Machine Learning & Information Theory for Model Benchmarking & Process Diagnostics

Grey Nearing
University of Alabama – Department of Geological Sciences
Motivation

1st Workshop on Information Theory in the Earth Sciences
Schneefernerhaus, KIT, Germany, 2016

2nd Workshop on Information Theory in the Earth Sciences
IH Cantabria, Santander Spain, 2018

Upcoming Summer School
Santander, Spain – June 03-08, 2019

CUAHSI
Universities Allied for Water Research

2019 SITES: Summer School on Information Theory in the Earth Sciences

Dear colleague,

It is our pleasure to invite you to the 2019 summer school “SITES: SUMMER SCHOOL ON INFORMATION THEORY IN THE EARTH SCIENCES”.

The course will be held in Santander, Spain, on 03-08 June 2019, at the Environmental Hydraulic Institute (IHCantabria) of the Universidad de Cantabria. The course is designed for PhD / Master students and postdocs interested in the fundamentals of Information Theory and its application in the Earth System Sciences. The course combines theory, practical exercises and discussion.

For further information, please see http://sites-2019.hcanabria.com/ or the attached flyer. The course fee EUR 600 and the deadline for course registration is 03 March 2019. The website also includes the application form. In case of further questions, please send an email directly to the course organizers cristina.prieto@unican.es.

Please feel free to forward this announcement to interested colleagues.

We look forward to hearing from you!

Kind regards,

Cristina Prieto on behalf of the organizing committee:

Cristina Prieto, Uwe Ebert, Hoshin Gupta, Praveen Kumar, Grey Nearing, Allison Goodwell, Ben Ruddell, Florian Waldman, Rui Pedigao and Steven Weiss

IT & ML for Model Diagnostics | G.S. Nearing – University of Alabama (2)
1995: ANNs out-perform a calibrated conceptual rainfall-runoff model.

2018: Land models use roughly half of the information contained in meteorological boundary conditions about half-hourly surface fluxes.

2015: Regressions with no state memory beat any of the major land surface models.

Observation data from diverse ecoclimates
Motivating Results: Hydrometeorology

2015: Regressions with no state memory beat any of the major land surface models. Observation data from diverse ecoclimates.

2018: Land models use roughly half of the information contained in meteorological boundary conditions about half-hourly surface fluxes.

Conceptual Model vs. Neural Network

IT & ML for Model Diagnostics | G.S. Nearing – University of Alabama
2019: Neural Networks make better streamflow predictions, on average, in **ungauged** basins than the best hydrology model in **gauged** basins.

Observation data from diverse catchments.
The success of machine learning relative to process-based modeling indicates that there is unused information in observation (training) data.
Why Not Replace Models with Machine Learning?

1) Oceanic Nino Index as a quantifiable externality.

2) ~70 Fluxnet towers globally with sufficient data to train/test neural networks for daily flux estimation (Qe, Qh, NEE, Soil Moisture).

3) At each FluxNet site, measure the strength of relationship between each ANN input and the flux response.

4) Are there differences in observed strengths of relationships between El Nino and La Nina periods? Can the ANNs capture these differences?

\[ \Delta_{\text{ElNino,LaNina}} = \frac{I_{\text{ONI} \geq 0.5}(y_{t+s}; x_t | y_t) - I_{\text{ONI} < 0.5}(y_{t+s}; x_t | y_t)}{I_{\text{all}}(y_{t+s}; x_t | y_t)} \]
Why Not Replace Models with Machine Learning?

Models were trained for each period (Neutral, El Nino, La Nina) using all data from the other two periods, and performance statistics were calculated out-of-sample (in the ENSO period not used for training).

Statistical significance ($\alpha = 0.05$) calculated against 200 random periods ($N = 500$) from the whole data record.
Although in some cases (e.g., streamflow), ML models out-perform process models even out-of-sample, but there are also cases where ML models fail to capture changes in process relationships.

How can we have the best of both worlds?
What's Wrong with the Models?

Predictive Performance: How well do simulated variables match observed variables?

Functional Performance: How well does the model simulate connections between different variables?

Claim: Pareto Tradeoffs between these two indicate model structural error.

Ruddell, Drewry, Nearing (in review), Information theory for model diagnostics: tradeoffs between functional and predictive performance in ecohydrology models, Water Resources Research.
What’s Wrong with the Models?

Observations

Model

Blue = Water Components
Warm = Energy Components
Green = Carbon Components

Difference

Underestimation

Overestimation

Timescale = 0.5 hours
What’s Wrong with the Models?

Ruddell, Drewry, Nearing (in review), Information theory for model diagnostics: tradeoffs between functional and predictive performance in ecohydrology models, Water Resources Research.

The stomatal slope parameter (mslope) represents the scaling of stomatal function with photosynthesis and ambient conditions at the leaf surface, and can vary as a function of vegetation type and within a growing season as a function of environmental conditions.

Comparison between information transfers in ML-Can vs. a Neural Network
2018: Land models use roughly half of the information contained in meteorological boundary conditions about half-hourly surface fluxes.
What’s Wrong with the Models?

Observation data from diverse ecoclimates.

Results:
All models appear to be wrong in the same ways at each site. Models are wrong for different reasons at different sites.
What about Data Assimilation?

Model:

\[ dx = \mu(x, u)dt + \sigma(x, u)dW_t \]

Data Assimilation:

\[ p(x_t | y_t) \propto h(y_t | x_t)m(x_t | u_{1:t}) \]

We lose all information from the observations at long forecast lead times.

Assimilation is good (at best) for improving initial states – it does not mitigate model structural uncertainty.
What about Data Assimilation?

Table 3
Breakdown of the Information-Use Efficiency Metrics From an EnKF Assimilation of LPRM Retrievals Into the Noah-MP Land Surface Model as Evaluated Against SCAN Data

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information in model simulations(^1)</td>
<td>(\frac{H(Z)}{H(Y)})</td>
<td>0.13</td>
</tr>
<tr>
<td>Information in retrievals(^1)</td>
<td>(\frac{H(Z)}{H(Y)})</td>
<td>0.08</td>
</tr>
<tr>
<td>Conditional information in retrievals(^1)</td>
<td>(\frac{H(Z</td>
<td>Y)}{H(Z)})</td>
</tr>
<tr>
<td>Total information from model and retrievals(^1)</td>
<td>(\frac{H(Z,Y)}{H(Z)})</td>
<td>0.18</td>
</tr>
<tr>
<td>Fraction of retrieval information lost via CDF-matching</td>
<td>(1 - \frac{H(Z</td>
<td>Y)}{H(Z)})</td>
</tr>
<tr>
<td>Information from EnKF</td>
<td>(\frac{H(Z</td>
<td>X)}{H(Z)})</td>
</tr>
<tr>
<td>Efficiency of EnKF ((\kappa_{DK}))</td>
<td>(\frac{H(Z)}{H(Z</td>
<td>X)})</td>
</tr>
<tr>
<td>Efficiency of EnKF ((\kappa_{r}))</td>
<td>(\frac{H(Z)}{H(Z,Y)})</td>
<td>0.03</td>
</tr>
</tbody>
</table>

\(^1\)Information metrics are normalized by the total entropy of the evaluation data \(Z\) so that their values range between 0 and 1. They are interpreted as, for example, “the fraction of total uncertainty about measurements \(Z\) that can be resolved given \(Y\) and \(X\).”

I see three major classes of methods to merge ML with physics:

1. **Learn physics from ML** – requires:
   - Deriving PDEs from ML models (long history of effort, no success)
   - Changing the way we think about physics (I see this starting to happen now – will require deep evolution in the philosophy of science, which we will likely do unconsciously)

2. **Impose physics on ML models**
   - Regularization in the loss function (penalize violating physical constraints)
   - Semi-supervised learning (e.g., adversarial networks)
   - Hierarchical constraints in model architecture

3. **Combine ML with physics models to solve PDEs**
   - System identification
   - Graph Analytics
Integrating Machine Learning with Process Modeling

\[
\frac{dx}{dt} = f(x, u, \theta) + \mathcal{GP}_\mu(x, u, \theta) + \mathcal{GP}_\sigma(x, u, \theta)
\]

Biogeophysics lives here
Structured information from observations goes here
Uncertainty lives here
Integrating Machine Learning with Process Modeling

Nearing, Gupta (2015), The quantity and quality of information in hydrologic models, *Water Resources Research*

HyMod: Calibration, Assimilation, and System Identification

**System ID: Convergence of the Mean Estimate**

**Streamflow Prediction Assessment**
Integrating Machine Learning with Process Modeling

Difference Between Process Connectivity in the Model before vs. after Assimilating Streamflow Obs.

- Streamflow
- Subsurface routing 3
- Subsurface routing 2
- Subsurface routing 1
- Surface routing
- Soil moisture
- Precipitation
- Potential evap.
- Soil moisture
- Surface routing 1
- Subsurface routing 2
- Subsurface routing 3

\[ I(y_{t+1}; x_t | y_2) = \int_{\mathbb{R}^n} \ln \left( \frac{p(y_{t+1} | x_t, y_2)}{p(y_{t+1} | y_2)} \right) \mu_{x_t=x_0,y_2} \]

IT & ML for Model Diagnostics | G.S. Nearing – University of Alabama (20)
System Identification can add nonlinearities to the internal model response surfaces.

Nearing, Gupta (2015), The quantity and quality of information in hydrologic models, Water Resources Research
I see three major classes of methods to merge ML with physics:

1. **Learn physics from ML** – requires:
   - Deriving PDEs from ML models (long history of effort, no success)
   - Changing the way we think about physics (I see this starting to happen now – will require deep evolution in the philosophy of science, which we will likely do unconsciously)

2. **Impose physics on ML models**
   - Regularization in the loss function (penalize violating physical constraints)
   - Semi-supervised learning (e.g., adversarial networks)
   - Hierarchical constraints in model architecture

3. **Combine ML with physics models to solve PDEs**
   - System identification
   - Graph Analytics
Objective: To understand where/why models aren’t perfect & improve them

• Uncertainty Assessment
• Hypothesis Testing
• Process Diagnostics
Uncertainty Assessment


**Data Processing Inequality**

\[ I(X; Y) \geq I(f(X), Y) \]

Results in an upper bound on the uncertainty due to model inputs and a lower bound on the uncertainty due to model structure.

**Information Reduces Entropy**

\[ H(X|Y) = H(X) - I(X; Y) \]

ML gives a lower bound on this quantity.
Information Reduces Entropy

\[ H(X|Y) = H(X) - I(X;Y) \]
Hypothesis Testing

\[ \mathcal{E} = I(z; u) - I(z, y) \]
\[ \hat{\mathcal{E}} = I(z; r(u)) - I(z, y) \]
\[ I(z; r(u)) \leq I(z; u) \rightarrow \hat{\mathcal{E}} \leq \mathcal{E} \]

Figure 2. Illustration of the concepts discussed in section 3.2. The leftmost bar indicates the amount of information \( I(Y_{\text{est}}) \) required to provide estimates of the system output. The middle bar indicates the amount of such information \( I(X_{\text{est}}, Y_{\text{est}}) \) contained in the available system input variables \( X_{\text{est}} \). The rightmost bar indicates the amount of explanatory information \( I(Y_{\text{est}}; Y_{\text{est}}) \) contained in the estimates \( Y_{\text{est}} \) obtained by use of the model hypothesis \( Y = f(X) \). The difference between length of the left and middle bars represents unresolved \( \text{AU} \). The difference between length of the middle and right bars represents potentially resolvable \( \text{EU} \).
Process Diagnostics

Over-Random  Over-Deterministic

Predictive Performance $A_p(y_j)$

Ideal

Functional Performance $A_f(y_i \rightarrow y_j)$

Difference Between Process Connectivity in the Model before vs. after Assimilating Streamflow Obs.

- Streamflow
- Subsurface routing 3
- Subsurface routing 2
- Subsurface routing 1
- Surface routing
- Soil moisture

IT & ML for Model Diagnostics | G.S. Nearing – University of Alabama
Big Picture: There is more to learn about hydrological processes.

ML tells us that there are patterns and information in hydrological data that we have either not noticed or not incorporated in our models.

My Prediction: data mining will take unless we learn how to marry the useful information from physical insight with the now-undeniable power of ‘universal approximators’.

---

Thank you.

The Monte Carlo Paradox
Why Information Theory?

What is IT Used For?
- Data Analysis
  - Spatial
  - Time Series
  - Data Uncertainty
- Complex Systems Analysis
- Model Evaluation
  - Benchmarking & Evaluation
  - Calibration Objective Functions
- Physical or Organizing Principle
  - Maximum Entropy
  - Maximum Entropy Production

Why IT vs. Other Metrics?
- Nonparametric
- Additive / Linear
- Multi-Dimensional
- Unique Epistemological Justification

Linearity
- $H(y)$
- $I(z;y)$
- $H(z|y)$
- $H(z)$

IT & ML for Model Diagnostics | G.S. Nearing – University of Alabama
Why Information Theory?

Before running the experiment:

After running the experiment:

\[ p(z_1 \land z_2) = p(z_2 | z_1) \times p(z_1) \]

\[ h(z_1 \land z_2) = h(z_2 | z_1) + h(z_1) \]
What about Data Assimilation?

\[ M \quad \text{Model Prediction} \]
\[ D \quad \text{Evaluation Data} \]

Model \[ H(M) \]
Eval. \[ H(D) \]

What about Data Assimilation?

\[ M \quad \text{Model Prediction} \]
\[ D \quad \text{Evaluation Data} \]

Model \( H(M) \)

Eval. \( H(D) \)

\[ \text{Information added by the retrieval.} \]

\[ \text{Assim. Obs.} \quad H(O) \]

What about Data Assimilation?

\[
\begin{align*}
M & \quad \text{Model Prediction} \\
D & \quad \text{Evaluation Data}
\end{align*}
\]

\[
\begin{align*}
\text{Model} & \quad H(M) \\
\text{Eval.} & \quad H(D)
\end{align*}
\]

Total information about the evaluation data

What about Data Assimilation?

\[ \epsilon_{DA} = \frac{I(D; M^+)}{I(D; (O, M))} \]

What about Data Assimilation?

Model State \((x)\)

- Sample of Model States
- Model Variance
- Observation
- Observation Error Variance

Time →
How to measure information supplied by model inputs?

The problem is dimensionality: The ‘inputs’ required to predict the state of a dynamical system at any point in time are typically multivariate time series (very high dimensional). Cannot directly estimate the joint probability density functions required to measure mutual information.

Instead, we approximate this with a ‘universal approximator’.

\[ I(Z; Y) = \int p(z, y) \ln \left( \frac{p(z, y)}{p(z)p(y)} \right) dzdy \]