

# Information Analysis of Water Balance Modelling at Different Time Scales

Pan Baoxiang & Cong Zhentao

December 2, 2014

- Introduction
- Water Balance Modelling at Different Time Scales
- Information Analysis of Hydrological Simulation
- Data & Method
- Results
- Discussion

- Problems for Different Time Scale Hydrological Modelling
  - Ambiguity of Time Scale Divisions
  - Interpretation of Coarser Scale Constitutive Functions
- Information Theory Applied for Hydrological Model Evaluation

# Water Balance Modelling at Different Time Scales

Classification	Time Scale	Philosophy	Advantage	Disadvantage
Distributed (S.H.E.,etc.)	min/h	Integrate by parts along temporal and spatial paths	Physical Sufficient Information	Data Requirement No General Picture
Conceptual (XAJ.,etc.)	h/daily	Simplified Runoff Generation Convergence Theory	Easy to Calculate	Weak Theory Foundations
(TPWB,etc.)	monthly	Runoff Evapotranspiration Supply-Demand Framework	Easy to Calculate Simple Form	Weak Theory Foundations
(SARMA,etc.)	monthly	Auto Regression Mean Average Characteristic of Runoff	No Meteorological Data Required	Weak Theory Foundations
(Budyko,etc.)	annul	Water-Heat Correlation	Simple Form General Picture	Coarse Scale

We need to solve the following questions toward a self-consistent theory that could account for the existent multiscale phenomenological patterns:

- The Applicable Temporal Scales of Different Models
- Are The Models Compatible at the Margin Scales
- The Hydrological Information Provided by the Inputs
- How the Constitutive Functions of Different Models Capture This Information

# Focalize the Problem

Besides the consideration of practical significance, We focus especially on the monthly water balance models because they act as:

- Interim from Physical Perspective to Systematic Perspective
- Interim from Single-Phenomenon-Focused Model to Multi-Phenomena-Focused Model
- Interim from Iterative Model Structure to Non-iterative Structure

We try to use the information method introduce below to explain the gaps we clarified above.

- Basic Conceptions
- Theoretical Consideration
- Methodological Consideration

# Basic Conceptions

Term	Implication	Discrete Form	Continuous Form
Entropy	Information Content	$H(X) = -\sum p(x) \log p(x)$	$H(X) = -\int f(x) \log f(x) dx$
Conditional Entropy	Y's Information Contribution to X	$H(X Y) = -E \log p(x y)$	$H(X Y) = -\int f(x, y) \log f(x y) dx dy$
Mutual Information	X's Uncertainty Decrease due to Y	$I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$	$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy$

For both discrete and continuous forms,

$$I(X; Y) = I(Y; X) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y) \geq 0$$

## Data Processing Inequality

$$I(X; Y) \geq I(X; Z)$$

If  $X$  and  $Z$  are conditional independent given  $Y$



Gong(2012) defines aleatory uncertainty (AU) and epistemic uncertainty (EU) based on the terms introduced above:

$$AU = H(Observation) - I(Observation; Input) = H(Observation|Input)$$

$$EU = I(Observation; Input) - I(Observation; Simulation)$$

AU depicts the simulation uncertainty caused by insufficient data.  
EU depicts the simulation uncertainty caused by imperfect data processing methods applied on the input.

# Theoretical Consideration

# Theoretical Consideration

Any definition corresponds with our common sense on a specific issue and allows us to deal with that issue quantitatively. Suppose  $I$  represents data;  $M$  represents model,  $U$  represents the whole uncertainty,  $AU$  represents the aleatory uncertainty,  $EU$  represents the epistemic uncertainty, the common sense requires:

- $\frac{\partial U}{\partial AU} > 0$
- $\frac{\partial U}{\partial EU} > 0$
- $\frac{\partial AU}{\partial M} = 0$
- $\frac{\partial U}{\partial M} = \frac{dEU}{dM} AU$

# Theoretical Consideration

The differential entropy  $H(X)$  and discrete entropy  $H(X^\Delta)$  obey the following relation:

$$H(X^\Delta) \rightarrow H(X) - \log \Delta \quad \text{when } \Delta \rightarrow 0$$

if  $X^\Delta$  is the discrete stochastic variable scattering  $X$  into boxes with length of  $\Delta$ .

Thus,  $h(X) - \log \Delta$  depicts the information content required to describe  $X$  to  $(-\log \Delta)$ -bit accuracy. The differential entropy itself can not represent the average uncertainty of the information resource nor the average information provided by each datum. AU, which is the differential conditional entropy of the observation given the inputs, could not represent the remaining information content of the observation given the information of the inputs.

In mapping the hydrological pattern to the information space, we are trying to construct a normalized information criterion of the hydrological simulation set.

The Nash Sutcliffe Efficiency is a criterion paragon for offering the best (no bias) and worst (mean of the calibrate simulation series) benchmarks.

However, there is no benchmark for comparing the aleatory uncertainties of different models.

# Theoretical Consideration–EU

The continuous mutual information  $I(X; Y)$  has the distinction of retaining its fundamental significance as a measure of discrete information since it is actually the limit of the discrete mutual information of partitions of  $X$  and  $Y$  as these partitions become finer and finer. Thus it still represents the amount of discrete information that can be transmitted over a channel that admits a continuous space of values. The data process inequality also suits here.

The original EU maintains its implication of representing the simulation uncertainty caused by imperfect data processing methods applied by the model.

EU faces the same problem of no normalization.

$$AU_{normalized} = \frac{H(Observation, Input)}{H(Observation) + H(Input)}$$

$$EU_{normalized} = \frac{I(Observation; Input)}{I(Observation; Simulation)}$$

Properties:

- Range
  - $AU_{normalized} \in [0, 1]$  for  $H(Observation, Input) \geq 0$   
(non-negativity of mutual information)
  - $EU_{normalized} \in [1, \infty]$  (data processing inequality)
- Larger values, larger uncertainty.
- $AU_{normalized} \rightarrow 1$  when the the required accuracy  $\Delta \rightarrow \infty$ .
- Observation information content invariant.



# Methodological Consideration

Most of the hydrological models share an iterative pattern for the impact of soil moisture on the hydrological respond to the meteorological inputs. Thus, the input of a single calculating unit should include the hydrological terms of the former period, which would bring the dimension disaster in calculating the input's entropy.

The method Gong(2012) adapts for dealing with the curse of dimensionality is to take the ICA transformation, the steps are as follows:

- Implement FastICA to transform the original data Matrix  $X$  into independent components  $Y$ , where  $Y = A * X$ ;
- Calculate the entropy of each independent signal using a bin-counting method and sum them to obtain  $H(Y)$ ;
- $H(X) = H(Y) + \log|\det(A)|$ .

In the general formulation of ICA, the purpose is to transform an observed random vector  $X$  linearly into a random vector  $Y$  whose components are statistically as independent from each other as possible (Hyvarinen, 1997), thus, the above method would overrate the entropy for using  $H(y_i)$  to represent  $H(y_i|y_{rest})$  because the former is larger.

In application, the fast ICA could not be implemented for a large part of the hydrological series.

# Methodological Consideration

## Data

- MOPEX experimental basin

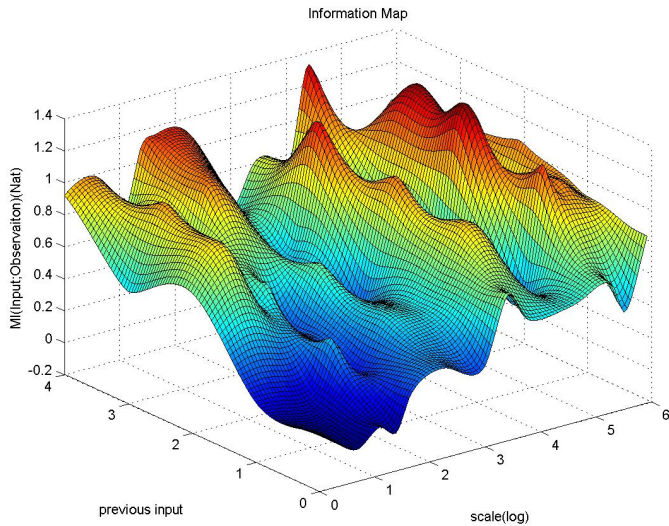
## Method

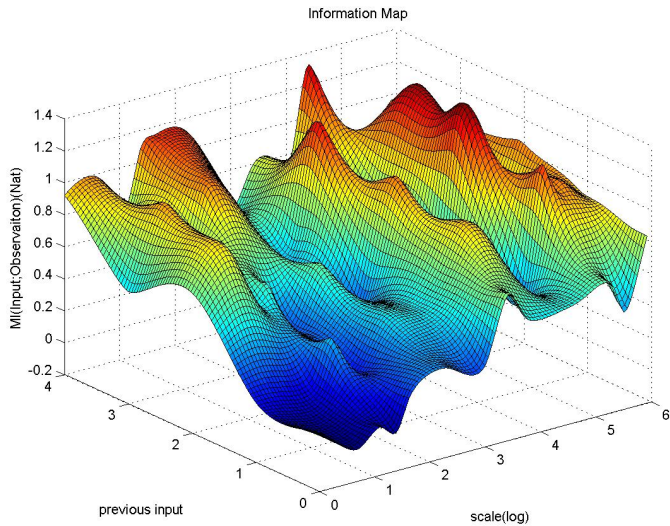
- Re-cluster the original hydrological data (daily precipitation potential evapotranspiration and runoff) into different time scale terms.
- Calculate the aleatory uncertainty at these time scales.
- Implement hydrological simulation and calculate the epistemic uncertainty.

Table: Calculated Terms

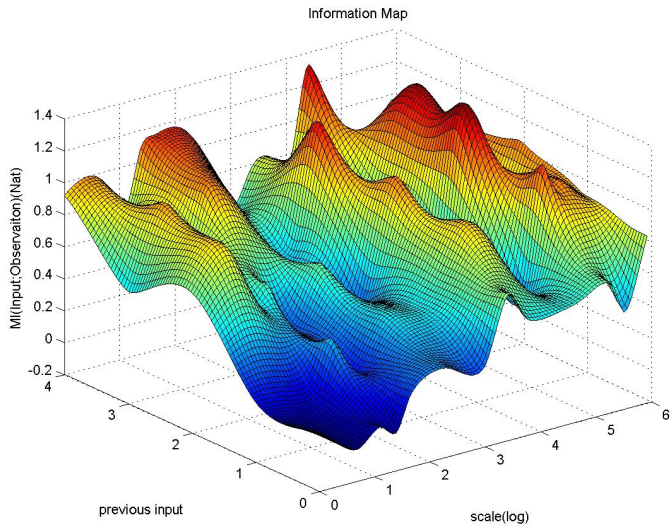
Terms variable	Model Invariant			Model Correlated		
	$MI(In; Obs)$	$AU$	$AU_n$	$MI(Simu; Obs)$	$EU$	$EU_n$
model scale	✓	✓	✓	✓	✓	✓
previous input	✓	✓	✓			
time period	✓	✓	✓	✓	✓	✓
current input & state variable	✓	✓	✓			

# $MI(\text{Input}; \text{Observation})$

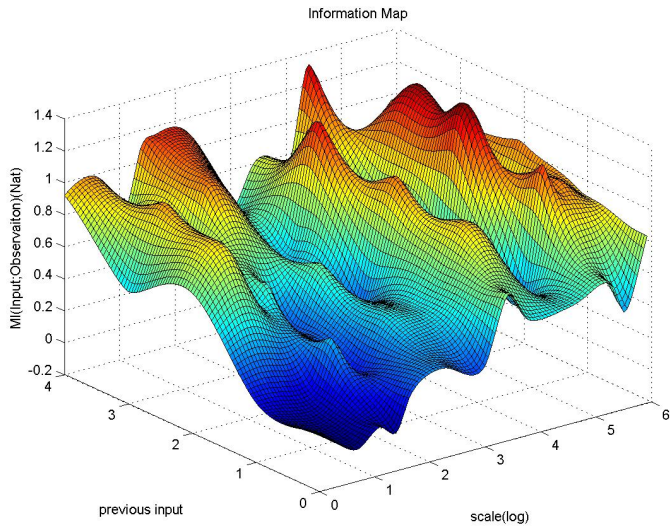


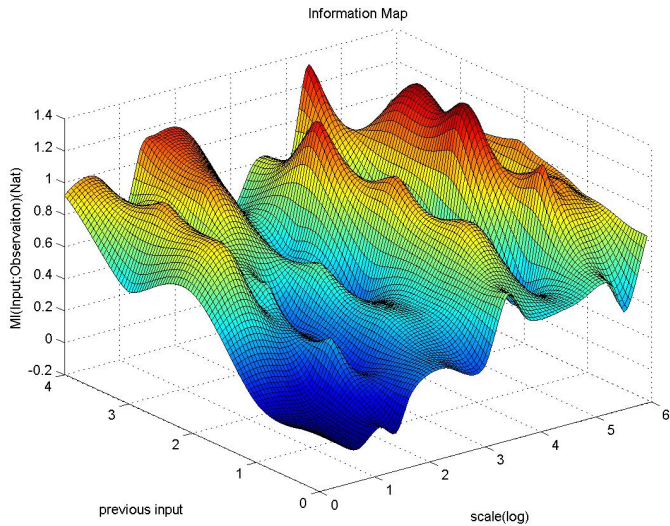


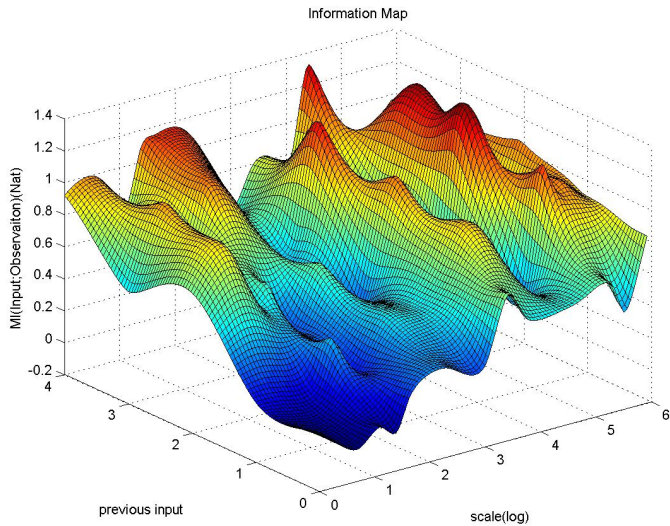




# MI(Simulation; Observation)









# Simulation\_Time-Mutual Information Slice

The uniformity of the Simulation\_Time-Mutual Information Slice represent a possible constant performance of an ideal model when applied in different atmospheric situations.

# Previous\_ Input-Mutual Information Slice

The first stationary point of the Previous\_ Input-Mutual Information Curve of small simulation scale represent the convergent time.

That of the larger simulation scale represent how the former hydrological condition effects the water movement.

The previous input value of the first stationary point becomes smaller and smaller as the simulation scale expands, disappears at the scale of 40 at this watershed. This is the time-scale where non-iterative model structure could provide satisfactory results.

# Simulation\_ Scale-Mutual Information Slice

The first stationary point of the Simulation\_ Scale-Mutual Information Curve of no previous input represent point when Budyko water-heat correlation dominates the hydrological circulation.



The result(which one?) cast doubt on the explanation that the failure of water-heat correlation at small time scales is due to the exclusion of soil moisture change. But for the generalized ignorance of different runoff-generation processes.

Mathematical hydrological models are implemented in the following two steps:

- Data Collection
- Data Processing

The aleatory uncertainty clarifies the uncertainty caused in the first step, the epistemic uncertainty accounts for the uncertainty of the latter step.