Sheryar Choudhry
CLIM 680 Class Project

Stochastic Analysis of Observational and Reanalysis Precipitation Using a VARX Model

Contents

Introduction:

Goal :

I want to compare an observational dataset and a reanalysis product to see if originate from similar stochastic processes.

More specifically, I would like to fit each grid point to a vector autoregressive model with exogenous variables (VARX).

Then using the model fits at each grid point, I can use a likelihood ratio statistic to see if my two datasets are generated from similar statistical processes.

Hypothesis:

I believe that the VARX analysis will show that the reanalysis and observational dataset stem from different statistical processes.

Introduction:

Datasets:

1. ECMWF ERA5 monthly averaged data on single levels from 1940 to present

https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-monthly-means?tab=overview

- Variables: Mean Monthly Precipitation Rate [m/day]
- Spatial Resolution: 0.5° x 0.5° [721 x 1440]
- Spatial extent: Global
- Temporal frequency: Monthly
- Temporal extent: 1940 to Present

2. The Global Precipitation Climatology Project (GPCP) Monthly Analysis Product

https://psl.noaa.gov/data/gridded/data.gpcp.html

- Variables: Mean Monthly Precipitation Rate [mm/day]
- Spatial Resolution: 2.5 degree latitude x 2.5 degree longitude global grid [72 x 144]
- Spatial extent: Global
- Temporal frequency: Monthly
- Temporal extent: 1979 to Present

Citations:

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2023): ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.f17050d7Links to an external site. (Accessed on 10-09-2025)

Adler, R.F., G.J. Huffman, A. Chang, R. Ferraro, P. Xie, J. Janowiak, B. Rudolf, U. Schneider, S. Curtis, D. Bolvin, A. Gruber, J. Susskind, and P. Arkin, 2003: The Version 2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979-Present). J. Hydrometeor., 4,1147-1167. (Accessed on 10-09-2025)

Preparing the Data:

1) Made sure both datasets cover the same time range
2) Convert the units for each dataset to be total accumulated monthly precipitation in mm
3) ERA5 was much higher resolution, so I made the dataset coarser to match GPCP

Vector Autoregressive Model with Exogenous Variables (VARX):

$$Y_t = Constant + Y_{t-1} + Annual\ Cycles + External\ Forcing\ Trends + Noise\ (\epsilon)$$

Constant: Base state/ y-intercept

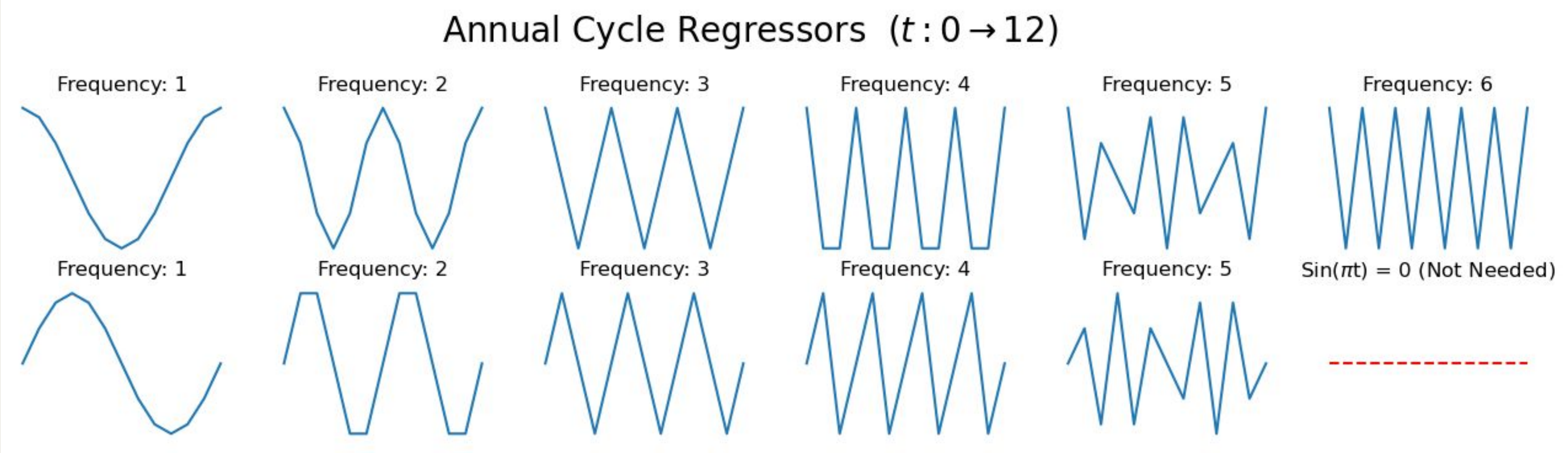
Y_t-1: The previous timestep of our time series.

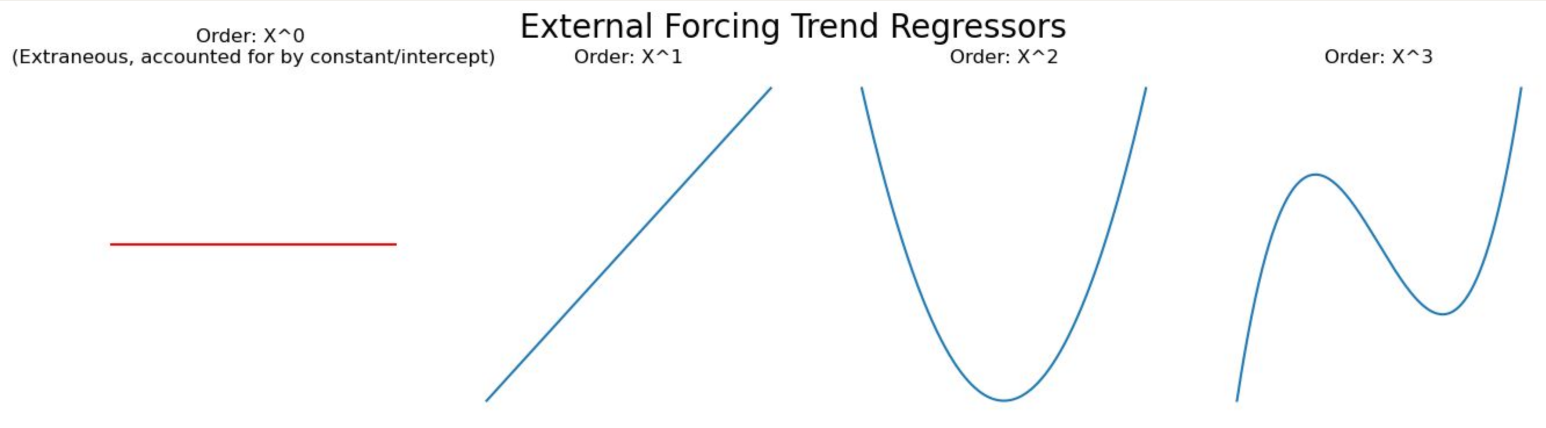
Annual Cycles: seasonal/annual behavior of our time series

External Forcing: Trends caused by greenhouse gases, volcanic activity, forest fires, etc.

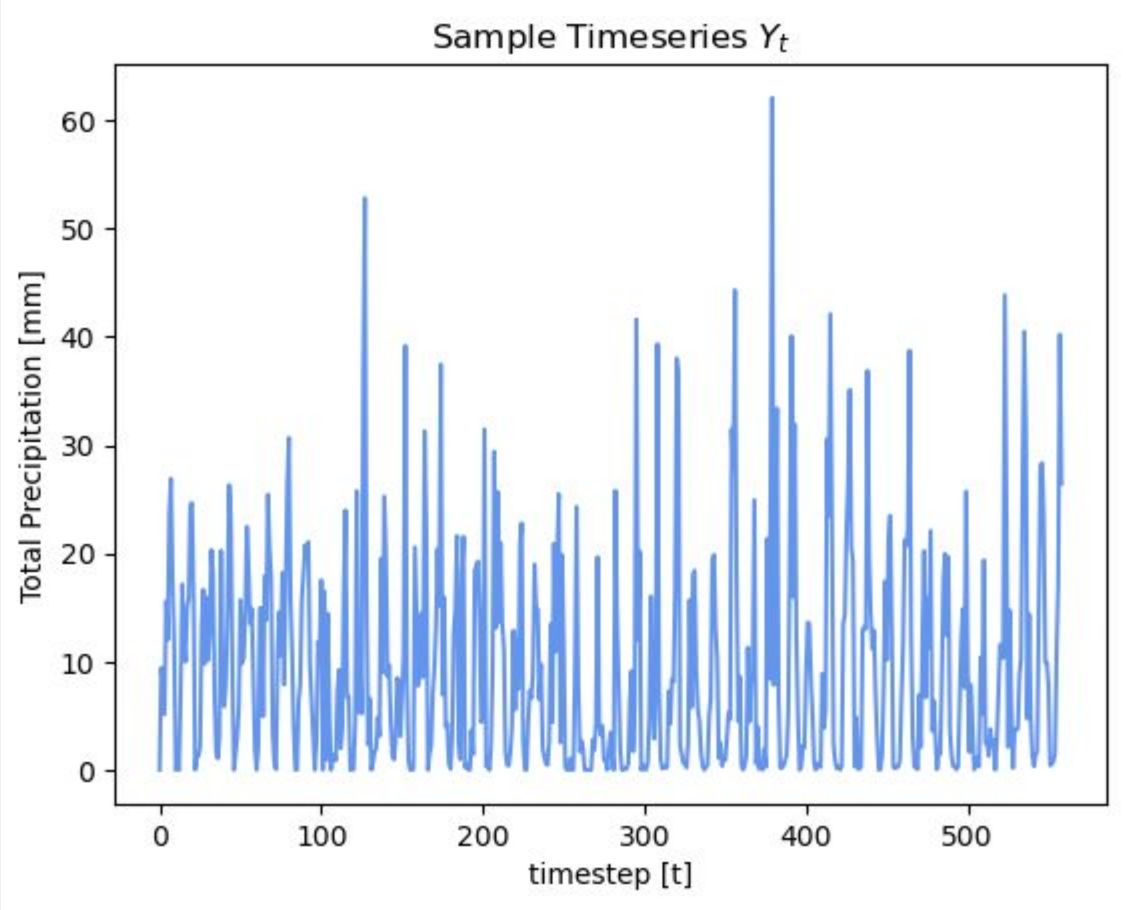
Noise: Contains the residuals/errors the model was not able to fit

Vector Autoregressive Model with Exogenous Variables (VARX):





Vector Autoregressive Model with Exogenous Variables (VARX):



Vector Autoregressive Model with Exogenous Variables (VARX):

$$Y_t = Constant + Y_{t-1} + Annual\ Cycles + External\ Forcing\ Trends + Noise\ (\epsilon)$$

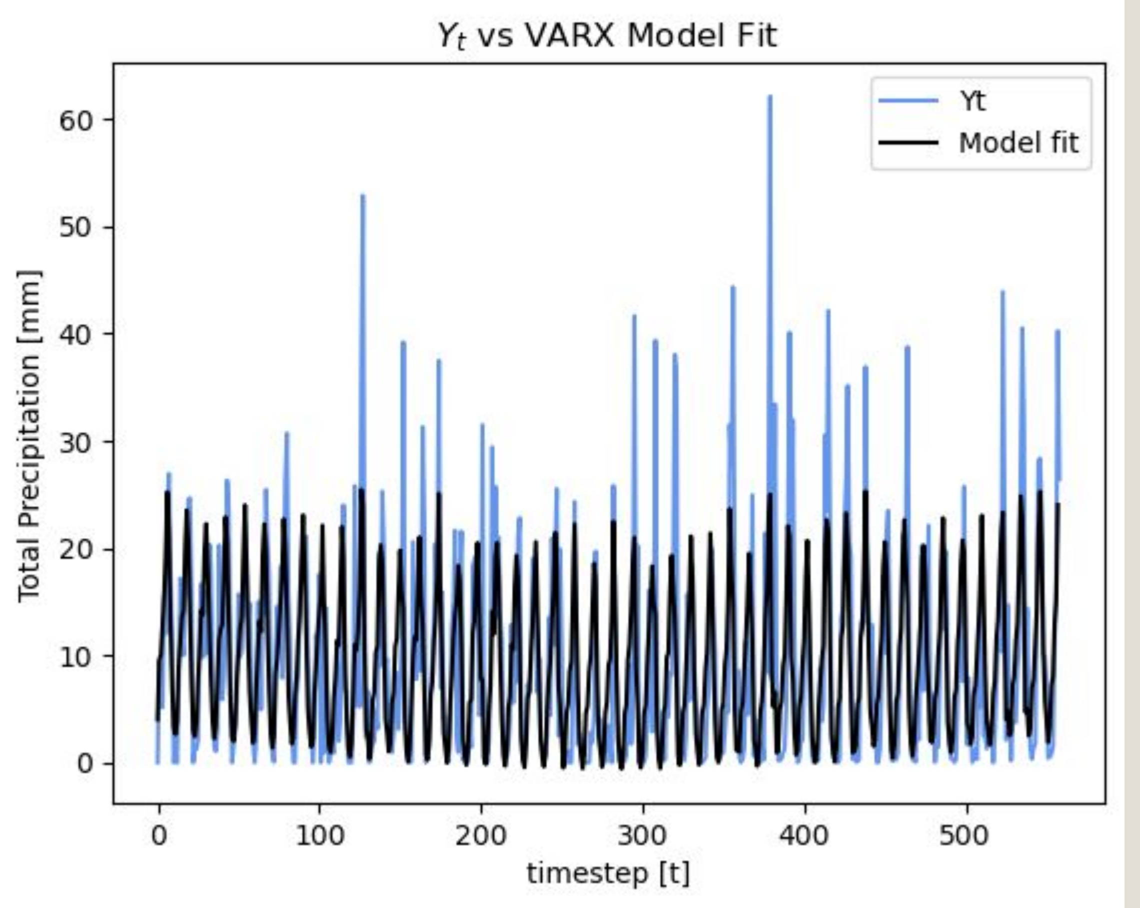
- 16 Total Regressors
- Create a Z matrix containing all the regressors that can be fed into the Ordinary Least Squares Solver
- Note, we skip the first value for Yt and omit the end values of our Z matrix

$$Y_t[1] \Rightarrow Y_t[T]$$

$$Z[0] \Rightarrow Z[T-1]$$

Vector Autoregressive Model with Exogenous Variables (VARX):

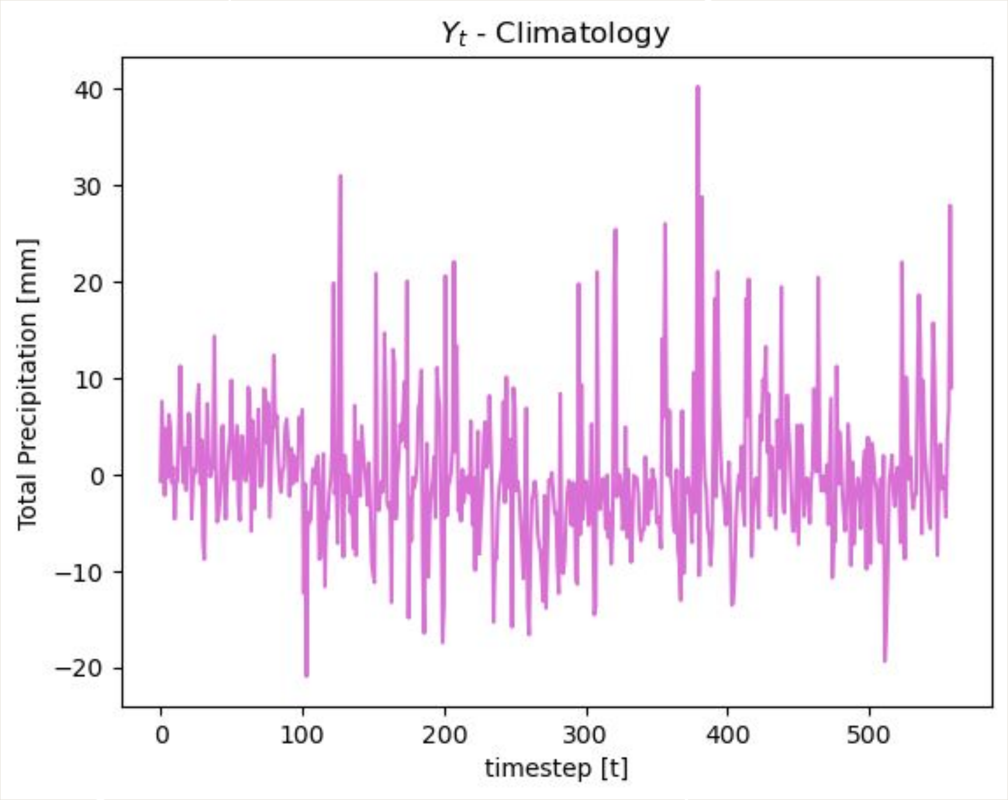
	coef	std err	t	P>|t|	[0.025	0.975]
const	7.7689	0.508	15.279	0.000	6.770	8.768
Yt-1	0.1817	0.042	4.298	0.000	0.099	0.265
Hc1	-7.1267	0.576	-12.374	0.000	-8.258	-5.995
Hc2	2.2357	0.443	5.043	0.000	1.365	3.107
Hc3	-1.0801	0.444	-2.435	0.015	-1.951	-0.209
Hc4	0.3632	0.444	0.819	0.413	-0.508	1.235
Hc5	-0.3181	0.443	-0.718	0.473	-1.189	0.553
Hc6	-0.1596	0.313	-0.509	0.611	-0.775	0.456
Hs1	2.7658	0.454	6.091	0.000	1.874	3.658
Hs2	0.9643	0.457	2.110	0.035	0.067	1.862
Hs3	-0.1397	0.445	-0.314	0.754	-1.015	0.735
Hs4	0.2972	0.443	0.671	0.503	-0.574	1.168
Hs5	0.4873	0.444	1.098	0.273	-0.384	1.359
P1	-0.2995	0.543	-0.551	0.582	-1.367	0.768
P2	1.9439	0.708	2.747	0.006	0.554	3.334
P3	-0.0810	0.829	-0.098	0.922	-1.710	1.548



Vector Autoregressive Model with Exogenous Variables (VARX):

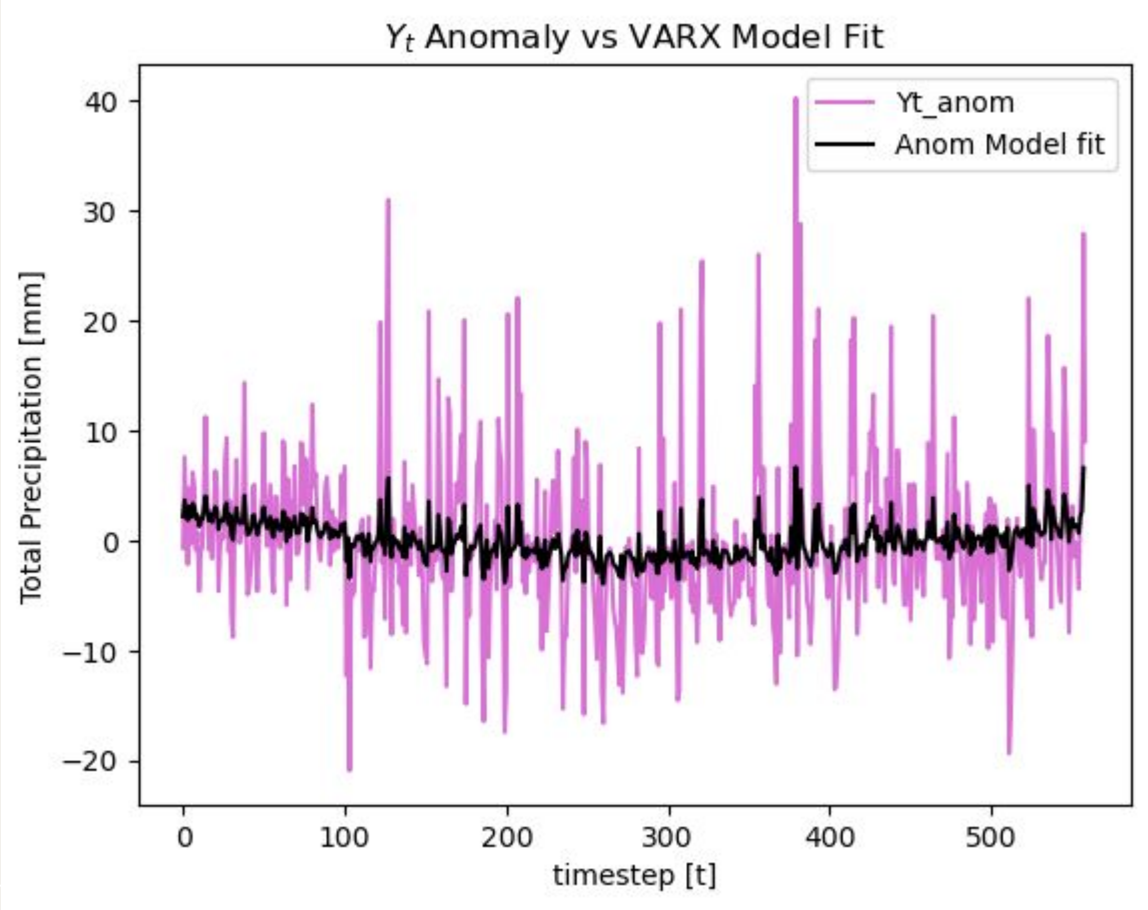
<ins>Potential Question?:</ins>

What if we remove the climatology from our time series. Will the VARX model still fit annual cycle coefficients?



Vector Autoregressive Model with Exogenous Variables (VARX):

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0038	0.313	0.012	0.990	-0.612	0.619
Yt-1	0.1817	0.042	4.298	0.000	0.099	0.265
Hc1	-0.0107	0.443	-0.024	0.981	-0.881	0.860
Hc2	0.0057	0.443	0.013	0.990	-0.865	0.876
Hc3	-0.0129	0.443	-0.029	0.977	-0.884	0.858
Hc4	0.0032	0.443	0.007	0.994	-0.868	0.874
Hc5	-0.0151	0.443	-0.034	0.973	-0.886	0.856
Hc6	0.0010	0.313	0.003	0.997	-0.615	0.617
Hs1	-0.0273	0.443	-0.062	0.951	-0.898	0.844
Hs2	-0.0046	0.443	-0.010	0.992	-0.875	0.866
Hs3	-0.0095	0.443	-0.021	0.983	-0.880	0.861
Hs4	-0.0030	0.443	-0.007	0.995	-0.874	0.868
Hs5	-0.0031	0.443	-0.007	0.994	-0.874	0.868
P1	-0.2995	0.543	-0.551	0.582	-1.367	0.768
P2	1.9439	0.708	2.747	0.006	0.554	3.334
P3	-0.0810	0.829	-0.098	0.922	-1.710	1.548

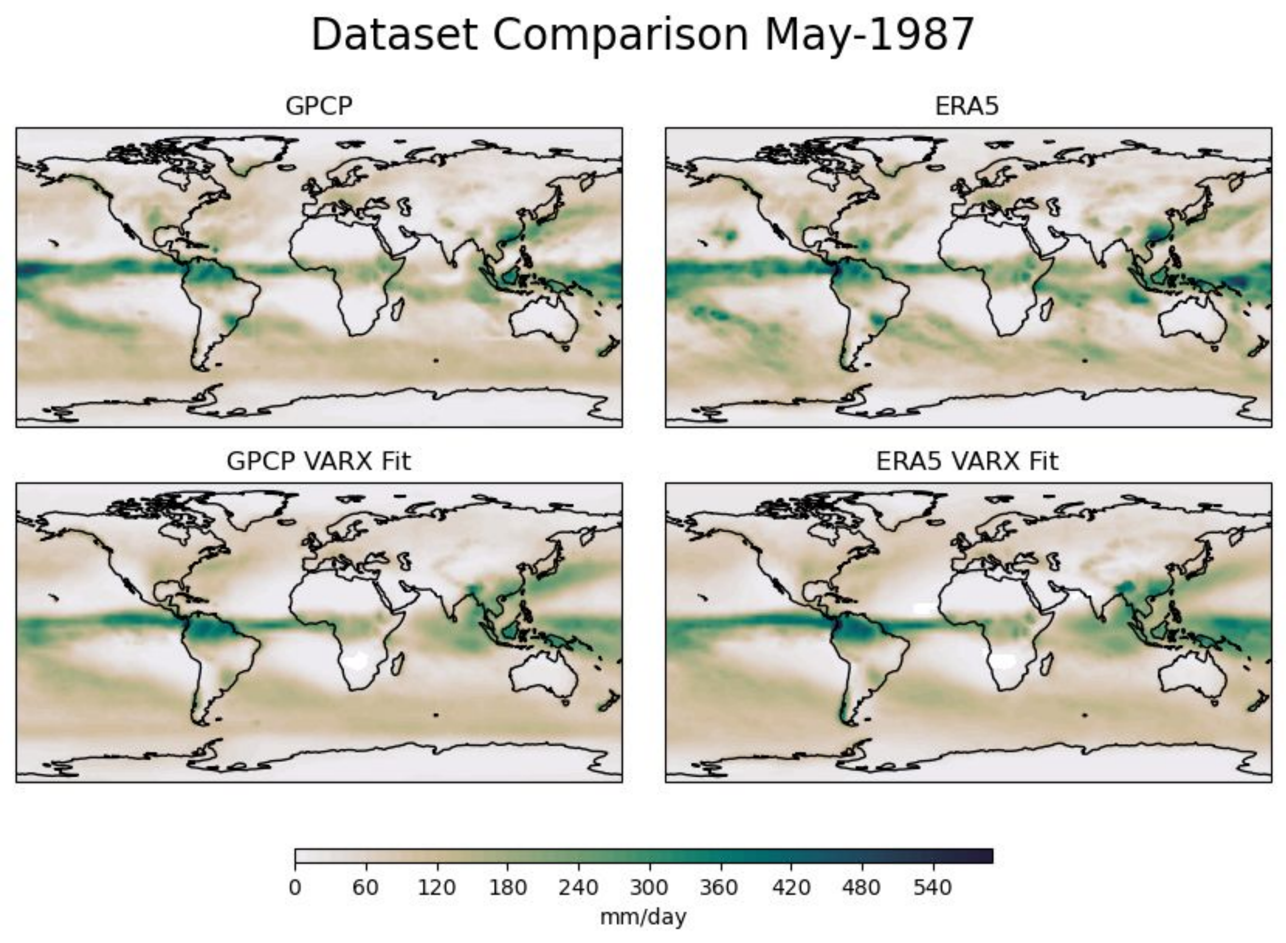


Vector Autoregressive Model with Exogenous Variables (VARX):

Procedure:

- Extract the time series for each grid point
- Define a Z matrix with all the regressors
- Create and run the OLS fit and store the results in a separate matrix

Vector Autoregressive Model with Exogenous Variables (VARX):



Stochastic Analysis:

Maximized Log-likelihood Ratio Test

$$\Lambda = e^{\ell_{H_0} - \ell_H}$$

$$\ln(\Lambda) = \ell_{H_0} - \ell_H$$

$$\ell_{H_0} = \ell(Y_{t1} + Y_{t2})$$

$$\ell_H = \ell(Y_{t1}) + \ell(Y_{t2})$$

Stochastic Analysis:

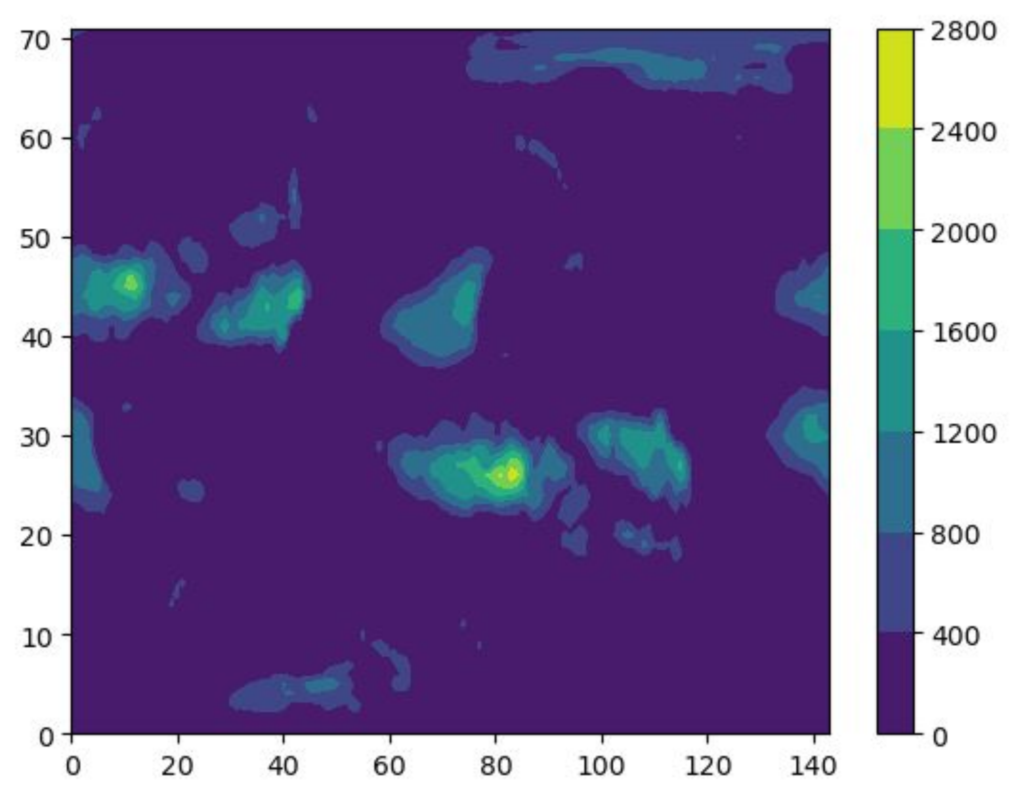
Maximized Log-likelihood Ratio Test

$$\ell_{H_0} = -\frac{T_1 + T_2}{2}[1 + \log(2\pi) + \log(\frac{S_1 + S_2}{T_1 + T_2})]$$

$$\ell_H = \ell(Y_{t1}) + \ell(Y_{t2})$$

Stochastic Analysis:

Sad Plot 😔



Conclusion:

- ❖ VARX Model can fit observational and reanalysis data fairly well. (May exclude extreme values/ smooths out spikes).
- ❖ Using the VARX Model, we were able to apply the Maximized Log-likelihood to show that both datasets are generated by different statistical processes.
- ❖ May be a good candidate for quickly comparing the statistic of two different datasets.


That's all Folks!