PSTAT 234 Final Project Report

Ivan Li

Introduction and Overview

Paper Summary

The paper introduces an original statistical framework to analyze dynamic relationships in multivariate functional data, proposing a Bayesian approach to infer time-varying conditional dependencies between functional data that incorporates changepoints. They take this methodology and apply it to analyze sea surface temperature (SST) data, as well as simulated functional graphical data in order to model connectivity patterns that change over time.

In practice, graphical models are commonly used to find dependencies among random variables, but traditional approaches assume static relationships, which can be problematic for time dependent processes. The authors extend functional graphical models (FGMs) by integrating Bayesian inference and changepoints, which allow for the identification of both time-invariant and time-varying connectivity structures.

The proposed model, which they call the “Dynamic Bayesian Functional Graphical Model”, represents functional data using basis expansions, such as B-splines and Fourier basis functions, that convert continuous functions into a finite and discrete representation. Conditional dependencies are estimated using a sparse precision matrix estimation, which keeps only the most prominent edges in the inferred graphical structure. A changepoint is also incorporated to allow for shifts in the graphical structure over time in an attempt to extend the analysis to the modeling of dynamic data. A block-structured spike-and-slab prior is employed for sparsity, which improves interpretability by grouping related variables. Markov Chain Monte Carlo (MCMC) methods are used for posterior sampling and graph estimation.

The expansion of each function can be written as

$$X_j(t) = \sum_{k=1}^{K} c_{j,k} f_k(t) + \epsilon_j(t),$$

where $f_k(t)$ are basis functions, $c_{j,k}$ are coefficients, and $\epsilon_j(t)$ represents random noise. The precision matrix Ω captures conditional dependencies with a block-wise sparsity prior written as

$$p(\Omega|G, v_0, v_1, \lambda) \propto \prod N(\omega_{ij}|0, v_{g_{ij}}^2) \cdot \prod \mathrm{Exp}(\omega_{ii}|\lambda/2),$$

where G is a latent adjacency matrix, having elements G_{ij} which indicate whether an edge exists between nodes i and j. The prior allows elements of Ω to be close to zero when $G_{ij} = 0$, and when $G_{ij} = 1$, values can be nonzero, which results in sparsity. A uniform prior is placed on a defined changepoint τ, allowing for segmentation into different graphical structures.

The model parameters are estimated using Gibbs sampling, where the simulation algorithm iteratively updates the graphical structure and the precision matrix. Given the current estimate of precision matrix Ω, the edge indicators G_{ij} are sampled using a Bernoulli posterior distribution written as

$$P(G_{ij} = 1|\Omega, \pi_{ij}) = \frac{\pi_{ij}\mathcal{N}(\omega_{ij}|0, v_1)}{\pi_{ij}\mathcal{N}(\omega_{ij}|0, v_1) + (1 - \pi_{ij})\mathcal{N}(\omega_{ij}|0, v_0)}$$

where v_0 and v_1 control the sparsity of edges in the inferred graphical structure. The precision matrix is then updated given the sampled graph structure, following a Gaussian conditional posterior written as

$$\Omega|G, Y \sim \mathcal{N}(\mu_\Omega, \Sigma_\Omega)$$

This ensures that the inferred precision matrix is positive definite. The posterior distribution of τ is computed over a discrete set of possible values, using a likelihood-weighted approach written as

$$P(\tau|Y, \Omega) \propto P(Y|\tau, \Omega)P(\tau)$$

where $P(Y|\tau, \Omega)$ represents the likelihood of the data under a changepoint segmentation, and $P(\tau)$ is assumed to be uniform over possible time points. The basis expansion coefficients are updated using a Normal conditional posterior, and the errors are sampled from an Inverse-Gamma distribution. This Gibbs sampling algorithm is repeated until convergence, and this produces posterior distributions over graph structures, changepoints, and functional relationships.

Model performances used in the paper are quantified using simulation studies and real-world SST data. In simulations, the authors generated data from a dynamic FGM with $p = 15$ functions, $n = 50$ replicates, and a true change-point at $t = 129$, representing the days after March in observed SST data in part of the Atlantic ocean. The model is compared against frequentist and Bayesian alternatives, such as the partially separable graphical model (PSFGM), Bayesian Lasso Functional Graphical Model (BL-FGM), and the Bayesian Gaussian Graphical Model (BGGM). They demonstrate that the proposed model, on average, shows superior true positive rates (TPR), false positive rates (FPR), and Matthews correlation coefficients (MCC), where TPR measures the proportion of correctly identified edges among true edges, FPR measures the proportion of falsely identified edges among non-existent edges, and MCC provides an overall edge classification success measure that can be written as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

This study shows the proposed model’s use cases in climatolgy, with a framework that allows for application to other related fields such as finance.

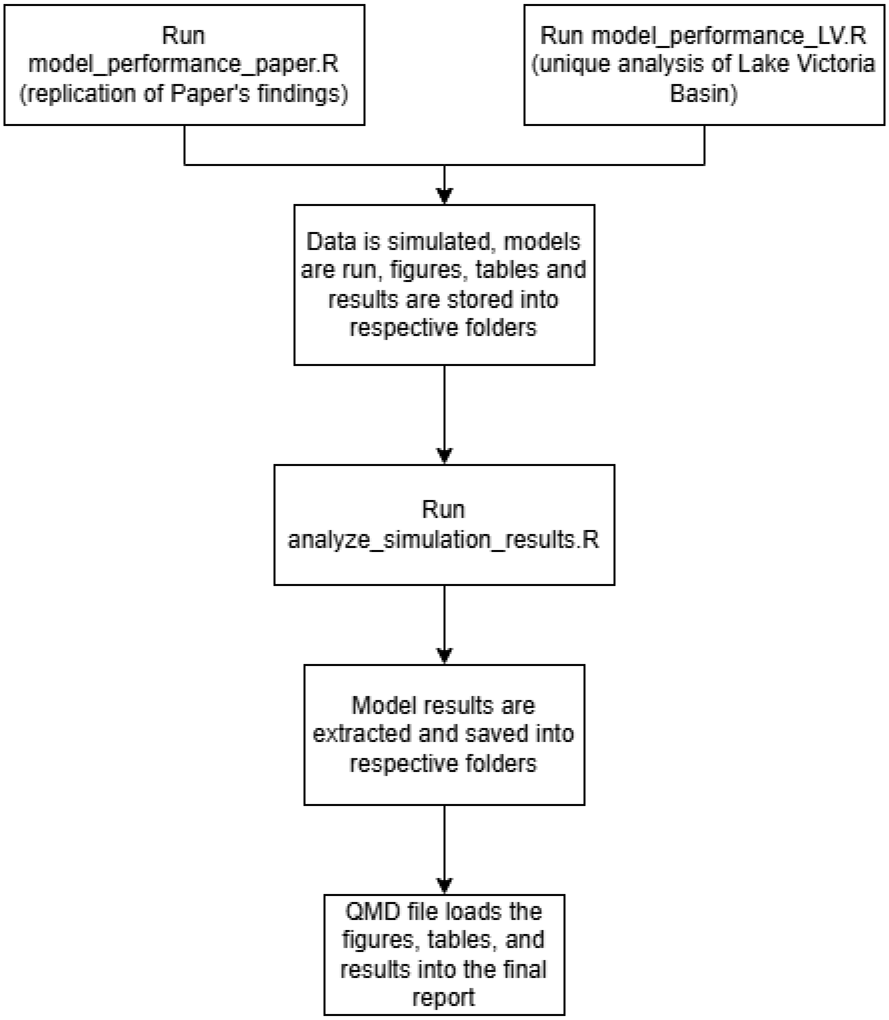
Report Goals

This report aims to evaluate and extend the methodology proposed in the referenced paper by replicating its findings, reproducing key figures, and conducting a unique data analysis scenario on simulated sea surface temperature (SST) data, as the original analysis employed a general simulation framework with a defined changepoint, but lacked a detailed simulation scenario. To address this, I will apply the paper’s methodology to simulated SST patterns representative of Lake Victoria Basin, a region of interest due to its seasonal variability and implications for food security. The extension of the analysis will focus on computational performance by conducting and documenting runtime on a Jetstream2 m3.medium instance.

Considerations

Reproduction Flowchart

The flowchart below shows how to reproduce the figures and findings of this report. However, some figures are reproduced separately with scripts in the scripts folder of the GitHub repository.


Run model_performance_paper.R (replication of Paper's findings)
Run model_performance_LV.R (unique analysis of Lake Victoria Basin)
Data is simulated, models are run, figures, tables and results are stored into respective folders
Run analyze_simulation_results.R
Model results are extracted and saved into respective folders
QMD file loads the figures, tables, and results into the final report

The authors of the paper maintain a publicly available GitHub repository that provides various functions supporting their methodology, including data simulation and Markov Chain Monte Carlo (MCMC) algorithms. This report utilizes select functions from that repository.

This report will analyze three different models; the proposed model (DBFGM), PSFGM, and BGGM. This is done for computational efficiency, as the other model that the paper explored (BLFGM) would take about 6 hours to run.

Regarding the GitHub repository that stores the analyses and reproductions described in this report, the RData objects generated during performance evaluation of the proposed model are approximately 200 MB each, which exceed the file size limit of GitHub. This is due to the computational nature of the proposed model, so these data files are not included in the repository but can be generated by running the *model_performance_paper.R* and *model_performance_simulation.R* scripts. However, executing these scripts requires significant computational resources, taking approximately three hours on an *m3.medium* Jetstream2 instance to run. The number of repetitions (*nreps*) can be adjusted to a lower value for each respective model to reduce computational time. The performance evaluation results presented in the report, including tabular and visual summaries, were generated from these scripts, stored in RData objects using the *analyze_simulation_results.R* script, and hosted on GitHub for convenient reproduction of this report. Similarly, the observed SST dataset analyzed in the paper (Figure 3) exceeds 50 GB and cannot be hosted on GitHub.

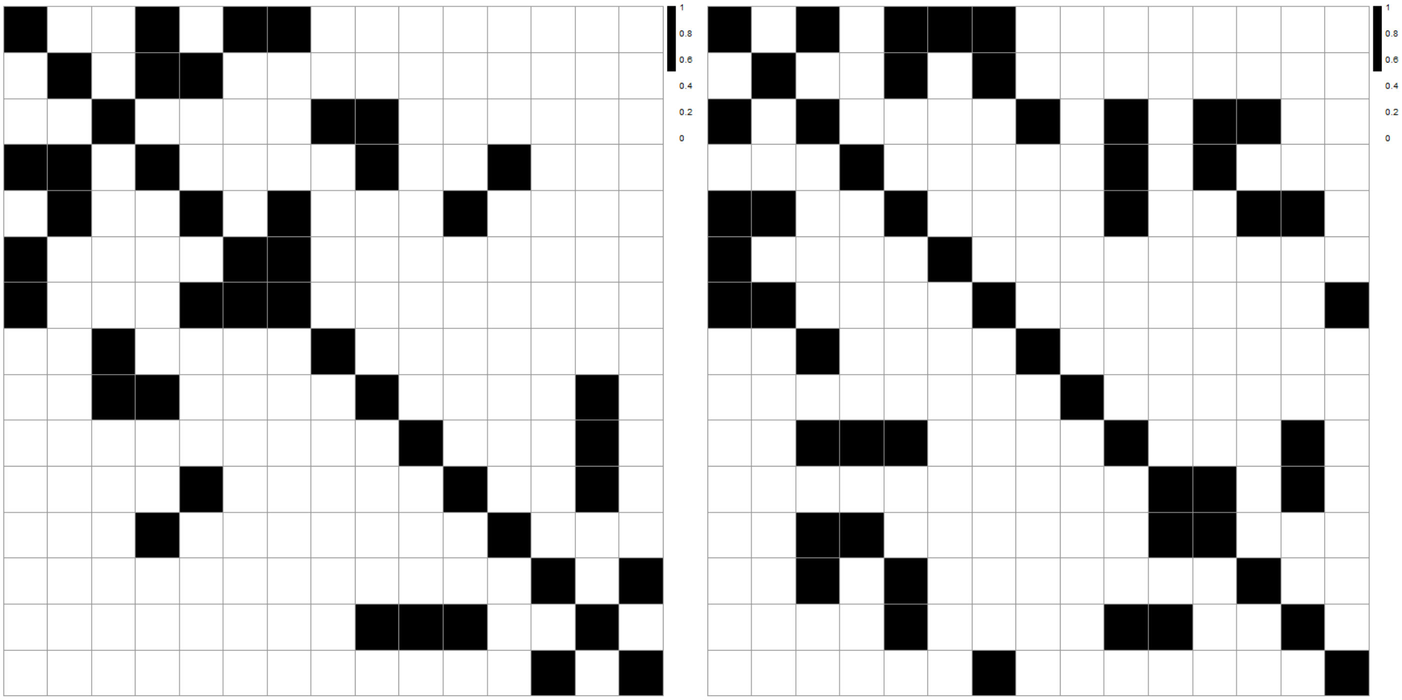
Re-implementation of Paper

This section aims to replicate the findings and figures presented in the referenced paper by re-implementing their methodology, as well as provide additional data visualizations not included in the original paper. Below is a general outline describing what will be replicated from the paper.

1. Generate simulated data with a defined constant changepoint that includes random graphical connections between functions in the simulated data, and show figures representing the graphical and functional relationship of that simulated data (Figure 2 from the paper)
2. Visualize observed real-world SST data (Figure 3 from the paper)
3. Run PSFGM, DBFGM, and BGGM on 50 replicates of simulated data to estimate graphical relationships
4. Analyze and visualize performance of all models using TPR, FPR, and MCC

Figure 2 reproduction

Initially, the authors simulate a functional graphical dataset with 15 functions, and a constant changepoint of τ = 129. After simulating this graphical data, the paper visualizes the graphical relationships before and after the changepoint using figures similar to the ones shown below.

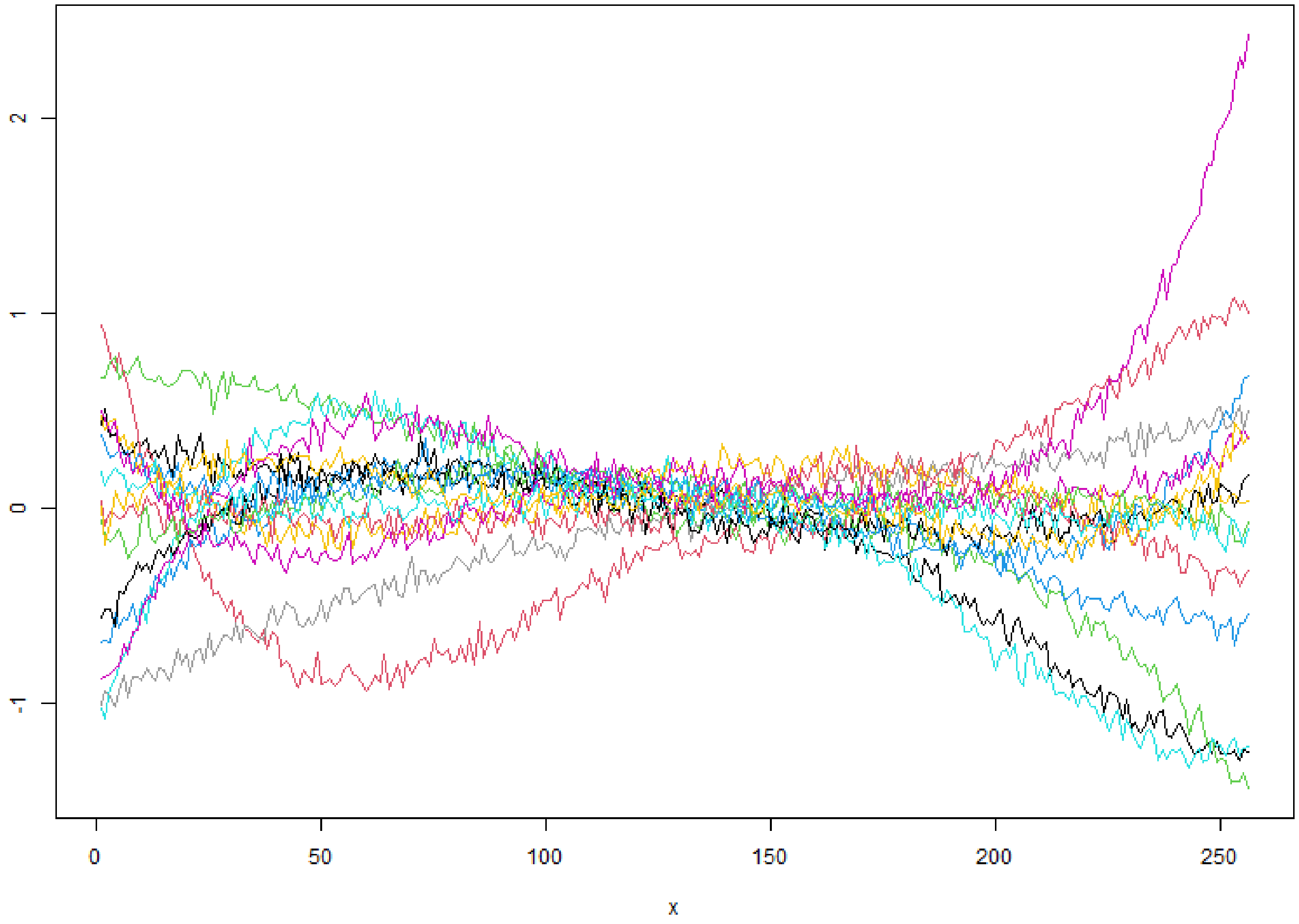


These figures show the graphical relationship between the simulated data before (left) and after (right) the changepoint.

Each square in the plot represents the relationship between two functions. The top-left square of the left figure, for instance, illustrates the relationship between Function 1 and itself. It is colored black, indicating a connection, as a function is inherently related to itself. Similarly, the square positioned three places to the right of the top-left represents the relationship between Function 1 and Function 4. The black coloring indicates a graphical relationship between these two functions. Conversely, white squares indicate no connection, where those corresponding functions do not exhibit any graphical relationship.

The replication of the figures differs in appearance from those shown in the original paper. This is likely due to the authors setting a different seed during their analysis, which remains unknown. However, this discrepancy does not affect the overall results of this analysis reproduction, as the simulation done in the paper and in the replication here accounts for a wide range of random graphical scenarios, which provides a sufficient number of samples to accurately assess the model, and the figures above represent just one of the simulated datasets and their graphical representation.

To provide further insight into how this simulated graphical data is structured, a time series visualization was plotted. This figure was not included in the original paper.

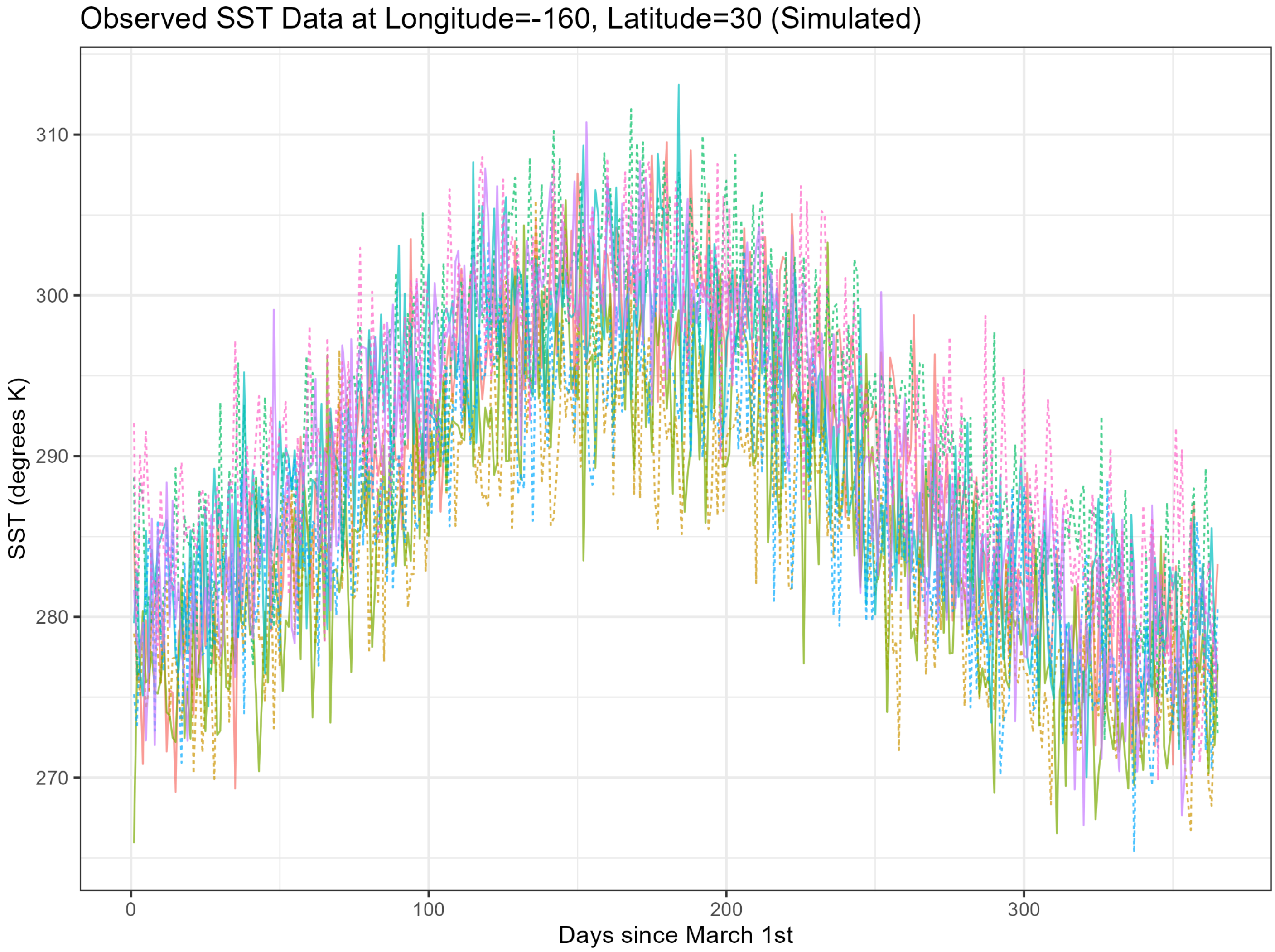


This shows the general functional graphical relationship over time, which in theory, can be extended to any data exhibiting such characteristics. On the x-axis, a changepoint of 129 is observed, and we see that each function transitions over time to a different state compared to before the changepoint. As seen in Figure 2, this transition due to the defined changepoint also alters the graphical relationships between the functions.

Figure 3 reproduction

Before conducting detailed simulation studies, the paper visualizes real-world SST data spanning from 1979 to 2022. Due to the original dataset’s large size (over 50 GB), I will instead simulate the observed data, with a changepoint at τ = 129 to replicate the seasonality trend shown in the paper for a specific region in the Atlantic Ocean.

The original dataset used in the paper is from the ERA5 reanalysis, and can be accessed via the Climate Data Store (CDS) at this link (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=download).



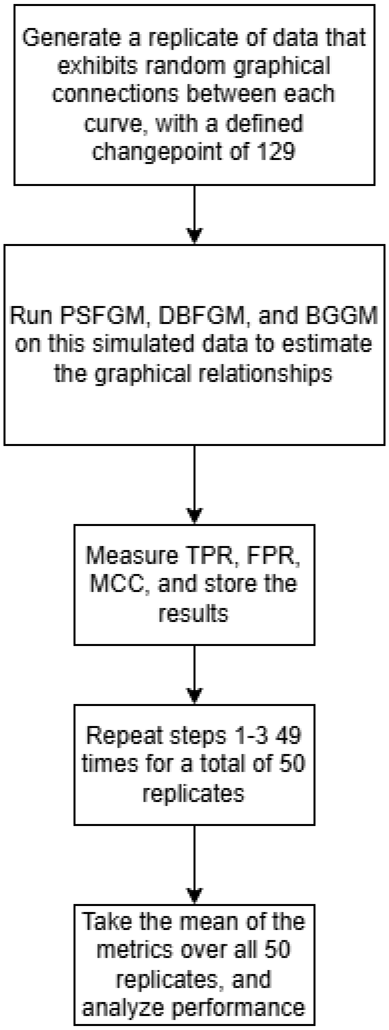
The plot shows SST in Kelvin over time for a specific region in the Atlantic Ocean, with each curve representing a different year of data. A noticeable spike occurs around August, which is a recurring seasonal pattern characteristic of this region.

Findings

Here, the numerical simulations done by the paper will be replicated and analyzed.

Analysis Flowchart

Below is a flowchart outlining the simulation and anlaysis procedures described in the paper. Each replicate of data has the same changepoint, but a varying graphical structure.



Generate a replicate of data that exhibits random graphical connections between each curve, with a defined changepoint of 129

Run PSFGM, DBFGM, and BGGM on this simulated data to estimate the graphical relationships

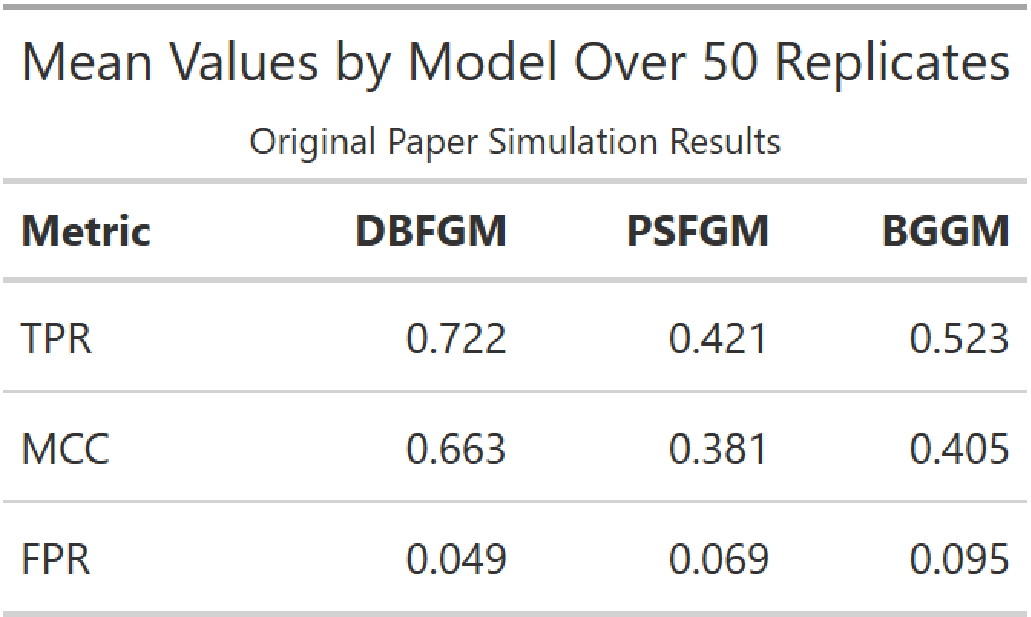
Measure TPR, FPR, MCC, and store the results

Repeat steps 1-3 49 times for a total of 50 replicates

Take the mean of the metrics over all 50 replicates, and analyze performance

Results Summary

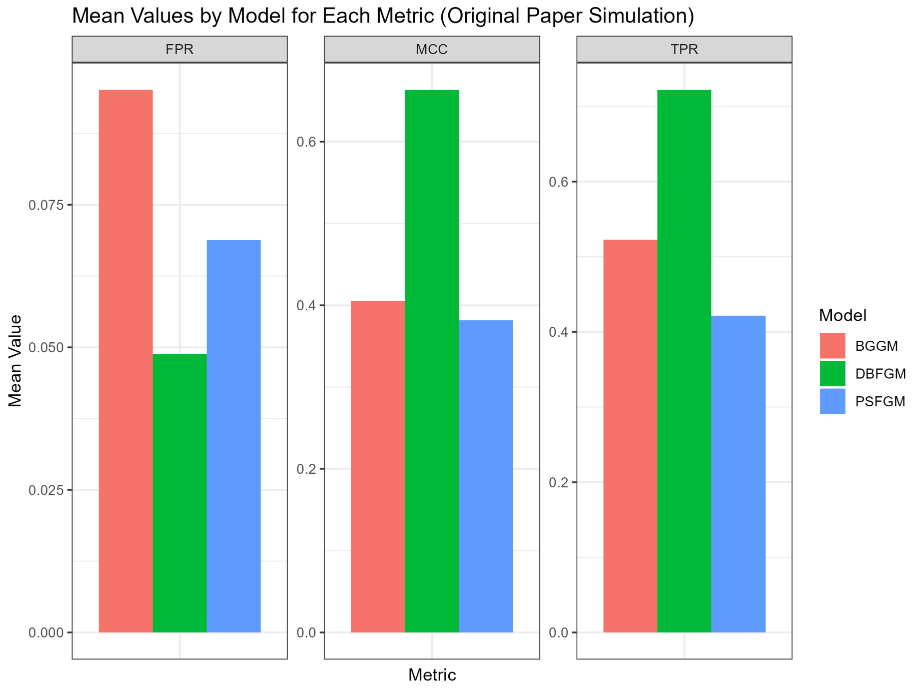
Here, the average results of all 50 replicates of each model are compiled into tabular and graphical summaries.



Mean Values by Model Over 50 Replicates

Original Paper Simulation Results

Metric	DBFGM	PSFGM	BGGM
TPR	0.722	0.421	0.523
MCC	0.663	0.381	0.405
FPR	0.049	0.069	0.095



The results indicate that the proposed model, DBFGM, demonstrates superior performance compared to PSFGM and BGGM. DBFGM has a lower false positive rate while maintaining a higher Matthews correlation coefficient and true positive rate. This shows that DBFGM more effectively identifies true graphical relationships, while keeping misclassifications low. However, it is important to note that these comparisons were conducted within the general simulation framework used in their study, with evaluations limited to a changepoint value of 129. Further simulations and analysis would be necessary to confirm the general application of these findings.

Extension of Numerical Section: Application to Lake Victoria Basin

In this section, the proposed model (DBFGM) will be applied to a specific data analysis scenario to investigate its effectiveness in capturing the seasonality trends of Lake Victoria Basin SST. To achieve this, the original simulation framework will be modified by adjusting the sample size, changepoint value, and functions used in generating the data. This setup aims to better simulate the OND (October–November–December) seasonal variations of Lake Victoria Basin SST, where significant climatic transitions typically occur.

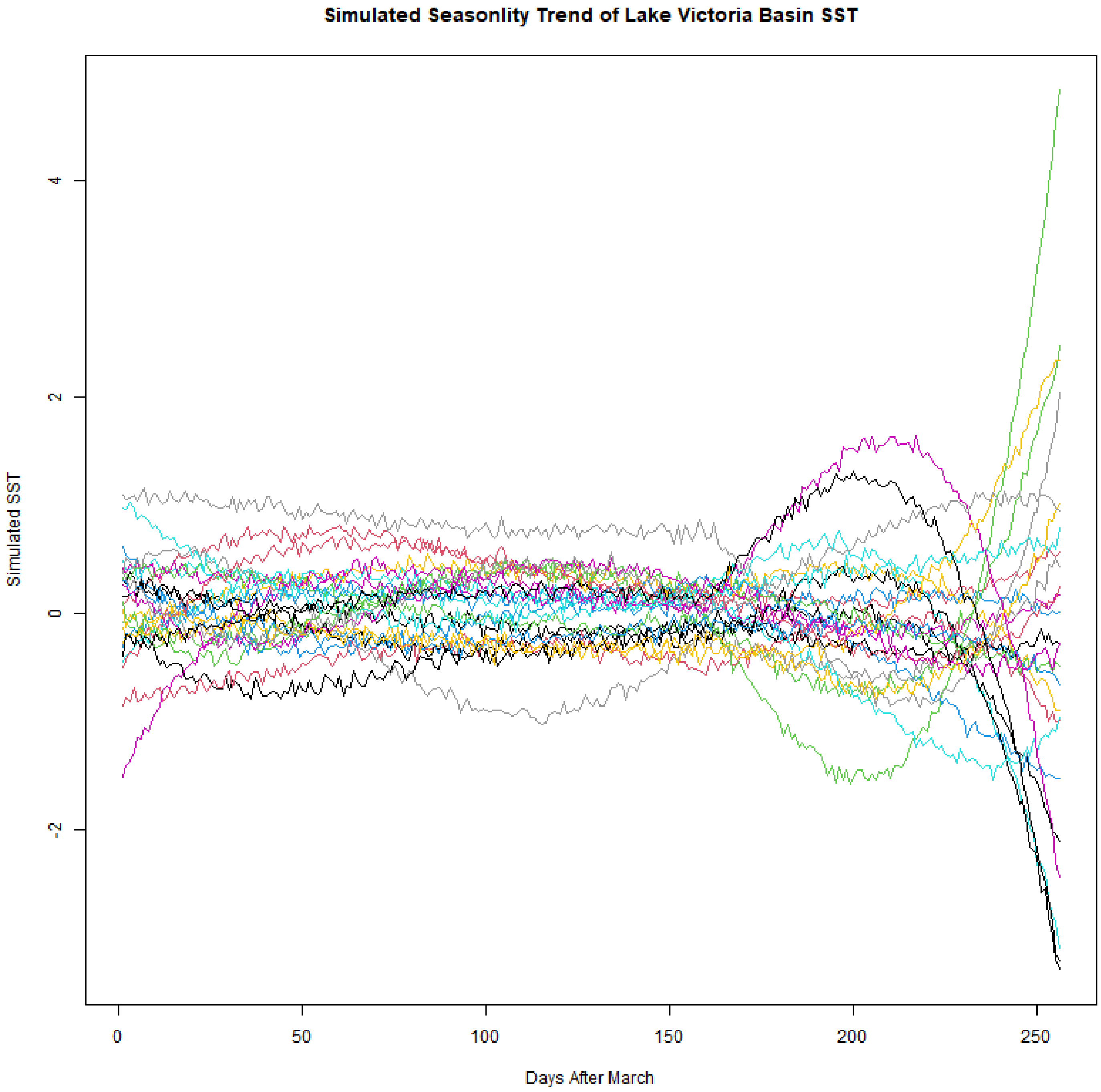
After generating this new simulated dataset, DBFGM, PSFGM, and BGGM will be run on the data to estimate the underlying graphical relationships. The models' performances will then be compared and analyzed to determine the extent to which each method accurately captures the seasonal changepoints and dependencies in the data.

Below is a general outline of the extension analysis

1. Generate a replicate of data with a changepoint at time 160, with random graphical connections between each function of the data to simulate the OND seasonality trends of Lake Victoria Basin's SST.
2. Show figures visualizing one replicate of data.
3. Run PSFGM, DBFGM, and BGGM on 50 replicates of simulated data to estimate graphical relationships
4. Analyze and visualize performance of all models using TPR, FPR, and MCC

Simulated OND Seasonality Plot

This figure illustrates one replicate of the data, incorporating a changepoint at time 160, with an increased sample size and additional functions that capture the OND seasonality trends of Lake Victoria Basin’s sea surface temperature (SST).

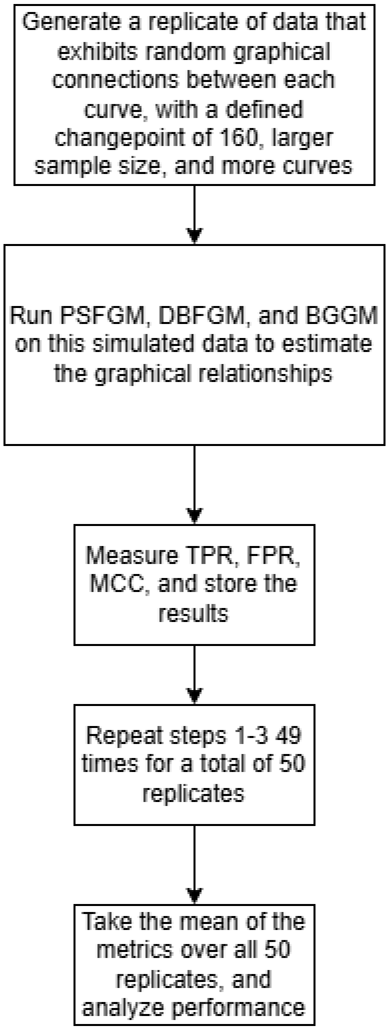


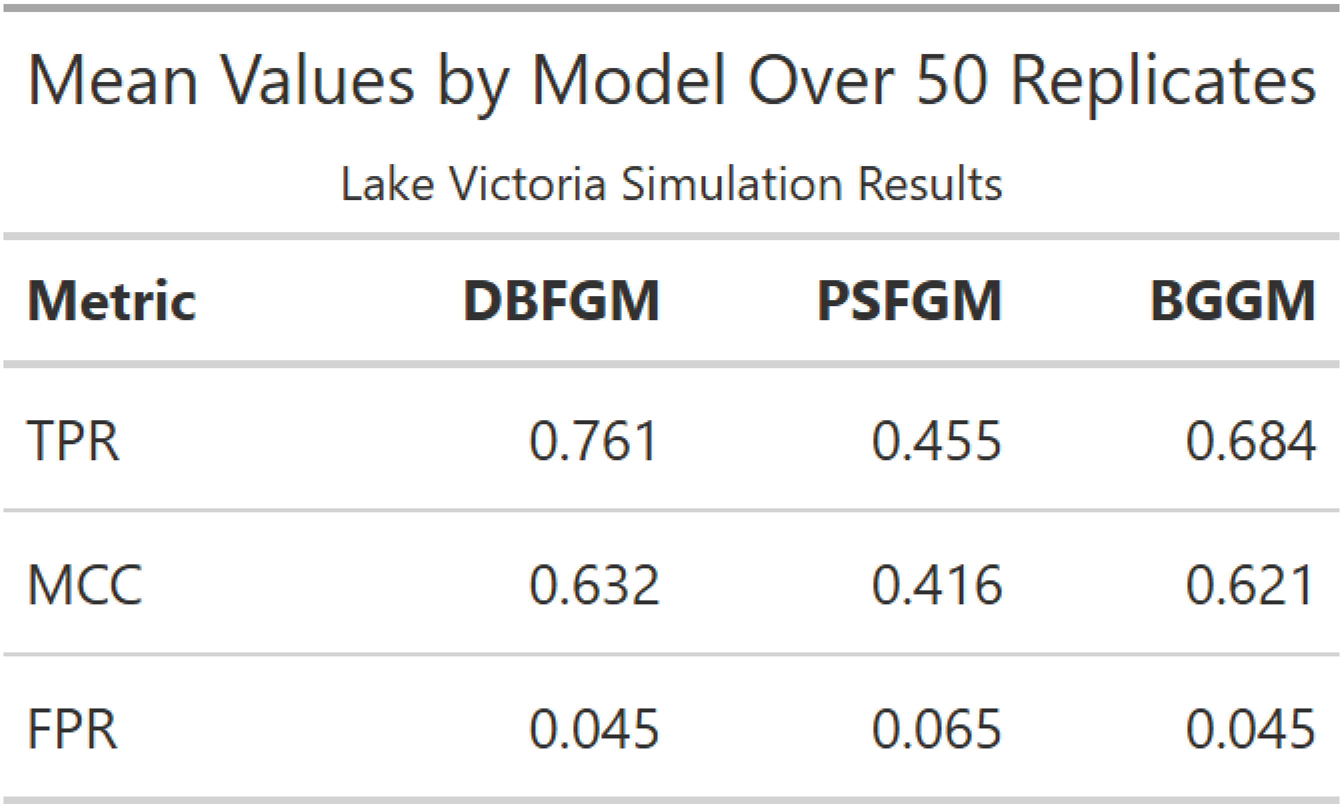
The simulated SST data here reflects both regular seasonal trends and irregular climate events like El Niño and La Niña. El Niño events are cause overall warming trends, while La Niña events show cooling trends. The simulated data account for years indicative of these climate phenomena.

Findings

Analysis Flowchart

Similar to the previous flowchart describing the paper’s analysis, this outlines the simulation and anlaysis procedures taken for the Lake Victoria Basin simulation. Each replicate of data has the same changepoint, but a varying graphical structure with more functions and more samples.


Generate a replicate of data that exhibits random graphical connections between each curve, with a defined changepoint of 160, larger sample size, and more curves
Run PSFGM, DBFGM, and BGGM on this simulated data to estimate the graphical relationships
Measure TPR, FPR, MCC, and store the results
Repeat steps 1-3 49 times for a total of 50 replicates
Take the mean of the metrics over all 50 replicates, and analyze performance

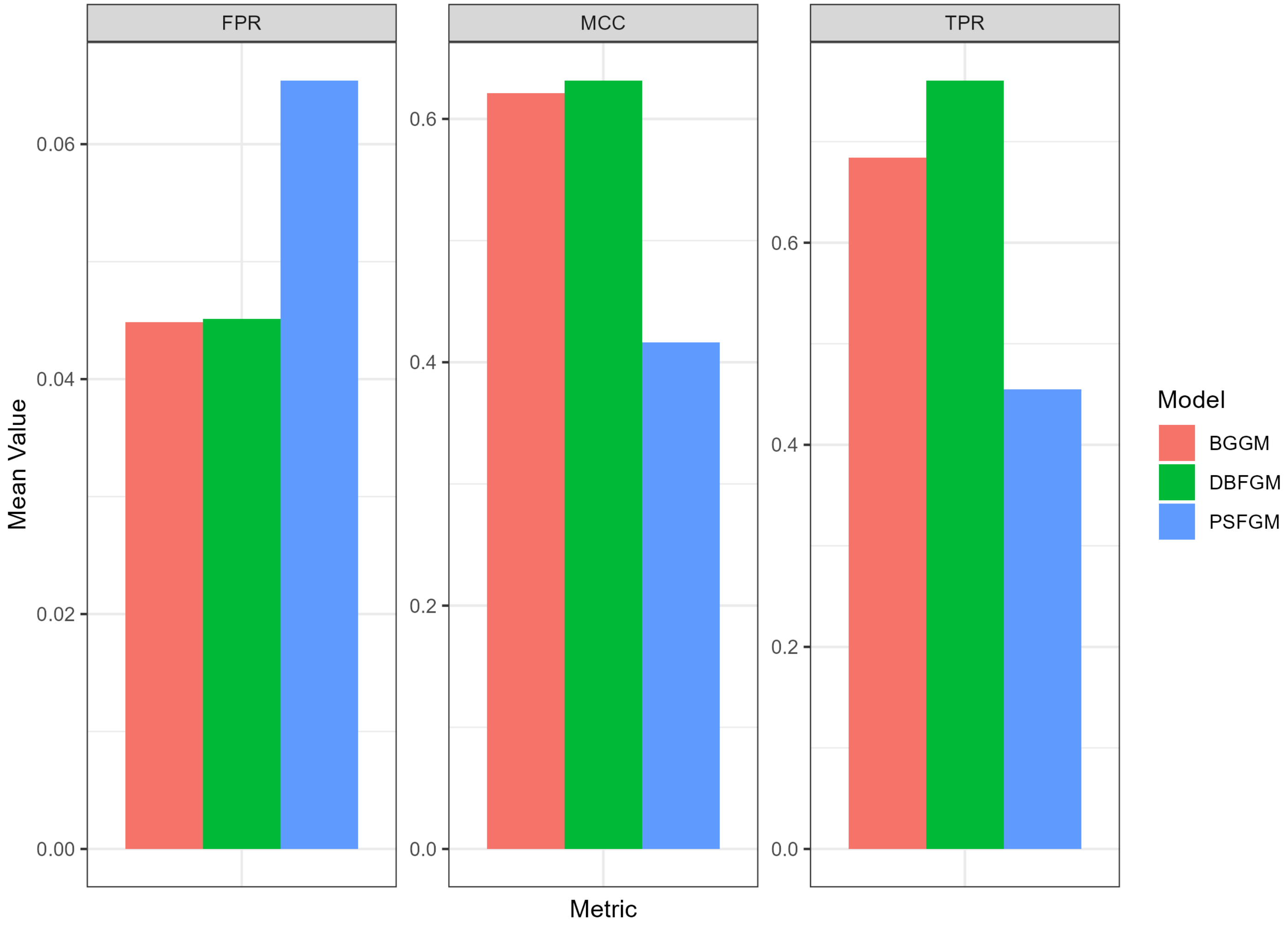


Mean Values by Model Over 50 Replicates

Lake Victoria Simulation Results

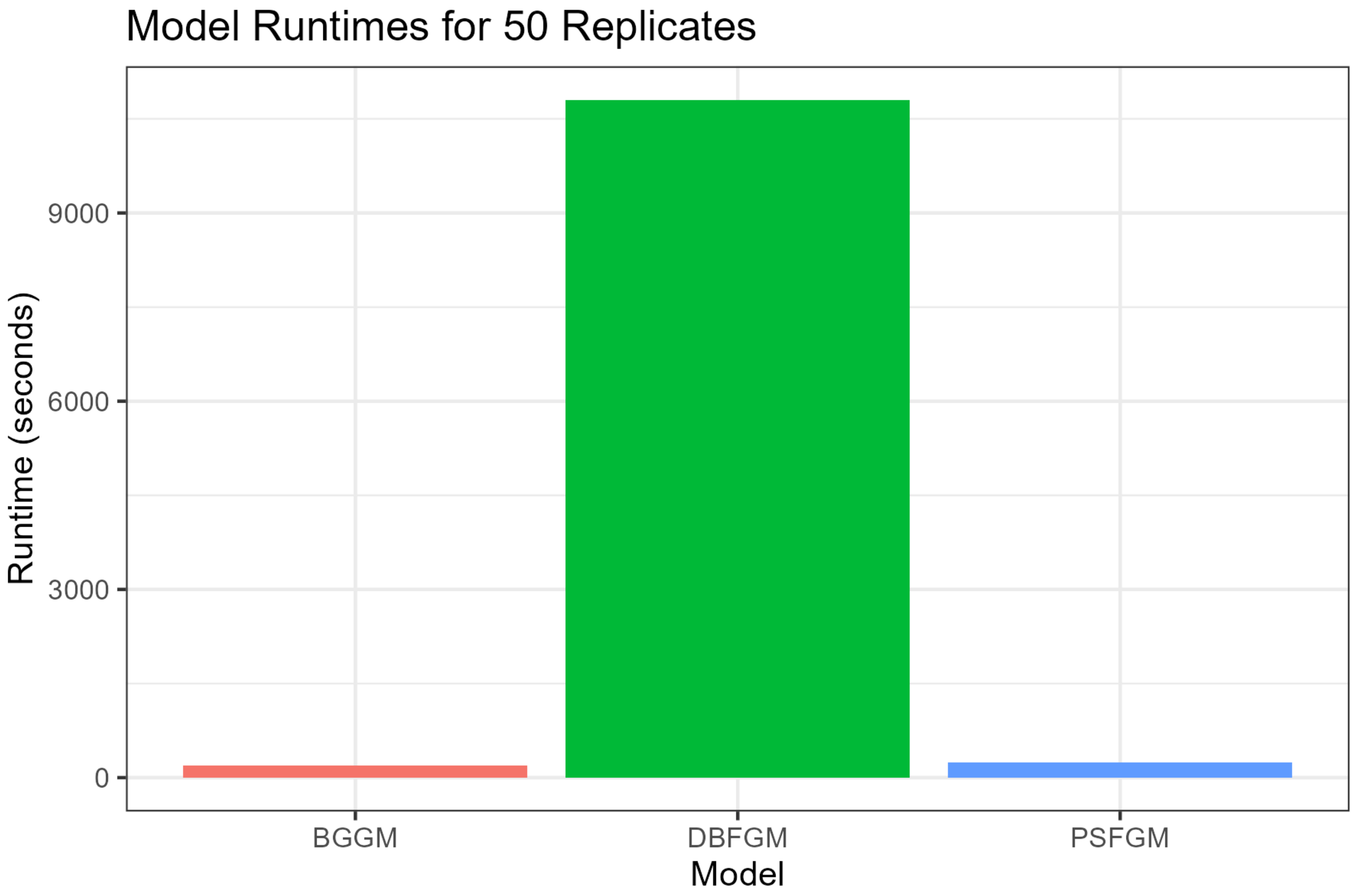
Metric	DBFGM	PSFGM	BGGM
TPR	0.761	0.455	0.684
MCC	0.632	0.416	0.621
FPR	0.045	0.065	0.045

Mean Values by Model for Each Metric (Lake Victoria Basin Simulation)



Here, we see that DBFGM and BGGM have comparable performance across all three metrics, while PSFGM struggles. DBFGM has a slightly higher FPR than BGGM but a slightly higher MCC and TPR. However, this slight improvement in performance is coupled with significantly reduced computational efficiency, as the proposed model is much more computationally intensive than the other two models.

This difference is visualized in the bar chart below.



For both the replication of the paper and the Lake Victoria Basin simulation, BGGM and PSFGM were executed in approximately 3-4 minutes, whereas DBFGM required over 3 hours to complete on a Jetstream2 m3.medium instance. Given the comparable accuracy in the simulated Lake Victoria Basin scenario, the proposed model's benefits do not outweight the costs.

Conclusion

The paper introduces a novel approach to modeling dynamic dependencies in multivariate functional data, proposing the Dynamic Bayesian Functional Graphical Model (DBFGM) that integrates Bayesian inference with changepoints. The DBFGM shows promise in specific scenarios, excelling in situations where the explicit definition of changepoints is present in the data, as demonstrated through its application to simulated SST data.

However, in simulations, the DBFGM was found to require heavy computational resources. In the Lake Victoria Basin and the paper's original simulation scenario, the DBFGM required over 3 hours to complete on a Jetstream2 m3.medium instance, significantly more time compared to the BGGM and PSFGM models, which completed in just 3-4 minutes. While DBFGM shows a slight improvement in performance metrics (TPR, FPR, MCC) over the other models, the difference in accuracy is not substantial enough to justify the excessive computational time, especially in time sensitive or limited resource scenarios.

Furthermore, the computational burden increases when scaling the model to higher-dimensional data, as would be the case in large-scale climate models or more complex real-world applications. In these situations, the proposed model may struggle to perform efficiently. For applications involving large-scale data, concepts such as parallelization or optimization of the MCMC sampling process, could be explored to help mitigate the computational challenges associated with DBFGM.