Studying Dynamical Process Networks with Information Theory (a Tutorial)

Benjamin L. Ruddell[1]

[1] Arizona State University, bruddell@asu.edu

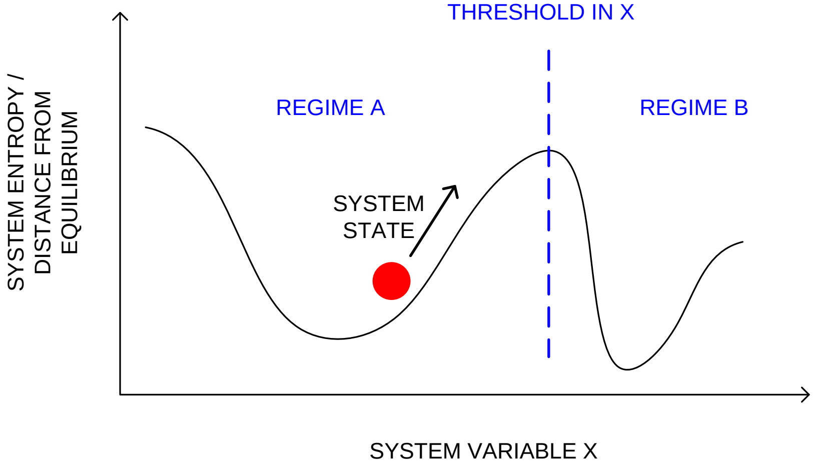
[1]Assistant Professor, Department of Engineering
Senior Sustainability Scientist, Global Inst. of Sustainability
Arizona State University at the Polytechnic Campus

bruddell@asu.edu
480-727-5123

Challenges for Observing and Modeling Complex Earth Systems: Feedback, Nonlinearity, and Scale

The **reductionist approach**, which reduces a system to isolated cause-effect pairs, struggles with complex systems because of:

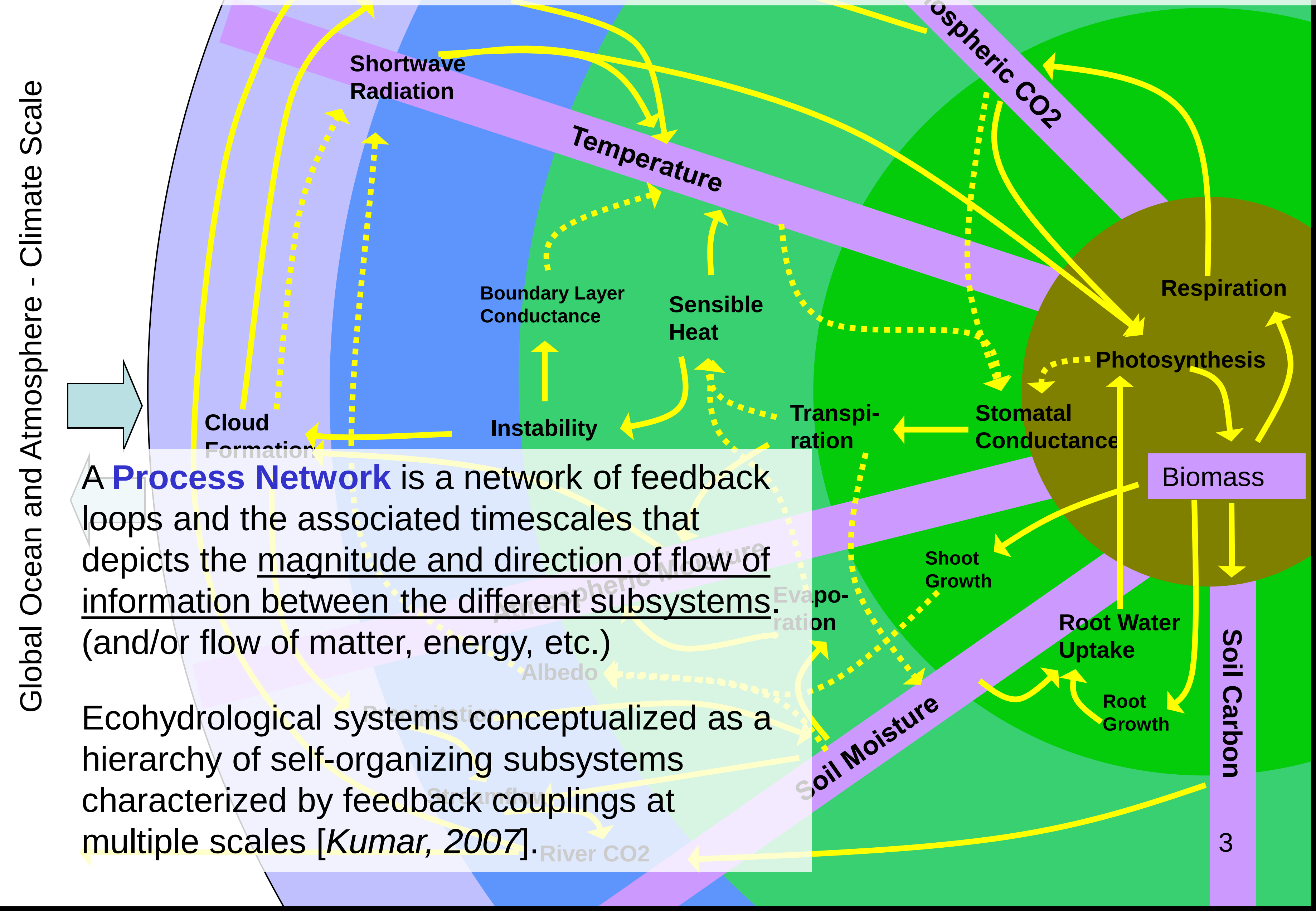
- **Feedback**: cause & effect lose meaning when processes becomes *circular* or *self-organizing*.
- **Thresholds** separate qualitatively different system states. Example: Plant Stomata close when soil moisture is < 20%. Nonlinearity is dominant at subdaily timescales [*Baldocchi*, 2001a].
- **Scale**: One scale cannot be isolated; all interact.
- **Variability**: system structure is dynamic (changing in time), and can be controlled by the *variability* of the system itself. Example: fish populations in Illinois River crash when new dams reduce flow variability [*Koel and Sparks*, 2002].
- **Observation**: It is difficult to observe all physical components of the system at all relevant scales.



Stomata Transpiration
SWC > 0.2, well-watered

Stomata Transpiration Ceases
SWC < 0.2, dry

Conceptualize the Complex System as a Process Network of Couplings between Observable and/or Modeled Subsystems



A **Process Network** is a network of feedback loops and the associated timescales that depicts the magnitude and direction of flow of information between the different subsystems. (and/or flow of matter, energy, etc.)

Ecohydrological systems conceptualized as a hierarchy of self-organizing subsystems characterized by feedback couplings at multiple scales [*Kumar, 2007*].

Q: What is Information?

Q: What is Information?

A: Information is the Answer to a Question

Q: What is Information?

A: Information is the Answer to a Question

Q: What is our Question?

Q: What is Information?

A: Information is the Answer to a Question

Q: What is the right Question for dynamical systems study?

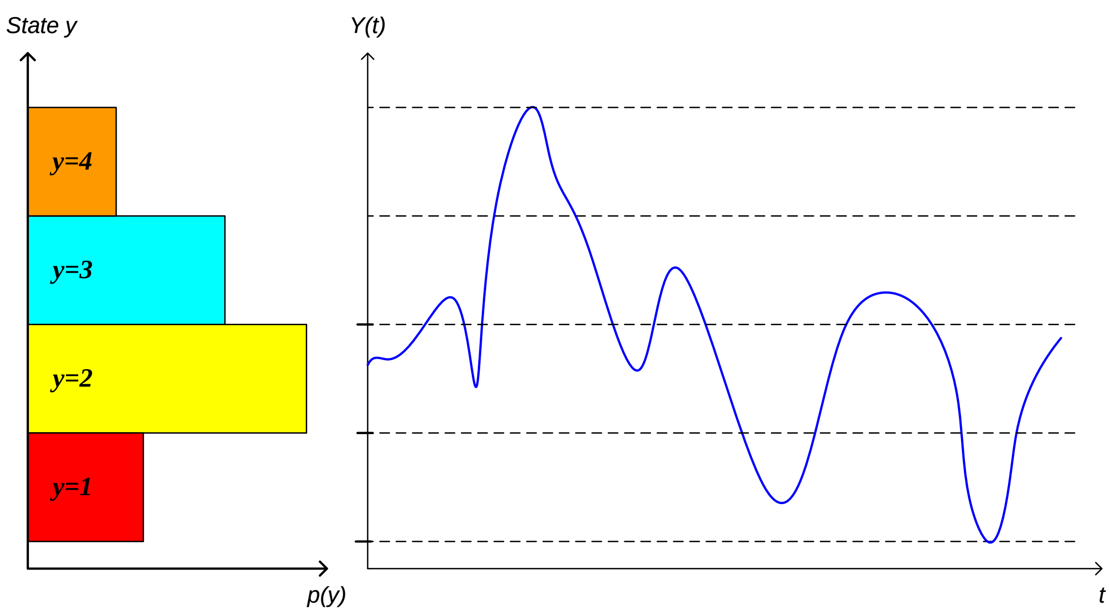
A: "What will be the Future State of Timeseries Variable *Y(t)*?"

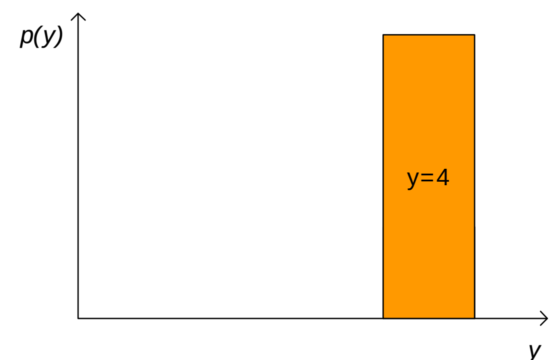
Shannon Entropy: The fundamental measure of uncertainty and information

p(y) is the prior probability that discrete variable *Y* takes state *y.*

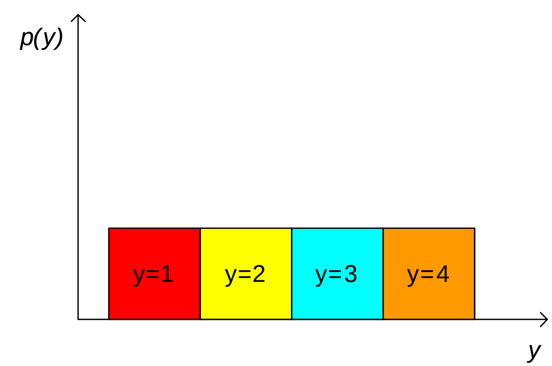
$$H(Y_t) = -\sum_{y \in Y_t} p(y) \cdot \log p(y)$$

$H(Y_t)$, the Shannon Entropy, measures the size of the question of state; this is also the amount of information we gain when we learn the answer to the question.





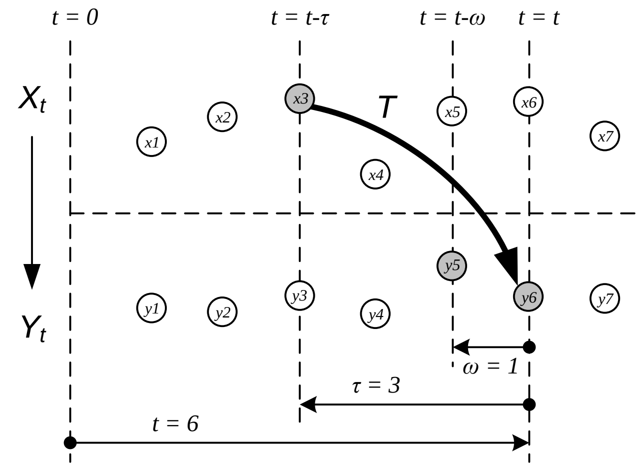
Minimum H for 4 discrete states
H(Y) = 0 bits, no uncertainty



Maximum H for 4 discrete states
H(Y) = $\log_2(4)$ = 2 bits

How to Measure Information Flow?

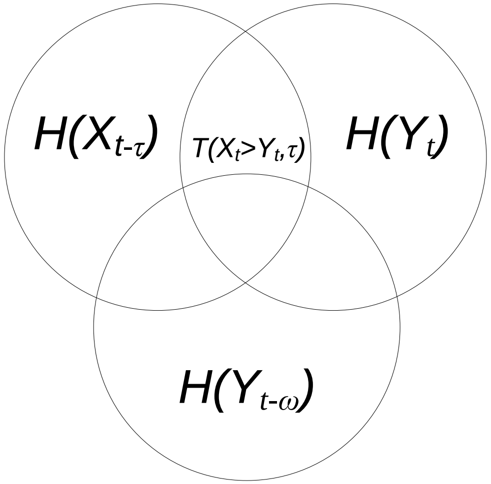
Transfer Entropy!



- To measure directional information flow and assess timescales of flow, we need an asymmetric measure of information flow

- Thomas Schreiber [2000] introduces **Transfer Entropy** *T*, conditioning information shared by X_t and Y_t on Y_t's history

$$T(X_t > Y_t, \tau) = \sum_{y_t, y_{t-1}, x_{t-\tau}} p(y_t, y_{t-1}, x_{t-\tau}) \log \frac{p(y_t \mid (y_{t-1}, x_{t-\tau}))}{p(y_t \mid y_{t-1})}$$



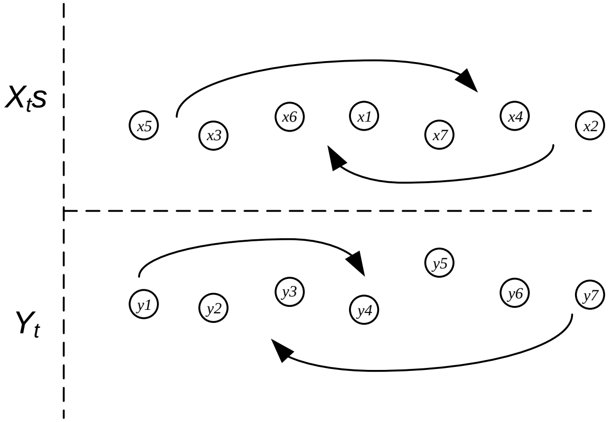
- *T* measures additional information contributed by X_t across at time lag τ. Entropy reduced = information produced.

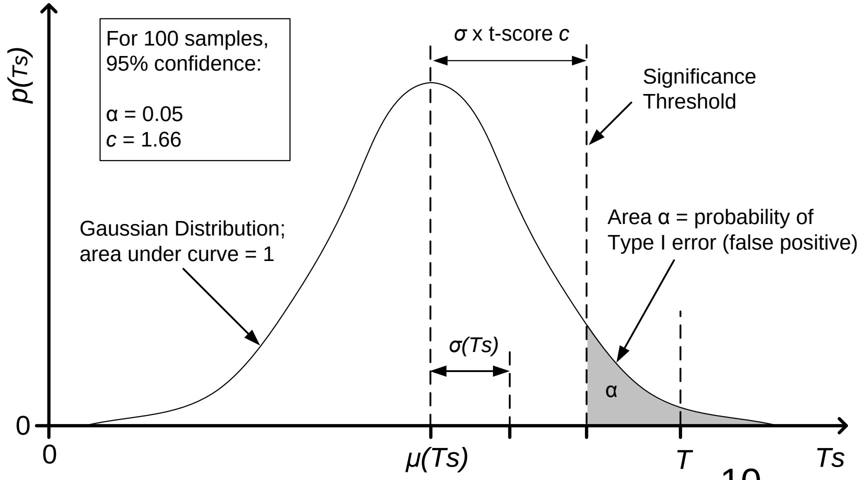
- By computing *T* across many time lags, we can assess the time scale of directional coupling from X_t to Y_t

$$T(X_t > Y_t, \tau) = H(X_{t-\tau}, Y_{t-1}) + H(Y_t, Y_{t-1}) - H(Y_{t-1}) - H(X_{t-\tau}, Y_t, Y_{t-1})$$

Establish Statistical Significance of Information Flow between X_t and Y_t

- How do we decide whether *T* is large enough to represent a significant flow of information?
- Compare measured *T* against *Ts*, which is the information flow using a time-shuffled X_t and Y_t "bootstrapping".
- When $T > Ts$, a significant information flow exists; X_t contributes significantly to our ability to answer questions about future states of Y_t.
- Robustness of results additionally ensured by quality control including testing on coupled Logistic maps, and with various *N*, *m*, and binning schemes.





Is strong *I* due to feedback synchronization or forcing?

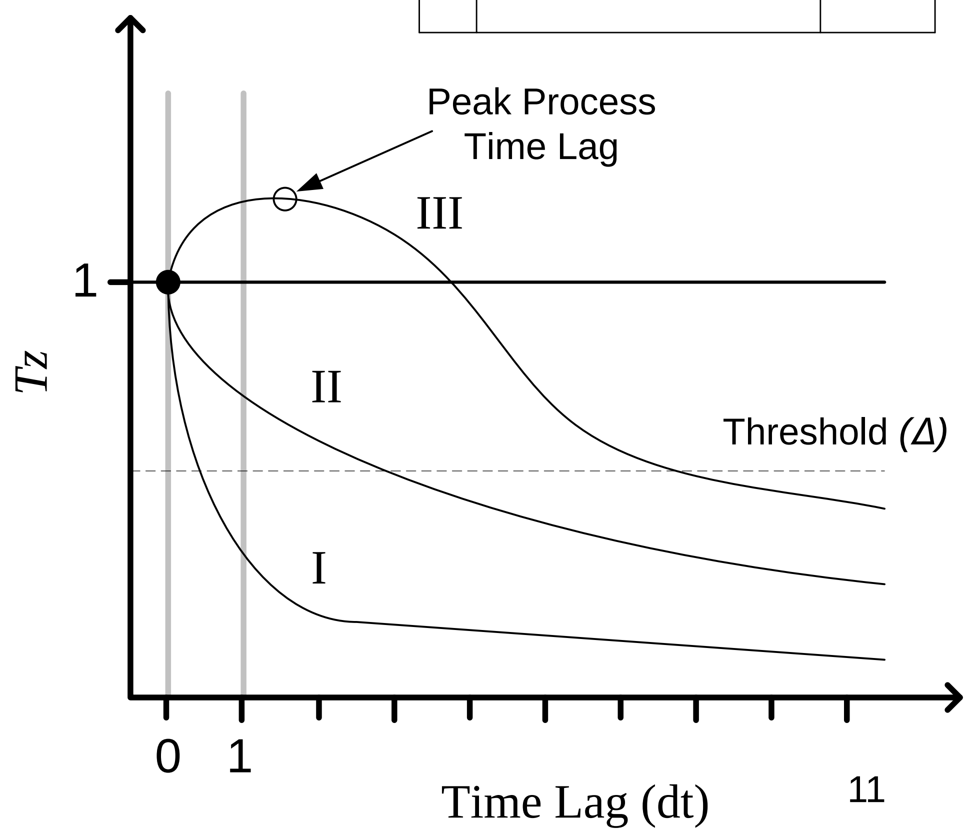
		$T(Y_t>X_t)$ I	II	III	IV
$T(X_t>Y_t)$	I	√	√	√	×
	II		√	√	×
	III			√	?
	IV				?

When two variables share a lot of information at a given time scale, WHY?

1. **Type-III (Forcing-Dominated)**: *T* > *I*, so information flow dominates shared information.
2. **Type-II (Feedback-Dominated):** *T* < *I*, shared information dominates information flow
3. **Type-I (Synchronization-Dominated)**: significant *I*, but not *T*. No flow.

Eight **canonical coupling types** are formed from pairs of these couplings.

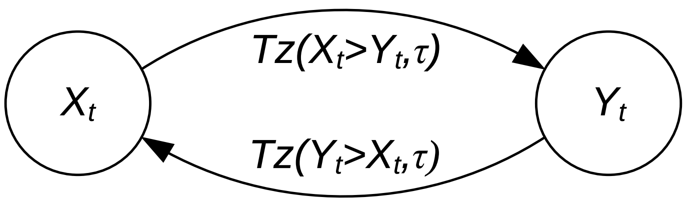
$$Tz(X_t > Y_t, \tau) = \frac{T(X_t > Y_t, \tau)}{I(X_t, Y_t)}$$

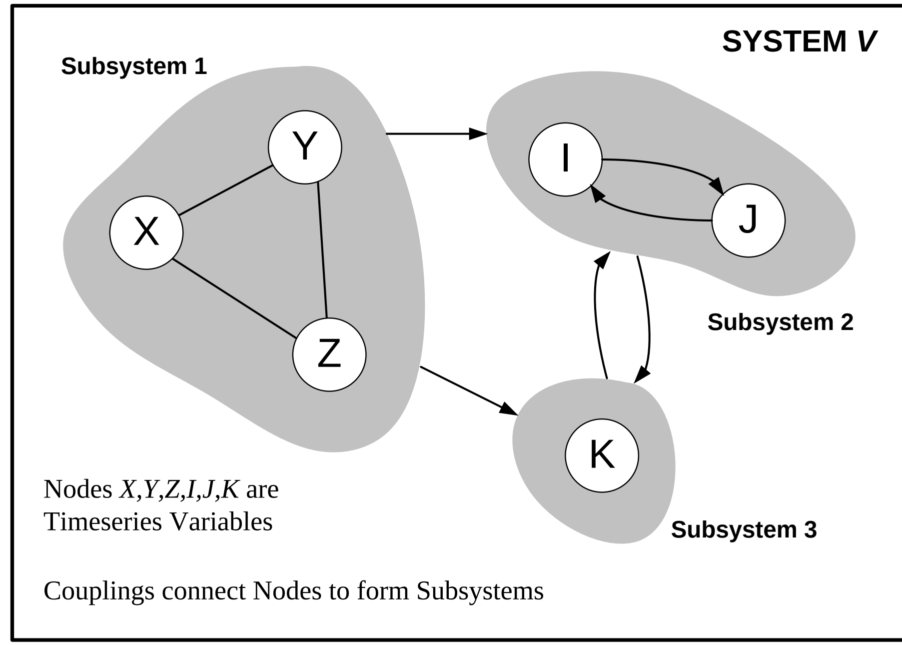


Build a Process Network using *Tz*

Procedure:

- Compute *T* between all pairs of variables at multiple time lags
- Assess statistical signficiance of each information flow coupling
- Compute *Tz*
- Identify Characteristic Time Lag
- Construct Process Network





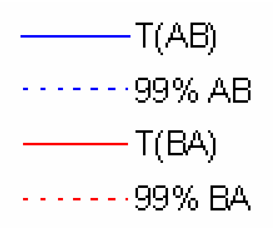
ATz*(i,j)* , 'x' means Type-I coupling, bold means Type-II coupling, otherwise Type-III coupling

	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

Testing with Coupled Logistic Chaotic Time-Series

- Construct two **synthetic time-series** with time-lag relationships
  - Coupled Autoregressive Noise
  - Coupled Logistic Maps
  - lagAB = 1, lagBA = 7
  - *r*=3.99 [Logistic Map chaotic range]

- **Characteristic Lag τ** is the first significant local peak to occur in a spectrum (circled in green). It is desirable to reduce dimensionality of the problem by picking just one τ.

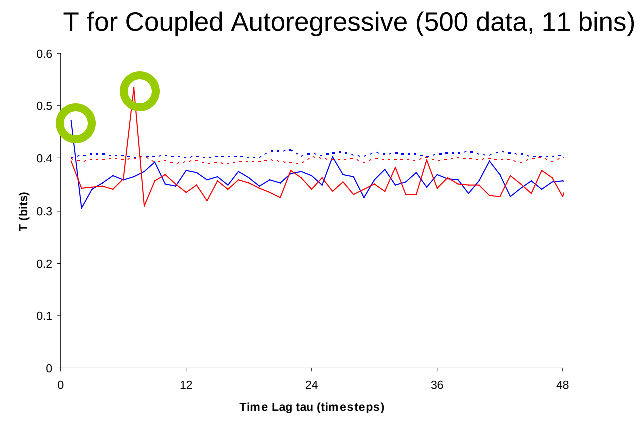


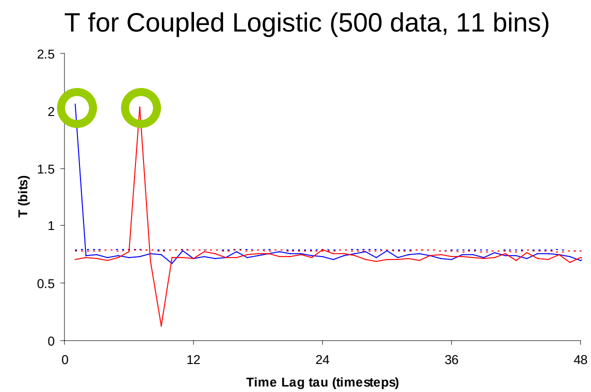
$$AR_A(t) = C \cdot AR_B(t - lag_A) + NormalGaussianNoise$$

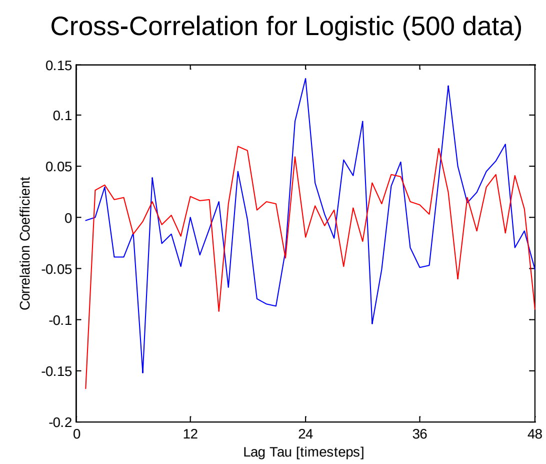
$$AR_B(t) = C \cdot AR_A(t - lag_B) + NormalGaussianNoise$$

$$Logistic_A(t) = r \cdot Logistic_B(t - lag_A) \cdot (1 - Logistic_B(t - lag_A))$$

$$Logistic_B(t) = r \cdot Logistic_A(t - lag_B) \cdot (1 - Logistic_A(t - lag_B))$$

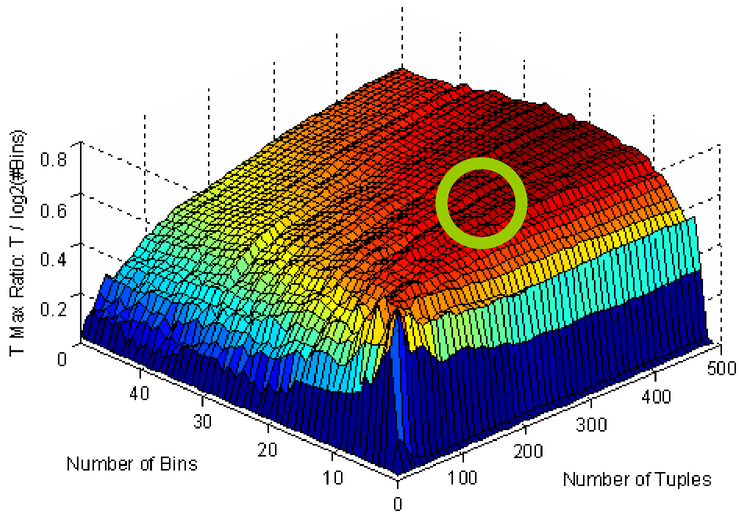


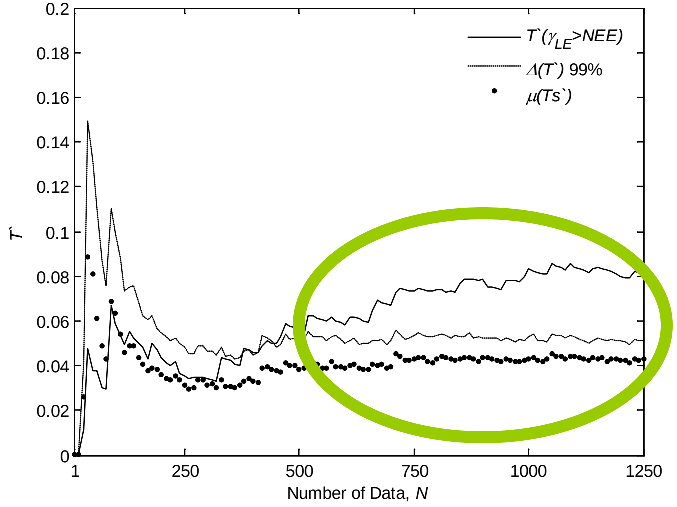


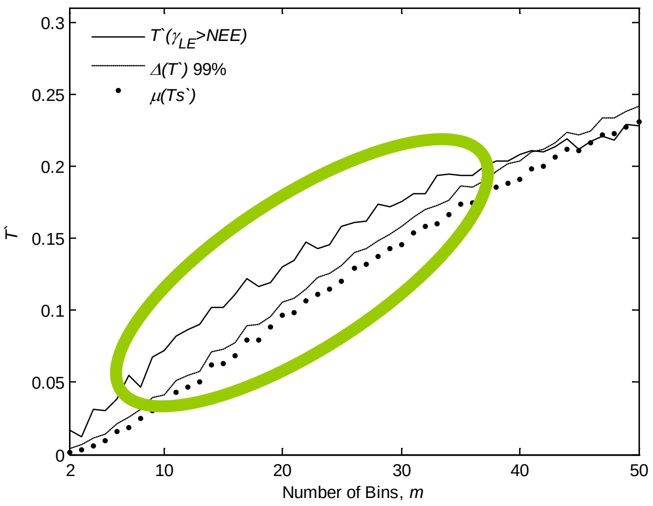


Estimation Issues: How Many *Bins* and *Data*?

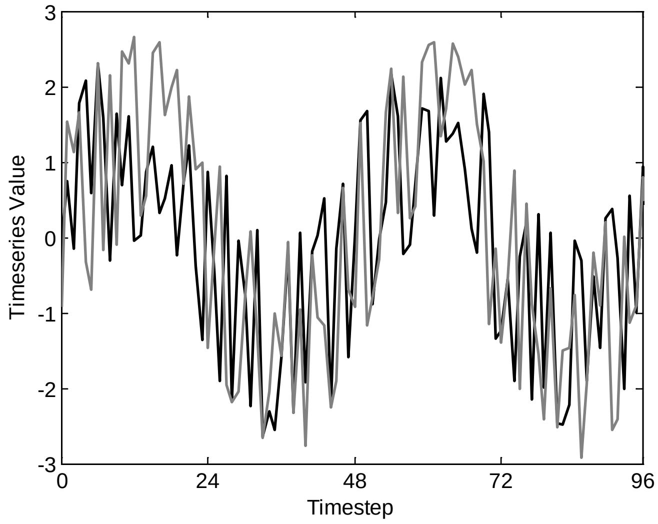
- Plot peak-lag *T* vs. #Bins, #Data used. Too few bins or data causes negative bias in *T.* Too many bins does the same.
- For the coupled logistic map, 10-20 bins, 200+ samples are adequate to provide a fully mature estimate of *T.*
- For the γ_{LE} and NEE coupling, 500+ data and 3-35 bins achieve a qualitatively accurate estimate, but one which is not fully mature. We lack sufficient data and must make a compromise.

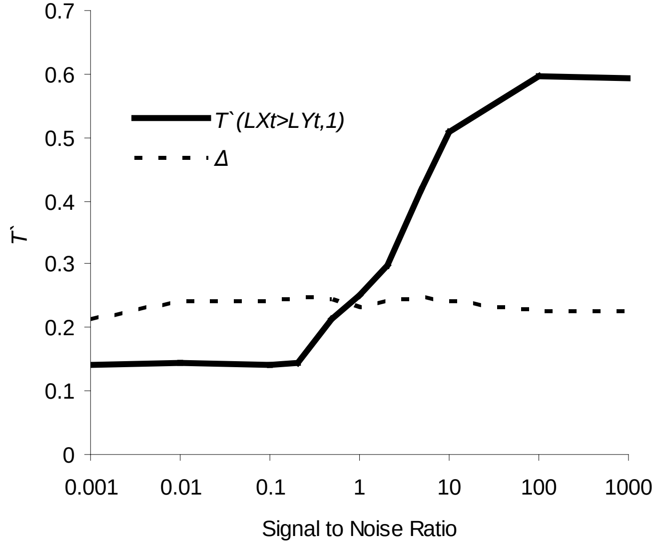






Periodic Noise Sensitivity (P1)

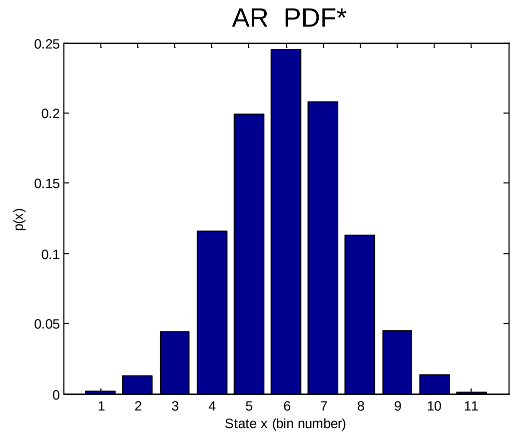


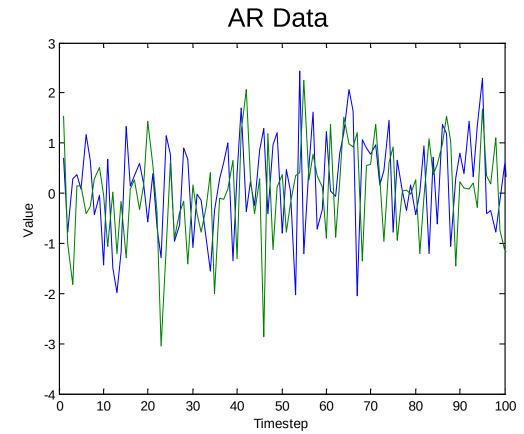


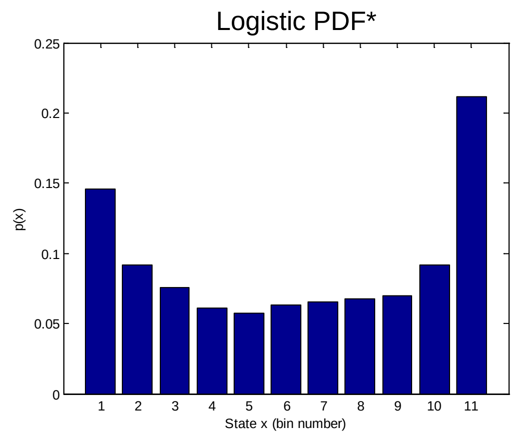
Proving T
(I) Toy Data

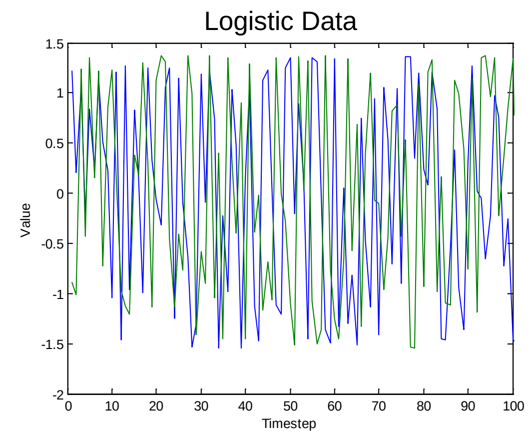
- Construct two toy datasets with directional time-lag relationships
  – Coupled Autoregressive Noise (AR)
  – Coupled Logistic Maps
  – lagAB = 1, lagBA = 7
  – r=3.99 [Logistic Map chaotic range]
  – C=0.5 [moderate coupling]

- AR has a positive linear correlation at specified lags, but Logistic Map has no linear relationship at all









$$AR_A(t) = C \cdot AR_B(t - lag_A) + NormalGaussianNoise$$

$$AR_B(t) = C \cdot AR_A(t - lag_B) + NormalGaussianNoise$$

$$Logistic_A(t) = r \cdot Logistic_B(t - lag_A) \cdot (1 - Logistic_B(t - lag_A))$$

$$Logistic_B(t) = r \cdot Logistic_A(t - lag_B) \cdot (1 - Logistic_A(t - lag_B))$$

* For 10000 data, using 11 discrete equal-interval states

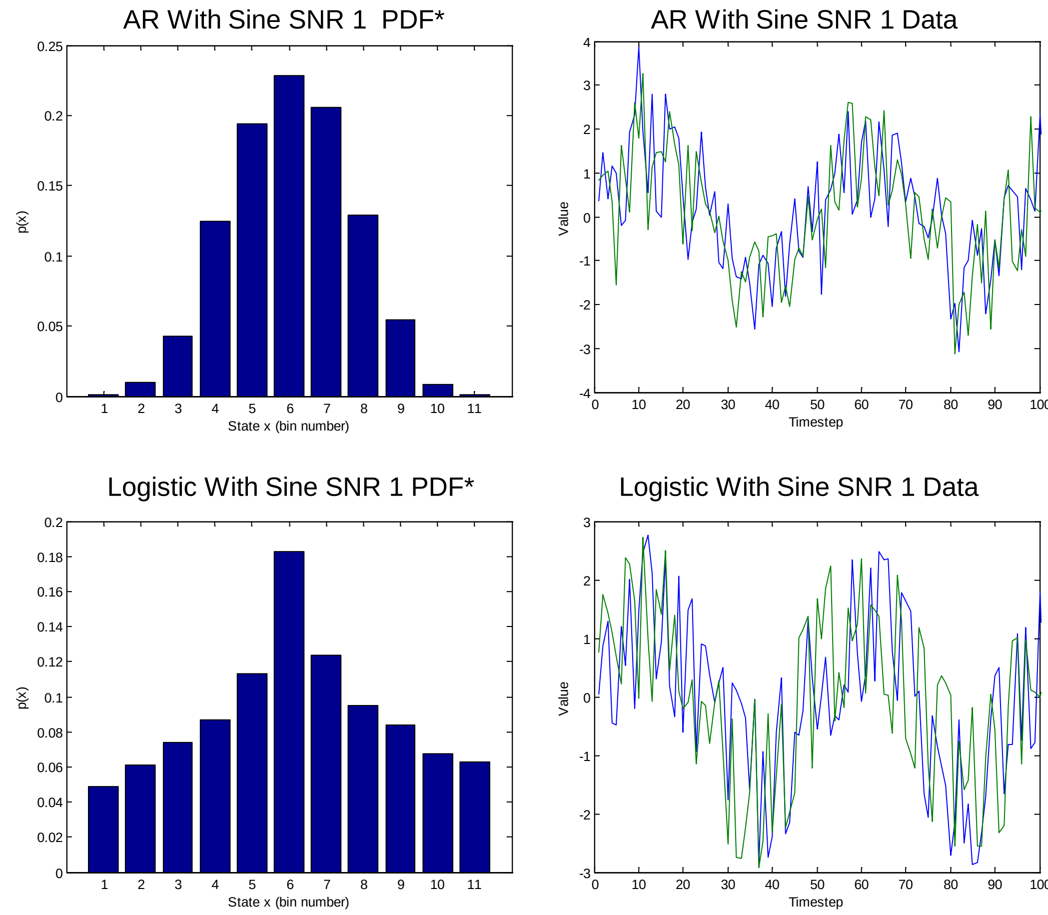
Proving T
(II) Add Periodics to Toy Data

•Using environmental timeseries data, all signals are dominated by periodics (diurnal, synoptic, seasonal oscillations) which are noise for our purposes.

•Add sine-wave periodic “noise” to our toy signals to test the methods

•Normalize sine wave and toy signals to mean 0 and Standard Deviation 1 then add them together in proportion to a Signal-to-Noise ratio (SNR)

•Plots to right have SNR=1



$$PeriodicData = (ToyData \cdot SNR) + SineWave$$

* For 10000 data, using 11 discrete equal-interval states

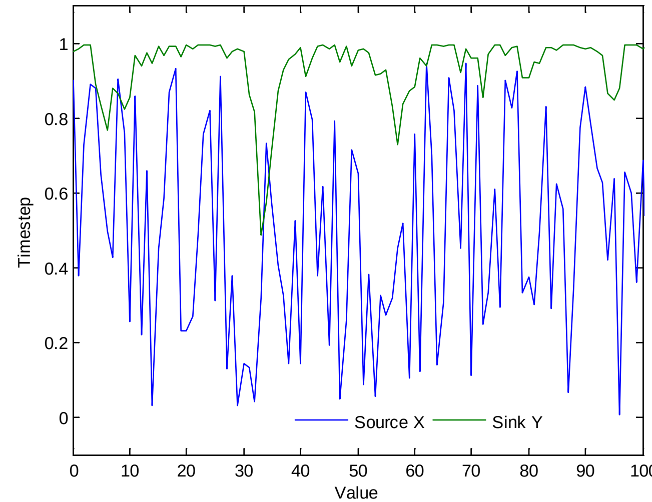
Proving T
(III) Inter-Scalar Relationships

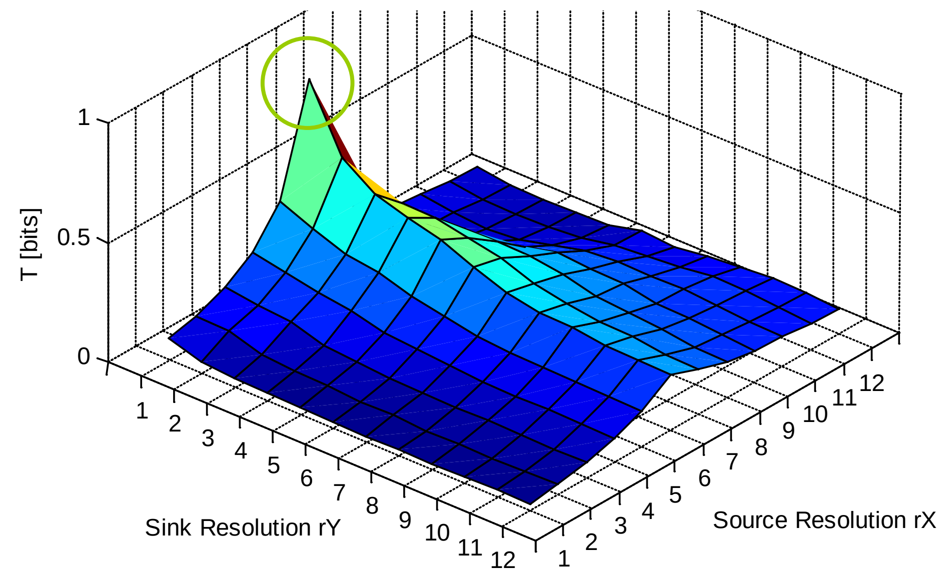
- Use T to resolve asymmetric information flow between scales

- Test using reformulated Coupled Logistic Map, where values in *Y* are mapped based on the mean of the previous *r* number of timesteps in *X* (X is transformed using a moving average of length *r*)

- Plotting T against the source (X) and sink (Y) ***process scale***, this method identifies a peak process scale of the coupling from X to Y at (*rX*=6,*rY*=1)

$$\bar{x}_t(r) = \mu(x_{t+w} : x_{t+w+r-1})$$

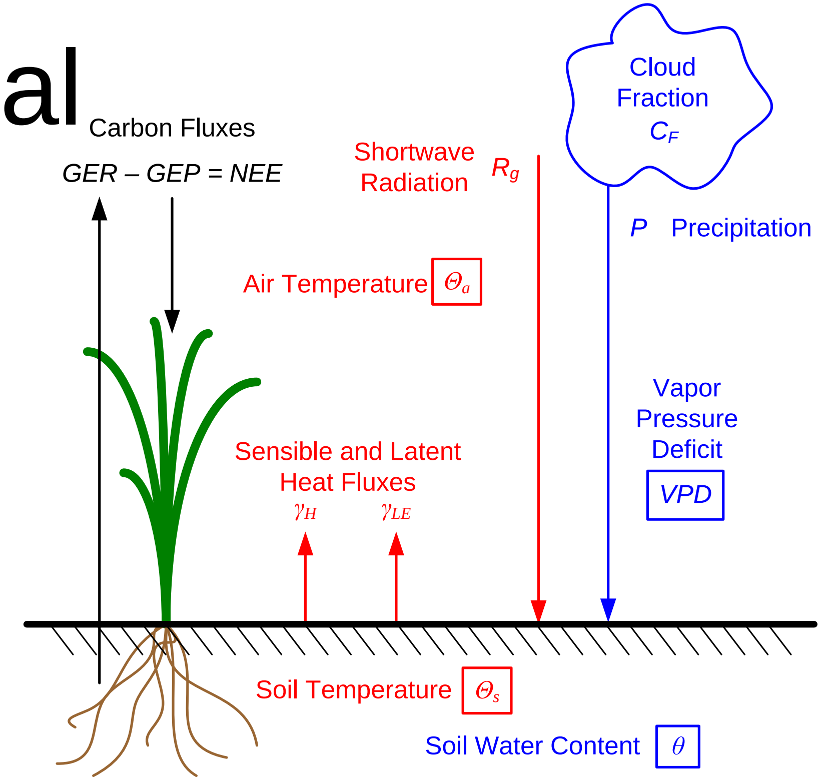




What makes these statistics different?

- Compared with correlations, etc., Information Flow resolves nonlinear and discrete relationships.
- Metrics are based on probability theory, so the results are directly related to predictability and uncertainty.
- *T* is asymmetric and conditioned on auto-information so it can distinguish one-way relationships from two-way relationships; *in other words it can distinguish two-way feedback-based synchronization of two subsystems from one-way forcing of one subsystem by the other.*
- By distinguishing synchronization from forcing relationships and identifying the time and space scales of these couplings, it is possible to logically delineate a hierarchy of physical/functional subsystems that share the same types couplings with the system.
- Information is conserved on the Process Network; this enables the calculation of system-average properties and the sensitivity of the system as a whole to changes in specific subsystems.

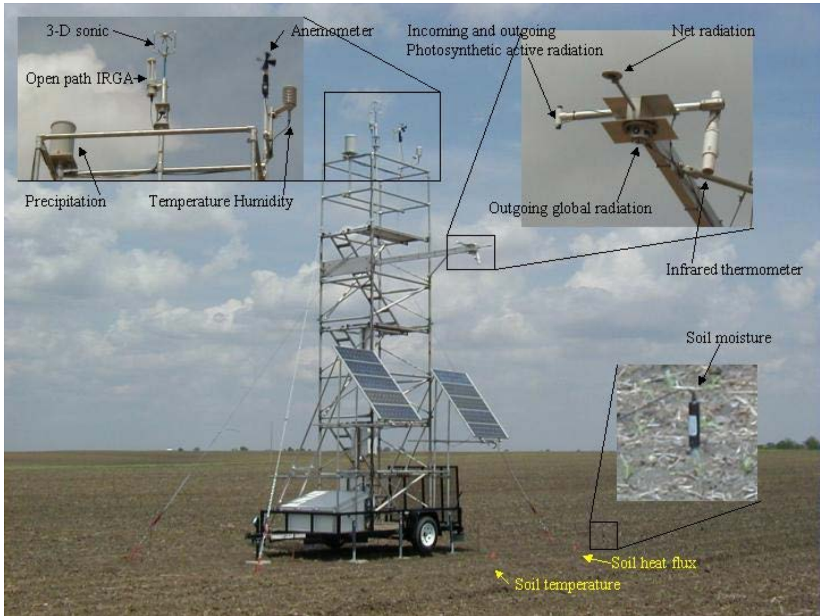
Example Experimental Framework



Natural Laboratory / Observatory Concept: an uncontrolled experiment consisting of many passive observation and data collection in the natural world.

FLUXNET Network: a global flux tower network collecting meteorological, hydrological, and environmental data, including carbon, water, and energy fluxes on the land surface [*Baldocchi*, 2001b].

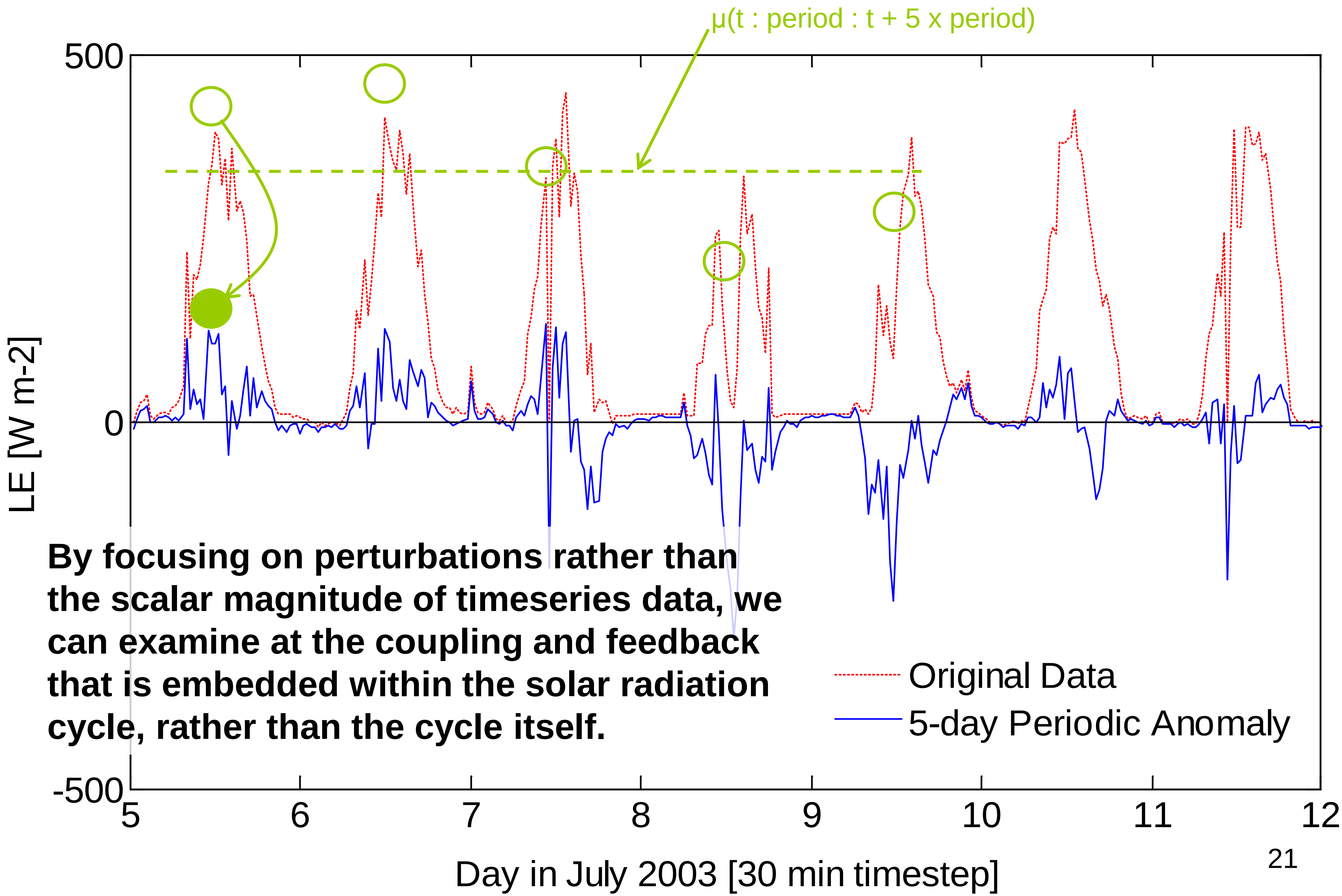
Timeseries Dataset*: several timeseries variables representing the most important ecohydrological processes are collected at eight North American FLUXNET sites, at a 30-minute averaging resolution, for the years 1998-2006. The Level-4 (L4) gap filled, quality controlled product is the best available product [*Reichstein et al.*, 2005]



*Data is preprocessed into a periodic anomaly, to emphasize change and filter out periodic cycles at scales greater than the subdaily (>24 hours).

Data

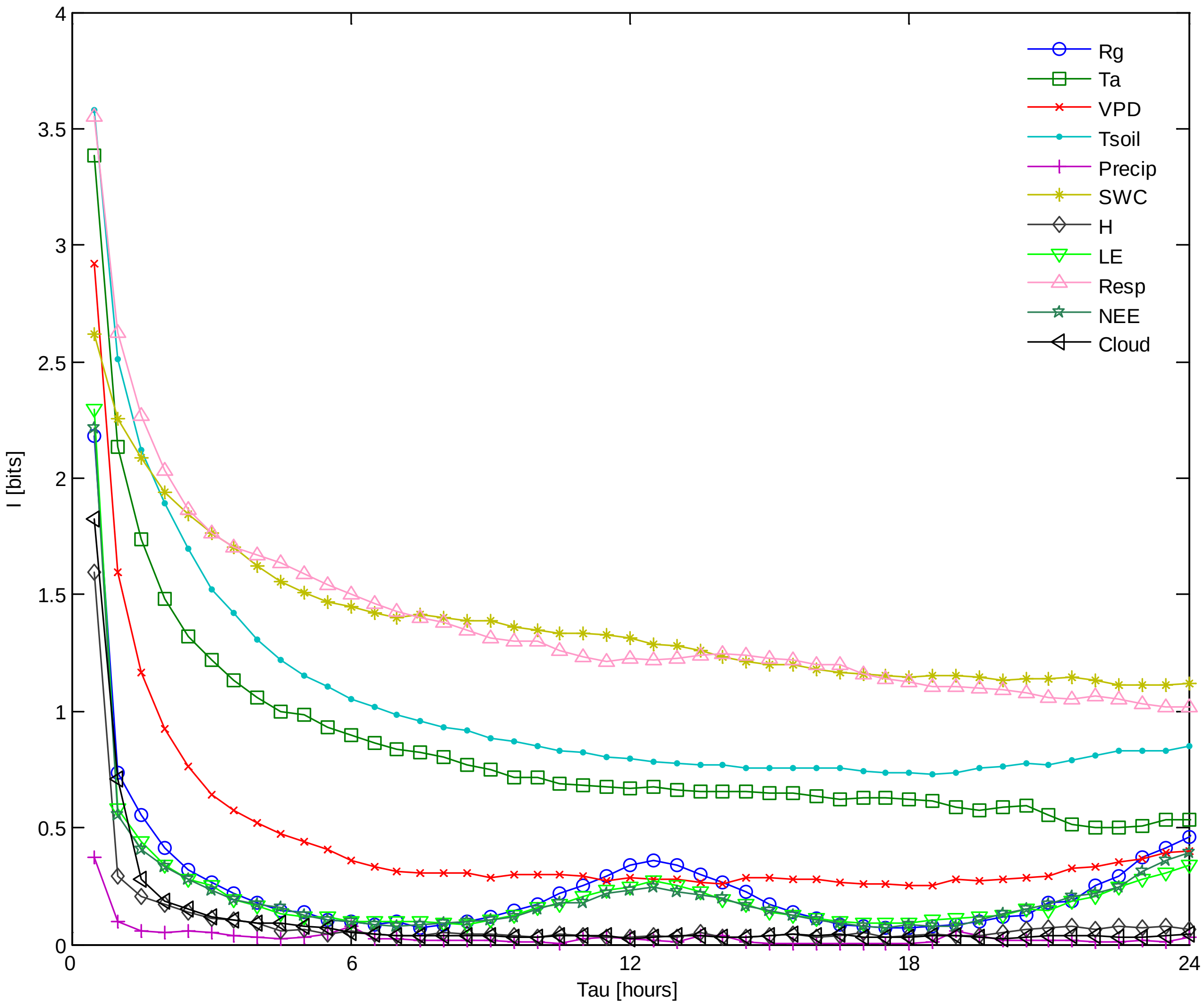
July 2003 γ_{LE} and NEE Data Before and After Taking Anomaly: Bondville, Illinois Ecohydrological System



By focusing on perturbations rather than the scalar magnitude of timeseries data, we can examine at the coupling and feedback that is embedded within the solar radiation cycle, rather than the cycle itself.

Auto-I for July 2003

- SWC, Resp, Tsoil, Tair, VPD show strong self-I results and slow decay with strong self-I to and past 24 hours… a synoptic-scale signature
- Precip, Cloud, and H show very weak I results, with little I past 6 hours… and no diurnal cycle
- Rg, LE, NEE show strong diurnal cycle signatures, decaying by 6 hours but echoing at 12 and 24 hours. Interestingly, H does not do this.

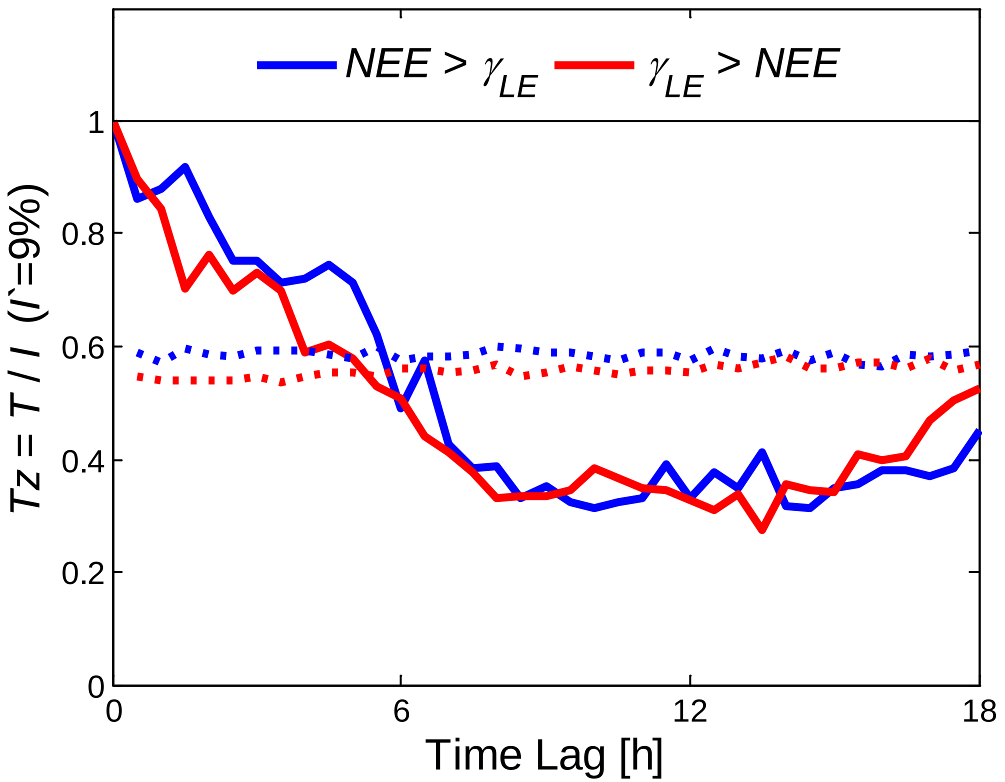


Coupling between Latent Heat Flux and Net Ecosystem Exchange

<30 min bidirectional feedback

Timescale corresponds to that known to control variance in turbulent canopy processes [Baldocchi et al. 2001].

γ_{LE} and NEE control each other, achieving a self-organized dynamic equilibrium via feedback at turbulent timescales



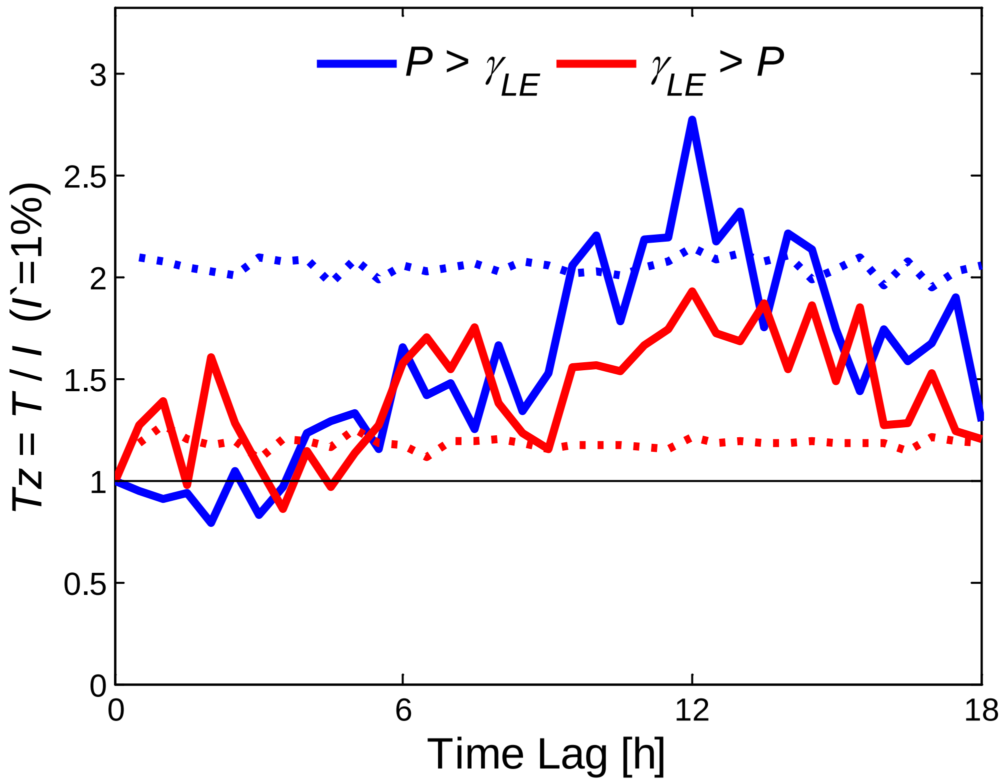
Coupling between Latent Heat Flux and Precipitation

2 to 16 hour feedback

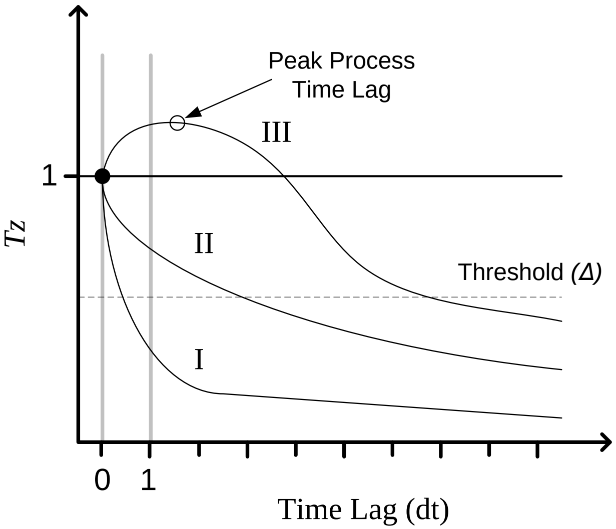
mechanism of moisture recycling in the ABL

NEE coupled to LE, so NEE indirectly controls ABL moisture recycling process

Individual plants’ photosynthetic processes combine to control ABL on a **regional scale**

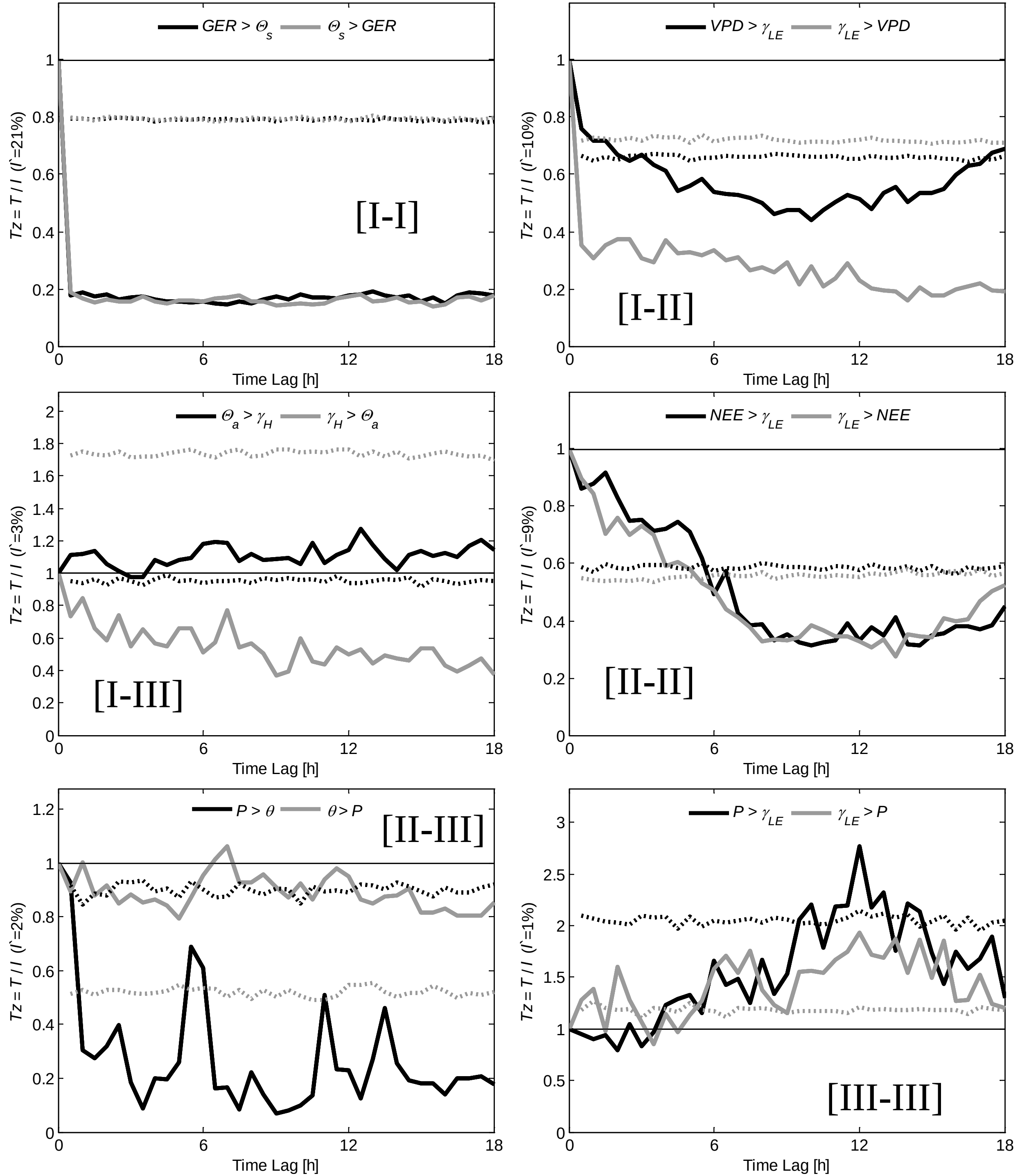


Canonical Couplings Illinois, July 2003



$T(Y_t>X_t)$

$T(X_t>Y_t)$	I	II	III	IV
I	√	√	√	×
II		√	√	×
III			√	?
IV				?



Example: Characterizing drought as a change in observed information feedback (in Illinois)

WATER RESOURCES RESEARCH, VOL. 45, W03419, doi:

Click Here for Full Article

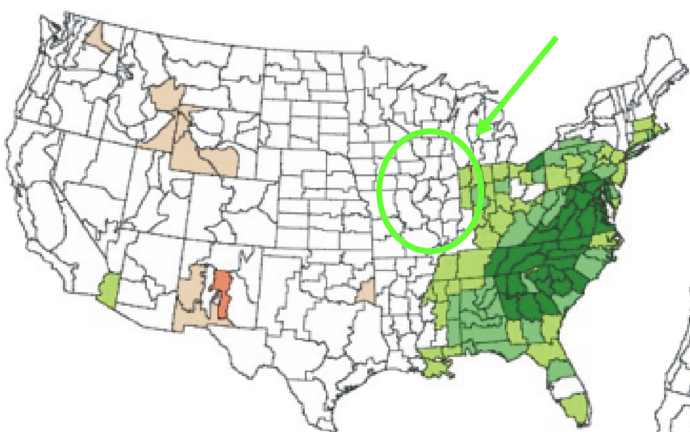
Ecohydrologic process networks: 1. Identification

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

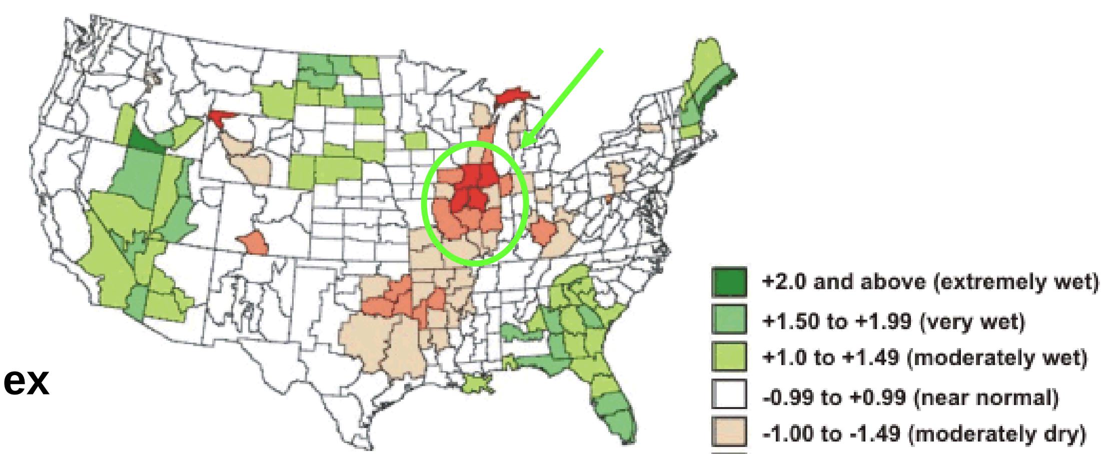
Received 11 July 2008; revised 25 November 2008; accepted 17 December 2008; published

Using the Bondville FLUXNET site; a corn-soybean ecosystem

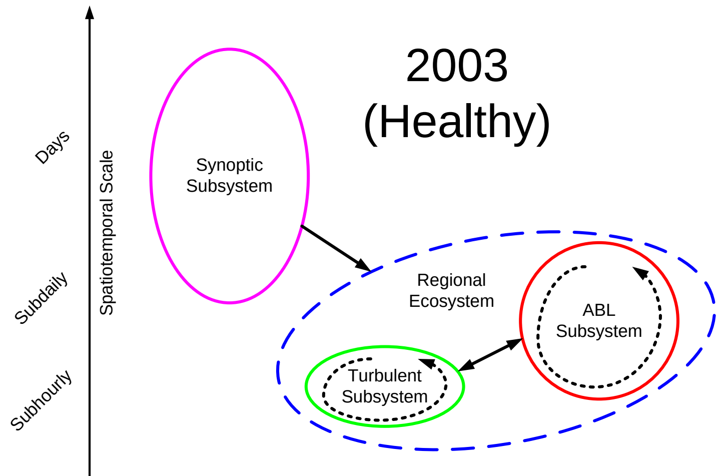
6-month SPI through the end of July 2003



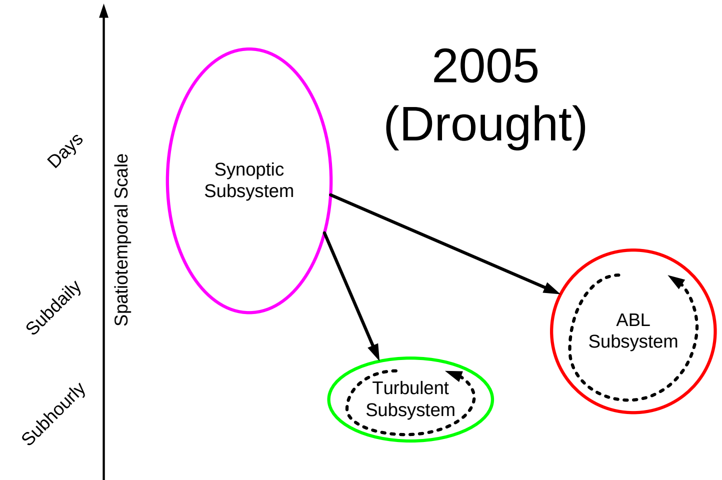
6-month SPI through the end of July 2005



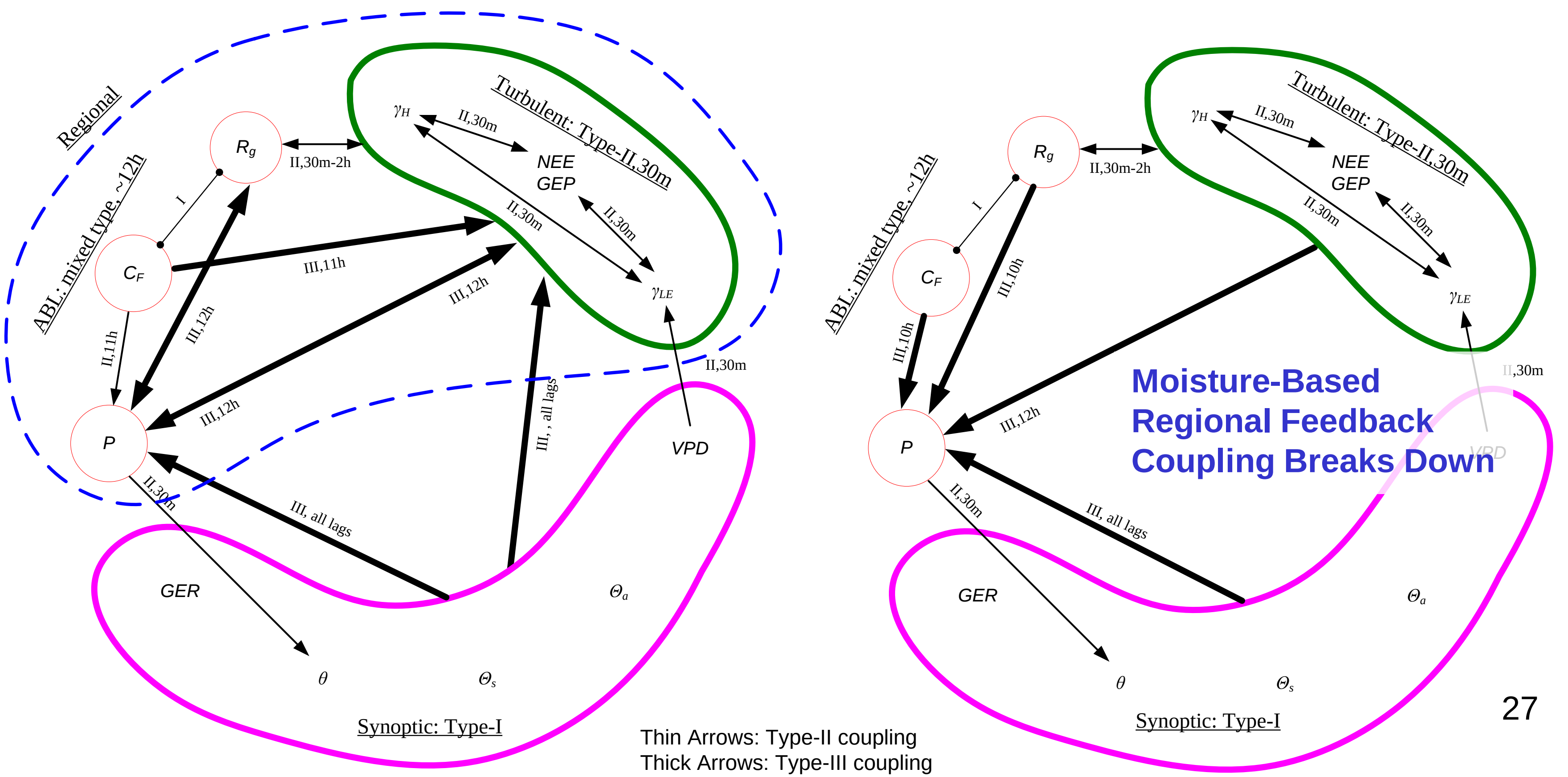
Standardized Precipitation Index
http://www.drought.unl.edu/



2003 July: Healthy System State

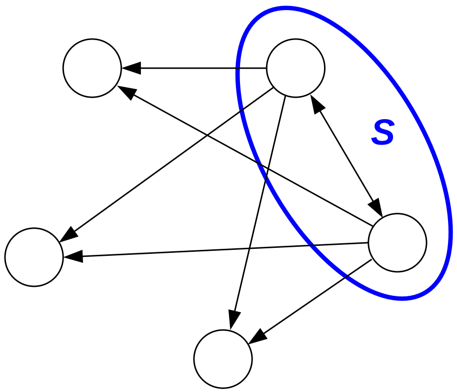


2005 July: Drought System State



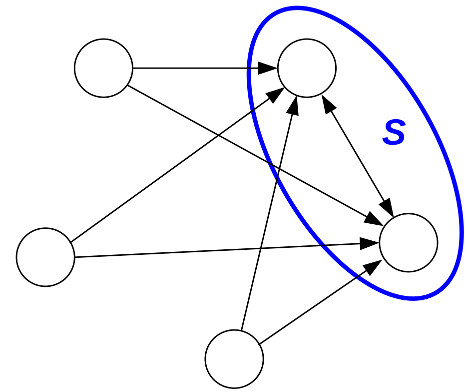
Computing Information Production and Shannon Entropy of a Process Network

Gross information production *T[+](S)*



$$T_S^{[+]}(\tau) = \sum_{\substack{i \in S \\ z \in V}} \mathbf{A}(i, z, \tau)$$

Gross information consumption *T[-](S)*

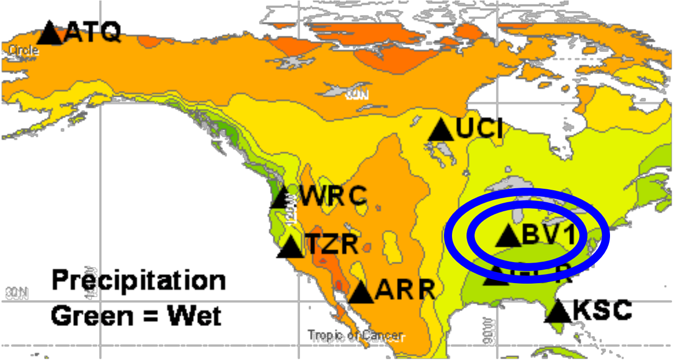


$$T^{[-]}(S, \tau) = \sum_{\substack{i \in S \\ z \in V}} \mathbf{A}(z, i, \tau)$$

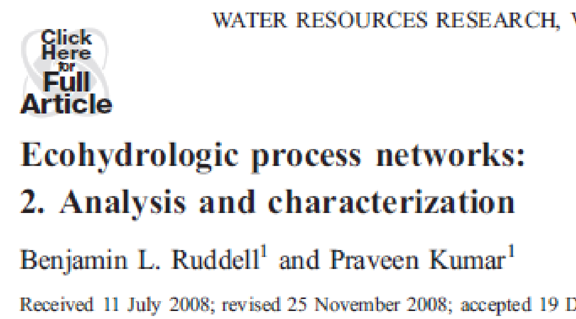
Net information production *T^net(S)* $T^{net}(S, \tau) = T^{[+]}(S, \tau) - T^{[-]}(S, \tau)$

Total information production *TST(V)* is the normalized sum of *T[+](S)* *across all subsystems S*

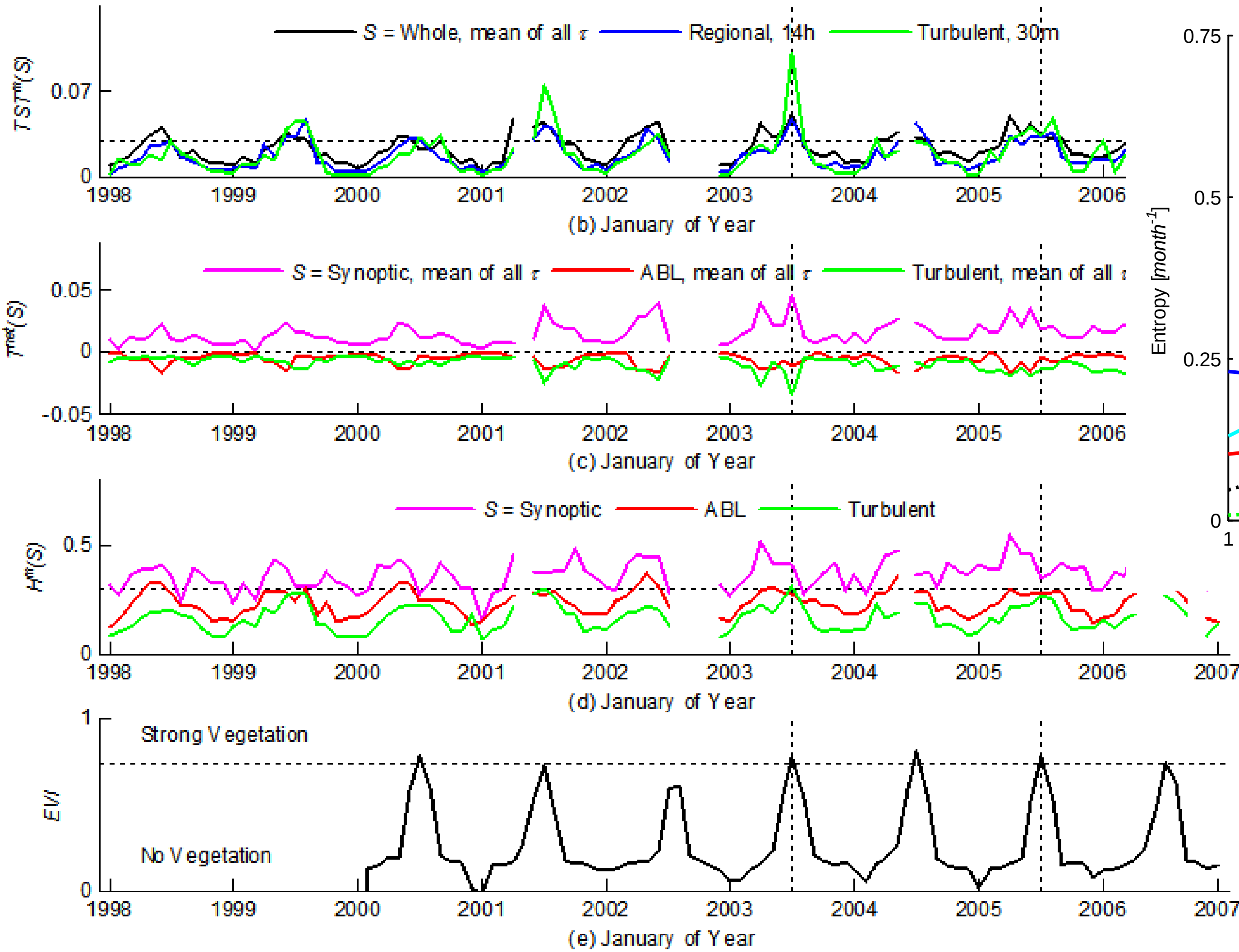
Mean System Shannon Entropy *H(V)* is the normalized average of all subsystem Shannon Entropies *H(S)*

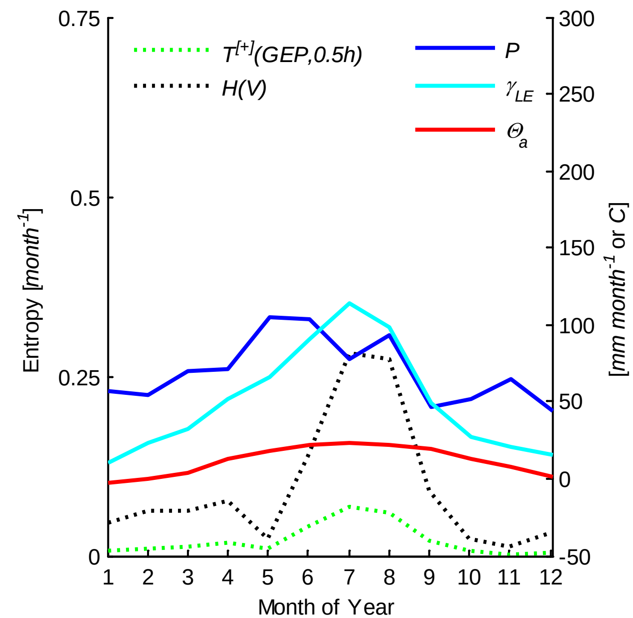


Network Information Flow Results


WATER RESOURCES RESEARCH, V

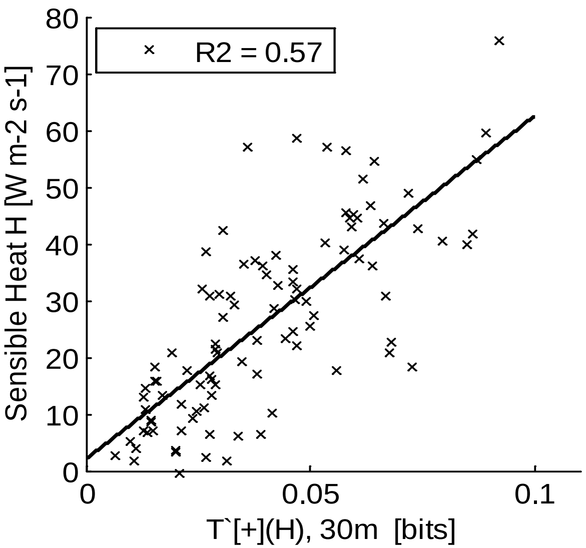
Ecohydrologic process networks:
2. Analysis and characterization
Benjamin L. Ruddell[1] and Praveen Kumar[1]
Received 11 July 2008; revised 25 November 2008; accepted 19 D

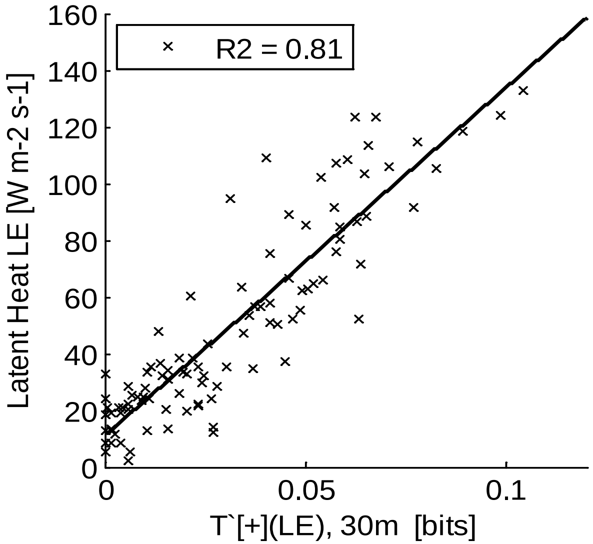


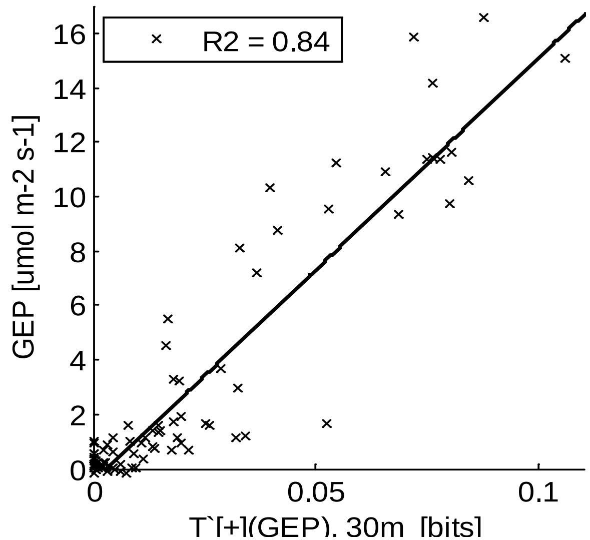


All data presented is 30-min L4 eddy-covariance measurements from FLUXNET sites (Baldocchi 2001, Reichstein et al. 2005)

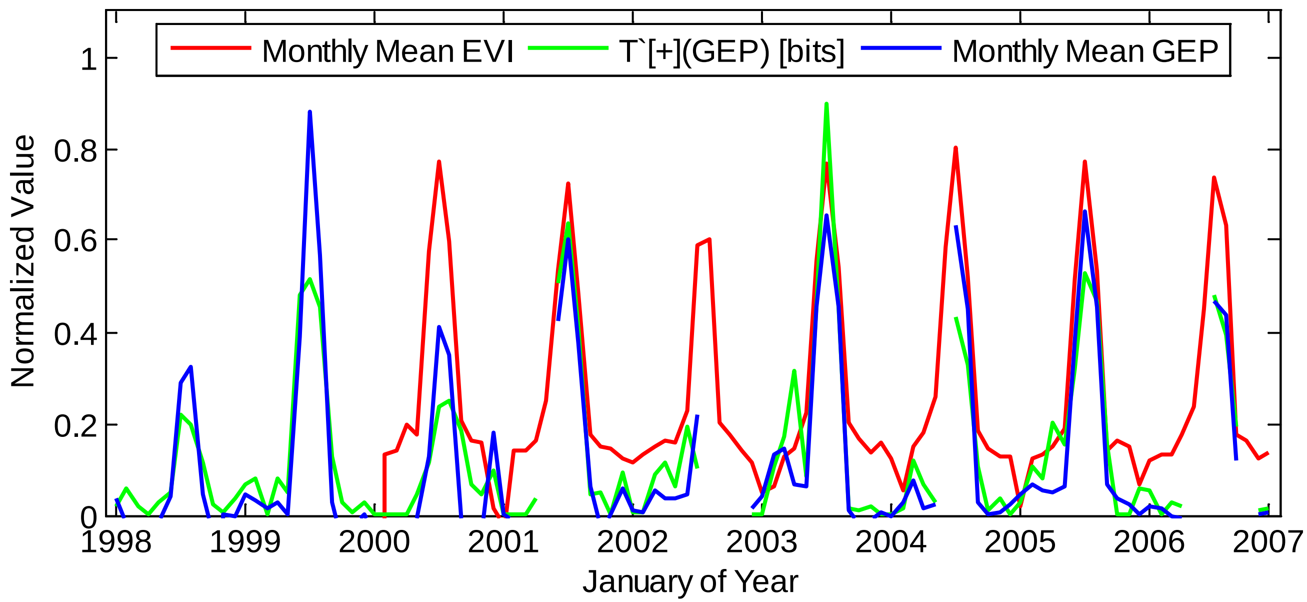
Information Production Related to Phenology and Productivity

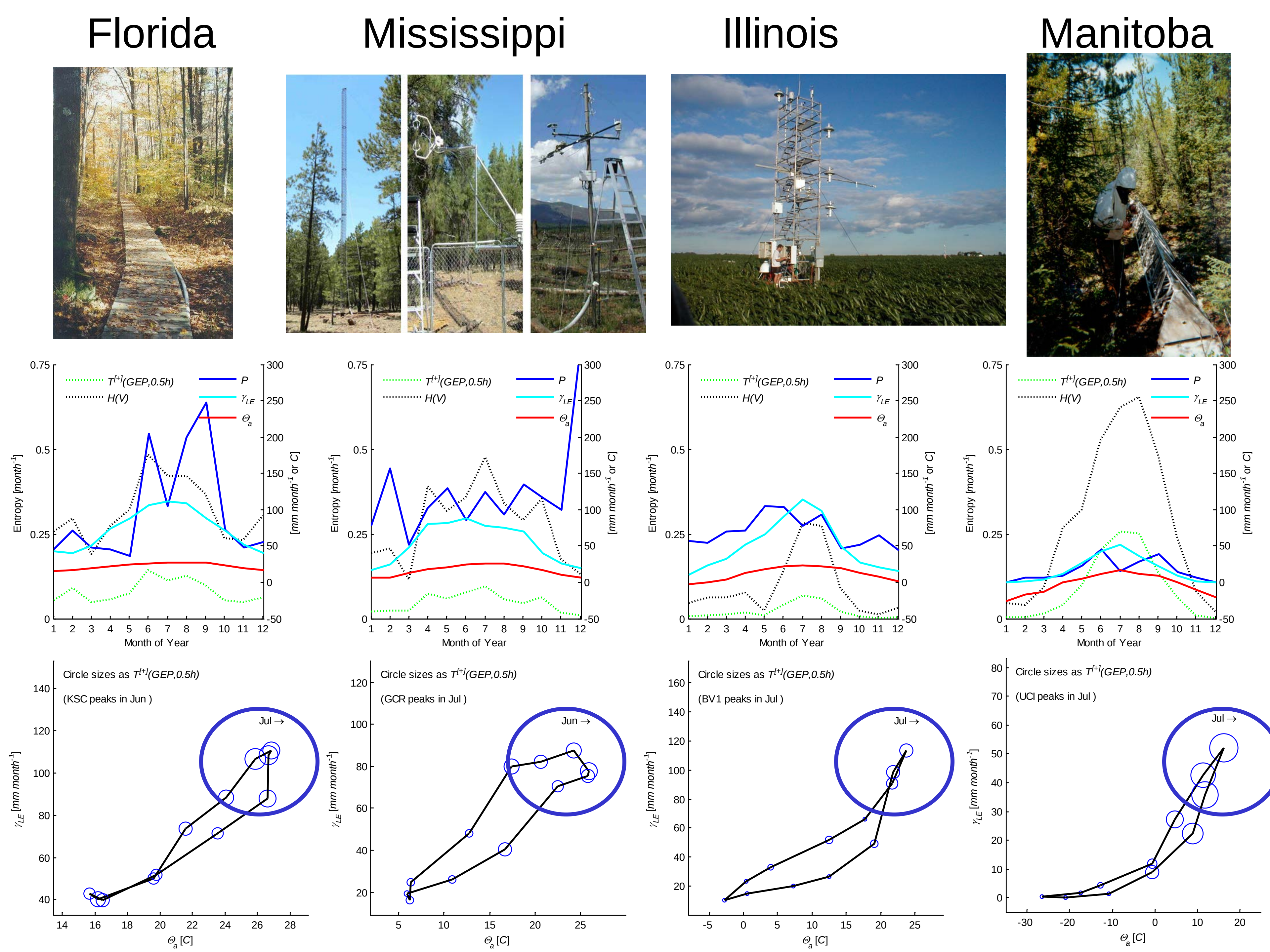


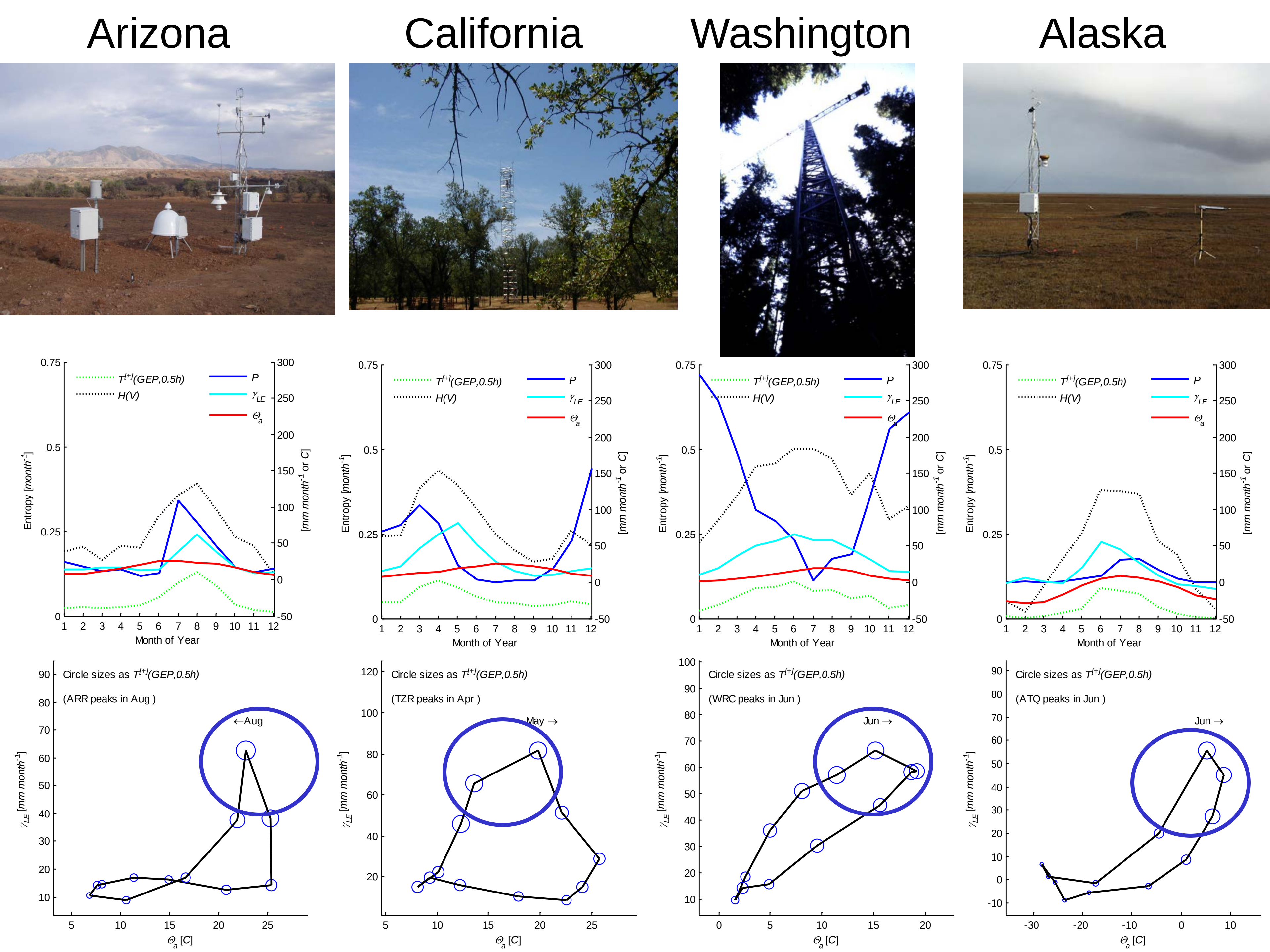




	T`net(X)	T`[+](X)	T`[-](X)
Tair	0.0097	0.17	**0.36**
Reco	0.058	0.49	**0.53**
Tsoil	0.043	0.21	**0.23**
VPD	0.13	0.76	**0.77**
SWC	**0.22**	0.14	0.039
GEP	0.74	**0.84**	0.37
LE	0.38	**0.81**	0.76
H	0.00029	**0.57**	0.4
Rshort	0.25	**0.63**	0.59
Precip	0.0086	**0.16**	0.12

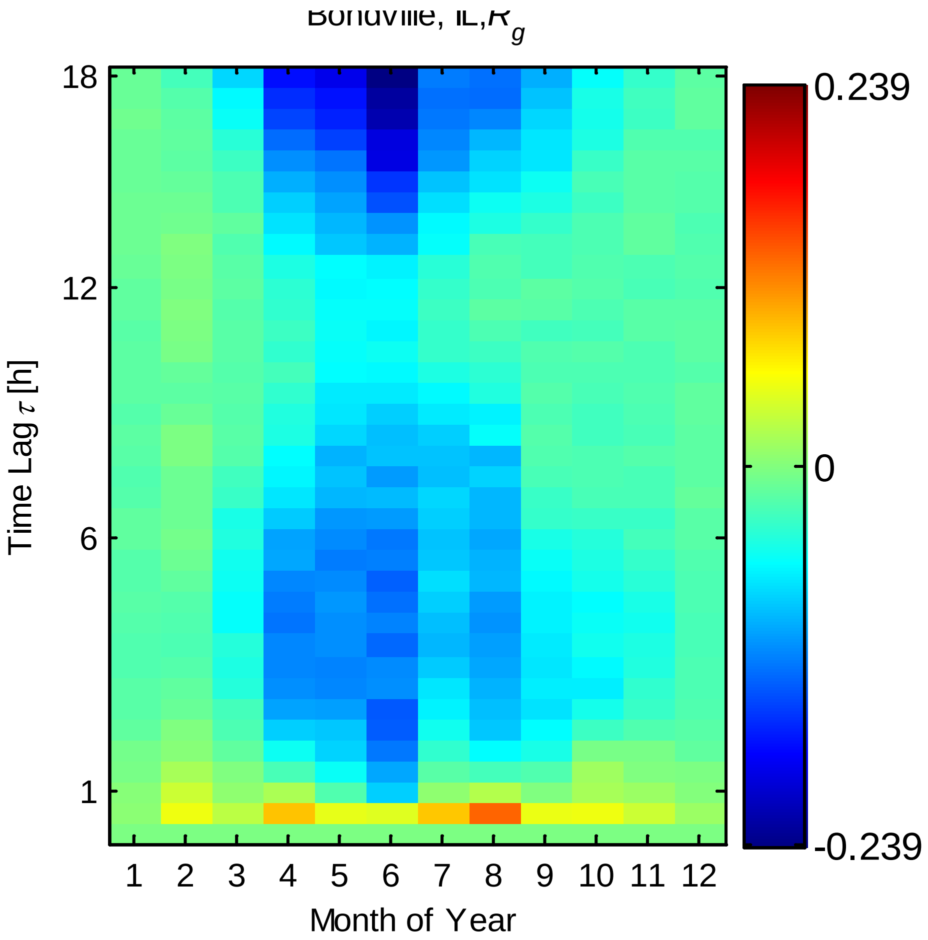



Florida
Mississippi
Illinois
Manitoba
$T^{[+]}(GEP,0.5h)$
$H(V)$
P
γ_{LE}
Θ_a
Entropy [month^{-1}]
[mm month^{-1} or C]
Month of Year
Circle sizes as $T^{[+]}(GEP,0.5h)$
(KSC peaks in Jun)
(GCR peaks in Jul)
(BV1 peaks in Jul)
(UCI peaks in Jul)
Jul →
Jun →
γ_{LE} [mm month^{-1}]
Θ_a [C]



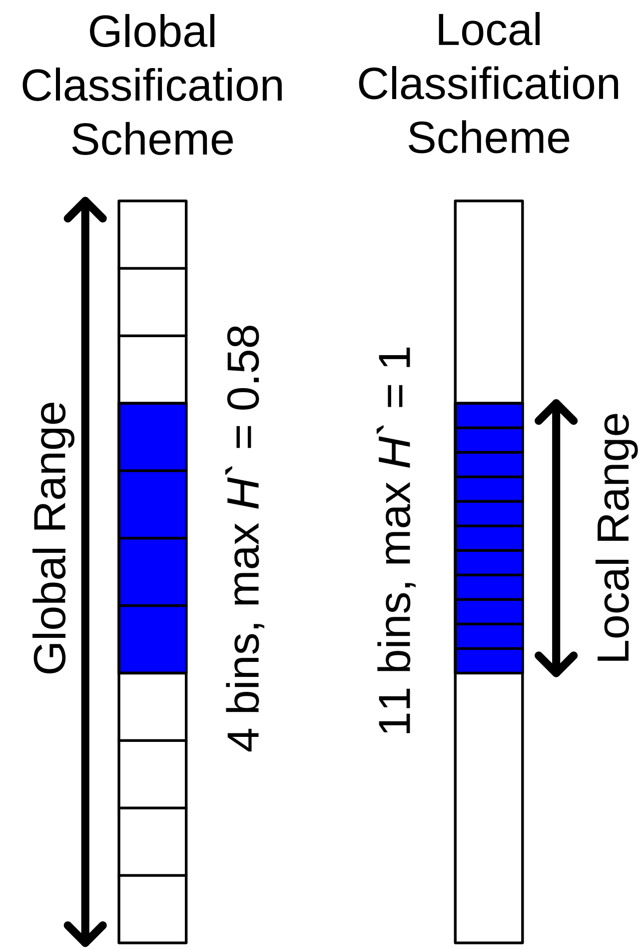
Arizona
California
Washington
Alaska
T[+](GEP,0.5h)
H(V)
P
γLE
Θa
Entropy [month-1]
[mm month-1 or C]
Month of Year
Circle sizes as T[+](GEP,0.5h)
(ARR peaks in Aug)
←Aug
(TZR peaks in Apr)
May →
(WRC peaks in Jun)
Jun →
(ATQ peaks in Jun)
Jun →
γLE [mm month-1]
Θa [C]

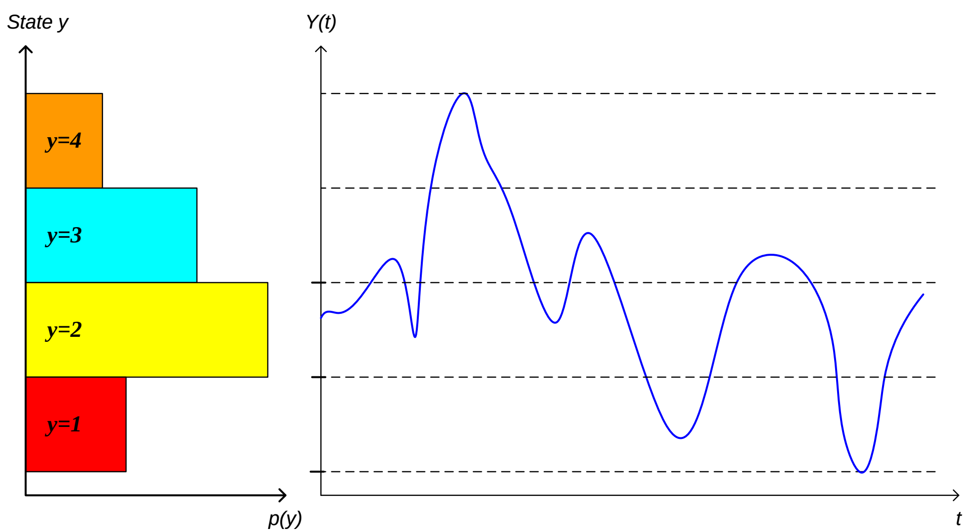
Net Information Flow Spectra



Local vs. Global Classes

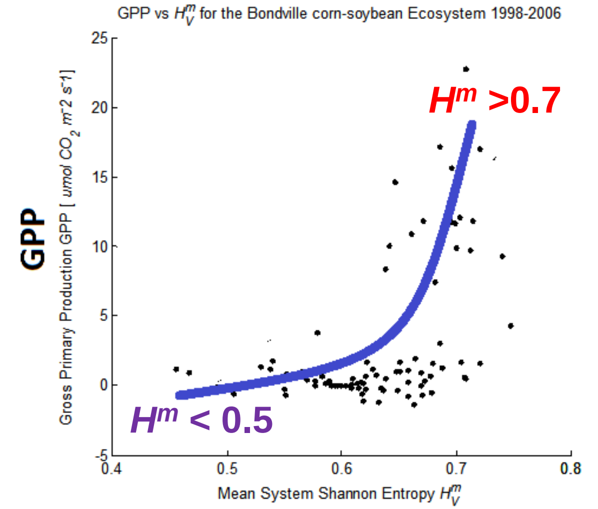
- Local scheme computes entropies with respect only to variation within each monthly period.
- In global scheme, entropies can be much lower because a variable may only visit a subset of global states during each month.
- Local scheme increases *H* for months that visit only a few global states.
- Effects of the filter are lessened for months that visit more global states.



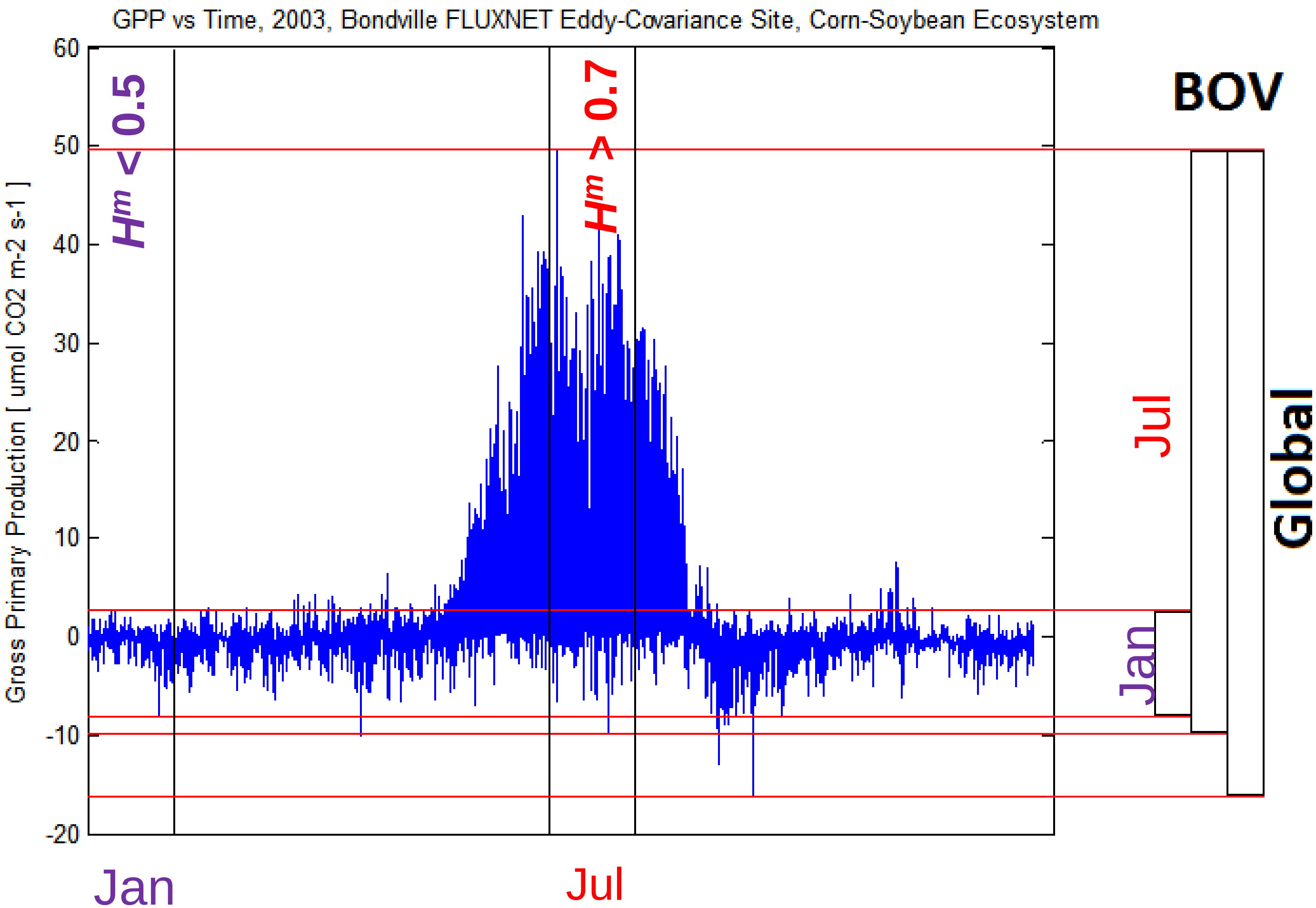


$$H(Y_t) = -\sum_{y \in Y_t} p(y) \cdot \log p(y)$$

Relationship of the Physical Bounds of Variability (BOV) to Shannon Entropy is Positive



Mean System Shannon Entropy



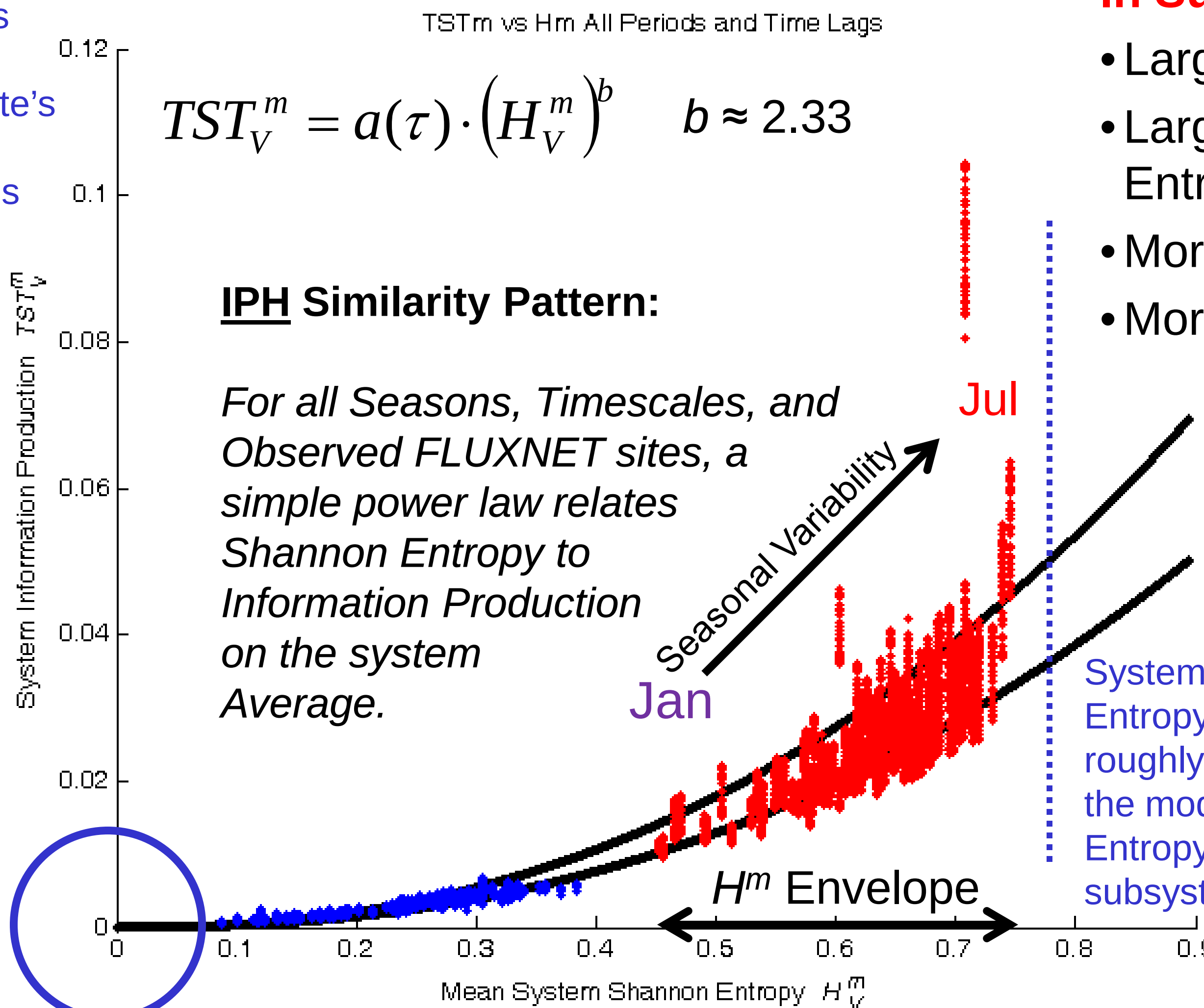
In Summer (Jul),

- Larger BOV
- Larger Shannon Entropy/ uncertainty
- More productivity
- More information

All data presented is 30-min L4 eddy-covariance measurements from FLUXNET sites (Baldocchi 2001, Reichstein et al. 2005)

For all systems and system states studied, System-Average Information Production is a function of System-Average Shannon Entropy

Each point on this plot represents a single system state's average process network properties (1 month = state)



In Summer (Jul),

- Larger BOV
- Larger Shannon Entropy/ uncertainty
- More information
- More productivity

a = 0.065, 1hr

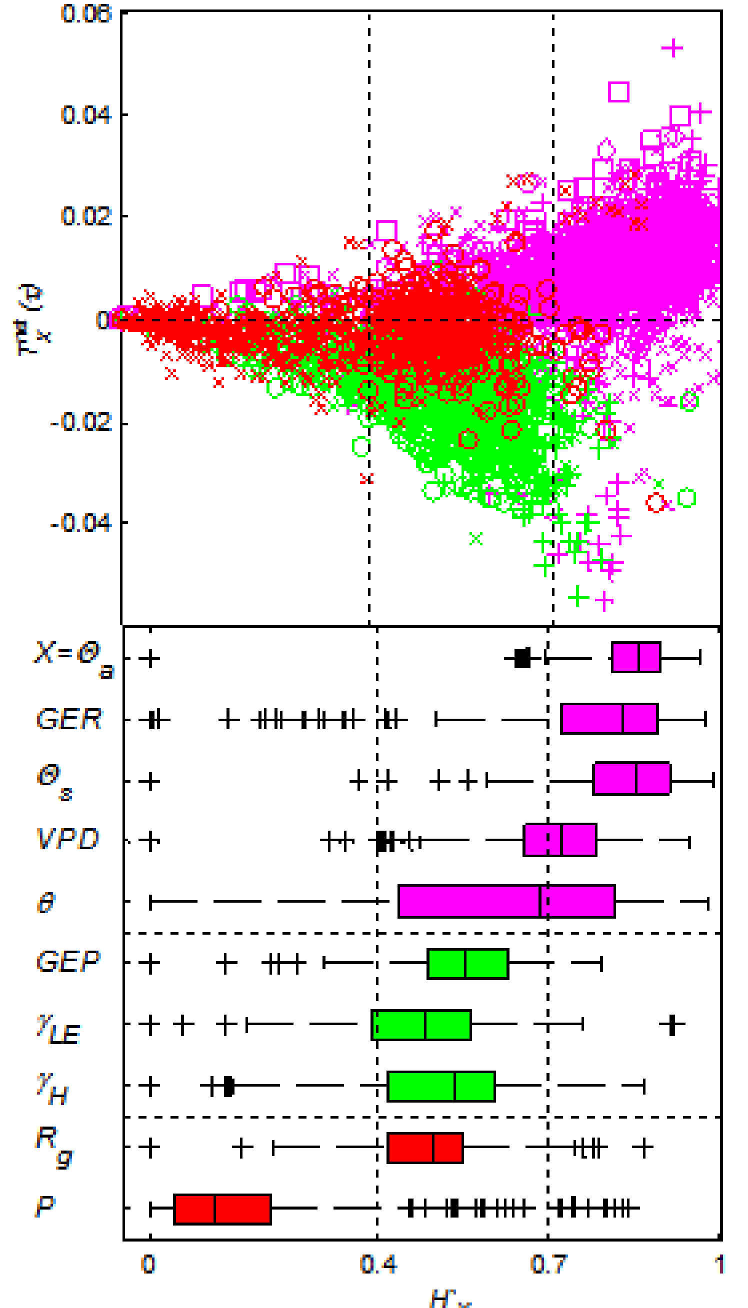
a = 0.090, 18hr

System average Shannon Entropy does not exceed roughly 0.75 because of the moderate Shannon Entropy of small-scale subsystems; see MEH.

Steady State Systems have H^m = 0 and do not produce any information.

Within a given system state, a second scale-free <u>MEH</u> similarity pattern holds

- Slower / Larger-Scale subsystems have higher Shannon Entropies and export net information to Faster/Smaller-Scale subsystems (e.g. synoptic).
- Faster / Smaller-Scale subsystems have moderate Shannon Entropies and are net consumers of information; these subsystems are those exhibiting Type-II self organizing logical relationships. (e.g. turbulent)
- Intermediate-Scale subsystems that loosely couple the large and small scales are informationally neutral and have low *H*. (e.g. ABL)
- Scale is relative; "small" means the system varies at the time and space scale at which the system is modeled or observed, in this case 30 minute dynamics of FLUXNET data.



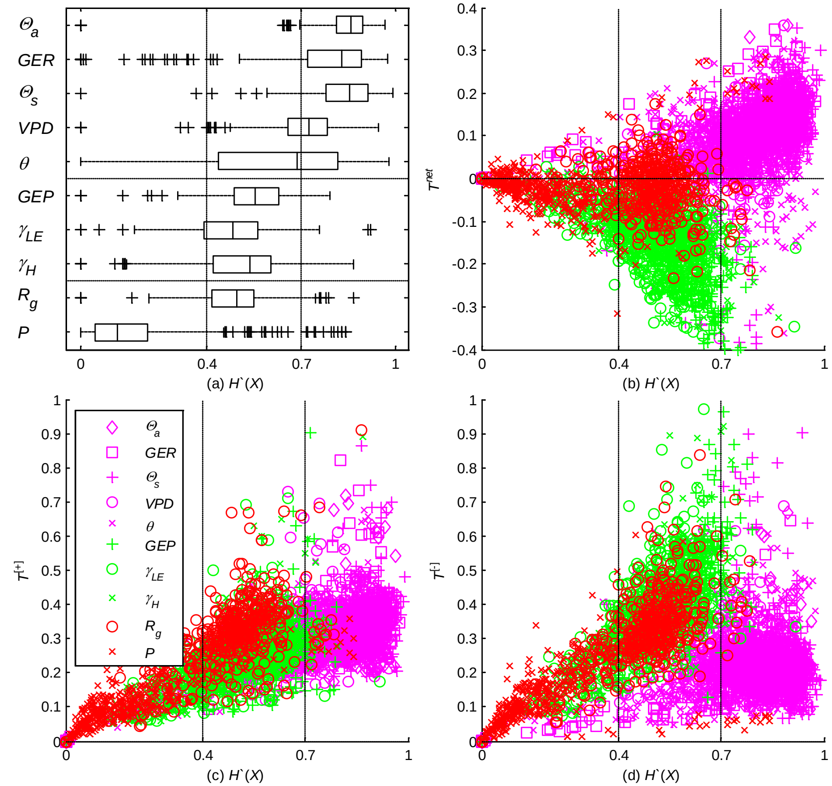
Observation: Variability Controls Organization in Open Dissipative Systems

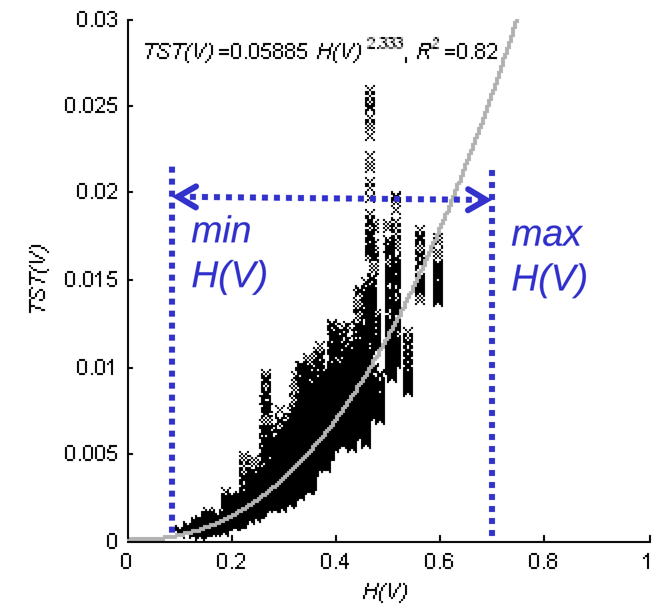
- System-averaged behavior: Total System Transport *TST(V)* of information is related to the mean Entropy of the system *H*(*V*) by a power law where a~0.058, b~2.33, using <u>global classes</u>.

- *H* is the control parameter and *TST* is the order parameter; <u>all</u> ecohydrological systems respond to variability in the same way. <u>Variability</u> itself relative to BOV is a universal organizing principle for dynamical system state definition and transition! [Kumar 2007, Kleidon 2007].

- Range of variability is relative to local climate; ecosystems adapt to the local variability regime as defined in the first order by Θ_a, *P*, and γ_{LE}.

- With <u>local classes</u>, $T^{net}(S)<0$, $H(X)<0.7$ for variables with local-scale feedback. High-entropy variables have $T^{net}(S)>0$, $H(X)>0.7$ are associated with larger temporal scales and forcing variables.





If time… Practice the lagged logistic map example using PNET 1.0