A Practical Approach to Spatio-Temporal Modeling

Diagnosing Model Performance in Capturing Local Behavior

Joonwon Lee
advisor: Michael Stein

Rutgers University, Statistics department

A Practical Apporach to Spatio-Temporal modeling Ozone

Motivation

The purpose of the study is to provide a systematic approach to modeling ozone data and a better answer to the question "why did you choose this model?" rather than simply citing a small mean squared prediction error or higher likelihood.

Data: Total Column Ozone

- Total column ozone measured by GEMS(Geostationary Korea Multi-Purpose Satellite).
- 2022.June $-$ present.
- Observation cycle: 8 times during the daytime, at 08:45, 09:45, ..., 16:45 local time.
- Spatial resolution is 3.5 km $\times$ 8 km.

Overview of the Geographical scope

Data for Local Behavior Analysis Purposes

We chose a small grid from North 5° to North 10° and East 110° to East 120°. Each orbit contains 20,000 to 28,000 data points in the grid.

Overview of the Geographical scope

By choosing a grid near equator, we aim to suppress seasonal effect because near equator where we expect constant sunlight throughout the year.

Data Pre-processing

To keep track of the same location over time, ozone values were averaged over $0.05^\circ \times 0.05^\circ$, ensuring that each orbit contains exactly 20,000 data points and every orbit shares the same locations.

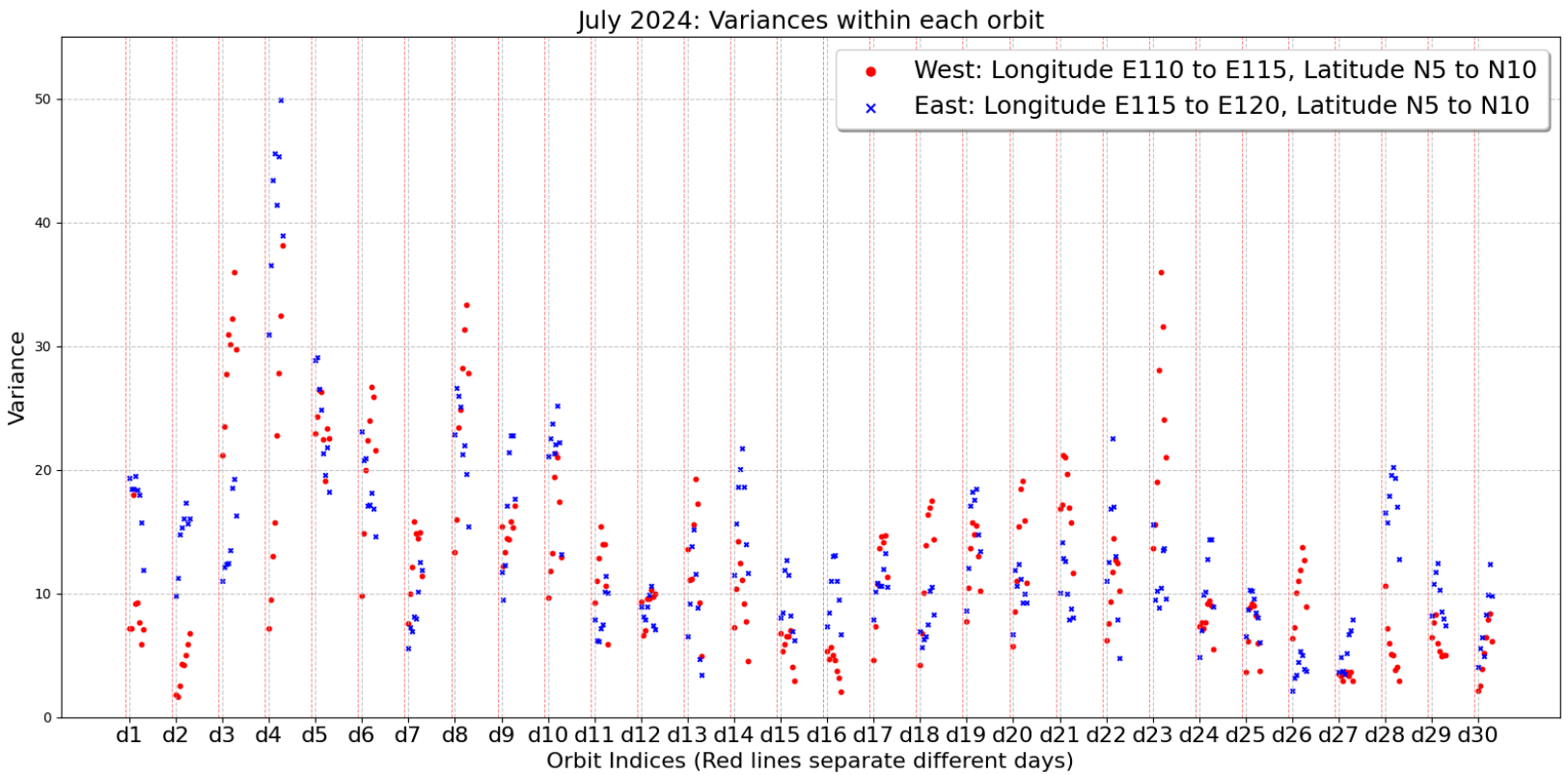
Diagnostic Tools in Spatial Modeling

- Stationarity
- Isotropy
- Symmetry in Cross-Variogram
- A diagnostic tool to evaluate whether a model is capturing the local behavior of the process.

1. Stationarity

Examining Variance Sufficient to Verify Stationarity?

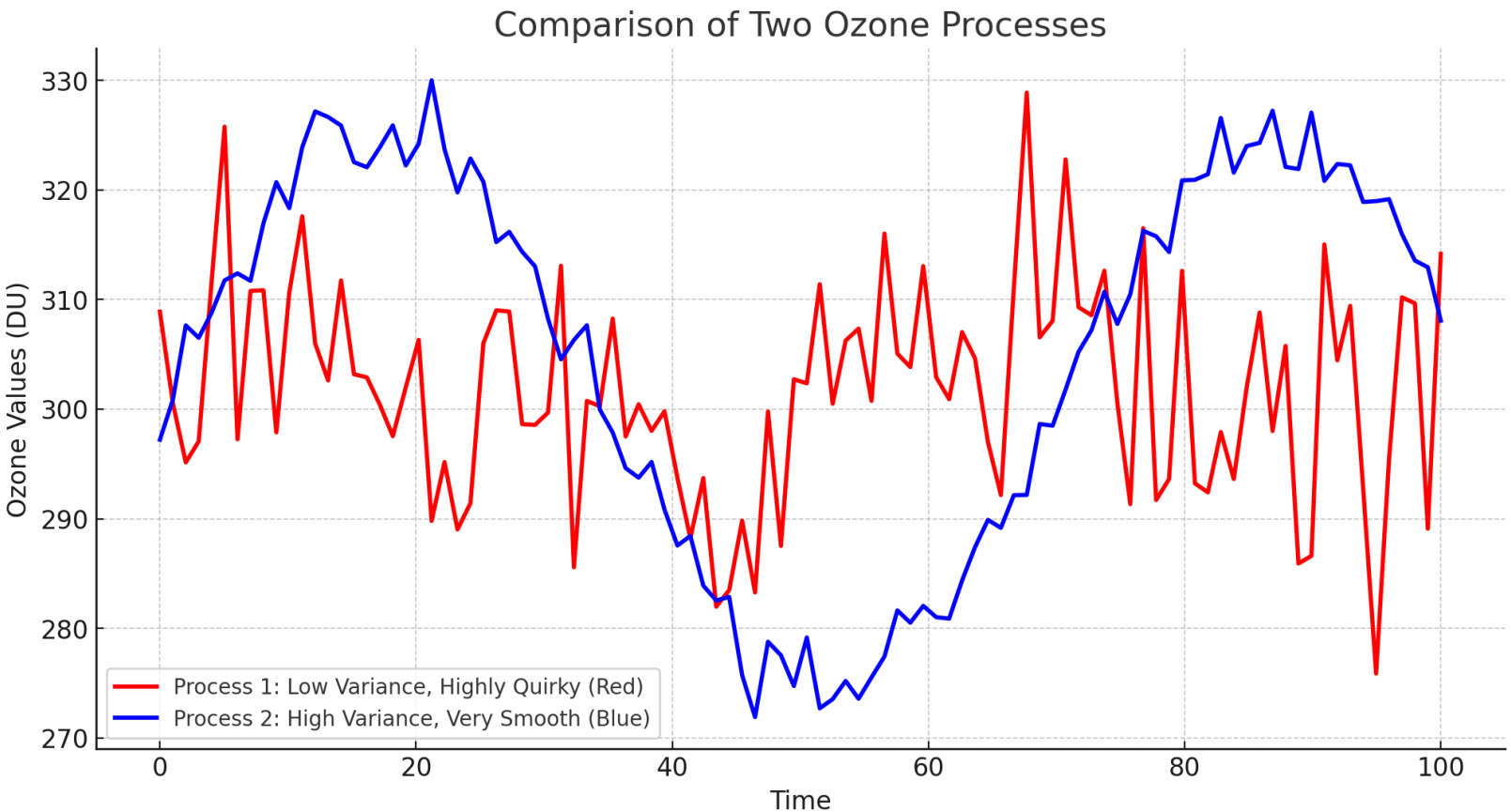
The Ozone process is not spatially stationary.



1. Stationarity

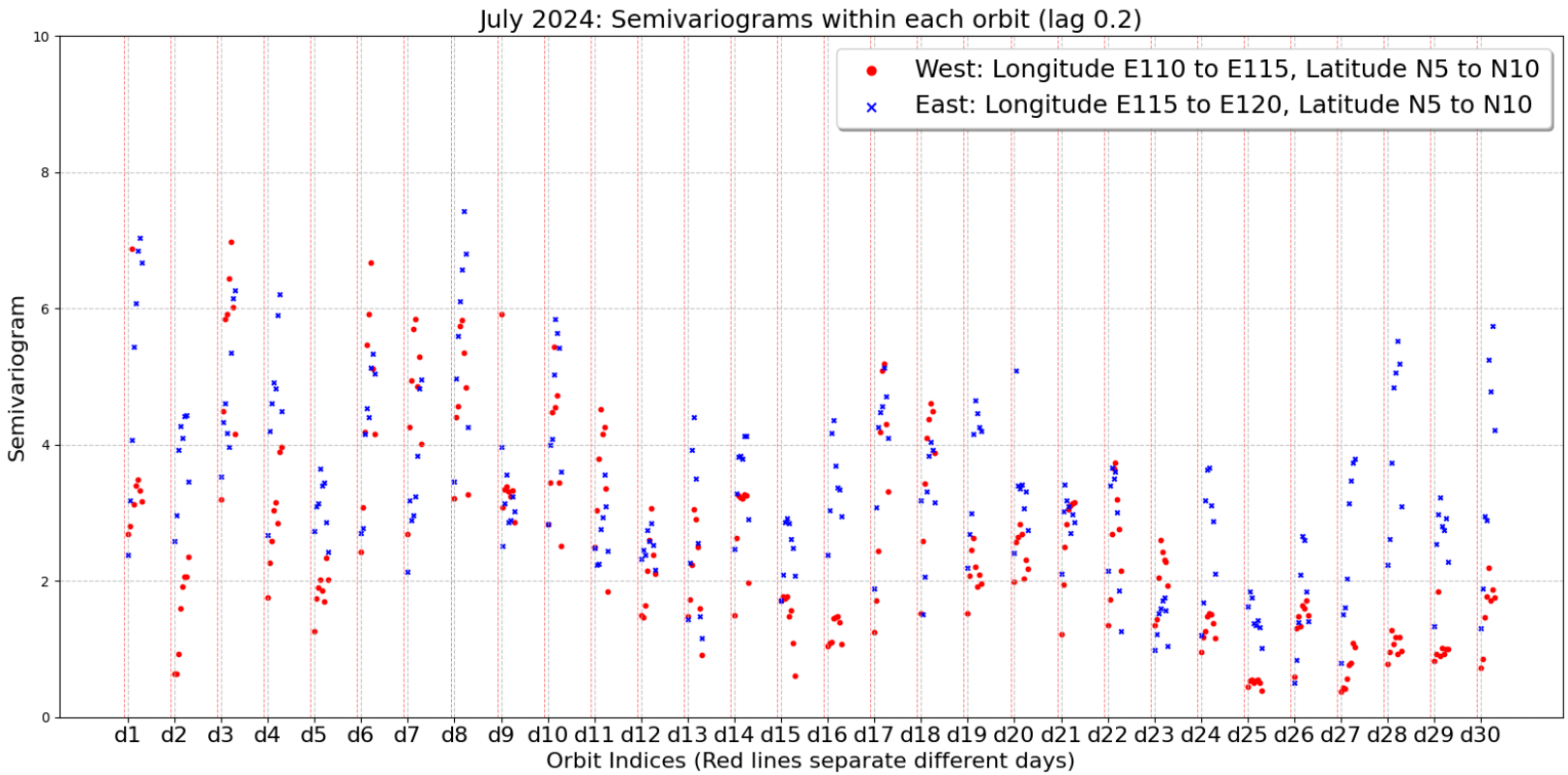
Examining Variance Sufficient to Verify Stationarity?

Semi-variogram : $\gamma(h) = 0.5 * \mathbb{E}[(\mathbf{Z}(x,t) - \mathbf{Z}(x+h,t))^2]$



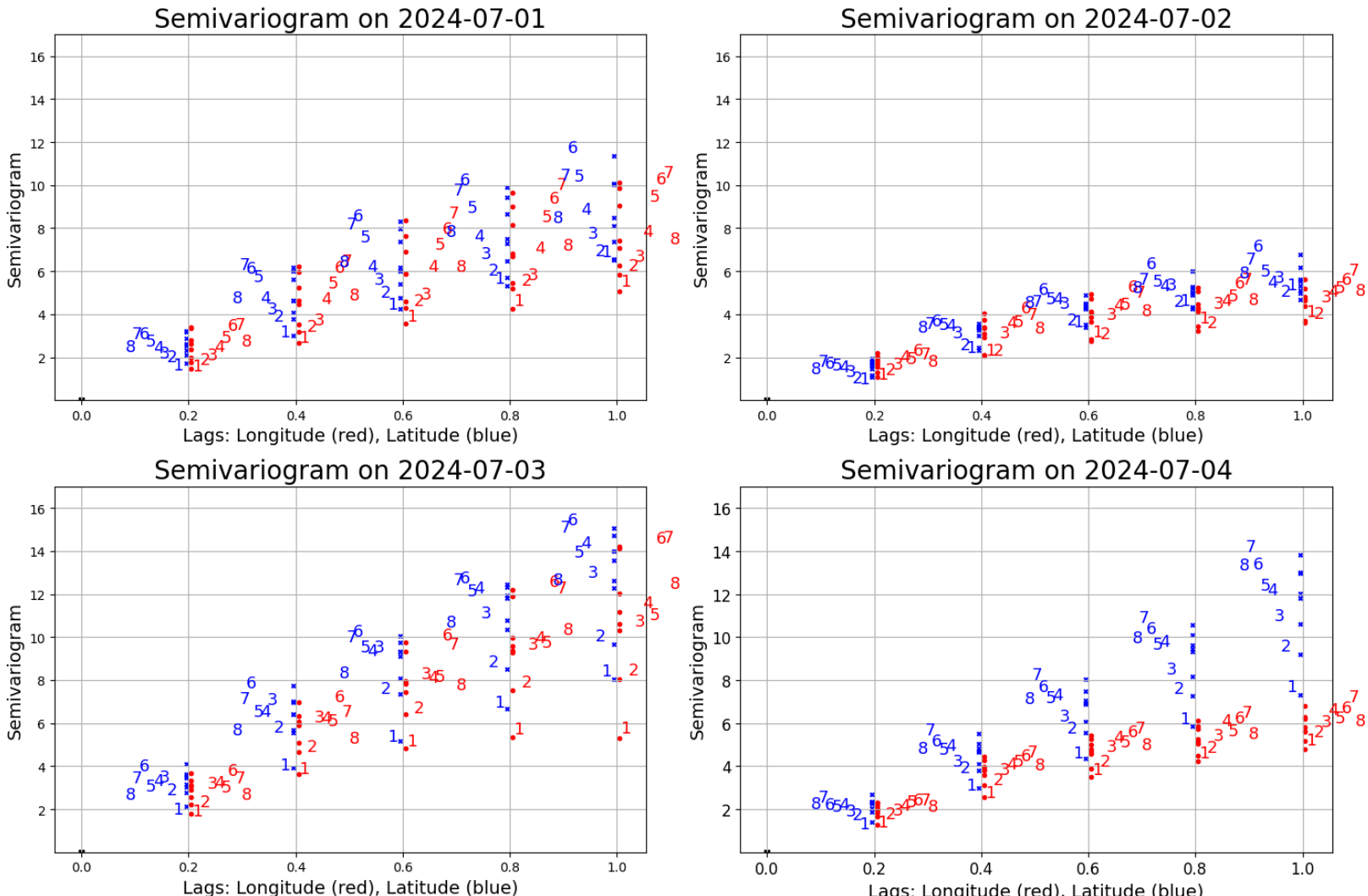
1. Stationarity

It's important to consider both variance and the short-lag semivariogram to model both overall variability and local variability.



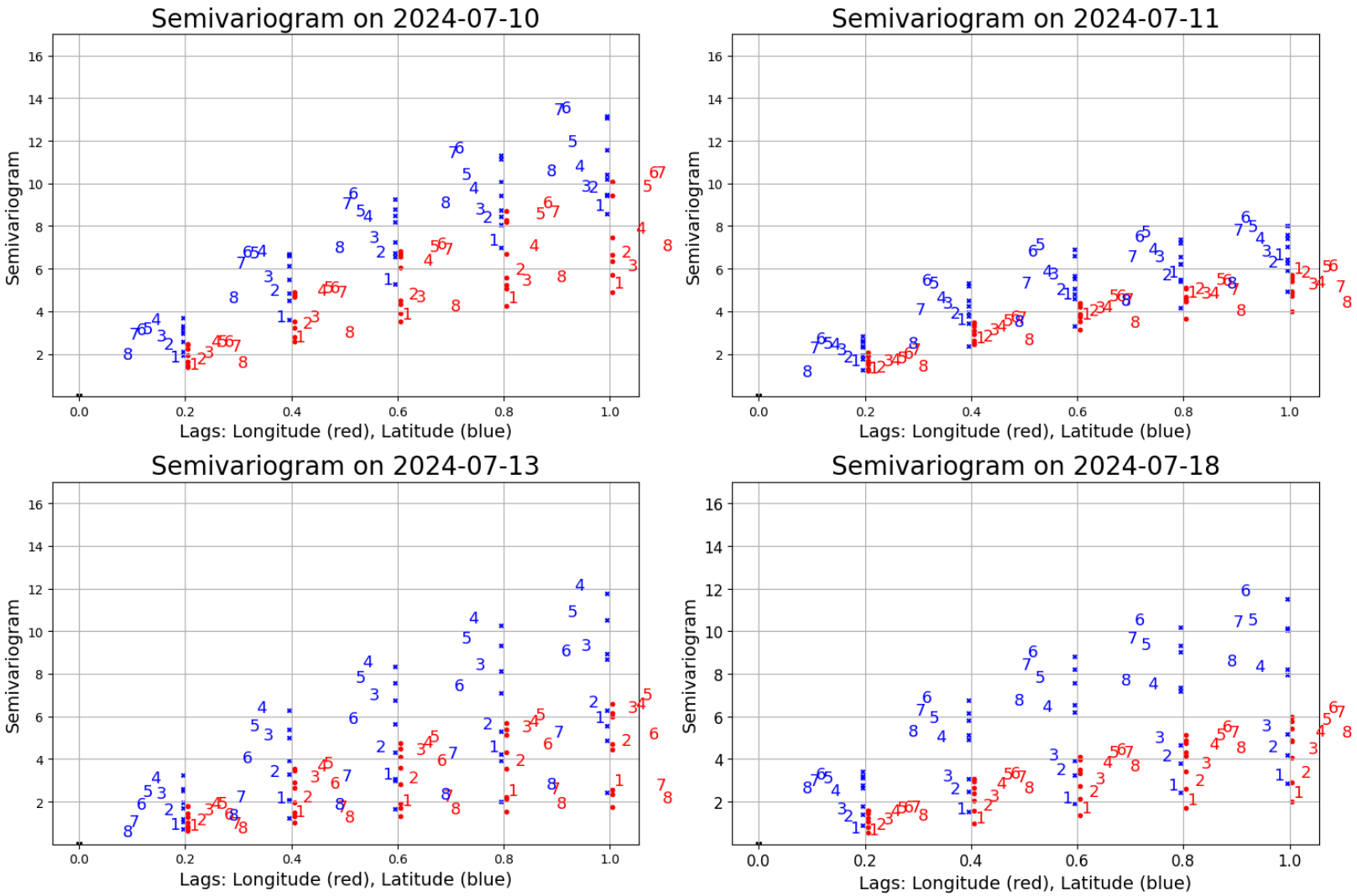
2. Isotropy or Anisotropy?

1. **Anisotropy**: Variability is greater along Latitude than Longitude.
2. **Distinct diurnal cycles** observed on some days.



2. Isotropy or Anisotropy?

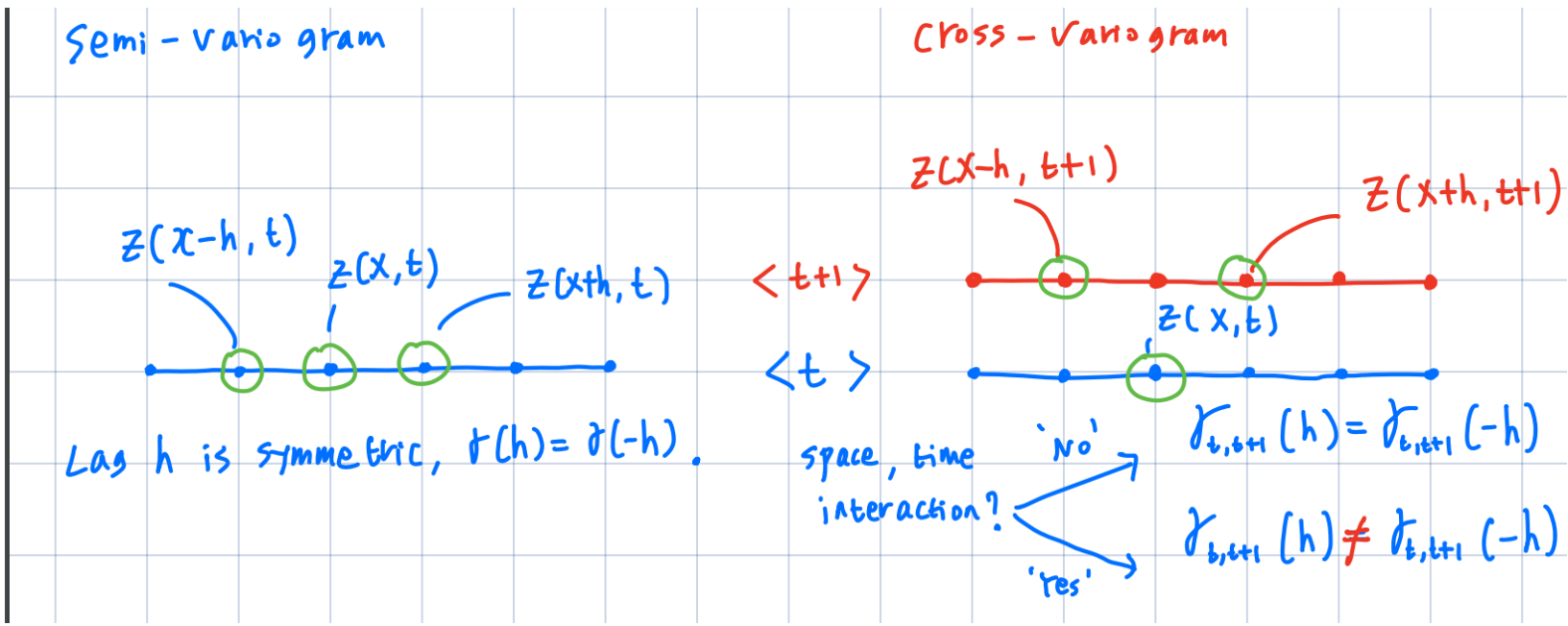
We would want any spatio-temporal model to capture these strong **diurnal** pattern in variograms



3. Symmetry in Cross Variogram (Space and Time Interaction)

Semi-variogram : $\gamma(h) = 0.5 * \mathbb{E}[(\mathbf{Z}(x,t) - \mathbf{Z}(x+h,t))^2]$

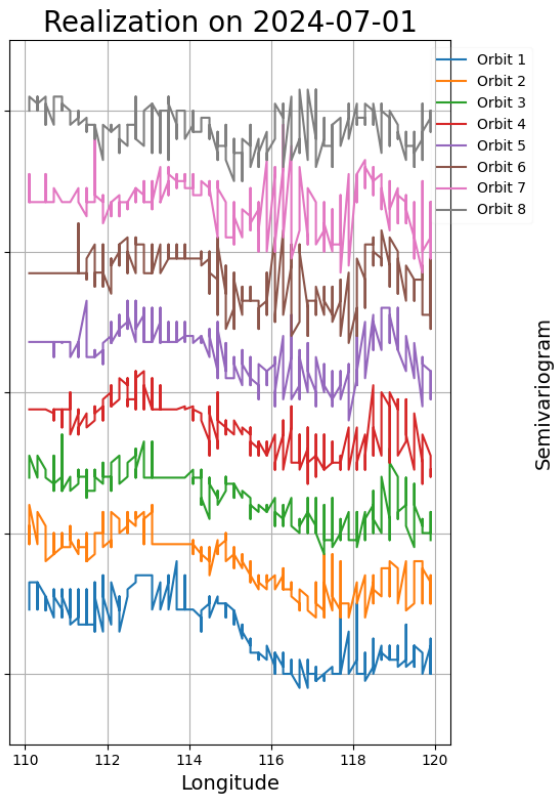
Cross-variogram : $\gamma_{t,t+1}(h) = 0.5 * \mathbb{E}[(\mathbf{Z}(x,t) - \mathbf{Z}(x+h,t+1))^2]$

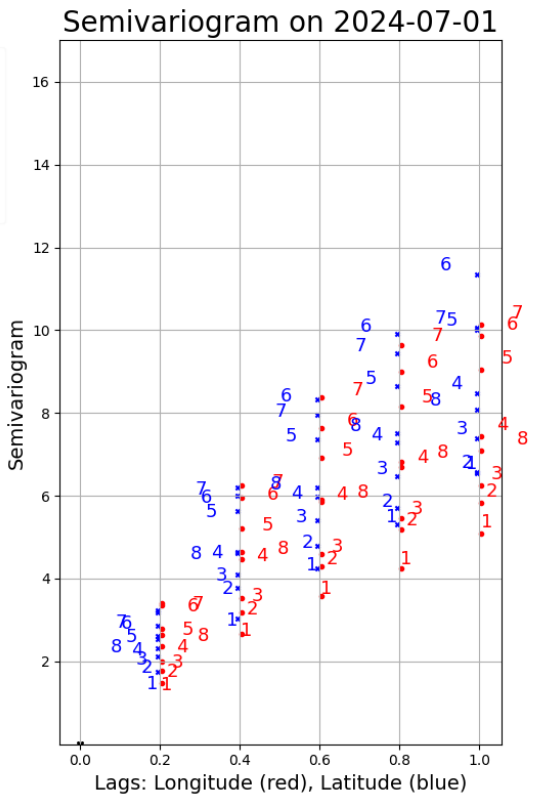


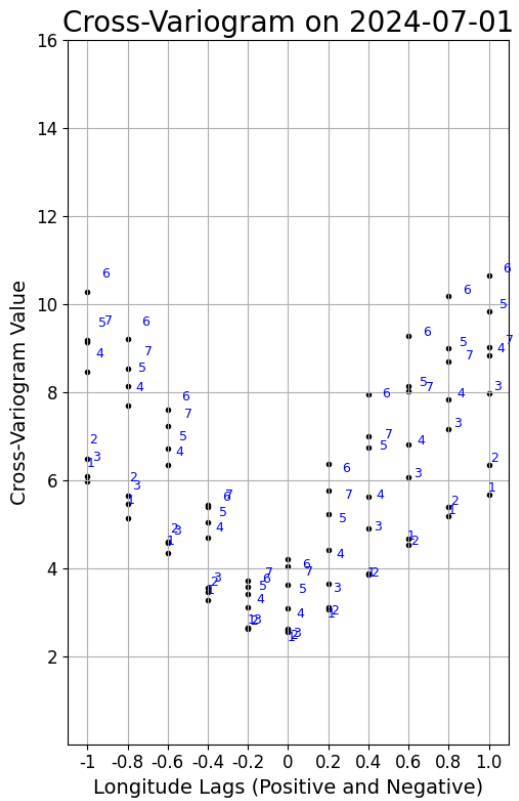
3. Symmetry in Cross Variogram (Space and Time Interaction)

Considering Longitude Lag for Cross-Variogram

Nearly symmetric on these days, implying that there is no strong interaction between space and time.



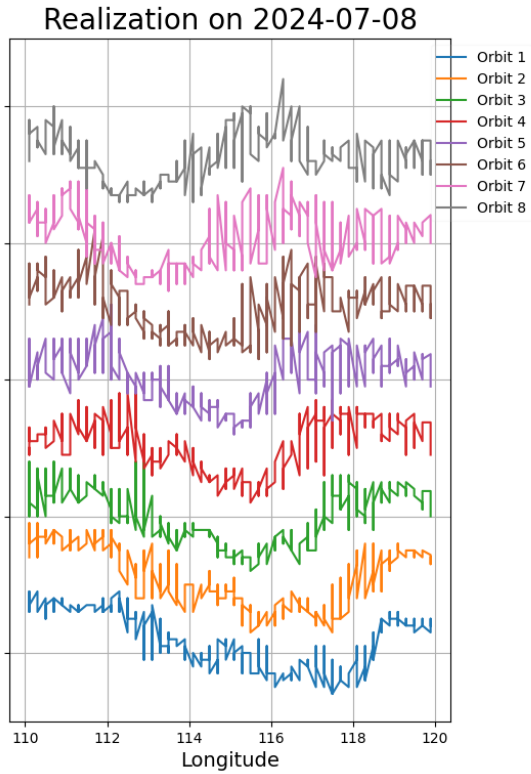


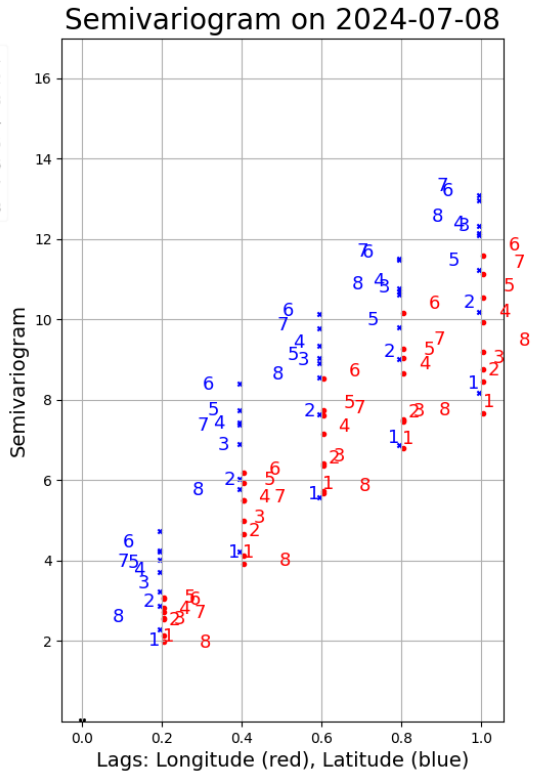


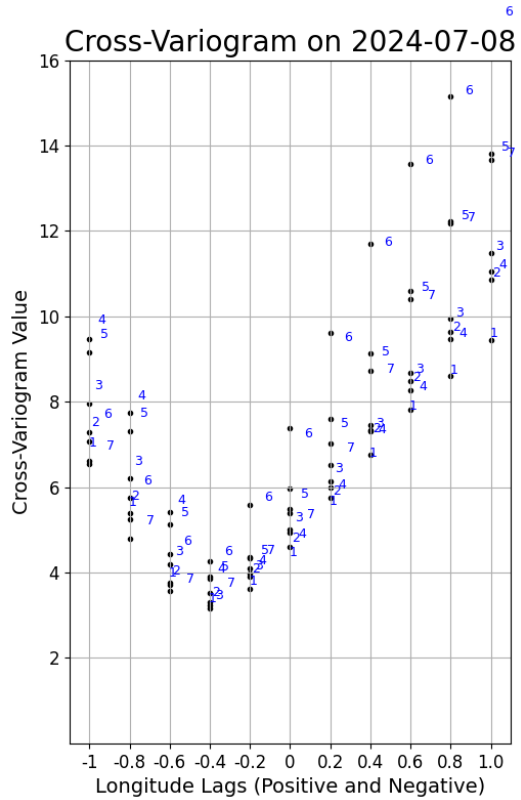
3. Symmetry in Cross Variogram (Space and Time Interaction)

Considering Longitude Lag for Cross-Variogram

Asymmetric, the wind may blow from east to west during this day. In fact, we expect easterly wind near equator.



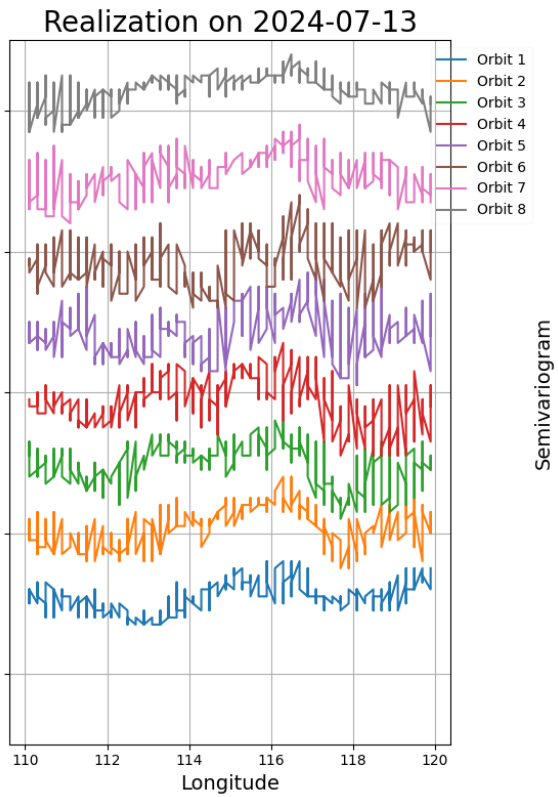


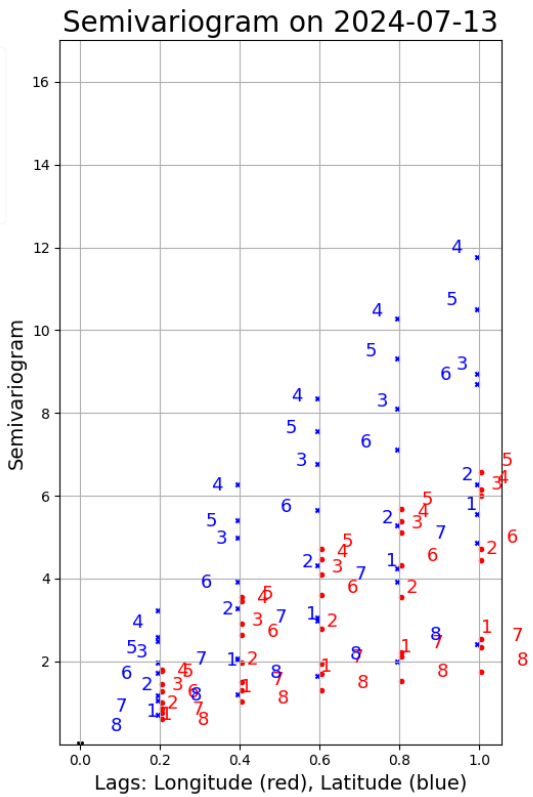


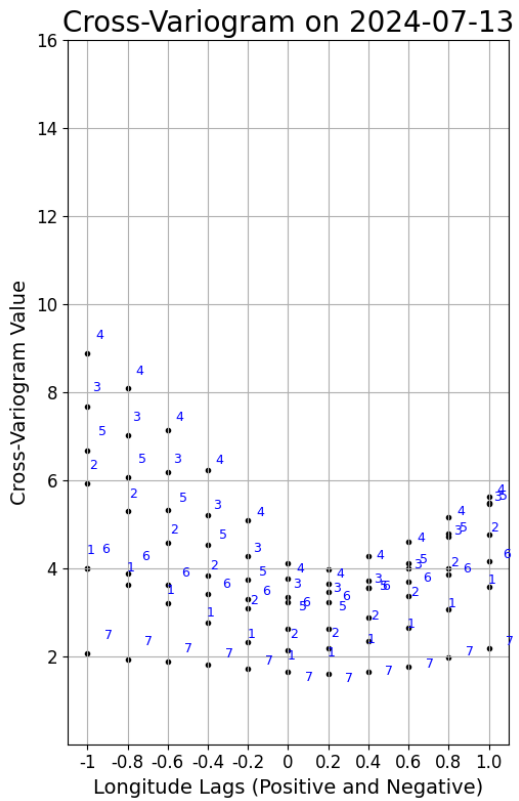
3. Symmetry in Cross Variogram (Space and Time Interaction)

Considering Longitude Lag for Cross-Variogram

Asymmetric, the wind may blow from west to east during this day.



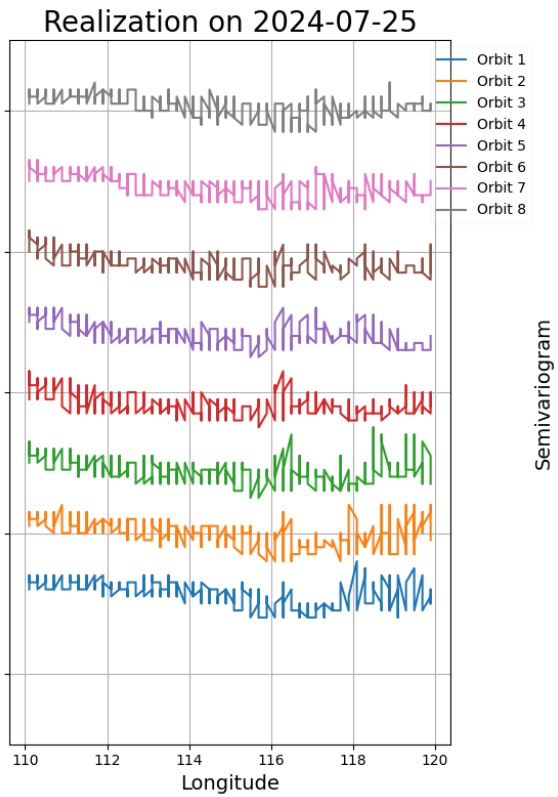


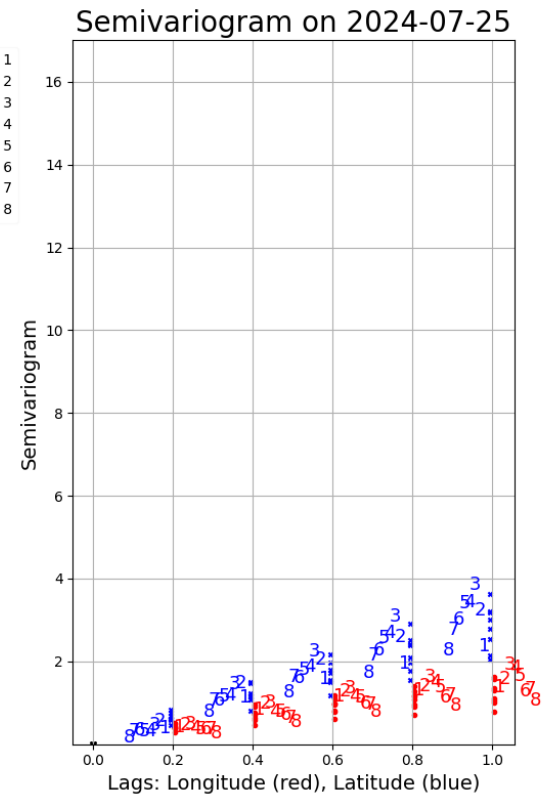


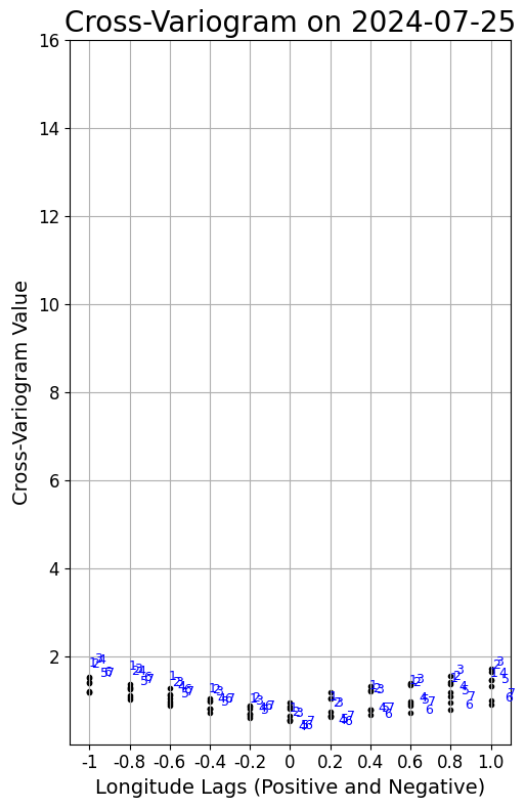
3. Symmetry in Cross Variogram (Space and Time Interaction)

Considering Longitude Lag for Cross-Variogram

Ozone process on July 25th, when it shows short tails.



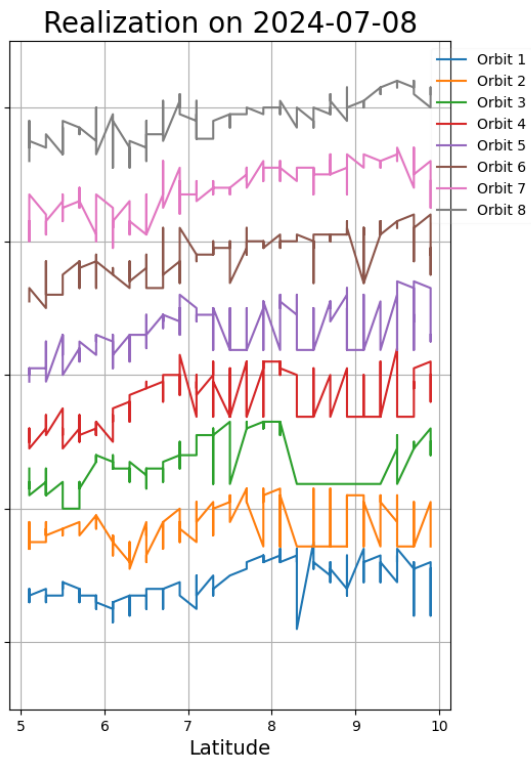


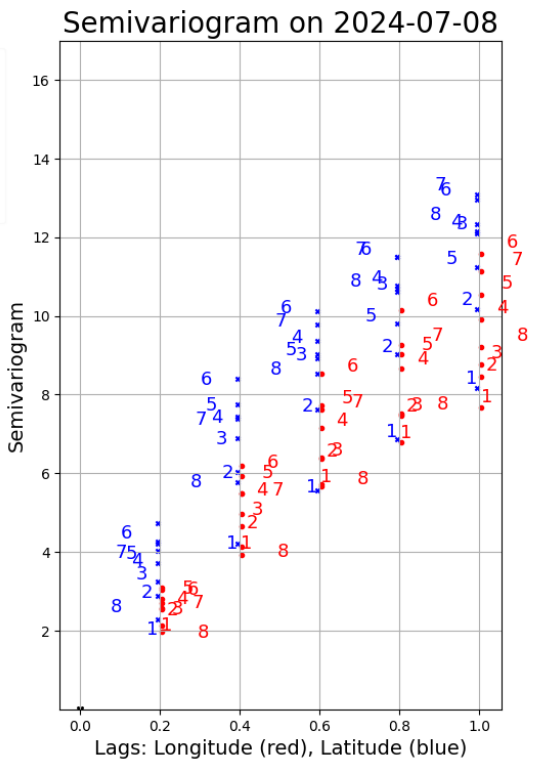


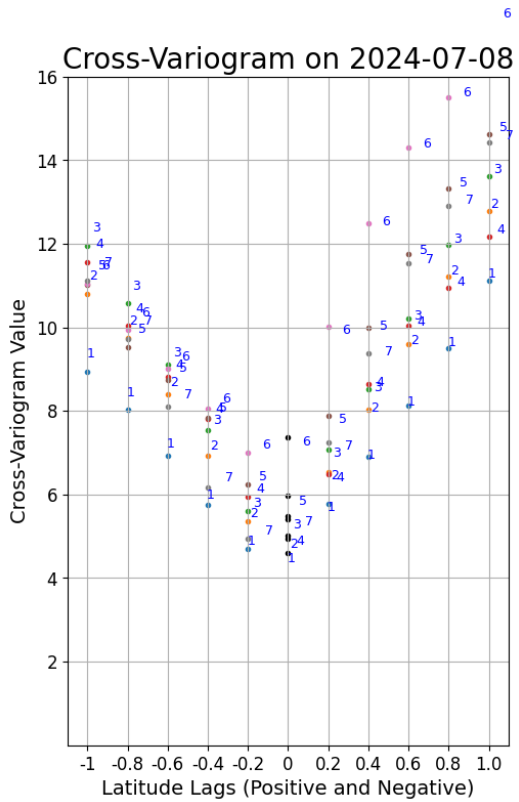
3. Symmetry in Cross Variogram (Space and Time Interaction)

Considering Latitude Lag for Cross-Variogram

Asymmetric on July 8th. The wind may blow from north to south on this day.



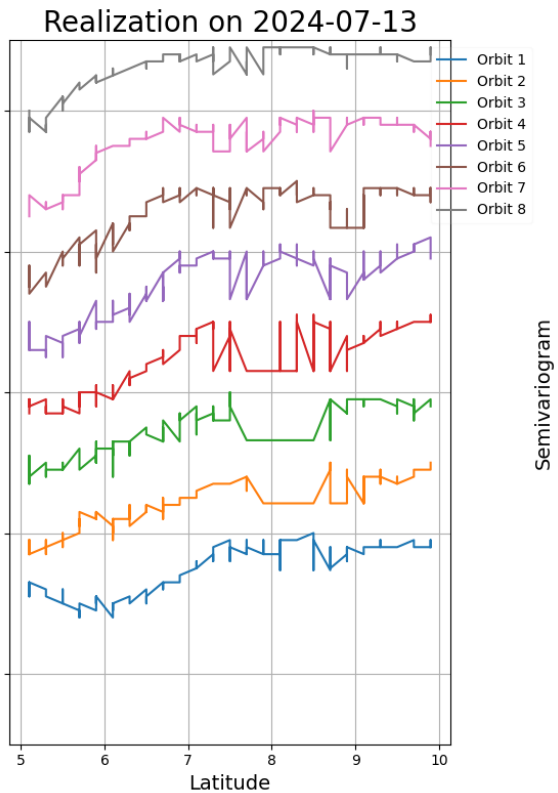


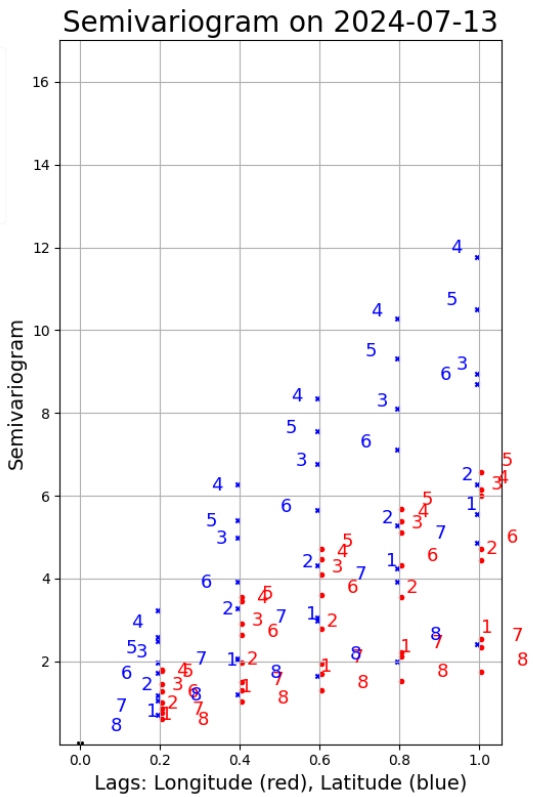


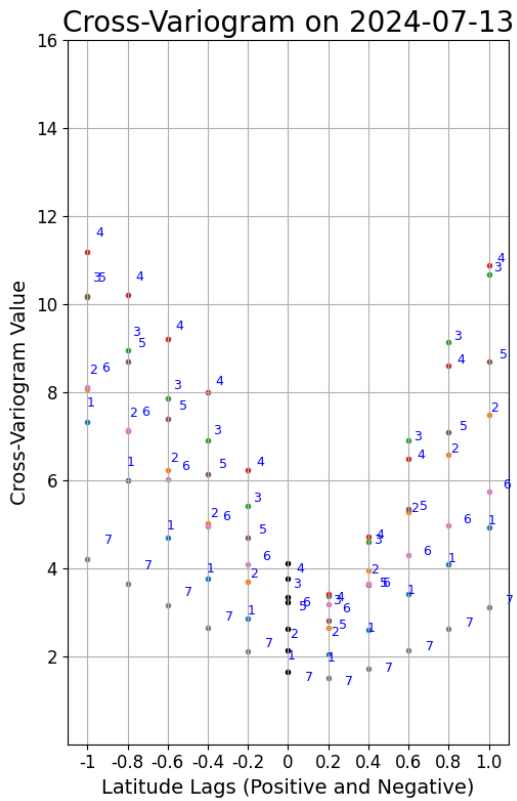
3. Symmetry in Cross Variogram (Space and Time Interaction)

Considering Latitude Lag for Cross-Variogram

Asymmetric on July 13th. The wind may blow from south to north on this day.







4. Smoothness, Local Property

Diagnosing Local Behavior

We propose a diagnostic method showing whether a model captures local behavior of the process. This method investigates how maximum likelihood estimators (MLEs) behave as the sample size varies in a spatial fixed-domain.

4. Why Local Property Has Been Important in Spatial Statistics?

Motivation

- Time series analysis: extrapolation, increasing domains.
- Spatial statistics: interpolation, hence fixed domains are considered.

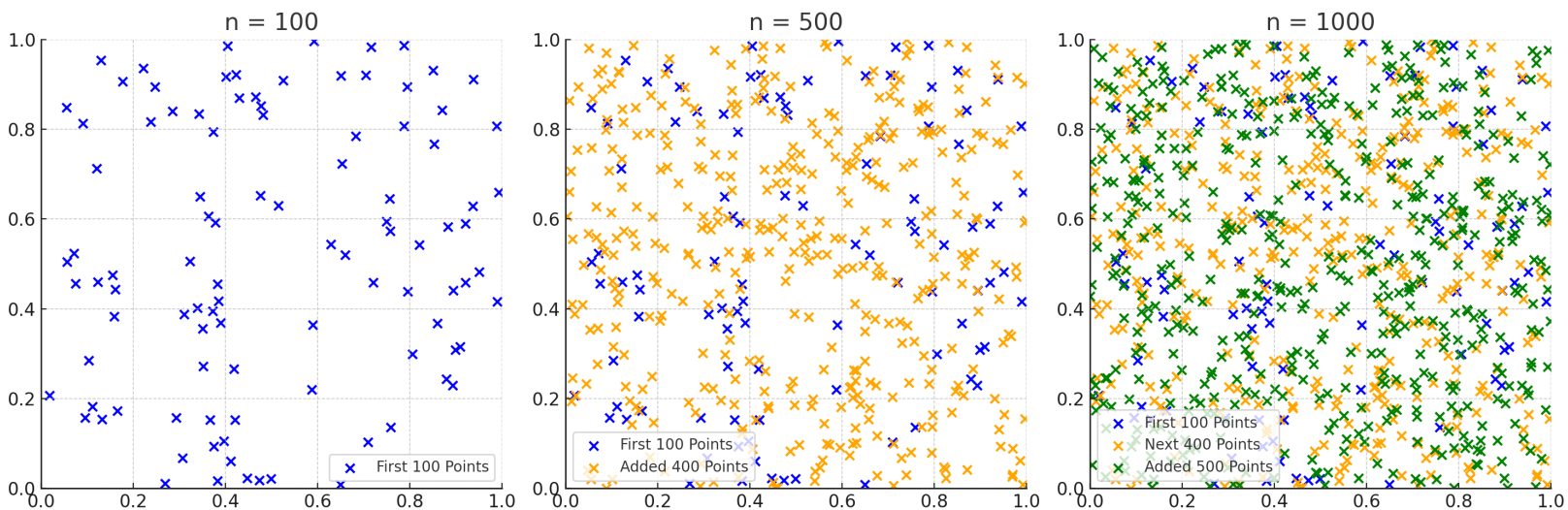
Stein(1999) showed that, in fixed-domain settings, most of the variance of best linear predictors are attributable to a set of high frequencies, which singificantly influences interpolation accuracy.

4. Smoothness, Local Property

Likelihood-Based Diagnostics to Capture Local Behavior

General idea: In a fixed domain, adding more data points provides more information on high-frequency behaviors.

If a model is good enough, especially in terms of capturing local behaviors, we expect maximum likelihood estimators(MLEs) obtained from a dense set not to be dramatically different from those obtained from a coarser set.



Challenge: 1. Size of data and MLEs require $O(n^3)$ for Cholesky decomposition.

Vecchia Approximation: Any joint distribution can be factored into a product of conditionals:

$$p(y) = p(y_1)p(y_2|y_1)p(y_3|y_2, y_1) \cdots p(y_n|y_{n-1}, ..., y_1).$$

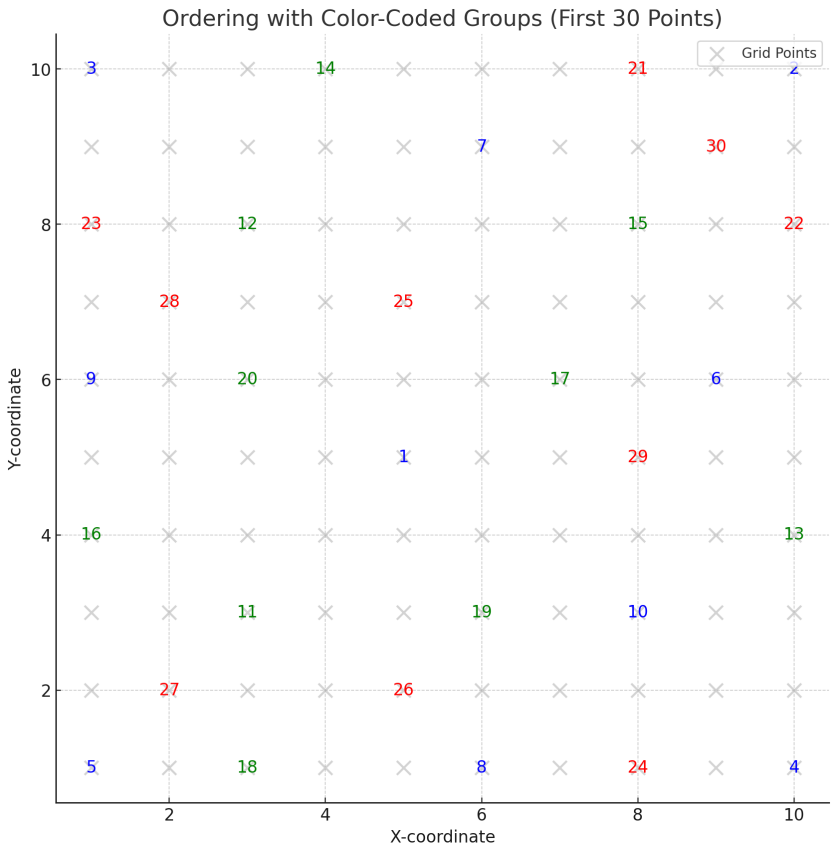
We want to choose a subset $c(n) \in (y_{n-1}, ..., y_1)$ in $p(y_n|y_{n-1}, ..., y_1)$ for approximation.

It is sequentially dependent, but there is no natural ordering of the spatial locations.

4. Smoothness, Local Behavior

Maximum Minimum Ordering and Nearest Neighbors in Space

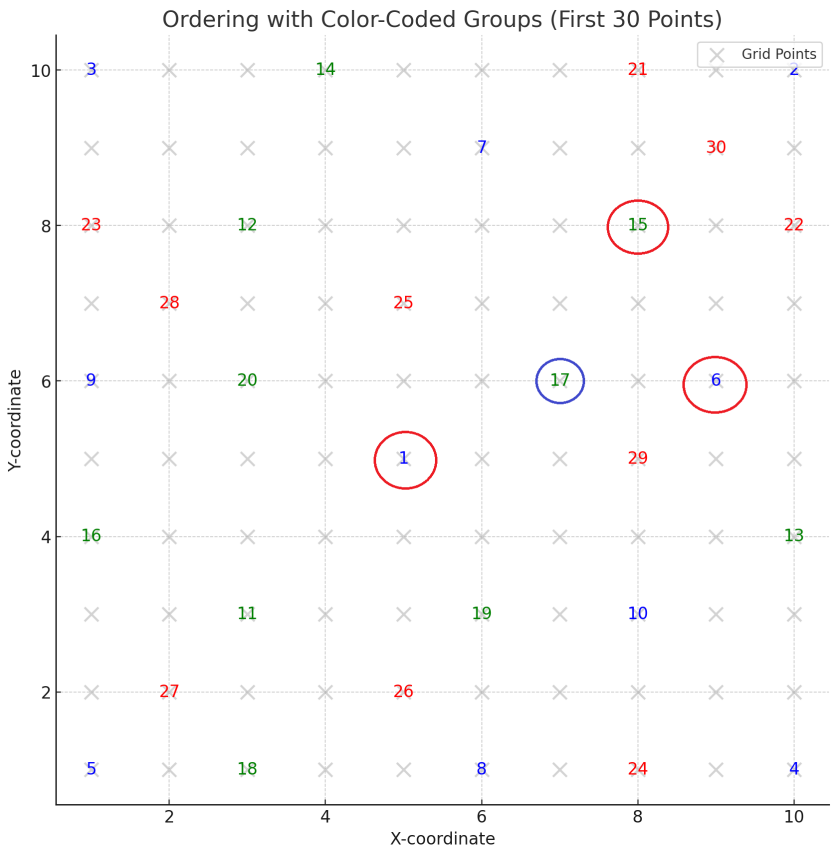
In each iteration, compute the minimum distance from the unselected (unordered) points to the selected (ordered) set. Then, choose the point that maximizes this minimum distance.



4. Smoothness, Local Behavior

Maximum Minimum Ordering and Nearest Neighbors in Space

1. The success of MM ordering in spatial statistics suggest that a diversity of frequency information is important.
2. High-frequency information has more influence than low-frequency information in a fixed domain setting.



Applying Vecchia approximation to Spatio-Temporal Setting

Ordering in Spatio-Temporal Setting is More Complex

Should spatial ordering be prioritized over temporal? For example, assign 1 to 20,000 order for time t, then 20,001 to 40,000 for t+1.

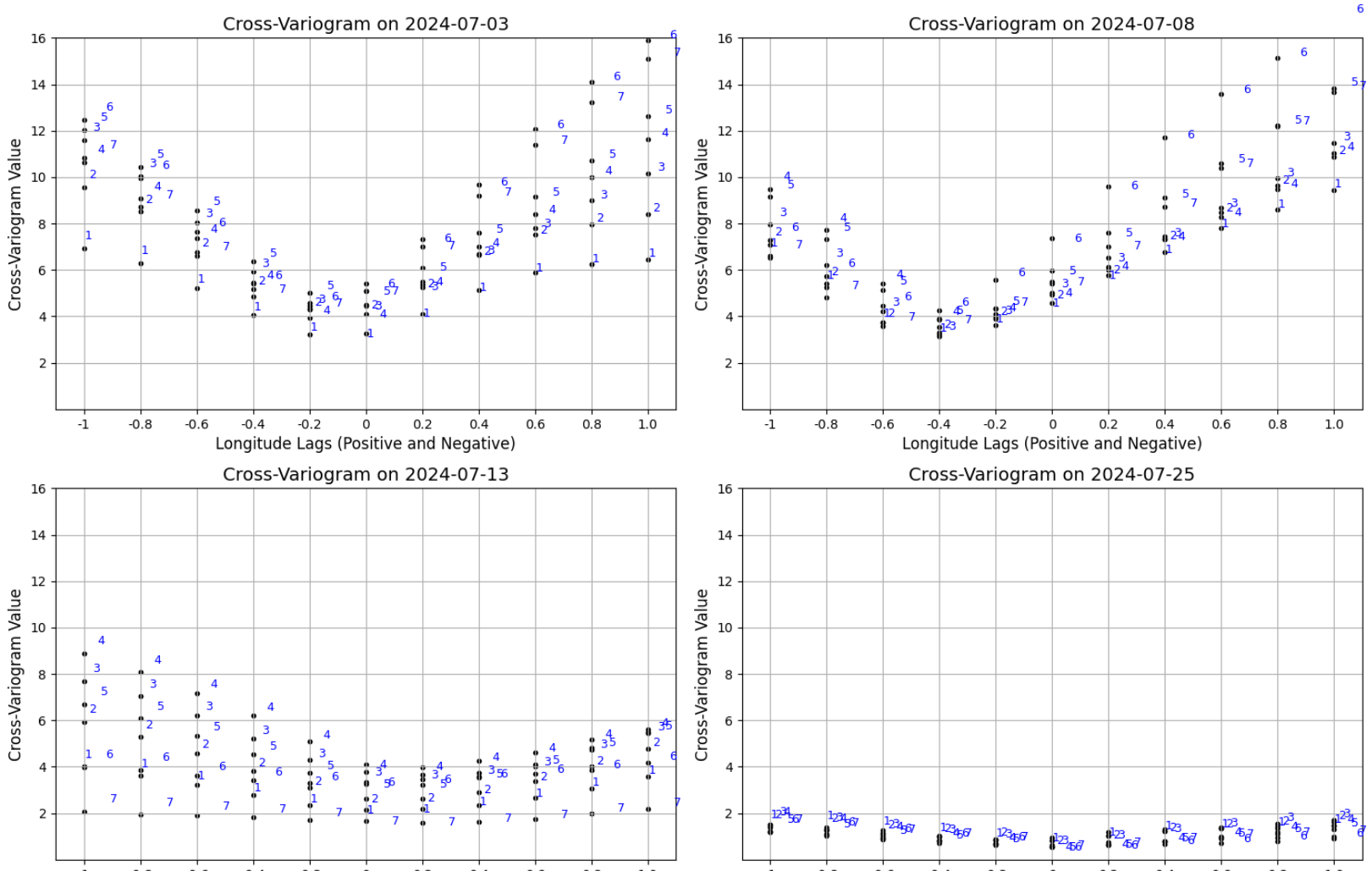
On some days, temporal variation is very small, but not always.

This makes it challenging to determine how data is is correlated across different locations and times.

Examples on next slide:

Changes in Spatio and Temporal Variations

This figure demonstrates how spatio-temporal variations can differ over time, showing the complexity of deciding on ordering strategies. More importantly, models will need to be nonstationary in time at least.



Spatial Modeling vs Spatio-Temporal Modeling

Does Spatio-temporal modeling inherit important properties of spatial models?

Purely spatial models:

- Interpolation
- Importance on local behavior
- Fixed domain properties
- Trade-off between scales: Focus on local behaviors, at the cost of some long-term information.

Spatio-temporal models:

- (Extrapolation property) Unlike purely spatial models, Spatio-temporal model may aim to predict the future, that is, outside the temporal domain.

Spatio-Temporal Modeling: Other Concerns

Spatio-temporal models may not be a simple extension of spatial models- they might possess distinct properties. To address these differences, we many need to include more mid to low frequency information, even at the expense of some high frequency details.

Spatio-Temporal Modeling: Other Concerns

Benefits:

- Averaging ozone values over 4 to 8 orbits in a day can help reduce short-term variations caused by temporary redistributions of ozone with no net gain or loss in a day.
- Selecting sparse locations enables us to expand the spatial scale (current region is under-dispersed with short tails) and to include data upto 1.5 year or 2 year, which is not possible with the current level of spatial density.