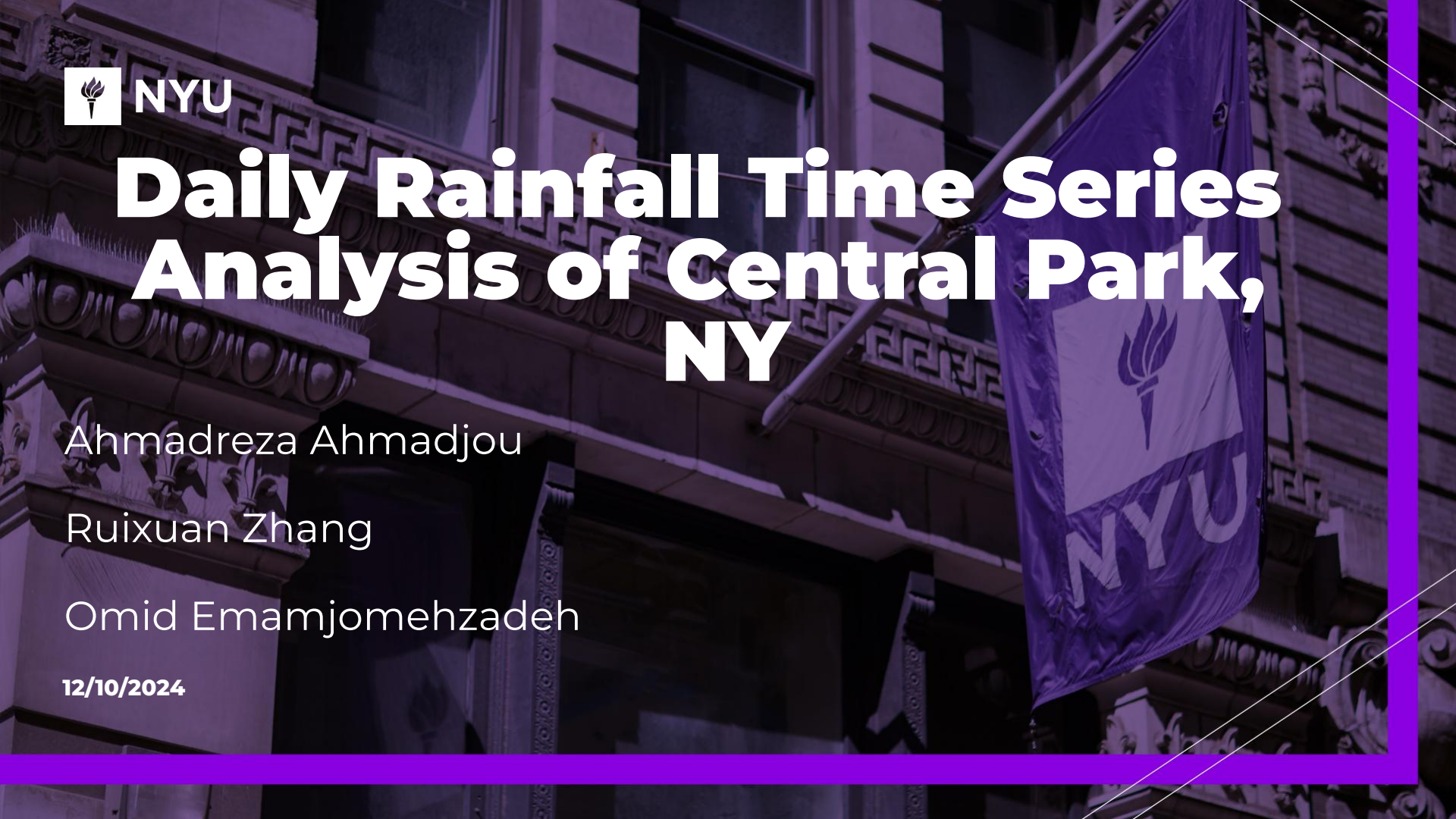


NYU

Daily Rainfall Time Series Analysis of Central Park, NY

Ahmadreza Ahmadjou

Ruixuan Zhang

Omid Emamjomehzadeh

12/10/2024

Goal

Evaluate time series analysis and prediction methods for precipitation, focusing on probabilistic models.

Significance:

Crucial for urban drainage system management.

Flood risk assessment.

Research Questions:

What are the problastic characteristics of precipitation?

Can probabilistic models predict future rainfall accurately?

How effective are these models in forecasting rare precipitation events?

What are the advantages and disadvantages of probabilistic models and deep learning models for predicting rainfall?

Dataset

Central Park daily weather dataset (1912–2022)

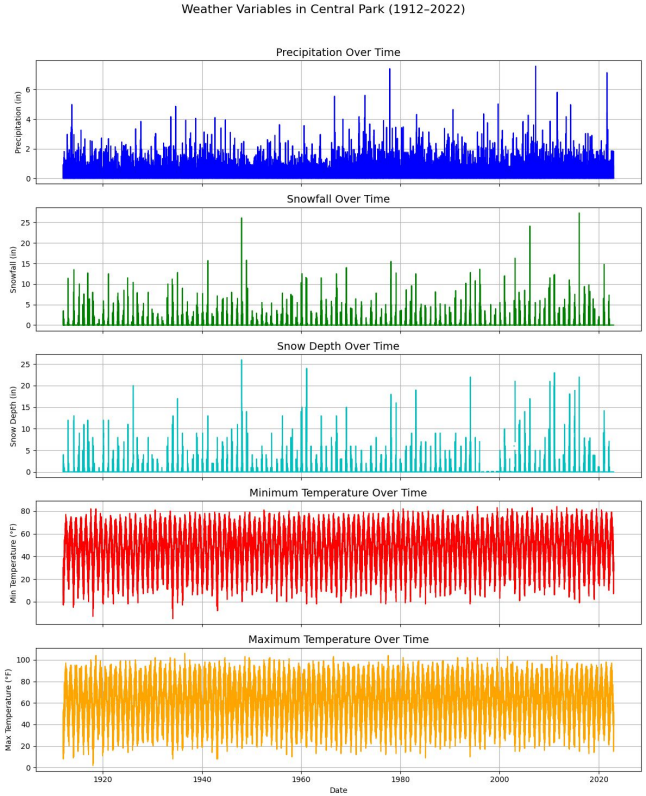
Features Included:

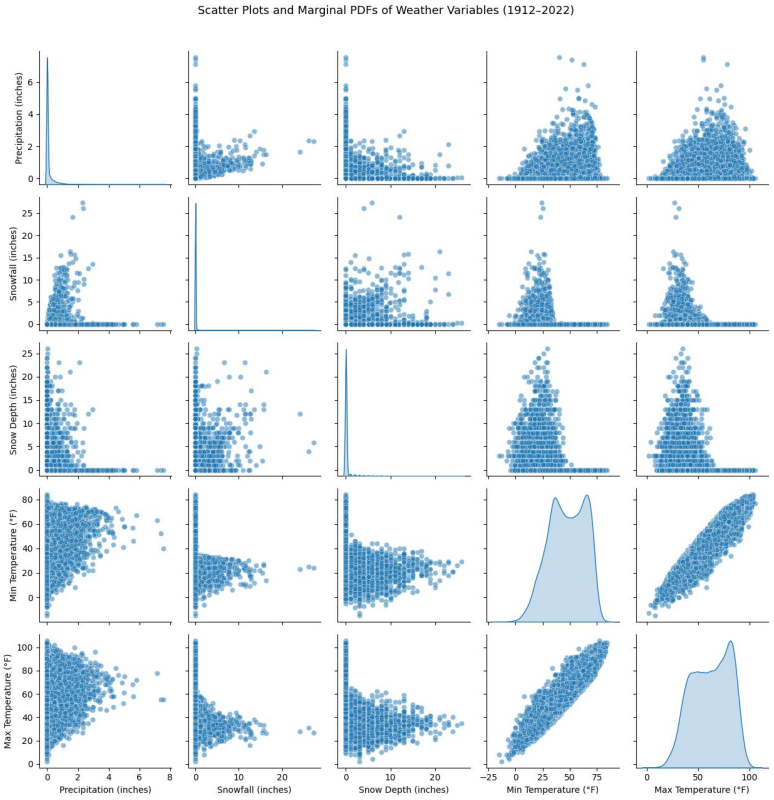
- Daily precipitation (PRCP).
- Snowfall (SNOW).
- Snow depth (SNWD).
- Minimum temperature (TMIN).
- Maximum temperature (TMAX).

Key Advantages:

- Covers over 100 years, enabling analysis of rare rainfall events.
- Publicly available on Kaggle.

Dataset visualization (daily)



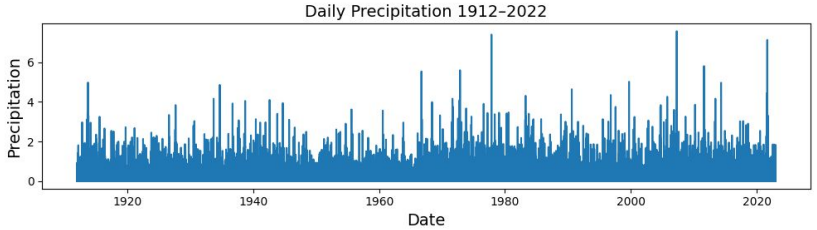


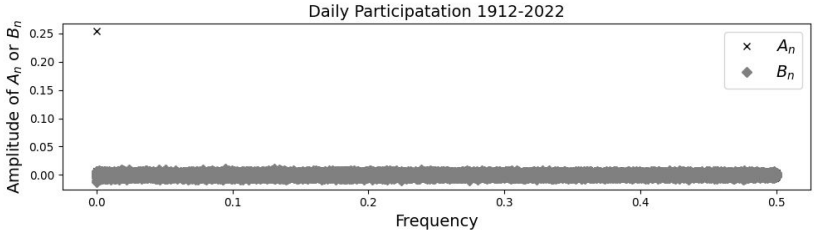
NYU

Frequency domain analysis (precipitation)

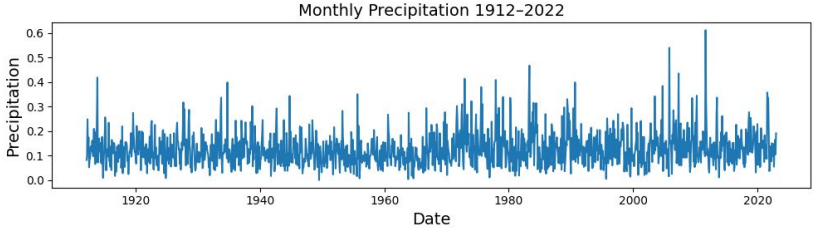
- Frequency: daily, monthly, annually

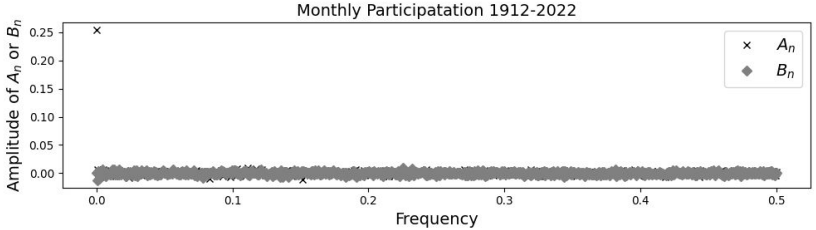
Daily



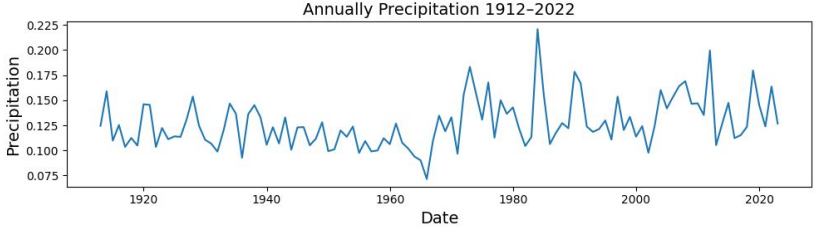


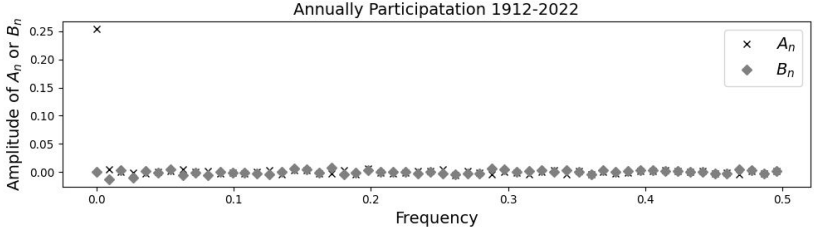
Monthly





Annually

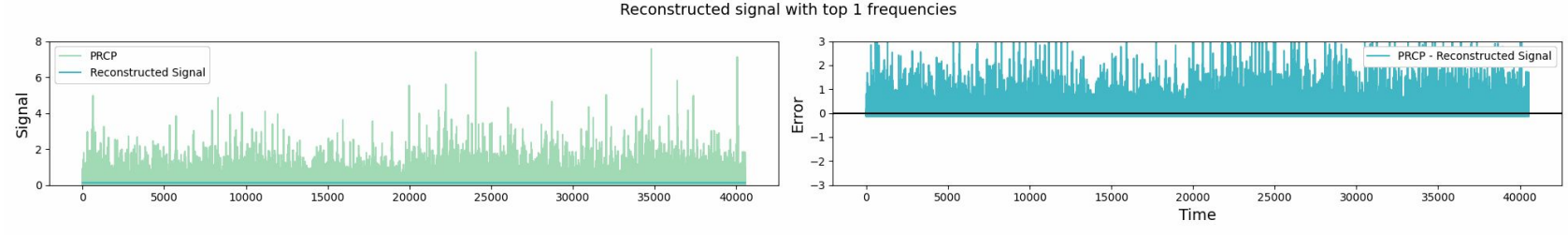




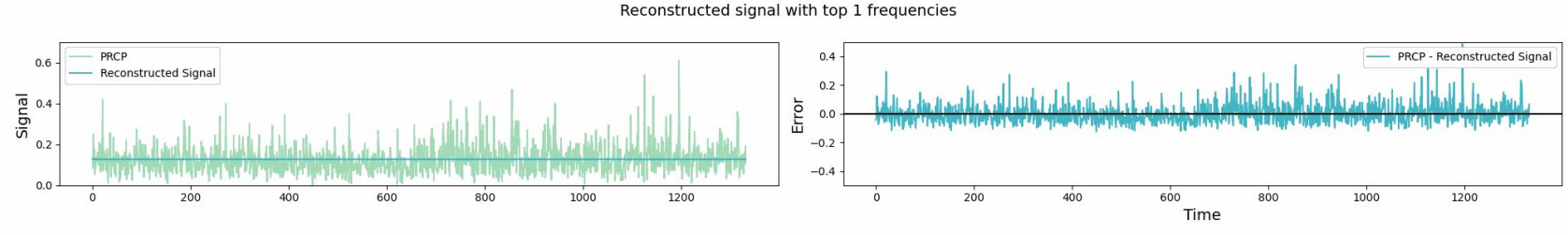
NYU

Frequency domain analysis (precipitation)

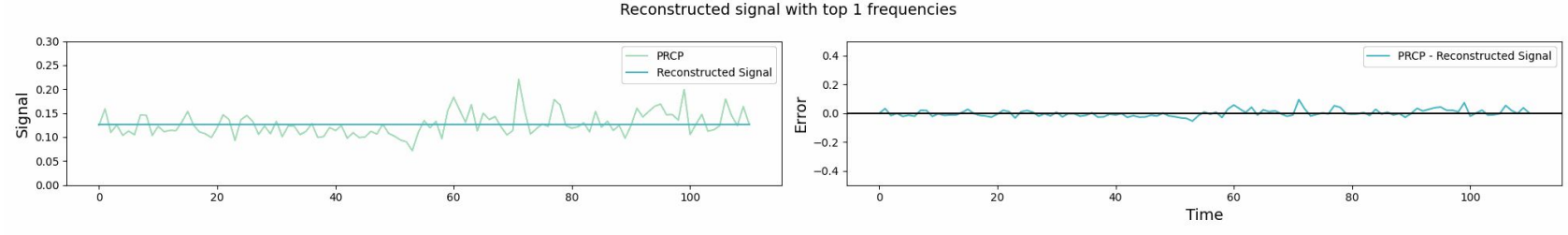
Daily



Monthly



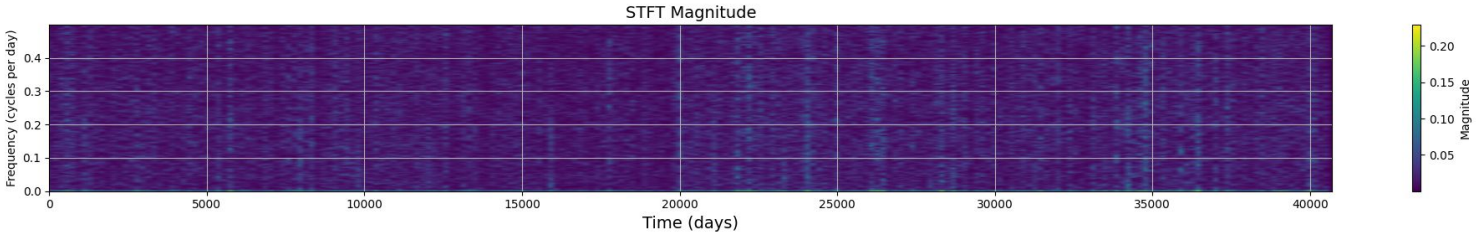
Annually



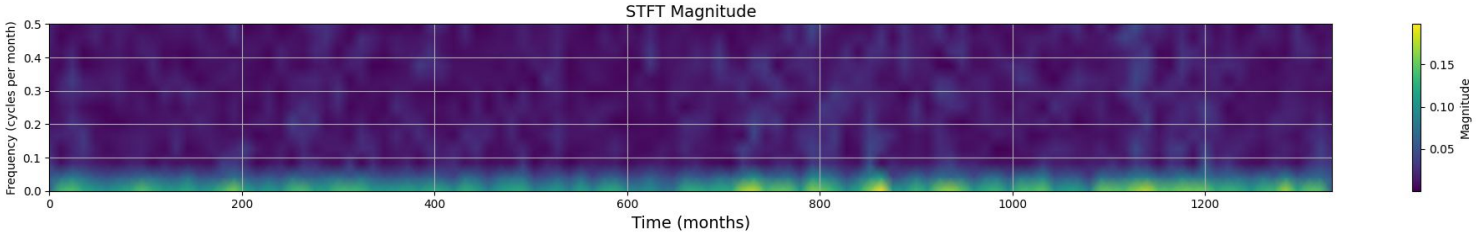
NYU

Short-time Fourier transformation (precipitation)

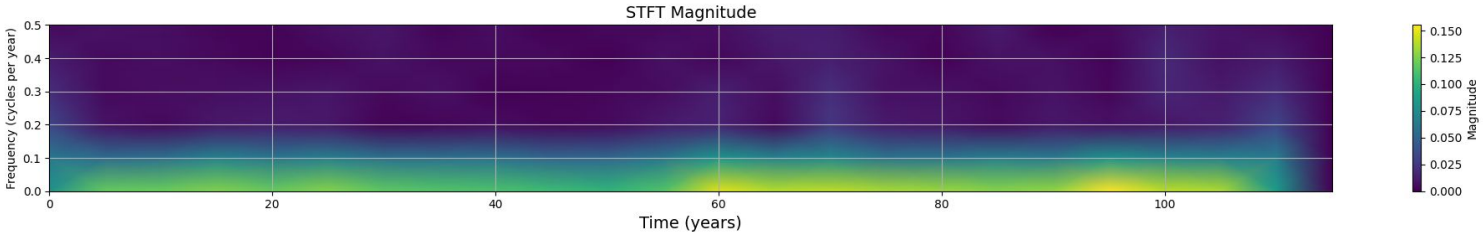
Daily



Monthly



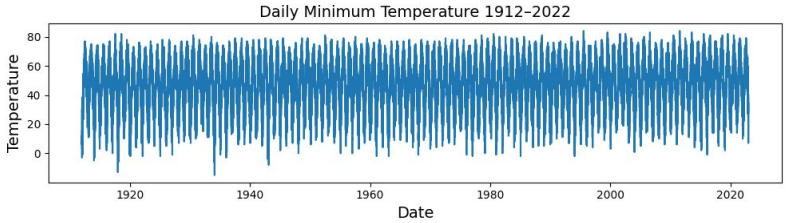
Annually

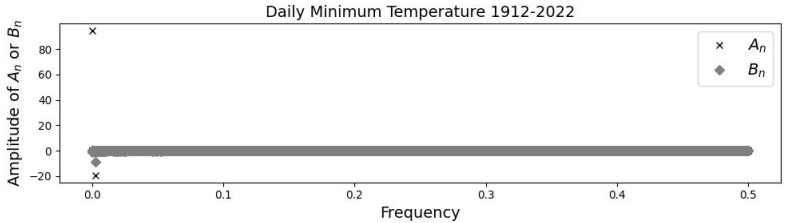


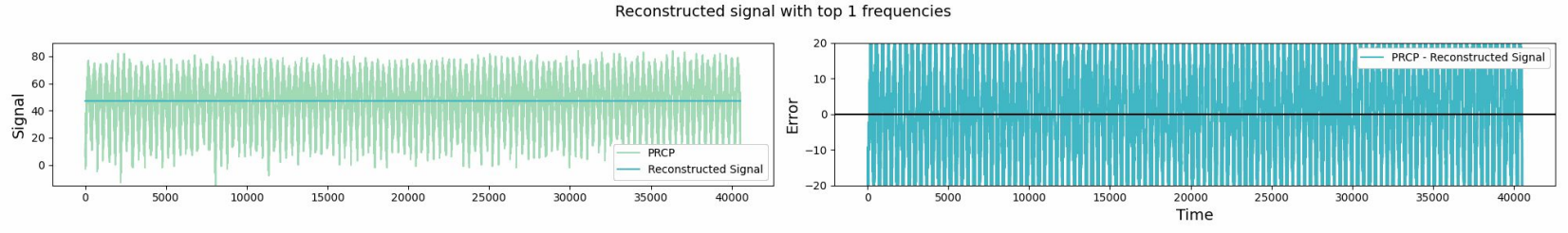
NYU

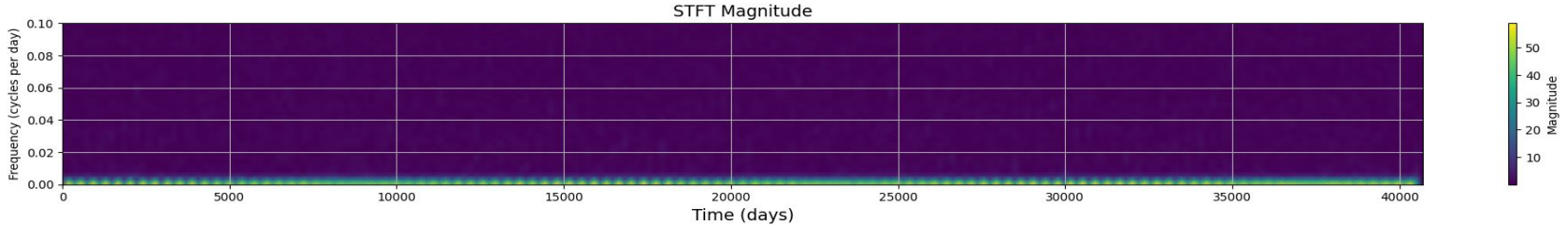
Comparison to periodic time series (min_temp)

Daily









SARIMAX

- $ARMA(p,q)$

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \longrightarrow \Phi_p(B) y_t = \Theta_q(B) \epsilon_t$$

- $ARIMA(p,d,q)$

$$\nabla^d y_t = y_t - y_{t-1} = (1-B)^d y_t \longrightarrow (1 - \phi_1 B - \cdots - \phi_p B^p)(1-B)^d y_t = (1 + \theta_1 B + \cdots + \theta_q B^q)\epsilon_t$$

$$\Phi_p(B) \nabla^d y_t = \Theta_q(B) \epsilon_t$$

- $SARIMA(p,d,q)(P,D,Q,s)$

$$\Phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D y_t = \Theta_q(B)\Theta_Q(B^s)\epsilon_t$$

Example: SARIMA(1,0,2)(2,0,1,5)

$$y_t = \phi_1 y_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \Phi_1 (y_5 - \phi_1 y_{t-6}) + \Phi_2 (y_{t-10} - \phi_1 y_{t-11}) + \Theta_1 (\epsilon_{t-5} + \theta_1 \epsilon_{t-6} + \theta_2 \epsilon_{t-7})$$

- $SARIMAX(p,d,q)(P,D,Q,s)$

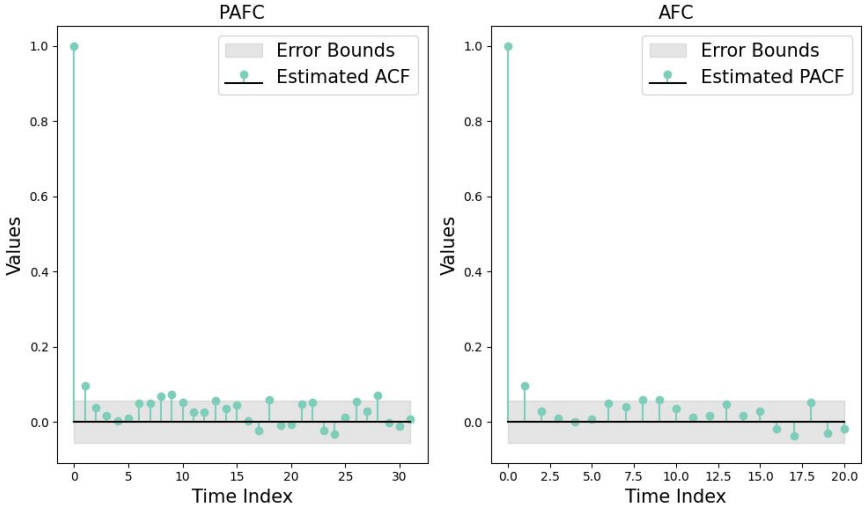
$$\Phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D(y_t - x_t^T \alpha) = \Theta_q(B)\Theta_Q(B^s)\epsilon_t$$

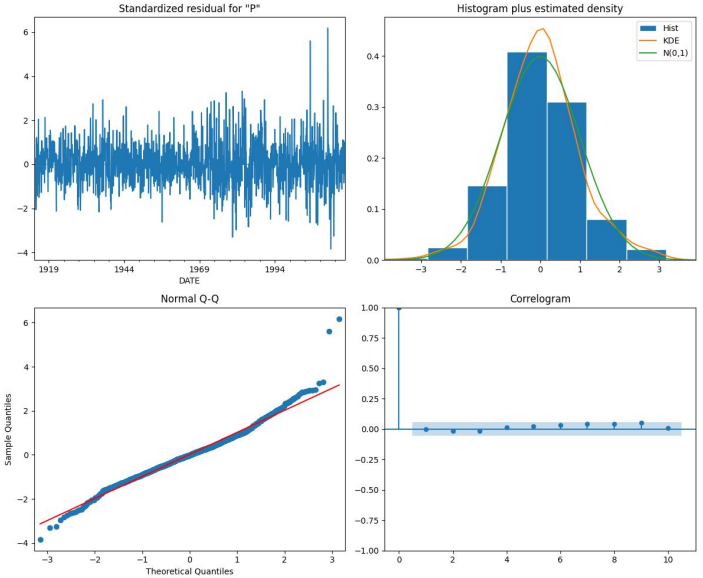
$$\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_k,]^T$$

NYU

SARIMAX (precipitation)

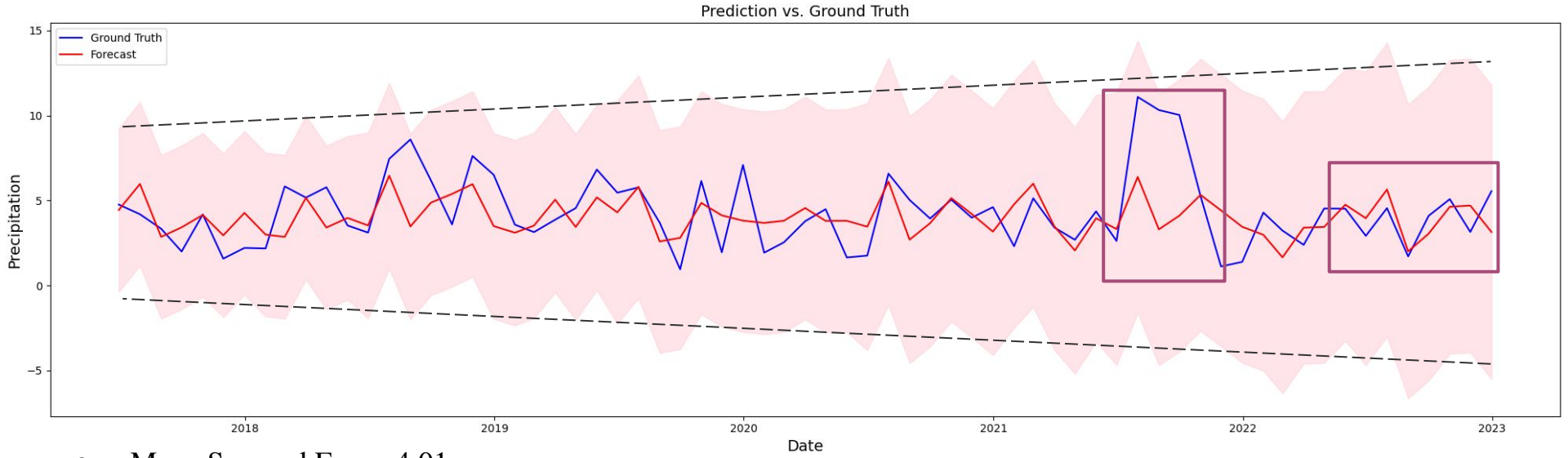
	MaxTemp	MinTemp	Snowfall	Snow depth	ar.L1	ma.L1	ar.S.L12
coef	-0.5083	0.5158	0.0032	0.0009	0.0492	-1.000	-0.4678
P>|z|	0.000	0.000	0.047	0.566	0.054	0.952	0.000





SARIMAX (precipitation)

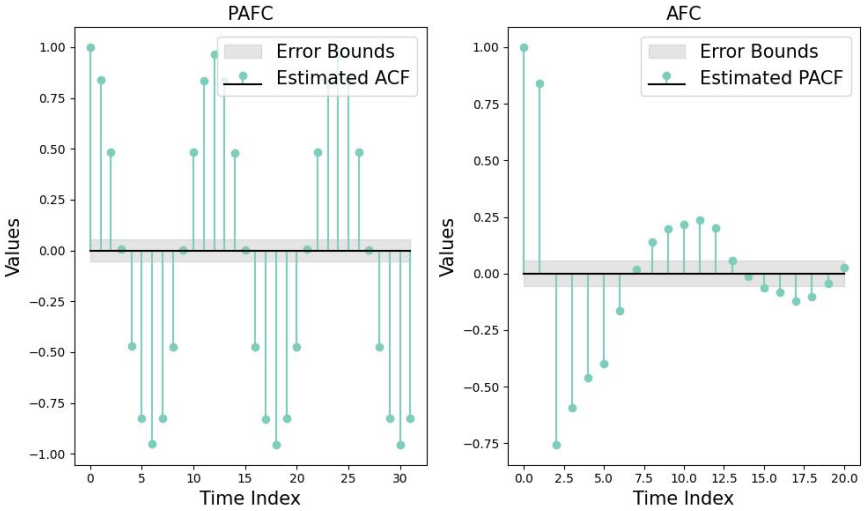
SARIMAX (1,1,1)(1,1,0,12)

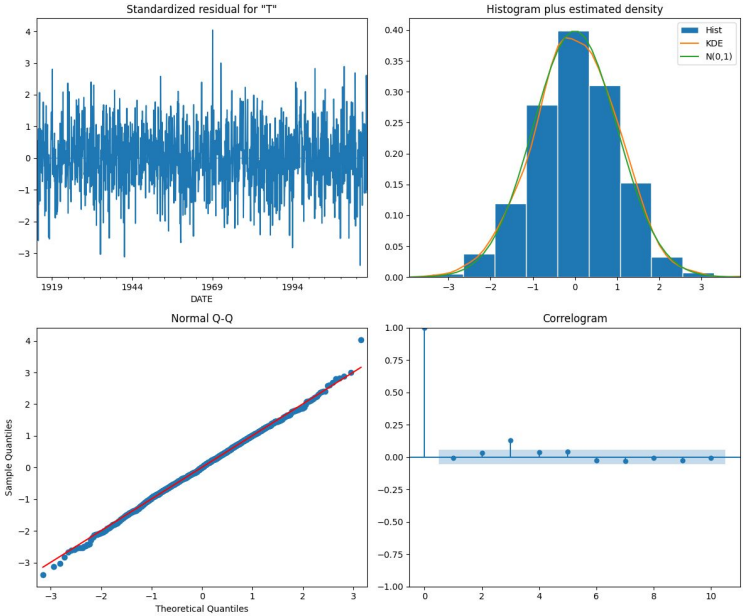


- Mean Squared Error: 4.01
- Mean Absolute Percentage Error: 42.8%

SARIMAX (min_temp)

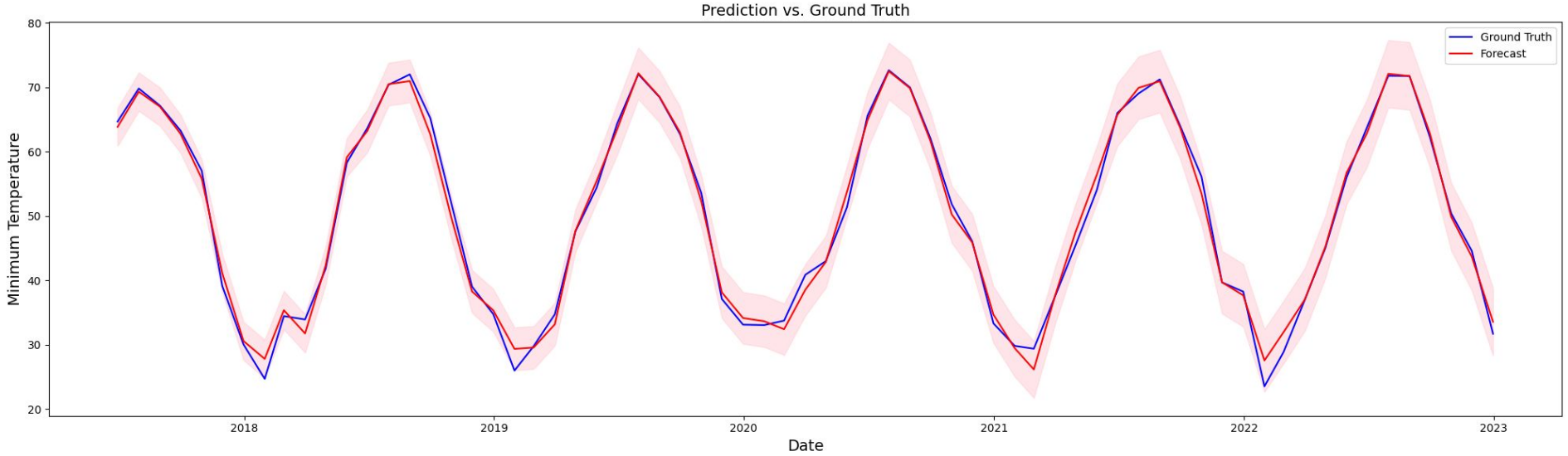
	MaxTemp	MinTemp	Snowfall	Snow depth	ar.L1	ma.L1	ar.S.L12
coef	0.8058	0.1944	0.0032	-0.0014	0.1407	-1.000	-0.5044
P>|z|	0.000	0.000	0.000	0.028	0.000	0.986	0.000





SARIMAX (min_temp)

SARIMAX (1,1,1)(1,1,0,12)

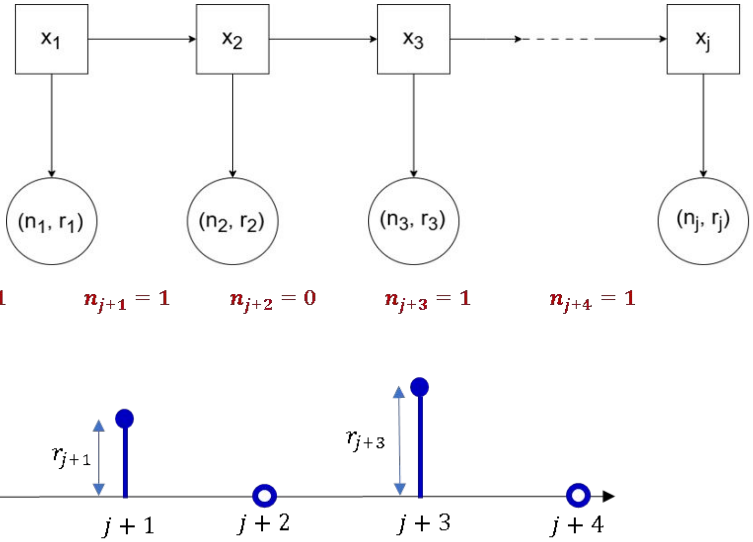


- Mean Squared Error: 2.02
- Mean Absolute Percentage Error: 2.66%

Marked Point Process Decoder

$$\hat{x}_j = \hat{x}_{j-1} + \epsilon_j \text{ where } \epsilon_j \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Hidden arousal state — Process noise



- Precipitation event: n_j with Bernoulli mass function

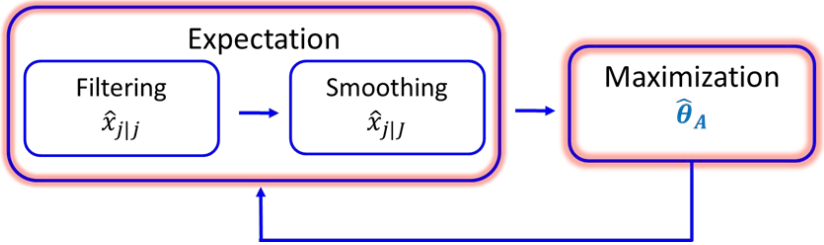
$$Prob(n_j = 1) = a_j \quad \text{where } a_j = \frac{1}{1 + e^{-(\hat{x}j + \beta)}}$$

- Precipitation intensity level: r_j

$$r_j = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{x}_j + \nu_j, \qquad \nu_j \sim \mathcal{N}(0, \sigma_\nu^2)$$

Parameter Vector
$\widehat{\theta}_A = \{\gamma_0, \gamma_1, \sigma_\nu^2, \sigma_\epsilon^2\}$

Observation Vector
$Y^J = \{(n_1, r_1), \ldots, (n_J, r_J)\}$



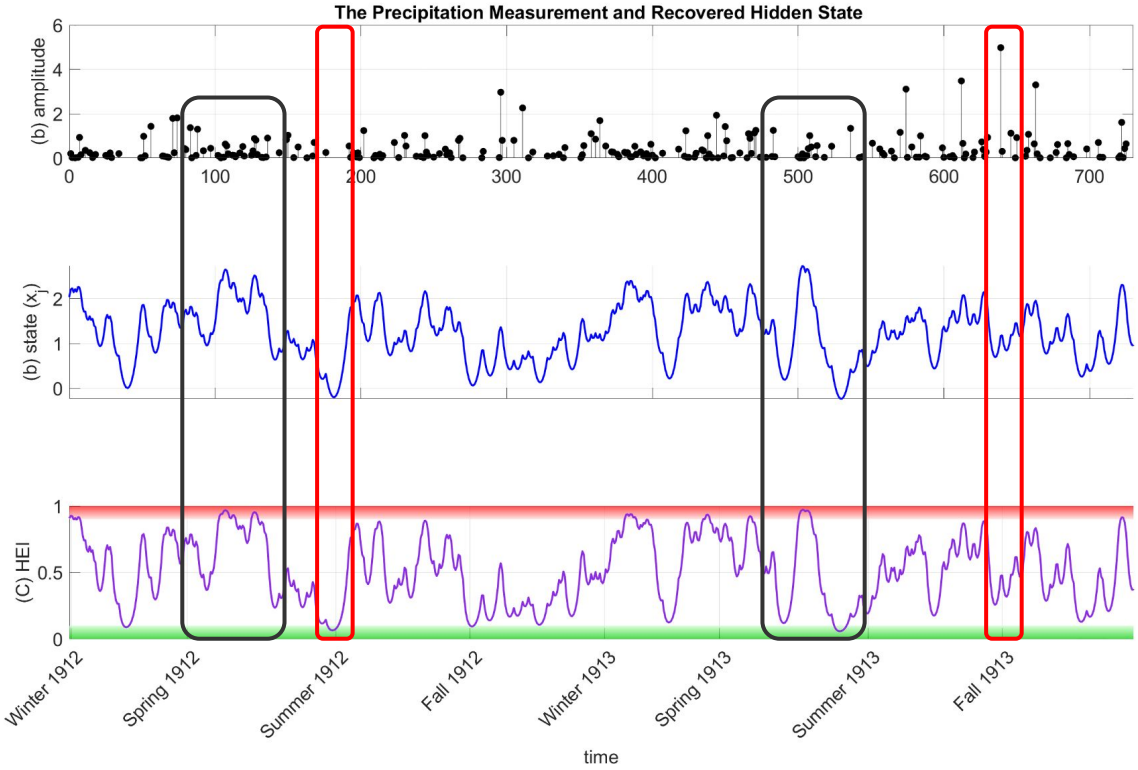
NYU

Wickramasuriya, D. S., & Faghih, R. T. (2020). A marked point process filtering approach for tracking sympathetic arousal from skin conductance. *IEEE Access*, *8*, 68499-68513.

Marked Point Process Decoder

Observations:

- The Hidden-State is both dependent on intensity and frequency
- The near-end/beginning of the season we have lowest values



LSTM and GP

Feature Selection & Scaling:

- Selected weather-related features including precipitation, snowfall, and temperature,and time related features including season, month, and day of the year.
- Standardized the features using StandardScaler.

Sequence Creation:

- Defined a function to generate input-output sequences for time series data.
- Used 30-day sequences as input for prediction tasks.

Data Splitting:

- Split sequences into training (80%) and test (20%) sets for model evaluation.

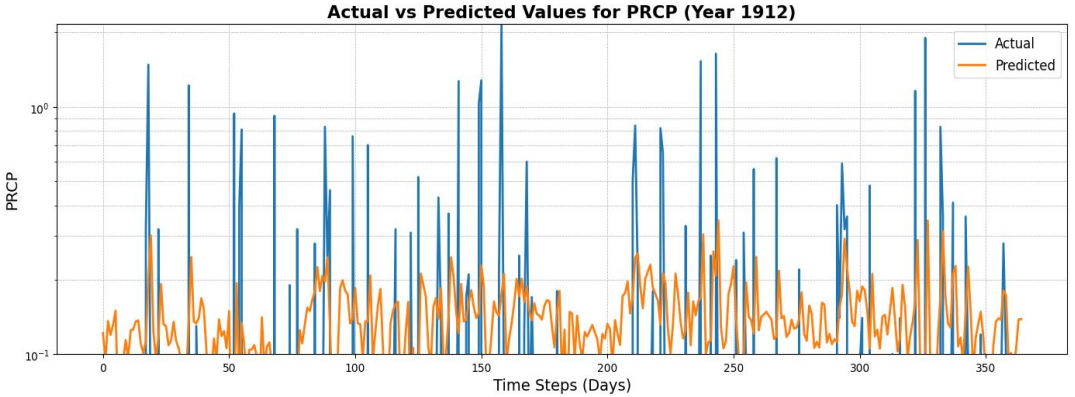
LSTM and GP

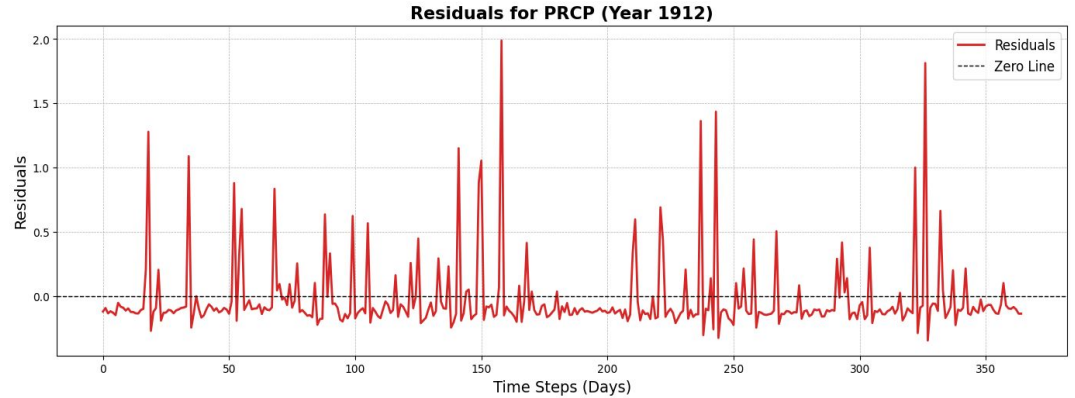
Hyperparameter Optimization and Training:

- Leveraged Optuna to tune critical hyperparameters (hidden size, layers, dropout, learning rate, batch size) and implemented a 50-epoch training loop with Adam optimizer, MSE loss, and OneCycleLR scheduler, dynamically updating DataLoaders for each trial.
- Best hyperparameters:
  hidden_size: 215,
  num_layers: 1,
  dropout': 0.23,
  learning_rate: 0.0123,
  batch_size: 16

Best Model Selection and Training:

- Identified optimal hyperparameters from 100 optimization trials and trained the final WeatherLSTM model using these parameters to maximize performance and accuracy on the validation dataset





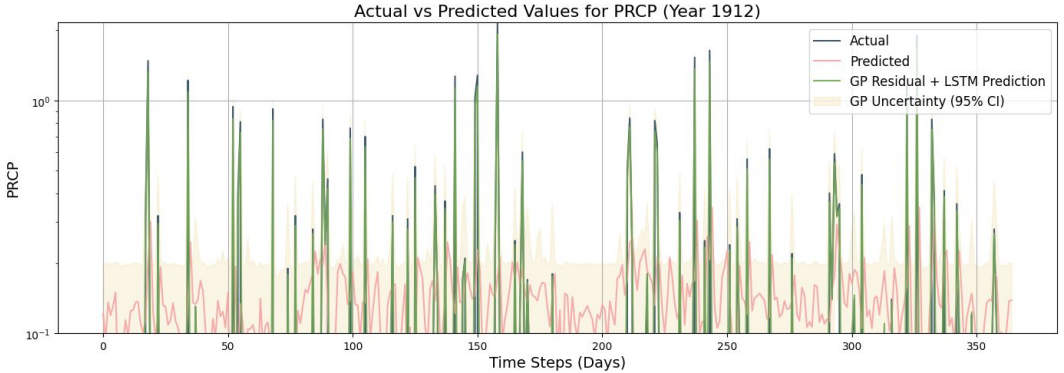
LSTM and GP

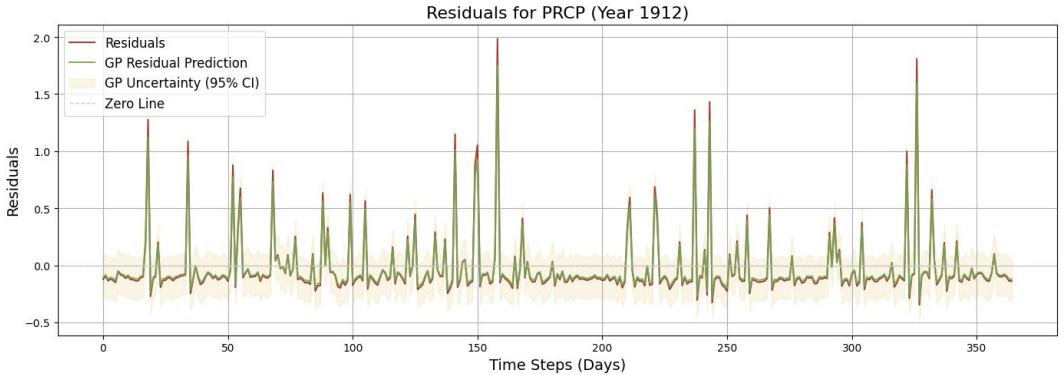
Data Preparation and Residual Calculation:

- Extracted the first year's actual and predicted values for a specific feature (e.g., precipitation), calculated residuals by subtracting predictions from actuals, and used time steps as input for further modeling.

Gaussian Process Regression for Residuals:

- Fitted a Gaussian Process with a custom kernel (Constant Kernel * RBF) to the residuals, optimizing the model to capture patterns and uncertainties, and generated predictions with 95% confidence intervals.





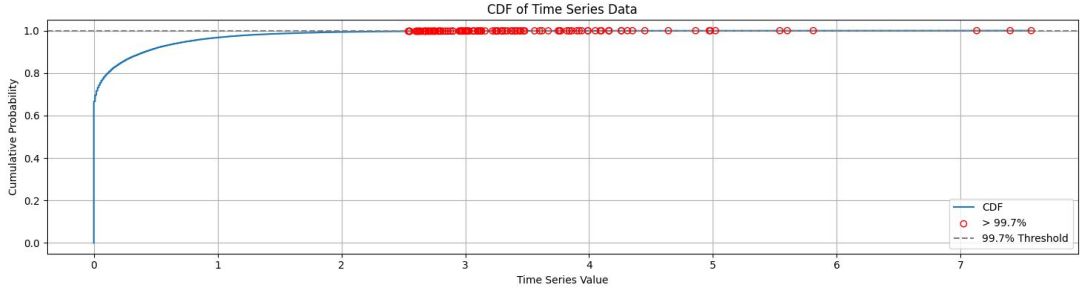
Thanks for your attention!

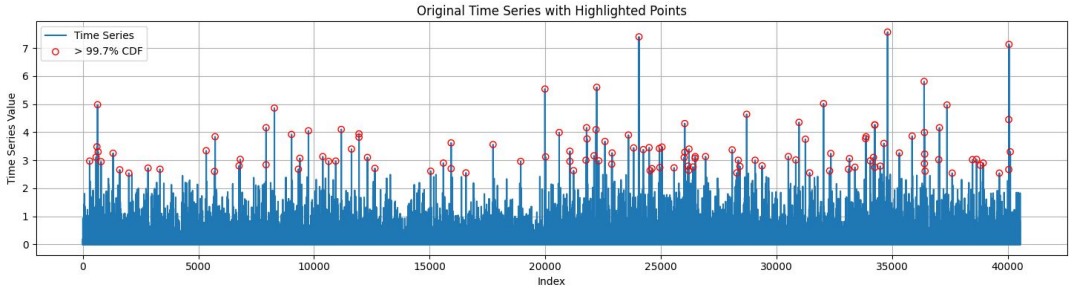
Questions?

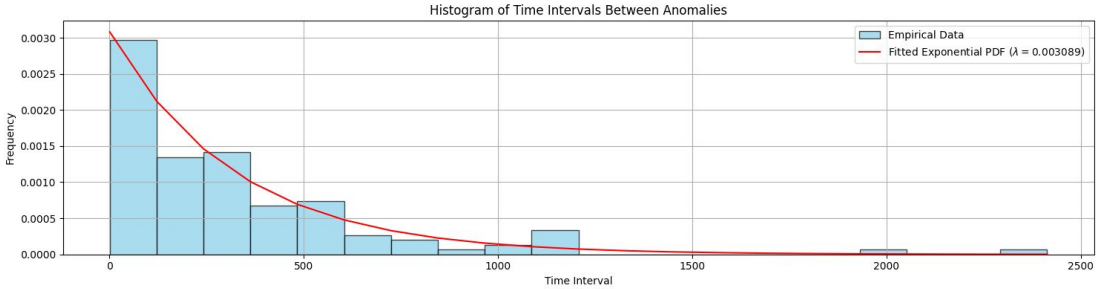
NYU

(Back-up slides)Extreme event statistics

- Participation at CDF >= 99.7%
- 124 events
- Arrival interval ~ Exp($3.09*10^{-3}$)







Extra Slides

$$p(n_j \cap r_j | x_j) = \begin{cases} 1 - p_j & \text{if } n_j = 0 \\ p_j \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{\frac{-(r_j - \gamma_0 - \gamma_1 x_j)^2}{2\sigma_v^2}} & \text{if } n_j = 1 \end{cases}$$

LSTM and GP

Model Architecture:

- Implemented a bidirectional LSTM with dropout and a fully connected layer.
- Designed for multi-feature time series prediction.

Training Setup:

- Used Mean Squared Error (MSE) loss, Adam optimizer, and a learning rate scheduler.
- Integrated gradient clipping to avoid exploding gradients.

Training Loop:

- Trained the model for 60 epochs with early stopping based on validation loss.
- Achieved adaptive learning rate adjustment using OneCycleLR.

Prediction & Evaluation:

- Generated predictions on the test set and reversed scaling for interpretability.
- Compared ground truth vs. predicted values and analyzed residuals.
- MSE of validation set is 0.25