

Guidelines for using the formatting and template of the Statistical Modeling and Data Science journal

Ahmad Baihaqi Rahman¹ , Chandra Dona Wibawa² , Edy F. Lee³ , Githa Hanum¹ 

¹ Department of Statistics and Data Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

² School of Data Science, University of Melbourne, Australia

³ Department of Statistics, National University of Singapore, Singapore

*Corresponding author, email: baihaqi.rahman@unnes.ac.id

March 9, 2026

Abstract

The abstract should provide a concise and self-contained summary of the paper with a maximum length of 250 words. It should briefly introduce the research background and clearly state the main objective or research question of the study. The abstract should then provide a short description of the methodology, modeling framework, or analytical approach used, without including excessive technical detail. Authors should summarize the main results and highlight the most important findings obtained from the analysis. Finally, the abstract should conclude with a brief statement describing the main contribution of the study and its potential implications within the broader context of statistical modeling, data science, or the relevant application domain.

Keywords: Bayesian hierarchical models, spatio-temporal statistics, environmental data science, air pollution modeling

1 Introduction

Statistical modeling of environmental systems has become increasingly important in the era of large-scale sensor networks and open environmental data. The rapid development of monitoring technologies has enabled the collection of high-resolution observations across space and time, creating new opportunities to better understand complex environmental processes. However, environmental data often exhibit substantial spatial and temporal variability, measurement uncertainty, and heterogeneous structures, which pose significant challenges for conventional statistical methods. Recent advances in statistical and computational approaches have provided new tools for analyzing such complex data, including methods from spatial statistics, spatio-temporal modeling, and modern data science. Motivated by these developments, this study aims to develop a statistical framework for modeling environmental processes that accommodates complex data structures while maintaining computational efficiency. The proposed approach is intended to provide reliable inference and improved predictive performance in environmental monitoring applications. The remainder of this paper is organized as follows. Section 2 presents the methodological framework. Section 3 reports the empirical results and experimental evaluation. Section 4 discusses the main findings and implications. Finally, Section 5 concludes the paper.

This section should introduce the research topic and explain its relevance within the broader context of statistical science and data analysis. The author should provide sufficient background to motivate the study, highlighting the importance of the problem and the challenges associated with modeling or analyzing the phenomenon under consideration. Relevant literature may be briefly reviewed to position the study within existing research and to identify gaps that motivate the present work. The introduction should then clearly state the objective of the study and summarize the main contributions of the paper, including methodological, theoretical, computational, or applied aspects. Finally, the section may conclude with a brief outline of the organization of the paper to guide the reader through the subsequent sections.

2 Methodology

Let $Y(s, t)$ denote pollutant concentration at location s and time t . The hierarchical model is written as

$$Y(s, t) = X(s, t)^T \beta + w(s, t) + \epsilon(s, t)$$

where $w(s, t)$ represents spatial dependence and $\epsilon(s, t)$ denotes measurement error [1, 2].

2.1 Model Estimation

Parameter estimation is conducted using a Markov Chain Monte Carlo (MCMC) sampling scheme. In the Bayesian framework, inference is based on the posterior distribution of the model parameters given the observed data. Let $\theta = (\beta, w, \sigma^2)$ denote the collection of unknown parameters. The posterior distribution can be expressed using Bayes' theorem as

$$p(\theta | Y) \propto p(Y | \theta) p(\theta), \quad (1)$$

where $p(Y | \theta)$ denotes the likelihood function and $p(\theta)$ represents the prior distribution of the parameters. As indicated in Equation (1), the posterior distribution combines information from the observed data and the prior specification. Since the posterior distribution is generally not available in closed form, samples are generated using an MCMC algorithm to approximate posterior summaries and conduct statistical inference.

2.2 Illustrative Figure

To include figures in the manuscript, authors should provide high-quality image files that are clear, readable, and not blurred or pixelated. Figures should ideally be generated using standard scientific plotting or graphics software to ensure a professional appearance. Vector-based graphics are strongly preferred because they maintain resolution when scaled or printed; therefore, formats such as **.eps** and **.pdf** are highly recommended. If vector graphics are not available, high-resolution raster images such as **.png** may be used, provided that the resolution is sufficiently high (typically at least 300 dpi). For consistency and easier management during the submission and production process, figure files should be named sequentially according to their order of appearance in the manuscript, for example **Fig1-Name1**, **Fig2-Name2**, and so forth.

Figure 1 shows a comparison between observed passenger counts and predictions from six models (M1–M6) over the test period. Most models capture the overall upward trend and seasonal fluctuations, although some deviate during peak periods. Among them, models M4 and M5 produce predictions that most closely follow the observed data across most of the time horizon.

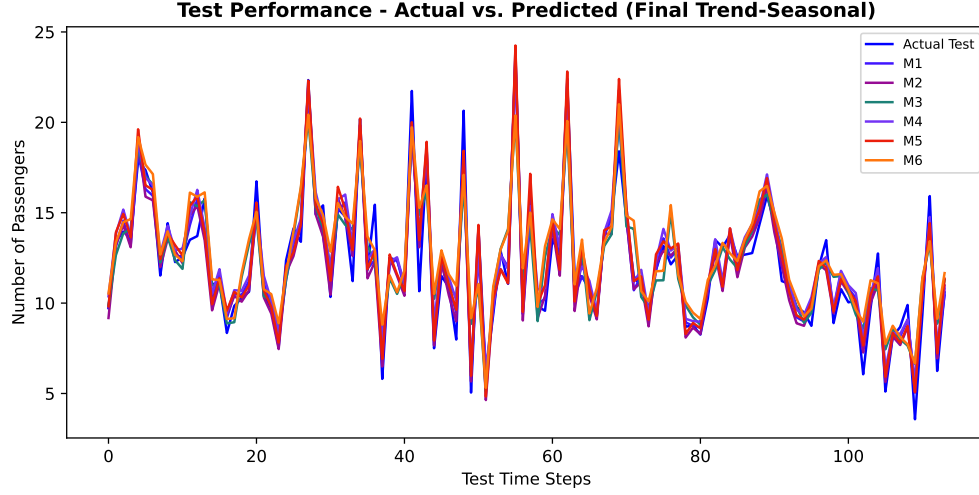


Figure 1: Comparison of observed and predicted passenger counts from six forecasting models

3 Empirical Results/Experiments

Table 1 summarizes the predictive performance of several models evaluated on the air quality dataset. The comparison includes traditional statistical methods as well as machine learning approaches. Performance is assessed using multiple evaluation metrics, including the root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). Lower values of RMSE, MAE, and MAPE indicate better predictive accuracy, while higher values of R^2 indicate a better model fit. As shown in Table 1, the proposed hybrid Bayesian model achieves the best overall performance across most evaluation metrics, indicating its ability to capture both spatial dependence and nonlinear patterns in the data.

References should be managed using a BibTEX bibliography file named `SMDSBib.bib`. Authors should store the metadata of each reference (such as author, title, journal, year, doi, and other publication details) in this file, where each entry is identified by a unique citation key. To cite a reference in the manuscript, use the command like `[3, 4]`, where `Hjort2025` and `Mendler2025` corresponds to the identifier defined in the `SMDSBib.bib` file. The bibliography will then be automatically generated at the end of the document according to the selected bibliography style. Example: Recent studies show that combining spatial statistical models with machine learning improves air quality prediction accuracy [5, 6].

Table 1: Comparison of predictive performance across different models

Model	RMSE	MAE	MAPE (%)	R^2	Time (s)	Parameters
Linear Regression	12.6	8.4	18.5	0.72	1.2	5
Random Forest	10.3	7.1	15.2	0.81	15.4	150
Gradient Boosting	9.5	6.8	14.3	0.84	22.7	200
Bayesian Hierarchical Model	9.2	6.5	13.8	0.86	48.6	12
Proposed Hybrid Model	8.7	6.1	12.9	0.89	55.3	18

4 Discussion

This section interprets the results obtained in the empirical analysis, discusses the implications and limitations of the proposed approach, and outlines potential directions for future research.

5 Conclusion

This section summarizes the main contributions of the study and highlights the key findings. It may also briefly discuss broader implications and potential avenues for future work.

Acknowledgments

The authors thank the anonymous reviewers for their constructive comments.

Funding

This research was supported by the International Data Science Research Grant Program.

References

- [1] P. Wang et al. “A Gaussian Process Method with Uncertainty Quantification for Air Quality Monitoring”. In: *Atmosphere* 12 (10 Oct. 2021), p. 1344. ISSN: 20734433. DOI: [10.3390/atmos12101344](https://doi.org/10.3390/atmos12101344).
- [2] Y. Kiouvrekis et al. “Development of electromagnetic pollution maps utilizing Gaussian process spatial models”. In: *Science of The Total Environment* 955 (Dec. 2024), p. 176907. ISSN: 18791026. DOI: [10.1016/j.scitotenv.2024.176907](https://doi.org/10.1016/j.scitotenv.2024.176907).
- [3] A. Hjort et al. “Uncertainty quantification in automated valuation models with spatially weighted conformal prediction”. In: *International Journal of Data Science and Analytics* (July 2025), pp. 1–18. ISSN: 23644168. DOI: [10.1007/S41060-025-00862-4/FIGURES/12](https://doi.org/10.1007/S41060-025-00862-4/FIGURES/12).
- [4] F. Mendler et al. “Evaluation of spatial clustering methods for regionalisation of hydrogen ecosystems”. In: *Energy Strategy Reviews* 57 (Jan. 2025), p. 101627. ISSN: 2211-467X. DOI: [10.1016/J.ESR.2024.101627](https://doi.org/10.1016/J.ESR.2024.101627).
- [5] Y. Wang et al. “Enhancing Air Quality Forecasting: A Novel Spatio-Temporal Model Integrating Graph Convolution and Multi-Head Attention Mechanism”. In: *Atmosphere* 15 (4 Mar. 2024), p. 418. DOI: [10.3390/atmos15040418](https://doi.org/10.3390/atmos15040418).
- [6] K. S. Rautela and M. K. Goyal. “Spatio-temporal analysis of extreme air pollution and risk assessment”. In: *Journal of Environmental Management* 373 (2025), p. 123807. DOI: [10.1016/j.jenvman.2024.123807](https://doi.org/10.1016/j.jenvman.2024.123807).

A Supplementary Mathematical Derivations

Additional derivations supporting the proposed statistical model.