# Information Analysis of Water Balance Modelling at Different Time Scales

Pan Baoxiang & Cong Zhentao

January 16, 2015

# Outline

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

# Introduction

- 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<br>along temporal and spatial paths | Physical<br>Sufficient Information | Data Requirement<br>No General Picture |
| Conceptual (XAJ.,etc.) | daily | Simplified Runoff Generation<br>Convergence Theory | Easy to Calculate | Weak Theory Foundations |
| (TPWB,etc.) | monthly | Runoff Evapotranspiration<br>Supply-Demand Framework | Easy to Calculate<br>Simple Form | Weak Theory Foundations |
| (SARMA,etc.) | monthly | Auto Regression Mean Average<br>Characteristic of Runoff | No Meteorological Data Required | Weak Theory Foundations |
| (Budyko,etc.) | annul | Water-Heat Correlation | Simple Form<br>General Picture | Coarse Scale |

# Gap

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 introduced below to explain the gaps we clarified above.

# Information Analysis of Hydrological Simulation

- Basic Conceptions
- Theoretical Consideration
- Methodological Consideration

# Basic Conceptions

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

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 $Y$

# Basic Conceptions

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–Probability Space Definition

Considering seasonal fluctuation, the length of stationary hydrological series is not long enough to support an accurate estimation of high dimensional information terms, either with plug-in or non-plug-in method.

However, the information-based criteria is to evaluate the possible and achieved performance of models over the whole time domain. The probability space we study here is built on the measure terms of all the meteorological and hydrological conditions. Explicitly, we converge all the observation into one space and calculate the information terms.

# Theoretical Consideration–Probability Space Definition

By converging all the data generated in different seasons into one sample space, we will lose the information of their changes(period, trend, catastrophe) along time domain on one hand, but on the other hand, the information terms calculated over this probability space provide a general evaluation criterion of the observation system and simulation mechanism.

The definition of Nash Sutcliffe Coefficient (NSC) offers an analogy.The benchmark of a model's worst performance is represented by the mean of the observation during the calibration period without considering further information provided by that series.

# Theoretical Consideration–AU

The differential entropy $H(X)$ and discrete entropy $H(X^{\Delta})$ are constrained by the following relation:

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

where $X^{\Delta}$ 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.

# Theoretical Consideration–AU

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.

# Theoretical Modification

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

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

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.
- $AU_{normalized} \to 1$ when the the required accuracy $\Delta \to \infty$.
- 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.

# Methodological Consideration

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 $H(Y)$;
- $H(X) = H(Y) + log|det(A)|$.

# Methodological Consideration

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 Modification

We adapt a widely accepted non-plug-in method and make some improvements in order to estimate high dimensional hydrological terms' mutual information. The original method is derived from the *k* nearest neighbour entropy estimation approach (Alexander Kraskov, 2004):

$$I(X, Y) = \psi(k) - N^{-1} \sum_{i=1}^{N} [\psi(n_x(i) + 1) + \psi(n_y(i) + 1)] + \psi(N)$$

Here $\psi(x)$ is the digamma function, $\psi(x) = \Gamma(x)^{-1} d\Gamma(x)/dx$. k is order of nearest neighbour, $n_x(i)$ and $n_y(i)$ are the numbers of samples that are within the $k - th$ nearest criss-cross surrounding sample point $i$.

# Methodological Modification

An intuitive understanding of this method is that we select typical windows around each sample point to examine the concentration of that cluster space.

However, we should notice that the width of the windows is determined by the ordered *distance functions* we select to define the distances between samples. Since each dimension of a single sample represent different hydrological terms, the widely accepted *norms* could not reflect the *geodesic distances* of our hydrological modelling space.

# Methodological Modification

In order to make a justifiable distance between two samples in the modelling space, we define the *SVM Metric* as follows:

$$SVM\_Metric(x_1, x_2) = |f(x_1) - f(x_2)|$$

Here $x_1$ and $x_2$ denotes the input part of a sample, $f(x)$ is the support vector machine function that fit the input to the output of the sample. Evidently the definition satisfies the standards of *metric* (which are: non-negativity, identity of indiscernibles, symmetry, triangle inequality).

# Data & Method

Data

- MOPEX experimental basins

Method

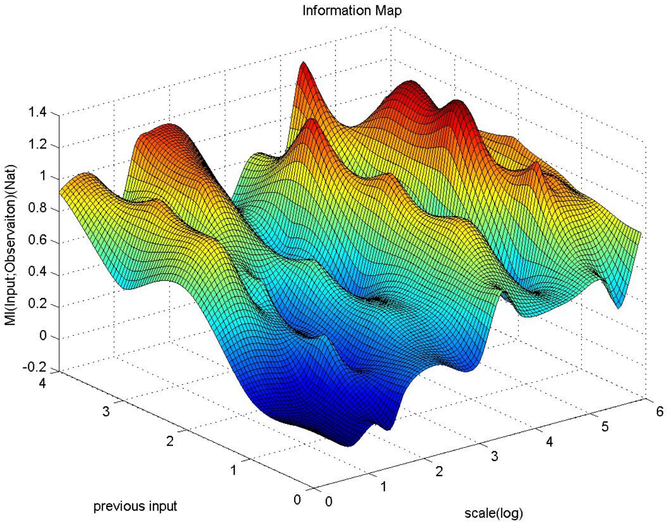
- Re-cluster the original hydrological data (daily precipitation potential evapotranspiration and runoff) into different time scale terms.
- Calculate the aleatory uncertainty at these time scales.
- Implement hydrological simulation and calculate the epistemic uncertainty.

# Results

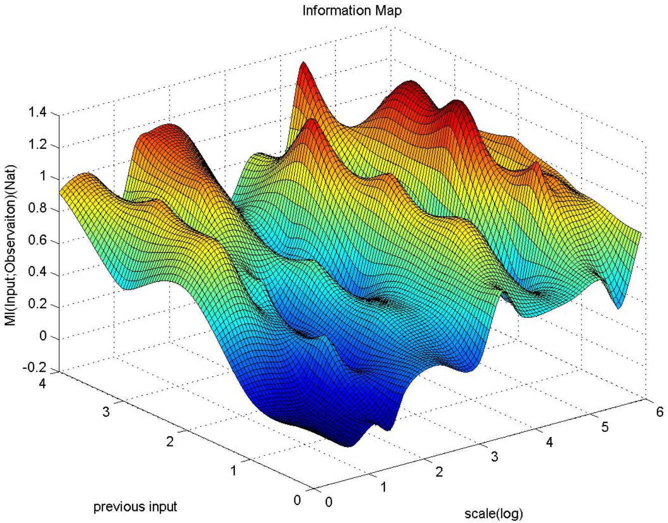
Table: Calculated Terms

| Terms<br>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<br>& state variable | √ | √ | √ | | | |

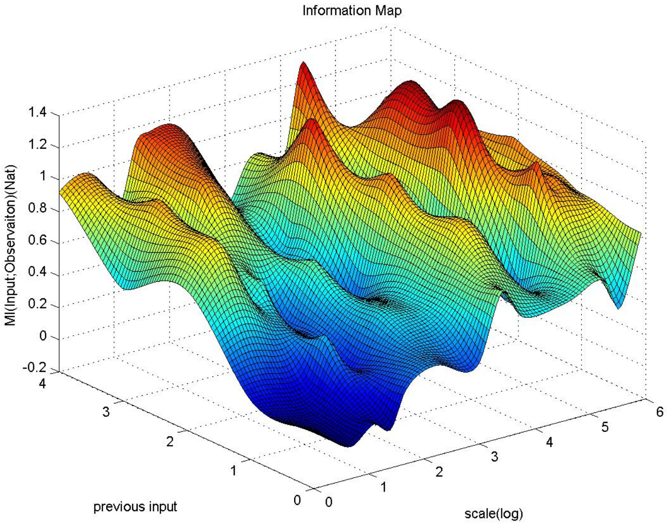
# $MI(Input; Observation)$



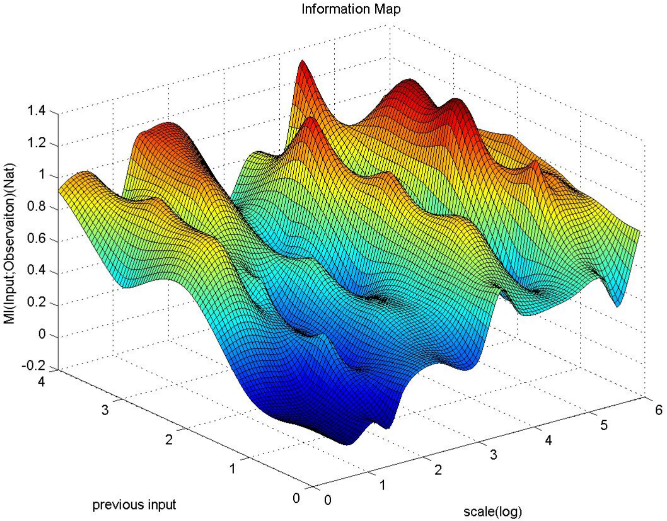
# AU


Information Map
MI(Input;Observaiton)(Nat)
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
4
3
2
1
0
previous input
0
1
2
3
4
5
6
scale(log)

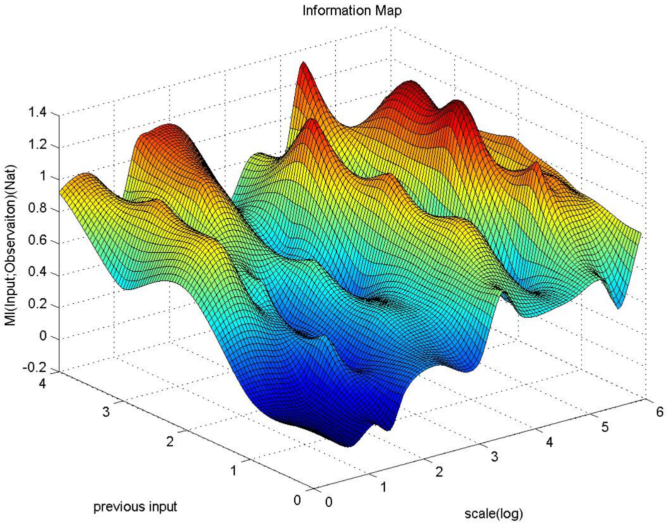
# $AU_{normalized}$



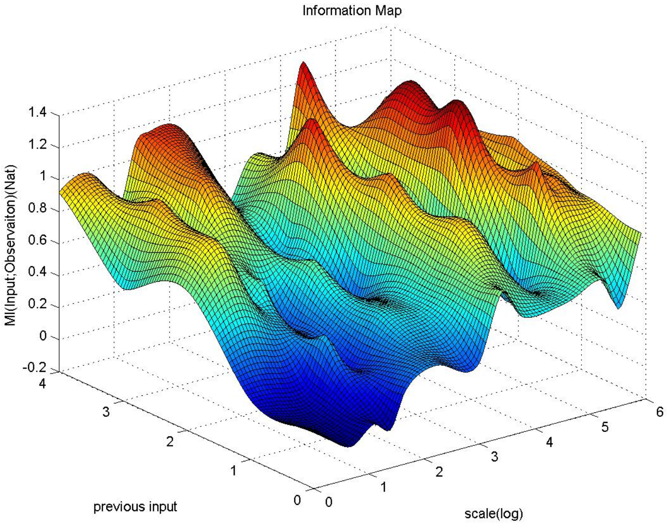
# $MI(Simulation; Observation)$



# *EU*


Information Map
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
MI(Input;Observaiton)(Nat)
4
3
2
1
0
previous input
0
1
2
3
4
5
6
scale(log)

# $EU_{normalized}$



# Discussion

# 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.

# Discussion

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.

# Discussion

If there are enough samples that enable us to contempt the dimension curse, a simple bin-counting method would support an accurate estimation of information terms that flow from the inputs to the outputs. Otherwise, we have to employ our knowledge to replenish the gaps and holes of the missing data. In this research, this is done by applying the support vector machine to find the "true" distance between two sample points, and use the distances to estimate the samples' uncertainty, or by the other word, information.