Towards A General Theory of Learning with Models & Data

“An Information Theory Perspective on Uncertainty & Learning”

Hoshin V Gupta (University of Arizona)

2017 RE Horton Lecture, Annual Meeting of the American Meteorological Society
Seattle, Washington, Jan 25
Thank You
Information Theory Perspective on System Identification

Water Resources Research

COMMENTARY
Debates—The future of hydrological sciences: A (common) path forward? Using models and data to learn: A systems theoretic perspective on the future of hydrological science

Hoshin V. Gupta¹ and Grey S. Nearing²,³

RESEARCH ARTICLE
The quantity and quality of information in hydrologic models

Grey S. Nearing¹ and Hoshin V. Gupta²
Some Main Collaborators

Grey Nearing           Steven Weijs              Uwe Ehret

Beth Jackson           Gong Wei                 Jasper Vrugt
Martyn Clark           Gab Abramowitz          Ben Ruddell
Shervan Gharari
In Particular

Grey Nearing

HYDROLOGICAL SCIENCES JOURNAL – JOURNAL DES SCIENCES HYDROLOGIQUES, 2016
VOL. 61, NO. 9, 1666–1678
http://dx.doi.org/10.1080/02626667.2016.1183009
Special Issue: Facets of Uncertainty

A philosophical basis for hydrological uncertainty
Grey S. Nearing\textsuperscript{ab}, Yudong Tian\textsuperscript{bc}, Hoshin V. Gupta\textsuperscript{d}, Martyn P. Clark\textsuperscript{a}, k
My Background

SYSTEMS THEORY

An interdisciplinary/multiperspectival scientific domain that aids to derive and formulate those principles that are generic to all fields of scientific inquiry. This pursuit is based on several fundamental assumptions. First, all phenomena can be viewed as a web of relationships, not as isolated objects, events, or a system. Second, a number of these are interrelated, such that the logical or social, or both.
My Interest is the “Learning” Problem

✗ Models as a Strategy for Predicting Behaviors or Events

✔ Models as a Strategy for Learning
\[ \frac{dX(t)}{dt} = U(t) - Y(t) \]

\[ Y(t) = K \cdot X(t) \]
model

\[
\frac{dX(t)}{dt} = U(t) - Y(t)
\]

\[
Y(t) = K \cdot X(t)
\]
OUTLINE

1. Models - Dynamical Environmental Systems (DES)
   - DES models as Tools for Scientific Investigation

2. Information - Its Fundamental Nature & Different Kinds
   - Data, Models and Assumptions are different, but related, kinds of (Uncertain) Information

3. Learning - How Information & Uncertainty are Related
   - Learning involves “Change”. Can Information be “Bad”? 

4. Structure of Information - How DES Models Encode Knowledge
   - Information is encoded as a Structured Hierarchy of Hypotheses

5. Model “Failure” - What Could Possibly Go Wrong?
   - Bias and Overconfidence

6. Inference – Learning From Our Mistakes
   - Different kinds of model-based Learning

7. Maximum Entropy Approach – Injecting Rigor into Learning
   - Dealing with Uncertainty in System Architecture & Process Parameterization

8. What Next? – Improving The Way We Use Models in Science
Models
Dynamical Environmental System Models (DESMs)

Working Definition:

A DESM is a simplified representation of the structure & function of a dynamical system that:

1. Enables
   (a) Simulations that are acceptably accurate
   (b) Testable Predictions under new circumstances
   (c) Reasoning within an idealized framework

2. By Encoding Knowledge about
   (a) Physics (conservation, thermodynamics, etc.)
   (b) System Properties (Geometry & Materials etc.)
   (c) Uncertainty (What we know that we don’t know)

Why Simplified?

a) Knowledge is Incomplete & Uncertain
b) Real system is Infinite Dimensional
c) Need to “compute” in Finite Time using Finite Resources
We Use DESMs for Scientific Investigation

Intuitively

We understand that “Models” & “Data” codify Knowledge about the World ... in the form of Information

How can I use this insight to improve my models?

How does “Learning” happen?
Information
What is The Nature of Information?

Info is the answer to questions such as:
When, What, Where, How, Why ...
& How Much ... etc.
Information is Relational / Contextual

Info is always “about” something

Context Matters

DATA is not Info ... until viewed in context

3.14

What?
Info is always “about” something

Context Matters

DATA is not Info … until viewed in context

Streamflow

(\text{mt}^3/\text{sec})

3.14
So Information is always “ABOUT” Something ...

In the Context of Model-Data Learning

Info is always about
- Values (Y)
- Relationships (R: X \( \rightarrow \) Y)
- Constraints (C) ...
  Assumptions are a kind of Constraint

A single DATA Point encodes Info about:
A Value of “Something”
So Information is always “ABOUT” Something ...

In the Context of Model-Data Learning

Info is always about

- Values (Y)
- Relationships (R: X → Y)
- Constraints (C) ... Assumptions are a kind of Constraint

A set of DATA Points encodes Info about:
The **Distribution** of Values
So Information is always “ABOUT” Something ...

**In the Context of Model-Data Learning**

**Info is always about**
- Values (Y)
- Relationships (R: X → Y)
- Constraints (C) ... *Assumptions are a kind of Constraint*

A set of DATA Points encodes Info about:

The space-time-ordered **Relationships** among those values
So Information is always “ABOUT” Something ...

In the Context of Model-Data Learning

Info is always about
- Values (Y)
- Relationships (R: X → Y)
- Constraints (C) ... Assumptions are a kind of Constraint

A MODEL encodes Info about:
- The space-time-ordered Relationships between variables ‘behind’ those values

Mapping

\[ R : p(Y \mid X) \]
So Information is always “ABOUT” Something ...

In the Context of Model-Data Learning

Info is always about

- Values (Y)
- Relationships (R: X → Y)
- Constraints (C) ... Assumptions are a kind of Constraint

A MODEL encodes Info about:

Values \( p(Y|X,M) \) conditional on Relationships and Data
Learning
Simply put ... Learning Involves

Converting Data-Info \(\rightarrow\) Model-Info

---

** Info can be added in/by the “Conversion Process” (have to be very careful here)

** The conditioning role of Assumptions (A) is very strong
And Occurs When …

Our **Prior Uncertainty** is Changed
due to the Assimilation of New Information

\[ p_b(R) \Rightarrow p_a(R) \]

“Changed” … not Reduced
Kinds of Learning ...

Learning occurs when our **Prior Uncertainty is Changed** due to the Assimilation of New Information

“**Changed**” ... not Reduced
Can Information Can Be Bad?

There is a view (advanced in the literature) that Information can be “Bad” (so called “Mis-Information”)

But Information is Simply What it Is ... What Can be Suspect is Our Interpretation

What can this new information tell me?

How to Handle Different Kinds of Information is the focus of ESTIMATION THEORY
The Structure of Information (in DES Models)
DESM’s Encode Information As a Hierarchical Sequence of Decisions

1. Control Volume, Physics, Processes to Include, System Geometry & Material Properties

2. Scale, Dimension & 3D Spatial Structure

3. Process Relationships

4. Uncertainty

5. Solution Methodology
Step One – Conservation Laws

**Information About:**

1. Physics (conservation, thermodynamics), Physical Processes, System Geometry & Material Properties to include

**Question:** What knowledge to encode and what to ignore?

**Result:** A System Diagram and Conservation Law Hypothesis - $H^{CL}$
Step Two – System Architecture

Information About:

2. Scale, Dimension and 3D Spatial Structure of the State-Space (elements), to enable finite computation

**Question:** What is a sufficiently complex, finite dimensional, spatially organized representation of sub-system architecture?

**Result:** A System Architecture Hypothesis - $H^{SA}$
Step Three – Process Parameterization

Information About:

3. Process Relationships via Equations, that account for Sub-Element Process & Material Heterogeneity

**Question:** What mathematical forms to use for the Process Parameterization equations, at the architectural scale of interest?

\[ Y_{ij} = f_{xy}(\Delta X_{ij} \mid \theta_{xy}) \]

*Parameters are artifacts of equation choice*

**Result:** A Process Parameterization Hypothesis - \( H_{PP}^{PP} \)
Step Four – Uncertainty

Information About:

4. What we KNOW that we do not know precisely (or at all)

Question: What uncertainties are important, and how to represent them mathematically?

Result: An Uncertainty Hypothesis - $H^{UN}$
Step Five – Solution Procedure

Information About:
5. Procedure for ‘Solving’ the resulting Mathematical Model

Question: How to Integrate (in space & time) the resulting system of (stochastic) differential equations?

Result: A Computational Model

→ Practical manifestation of the Overall System Hypothesis - $H^{OS}$
→ Structured hierarchy of Conservation Law, System Architecture, Process Parameterization, and Uncertainty Hypotheses - $H^{OS} = \{H^{UN} | H^{PP} | H^{SA} | H^{CL}\}$

2. System Architecture (a) further restricts trajectories & (b) determines spatial variability

3. Process Parameterization (a) further restricts trajectories & (b) introduces “tunable” parameters

4. Specification of Uncertainty characterizes and quantifies “known unknowns”

5. Solution Procedure converts Model Info & Input Info into specific (uncertain) X-Y trajectories
Model “Failure”
What Could Possibly Go Wrong?
What Could Go Possibly Wrong?

1. Problem Becomes Over-Constrained
   
   Due to Hypotheses that are Unjustifiably Strong
   
   a) Neglect Heterogeneity that is important
   b) Over-simplify the System Architecture
   c) Incorrect Process Equations forms
   d) Deterministic Process Parameterizations (instead of Stochastic)

2. Problem Remains Under-Constrained
   
   Due to Lack of Knowledge
   
   a) Do not know Process Physics at the scale of system elements
   b) Do not know Heterogeneity of Material Properties and Geometry at scale of system elements
   c) Do not know (& account for) Heterogeneity of Material Properties and Geometry at scales smaller than the system elements
What Can Go Wrong?

1. Problem Becomes Over-Constrained
   Due to Hypotheses that are Unjustifiably Strong

   Usually some Combination of Both

2. Problem Becomes Under-Constrained
   Due to Lack of Knowledge
Conceptual Illustration

IF ALL GOES WELL WE CONVERGE AROUND THE DATA
Conceptual Illustration

IF OUR HYPOTHESES ARE “WRONG” ...
Inference ...
Learning by
“Try ... Assess ... & Try Again”
Make Hypothesis

Model Hypothesis

System
Make Hypothesis ➔ Collect System Data

\[ Y = Y_{\text{obs}} \]

Model Hypothesis

System

Data

\[ U_{\text{obs}} \]

\[ Y_{\text{obs}} \]
Make Hypothesis → Collect System Data → Assess Likelihood

Add Hypothesis on Data Uncertainty
Select Inference Rule

Compute “Likelihood”
Try to Improve ("Fix") the Hypothesis

Data

System

Model Hypothesis

Add Hypothesis on Data Uncertainty

Select Inference Rule

Compute "Likelihood"

Y = Y_{obs}

Improve (Correct) the Model Hypothesis

© Hoshin Gupta, The University of Arizona
1. Basic → Check Assumptions about Data

Make Sure that the “Likelihood” Metric Accurately Reflects the Stochastic Nature of the Information Provided by the Data

Maximize $L(\text{Model} | \text{Data})$
2. Easiest → Parameter/State Estimation

\[ Y_{\text{obs}} = Y \]

\[ U_{\text{obs}} \]

System

\[ X_1, X_2, X_3, X_5 \]

Model

Hypothesis

Maximize

\[ L(X, \theta \mid H^{\text{UN}}, H^{\text{PP}}, H^{\text{SA}}, H^{\text{CL}}, \text{Data}) \]

Assumed Correct

Search over all feasible values
3. Harder $\Rightarrow$ Process Parameterization Estimation

System

Data

$U_{obs}$

$Y_{obs}$

Maximize $L(H^{PP} \mid H^{UN}, H^{SA}, H^{CL}, Data)$

Assumed Correct

Re-Read Literature

More Field Work

Try other Guesses
4. Harder Still → System Architecture Estimation

How to Distinguish the effects of Architecture from Parameterization?

Maximize $\mathcal{L}(H^{SA}, H^{PP} | H^{UN}, H^{CL}, Data)$

Assumed Correct
“Maximum Entropy”

The “ME” Approach to DESM Learning

Grey Nearing
Uwe Ehret
Shervan Gharari
What is a Maximum-Entropy (ME) Approach?

DESM’s Code Information as a Hierarchical Sequence of Decisions

In ME → At each stage, we adopt an Informationally Justifiable approach

In other words ...

Try to not build Overly Strong Assumptions into either the Model or the Inference Procedure

Use “Maximal Entropy” Assumptions

That add only as much Info to the Model as is justified by available evidence at the model relevant space-time scale
Example → ME Process Parameterization

1. Any Flux Parameterization must have the general form

\[ Y_t = K_t^{xy} \cdot X_t \]

where \( X_t \) is the gradient to be dispersed and \( K_t^{xy} \) is the conductivity of the medium.

Basic Principle of Thermodynamics

Assumes: Medium is “homogenous”
Gradient established “instantaneously”
Example → ME Process Parameterization

1. A Flux Parameterization must have the general form

\[ Y_t = K_t^{xy} \cdot X_t \]

where \( X_t \) is the gradient to be dispersed and \( K_t^{xy} \) is the conductivity of the medium.

2. Condition \( 0 \leq Y_t \leq X_t \) must hold to preserve mass balance, implying that \( 0 \leq K_t^{xy} \leq 1 \).
Example → ME Process Parameterization

1. A Flux Parameterization must have the general form

\[ Y_t = K_t^{xy} \cdot X_t \]

where \( X_t \) is the gradient to be dispersed and \( K_t^{xy} \) is the conductivity of the medium.

2. Condition \( 0 \leq Y_t \leq X_t \) must hold to preserve mass balance, implying that

\[ 0 \leq K_t^{xy} \leq 1 \]

3. \( K_t^{xy} \) is a monotonic non-decreasing or constant function of \( X_t \).

Consistent with physical principle that

Larger gradients → Larger fluxes
Example ➔ ME Process Parameterization

1. A Flux Parameterization must have the general form

\[ Y_t = K_{t}^{xy} \cdot X_t \]

where \( X_t \) is the gradient to be dispersed and \( K_{t}^{xy} \) is the conductivity of the medium

Because sub-element conditions will generally be different (in an unknown manner)
each time the gradient \( \Delta X \) is applied

4. \( K_{t}^{xy} \) is a Probabilistic function of \( X_t \)

\[ K_{t}^{xy} \sim p(K_{t}^{xy} \mid X_t) \]
Example $\rightarrow$ ME Process Parameterization

1. A Flux Parameterization must have the general form
   \[ Y_t = K_{tx}^{xy} \cdot X_t \]
   where $X_t$ is the gradient to be dispersed and $K_{tx}^{xy}$ is the conductivity of the medium.

2. Condition $0 \leq Y_t \leq X_t$ must hold to preserve mass balance, implying that $0 \leq K_{tx}^{xy} \leq 1$.

3. $K_{tx}^{xy}$ is a monotonic non-decreasing or constant function of $X_t$.

4. $K_{tx}^{xy}$ is a Probabilistic function of $X_t$.
   \[ K_{tx}^{xy} \sim p(K_{tx}^{xy} | X_t) \]
In the Common Modeling Approach ...

**Deterministic Representations** of the Process Parameterization Equations impose **Overly Strong Assumptions** about what we really know regarding the actual nature of the process relationships **at the modeling scale**.

Invariably based on small scale field or lab studies.
Maximum-Entropy Parameterization

Imposes only the minimal information required by Physics

PPE’s are randomly selected from a Maximum-Entropy Distribution of Random Functions that is Constrained to obey Physical Principles
This Actually Enables Better Inference of System Architecture

By minimizing the confounding effect of incorrect assumptions regarding the PPE’s

Run e.g., 10,000+ random cases (Monte Carlo on “Functions”)

Simulation Uncertainty constrained mainly by info in System Architecture $U(Q | H_{SA}, H_{CL}, P)$
Inferring System Architecture

Evaluate Performance using Data

Run e.g., 10,000+ random cases

Compute the (temporal) Likelihood of the Hypothesis given the Data

\[ L(H_{SA} | D) \]
Inferring System Architecture

Evaluate Remaining Uncertainty

Run e.g., 10,000+ random cases

Uncertainty

Compute the (temporal) Entropy of the Model Ensemble Simulations

\[ U(Y \mid H_{SA}) \]
Inferring System Architecture

Bootstrap & Plot the Results
For different System Architectures

- Uncertainty $U$
- Performance $-\ln L$
- Max $(U)$
- Max $(\ln L)$

Target

© Hoshin Gupta, The University of Arizona
Inferring PPE Form

Use Bayesian Updating to Select Equation Forms with Highest Performance (conditional on a selected Architecture)
Inferring PPE Form

Propose Mathematical Forms for Equations
(conditional on a selected Architecture)

$R : p(Y \mid X, \theta)$

Mapping

$\text{Max}$

$Y$

$Y$

$0$

$0$

$X$
Inferring PPE Form

Proceed with Parameter Estimation
(conditional on a selected Architecture & Parameterization Form)

\[ R : p(Y \mid X, \theta) \]

Mapping
Inferring PPE Form

Evaluate Remaining Uncertainty
(conditional on a selected Architecture & Parameterization Form)

Red: mapping $R: p(Y | X, \theta)$

Less certain about functional form

More certain about functional form
The Result

**Strategy to investigate**

**Model Structural Hypotheses**

*Without the need to make Strong Assumptions Regarding Process Parameterizations (Equations)*

*In Principle a similar approach could be used to investigate value of different Conservation Laws*
More Generally

Bring more “Honesty/Rigor” into the Model Building Process

1) Build Into the Model Clarity Regarding What We Feel Certain/Uncertain About

2) Be Clear about “What is Known” versus “What is Hypothesis / Assumption”

“Maximum Entropy Approach” To Model Building
Some Comments in Conclusion

1. Models & Data codify Information about the world

2. Information implies Change in Uncertainty about Something

3. Models are Hierarchical Assemblages of Hypotheses
   1. Conservation Laws
   2. System Architecture
   3. Process Parameterization
   4. Uncertainty

4. Model Hypotheses can be:
   1. Over-Constrained by un-justifiably strong hypotheses
   2. Under-Constrained by lack of knowledge about
      a) Scale-dependence of process relationships
      b) Sub-element heterogeneity

5. Model Structural Inference can be done using Max-Entropy PP’s

6. Process Equation Inference can be done using Bayes’ Law
And Finally ...

An **Information Theory** perspective can help improve the way we use models as Hypotheses for **Scientific Investigation**

How does this system function?
Thank You

Predictive Uncertainty

Learning

SYSTEMS THEORY

© Hoshin Gupta, The University of Arizona