The Earth Data Analytic Services (EDAS) Framework

Thomas Maxwell[1], Daniel Duffy[2], Laura Carriere[3], and Gerald Potter[1]

[1]NASA Goddard Space Flight Center
[2]NASA Center for Climate Simulation
[3]NCCS, NASA Goddard

November 26, 2022

Abstract

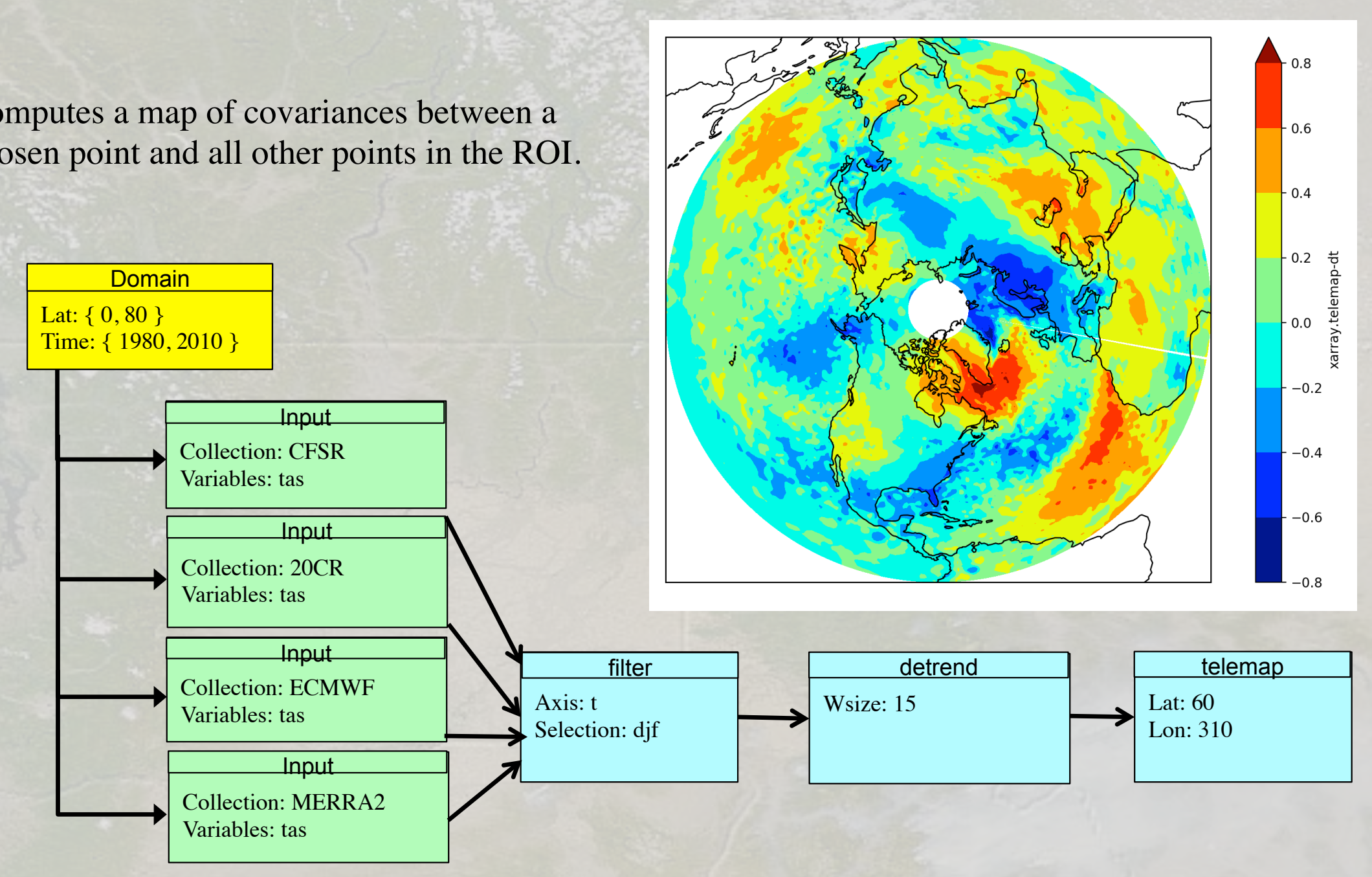
Faced with unprecedented growth in earth data volume and demand, NASA has developed the Earth Data Analytic Services (EDAS) framework, a high performance big data analytics and machine learning framework. This framework enables scientists to execute data processing workflows combining common analysis and forecast operations close to the massive data stores at NASA. The data is accessed in standard (NetCDF, HDF, etc.) formats in a POSIX file system and processed using vetted tools of earth data science, e.g. ESMF, CDAT, NCO, Keras, Tensorflow, etc. EDAS utilizes high performance parallel data access, a custom distributed array framework, and a streaming parallel in-memory workflow for efficiently processing huge datasets within limited memory spaces with interactive response times. EDAS services are accessed via a WPS API being developed in collaboration with the ESGF Compute Working Team to support server-side analytics for ESGF. The API can be accessed using direct web service calls, a Python script, a Unix-like shell client, or a JavaScript-based web application. New analytic operations can be developed in Python, Java, or Scala (with support for other languages planned). Client packages in Python, Java/Scala, or JavaScript contain everything needed to build and submit EDAS requests. The EDAS architecture brings together the tools, data storage, and high-performance computing required for timely analysis of large-scale data sets, where the data resides, to ultimately produce societal benefits. It is currently deployed at NASA in support of the Collaborative REAnalysis Technical Environment (CREATE) project, which centralizes numerous global reanalysis datasets onto a single advanced data analytics platform. This service enables decision makers to compare multiple reanalysis datasets and investigate trends, variability, and anomalies in earth system dynamics around the globe. EDAS services include configurable high performance neural network learning modules designed to operate on the products of EDAS workflows. As a science technology driver we have explored the capabilities of these services for long-range forecasting of the interannual variation of important regional scale seasonal cycles. Neural networks were trained to forecast All-India Summer Monsoon Rainfall (AISMR) one year in advance using (as input) the top 8-64 principal components of the global surface temperature and 200 hPa geopotential height fields from NASA's MERRA2 and NOAA's Twentieth Century Reanalyses. The promising results from these investigations illustrate the power of easily accessible machine learning services coupled to huge repositories of earth science data.

The Earth Data Analytic Services (EDAS) Framework

Thomas Maxwell, Dan Duffy, Laura Carriere, Jerry Potter
NASA GODDARD SPACE FLIGHT CENTER, GREENBELT, MD

Teleconnection Maps

Computes a map of covariances between a chosen point and all other points in the ROI.



Domain: Lat: { 0, 80 } Time: { 1980, 2010 }

Input: Collection: CFSR Variables: tas

Input: Collection: 20CR Variables: tas

Input: Collection: ECMWF Variables: tas

Input: Collection: MERRA2 Variables: tas

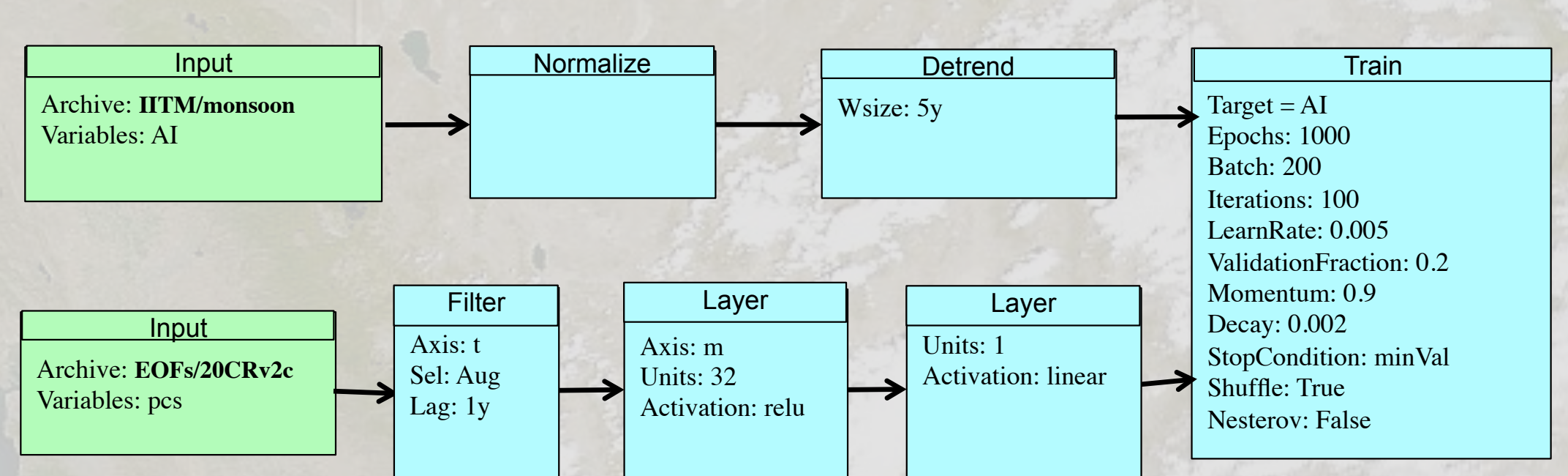
filter: Axis: t Selection: djf

detrend: Wsize: 15

telemap: Lat: 60 Lon: 310

Machine Learning Workflow

- Predict All-India Monsoon rainfall accumulation one year in advance
- Use a two-layer neural network
- Inputs: First 32 PCs of global surface temperature, 1 year lag time



Input: Archive: **IITM/monsoon** Variables: AI

Normalize

Detrend: Wsize: 5y

Train: Target = AI, Epochs: 1000, Batch: 200, Iterations: 100, LearnRate: 0.005, ValidationFraction: 0.2, Momentum: 0.9, Decay: 0.002, StopCondition: minVal, Shuffle: True, Nesterov: False

Input: Archive: **EOFs/20CRv2c** Variables: pcs

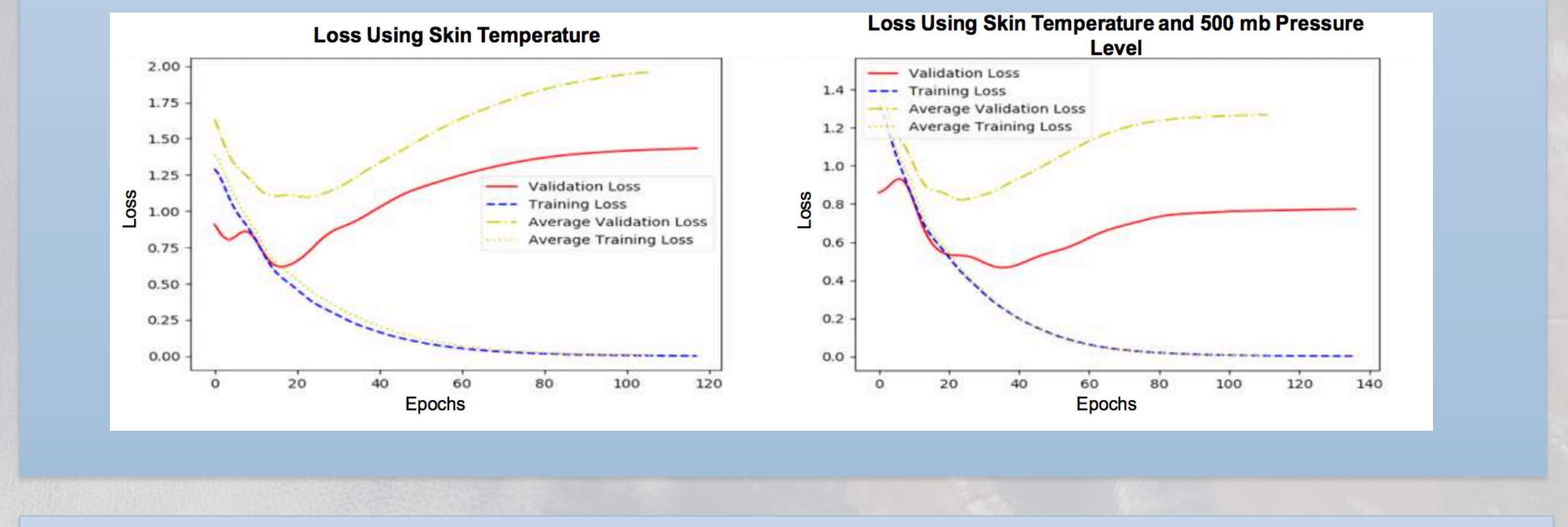
Filter: Axis: t Sel: Aug Lag: 1y

Layer: Axis: m Units: 32 Activation: relu

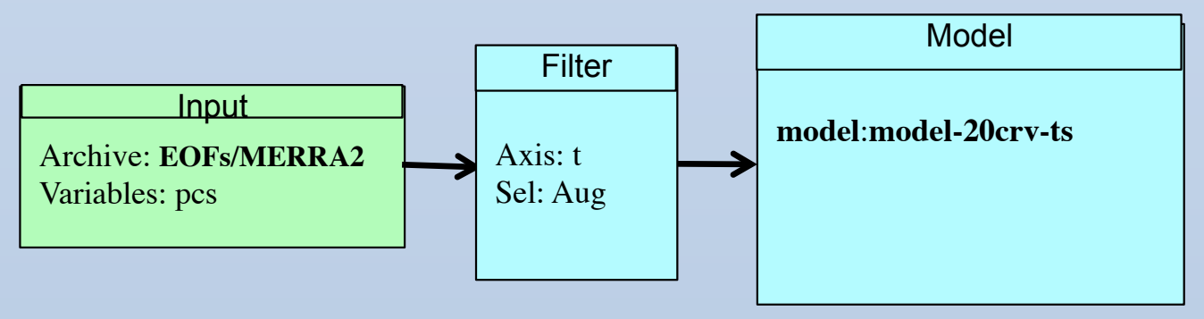
Layer: Units: 1 Activation: linear

Training Performance

- Loss Function: Mean square error
  - Output node results vs. IITM-AI timeseries
- Last 20% of data reserved for validation
- Choose model with minimum error on validation data



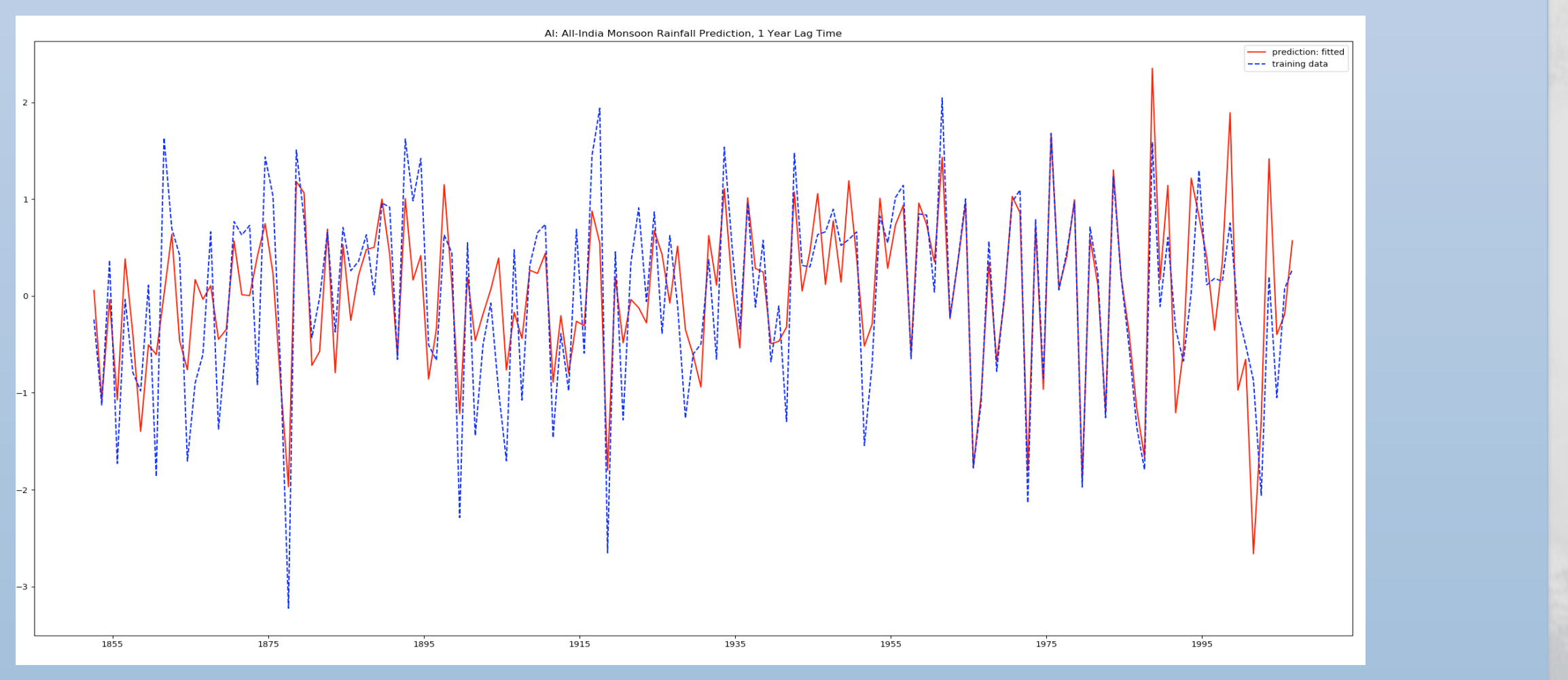
Prediction



Input: Archive: **EOFs/MERRA2** Variables: pcs

Filter: Axis: t Sel: Aug

Model: **model**:**model-20crv-ts**



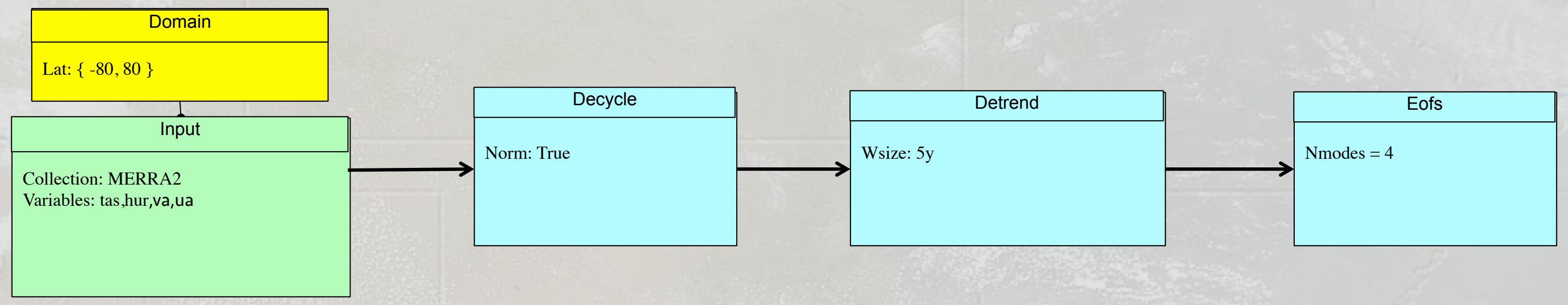
- Comparison of predicted (red) to actual (blue) monsoon precipitation

ABSTRACT

Faced with unprecedented growth in earth data volume and demand, NASA has developed the Earth Data Analytic Services (EDAS) framework, a high performance big data analytics and machine learning framework. This framework enables scientists to execute data processing workflows combining common analysis and forecast operations close to the massive data stores at NASA. The data is accessed in standard (NetCDF, HDF, etc.) formats in a POSIX file system and processed using vetted tools of earth data science, e.g. ESMF, CDAT, NCO, Keras, Tensorflow, etc. EDAS facilitates the construction of high performance parallel workflows by combining canonical analytic operations to enable processing of huge datasets within limited memory spaces with interactive response times. EDAS services are accessed via a WPS API being developed in collaboration with the ESGF Compute Working Team to support server-side analytics for ESGF. Client packages in Python, Java/Scala, or JavaScript contain everything needed to build and submit EDAS requests.

EDAS services include configurable high performance neural network learning modules designed to operate on the products of EDAS workflows. As a science technology driver we have explored the capabilities of these services for long-range forecasting of the interannual variation of important regional scale seasonal cycles. Neural networks were trained to forecast All-India Summer Monsoon Rainfall (AISMR) one year in advance using (as input) the top 8-64 principal components of the global surface temperature and 200 hPa geopotential height fields from NASA's MERRA2 and NOAA's Twentieth Century Reanalyses. The promising results from these investigations illustrate the power of easily accessible machine learning services coupled to huge repositories of earth science data. The EDAS architecture brings together the tools, data storage, and high-performance computing required for timely analysis of large-scale data sets, where the data resides, to ultimately produce societal benefits.

EOFs Workflow



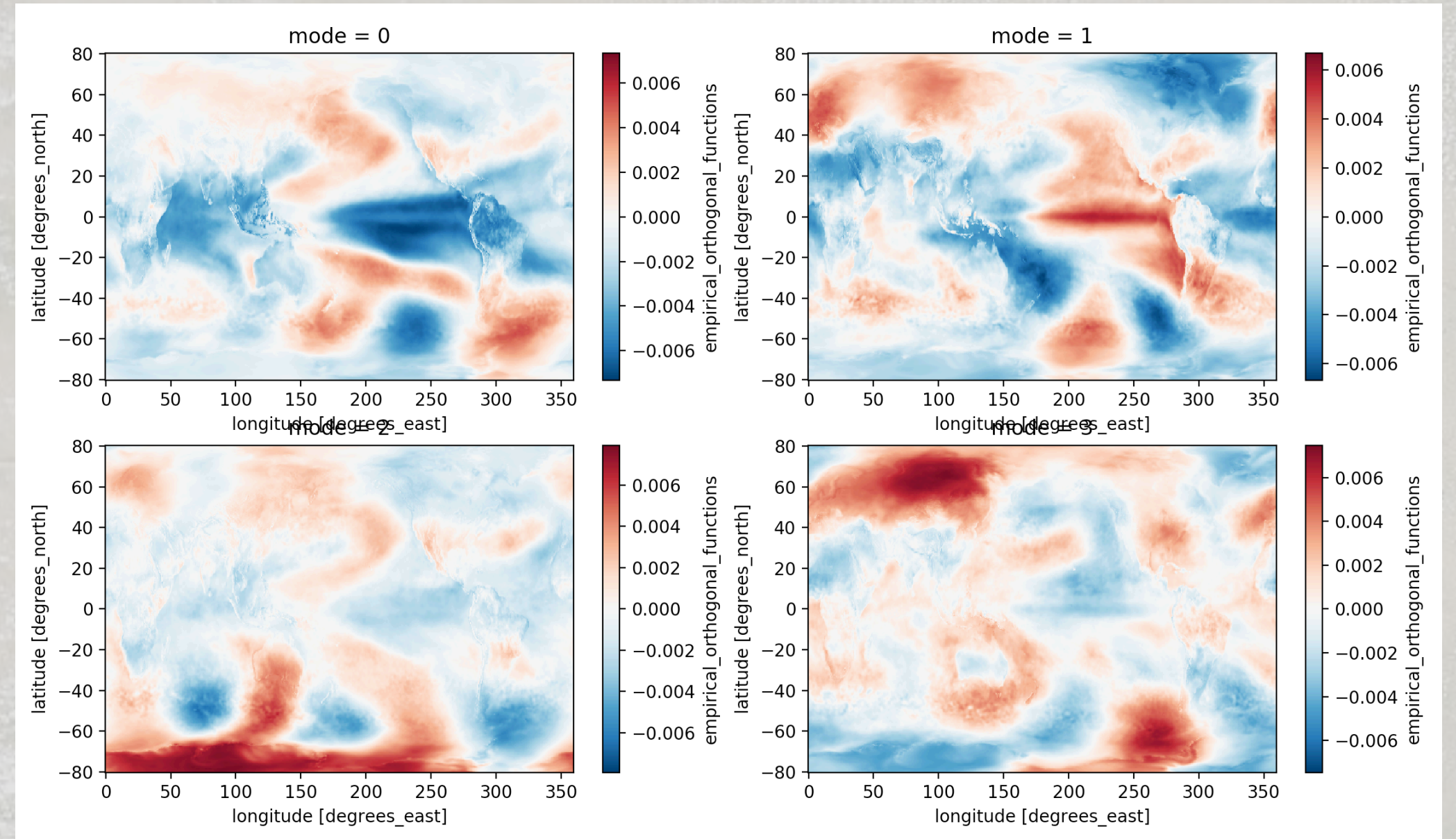
Domain: Lat: { -80, 80 }

Input: Collection: MERRA2 Variables: tas,hur,va,ua

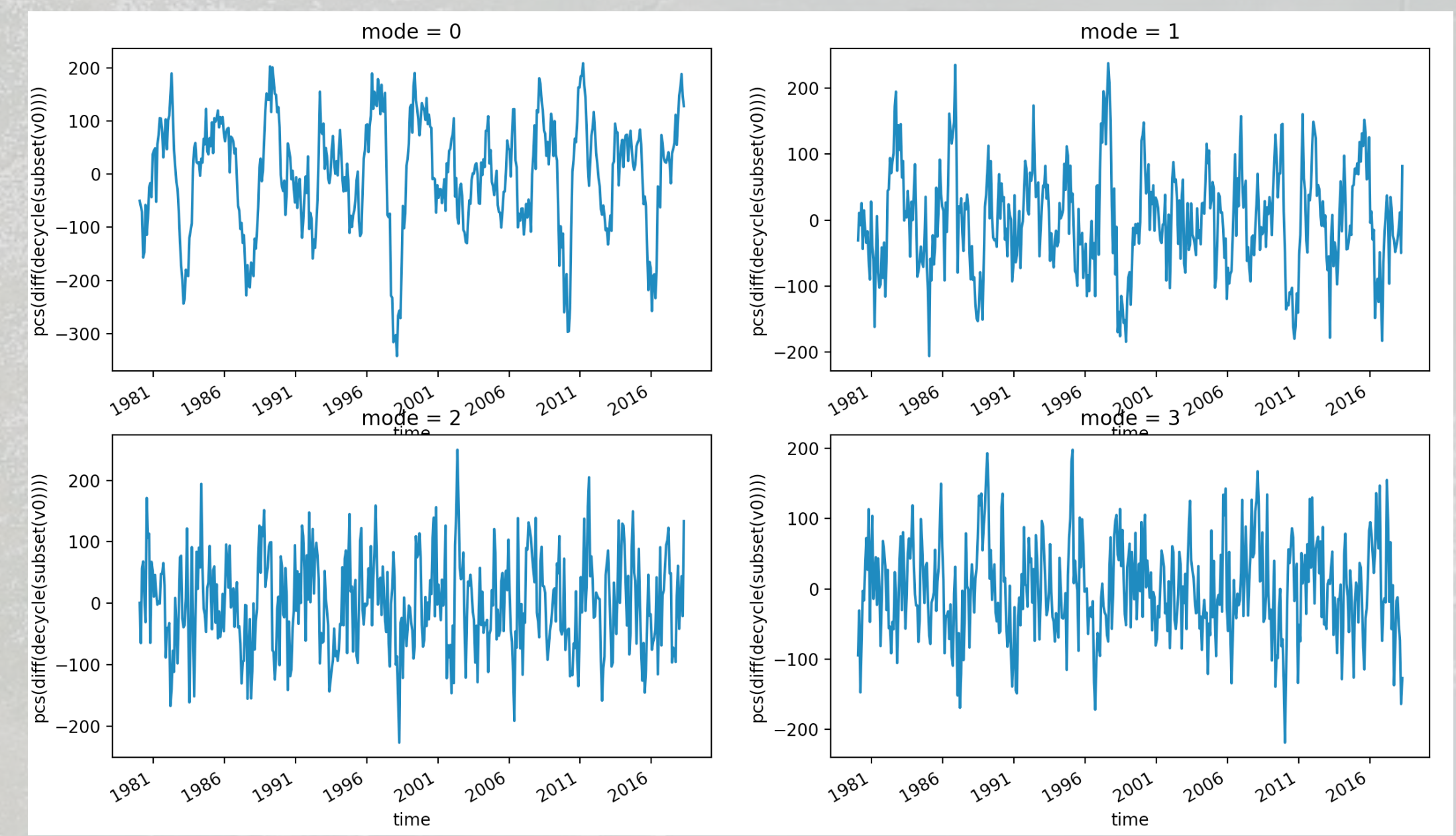
Decycle: Norm: True

Detrend: Wsize: 5y

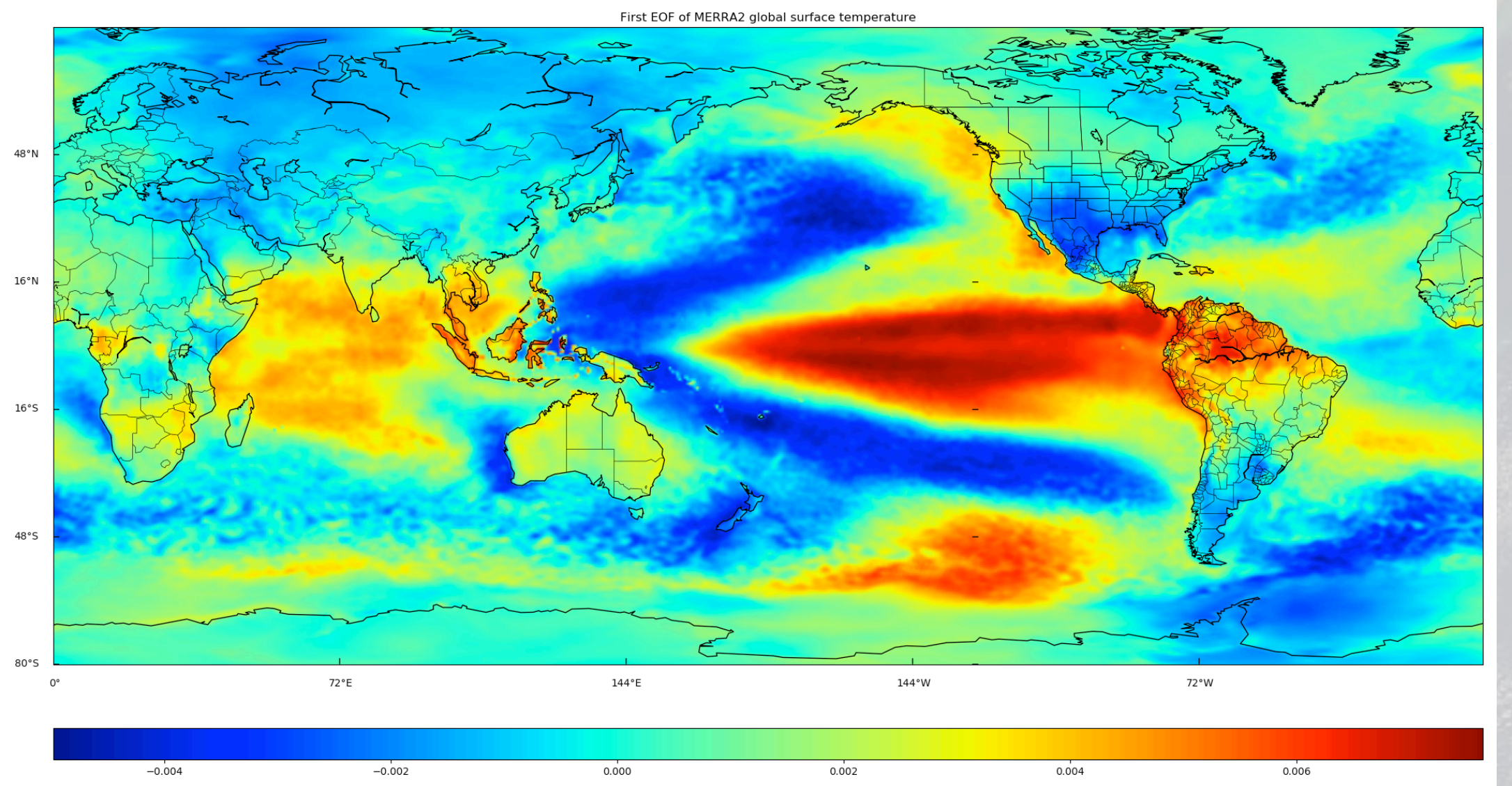
Eofs: Nmodes = 4



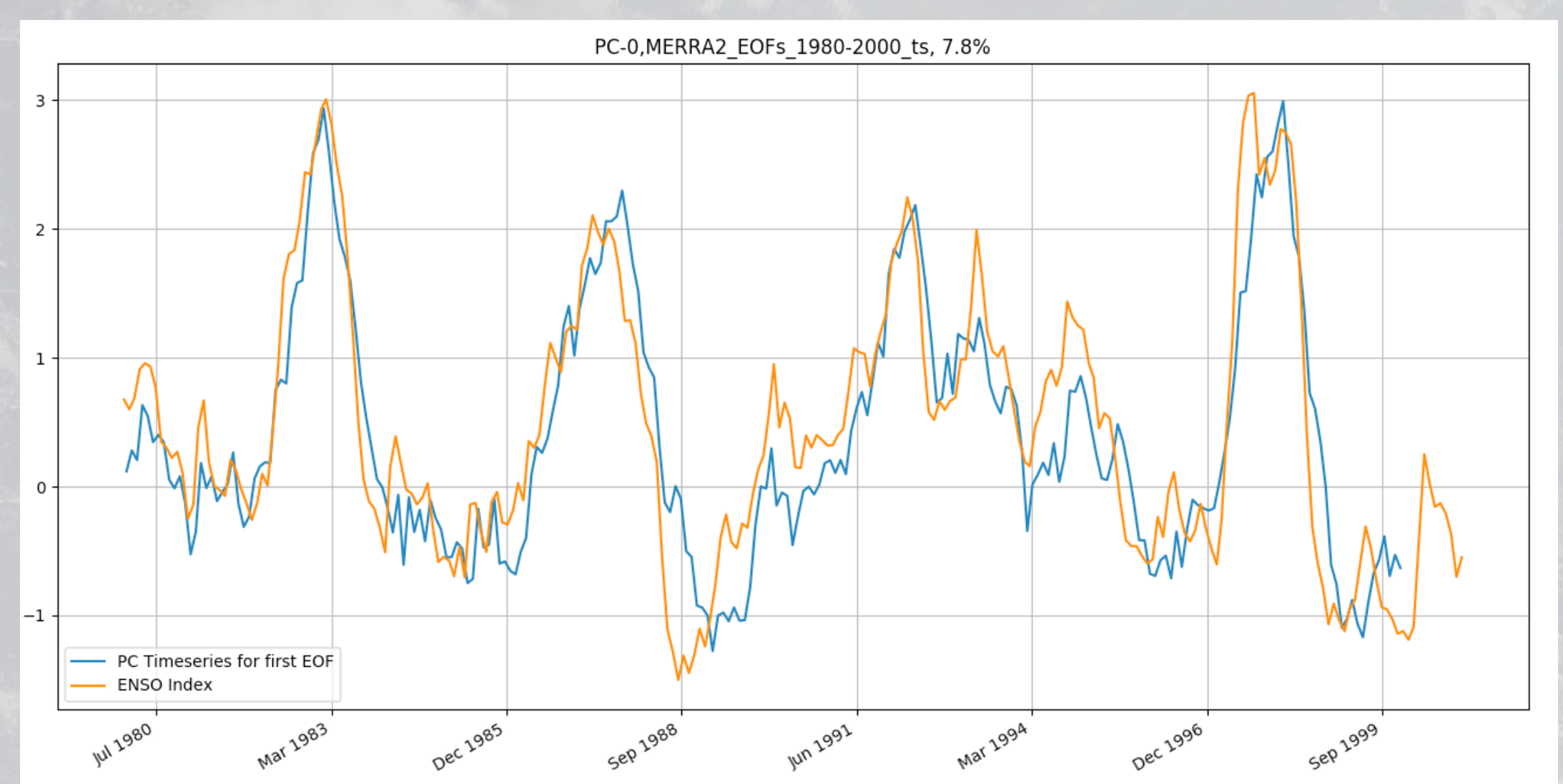
First four EOFs of MERR2 surface temperature



First 4 PCs of MERRA2 surface temperature

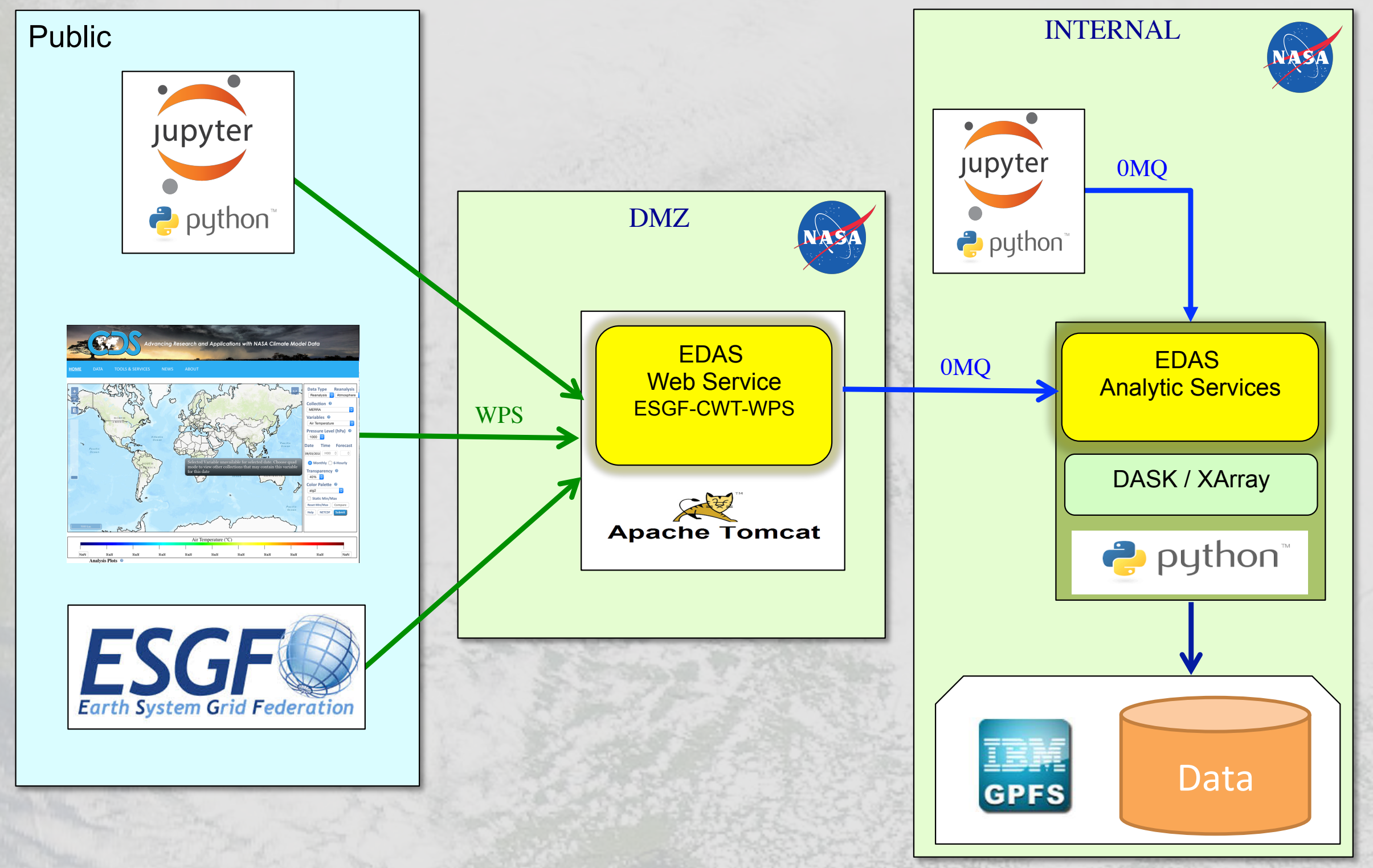


First EOF of surface temperature, capturing ENSO



Comparison of first PC with ENSO Index

EDAS Infrastructure



EDAS Canonical Operations

- Data access & subset
- Weighted Average
- Maximum
- Minimum
- Sum
- Difference
- Product
- Standard Deviation
- Variance
- Anomaly
- Median
- Norm
- Filter
- Decycle
- Highpass/Detrend
- Lowpass/Smooth

Specialized operations:
- EOF
- PC
- TeleconnectionMap

Neural Network Operations:
- Layer
- Trainer
- Model

Canonical Operation Options

- Domain: subset to region of interest
- Axes: reduce over axes
  - X (latitude), Y (longitude) , Z (levels), T (time), E (ensemble)
- Groupby: split-apply-combine
  - Custom or existing Axis, supports Pandas groups
- Resample: upsampling and downsampling
  - Pandas resample API

Example (for 10 years of data):

Operation	Interpretation	Size
ave(axis: t)	Time average	1
ave(axis: te)	Time ensemble average	1
ave(axis: t, groupby: t.month)	Monthly climatology	12
ave(axis: t, resample: t.month)	Monthly means	120

Why is this approach distinctive?

- **Server-side analytics**
  - *Performs the analytics close to the data.*
- **Direct access to reanalysis data archives via disk or OpenDAP:**
  - *Alleviates the need to download data to a local computer.*
- **Deploys existing (Python) climate data analysis tools:**
  - *Utilizes UVCDAT, ESMF, and other python analytic toolkits.*
- **High performance analytics:**
  - *Optimizes decomposition over processors for the current task.*
- **Modular structure:**
  - *Build new workflows by composing canonical operations.*
- **ESGF CWT WPS API Compliant:**
  - *Designed to operate as a compute engine for the ESGF.*
- **Pre-packaged analytics:**
  - *Simple composition of complex analysis and AI/ML operations.*