# Feature Selection for High Dimensional Time Series Forecasting with Artificial Neural Networks

Paul Tarpey (Cornell University), Yuki Hamada (Argonne National Laboratory)

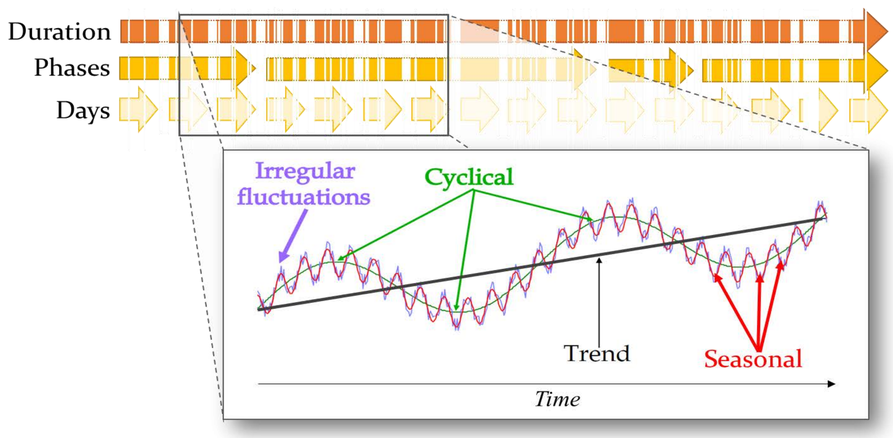
## Motivations

1. Many real world time-series data sets are BIG, COMPLEX, and MESSY!
   How can we analyze data having systematic gaps as well as randomly missing observations in hierarchically nested temporal patterns in conjunction with traditional time-series analysis?
2. The 'black box' nature of artificial neural network (ANN) models make it difficult for us to understand the mechanism or phenomena under investigation.
   How can we effectively open the 'black box' to better interpret model results?

## Goal

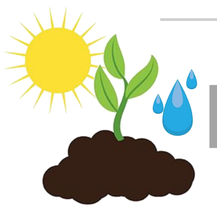
Demonstrate the use of conditional inference (CI) trees as a knowledge-assisted feature selection method for high-dimensional time series forecasting using ANNs with a focus on optimizing model interpretability and prediction accuracy.

## Big, Complex, and Messy Time Series Data



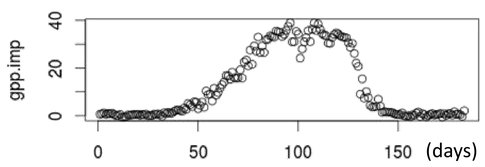

## What and Why? EcoSpec Project

At Argonne National Laboratory, Dr. Yuki Hamada investigates how land surface responds and contributes to climate change using hyperspectral remote sensing and field observations at a high temporal frequency and local scale. The EcoSpec Project attempts to increase our understanding of fine scale phenomena in order to fill the knowledge gap in regional/global climate modeling and improve future climate forecasting. The 'big, complex, and messy' time-series data collected for the project contains varying temporal, spatial, and spectral continuity, contiguity, and intermittence that necessitated the development of new analytical methods.

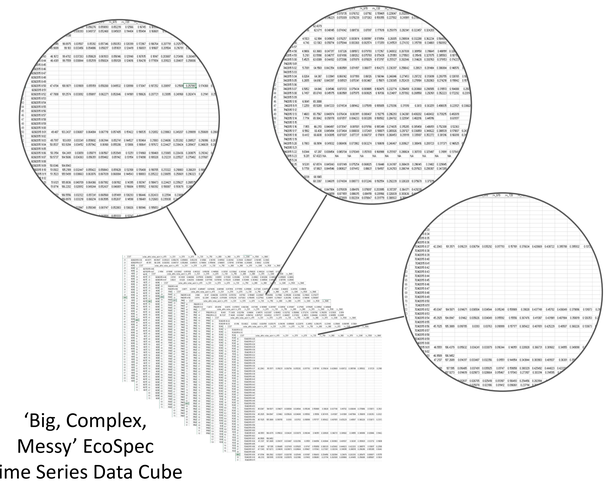


**Data Prep** → **Feature Selection** → **Artificial Neural Networks**

Change point analysis to define phases.



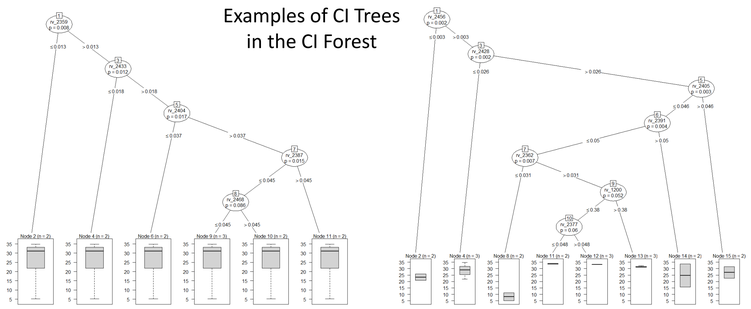

Data sub-setting to obtain temporally continuous input for CI trees.



'Big, Complex, Messy' EcoSpec Time Series Data Cube

**Step 1** Build a forest of CI trees using temporally continuous data.

Examples of CI Trees in the CI Forest



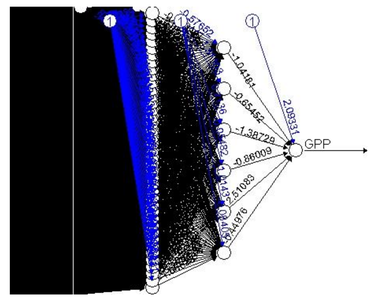
**Step 2** For each tree:

- Determine the correlation for each predictor variable with all other predictor variables.
- If this correlation exceeds some threshold value, then permutate the values of this variable in a separate temporally continuous test set conditioned upon the partition of the feature space defined by the tree using all cutpoints as bisectors of the sample space.

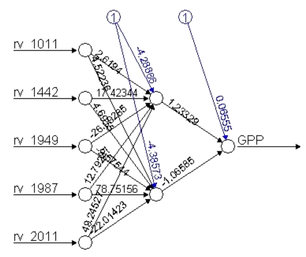
**Step 3** If the variable of interest has missing values in the test set, then randomly assign all values to the left and right child nodes of the primary split of the variable according to their corresponding relative frequencies in the original split.

**Step 4** Determine the difference between the prediction accuracy of the tree for out-of-bag (OOB) observations before and after the permutation and assign each variable an importance measure averaged over all trees to be used in feature selection for the ANN.

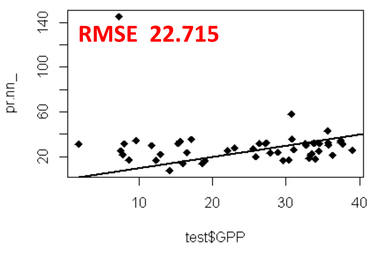
### Black Box



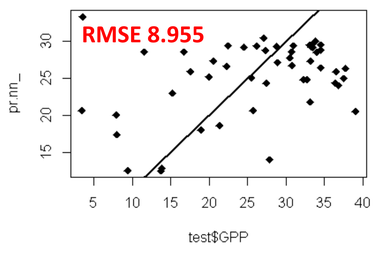

### Feature Selection




### Actual vs Predicted BB




### Actual vs Predicted FS




## What's Next?

1. Tune 'Feature Selection' ANNs to improve prediction accuracy of gross primary production of ecosystems using hyperspectral data from the EcoSpec project while providing insight into the components of the model.
2. Author a R package for the automation of conditional variable importance selection with missing data values.