Land hybrid modeling
Pierre Gentine – Columbia University
Wenli Zhao, Xu Lian - Columbia
Hybrid land modeling

Land-surface carbon and water fluxes

Sparse observations in time/space of carbon and water fluxes
Need global obs. to constrain surface processes in models
Hybrid land modeling

Land-surface carbon and water fluxes

Idea: use ML to define global mapping of surface fluxes (CO₂, water and energy) based on routine weather station data

![Diagram showing the process of using weather/satellite data through an ML algorithm to predict fluxes.

Weather/satellite data → ML algorithm (NN, RF) → Fluxes

Difficulty in assessing extremes!
Out-of-sample generalization issue

ML algorithm

NN, RF

Figure 3. (a) Scatterplots of observed data by eddy-covariance (y axis) and the median ensemble of modeled fluxes by RS setup (x axis). The panels from left to right were the 8-day predictions, the across-site variability, the mean seasonal cycle and the 8-day anomalies. The fluxes considered here were the gross primary production estimated following Lasslop et al. (2010), GPP_L (first row); the total ecosystem respiration estimated following Reichstein et al. (2005), TER_R (second row); the sensible heat, H (third row); and the latent heat, LE (fourth row). The reference units were g C m⁻² d⁻¹ and MJ m⁻² d⁻¹ for CO₂ fluxes (GPP_L and TER_R) and energy fluxes (H and LE), respectively.

(b) As in Fig. 3a, but the predictions (x axis) were obtained by the RS+METEO setup.

Another common issue with eddy-covariance data is the gaps generated by the data exclusion rules. Data exclusion strikes strongly the nighttime period (primarily for the low turbulence condition), affecting the representativeness of the diurnal cycle, and hence the quality of the averaged daily/8-day eddy-covariance fluxes, in particular CO₂. To reduce the risk of biased estimates, half-hourly data gaps are filled by models. In our study NEE data were gap filled using site-specific empirical relationships between meteorological data and net CO₂ ecosystem exchange (the MDS method, Reichstein et al., 2005) that produce small biases when short gaps were encountered (Moffat et al., 2007). This has a limited effect in this study as only a very small percentage of high-quality gap-filled data is used. We also minimize the bias in estimates of gross CO₂ fluxes (GPP and TER) by using two different partitioning methods that yield very consistent results.

4.4 Data quantity and representativeness

The mismatch between prediction and eddy-covariance estimation was also influenced by data representativeness. FLUXNET sites are not uniformly distributed over the globe and not all climates and PFTs are well represented. Very few sites are currently distributed in tropical forest, and data availability over the record is fragmented. Similarly, very few sites are located in the poorly predicted extreme environments, e.g., cold and dry climates. There was a clear pattern in our cross-validation results where more accurate predictions were obtained for the better represented vegetation types and climates (e.g., temperate and boreal forests). Therefore, increasing the number of study sites in less represented environments (e.g., the tropics and in the extreme climates) could improve the prediction by ML and models in general (Papale et al., 2015).
Hybrid land modeling

However: extremes are key for land carbon and water cycle

“A few extreme events dominate global interannual variability in global photosynthesis”

![Graph showing extreme events in global photosynthesis](image)

![Map showing changes in carbon uptake](image)
Issues with brute force ML

1. Do not respect physical laws
e.g. conservation of energy and mass
→ strict requirement

2. Issue with out-of-sample generalization
Important for many climate applications
 e.g. extremes, climate change

[Diagram showing ML algorithm input and output for within range prediction and out-of-sample prediction]
Potential Overcoming Strategies

Use of data

Data-driven ML
- Interpolation
  - Data Rich

Hybrid
- Moderate Data

Knowledge-Driven
- Extrapolation
  - Data Poor

Use of domain knowledge
Hybrid approaches

1. Physical constraints

Energy and mass conservations

Impose them within NN as function of inputs \((x)\) and outputs \((y)\):

\[
\{ C \left[ \begin{array}{c} x \\ y \end{array} \right] = 0 \}
\]

2 equations: reduce NN degrees of freedom to \(n-2\) degrees of freedom
Hybrid approaches

Constraining physics within ML

2. Land surface latent heat flux (LE) i.e. evaporation

Objective: predict LE from environmental variables

- Pure ML (feedforward NN) performs well
- But **does not conserve surface energy budget**
  \[ R_n - G \neq H + LE \]
- Generalization + out-of-sample issue
Hybrid approaches

Constraining physics within ML for better extrapolation

Land surface latent heat flux (LE) i.e. evapotranspiration

Objective: predict LE + conserve energy + respect diffusion

\[ LE = \rho \frac{e_s - e_a}{r_s + r_a} \]

- Hybrid ML performs as well as pure ML
- **Conserves surface energy balance 😊**
  \[ R_n - G = H + LE \]
- Can be used to learn physics: vegetation height and soil moisture primary controlling variables
Hybrid approaches

Constraining physics within ML for better extrapolation

Land surface latent heat flux (LE) i.e. evapotranspiration

Out-of-sample generalization/extremes

Test 1 and 99 percentiles

Hybrid systematically outperforms pure ML for extremes 😊
Hybrid approaches

Estimate of canopy interception

Canopy interception = LE\textsubscript{hybrid} (rainy model) - LE\textsubscript{hybrid} (rainless model)

Lian, Zhao, Gentine in review
Hybrid approaches

Estimate of canopy interception

mixed site-hour samples

rain events

Lian, Zhao, Gentine in review
Hybrid approaches

Global upscaling
Hybrid approaches

Physics Guided Machine Learning

Streamflow

Hybrid = VIC-CAMA Flood + LSTM (weather)

Physical model  Machine Learning model

Hybrid systematically outperforms physical model
Machine learning is an appealing approach for land process parameterizations

Issues:
1. Conservations, physical invariances, physical laws
2. Generalization, out-of-sample generalization

Solution:
**Hybrid physics + ML** approaches appear as powerful tool to tackle this
THANK YOU

Questions?

pg2328@columbia.edu
@PierreGentine

leap.columbia.edu