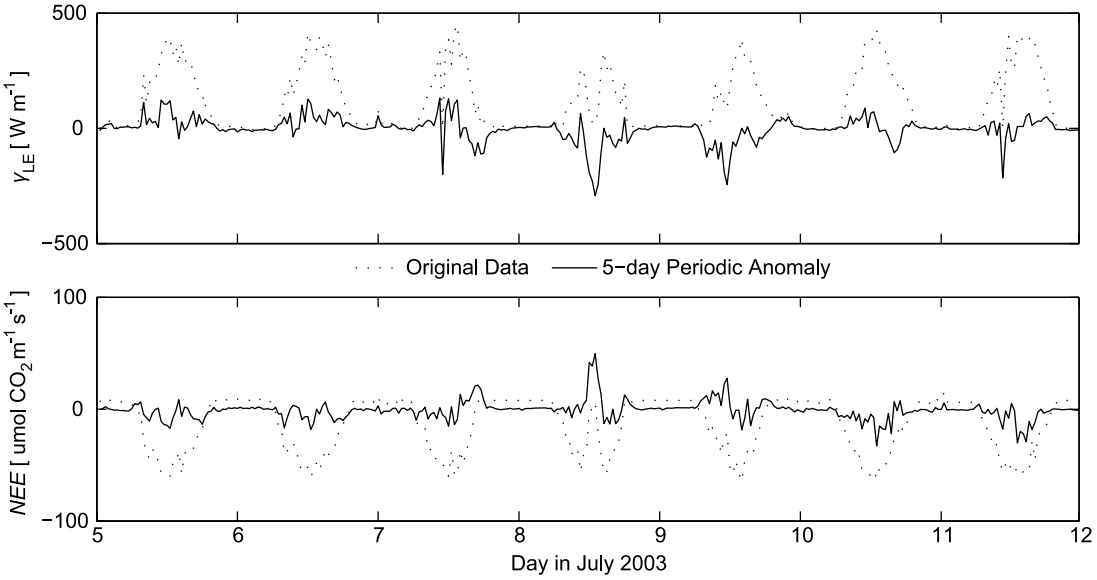

Figure 5. Illustration of the original data time series and transformed 5 day periodic anomaly time series for γ_{LE} and NEE variables. The anomaly removes the diurnal cycle, rendering each variable as a departure from the norm. During the fourth and fifth day of this week, an event occurs which causes a below-normal γ_{LE} anomaly and an above-normal NEE anomaly to occur. Statistics computed using anomaly transformed data are sensitive to departures from the normal pattern and are therefore more sensitive to the coupling processes that caused the departure from the normal pattern.

resolution of the method is equal to the length of time covered by the data set analyzed (one month in this case), because the system state is averaged across this length of time. It is difficult to analyze system states at a resolution finer than one month, when 30 min data is being used, because the sample size will become too small. This approach to the computation of transfer entropy requires at least 500–1000 data (see section A1).

4. Results

[40] In section 2.3 four types of coupling between variables are described. Since information may flow asymmetrically both in magnitude and temporal lag between any pair of variables, there are a number of ways in which two variables may interact with each other. These canonical couplings are described in section 4.1 below and an interpretive framework is developed. These are illustrated with examples derived from the observation from the study site. The goal of the process network approach is to go beyond pairwise couplings, and define the state of the system as a hierarchical pattern of coupling and feedback between subsystems. To do this, it is necessary to identify the directionality and time scale of all pairwise couplings in the system, and then to assemble a holistic picture of system structure using a hierarchy of subsystems logically constructed from these pairwise couplings. The synthesis of a process network using the conjugation of pairwise couplings for the study site is described in section 4.2. In section 4.3 the process network for a healthy and drought stressed corn-soybean ecosystem are compared to illustrate the ability of the process network concept to distinguish between the coupling and feedback structure of the two contrasting states.

4.1. Canonical Coupling Patterns

[41] The adjacency matrix for the 12 variables (Table 2) results in 132 potential pairwise couplings, out of which 60 (or 46%) were found to be statistically significant such that $T' > \Delta(T')$ at one or more time lags. Six typical combinations or "canonical coupling types" are observed in the July 2003 network, representing each of the three types of couplings (types 1–3) outlined in section 2.3, taken in pairs.

[42] Figure 6 illustrates examples of each canonical coupling type. In the list below, when referring to time scales, "short" means time scales close to the data set resolution r (which is 30min in this case), "very short" means time scales much finer than the data set resolution, and "long" means time scales coarser than the data set resolution.

[43] 1. In coupling type 1–1 (Figure 6a), air temperature Θ_s and ecosystem carbon respiration GER are highly synchronized (I' = 21%) at very short time scales, but are so similar that there is no significant transfer of information between them at lagged time scales. These variables are effectively fully synchronized at the 30 min time scale resolved by the data. Soil microbial respiration is known to be dominated by soil temperature which governs microbial metabolism [*Stoy et al.*, 2007], and the L4 algorithm which computes GER is also based in part on the soil temperature [*Reichstein et al.*, 2005]. Consequently, Figure 6a shows a pattern that is expected in the L4 data set.

[44] 2. In coupling type 1–2 (Figure 6b), latent heat flux γ_{LE} and vapor pressure deficit VPD are partially synchronized (I' = 10%) at very short time scales, but there is a significant one-way (unidirectional) flow of information from VPD to γ_{LE}. Their synchronization is partially due to forcing of γ_{LE} by VPD at the <30 min time scale. However, because mutual information is larger than transfer

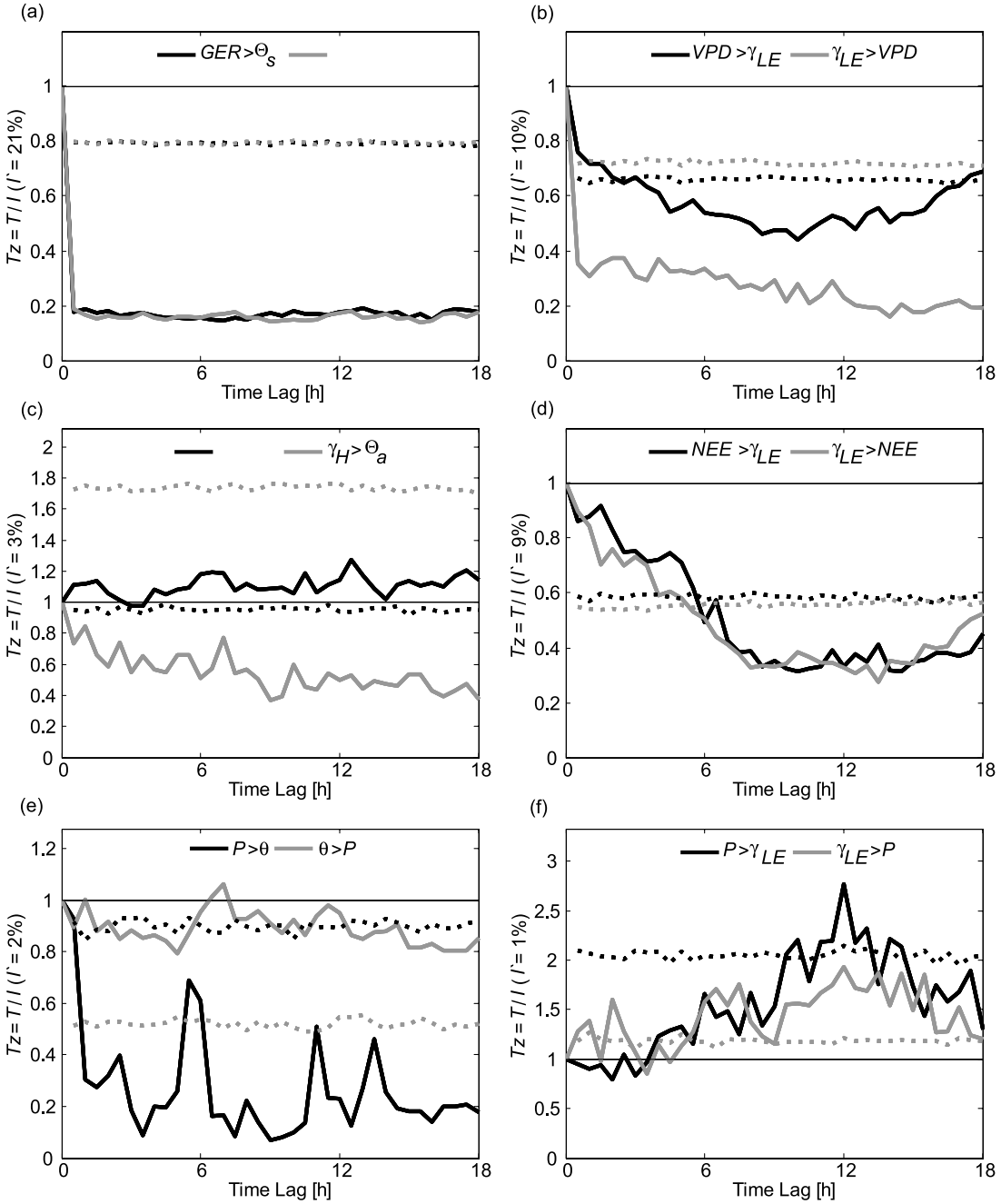

Figure 6. Selected *Tz* lag plots for July 2003; each demonstrates one of the six observed canonical types of coupling. Variable abbreviations are given in Table 2. The normalizing quantity I' is listed on the vertical axis; larger I' makes a large *Tz* more meaningful. Dotted lines are significance thresholds Δ, above which information flow coupling is statistically significant.

entropy, this is a synchronization-dominated information flow relationship since $Tz < 1$. Vapor pressure deficit affects latent heat flux by determining the ability of the air to hold moisture [*Katul et al.*, 2007a].

[45] 3. In coupling type 1–3 (Figure 6c), air temperature Θ_a and latent heat flux γ_H are not very synchronized ($I' = 3\%$) at very short time scales, but there is a significant unidirectional flow of information from Θ_a to γ_H which is substantially larger than the mutual information. Air Temperature is the primary driver of sensible heat because of its gradient near the land surface [*Baldocchi et al.*, 2001a] and its effect persists for a long time, resulting in a type 3 coupling. Consequently, the impact of the synoptic weather event is effective at the 30 min time scale, but its effects persist across all synoptic time scales. This is a forcing-dominated information flow relationship since $Tz > 1$. However, the sensible heat flux does not alter the air temperature, and is dissipated through turbulent mixing at the short time scale resulting in a type 1 coupling.

[46] 4. In coupling type 2–2 (Figure 6d), net ecosystem carbon exchange NEE and latent heat flux γ_{LE} are partially synchronized ($I' = 9\%$) at very short time scales, but they

also transfer significant information to each other (feedback) at short time scales. These variables have a synchronization-dominated relationship since $Tz < 1$, driven by information feedback at short time scales. Type 2–2 relationships indicate that these variables form a self-organizing subsystem, where substantial synchronization occurs because of significant information feedback at short time scales. Ecosystem carbon flux is dominated in July by plant photosynthetic processes, which comes at the cost of substantial transpiration of water through the open leaf stomata [*Farquhar and Sharkey*, 1982; *Stoy et al.*, 2006].

[47] 5. In coupling type 2–3 (Figure 6e), precipitation P and soil water content θ are not very synchronized ($I' = 2\%$) at very short time scales, but there is significant feedback information flow. The information flow is asymmetric, with the type 2 flow from P to θ occurring at short time scales, and the type 3 flow from θ to P occurring at larger scales. This is an uncommon hybrid type of feedback relationship which can only exist when both the mutual information and transfer entropy are weakly significant. Because feedback exists, weak self-organizing behavior may occur in the relationship. Precipitation is the foremost determinant of soil moisture (this happens on short time scales as rain saturates the soil), but soil moisture is also known to impact precipitation over longer subdaily time scales by modifying the land surface energy balance (γ_H and γ_{LE}) and, in turn, the atmospheric boundary layer processes that govern cloud formation and convective rainfall [*Juang et al.*, 2007a, 2007b].

[48] 6. In coupling type 3–3 (Figure 6f), precipitation P and latent heat flux γ_{LE} are not very synchronized ($I' = 1\%$) at very short time scales, but there is significant feedback information flow of the third type centered on a time scale of 10–14 h. These variables have a forcing-dominated relationship since $Tz > 1$. Because feedback exists, weak self-organizing behavior may occur in the relationship. In the summer, rainfall and latent heat flux are coupled at subdaily time scales via atmospheric boundary layer processes [*Juang et al.*, 2007a, 2007b].

[49] Two additional coupling types, 3-4 and 4-4, are theoretically possible but are not observed in this data set; these couplings represent a one-way coupling without significant synchronization (3-4) and a complete decoupling (4-4) between the pair of variables, respectively. Coupling types 1–4 and 2-4 are logically impossible, because the mutual information is a symmetric quantity.

[50] Type 1 couplings are indistinguishable from perfect synchronization at the time scales examined. This might happen because the two nodes are actually measuring the same process in different ways, or it might happen because their coupling processes operate at such short or long time scales that time-lagged dependencies in the variables' perturbations are not discernable, given the finite resolution r of this data set and finite range of τ lags analyzed. Type 2 couplings may indicate that the coupling process operates at short time scales near or slightly below the resolution of the data (30 min), resulting in the appearance that the two variables are near a state of full synchronization but still exchanging weakly significant information flows (Contrast this with type 1 couplings, where coupling processes operate at time scales far shorter than the resolution of the data). Accordingly, type 2 couplings will peak at short time scales. Type 3 couplings occur when the two nodes are coupled by processes at the exact time scale observed, τ. The difference between feedback-coupled type 2–2 and feedback-coupled type 3–3 relationships may be an artifact of the data set resolution; 3-3 couplings will appear as 2-2 if the resolution is coarsened until it cannot resolve the process, and may appear as 1-1 couplings if the resolution is coarsened farther still. This relativism is appropriate, given the hierarchical and irreducible nature of all complex systems [*Haigh*, 1987; *Capra*, 1996].

[51] This ecohydrological system is embedded within a periodic and stochastic forcing structure, including the diurnal and annual cycles and the synoptic weather patterns, which regularly "reset" the ecohydrological system's quasi-stable daytime state by forcing it into an alternate state such as nighttime. By characterizing the system this way, the language of nonlinear self-organizing systems is employed [*Holling*, 1973; *Nicolis and Prigogine*, 1989]. Feedback-driven self-organized structures require time for iterative feedback to produce synchronization between the parts; this synchronization is what characterizes a quasi-stable system state. On the basis of these findings, the authors believe that these ecohydrological variables should be conceptually modeled as coupled chaotic oscillators, of which coupled lasers are an idealized physical example [*Haken*, 1988; *Kanter et al.*, 2007]. The peak value of T indicates the time scale of synchronization. The time scale of synchronization is proportional to the time scale of feedback in coupled synchronizing systems like this one; the time scale of synchronization must be much smaller than the time scale of the forcing which resets the system, or the variables that participate in the feedback will not achieve synchronization [*Rulkov et al.*, 1995; *Kurths et al.*, 2003, Kanter et al. 2007].

[52] The canonical type 2–3 and 3–3 couplings may reveal the presence of this sort of self-organizing structure; although information flow feedback is coupling these variables (P, θ, γ_{LE}), the feedback time scale (6–24 h) is too slow to produce substantial synchronization (measured as I') before the setting sun (24 h scale) or a weather front (3–5 day scale) pushes the system into a different state. From this perspective, type 2–2 couplings have a feedback time scale (1 h) short enough to produce substantial synchronization, but long enough for the process coupling to be captured by 30 min data. Type 1–1 couplings may have a feedback time scale so fast it is not discernable at the 30 min resolution, but that can only be revealed using higher-resolution data.

4.2. Building and Interpreting the Process Network for July 2003

[53] The goal of our process network systems approach is to go beyond pairwise couplings, and define the state of the system as a hierarchical pattern of coupling and feedback between subsystems. Subsystems are defined as a group of variables which are structurally equivalent, meaning that they share a common role in the larger system structure [*Lorrain and White*, 1971]. However, hierarchy is allowed in the subsystem structure, such that if two subsystems comprising structurally equivalent variables both relate to a third subsystem in the same way, then those two subsystems may be considered structurally equivalent, at the higher level of aggregation [*Haigh*, 1987]. How can process network and subsystems be delineated using Shannon entropy statistics? Subsystems are aggregations of individ-

Table 3a. Network Matrix **AI**(i,j)[a]

	R_g	Θ_a	VPD	Θ_s	P	θ	γ_H	γ_{LE}	GER	NEE	GEP	C_F
R_g	*55.7*	2.5	9.2	2.5	1.3	2.6	5.1	15.4	5.1	11.1	13.2	6.8
Θ_a		*85.2*	10.0	32.8	1.6	15.5	2.9	3.8	24.6	2.1	2.6	4.4
VPD			*72.0*	9.7	1.8	9.0	2.3	9.6	12.9	4.2	5.5	6.2
Θ_s				*91.1*	2.1	17.4	2.5	4.0	20.7	2.7	3.0	4.9
P					*9.8*	2.0	x	x	1.5	x	1.0	3.4
θ						*63.7*	2.4	2.5	20.1	2.5	3.3	7.0
γ_H							*35.5*	3.8	2.9	3.9	3.6	1.9
γ_{LE}								*62.6*	4.7	8.8	9.6	3.5
GER									*89.7*	4.1	4.8	6.7
NEE										*54.9*	33.0	3.2
GEP											*58.3*	3.0
C_F												*86.5*

[a]Matrix show the mutual information between pairs of variables at zero time lag. The cross means $I' < \Delta(I')$. Source variable X index i is in rows; sink variable Y index j is in columns. Matrix is symmetric. Italics indicate matrix diagonal. All values are in percent.

Table 3b. Network Matrix **AIr**(i,j)[a]

	R_g	Θ_a	VPD	Θ_s	P	θ	γ_H	γ_{LE}	GER	NEE	GEP	C_F
R_g	*100.0*	3.0	**12.8**	2.7	**13.4**	4.1	**14.2**	**24.7**	5.7	**20.3**	**22.7**	7.9
Θ_a	4.5	*100.0*	**13.8**	**36.0**	**16.6**	**24.4**	8.2	6.1	**27.4**	3.8	4.4	5.0
VPD	**16.5**	**11.7**	*100.0*	**10.6**	**18.1**	**14.1**	6.6	**15.3**	**14.4**	7.7	9.5	7.1
Θ_s	4.5	**38.5**	**13.4**	*100.0*	**21.7**	**27.3**	7.0	6.4	**23.1**	4.9	5.2	5.6
P	2.4	1.9	2.5	2.3	*100.0*	3.1	x	x	1.6	x	1.7	3.9
θ	4.7	**18.2**	**12.5**	**19.1**	**20.2**	*100.0*	6.7	4.0	**22.4**	4.5	5.7	8.1
γ_H	9.1	3.4	3.2	2.7	x	3.8	*100.0*	6.0	3.3	7.2	6.2	2.2
γ_{LE}	**27.7**	4.5	**13.3**	4.4	x	3.9	1.6	*100.0*	5.3	**16.1**	**16.5**	4.1
GER	9.2	**28.9**	**18.0**	**22.8**	**15.0**	**31.6**	8.3	7.6	*100.0*	7.4	8.2	7.8
NEE	**20.0**	2.4	5.9	3.0	x	3.9	**11.1**	**14.1**	4.5	*100.0*	**56.6**	3.7
GEP	**23.7**	3.0	7.7	3.3	**10.2**	5.2	**10.1**	**15.4**	5.3	**60.1**	*100.0*	3.4
C_F	9.9	4.1	6.9	4.3	**28.2**	8.9	4.3	4.6	6.1	4.8	4.1	*100.0*

[a]Matrix shows the percentage of uncertainty of each Y explained by X. Bold means > 10% of $X_t(j)$ explained by $X_t(i)$. The cross means $I' < \Delta(I')$. Source variable X index i is in rows; sink variable Y index j is in columns. Italics indicate matrix diagonal. All values are in percent.

ual nodes that share similar patterns of coupling type and time scale; they are analogous to the functional groups which ecologists use to describe ecosystem structural networks [*Folke et al.*, 2004]. Nodes in the same subsystem should share type 1 or type 2 synchronization-dominated couplings. Type 3 or type 4 couplings mean that coupled nodes do not belong to the same subsystem. When a network is built using pairwise couplings, the biggest risk is that of identifying false positive couplings that are caused by third-party effects, that is, nodes A and B are independent but driven by a common forcing node C, so A and B will falsely appear to drive each other [*Jorgensen et al.*, 2007]. By grouping nodes into subsystems this confusion can be reduced. The information needed (I', Tz', τ') to make these interpretations is contained in Tables 3a, 3b, 3c, and 3d, which presents the three network matrices explained in section 2.4 for the July 2003 Bondville corn-soybean ecosystem.

[54] First the zero-lag mutual information is presented in Table 3a, which presents I' between pairs of variables. The diagonal of this matrix gives the Shannon entropy of the source variable, $H'(X)$. Diagonal values range from 9.8% for precipitation, which resides in the "zero precipitation" state most of the time, to 91.1% for soil temperature, which resides in all eleven of its discrete states with roughly equal frequency. R_g, γ_H, γ_{LE}, GEP, and NEE have moderate entropy ranging from 35.5% to 62.6%, and Θ_a, VPD, Θ_s, θ, GER, and C_F have higher entropies ranging from 63.7% to 91.1%.

[55] Mutual information values between variables can never exceed $H'(X)$, but frequently approach a relative mutual information of 30%, which means that roughly 30% of the total Shannon entropy of one variable is explained by the other. In Table 3b the relative mutual information is presented, which clarifies which groups of variables are strongly synchronized with each other at very short time scales. The relative mutual information of GEP to NEE is 60.1%, so NEE and GEP are assumed to comprise a single node in the network, since 60.1% of NEE's entropy is explained (where explanation means mutual information) by GEP. This finding confirms that subdaily variability in the net carbon exchange at the peak of the growing season is dominated by photosynthetic activity, rather than respiration. In the same way, Θ_a, Θ_s, GER, VPD, and θ are strong explainers of each other and precipitation, and VPD also predicts γ_{LE}. R_g explains P, γ_H, γ_{LE}, NEE, and GEP. Furthermore, γ_H, γ_{LE}, NEE, and GEP explain much about each other. C_F explains more than 10% of P and of R_g. Nothing explains more than 10% about C_F, and γ_H and P do not explain more than 10% of any other variable.

[56] A careful study of Tables 3a–3d indicates that all variables fall into three cohesive groups, and R_g, γ_{LE}, and VPD are key variables that explain more than 10% in multiple groups. The first major group is the "atmospheric boundary layer" (ABL) group of R_g, C_F, and P, which are associated with ABL formation processes and convective activity [*Margulis and Entekhabi*, 2001; *Juang et al.* 2007a, 2007b; *Katul et al.* 2007a]. The second major group is the "turbulent" group associated with energy budget and photosynthetic processes on the land surface at plant canopy turbulent time scales [*Farquhar and Sharkey*, 1982; *Baldocchi et al.*, 2001a; *Katul et al.*, 2001; *Siqueira et al.*, 2006]. It includes γ_H, γ_{LE}, GEP, and NEE, which are variables

Table 3c. Network Matrix **ATz**(i,j)[a]

	R_g	Θ_a	VPD	Θ_s	P	θ	γ_H	γ_{LE}	GER	NEE	GEP	C_F
R_g	***0.10***	x	x	x	1.25	x	**0.74**	**0.53**	x	**0.73**	**0.63**	x
Θ_a	x	*x*	x	x	1.16	x	1.27	2.06	x	3.38	2.76	x
VPD	x	x	*x*	x	1.33	x	1.40	0.76	x	1.56	1.33	x
Θ_s	x	x	x	*x*	**0.90**	x	1.54	2.17	x	2.93	2.46	x
P	1.44	x	x	x	***0.15***	**0.93**	2.30	2.77	x	2.53	1.89	x
θ	x	x	x	x	1.06	*x*	1.27	2.42	x	2.13	x	x
γ_H	**0.62**	x	x	x	2.41	x	***0.09***	1.16	x	**0.92**	**0.94**	x
γ_{LE}	**0.43**	x	x	x	1.93	x	**0.89**	*0.15*	x	**0.90**	**0.84**	x
GER	x	x	x	x	1.35	x	1.25	1.87	*x*	1.98	1.70	x
NEE	**0.48**	x	x	x	1.83	x	**0.82**	**0.92**	x	***0.14***	**0.22**	x
GEP	**0.42**	x	x	x	1.48	x	**0.96**	**0.88**	x	**0.22**	**0.13**	x
C_F	x	x	x	x	**0.55**	x	1.75	2.16	x	2.21	2.36	x

[a]Matrix shows the ratio of the maximum lag T to mutual information for all significant couplings. Source variable X index i is in rows; sink variable Y index j is in columns. Italics indicate matrix diagonal. The cross means type 1 coupling, and bold means type 2 coupling; otherwise, values are type 3 coupling.

Table 3d. Network Matrix $\Gamma(i,j)$[a]

	R_g	Θ_a	VPD	Θ_s	P	θ	γ_H	γ_{LE}	GER	NEE	GEP	C_F
R_g	*.5–4(5),.5*	x	x	x	.5–15(22),13	x	.5–15(11),.5	.5–5(10),.5	x	.5–5(11),.5	.5–5(11),.5	x
Θ_a	x	*x*	x	x	.5–17(29),.5	x	.5–18(34),12	16–17(2),16	x	6–18(11),16	7–9(3),9	x
VPD	x	x	*x*	x	.5–18(32),5	x	.5–18(8),.5	.5–18(4),.5	x	.5–18(7),.5	.5–3(6),.5	x
Θ_s	x	x	x	*x*	.5–18(32),12	x	.5–18(35),4	16–18(4),18	x	4–17(9),8	4–8(2),4	x
P	12–13(3),12	x	x	x	*.5–13(12),5*	.5–.5(1),.5	10–18(9),11	10–14(5),12	x	12–12(2),12	12–12(1),12	x
θ	x	x	x	x	.5–18(36),7	*x*	.5–18(36),5	17–17(1),17	x	11–11(1),11	x	x
γ_H	.5–.5(1),.5	x	x	x	.5–13(21),8	x	*.5–3(5),.5*	.5–4(6),.5	x	.5–3(7),.5	.5–3(5),.5	x
γ_{LE}	.5–3(5),.5	x	x	x	.5–17(22),12	x	.5–5(9),.5	*.5–4(7),.5*	x	.5–5(10),.5	.5–5(11),.5	x
GER	x	x	x	x	.5–18(36),15	x	4–18(19),17	.5–16(7),2	*x*	.5–17(24),.5	.5–14(11),3	x
NEE	.5–4(7),.5	x	x	x	.5–15(8),6	x	.5–12(8),.5	.5–5(11),.5	x	*.5–5(9),.5*	.5–5(9),.5	x
GEP	.5–3(6),2	x	x	x	.5–17(13),6	x	.5–3(7),.5	.5–5(10),.5	x	.5–5(9),.5	*.5–5(10),.5*	x
C_F	x	x	x	x	.5–18(20),11	x	.5–14(13),9	10–12(4),11	x	5–15(15),12	5–15(12),9	*x*

[a]Matrix shows time lags of significant information flow on the interval [τ = 0.5 h . . .18 h], including the first significant lag, last significant lag, number of significant lags, and peak time lag. Source variable index X is in rows; sink variable index Y is in columns. Significant lag times are [first–last(number),max]. The cross means $T' < \Delta(T')$.

whose information transport was found in section 4.1 to peak at fast <30 min time scales. The third major group is the ''synoptic'' group of Θ_a, Θ_s, GER, VPD, and θ which are known to be strongly linked to the synoptic-scale weather patterns [*Baldocchi et al.*, 2001a; *Katul et al.*, 2001].

[57] By using the canonical couplings and the information in Table 3c, an arrangement of subsystems, information flow, feedback, and time scales can define the July 2003 state of the system. In Table 3c *Tz* is presented for every statistically significant information flow coupling in the network. In Table 3d key time lags are presented for each coupling: the first significant time lag, last significant time lag, number of significant time lags, and the characteristic time lag τ' of the greatest information transfer. The turbulent group variables share type 2 couplings at the short <30 min time scale, and therefore form a type 2 self-organizing subsystem. The synoptic group variables share type 1 couplings, and form a type 1 subsystem. The ABL variables do not form a logically consistent subsystem of a single type and time scale, but are connected by a rich pattern of type 3 couplings at longer time scales up to 18 h.

[58] The resulting process network in Figure 7 exhibits a rich pattern of information flow between three major subsystems. The synoptic subsystem, associated with large-scale weather patterns, forces the ABL and turbulent subsystems via type 3 coupling at all time scales, but the ABL and turbulent subsystems participate in a feedback loop with each other at subdaily time scales. The ABL subsystem exports information to the turbulent subsystem via a type 3 coupling at the 12 h time scale. The ABL subsystem is also a sink of information flow from the turbulent subsystem via connections through P and R_g. The ABL and turbulent subsystems are therefore connected via a feedback loop, forming a larger-scale hierarchical self-organizing subsystem which is termed the ''regional'' subsystem because of the longer subdaily time scales that characterize the feedback loop. The variables and time scales of this feedback loop are consistent with what is known about ABL formation processes [*Juang et al.* 2007a, 2007b; *Margulis and Entekhabi*, 2001; *Katul et al.* 2007a], and this feedback loop identifies the well-known process by which land surface variables affect ABL formation throughout the day, and in turn modify cloud cover, radiation, and convective precipitation later in the day.

[59] GEP participates in the organization of the turbulent subsystem through the control of photosynthesis on stomatal conductance and plant transpiration. Therefore, it follows that local plant photosynthetic processes exert an indirect control on the ABL and in turn on other parts of the land surface ecosystem via network feedback. The marginal entropies H' for the turbulent subsystem are substantially lower than those of the synoptic and ABL subsystems (35–62% for turbulent, versus 55–91% for others). This suggests that this feedback loop is negative, serving to stabilize and moderate the regional land surface ecosystem even as it receives forcing and feedback from subsystems with more variability. This is consistent with the findings of *McNaughton and Jarvis* [1991] that plant ecosystems inhabit negative or stabilizing feedback loops involving the ABL at scales from the leaf scale (very short time scale) to the regional scale (longer subdaily time scale).

[60] These findings are not without their limitations. In particular, it is difficult to unambiguously resolve the structure of the ABL subsystem. The authors believe the logical inconsistencies in the ABL subsystem of the process networks are caused by the spatial nature of information flow and feedback in the ABL (as explained by *McNaughton and Jarvis* [1991] and *Jacobs and De Bruin* 1992]). The use of a spatial data set in a future study could help bring this subsystem into clearer focus. Given that the FLUXNET data is collected at a single point in space, we will have to accept a lack of clarity regarding the ABL and precipitation. *Shuttleworth* [1988] found that the ABL is known to vary at the subdaily time scale with surface conditions at spatial scales on the order of 10 km to 100 km; the ''fetch'' or spatial measurement scale of the Bondville flux tower is much smaller, on the order of 100 m. This experiment effectively substitutes temporal data for spatial data; a review of studies that have taken a similar approach is provided by *Baldocchi* [2008].

4.3. Comparing the Process Network: Healthy Versus Drought States

[61] To compare a second system state with the July 2003 state (Figure 7), a process network is computed for the July

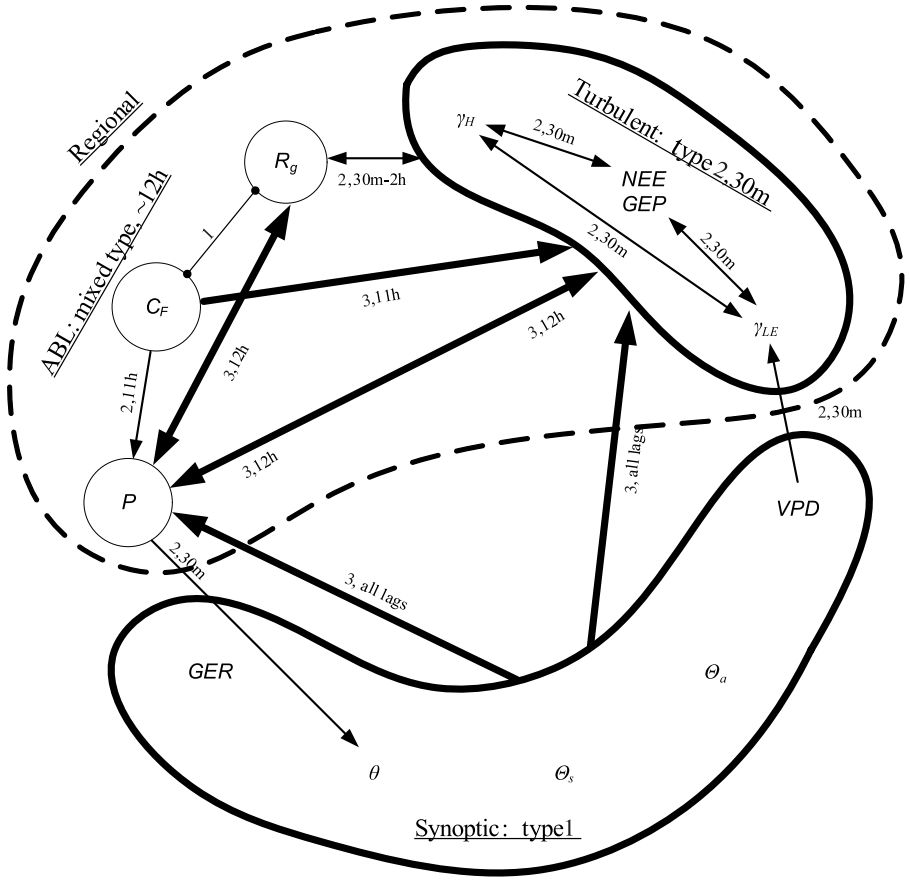

Figure 7. The process network for July 2003, a healthy system state. Types 1, 2, and 3 relationships result in the interpretation of the system as three subsystems linked at time scales ranging from 30 min to 12 h. Thin arrows represent type 2 couplings. Thick arrows represent type 3 couplings. A type 1 "synoptic" subsystem including GER, θ, Θ_s, Θ_a, and VPD forces the other subsystems at all studied time scales from 30 min to 18 h. A type 2 "turbulent" self-organizing subsystem including γ_H, γ_{LE}, NEE, and GEP exists with a feedback time scale of 30 min or less and inhabits a feedback loop with P and R_g at time scales from 30 min to 12 h. The P, C_F, and R_g variables form a loose subsystem of mixed types, which interact with each other on a time scale of roughly 12 h.

2005 drought-afflicted ecohydrological system. The resulting weighted-cut process network for July 2005 is visualized in Figure 8. The first salient observation is that the drought process has fewer couplings than the healthy process network; in fact, roughly half of the couplings disappear during the drought state (adjacency matrix results not shown). Not a single new coupling exists during the drought, which did not occur during healthy conditions. In general, then, the drought state is characterized by decoupling.

[62] The most important decoupling is that of the turbulent type 2 land surface subsystem from the other two subsystems. In the process network, neither the synoptic nor the ABL subsystems are coupled to the turbulent subsystem during drought conditions, with the same type 3 12 h couplings that existed during healthy conditions. Because less information is flowing between the subsystems, the surface energy balance and carbon flux processes are not being organized as strongly by the weather patterns and boundary layer processes. The "engine of variability" that is necessary for the land surface ecohydrological system to thrive [*Kumar*, 2007] appears to be broken down during drought because of insufficient information input from the synoptic weather patterns. The moisture fluxes which carry the information may be reduced below a key threshold during drought.

[63] The absence of information flow from the ABL subsystem to the turbulent subsystem means that the circular type 3 feedback between these two subsystems is broken during drought. The regional self-organizing type 3 subsystem that binds the turbulent and ABL subsystems into a larger hierarchical subsystem at longer time scales around 12 h dissolves because of the disappearance of the feedback that defined this structure during healthy conditions. A physical interpretation of this breakdown is that individual local-scale land surface ecosystems are not able to communicate with each other via the medium of the ABL, or to collectively organize and influence their atmospheric environment on a regional scale. The reduction of information flow and feedback between the turbulent and ABL subsystems on longer "regional" time scales is characteristic of drought.

[64] At one time it was believed that the plant ecosystem is passively forced by climate conditions, but as far back as

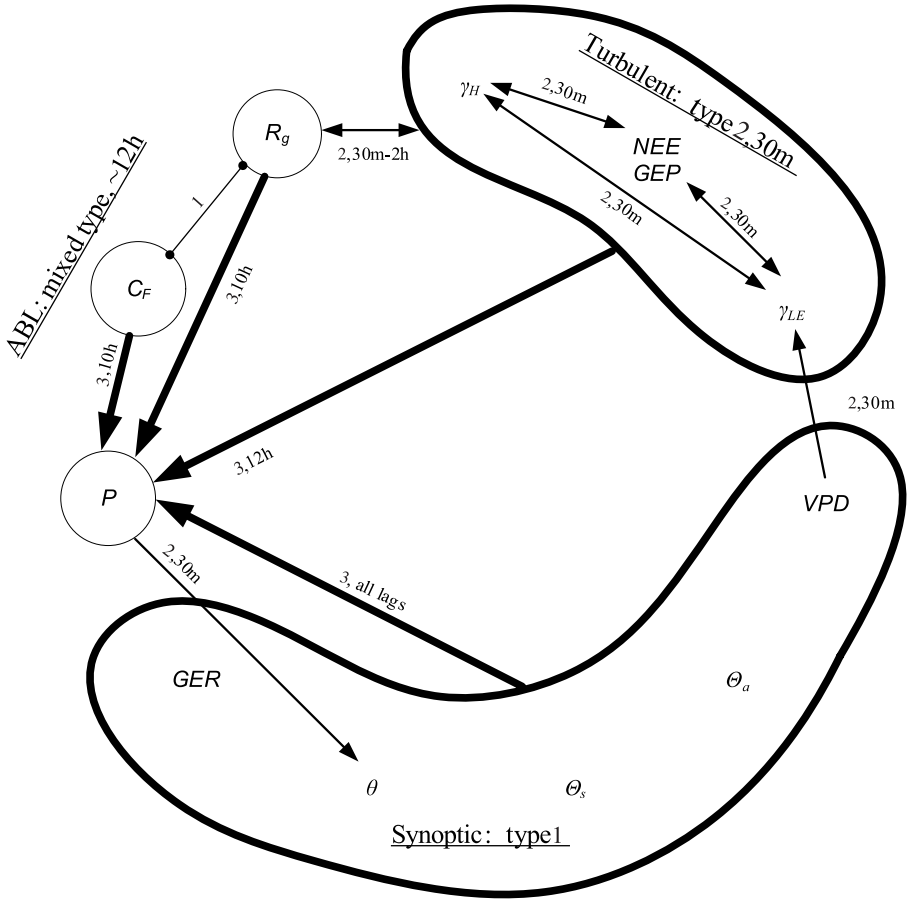

Figure 8. Same as Figure 7 but for July 2005, the peak growing month of a season impacted by severe drought. The same three subsystems exist as in a healthy system state (Figure 7), but fewer couplings exist between the subsystems; the network is substantially decoupled. Most notably, the synoptic subsystem is no longer coupled to the turbulent subsystem via a type 3 coupling, and the turbulent and ABL subsystems no longer participate in circular feedback at time scales up to 18 h. As a consequence, the regional self-organizing subsystem disappears in this drought system state.

1960 it was recognized that "drought begets drought" [*Namias*, 1960; *Monteith*, 1995]. In other words, drought can be understood as a self-organizing phenomenon involving feedback between the land surface and the atmosphere. There is a physical feedback between the land surface and atmospheric boundary layer, such that fewer clouds and precipitation cause insufficient soil moisture, reduced evapotranspiration, and in turn reduce atmospheric humidity, reducing the potential for future precipitation and moisture transport [*McNaughton and Jarvis*, 1991; *Dominguez and Kumar*, 2008a, 2008b]. On the other hand, reduced evapotranspiration results in increased sensible heat flux, which can cause moisture to rise more quickly to the lifting condensation level at which precipitation occurs [*Juang et al.*, 2007b; *Dominguez and Kumar*, 2008a, 2008b]. The feedback resulting from ecosystem can therefore impact the occurrence of convective precipitation [*Freedman et al.*, 2001; *Juang et al.*, 2007a; *Carleton et al.*, 1994].

[65] Although there is an agreement on how to recognize drought (negative ecosystem impacts due to insufficient moisture), there is no agreement on what exactly drought "is," or how to provide a robust definition for it. Some ecosystems apparently never experience drought even during periods of low precipitation because of a lack of moisture stress where soil moisture does not drop below the wilting point [*Jaksic et al.*, 2006; *Rodriguez-Iturbe*, 2000]. The impact of drought varies widely from one location to another, possibly due to plant adaptation and rooting patterns [*Calvet*, 2000]. These studies, especially that of *Carleton et al.* 1994], suggest that summertime drought may be characterized as a specific regime of feedback between the land surface vegetation and the convective precipitation activity in the atmosphere. The results we present define drought as a system state using a process network, where the time scales and subsystems involved in the feedback couplings are precisely established. The methods developed here shed more light into this issue because they are able to detect and characterize the asymmetry of information flow in the network.

5. Discussion

[66] FLUXNET data is derived from eddy covariance measurement towers, which suffer from several types of

error. In addition to the limited precision, random noise, instrument failures, downtime, and calibration errors which impact all physical measurements, measurements of fluxes suffer from unique biases. First, because atmospheric eddies and turbulent mixing occurs over an area of the landscape upwind of the tower (the area is called the ''footprint'' or ''fetch''), changes in wind speed and direction can alter the composition of the footprint [*Hollinger and Richardson*, 2005]. Heterogeneous vegetation types, canopy structures, soil moisture conditions, and obstructions in the vicinity of the tower will cause measurements to vary with the wind regime; in cases of extreme heterogeneity, this can mean that one tower is measuring two different ecohydrological systems, with the time-averaged data set reflecting a mixture of the two. Second, nighttime measurements of flux are notoriously inaccurate, because low wind speeds significantly reduce the area of the footprint, resulting in extreme variability in measurements and skewed (nonnormal) error [*Moffat et al.*, 2007; *Richardson et al.*, 2008]. Snow and ice fouling can affect precipitation and surface heat flux measurements during the winter, resulting in unrealistically low measurements and measurement variability. In addition, since hundreds of different researchers are involved in measuring and transcribing FLUXNET measurements before they reach the archives, it is inevitable that uneven quality control is applied during data collection and processing. This lack of quality control threatens to render data from different sites incomparable.

[67] Fortunately, ''gap filling'' reanalysis methods can address some of these problems, specifically those of (1) with quality control caused by nonhomogeneous field methods and transcription and (2) gaps caused by instrument failure and nighttime and wintertime downtime. The level 4 (L4) data product is such a gap-filled product. Multiple reanalysis models are used to identify unreliable data and replace it with more realistic estimates [*Moffat et al.*, 2007; *Juang et al.*, 2006; *Reichstein et al.*, 2005], resulting in a product that is free of gaps, has more reliable nighttime estimates of flux, is free of extreme outliers, and is generated in a uniform way across all FLUXNET sites. These reanalyzed data are optimized to give accurate daily and annual mean totals, and to reproduce the same spectrum of periodic variability as raw observed data [*Moffat et al.*, 2007], but the data are not optimized to reproduce the time correlation structure between variables. *Richardson et al.* [2008] found that models generally overestimate autocorrelation. If these findings hold for the level 4 data set, it would impact the quality of the results in this paper by causing us to underestimate the strength of information flow, and the data set variability (Shannon entropy) at short time scales. The present methods are robust against random measurement noise in the data and against errors in the magnitude of estimates, but are vulnerable to erroneous time-lagged covariation structures in the data. This sort of error has not been evaluated in FLUXNET data and may impact the results and interpretation presented here.

6. Summary and Conclusions

[68] The central conclusion of this paper is that robust process networks can be delineated from multivariate ecohydrologic time series data using information flow statistics. The flow of statistical information is an effective way to identify the direction and time scale of process couplings and an effective basis for the delineation of process networks in complex systems, when the appropriate methodology is applied. A methodology is developed for the construction and logical interpretation of information flow process networks, and a quality control framework is provided to ensure that entropy statistics are robustly computed from time series data. Process networks are an ideal way to define the state of a complex system, because in these systems the structure of the system as a whole is more important than the variability of its individual parts.

[69] A second conclusion is that mutual information and transfer entropy should be used to complement each other in the analysis of time series data. Mutual information measures the extent of synchronization between variables, but transfer entropy alludes to the cause of the synchronization, where the cause is either forcing or feedback coupling. New questions can now be answered. Are two variables similar because one variable forces the other, or because a third variable forces both, or do they synchronize in a self-organizing fashion by exchanging information? What is the time scale of each coupling process? Do the variables form subsystems, and do those subsystems interact with each other as organized units? The ''synchronization ratio'' of the information flow at the peak time lag to the zero-lag mutual information, Tz, is a fundamental dimensionless quantity that can be used to classify couplings and subsystems. Three types of couplings are logically defined using this ratio. Type 1 couplings are dominated by shared information (synchronization), type 3 couplings are dominated by information flow (forcing), and type 2 couplings represent the middle ground where substantial synchronization and forcing both exist. These three types, taken by pairs of variables, form six combinations; examples and discussion of each of these six ''canonical coupling types'' observed in the July 2003 process network. These canonical types are general to time-varying complex systems, and they should be observable in a wide variety of such systems in nature.

[70] Using a Tz-based network adjacency matrix, the healthy state July 2003 process network is interpreted as an arrangement of three subsystems that interact via types 1, 2, and 3 couplings. The process network is computed for a healthy corn-soybean Midwestern ecosystem at the peak of its growing season. A self-organizing ''turbulent'' subsystem is formed from the sensible heat flux, latent heat flux, net ecosystem exchange of carbon, and gross ecosystem carbon uptake, which share type 2 feedback couplings at the $<$30 min time scale. A ''synoptic'' subsystem is formed from the air temperature, vapor pressure deficit, soil temperature, soil water content, and ecosystem respiration of carbon, which share type 1 couplings at very short time scales. A loosely organized ''atmospheric boundary layer'' subsystem is formed from the incoming shortwave radiation and cloud cover, which share a variety of couplings at longer subdaily time scales up to 18 h. The synoptic subsystem exports significant information to the other two, and serves as a large-scale forcing on the system. The turbulent and atmospheric boundary layer subsystems are coupled via a feedback loop, and form a regional self-organizing subsystem that is a hierarchical aggregate of

those two subsystems. This regional-scale feedback subsystem forms the basis for plant ecosystems to self-organize at the regional scale through collective dynamics, at least during the peak of a healthy growing season. The resulting process network reveals couplings and time scales consistent with what is already established in the literature [*Baldocchi et al.*, 2001a; *Katul et al.*, 2007b; *Margulis and Entekhabi*, 2001], but is able to additionally classify the nature of one-way forcing and circular feedback between subsystems.

[71] The drought state July 2005 process network is constructed and compared with the healthy state July 2003 process network. This comparison reveals that key type 3 couplings to the turbulent subsystem at longer time scales up to 18 h turn off during drought. The result is that the process network is decoupled, and the regional hierarchical self-organizing system dissolves because of the lack of feedback between the ABL and turbulent subsystems. The disconnection of longer–time scale couplings to the turbulent subsystem is characteristic of the drought state. Ecosystem photosynthetic efficiency is known to increase with increased cloud cover because of a substitution of diffuse solar radiation for direct radiation [*Baldocchi*, 2008], and cloud cover generally increases with ecosystem transpiration in humid temperate ecosystems [*Juang et al.*, 2007a, 2007b]. Ecosystem transpiration decreases during drought. Accordingly, we believe that increased self-organization and feedback involving the ecosystem at the regional scale may be an evolved collective (or "emergent" [*Corning*, 2002]) ecosystem behavior that serves to increase photosynthetic efficiency in the regional ecosystem as a whole.

[72] This conclusion provides evidence to support the assertion of *Carleton et al.* [1994, p. 593] that "the vegetation-convective-cloud interaction may be a feature of Mid-west USA summer climate." This collective ecosystem behavior breaks down to a certain extent during drought. If feedback coupling between the various local ecosystems breaks down at the regional scale during drought, one should expect increased heterogeneity on the landscape, such that local ecosystems which are fortunate to receive ample rainfall continue to thrive, but are unable to aid surrounding areas via regional moisture recycling. In fact this increased landscape heterogeneity is exactly what *Carleton et al.* [1994] found, using satellite data: the summertime spatial variation of land cover data sets is higher during drought.

[73] These observations confirm that the process network is able to distinguish between drought and healthy ecosystem states and to provide insight into the nature of the differences between the states. As key couplings turn on and off, the organization and feedback in the system changes as the system enters a new state. Eight logical types of canonical couplings have been identified to classify the type of information flow between variables, six of which are observed in the Bondville July 2003 corn-soybean ecohydrological system. Using these canonical coupling types, a set of measured variables may be sorted into a hierarchy of subsystem groupings. Now that process networks can be constructed to describe the coupling of a number of observed time series variables, these exciting tools may be applied to the new field of complex network theory [*Strogatz*, 2001] to study the properties of the system as a whole. Applications of network theory using process ecohydrological process networks will be explored in part 2 of this paper [*Ruddell and Kumar*, 2009].

Appendix A

A1. Estimation of Entropy Statistics From Data

[74] The estimation of the Shannon entropy (H), mutual information (I), and transfer entropy (T) is dependent on the accurate estimation of probability densities from data. In fact, all statistical information is relative to the observer's choice of the state space in which to embed the dynamics of each observed variable [*Gershenson and Heylighen*, 2003], so it is important that we choose well. To estimate density from a finite discrete data set a number of approaches exist such as function fitting [*Wolpert and Wolf*, 1995; *Knuth et al.*, 2005], kernel estimation [*Scott*, 1979; *Kaiser and Schreiber*, 2002; *Sharma et al.*, 2000; *Sabesan et al.*, 2003; *Bauer et al.*, 2004; *Nichols*, 2006], and binning with fixed mass or fixed interval partitions [*Marschinski and Kantz*, 2002; *Kaiser and Schreiber*, 2002]. A good mathematical review is provided by *Paninski* [2003].

[75] The advantages of fixed interval bin counting are its simplicity, dependence on a single parameter for the number of bins, computational efficiency, well-understood biases in I and T with respect to the density estimation scheme, and its ability to handle both discrete and continuous data (convergence of discrete and continuous results is discussed by *Kontoyiannis and Antos* [2001]). The main downside of fixed interval binning schemes is that this method introduces arbitrary partitions to what is fundamentally a continuous system, resulting in "edge effects" as noted by *Kaiser and Schreiber* [2002], *Knuth et al.* [2005], and K. H. Knuth (Optimal data-based binning for histograms, unpublished manuscript, 2006).

[76] Because of edge effects, bin counting is sensitive to the number of partitions used. With too few partitions, edge effects become severe and H estimates are positively biased, resulting in underestimated I and T. If too many partitions are used, there will not be enough data to accurately estimate every partition's probability, resulting in a positive bias in H and a similar underestimation of I and T.

[77] In theory it is impossible to have too fine a discretization scheme or too many data; more detail and more data is better. However, data requirements grow with the number of bins used. In practice this means that we need to find the smallest number of bins, m, and number of data, N, under which relatively consistent and unbiased estimates of T may be obtained. The quality of the discretization scheme may be evaluated by plotting the entropy metric H, I, or T against the number of bins and data used that a plateau occurs in the vicinity of the chosen number of bins and data.

[78] For the synthetic coupled logistic map (section 2.2), Figures A1a and A1b show that T' is estimated consistently when more than $N = 250$ data are used, and I' requires fewer than 100 data. Ten or more bins appear to be adequate to achieve accurate results, but more bins require more data. A "sweet spot" exists for the coupled logistic map in the vicinity of $m = 10–20$ bins, where the fewest data are required to achieve accurate results.

[79] A similar analysis is produced for couplings between γ_{LE} and NEE (Table 2) at a 30min time lag. First entropy

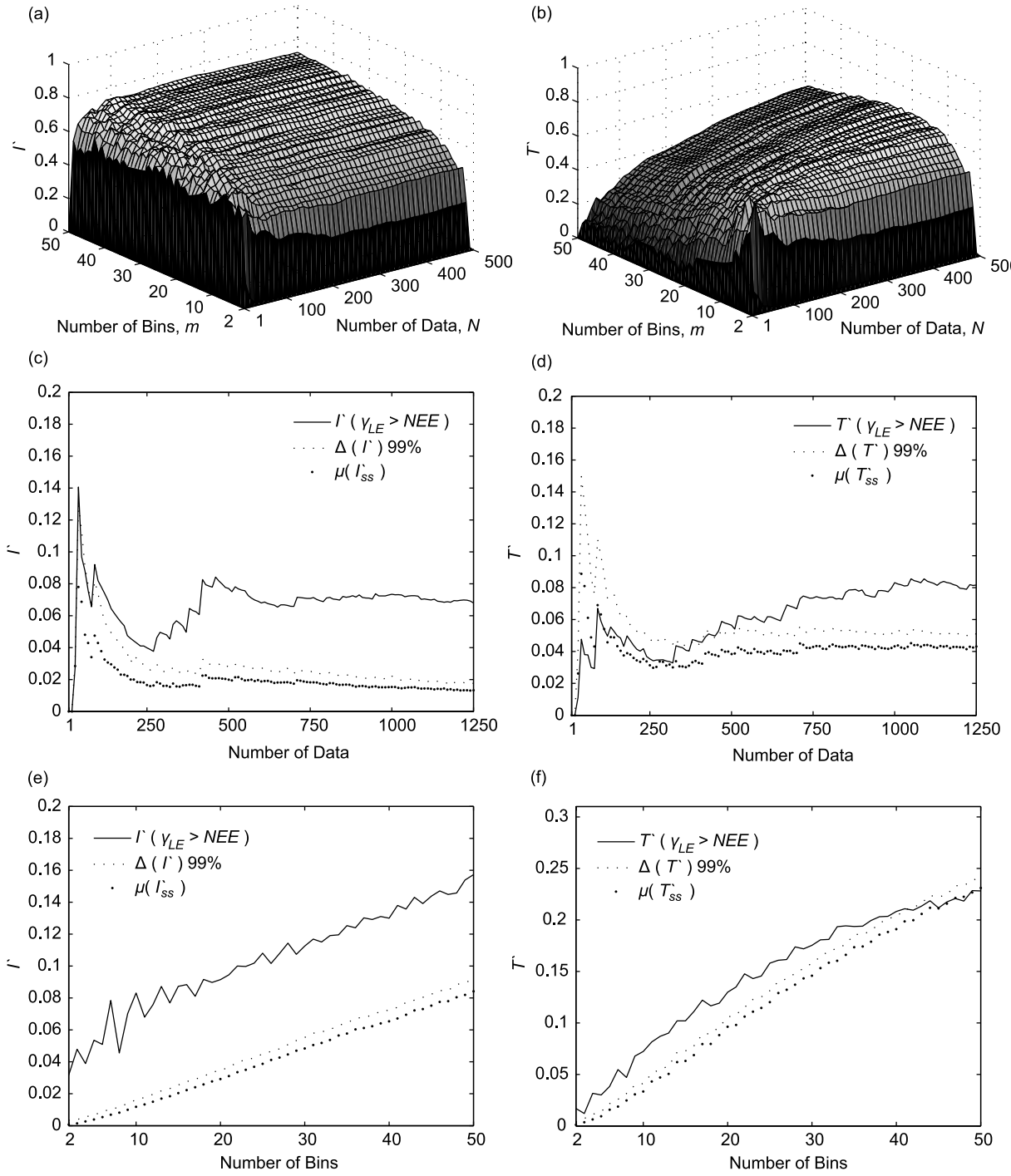

Figure A1. Estimation issues for mutual information and transfer entropy. (a) I' for the coupled logistic map, the estimate plateaus above 10 bins and 50–100 data. (b) T' for the same, the estimate plateaus above 10 bins and 250 data. More bins necessitate more data; a "sweet spot" requiring minimal data exists at 10–20 bins. (c, d) Observed July 2003 γ_{LE} > NEE, cross sections of number of data using 11 bins. I' plateaus at 100 data, but T' requires 500 data. (e, f) Same as Figures A1c and A1d but for number of bins using 1250 data; for T', 35 bins is too fine a resolution for the limited sample size.

statistics are computed for many sample sizes N, holding the bin count constant at m = 11. Figures A1c and A1d show that T' requires more data to achieve a consistent estimate (>500 data points), as compared with I' (>100 data points). This data set's sample size of N = 1250 data points is sufficient to estimate I' and T' when an m = 11 bin distribution is used.

[80] Next, entropy statistics are computed for many bin counts m, holding the sample size constant at N = 1250. Figures A1e and A1f show that m = 50 bins does not appear to be enough to achieve a consistent estimate of I' and T', as indicated by the lack of a plateau in the curves. This is unfortunate, since there is not enough data to satisfy

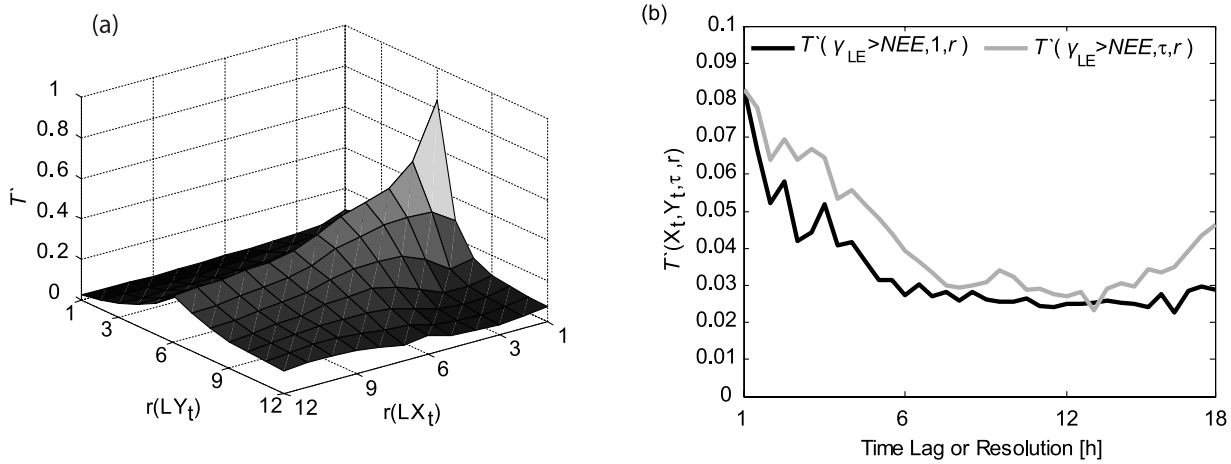

Figure A2. Examination of the relationship between time lag and averaging scale. (a) $T'(LX_t > LY_t)$ for the coupled logistic map computed against source and sink averaging scales $r(LX_t)$ and $r(LY_t)$. The methods identify the correct process averaging scale as a peak in the plot, where source $r(LX_t) = 6$ and sink $r(LY_t) = 1$. (b) $T'(\gamma_{LE} > \text{NEE})$ is computed using two different data sets: first, using an averaging scale $r(\gamma_{LE}) = r(NEE) = 30$ min and a time lag τ up to 18 h and second, using a time lag τ of 30 min and an averaging scale up to 18 h. The two results are similar, proving that the time lag is closely related to the averaging scale of the process.

a binning scheme with hundreds of bins. In fact, there is not enough data to satisfy a binning scheme with more than roughly 35 bins. In Figure A1f, T' attains a consistent difference from its significance threshold in the range of roughly 5 to 35 bins, but above $m = 35$ bins T' drops below the significance threshold, indicating that there is not enough data to populate such a detailed distribution. Fortunately, the $m = 11$ discretization scheme is within that range, so we can be assured of a robust judgment of statistical significance, if not perfect accuracy.

[81] These estimates of I' and T' show a consistent negative bias when 11 bins are used, because both statistics steadily increase as more than 11 bins are used. This means that an improved probability density estimation scheme could improve the accuracy of these results by reducing the bias. However, because the bias is consistent across the network of couplings, and because statistically significant couplings can be qualitatively distinguished from those that are not, it is concluded that this approach provides adequate robustness and accuracy for our purposes.

[82] Bin-counting methods have been found to require on the order of 1000 data points to adequately estimate T [*Bauer et al.*, 2004; *Nichols*, 2005]. However, as few as 150 data points may be required for the ideal case where data is Gaussian distributed, according to *Wolpert and Wolf* [1995] and Knuth (unpublished manuscript, 2006). The authors have found, on the basis of experience, that 10–20 bins give consistent results for I and T using a wide variety of time series data, when at least 500–1000 data are used. Because this experience is consistent with the findings of other authors who have applied transfer entropy [*Knuth et al.* 2005; *Nichols*, 2005], the authors propose that $m = 10$–20 bins and $N > 500$–1000 data is a good rule of thumb for the application of transfer entropy using finite-interval bin-counting probability density estimation schemes.

A2. Time Lag Versus Scale

[83] In a temporal system characterized by processes that operate at many scales, we have been working exclusively with time lags (see section 2.1). It is therefore important to ask the question "what is the relationship between time lag and time scale"? This question can be answered by computing the same statistics using various temporal averaging resolutions instead of time lags.

[84] This is done by modifying the coupled logistic map equations presented in section 2, replacing time lag with time averaging scale, such that LY_t is mapped to the average value of the last six time steps of LX_t, instead of being mapped to the single value at a fixed time length in the past. This produces a coupling process where the averaging scale $r(LX_t)$ of the source node is six, and the averaging scale $r(LY_t)$ of the sink node is one. This synthetic data set approximately represents a coarse-scale temporal process that drives a fine-scale temporal process. Figure A2a, shows that this method correctly identifies $r(LX_t) = 6$, $r(LY_t) = 1$ as the peak process coupling scale or time lag.

[85] To apply this variable-scale approach to a variety of data sets, its application is now generalized. Time series variable X_t is transformed using a standard moving average of window size r, such that the new time series $X_t^r = (X_t + X_{t-1} + \ldots + X_{t-r+1})/r$. When a temporally averaged variable is used to compute I or T, the notation presented in section 2 is extended to include the averaging resolution of the source and sink variables as, $I(X_t,Y_t,\tau,r(LX_t),r(LY_t))$ and $T(X_t,Y_t,\tau, r(LX_t),r(LY_t))$. When $r(LX_t) = r(LY_t)$ the fifth index is omitted.

[86] $T'(\gamma_{LE} > \text{NEE},30\text{min},r)$ is plotted along with $T'(\gamma_{LE} > \text{NEE},\tau,30\text{min})$, where resolution r is equal to the time lag τ, in Figure A2b. This allows comparison of information transfer at a time lag with information transfer at an equivalent averaging scale. When the time lag and resolution are both 30 min, the two estimates are identical. As the

time lag and averaging scale increase, the time-lagged version of T' remains similar to its time-averaged counterpart. It is concluded that time lag and averaging scale are closely related in this coupled system. The interpretation of time lag as a surrogate for scale is therefore justified.

[87] In general, uncertainty and therefore entropy, especially joint entropy between two variables, will decrease as the resolution (scale) of the data becomes coarser. Accordingly, mutual information between variables should increase when scale becomes coarser, as was found for comparisons of observed and modeled hydrological data by *Chapman* [1986] and for scaling in lagged rainfall fields by *Molini et al.* [2005]. The effect of scale on transfer entropy is more subtle; the transfer entropy will remain sensitive to the process coupling scale present in the data, and will therefore be strongest when both variables are averaged at the strongest process coupling scale.

Notation

$\mathbf{\Gamma}(i,j)$ adjacency matrix where indices store τ' values [Δt].

γ_H sensible heat flux [W m^{-2}].

γ_{LE} latent heat flux [W m^{-2}].

$\Delta(I)$, $\Delta(T)$ statistical significance threshold for I or T [bits].

Δt, dt discrete interval of time, the units of time lags and steps [T].

Θ soil water content of the surface layer [m^3 m^{-3}].

Θ_a air temperature [°C or K].

Θ_s soil temperature (surface layer) [°C or K].

τ time lag between variables X_t and Y_t [Δt].

τ' characteristic time lag of the coupling between two variables [Δt].

τ_{max} time lag of maximum information flow between two variables [Δt].

ω number of time lags skipped for variable Y_t's own history [Δt].

$\mathbf{A}(j,i,\tau)$ network adjacency matrix where indices store T' [arbitrary units].

$\mathbf{AI}(i,j)$ adjacency matrix where indices store I' values [fraction].

$\mathbf{AIr}(i,j)$ adjacency matrix where indices store $I(X_t,Y_t)/H(Y_t)$ [fraction].

$\mathbf{Ap}(S, \tau)$ adjacency matrix where indices store $T'/TST_S(\tau)$.

$\mathbf{ATz}(i,j)$ adjacency matrix where indices store Tz values at τ' lags [fraction].

C_F cloud fraction [fraction].

GEP estimated gross ecosystem production [μmol CO$_2$ m^{-2} s^{-1}].

GER estimated gross ecosystem respiration [μmol CO$_2$ m^{-2} s^{-1}].

$\mathbf{H}$ vector storing values of H' for each variable [fraction].

H' normalized Shannon entropy [fraction].

H_s^m mean normalized Shannon entropy of subsystem S [fraction].

$H(X_t)$ or $^{H_{X_t}}$ Shannon entropy of variable X_t [bits].

i, j matrix indices for X and Y [positive integer].

I' normalized mutual information [fraction].

$I(X_t,Y_t)$ mutual information of variables X_t and Y_t [bits].

k, l length of time series history used for variables X_t and Y_t [Δt].

LX_t, LY_t coupled logistic map variables [arbitrary units].

m number of states used to classify the data [positive integer].

N number of data points in the data set [positive integer].

NEE net ecosystem exchange [μmol CO$_2$ m^{-2} s^{-1}].

n_τ number of time lags τ being considered [positive integer].

n_S number of variables in the subsystem S [positive integer].

n_V number of variables in the system V [positive integer].

P precipitation [mm month^{-1}].

$p(x_t)$, $p(y_t)$ marginal probability distribution of variables X_t and Y_t [fraction].

$p(x_t,y_t)$ joint probability distribution of variables X_t and Y_t [fraction].

r resolution of the time series data set [time].

R_g total incoming shortwave radiation [W m^{-2}].

$T(X_t > Y_t, \tau)$ abbreviated version of T^G [bits].

T' normalized transfer entropy [fraction].

$T^G(X_t > Y_t, \tau, k, l, \omega)$ generalized time-lagged transfer entropy from X_t to Y_t [bits].

$T^S(X_t > Y_t)$ Schreiber's original transfer entropy from variable X_t to Y_t [bits].

$T_{ss}(X_{ss} > Y_{ss},\tau)$ surrogate transfer entropy using time-shuffled data X_{ss} and Y_{ss} [bits].

Tz synchronization ratio [ratio].

VPD vapor pressure deficit [kPa].

X_{ss}, Y_{ss} time-shuffled surrogates for X_t and Y_t [units of data].

X, Y source and sink variables, respectively [arbitrary units].

$X_t^{(i)}$, $X_t^{(j)}$ time series versions of X and Y [units of data].

X-corr linear cross correlation [fraction].

[88] **Acknowledgments.** This research is funded by the 2006–2009 NASA Earth Systems Science (ESS) Fellowship Program grant NNX06AF71H and NSF grant ATM 06–28687. The authors would like to acknowledge Kevin Knuth of the University at Albany, State University of New York, for helpful preliminary discussions on transfer entropy and probability-density estimation methodologies and Richard Robertson and Darren Drewry, who provided feedback and support.

References

Baldocchi, D. (2008), 'Breathing' of the terrestrial biosphere: Lessons learned from a global network of carbon dioxide flux measurement systems, *Aust. J. Bot.*, *56*, 1–26, doi:10.1071/BT07151.

Baldocchi, D., E. Falge, and K. Wilson (2001a), A spectral analysis of biosphere-atmosphere trace gas flux densities and meteorological variables across hour to multi-year scales, *Agric. For. Meteorol.*, *107*, 1–27, doi:10.1016/S0168-1923(00)00228-8.

Baldocchi, D., et al. (2001b), FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities, *Bull. Am. Meteorol. Soc.*, *82*, 2415–2434, doi:10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2.

Bauer, M., N. F. Thornhill, and A. Meaburn (2004), Specifying the directionality of fault propagation paths using transfer entropy, paper presented at 7th International Symposium on Dynamics and Control of Process Systems, Int. Fed. of Autom. Control, Cambridge, Mass., 5–7 July.

Calvet, J. C. (2000), Investigating soil and atmospheric plant water stress using physiological and micrometeorological data, *Agric. For. Meteorol.*, *103*, 229–247, doi:10.1016/S0168-1923(00)00130-1.

Capra, F. (1996), *The Web of Life*, Random House, New York.

Carleton, A. M., D. Travis, D. Arnold, R. Brinegar, D. E. Jelinski, and D. R. Easterling (1994), Climatic-scale vegetation-cloud interactions during drought using satellite data, *Int. J. Climatol.*, *14*(6), 593–623, doi:10.1002/joc.3370140602.

Chapman, T. G. (1986), Entropy as a measure of hydrologic data uncertainty and model performance, *J. Hydrol.*, *85*, 111–126, doi:10.1016/0022-1694(86)90079-X.

Corning, P. A. (2002), The re-emergence of "emergence": A venerable concept in search of a theory, *Complexity*, *7*(6), 18–30, doi:10.1002/cplx.10043.

Cover, T. M., and J. A. Thomas (2006), *Elements of Information Theory*, John Wiley, Hoboken, N. J.

Dominguez, F., and P. Kumar (2008a), Precipitation recycling variability and ecoclimatological stability—A study using NARR data. Part I: Central U.S. Plains ecoregion, *J. Clim.*, *21*, 5165–5186, doi:10.1175/2008JCLI1756.1.

Dominguez, F., and P. Kumar (2008b), Precipitation recycling variability and ecoclimatological stability—A study using NARR data. Part II: North American monsoon region, *J. Clim.*, *21*, 5187–5203, doi:10.1175/2008JCLI1760.1.

Falge, E., et al. (2001a), Gap filling strategies for long term energy flux data sets, *Agric. For. Meteorol.*, *107*, 71–77, doi:10.1016/S0168-1923(00)00235-5.

Falge, E., et al. (2001b), Gap filling strategies for defensible annual sums of net ecosystem exchange, *Agric. For. Meteorol.*, *107*, 43–69, doi:10.1016/S0168-1923(00)00225-2.

Farquhar, G. D., and T. D. Sharkey (1982), Stomatal conductance and photosynthesis, *Annu. Rev. Plant Physiol.*, *33*, 317–345, doi:10.1146/annurev.pp.33.060182.001533.

Foley, J. A., M. T. Coe, M. Scheffer, and G. Wang (2003), Regime shifts in the Sahara and Sahel: Interactions between ecological and climatic systems in northern Africa, *Ecosystems*, *6*, 524–539, doi:10.1007/s10021-002-0227-0.

Folke, C., S. Carpenter, B. Walker, M. Scheffer, T. Elmqvist, L. Gunderson, and C. S. Holling (2004), Regime shifts, resilience, and biodiversity in ecosystem management, *Annu. Rev. Ecol. Evol. Syst.*, *35*, 557–581, doi:10.1146/annurev.ecolsys.35.021103.105711.

Fraser, A. M. (1989), Information and entropy in strange attractors, *IEEE Trans. Inf. Theory*, *35*, 245–262, doi:10.1109/18.32121.

Fraser, A. M., and H. L. Swinney (1986), Independent coordinates for strange attractors from mutual information, *Phys. Rev. A*, *33*, 1134–1140, doi:10.1103/PhysRevA.33.1134.

Freedman, J. M., D. R. Fitzjarrald, K. E. Moore, and R. K. Sakai (2001), Boundary layer clouds and vegetation-atmosphere feedbacks, *J. Clim.*, *14*, 180–197, doi:10.1175/1520-0442(2001)013<0180:BLCAVA>2.0.CO;2.

Gather, U., F. Roland, I. Michael, and B. Claudia (2002), Patterns of dependencies in dynamic multivariate data, in *Pattern Detection and Discovery*, edited by D. J. Hand et al., pp. 214–226, Springer, Heidelberg, Germany.

Gershenson, C., and F. Heylighen (2003), When can we call a system self-organizing?, in *Advances in Artificial Life: European Conference, ECAL 2003*, *Lecture Notes Artificial Intell.*, vol. 2801, edited by W. Banzhaf et al., pp. 606–614, Springer, Berlin.

Gu, L., T. Meyers, S. G. Pallardy, P. J. Hanson, B. Yang, M. Heuer, K. P. Hosman, J. S. Riggs, D. Sluss, and S. D. Wullschleger (2006), Direct and indirect effects of atmospheric conditions and soil moisture on surface energy partitioning revealed by a prolonged drought at a temperate forest site, *J. Geophys. Res.*, *111*, D16102, doi:10.1029/2006JD007161.

Haigh, M. J. (1987), The Holon: Hierarchy theory and landscape research, in *Geomorphological Model—Theoretical and Empirical Aspects*, edited by F. Ahnert, *Catena Suppl.*, *10*, 181–192.

Haken, H. (1988), *Information and Self-Organization*, Springer, Berlin.

Hanway, J. J. (1966), How a corn plant develops, *Spec. Rep. 48*, Coop. Ext. Serv., Iowa State Univ., Ames, Iowa.

Hanway, J. J., and H. E. Thompson (1967), How a soybean plant develops, *Spec. Rep. 53*, Coop. Ext. Serv., Iowa State Univ., Ames, Iowa.

Heylighen, F. (2001), Cybernetics and second-order cybernetics, in *Encyclopedia of Physical Science and Technology*, 3rd ed., edited by R. A. Meyers, pp. 155–170, Academic, San Diego, Calif.

Holling, C. S. (1973), Resilience and stability of ecological systems, *Annu. Rev. Ecol. Syst.*, *4*, 1–23, doi:10.1146/annurev.es.04.110173.000245.

Hollinger, D. Y., and A. D. Richardson (2005), Uncertainty in eddy covariance measurements and its application to physiological models, *Tree Physiol.*, *25*, 873–885.

Hollinger, S. E., C. J. Bernacchi, and T. P. Meyers (2005), Carbon budget of mature no-till ecosystem in north central region of the United States, *Agric. For. Meteorol.*, *130*, 59–69, doi:10.1016/j.agrformet.2005.01.005.

Hubler, A. W. (2005), Predicting complex systems with a holistic approach, *Complexity*, *10*(3), 11–16, doi:10.1002/cplx.20071.

Jacobs, C. M., and H. A. R. De Bruin (1992), The sensitivity of regional transpiration to land-surface characteristics: Significance of feedback, *J. Clim.*, *5*, 683–698, doi:10.1175/1520-0442(1992)005<0683:TSORTT>2.0.CO;2.

Jaksic, V., G. Kiely, J. Albertson, R. Oren, G. Katul, P. Leahy, and K. A. Byrne (2006), Net ecosystem exchange of grassland in contrasting wet and dry years, *Agric. For. Meteorol.*, *139*, 323–334, doi:10.1016/j.agrformet.2006.07.009.

Jorgensen, S. E., B. D. Fath, S. Bastianoni, J. C. Marques, F. Muller, S. N. Nielsen, B. C. Patten, E. Tiezzi, and R. E. Ulanowicz (2007), *A New Ecology: Systems Perspective*, Elsevier, Amsterdam.

Juang, J.-Y., G. G. Katul, M. B. S. Siqueira, P. C. Stoy, S. Palmroth, H. R. McCarthy, H.-S. Kim, and R. Oren (2006), Modeling nighttime ecosystem respiration from measured CO2 concentration and air temperature profiles using inverse methods, *J. Geophys. Res.*, *111*, D08S05, doi:10.1029/2005JD005976.

Juang, J. Y., G. G. Katul, A. Porporato, P. C. Stoy, M. S. Siqueira, M. Detto, H. Kim, and R. Oren (2007a), Eco-hydrological controls on summertime convective rainfall triggers, *Global Change Biol.*, *13*, 1–10, doi:10.1111/j.1365-2486.2006.01315.x.

Juang, J.-Y., A. Porporato, P. C. Stoy, M. S. Siqueira, A. C. Oishi, M. Detto, H.-S. Kim, and G. G. Katul (2007b), Hydrologic and atmospheric controls on initiation of convective precipitation events, *Water Resour. Res.*, *43*, W03421, doi:10.1029/2006WR004954.

Kaiser, A., and T. Schreiber (2002), Information transfer in continuous processes, *Physica D*, *166*, 43–62, doi:10.1016/S0167-2789(02)00432-3.

Kanter, I., N. Gross, E. Klein, E. Kopelowitz, P. Yoskovits, L. Khaykovich, W. Kinzel, and M. Rosenbluh (2007), Synchronization of mutually coupled chaotic lasers in the presence of a shutter, *Phys. Rev. Lett.*, *98*, 154101, doi:10.1103/PhysRevLett.98.154101.

Kantz, H., and T. Schreiber (2000), *Nonlinear Time Series Analysis*, Cambridge Univ. Press, Cambridge, U. K.

Kantz, H., and T. Schurmann (1996), Enlarged scaling ranges for the KS-entropy and the information dimension, *Chaos*, *6*(2), 167–171, doi:10.1063/1.166161.

Katul, G. G., C. Lai, K. Schafer, B. Vidakovic, J. Albertson, D. Ellsworth, and R. Oren (2001), Multiscale analysis of vegetation surface fluxes: From seconds to years, *Adv. Water Resour.*, *24*, 1119–1132, doi:10.1016/S0309-1708(01)00029-X.

Katul, G. G., A. Porporato, and R. Oren (2007a), Stochastic dynamics of plant-water interactions, *Annu. Rev. Ecol. Evol. Syst.*, *38*, 767–791, doi:10.1146/annurev.ecolsys.38.091206.095748.

Katul, G. G., A. Porporato, E. Daly, A. C. Oishi, H.-S. Kim, P. C. Stoy, J.-Y. Juang, and M. B. Siqueira (2007b), On the spectrum of soil moisture from hourly to interannual scales, *Water Resour. Res.*, *43*, W05428, doi:10.1029/2006WR005356.

Khan, S., A. R. Ganguly, S. Bandyopadhyay, S. Saigal, D. J. Erickson, III, V. Protopopescu, and G. Ostrouchov (2006), Nonlinear statistics reveals stronger ties between ENSO and the tropical hydrological cycle, *Geophys. Res. Lett.*, *33*, L24402, doi:10.1029/2006GL027941.

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

Knuth, K. H., A. Gotera, C. T. Curry, K. A. Huyser, K. R. Wheeler, and W. B. Rossow (2005), Revealing relationships among relevant climate variables with information theory, paper presented at Earth-Sun System Technology Conference, NASA, Adelphi, Md.

Kontoyiannis, A., and A. Antos (2001), Convergence properties of functional estimates for discrete distributions, *Random Struct. Algorithms.*, *19*, 163–193.

Kumar, P. (2007), Variability, feedback, and cooperative process dynamics: Elements of a unifying hydrologic theory, *Geogr. Compass*, *1*(6), 1338–1360, doi:10.1111/j.1749-8198.2007.00068.x.

Kunkel, K. E., et al. (2006), The 2005 Illinois drought, *Ill. State Water Surv. Inf. Educ. Mat. 2006-03*, Champaign, Ill.

Kurths, J., S. Boccaletti, C. Grebogi, and Y.-C. Lai (2003), Introduction: Control and synchronization in chaotic dynamical systems, *Chaos*, *13*(1), 126–127, doi:10.1063/1.1554606.

Lorrain, F. P., and H. C. White (1971), Structural equivalence of individuals in social networks, *J. Math. Sociol.*, *1*, 49–80.

Ma, S., and H. J. Bohnert (2007), Integration of Arabidopsis thaliana stress-related transcript profiles, promoter structures, and cell-specific expression, *Genome Biol.*, *8*, R49, doi:10.1186/gb-2007-8-4-r49.

Margulis, S. A., and D. Entekhabi (2001), Feedback between the land surface energy balance and atmospheric boundary layer diagnosed through a model and its adjoint, *J. Hydrometeorol.*, *2*, 599–620, doi:10.1175/1525-7541(2001)002<0599:FBTLSE>2.0.CO;2.

Markowetz, F., and R. Spang (2007), Inferring cellular networks- a review, *BMC Bioinf.*, *8*, suppl. 6, S5, doi:10.1186/1471-2105-8-S6-S5.

Marschinski, R., and H. Kantz (2002), Analyzing the information flow between financial time series, *Eur. Phys. J. B*, *30*, 275–281, doi:10.1140/epjb/e2002-00379-2.

Mays, D. C., B. A. Faybishenko, and S. Finsterle (2002), Information entropy to measure temporal and spatial complexity of unsaturated flow in heterogeneous media, *Water Resour. Res.*, *38*(12), 1313, doi:10.1029/2001WR001185.

McNaughton, K. G., and P. G. Jarvis (1991), Effects of spatial scale on stomatal control of transpiration, *Agric. For. Meteorol.*, *54*, 279–301, doi:10.1016/0168-1923(91)90010-N.

Meyers, T. P. (2008), Bondville 30 minute L4 product, IL, U.S.A., 1998–2007, http://www.daac.ornl.gov, Oak Ridge Natl. Lab. Distrib. Active Arch. Cent., Oak Ridge, Tenn.

Moffat, A. M., et al. (2007), Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes, *Agric. For. Meteorol.*, *147, 209–232, doi:10.1016/j.agrformet.2007.08.011*.

Molini, A., P. La Barbera, and L. G. Lanza (2005), Correlation patterns and information flows in rainfall fields, *J. Hydrol.*, *322*, 89–104, doi:10.1016/j.jhydrol.2005.02.041.

Monteith, J. L. (1995), Accommodation between transpiring vegetation and the convective boundary layer, *J. Hydrol.*, *166*, 251–263, doi:10.1016/0022-1694(94)05086-D.

Namias, J. (1960), Factors in the initiation, perpetuation and termination of drought, *Publ. 51*, pp. 81–94, Int. Assoc. of Sci. Hydrol., Gentbrugge, Belgium.

Nichols, J. M. (2005), Inferences about information flow and dispersal for spatially extended population systems using time-series data, *Proc. R. Soc. London, Ser. B*, *272*, 871–876, doi:10.1098/rspb.2004.2889.

Nichols, J. M. (2006), Examining structural dynamics using information flow, *Probab. Eng. Mech.*, *21*, 420–433, doi:10.1016/j.probengmech.2006.02.003.

Nicolis, G., and I. Prigogine (1989), *Exploring Complexity*, W. H. Freeman, New York.

Opgen-Rhein, R., and K. Strimmer (2007), From correlation to causation networks: A simple approximate learning algorithm and its application to high-dimensional plant gene expression data, *BMC Syst. Biol.*, *1*, 37, doi:10.1186/1752-0509-1-37.

Paninski, L. (2003), Estimation of entropy and mutual information, *Neural Comput.*, *15*, 1191–1253, doi:10.1162/089976603321780272.

Percha, B., R. Dzakpasu, and M. Zochowski (2005), Transition from local to global phase synchrony in small world neural network and its possible implications for epilepsy, *Phys. Rev. E*, *72*, 031909, doi:10.1103/PhysRevE.72.031909.

Reichstein, M., et al. (2005), On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm, *Global Change Biol.*, *11*, 1424–1439, doi:10.1111/j.1365-2486.2005.001002.x.

Reiners, W. A., and K. L. Driese (2003), Transport of energy, information, and material through the biosphere, *Annu. Rev. Environ. Resour.*, *28*, 107–135, doi:10.1146/annurev.energy.28.050302.105452.

Richardson, A. D., et al. (2008), Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals, *Agric. For. Meteorol.*, *148*, 38–50, doi:10.1016/j.agrformet.2007.09.001.

Rodriguez-Iturbe, I. (2000), Ecohydrology: A hydrologic perspective of climate-soil-vegetation dynamics, *Water Resour. Res.*, *36*, 3–9, doi:10.1029/1999WR900210.

Roederer, J. G. (2005), *Information and Its Role in Nature*, Springer, Berlin.

Ruddell, B. L., and P. Kumar (2009), Ecohydrologic process networks: 2. Analysis and characterization, *Water Resour. Res.*, *45*, W03420, doi:10.1029/2008WR007280.

Rulkov, N. F., M. M. Sushchik, L. S. Tsimring, and H. D. I. Abarbanel (1995), Generalized synchronization of chaos in directionally coupled chaotic systems, *Phys. Rev. E*, *51*, 262–279, doi:10.1103/PhysRevE.51.980.

Sabesan, S., K. Narayanan, A. Prasad, A. Spanias, and L. D. Iasemidis (2003), Improved measure of information flow in coupled nonlinear systems, paper presented at International Conference on Modelling and Simulation, Int. Assoc. of Sci. and Technol. for Dev., Palm Springs, Calif., 24–26 Feb.

Schreiber, T. (2000), Measuring information transfer, *Phys. Rev. Lett.*, *85*, 461–464, doi:10.1103/PhysRevLett.85.461.

Scott, D. W. (1979), On optimal and data-based histograms, *Biometrika*, *66*, 605–610, doi:10.1093/biomet/66.3.605.

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

Sharma, A., K. C. Luk, I. Cordery, and U. Lall (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 2—Predictor identification of quarterly rainfall using ocean-atmosphere information, *J. Hydrol.*, *239*, 240–248, doi:10.1016/S0022-1694(00)00347-4.

Shuttleworth, W. J. (1988), Macrohydrology—The new challenge for process hydrology, *J. Hydrol.*, *100*, 31–56, doi:10.1016/0022-1694(88)90180-1.

Singer, B. H., M. Derchansky, P. L. Carlen, and M. Zochowski (2006), Lag synchrony measures dynamical processes underlying progression of seizure states, *Phys. Rev. E*, *73*, 021910, doi:10.1103/PhysRevE.73.021910.

Siqueira, M. B., G. G. Katul, D. A. Sampson, P. C. Stoy, J.-Y. Juang, H. R. McCarthy, and R. Oren (2006), Multiscale model intercomparisons of CO_2 and H_2O exchange rates in a maturing southeastern US pine forest, *Global Change Biol.*, *12*, 1189–1207, doi:10.1111/j.1365-2486.2006.01158.x.

Stoy, P. C., et al. (2006), Separating the effects of climate and vegetation on evapotranspiration along a successional chronosequence in the southeastern US, *Global Change Biol.*, *12*, 2115–2135, doi:10.1111/j.1365-2486.2006.01244.x.

Stoy, P. C., et al. (2007), Are ecosystem carbon inputs and outputs coupled at short time scales? A case study from adjacent pine and hardwood forests using impulse-response analysis, *Plant Cell Environ.*, *30*, 700–710, doi:10.1111/j.1365-3040.2007.01655.x.

Strogatz, S. (2001), Exploring complex networks, *Nature*, *410*, 268–276, doi:10.1038/35065725.

Vastano, J. A., and H. L. Swinney (1988), Information transport in spatiotemporal systems, *Phys. Rev. Lett.*, *60*, 1773–1776, doi:10.1103/PhysRevLett.60.1773.

Veeramani, B., K. Narayanan, A. Prasad, L. D. Iasemidis, A. S. Spanias, and K. Tsakalis (2004), Measuring the direction and strength of coupling in nonlinear systems—A modeling approach in the state space, *IEEE Signal Process. Lett.*, *11*(7), 617–620, doi:10.1109/LSP.2004.830120.

Wilhelm, T., and J. Hollunder (2007), Information theoretic description of networks, *Physica A*, *385*, 385–396, doi:10.1016/j.physa.2007.06.029.

Wolpert, D. H., and D. R. Wolf (1995), Estimating functions of probability distributions from a finite number of samples, *Phys. Rev. E*, *52*, 6841–6854, doi:10.1103/PhysRevE.52.6841.

Zhang, P., B. T. Anderson, and R. Myneni (2006), Monitoring 2005 Corn Belt yields from space, *Eos Trans. AGU*, *87*, 150, doi:10.1029/2006EO150003.

P. Kumar and B. L. Ruddell, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. (kumar1@illinois.edu)