Radiant MLHub
EARTH IMAGERY FOR IMPACT

Dataset Documentation

Dataset Name: CSU Synthetic Attribution Benchmark Dataset

Description

This is a synthetic dataset that can be used by members of the geospatial community who are interested in benchmarking methods of explainable artificial intelligence (XAI) for geoscientific applications. The dataset is specifically inspired from a climate forecasting setting (using seasonal timescales) where the task is to predict regional climate variability given global climate information lagged in time. The dataset consists of a synthetic input **X** (10^6 series of 2D arrays of random fields drawn from a multivariate normal distribution) and a synthetic output Y (scalar series) generated by using a nonlinear function F: R^d -> R.

The synthetic input aims to represent temporally-independent realizations of anomalous global fields of sea surface temperature, the synthetic output series represents some type of regional climate variability that is of interest (temperature, precipitation totals, etc.) and the function F is a simplification of the climate system.

Since the nonlinear function F that is used to generate the output given the input is known, we also derive and provide the attribution of each output value to the corresponding input features. Using this synthetic dataset users can train any AI model to predict Y given **X** and then implement XAI methods to interpret it. Based on the “ground truth” of attribution of F the user can assess the faithfulness of any XAI method.

NOTE: the spatial configuration of the observations in the NetCDF database file conform to the planetocentric coordinate system (89.5N - 89.5S, 0.5E - 359.5E), where longitude is measured in the positive heading east from the prime meridian.

License

CC-BY-4.0

Creator(s)

Colorado State University (CSU) , Cooperative Institute for Research in the Atmosphere (CIRA)

Contact

amamalak@colostate.edu

Publications

Mamalakis, A., Ebert-Uphoff, I., Barnes, E. A. (2022) Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset, *Environmental Data Science*, DOI: 10.1017/eds.2022.7

Data Properties

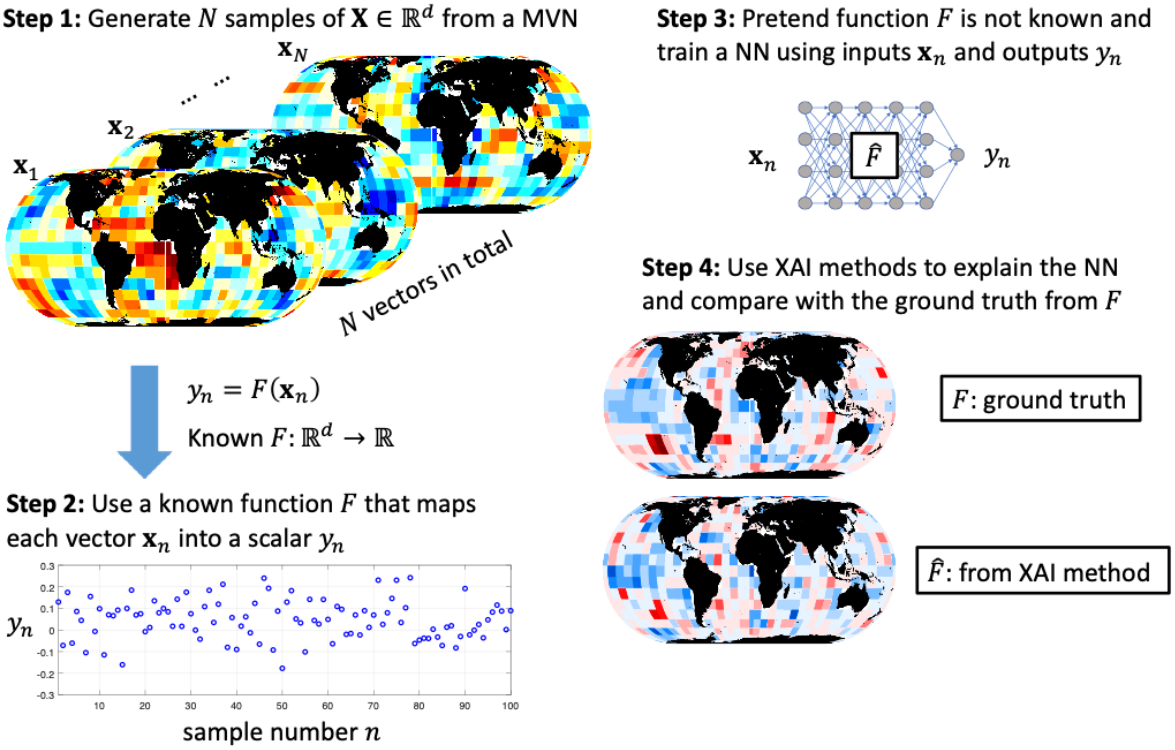
Property name	Property Description	Parameters/Allowed responses (optional)
SSTrand	A synthetic input **X** of N independent (in time) random maps that aim to represent monthly SST anomalies around the globe (i.e., honoring spatial dependence). Dimensions: 18 by 36 by 1000000.	
y	A synthetic output y = F(**X**), with F being the summation of local piecewise linear responses to **X**. See more details in our paper. The synthetic output is labeled "y" and is a time series of scalar values with dimensions: 1000000.	
Cnt	The ground truth of attribution for F (and for a zero baseline) labeled as "Cnt" (dim: 18 by 36 by 1000000). This is the true contribution of each of the grid points to each of the values of Y.	
W	The weights between the break points (in each linear segment) of the local piecewise linear functions that add up to F. The weights array is labeled as "W" and has dimensions of 18 by 36 by 6.	

Appendix F How was the dataset generated

We start by randomly generating independent realizations of an input vector $\mathbf{X} \in R^{d}$. Although arbitrary, the distributional choice of the input is decided with the aim of being a reasonable proxy of the independent variable of the physical problem of interest. Here, the input series represents monthly global fields of SST anomalies (deviations from the seasonal cycle) at a $10^{o}\times10^{o}$ resolution (fields of d = 458 variables; see step 1 in figure below). We generate the SST anomaly fields from a Multivariate Normal Distribution MVN($\mathbf{0},\mathbf{\Sigma}$), where $\mathbf{\Sigma}$ is the covariance matrix and represents the dependence between SST anomalies in different grid points (or pixels in image classification settings) around the globe (spatial dependence). The matrix $\mathbf{\Sigma}$ is set equal to the sample correlation matrix that is estimated from monthly SST observations.

We next create a nonlinear response $Y \in R$ to the synthetic input $\mathbf{X} \in R^d$ (see step 2 in figure below), using a real function F: R^d -> R. For any sample n, the response of our system y_n to the input $\mathbf{x}_n$ is given as $y_n=F(\mathbf{x}_n)$ or after dropping the index n for simplicity and relating the random variables instead of the samples, $Y=F(\mathbf{X})$. More details can be found in our paper: Mamalakis et al. (2022).

Appendix G How to use the dataset


Step 1: Generate N samples of $\mathbf{X} \in \mathbb{R}^d$ from a MVN
$\mathbf{x}_N$
...
$\mathbf{x}_2$
$\mathbf{x}_1$
N vectors in total
$y_n = F(\mathbf{x}_n)$
Known $F: \mathbb{R}^d \rightarrow \mathbb{R}$
Step 2: Use a known function F that maps each vector $\mathbf{x}_n$ into a scalar y_n
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
y_n
10
20
30
40
50
60
70
80
90
100
sample number n
Step 3: Pretend function F is not known and train a NN using inputs $\mathbf{x}_n$ and outputs y_n
$\mathbf{x}_n$
$\hat{F}$
y_n
Step 4: Use XAI methods to explain the NN and compare with the ground truth from F
F: ground truth
$\hat{F}$: from XAI method

Schematic overview of the general idea of the dataset. In step 1, we generate $N=10^6$ independent realizations of a random vector $\mathbf{X}$ from a multivariate Normal Distribution. In step 2, we generate a response Y to the synthetic input $\mathbf{X}$, using a nonlinear function F. In step 3, users can train an AI model to learn to predict Y given $\mathbf{X}$. Lastly, in step 4, users can compare the XAI results estimated from different XAI methods to the ground truth of F, in order to assess XAI fidelity (from Mamalakis et al., 2022).