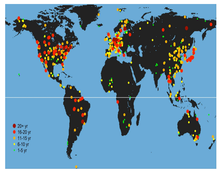
# Deep Learning to Address Data Sparsity in Climate Change Monitoring

**Shridhar Vashishtha,**
**PhD student mentor: Arvind Renganathan,**
**Faculty mentor: Vipin Kumar**

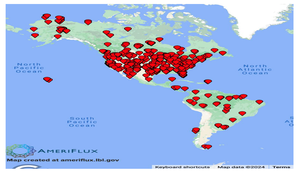
## Introduction

- Data sparsity is a challenge in climate change. There is a scarcity of flux towers that measure climate quantities such as GPP and $R_{eco}$. The motivation is to predict values for GPP and $R_{eco}$ with no labeled data since GPP and $R_{eco}$ are important quantities for carbon cycle predictions.
- The goal is to leverage knowledge from well-observed entities to improve predictions for unobserved ones, and to incorporate domain knowledge into machine learning models for better generalization.
- We present a deep-learning approach called Concatenation Long Short-Term Memory (CT-LSTM) to enable predictions for GPP and $R_{eco}$ using minimal labeled data.
- CT-LSTM leverages drivers (meta features/ characteristics) to transfer information from well-observed sites to unobserved ones which can then later be applied to these unobserved sites.



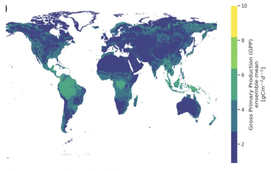
https://fluxnet.org/data/fluxnet2015-dataset/

**Fluxnet dataset of flux towers**



https://ameriflux.lbl.gov/about/about-ameriflux/

**Ameriflux dataset of flux towers**

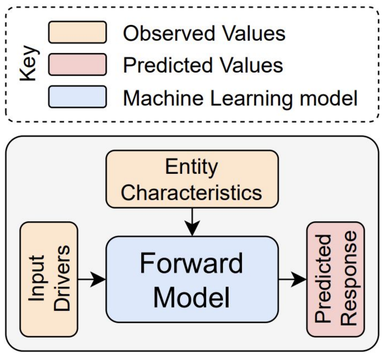


Nathaniel, Juan, et. al *Scientific Data* (2023)

**Entity**: Flux Tower
**Driver**: Weather and LAI
**Response**: GPP (Energy captured through photosynthesis)

## Method: CT-LSTM

- GPP and $R_{eco}$ are predicted using input drivers and entity characteristics by using the model.
- We trained a deep learning-based CT-LSTM on well-observed flux tower sites to predict GPP and $R_{eco}$.
- We trained a multi-loss-based CT-LSTM model on flux tower sites from the Fluxnet dataset of point-based flux measurements with weather drivers to predict GPP and $R_{eco}$ in unobserved sites from Fluxnet and Ameriflux datasets.
- 5 independent models were trained to account for deep learning based uncertainty due to model initialization. Then, an ensembled prediction of the 5 models were made as the final prediction.
- The trained model on Fluxnet sites was used to make predictions on test sites in Fluxnet and Ameriflux datasets.




**Entity Characteristics**: Latitude and Longitude, etc.
**Driver**: Weather and LAI
**Response**: GPP and $R_{eco}$.

## Results: FLUXNET and AMERIFLUX

- CT-LSTM multi-loss outperforms the primary baseline, Metaflux, by a substantial margin for both GPP and $R_{eco}$ flux predictions.
- CT-LSTM multi-loss performs 10% better in predicting GPP and 4% better in $R_{eco}$ flux predictions in Fluxnet testing.
- Better improvements were seen in Ameriflux testing by 32% for GPP and 35% for $R_{eco}$ predictions.
- Incorporating domain knowledge in CT-LSTM Physics didn't perform better than Metaflux.

| MODEL | Quantity predicted | |
|---|---|---|
| FLUXNET TESTING | GPP | $R_{eco}$ |
| CT-LSTM Multi-Loss | 3.15 | 2.75 |
| CT-LSTM Physics | 3.39 | 3.48 |
| METAFLUX | 3.50 | 2.89 |

Mean Ensemble Root Mean Square Error (RMSE).

| MODEL | Quantity predicted | |
|---|---|---|
| AMERIFLUX TESTING | GPP | $R_{eco}$ |
| CT-LSTM Multi-Loss | 2.28 | 1.58 |
| CT-LSTM Physics | 2.32 | 1.61 |
| METAFLUX | 3.39 | 2.46 |

Mean Ensemble Root Mean Square Error (RMSE).

## Conclusion and Next Steps

- CT-LSTM is a model to be flexibly transferred to new tasks with minimal data and it can be used for spatial downscaling. The method proposed is general and can be used in other applications (e.g. agriculture).
- In the next steps, driver data for every point on earth downloaded will be used along with the model to downscale for the world.
- Since incorporating domain knowledge in CT-LSTM Physics didn't perform much better than CT-LSTM multi-loss and Metaflux, in the future, better ways of incorporating knowledge will be performed.

## Acknowledgement

- MSI's agate cluster was used to perform pre-processing of the datasets, the creation of deep learning models, and to download the driver data for the whole world.
- Effective use of MSI computing nodes and GPUs resulted in training and testing of all the CT-LSTM models on the time-series based datasets. MSI resources such as high -performing GPUs resulted in the scientific investigation of the results produced by the CT-LSTM models.
- This work was funded by the NSF LEAP Science and Technology Center award 201962 and NSF Advancing Deep Learning for Inverse Modeling grant 2313174. Arvind Renganathan's work in this research is supported by the UMII MNDrive Graduate Fellowship.