Sandia National Laboratories

Pathway Detection using Random Forest Regressor Feature Importances

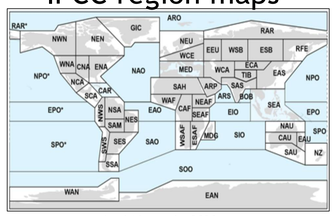
Matt Peterson, Meredith Brown, Irina Tezaur, Diana Bull

Motivation

The CLDERA Grand Challenge Project is identify causal pathways between various climate features and spatial components.

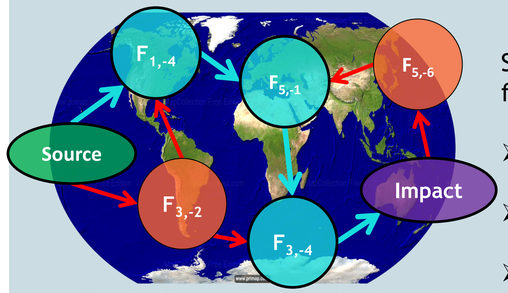
Goal: Generate a weighted directed graph where the edges indicate influence from one feature of interest to another. Chaining these edges together defines a pathway which we call a **Source to Impact pathway**.

IPCC region maps



Iturbide et al., 2020.
https://essd.copernicus.org/articles/12/2959/2020/

*These are **climate regions** around the globe that we want to be able to **identify how one region influences another region** and at what time scale.*



*Example of the type of graphs we will produce and how a potential **Source to Impact pathway** may look like. The most direct path is not always best path.*

Source to Impact pathways can help analysts identify which features of interest have the largest influence on their results

- *Which sources are the **largest contributors** to Arctic Sea Ice extent?*
- *What are the potential **downstream effects** of a drought in Western Europe?*
- *Find the **path of least resistance** from a source to impact?*
- *What are the **bottleneck features/nodes** between source and impact?*

Technical Approach

Random Forest Regressors (RFRs): Machine Learning (ML) predictive models

- Require training data comprised of pairs of inputs and outputs
- Once trained, can predict outputs given a set of input variables

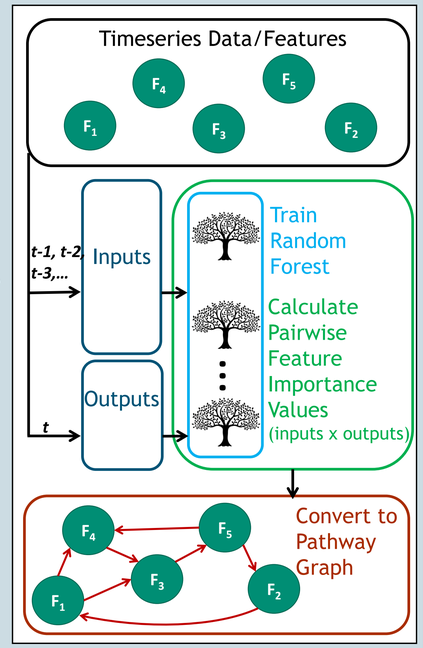
Feature Importances: a scalar value that indicates the predictive power that a particular input variable has on a given output variable

RFR is commonly used for and well-developed for regression, classification and prediction tasks.

- Our approach **extends RFR for purpose of discovering pathways**

Approach:

- Train a multi-variate RFR (*python package: sklearn*)
- Determine feature importances weights between inputs and outputs
- Convert weights into a weighted directed graph

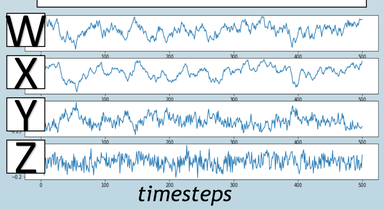


Results Summary

Example: Synthetic Dataset

- Set of coupled equations
- Dependencies are 1 timestep, the algorithm will look at 1 and 2 timesteps

$$W_t = 0.9W_{t-1} + \varepsilon_{W_t}$$
$$X_t = 0.8X_{t-1} + 0.5W_{t-1} + \varepsilon_{X_t}$$
$$Y_t = -0.9W_{t-1} + \varepsilon_{Y_t}$$
$$Z_t = 0.3X_{t-1} + 0.5Y_{t-1} + \varepsilon_{Z_t}$$



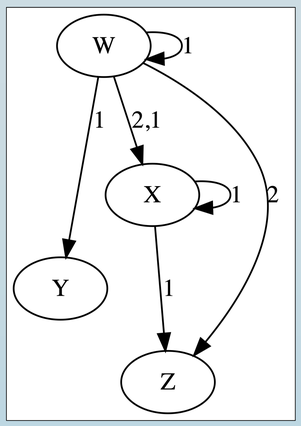
id	source	Target	Weight
17	W_-1	Y	0.219
10	W_-2	X	0.176
11	X_-1	X	0.176
9	W_-1	X	0.168
26	W_-2	Z	0.149
27	X_-1	Z	0.149
1	W_-1	W	0.211

Feature Importances

- Importance Weights calculated using *SHAP feature importance*
- Weights are pruned using a random variable as a cutoff point

RFR Feature Pathway Graph

- Edges indicate influence from one feature to another
- The *numbers are time lags* associated with the path
- *Mostly correct*, but contains two extra edges and one missing edge
- These extra pathways are *'Not wrong, but not correct'* $(W_{t-2}, Z_t) \approx (W_{t-2}, X_{t-1}) \rightarrow (X_{t-1}, Z_t)$



Initial development of an AI-based surrogate method for mining causal relationships from climate data that can be used to validate and improve upon computational models

Impacts & Successes to date

Contributions to CLDERA

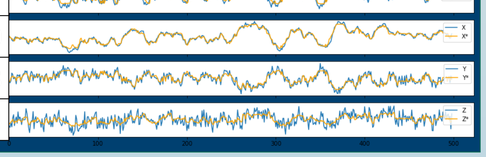
- Developing a **suite of verification metrics** for feature pathway graphs
- Used as **tool for explaining** model outputs from other thrusts
- Identify key **pathway differences** between simulation runs

Presentations / Papers

- Poster at American Geophysical Union (AGU) 2022
- In Progress of writing a journal paper

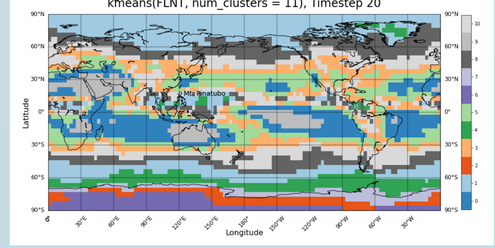
Variable	R^2
W	0.812
X	0.940
Y	0.808
Z	0.484

Verification metrics



Example of a verification metric to determine if our directed graphs are trustworthy in the absence of ground truth. In this case we are checking to see if the RFR can predict our data well

Signature-based Clusters



Example of clustering analysis from another subthrust in CLDERA. We will be using these types of clusters as inputs to identify how these moving clustering influence one another

0.5 FTE for Subthrust in CLDERA Grand Challenge

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology &Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. **SAND2023-00782C**

Sandia National Laboratories