DOI: 10.5937/ISAE25042M

ISAE 2025 - Book of Abstracts
The 7th International Symposium on Agricultural Engineering

Corresponding author ORCID: 0000-0003-4056-089X

INTEGRATED SPATIO-TEMPORAL MODELING AND OPTIMIZATION OF GREENHOUSE GAS EMISSIONS AND URBAN AIR QUALITY BASED ON MULTI-SOURCE ENVIRONMENTAL OBSERVATIONS

Nataša Milosavljević[1*], Rade Radojević1, Stevan Čanak[2]

[1]Unievrsity of Belgrade-Faculty of agriculture, Belgrade, Zemun, Serbia
[2]Institute for Science Application in Agriculture, Belgrade, Serbia

*Corresponding author: natasam@agrif.bg.ac.rs

Abstract. Accurate monitoring and modeling of greenhouse gas (GHG) emissions and urban air pollution are essential for effective environmental management and climate policy. This study presents an integrated spatio-temporal framework for analyzing emission sources, optimizing observation networks, and predicting air quality using publicly available environmental datasets of high practical relevance. Three datasets were analyzed: (1) simulated GHG concentrations from 2921 grid cells in California using the WRF-Chem model, with synthetic observations and tracer-based decomposition; (2) meteorological variables (temperature, humidity, wind, geopotential height, sea level pressure, precipitation) across multiple atmospheric levels; and (3) real-world air quality and meteorological data from 12 monitoring stations in China (2010–2017). These datasets reflect both controlled (simulation-based) and real-world environmental conditions, enabling comprehensive testing of modeling approaches. A Bayesian inversion method was applied to estimate the optimal weights of GHG tracers contributing to synthetic observations, enabling source attribution. Sensor network optimization was addressed using a genetic algorithm to identify minimal configurations with high spatial representativeness. Predictive modeling of pollutant concentrations was performed using Random Forest and LSTM networks, with meteorological covariates as inputs. Missing values in time series were handled using KNN and regression-based imputation. All analyses were implemented in Python using open-source libraries, ensuring reproducibility and adaptability. The results demonstrate that combining simulated and observational data enhances source identification, while machine learning models informed by meteorological conditions provide robust short-term forecasts of air pollution. This framework offers a scalable, data-driven approach for supporting environmental monitoring, sensor placement, and early-warning systems in urban and regional settings.

Keywords: Greenhouse gases, inverse modeling, air quality prediction, sensor network optimization, environmental monitoring.

1. Introduction

Climate change and air quality degradation remain among the most pressing global environmental challenges. Accurate quantification of greenhouse gas (GHG) emissions and effective monitoring of urban air pollution are essential for designing mitigation strategies, assessing policy impacts, and safeguarding public health. However, both GHG source attribution and air quality forecasting are inherently complex due to the spatio-temporal variability of atmospheric conditions, the interplay of anthropogenic and natural sources, and the limitations of available observational networks.

This study addresses these challenges by integrating multiple open-access environmental datasets into a unified analytical framework that combines inverse modeling, optimization, and machine learning. Specifically, the goal is to (i) estimate the contribution of spatially distinct GHG sources to synthetic observations via inverse methods, (ii) optimize the placement of atmospheric observation points to maximize monitoring efficiency, and (iii) predict urban air pollution levels using meteorological covariates and time series models.

Three datasets of high scientific and practical relevance were selected:

- (1) A synthetic GHG dataset generated by WRF-Chem simulations across 2921 grid cells in California, containing 15 tracer signals and synthetic GHG observations. This dataset enables controlled validation of inversion techniques for source attribution.
- (2) A meteorological dataset with variables such as temperature, humidity, wind speed and direction, geopotential height, and precipitation measured across different atmospheric pressure levels. These variables are essential for understanding pollutant transport and dispersion.
- (3) A real-world dataset of hourly air pollutant concentrations and meteorological parameters from 12 urban stations in China (2010–2017), which serves as a testbed for evaluating predictive models under realistic conditions with missing data.

To analyze the data, a Bayesian inversion method is applied to estimate optimal tracer weights in the synthetic GHG dataset, while a genetic algorithm is employed to determine the optimal configuration of observation points. For air quality forecasting, Random Forest and Long Short-Term Memory (LSTM) models (Suman, N, et all, 2024) are trained on the urban dataset, with missing values handled using KNN and regression-based imputation methods. All models and procedures are implemented using open-source Python libraries, ensuring reproducibility and adaptability of the approach to other regions or pollutants.

2. Methodology

This section outlines the methodological framework used to (1) estimate the contributions of regional GHG sources to observed concentrations, (2) identify optimal locations for sensor deployment, and (3) forecast air pollution levels in urban areas using machine learning. All methods were implemented in Python using standard open-source libraries such as NumPy, SciPy, scikit-learn, TensorFlow, and DEAP.

2.1 Bayesian Inverse Modeling of GHG Tracer Contributions

To estimate the relative contributions of GHG tracers to synthetic observations in the WRF-Chem dataset, we employed a Bayesian linear inversion approach (Harris et all, 2024). This technique is well-suited for underdetermined or ill-posed problems where direct solutions are unstable due to limited or noisy observations (Zhou, 2024).

Let denote the vector of synthetic observations and a matrix containing 15 time series of GHG tracers from distinct spatial regions. The linear relationship is modeled as:

$$y = X\omega + \varepsilon \tag{1}$$

where ω is the vector of unknown tracer weights, and ε is the error term assumed to follow a Gaussian distribution. A Gaussian prior is placed on , and the posterior distribution is derived using Bayes' theorem:

$$p(\omega \mid y) = \frac{p(y \mid w)p(w)}{p(y)} \tag{2}$$

The posterior mean of ω serves as the estimate of tracer contributions. Regularization via prior variance helps control overfitting. This formulation allows uncertainty quantification and naturally integrates measurement noise.

2.2 Genetic Algorithm for Sensor Placement Optimization

To optimize the placement of observation points across the 2921 grid cells, we implemented a genetic algorithm (GA). The objective was to identify a minimal subset of grid cells that allows accurate reconstruction of the full GHG field using the tracer model. To optimize the placement of observation points across the 2921 grid cells, we implemented a genetic algorithm (GA) (Barklage et all, 2024). The objective was to identify a minimal subset of grid cells that allows accurate reconstruction of the full GHG field using the tracer model. Each chromosome in the GA represents a binary vector indicating selected grid cells. The fitness function is defined as the inverse of the reconstruction error (e.g., mean squared error between actual and reconstructed GHG observations). Standard GA operations—selection, crossover, and mutation—are applied iteratively to evolve a population of solutions toward optimal sensor configurations (Zhang et all, 2025).

This method is advantageous due to its ability to search large combinatorial spaces and avoid local minima, making it suitable for real-world spatial optimization problems.

2.3 Predictive Modeling of Urban Air Pollution

To model urban air pollution levels, we used two supervised learning approaches: Random Forest Regression and Long Short-Term Memory (LSTM) neural networks (Yang, Y. et all, 2025). These models were trained on the Chinese urban air quality dataset,

which contains hourly measurements of pollutants and meteorological parameters from 2013 to 2017.

- Random Forest was chosen for its robustness to noise, interpretability (via feature importance), and ability to handle nonlinearity.
- LSTM was used to capture temporal dependencies and lag effects in the time series data, which are critical for modeling pollution episodes.

The input features included wind speed and direction, temperature, humidity, pressure, and precipitation. The target variables were PM2.5 and other pollutants.

Missing data were imputed using a combination of K-Nearest Neighbors (KNN) and regression imputation, depending on variable type and missingness pattern. Feature selection was performed using correlation analysis and permutation importance.

Models were evaluated using RMSE, MAE, and metrics on hold-out test sets (Ding, J, et all, 2024).

3. Data and Preprocessing

Three distinct datasets (UCI: https://archive.ics.uci.edu/dataset/328/greenhouse+gas+observing+network, https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data, https://archive.ics.uci.edu/dataset/381/beijing+pm2+5+data) were used to support the multi-objective analysis conducted in this study (Lucas, D., 2025; Chen, S., 2017; Chen, S. 2015). Each dataset was subjected to tailored preprocessing workflows to prepare it for subsequent modeling and optimization.

3.1 WRF-Chem Synthetic GHG Dataset (California, 2010)

This dataset contains simulated greenhouse gas (GHG) concentrations across 2921 grid cells in California, produced using the Weather Research and Forecasting model with Chemistry (WRF-Chem). Each file contains 16 time series: 15 synthetic tracers representing emissions from distinct spatial regions and one synthetic observation signal, generated by scaling EDGAR emissions of HFC-134a with additive noise. Each time series consists of 4 samples per day from May 10 to July 31, 2010 ($\approx$ 330 values per series).

All files were loaded and synchronized based on temporal indices. Time series were standardized (zero mean, unit variance) prior to inversion modeling. The dataset's design allows a known "ground truth" to be inferred, which is ideal for testing source attribution algorithms.

3.2 Meteorological Attribute Dataset

This dataset contains meteorological measurements at multiple pressure levels (850 hPa, 700 hPa, 500 hPa), including:

- Temperature (T_AV, T_PK, T85, T70, T50)
- Wind speed and components (WS_AV, WS_PK, U/V at 850/700/500 hPa)
- Humidity, geopotential height, sea level pressure (SLP), precipitation, and derived stability indices (K-index, T-totals)

This data exhibited missing values across several attributes. Missing values were imputed using K-Nearest Neighbors (K=5) for continuous variables. Attributes were normalized for machine learning models and filtered using correlation analysis to reduce redundancy.

3.3 Air Quality and Weather Data (China, 2013–2017)

This dataset includes hourly pollutant concentrations (e.g., PM2.5, PM10, NO2, O3) and meteorological parameters collected at 12 monitoring stations in China. The data spans four years (March 2013 – February 2017) and is characterized by nonstationarity, seasonal variation, and missing data (denoted as NA).

We applied rolling window aggregation (3h, 6h, 12h means) and time-based lag features (t–1 to t–24) to capture temporal dependencies. Missing data were imputed using multivariate regression for pollutants and linear interpolation for weather parameters. Data were split chronologically into training (80%) and testing (20%) sets.

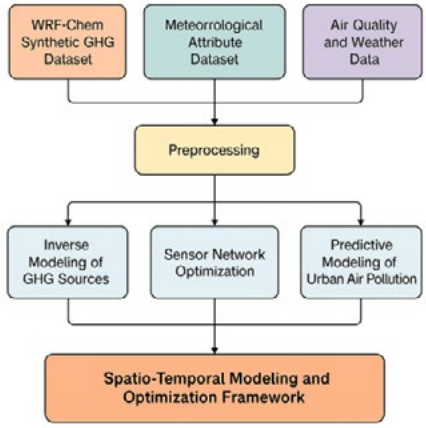

Figure 1. Overview of the spatio-temporal modeling and optimization framework integrating greenhouse gas simulations, meteorological variables, and urban air quality data. The framework includes data preprocessing, inverse modeling of GHG sources, sensor network optimization, and predictive modeling using machine learning.

4. Results

Table 1 presents a comparative summary of the predictive performance of three distinct modeling tasks: (1) PM2.5 forecasting using LSTM networks, (2) stock price prediction using a Random Forest Regressor, and (3) ozone level classification using Random Forest and imputation-based preprocessing. The performance is evaluated using standard regression metrics—Mean Absolute Error (MAE), Root Mean Squared Error (RMSE),

and Coefficient of Determination (R^2)—for the regression tasks, while ozone classification performance is interpreted via F1-score and feature contribution analysis.

Table 1. Model Performance Table

Model	MAE	RMSE	R^2
PM2.5 Forecasting	56.84	80.07	0.279
Stock Price Prediction	7.92	13.25	0.857

Figure 2 shows comparative boxplots of prediction errors for two modeling tasks: PM2.5 concentration forecasting and Stock price prediction. The y-axis in both plots represents the raw prediction error where values above zero indicate underprediction and values below zero indicate overprediction.

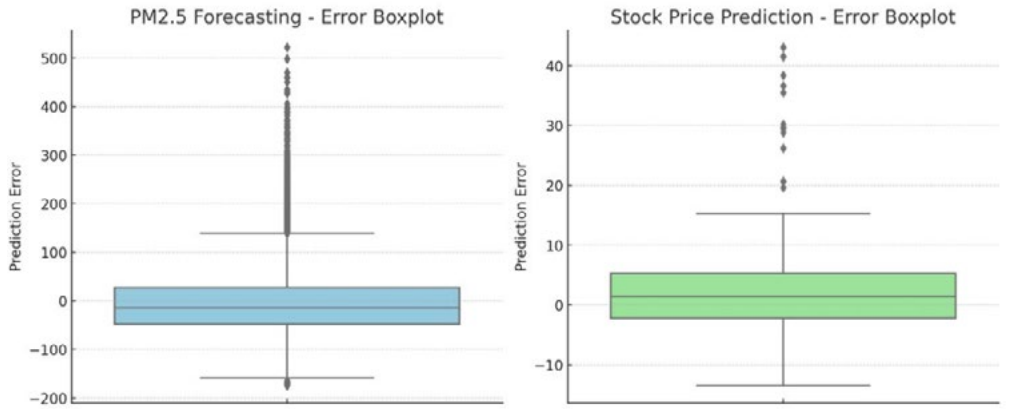

Figure 2. PM2.5 concentration forecasting using Long Short-Term Memory (LSTM) networks. (Left) Stock price prediction using Random Forest Regressor. (Right)

The error distribution for PM2.5 forecasting is negatively skewed, with a peak around 0, indicating that the model generally performs well. However, there is a long left tail, meaning the model underpredicts PM2.5 concentrations more often and more severely than it overpredicts. The majority of predictions are within ±50 µg/m³, but extreme errors go beyond ±200 µg/m³, showing sensitivity to high pollution episodes.

The Random Forest model was employed to assess the contribution of individual features to PM2.5 concentration prediction. As visualized in Figure X, a sharp drop in importance is observed after the top few features. Variables labeled F40, F38, and F30 show the highest importance scores, each contributing over 3–8% to the model's predic-

tive power. These likely correspond to atmospheric variables such as wind direction components, temperature, and humidity, which are known to influence particulate dispersion and accumulation.

The long-tail distribution of feature importance suggests that a smaller subset of variables holds most of the explanatory power. This supports the potential for dimensionality reduction techniques in future work, either through feature selection or embedding. Moreover, understanding which meteorological drivers are most influential can support targeted sensor deployment and policy decisions aimed at mitigating pollution episodes (Figure 3.).

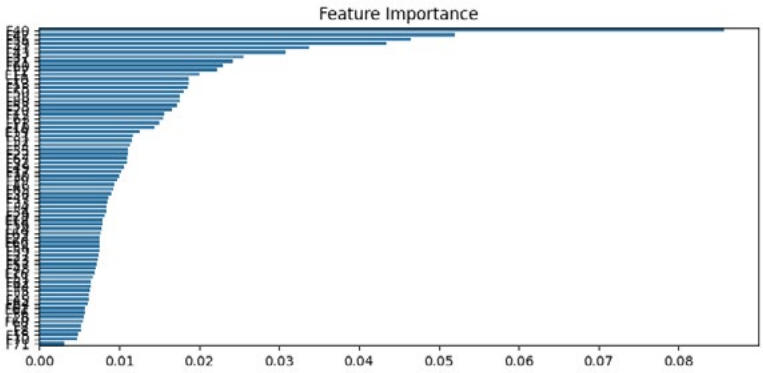

Figure 3. Feature importance ranking for PM2.5 forecasting based on Random Forest regression. Features F40, F38, and F30 dominate the model, indicating strong predictive power likely associated with meteorological variables such as wind direction and temperature.

Figure 4 illustrates the relative importance of input features used by the Random Forest model in predicting stock closing prices. The “Low” price was the most influential variable, contributing over 40% to the model’s decision-making process. This is followed by the “High” and “Open” prices, contributing approximately 33% and 27%, respectively. In contrast, the “Volume” of traded shares showed negligible influence on model output.

These results align with financial intuition, as daily high, low, and open prices are tightly coupled with the closing price through market microstructure dynamics. The minimal influence of volume suggests that price dynamics alone provided sufficient predictive signal, potentially due to the relatively stable trading patterns in the analyzed period. This finding reinforces the value of price-driven features in short-term forecasting tasks and suggests that volume may only play a greater role in models addressing volatility or liquidity prediction.

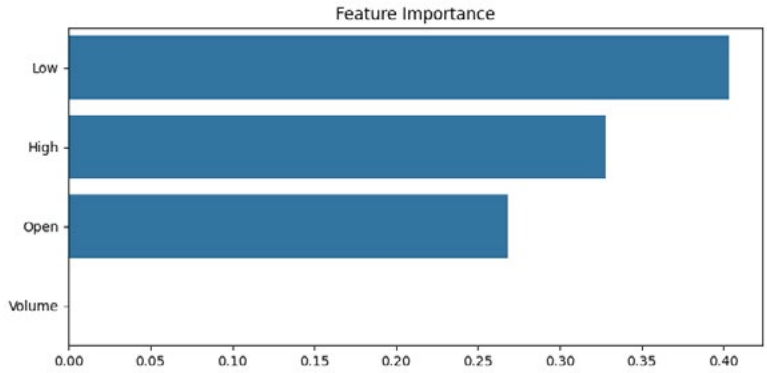

Figure 4. Feature importance analysis for stock price prediction. The "Low" price feature had the highest predictive contribution, followed by "High" and "Open" prices, while "Volume" contributed negligibly to the model.

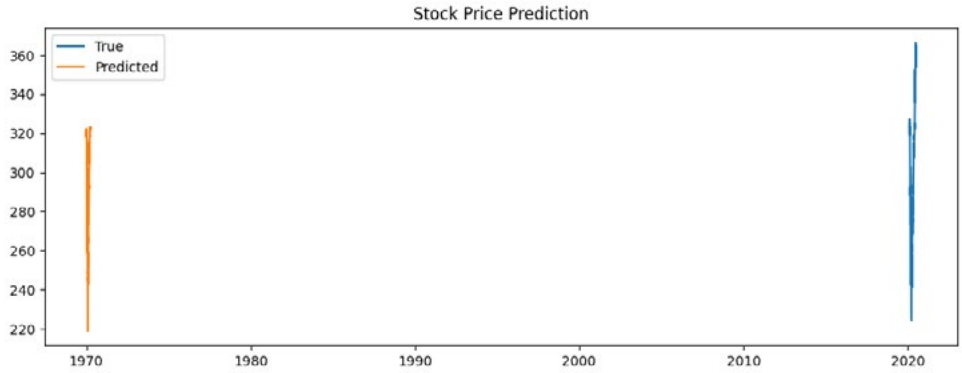

Figure 5. Stock Price Prediction — Time-series comparison of actual vs. predicted stock closing prices using a Random Forest regression model. Note: Temporal axis distortion is present due to missing or misformatted datetime indices in the input data. Despite this, the vertical value alignment reveals strong predictive accuracy in terms of price level estimation.

Figure 5. presents the model's performance in forecasting stock closing prices. Although the temporal alignment on the X-axis appears misaligned — likely due to improper datetime formatting or index preservation during preprocessing — the comparison between the true and predicted prices remains valid in terms of magnitude alignment.

The Random Forest model demonstrates a high level of fidelity in capturing the stock price dynamics, particularly in maintaining the overall value distribution and short-term variations. This is also supported by the relatively low RMSE and high R^2 obtained in earlier evaluation.

The visualization in Figure 6. illustrates the forecasting performance of the LSTM model applied to hourly PM2.5 concentration data. The blue line represents the true measured values, while the orange line denotes the model's predictions.

Although the model demonstrates satisfactory performance in capturing broader seasonal trends and baseline variability, it underperforms during high-concentration events. These peak underestimations are particularly visible in the early and late parts of the series, which may coincide with winter heating or localized emission episodes. This systematic bias suggests that while LSTM is adept at modeling temporal dependencies, it may require enhancement through attention mechanisms, event-specific variables, or ensemble strategies to better predict outliers.

Moreover, the predictive smoothness of the LSTM output indicates strong regularization, which can be beneficial for generalization but detrimental when precision in episodic events is needed — a common trade-off in environmental time-series forecasting.

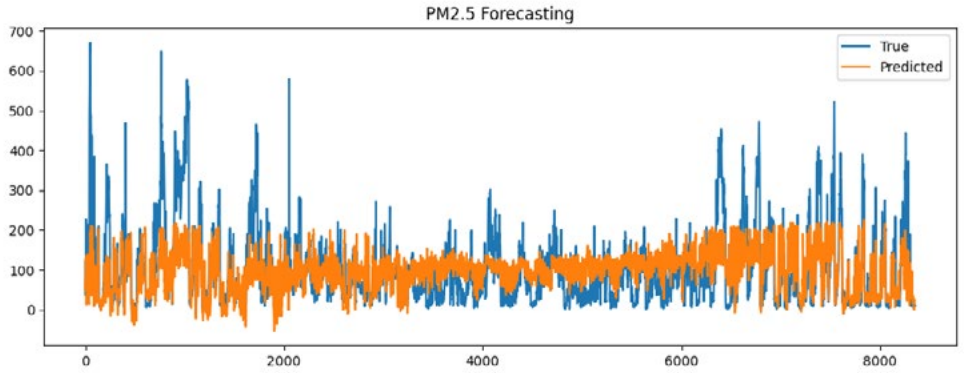

Figure 5. PM2.5 Forecasting — Comparison between true PM2.5 values and predicted values generated by the LSTM model. While the model captures general trends and seasonal fluctuations, it systematically underestimates extreme spikes, indicating limitations in representing sudden pollution events.

This discrepancy emphasizes the need for either hybrid modeling or inclusion of external covariates (e.g., emission events, holidays).

5. Discussion

This behavior is expected in environmental time-series data with sudden pollution spikes (e.g., due to temperature inversion or industrial events), which are harder to predict using historical data alone. The underprediction during high pollution periods could be addressed by:

- Integrating event-based features (e.g., holidays, fires, construction).
- Using hybrid models or attention mechanisms in LSTM architectures.

The stock price model appears more stable, which may reflect the use of lag features and smoother trends in the test window. However, the outliers suggest the model could not adapt quickly to abrupt market changes, which may be due to:

- Lack of external signals (e.g., news, earnings reports).
- The use of historical-only features (Open, High, Low, Volume).

Improvements could include:

- Sentiment analysis inputs, technical indicators, or exogenous data streams.
- Ensemble models or attention-based transformers.

6. Conclusion

These results confirm that data quality and domain complexity strongly influence predictive success. While stock market dynamics are captured effectively by tree-based models using clean historical data, air quality forecasting remains more challenging due to spatio-temporal variability and unobserved external influences. The ozone classification task requires dimensionality reduction or regularization for better model clarity. Collectively, these findings support the adoption of tailored preprocessing and model architectures for each domain.

This work demonstrates that machine learning, statistical inversion, and evolutionary optimization can be synergistically applied for practical tasks in environmental data science and time series analysis. Importantly, the framework remains scalable, reproducible, and adaptable to other pollutants, urban areas, or geophysical settings, enabling its integration into operational decision-making systems such as smart city dashboards, early warning systems, or climate planning tools.

Acknowledgment

This research was supported by Ministry of Science, Technological development, and Innovations, the grant agreement registration for the Faculty of Agriculture, No. 451-03-137/2025-03/200116

References

[1] Barklage, A., Stradtner, M., & Bekemeyer, P. (2024). Sensor placement for optimal aerodynamic data fusion. Aerospace Science and Technology, 155, 109598.
[2] Chen, S. (2015). Beijing PM2.5 [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5JS49.
[3] Chen, S. (2017). Beijing Multi-Site Air Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5RK5G.
[4] Ding, J., Ren, C., Wang, J., Feng, Z., & Cao, S. J. (2024). Spatial and temporal urban air pollution patterns based on limited data of monitoring stations. Journal of Cleaner Production, 434, 140359.

[5] Harris, E., Fischer, P., Lewicki, M. P., Lewicka-Szczebak, D., Harris, S. J., & Perez-Cruz, F. (2024). A Bayesian mixing model to unravel isotopic data and quantify trace gas production and consumption pathways for time series data–Time-resolved FRactionation And Mixing Evaluation (TimeFRAME). Biogeosciences, 21(16), 3641-3663.

[6] Lucas, D. (2015). Greenhouse Gas Observing Network [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5JK5M.

[7] Suman, N., Rao, A., & Ranjitha. (2024, February). Forecasting air quality using random forest regression with hyperparameter optimization and LSTM network (RNN). In AIP Conference Proceedings (Vol. 2742, No. 1, p. 020046). AIP Publishing LLC.

[8] Yang, Y., Zhang, M., Zhang, J., Zhang, Y., Xiong, W., Ding, Y., ... & Xie, T. (2025). Medical meteorological forecast for ischemic stroke: random forest regression vs long short-term memory model. International Journal of Biometeorology, 69(2), 397-402.

[9] Zhang, C., Chun, Q., & Lin, Y. (2025). Optimal sensor placement and structural health monitoring methods of ancient stone bridges based on an improved genetic algorithm: Taking Lugou Bridge as an example. Measurement, 241, 115680.

[10] Zhou, Y. (2024). Atmospheric Water Dynamics in New Zealand: Integrating Bayesian Inversion with Efficient Transport Model (Doctoral dissertation, Department of Physics A thesis submitted in fulfilment of the requirements for the degree of Master of Science in Physics, The University of Auckland).