

COMPARISON OF DAILY OZONE CURVE FORECASTING BETWEEN MULTI-STEP AND ITERATIVE ONE-STEP AHEAD METHODS WITH DIFFERENT BASIS FUNCTIONS

(Perbandingan Ramalan Lengkuk Harian Ozon antara Kaedah Pelbagai Langkah dan Lelaran Satu Langkah ke Depan dengan Fungsi Basis Berbeza)

ROS RASYIAH ROSSLAN, NORSHAHIDA SHAADAN* & SAYANG MOHD DENI

ABSTRACT

Ground-level Ozone pollution causes several potential environmental and human health issues. Forecasting the concentration levels helps to mitigate the problem. This study compared two functional time-series (FTS) forecasting methods, Multistep-ahead and Iterative One-step ahead, for forecasting daily Ozone curves. Beyond comparing the relative performance of the two methods, this study investigates the impact of different basis functions (B-spline, Fourier, and Polygonal) on the forecasting performance. Hourly recorded Ozone data obtained from the Department of Environment (DOE) Malaysia, involving Shah Alam, Batu Muda and Klang air quality monitoring stations were utilized. The methodology began by transforming real observed hourly Ozone data from January 1 to November 6, 2019, to daily Ozone curves using the basis expansion method, followed by FTS modelling. The models' forecasting performance was evaluated using error measures including Mean Squared Error (MSE), Mean Absolute Error (MAE), Integrated Squared Error (ISE), and the coefficient of determination (R^2). The analysis results indicate that variations in the types of basis functions influence forecasting performance across stations; however, no significant differences are observed within the same station. Overall R^2 of Multi-step ahead method is found to be higher than the Iterative One-step ahead method in forecasting daily Ozone curves at the study locations, with (0.8017, 0.8214, 0.629) and (0.7831, 0.8208, 0.6154), respectively. The findings suggest that, while the Multi-step ahead approach offers a somewhat improved functional representation of diurnal Ozone dynamics, its predictive performance remains comparable to that of the Iterative one-step ahead approach, indicating only a modest superiority rather than absolute.

Keywords: Ozone; functional data analysis; functional time series; air quality; forecasting

ABSTRAK

Pencemaran Ozon permukaan bumi menyebabkan pelbagai kemungkinan isu alam sekitar dan kesihatan manusia. Peramalan aras kepekatan Ozon dapat membantu menangani masalah tersebut. Kajian ini membandingkan dua pendekatan model ramalan lengkung (FTS), kaedah pelbagai langkah ke depan dan lelaran satu langkah ke depan dalam meramal lengkung harian kepekatan Ozon. Bagi mengenal pasti kaedah terbaik, kesan tiga jenis fungsi asas yang berbeza (B-spline, Fourier, dan Poligonal) terhadap prestasi model FTS telah dikaji. Kajian ini menggunakan data setiap jam yang diperoleh daripada Jabatan Alam Sekitar (DOE) Malaysia, meliputi tiga stesen pemantauan kualiti udara: Shah Alam, Batu Muda dan Klang. Metodologi kajian bermula dengan menukarkan cerapan Ozon setiap jam bagi tempoh 1 Januari hingga 6 November 2019 kepada data lengkung Ozon harian menggunakan kaedah pengembangan fungsi asas, diikuti dengan pemodelan FTS. Prestasi ramalan model dinilai menggunakan ukuran ralat termasuk Ralat Kuasa Dua Min (MSE), Ralat Mutlak Min (MAE), Ralat Kuasa Dua Bersepadu (ISE), serta pekali penentuan (R^2). Hasil analisis menunjukkan bahawa variasi jenis fungsi asas mempengaruhi prestasi ramalan antara stesen, namun tidak terdapat perbezaan yang signifikan dalam impaknya terhadap prestasi ramalan dalam stesen yang sama. Berdasarkan purata keseluruhan nilai R^2 , kaedah pelbagai langkah ke depan menunjukkan prestasi yang lebih baik berbanding kaedah lelaran satu langkah ke depan dalam meramal lengkung Ozon harian di

lokasi kajian, dengan nilai (0.8017, 0.8214, 0.629) dan (0.7831, 0.8208, 0.6154) masing-masing. Dapatan kajian menunjukkan bahawa walaupun pendekatan pelbagai langkah ke depan memberikan perwakilan fungsi yang sedikit lebih baik, prestasi ramalannya kekal setara dengan pendekatan lelaran satu langkah ke depan dengan kelebihan yang sederhana dan bukannya mutlak

Kata kunci: Ozon; analisis data fungsian; siri masa berfungsian; kualiti udara; ramalan

1. Introduction

High concentration levels of air pollutants, including ground-level Ozone, contribute to air pollution and pose health risks to humans and the environment. In Malaysia, ground-level Ozone concentrations have become a significant concern due to the tropical climate and urbanization. High temperatures and ample sunlight, combined with emissions from vehicles and industrial activities, create favorable conditions for Ozone formation (Cao *et al.* 2024). The Department of Environment (DOE) of Malaysia monitors air quality across the country, noting that elevated Ozone levels can lead to poor air quality indexes, which have implications for public health, including respiratory problems and exacerbation of conditions such as asthma. Efforts to address Ozone pollution in Malaysia include regulatory measures to control emissions and public awareness campaigns to mitigate its health impacts. Understanding and managing Ozone levels are vital for ensuring both environmental and public health.

Forecasting Ozone concentration is important because it helps authorities prepare for periods of poor air quality and issue early warnings to the public. Exposure to high Ozone levels can affect the respiratory and cardiovascular systems and harm plants and ecosystems (De Marco *et al.* 2022; Merz *et al.* 2020). Traditional forecasting models, such as Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), and volatility-based models such as GARCH, have been widely employed in time-series modelling and forecasting applications (Mohd Ali *et al.* 2022). However, these conventional approaches generally model observations as scalar time-series values rather than complete functional trajectories. However, Ozone levels change throughout the day, typically rising after sunrise, peaking in the afternoon, and decreasing in the evening (Ahmad *et al.* 2023). This means that predicting the full daily pattern, rather than a single value, can provide more useful information. To enable the analysis, the application of functional data analysis (FDA) provides the solution. In the FDA, each observation is treated as a curve instead of a single point value. In the context of Ozone data, each day is considered as one curve comprising continuous 24-hours Ozone levels throughout the day. Notably, in the literature, the application of the FDA for forecasting Ozone curves in the Malaysian environment, or for other pollutants, remains limited.

Very few FDA research specially on curve forecasting, were found. Among several popular applications were the study by Hyndman and Shang (2009) on sea surface temperatures in climatology and the study of Hyndman and Ullah (2007) to forecast Australian fertility curves. They highlight that their application of an FDA-based forecasting approach was effective. Another study by Xie *et al.* (2024) on enhancing the accuracy of 10-day forecasts of hourly Ozone concentrations by integrating chemical transport models (CTMs) with machine-learning approaches. The output helps overcome the limitations and uncertainties associated with traditional CTMs, which can be influenced by factors such as meteorological fields, physicochemical parameterization schemes, and emission inventories that contribute to inaccuracies in Ozone forecasts. A study by Shang *et al.* (2021) used the FDA to develop a

robust forecasting method for population data using multilevel functional data analysis techniques. The focus is on improving the accuracy of population forecasts by capturing complex demographic patterns over time. The contribution shows that this method provides more accurate forecasts compared to traditional demographic forecasting methods.

In Malaysia, Din and Shaadan (2022) utilized Functional Time Series (FTS) models to forecast monthly diurnal maximum Air Pollution Index (API) curves for Shah Alam and Pasir Gudang. The study compared two methods that were introduced by Hyndman and Shang (2009), namely the Multi-step ahead and the Iterative One-step ahead approaches. The Multi-step ahead model demonstrated superior performance, achieving lower error rates compared to the Iterative One-step ahead model in their study. The findings from research conducted by Din and Shaadan (2022) indicate that the FTS model effectively predicts continuous API levels over 24 hours, with API levels peaking at 3:00 pm and increasing after sunrise, reflecting higher vehicle emissions in the morning. However, the application of the FTS model in their study did not address whether the choice of basis functions in the model development affects the forecasting performance. This lack represents a gap in the existing research analysis.

Within the FTS framework, Functional Principal Component Analysis (FPCA) plays a central role by decomposing functional data into orthogonal components that capture dominant patterns and variability, alongside the integration of univariate time series forecasting models. The process began with data conversion into curves using basis functions expansion with a regression smoothing method (Bosq 2000). Different basis functions have different strengths (Ramsay & Silverman, 2005). It might affect forecasting performance. Spline functions are flexible and can capture local changes, Fourier functions are suitable for smooth and repeating patterns, and polygonal functions can represent sharp changes more clearly. In the context of the daily Ozone dynamic, the cycle may not always be perfectly smooth; it is not clear which basis function works best. Fourier functions may perform well because of the cyclical pattern, while spline and polygonal functions may better capture irregular changes during the day. Therefore, it is important to compare these approaches and their effect on the forecasting performance. However, not many studies have explored how this choice influences forecasting results.

To further extend the study of Din and Shaadan (2022) and to fill the gaps, this study aims to examine the application of the two FTS forecasting methods, the Multi-step ahead and the Iterative One-step ahead, to forecast daily Ozone curves at Shah Alam, Batu Muda, and Klang locations. At the same time, the effects of three different basis functions, including B-spline, Fourier, and Polygonal, on forecasting performance are also investigated.

2. Methodology

2.1. Data and Study Location

For experimental purposes, this study utilized an available historical dataset obtained from the DOE Malaysia, comprising hourly Ozone concentration (ppm) records collected between January 1 and November 6, 2019, from three air quality monitoring stations in the Klang Valley region: Shah Alam, Batu Muda, and Klang. The Klang Valley spans approximately 2,832 km² and had become highly urbanized by 2014, with a population exceeding 4 million, representing around 16% of Malaysia's total population (Naji *et al.* 2014). According to the Department of Statistics Malaysia's census in July 2023, the population of Klang was estimated at 1,132,800 and is expected to continue growing. The Klang air quality monitoring station is strategically located in the heart of the city, adjacent to a busy industrial zone and major transportation routes, including Port Klang, which handled nearly half of Malaysia's seaborne container trade

in 2013, highlighting the area's economic significance. The Batu Muda monitoring station is positioned at Sekolah Kebangsaan Batu Muda, within a locality that houses a population of approximately 1.6 million. Meanwhile, the Shah Alam station is located at Taman Tun Dr. Ismail (TTDI) Jaya Primary School in Shah Alam, Selangor. Shah Alam, a key district within the Petaling region, is densely populated with around 2.3 million residents

Table 1: Information on the Study Locations

Station Name	Latitude, Longitude	Background	Category
Shah Alam	3°6'17.03"N, 101°33'21.66"E	Urban	Residential
Batu Muda	3°12'45.08"N, 101°40'56.47"E	Urban	Residential
Klang	3°0'53.72"N, 101°24'47.02"E	Urban	Industrial

2.2. Research Framework

Figure 1 illustrates the research framework designed to achieve the study objectives. This includes several key stages: starting with data acquisition, followed by data preparation and cleaning, discrete data conversion into daily curves, modelling daily functional time series and finally performance assessment.

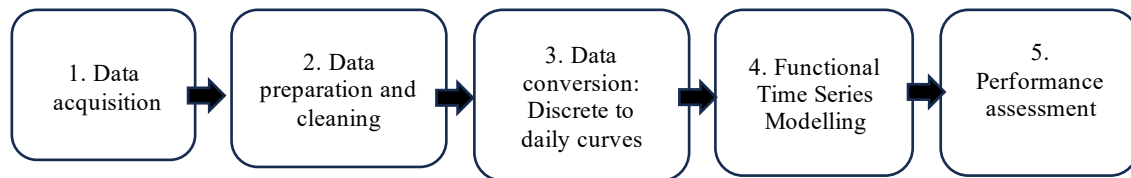


Figure 1: Research framework

Following data acquisition in stage 1, the data are prepared and undergo a cleaning process in stage 2. To preserve the temporal element, the missing values were treated with hourly means (column means) (Tabachnick & Fidell 2007). Moreover, the percentage of missing values in the data set was small, which is about 1%. Previous studies suggest that when the percentage of missing data is minimal (< 5%), simple imputation techniques such as mean substitution have little to no influence on results and inference, thus acceptable for replacing missing values (Acuna & Rodriguez 2004).

Next, the data were organized into a matrix or data frame format, with hours 1 to 24 as the variables and days as the rows or cases. The data must be in this format so that a set of daily Ozone curves for the analysis can be created using a basis function approach in stage 3. In stage 4, to begin with the FTS modelling, the converted data set, in the form of daily curves, was divided into two. The daily Ozone curves from 1 Jan 2019 to 27 Oct 2019, consisting of 300 functional observations, were used for developing the FTS model. Meanwhile, the remaining daily curves from 28 Oct 2019 to 6 Nov 2019 were used for validation and assessment at stage 5.

In the FTS modelling and development stage 4, FPCA was applied to obtain the optimal number of PCs and the scores. The resulting principal component scores are modelled as time series and forecasted using two approaches: Multi-step ahead and Iterative One-step ahead approaches. The forecasted curves were then reconstructed, and the forecasting performance was evaluated and compared using pointwise forecast accuracy measures such as Mean Squared Error (MSE), Mean Absolute Error (MAE), as well as a functional-based method, the Integrated Squared Error (ISE). R^2 , which is the measure of the percent variation explained by the model,

was also computed to assess the capability of the model to capture the daily diurnal pattern. The analysis for this study was performed using R software (R Development Core Team, 2008).

2.3. Basis Functions

Basis expansion is a technique used in statistics and machine learning to transform data into a higher-dimensional space, making it easier to capture complex relationships and improve model performance. In many applied studies, the selection of basis functions (e.g., Fourier, B-spline, wavelet, Polynomial) is made based on convention or convenience rather than systematic evaluation. However, different bases may lead to substantially different forecasts, even when the same forecasting model is used. Regression smoothing is a technique used to fit a smooth curve to data points, helping to capture underlying trends while reducing the impact of noise. Curve development often employs regression smoothing with different types, each with its strengths and use cases. Although Ozone diurnal curves may exhibit seasonal patterns, the current study did not explicitly account for seasonality in the comparison of basis functions. Instead, the focus was on evaluating overall curve representation using B-spline, Fourier, and polygonal bases across the observed data.

2.3.1. Spline basis

Spline basis functions, including B-splines and natural splines, are used to model smooth functions and are particularly useful in functional data analysis and curve fitting. Recent research by Schuster *et al.* (2022) highlighted the use of B-splines to improve the flexibility and interpretability of functional models in epidemiology. Spline bases are a powerful tool in regression and data smoothing. They provide a flexible method for fitting smooth curves to data by partitioning the data range into intervals and using polynomials within those intervals. The B-splines are a specific type of spline function that allows for efficient computation and flexibility. A B-spline function is a type of piecewise polynomial used to represent smooth curves. B-spline builds a curve by joining several low-degree polynomials smoothly and stably at certain points called knots. A polynomial function is defined by:

$$B_{i,d}(x) = \frac{x-t_i}{t_{i+d}-t_i} B_{i,d-1}(x) + \frac{t_{i+d+1}-x}{t_{i+d+1}-t_{i+1}} B_{i+1,d-1}(x), \quad (1)$$

with the initial conditions:

$$B_{i,0}(x) = \begin{cases} 1 & \text{if } t_i \leq x < t_{i+1}, \\ 0 & \text{otherwise,} \end{cases}$$

where t_i are the knots and d is the degree of freedom.

2.3.2. Fourier basis

Fourier basis functions, including sines and cosines, are used to decompose periodic functions or signals into their frequency components. They are particularly useful in time series analysis and signal processing. A recent study by Xie *et al.* (2024) explored the application of Fourier basis functions in handling complex periodic data and improving forecasting accuracy in financial markets. Here are the key equations associated with the Fourier bases. For a periodic function $f(x)$ with period T , the Fourier series expansion expresses $f(x)$ as a sum of sines and cosines:

$$f(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos\left(\frac{2\pi nx}{T}\right) + b_n \sin\left(\frac{2\pi nx}{T}\right)), \quad (2)$$

where:

- n is an integer representing the frequency index (e.g., $n = 0, 1, 2, \dots$) ~ number of k ,
- x is the time (day) ~ $x = 0, 1, 2, \dots, 239$,
- T is the period of the function, 24 hours for daily cycles,
- T is equal to 3.142,
- a_0 is the average value of the function over one period:

$$a_0 = \frac{1}{T} \int_0^T f(x) dx,$$

- a_n and b_n are the Fourier coefficients:

$$a_n = \frac{2}{T} \int_0^T f(x) \cos\left(\frac{2\pi nx}{T}\right) dx,$$

$$b_n = \frac{2}{T} \int_0^T f(x) \sin\left(\frac{2\pi nx}{T}\right) dx.$$

2.3.3. Polygonal basis

Polygonal basis functions are a type of basis function used in various applications such as approximation, interpolation, and numerical analysis. They are particularly useful in representing piecewise linear functions, which are advantageous for their simplicity and computational efficiency. The fundamental equation for a polygonal basis function can be expressed as:

$$f(x) = \sum_{i=1}^k \alpha_i \phi_i(x) + \sum_{j=1}^k \beta_j \psi_j(x), \quad (3)$$

where:

- α_i is the coefficient for the daily piecewise linear basis functions,
- β_j is the coefficient for the trend basis functions,
- x represents a point within the polygonal domain,
- $\phi_i(x)$ are piecewise linear functions for the daily cycles,
- $\psi_j(x)$ are trend functions for longer-term trends,
- k is the number of basis functions.

2.4. Data Conversion

FTS models are designed to forecast future diurnal hourly Ozone concentration curves. To develop the FTS model, a series of daily Ozone concentration curves is required. Thus, the process begins by converting the observed data into curves using the basis expansion method. As mentioned earlier, there were three different basis functions involved (i.e., B-spline, Fourier and Polygonal) in this study. Assuming there is a set of discrete observations for day t , denoted as $y_t = [y_{t,1}, y_{t,2}, \dots, y_{t,24}]$ recorded across n days $t = [1, 2, \dots, n]$ and time points $i = [1, 2, \dots, 24]$, data conversion involves transforming the discrete data y into a continuous function $y_t(x_i)$ that can be computed at any time point i using the following equation:

$$y = y_t(x_i) + \varepsilon, \quad (4)$$

where the term ε denotes random noise and $y_t(x_i)$ taken to be a random smooth function. The following equation is used to establish the function $y_t(x_i)$:

$$y_t(x_i) = \sum_{k=1}^K C_k \varphi_k(x_i), \quad (5)$$

where term φ_k represents the basis system consisting of k basis functions. C_k refers to the basis coefficients estimated using the least-squares method. The optimum k is determined by the smallest Bayesian Information Criteria, BIC (Din and Shaadan 2022), with RSS being the residual sum of squares. The formula is as follows:

$$\text{BIC} = n \ln(n \text{Rss}) + k \ln(n). \quad (6)$$

Based on the prior analysis shown in Table 2, within the range between 10 and 20 (of 24-hour points), the number of Spline basis for Batu Muda and Shah Alam is 18, and for Klang, $k = 20$ was used in this study. The identified k values for each location data set were used consistently in the FTS modelling to ensure a fair comparison between models, while controlling model complexity and avoiding bias from differences in parameter counts.

Table 2: Summary of BIC value

Shah Alam			Batu Muda			Klang		
k	RSS	BIC	k	RSS	BIC	k	RSS	BIC
10	2.24E-05	-11.42	10	3.29E-05	-11.0356	10	1.96E-05	-11.5535
11	2.02E-05	-11.7632	11	3.48E-05	-11.2192	11	2.78E-05	-11.4438
12	1.78E-05	-12.097	12	3.15E-05	-11.5262	12	1.62E-05	-12.1912
13	1.84E-05	-12.2458	13	2.55E-05	-11.9195	13	1.77E-05	-12.2846
14	1.78E-05	-12.4404	14	2.60E-05	-12.0615	14	1.68E-05	-12.4982
15	1.60E-05	-12.585	15	2.47E-05	-12.2574	15	1.34E-05	-12.869
16	1.50E-05	-12.8869	16	1.73E-05	-12.7443	16	1.64E-05	-12.7977
17	1.66E-05	-12.9046	17	2.38E-05	-12.5443	17	1.31E-05	-13.1414
18	1.15E-05	-13.5744	18	1.37E-05	-13.2037	18	1.02E-05	-13.5007
19	1.22E-05	-13.4222	19	1.53E-05	-13.1958	19	1.47E-05	-13.2358
20	1.12E-05	-13.4072	20	1.73E-05	-13.166	20	6.55E-06	-14.1373

2.5. Modelling Daily Ozone Functional Time Series and Forecasting

The FTS model developed to predict daily Ozone concentration curves in this study is based on the approach proposed by Hyndman and Shang (2009), who first applied it to forecast annual mortality curves. The methodology combines Functional Principal Component Regression (FPCR) with FPCA. Mathematically, let $y_t(x_i)$ represent the Ozone function for the continuous hour variable x_i in day t . It is assumed that there exists an underlying smooth function $f_t(x_i)$ which is observed with errors at the discretized grid points of x . In this study, the concentration of Ozone level readings for $t = 1, 2, \dots, n$ days and $i = 1, 2, \dots, 24$ point hours are observed by the researcher as $\{x_i, y_t(x_i)\}$. From these observations, a smooth function $f_t(x_i)$ is obtained and is provided by the following equation.

$$y_t(x_i) = f_t(x_i) + \sigma_t(x_i) \varepsilon_{t,i}, \quad (7)$$

where $\varepsilon_{t,i}$ is an independent and identically distributed i.i.d standard random variable $\sigma_t(x_i)$ represents the variability of noise at x_i , and $\{x_1, x_2, \dots, x_{24}\}$ denotes a set of discrete data points. The function $f_t(x_i)$ is defined as follows:

$$f_t(x_i) = \mu(x_i) + \sum_{p=1}^P \beta_{t,p} \phi_p(x_i) + \varepsilon_t(x_i), \quad (8)$$

where $\mu(x_i)$ denotes the mean function, $\phi_p(x_i)$ represents the p -th functional principal component, and $\beta_{t,p}$ are the corresponding principal component scores for $p = 1, 2, 3, \dots, P$ where $P < k$. To forecast the daily Ozone curve for the next day, when $t = n+1$, the function is defined as follows:

$$y_{n+1}(x_i) = f_{t+1}(x_i) + \sigma_t(x_i) \varepsilon_{t+1,i}. \quad (9)$$

The mean function $\mu(x_i)$ is estimated to be using the weighted average method such that

$$\hat{\mu}(x_i) = \sum_{t=1}^n w_t \hat{f}_t(x_i), \quad (10)$$

where $\hat{f}_t(x_i)$ denotes the smooth curve estimated from y_t , and w_t are geometrically decreasing weights defined as:

$$w_t = \rho(1 - \rho)^{n-t}, \quad 0 < \rho < 1. \quad (11)$$

Finally, the forecasted Ozone curve $\hat{y}_{n+1}(x_i)$ is obtained by combining the estimated mean function $\hat{\mu}(x_i)$, and the multiplicative sum of forecasted principal component scores $\hat{\beta}_{n+1,p}$ with the estimated eigenfunction $\phi_p(x_i)$ illustrated in the equation below:

$$\hat{y}_{n+1}(x_i) = \hat{f}_{n+1}(x_i) = \hat{\mu}(x_i) + \sum_{p=1}^p \hat{\beta}_{n+1,p} \phi_p(x_i). \quad (12)$$

Following Hyndman and Ullah (2007), each principal component score series $\{\beta_{t,p}\}$ for $p = 4$, were forecasted separately to obtain $\hat{\beta}_{n+1,p}$ using the univariate time series ARIMA model. This is due to the uncorrelated properties of the principal component score. Based on the results of prior analysis shown in Table 3, the number of retained principal components was set to $p = 4$ based on the proportion of variance explained. The first four components accounted for a substantial proportion of the total variability, exceeding 85% across stations, while additional components contributed marginal gains. For consistency and comparability, the same value of p was applied across all monitoring stations and basis function types.

Table 3: Summary of cumulative % of PC variation

Station	Shah Alam		Batu Muda		Klang	
PC	% Variance	Cumulative %	% Variance	Cumulative %	% Variance	Cumulative %
PC1	57.43	57.43	56.59	56.59	60.47	60.47
PC2	16.27	73.70	17.91	74.50	13.40	73.87
PC3	9.13	82.83	8.58	83.08	7.82	81.69
PC4	5.12	87.95	4.23	87.31	5.40	87.09
PC5	3.60	91.55	3.90	91.21	3.38	90.47
PC6	2.37	93.92	2.61	93.82	2.07	92.54
PC7	1.44	95.36	1.72	95.54	1.65	94.19
PC8	1.28	96.64	1.01	96.55	1.52	95.71
PC9	0.96	97.60	0.78	97.33	1.18	96.89
PC10	0.72	98.32	0.71	98.04	0.68	97.57

2.5.1. Multi-step ahead forecasting

In this study, two forecasting strategies were employed, namely direct Multi-step ahead forecasting and Iterative One-step ahead forecasting, within the FTS framework. The key difference between these two approaches is how future curves are forecasted. In the Multi-step ahead approach, separate forecasting models are estimated for each forecast horizon h . Within the FTS framework, the principal component scores are modelled as time series, and distinct models are constructed to directly forecast the scores at each future time point using ARIMA. The predicted scores for each horizon are then used to reconstruct the corresponding functional observations using the fixed basis functions.

For a specified forecast horizon h , the direct multi-step ahead forecast of the future Ozone curve is defined as:

$$\hat{y}_{n+h|n}(x) = E[y_{n+h}(x)|F_n], \quad (13)$$

where F_n denotes the information set available up to time n , including the estimated functional data and principal component structure. This forecast is constructed as:

$$\hat{y}_{n+h|n}(x) = \hat{\mu}(x_i) + \sum_{p=1}^p \hat{\beta}_{n+1,p} \phi_p(x_i), \quad (14)$$

where $\hat{\beta}_{n+h|n,p}$ are the direct h -step ahead forecasts of the principal component scores. In this approach, each score series $\{\hat{\beta}_{t,p}\}$ is modelled using a univariate time series model (ARIMA), and the h -step ahead forecasts are obtained directly from the fitted model, using the anticipated value of the current time step to calculate its value in the following time step without recursion. For example, if $n = 300$ and $h = 3$, the 3-step ahead forecast is given by:

$$\hat{y}_{303|300}(x) = \hat{\mu}(x) + \sum_{p=1}^p \hat{\beta}_{303|300,p} \phi_p(x_i).$$

2.5.2. An iterative one-step ahead forecast or predicting

In this approach, the FPCA scores are modelled as a time series and forecasted using an ARIMA one-step-ahead model. The predicted scores are then recursively used for subsequent forecasts, and the corresponding functional observations are reconstructed using the fixed basis functions.

The Iterative One-step ahead forecast is defined as:

$$\hat{y}_{n+1} = \hat{y}_{n+1|n}(x) = E[y_{n+1}(x)|f(x), B]. \quad (15)$$

For example, if $n = 300$, begin with one-step ahead $n+1 = 301$. To forecast for horizon $h = 3$, the Iterative One-step ahead forecast is illustrated as follows:

Iteration 1:

$$\hat{y}_{301} = \hat{y}_{300+1|300}(x) = E[y_{300+1}(x)|f(x), B] = \hat{\mu}(x) + \sum_{p=1}^p \hat{\beta}_{300+1|300,p} \hat{\phi}_p(x).$$

Iteration 2:

$$\hat{y}_{302} = \hat{y}_{301+1|301}(x) = E[y_{301+1}(x)|f(x), B] = \hat{\mu}(x) + \sum_{p=1}^p \hat{\beta}_{301+1|301,p} \hat{\phi}_p(x).$$

Iteration 3:

$$\hat{y}_{303} = \hat{y}_{302+1|302}(x) = E[y_{302+1}(x)|f(x), B] = \bar{f}(x) + \sum_{p=1}^P \hat{\beta}_{302+1|302,p} \hat{\phi}_p(x).$$

2.6. Performance Indicator

The data used are the discretized points of the actual and predicted values in calculating the performance indicators. The accuracy measurements that can be computed from the residuals of a prediction model in the FDA are explained in this subsection. Functional faults typically rely on the application and may not have a uniform criterion. Functional accuracy is calculated using pointwise Mean Squared Error (MSE), Mean Absolute Error (MAE) and coefficient of determination (R^2). MSE measures the average squared error and heavily penalizes large deviations, whereas MAE measures the average absolute error and treats all deviations equally, making it less sensitive to outliers.

The formula can be written as:

$$MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2, \quad (16)$$

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|, \quad (17)$$

where the R^2 value provides information about the goodness of fit of a model, and measures the variability of Ozone curves predicted by the model. The value indicates how closely the predicted Ozone curve follows the observed curve. The equation can be written as:

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}, \quad (18)$$

where

- Σ is a symbol that means “sum”,
- n is the sample size,
- y_i is the actual data value,
- $\hat{y}_i$ is the predicted data value.

Lower error measures (MSE and MAE) indicate that the model's predictions are closer to the observed values. A higher R^2 indicates that the model explains a larger proportion of the variability in the data. Therefore, in the context of this study, the R^2 is interpreted as a measure of how well the model captures overall patterns. The optimal model is selected by considering both criteria simultaneously, ensuring that the model not only provides accurate predictions but also captures the underlying data patterns effectively.

In addition to MSE, MAE, and R^2 , the Integrated Squared Error (ISE) was also employed to evaluate the accuracy of the predicted ozone concentration curves. ISE measures the total squared difference between the observed curve and the forecasted curve over the entire time domain (Bosq 2000). It is defined as:

$$ISE = \int [y_t(x) - \hat{y}_t(x)]^2 dx \quad (19)$$

In this study, since the Ozone data are observed at discrete hourly time points (24 hours), the ISE is approximated as:

$$ISE \approx \sum_{i=1}^{24} (y_t(x_i) - \hat{y}_t(x_i))^2 \quad (20)$$

The inclusion of ISE helps in adding an important dimension to the evaluation by capturing the cumulative error across the entire curve rather than pointwise deviations alone. The ISE results largely support the findings from MSE and MAE, but also highlight subtle differences in overall curve fitting. For instance, although some models show similar MSE values, their ISE values reveal differences in how well the full trajectory of the data is captured.

3. Results of Analysis

This section discusses the results and findings of this research study. The daily Ozone concentration levels for three selected air monitoring stations involving Shah Alam, Batu Muda, and Klang were analyzed using R programming. Initially, the data underwent a preliminary investigation, which involved describing it using graphs, means, and standard deviations. Subsequently, analyses were conducted to address all study objectives. The results and discussions of these analyses will be elaborated upon in the following sections.

3.1. Descriptive Statistics of Diurnal Ozone

From Table 4, the pointwise descriptive statistics indicate that the mean and median are different within and between stations. The station that has the highest mean and median daily Ozone level is Shah Alam. Klang has the highest median. Based on the median, the average daily diurnal Ozone level is the highest in Klang compared to Shah Alam and Batu Muda. The highest maximum level was observed in Batu Muda. Based on the standard deviation, the Ozone variation is similar at Batu Muda and Klang, with the value of 0.0057 ppm; meanwhile, a larger variation occurred in Shah Alam with 0.0096 ppm.

Table 4: Pointwise summary statistics of diurnal Ozone (ppm)

Air quality monitoring sites	Mean	Median	[Min, Max]	Standard deviation
Shah Alam	0.0071	0.0025	[0.0000, 0.0913]	0.0096
Batu Muda	0.0043	0.0018	[0.0000, 0.1000]	0.0057
Klang	0.0050	0.0029	[0.0000, 0.0401]	0.0057

Figure 2 portrays the summary of functional statistics of the diurnal Ozone levels at the study locations. The descriptive analysis, including (a) the functional description of daily Ozone concentration levels, (b) the mean, and (c) the standard deviation, Shah Alam, Batu Muda and Klang. The functional description graphs for daily Ozone concentration levels across all three basis types exhibit a similar pattern, with a peak between 3 p.m. and 4 p.m. The mean curve fluctuates between 5 a.m. and 9 a.m., reaching its highest point of 0.06 ppm at 3 p.m. The standard deviation curve (a) indicates that the Ozone in Shah Alam has the highest variation, followed by Batu Muda and Klang during the peak time.

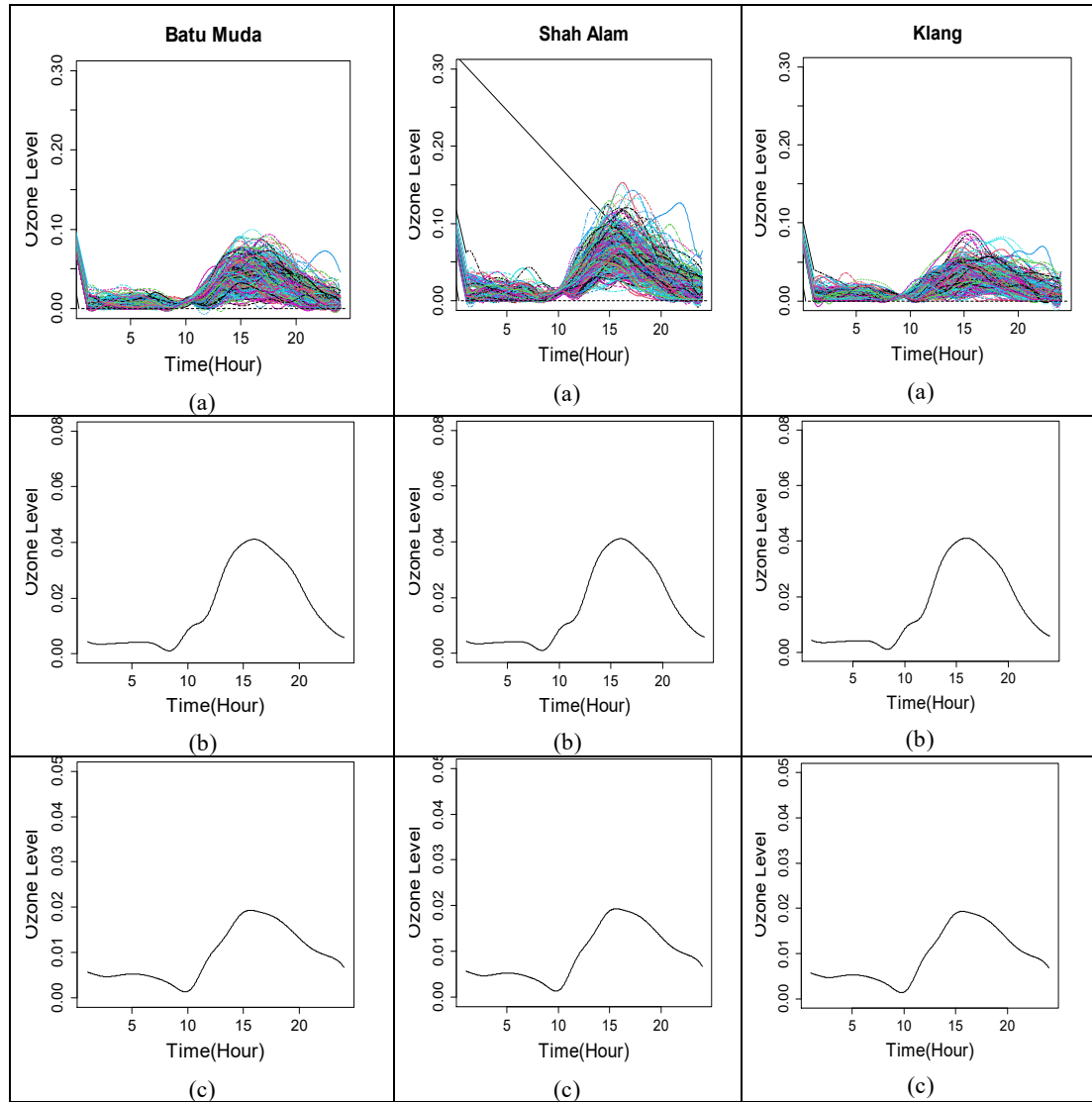


Figure 2: Functional summary statistics of (a) daily ozone concentration, (b) mean and (c) standard deviation curves

3.2. Assessing the Effect of Different Basis Functions on the Forecasting Performance

For experimental purposes and due to data availability, this evaluation period is treated as an independent future forecasting window to ensure temporal ordering is preserved. Although the evaluation horizon is relatively short, with about 10 future day curves, it provides a realistic assessment of short-term predictive performance under unseen conditions (Hyndman & Shang 2009). The computation of the performance indicators was based on 24-hour discretized points of the 10-day actual future curve (short-term unseen curves) (Din & Shaadan 2022), involving a total of 240 tested points.

3.2.1. Multi-step ahead forecasting method

To examine the effect of different basis functions (B-spline, Fourier and Polygonal) on the forecasting performance, the following Figures 3 and 4 illustrate the results of MSE, MAE, ISE and R^2 for the data set at the three different locations.

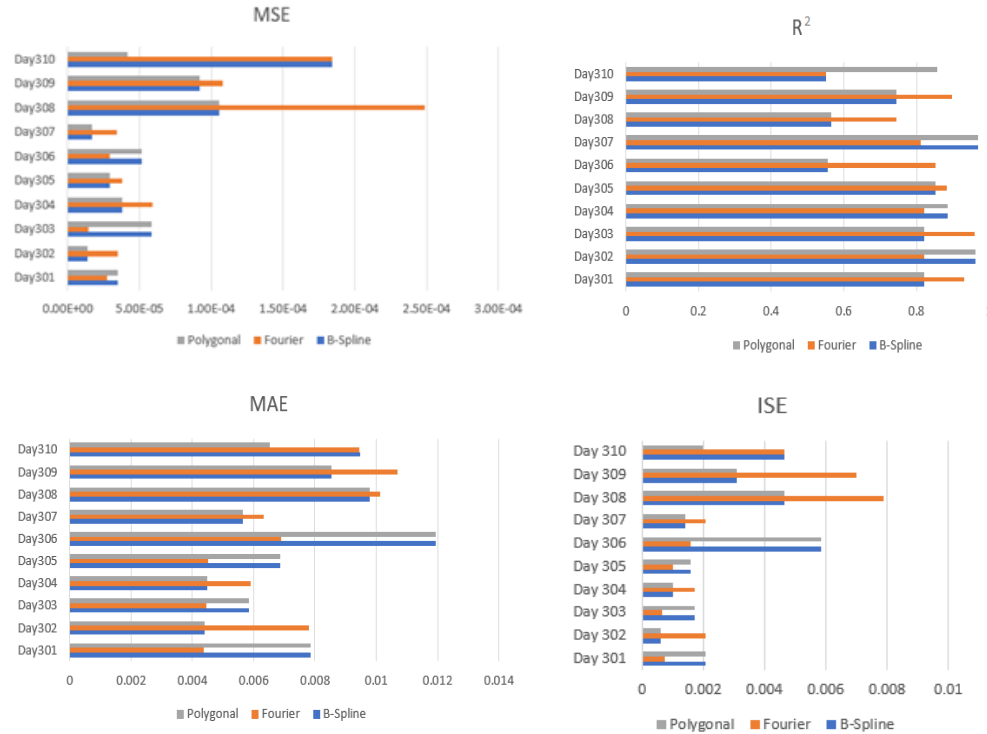
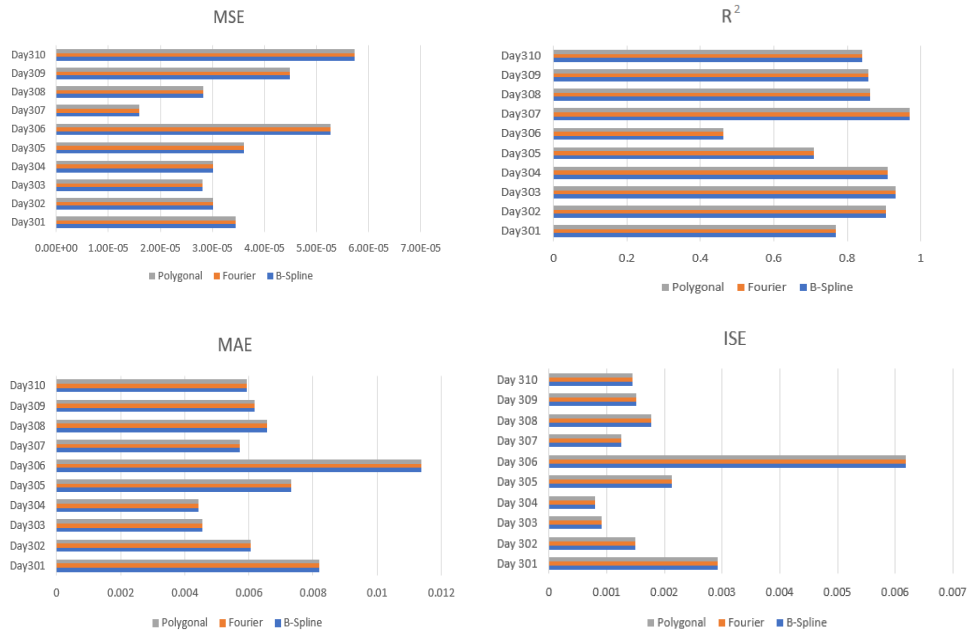
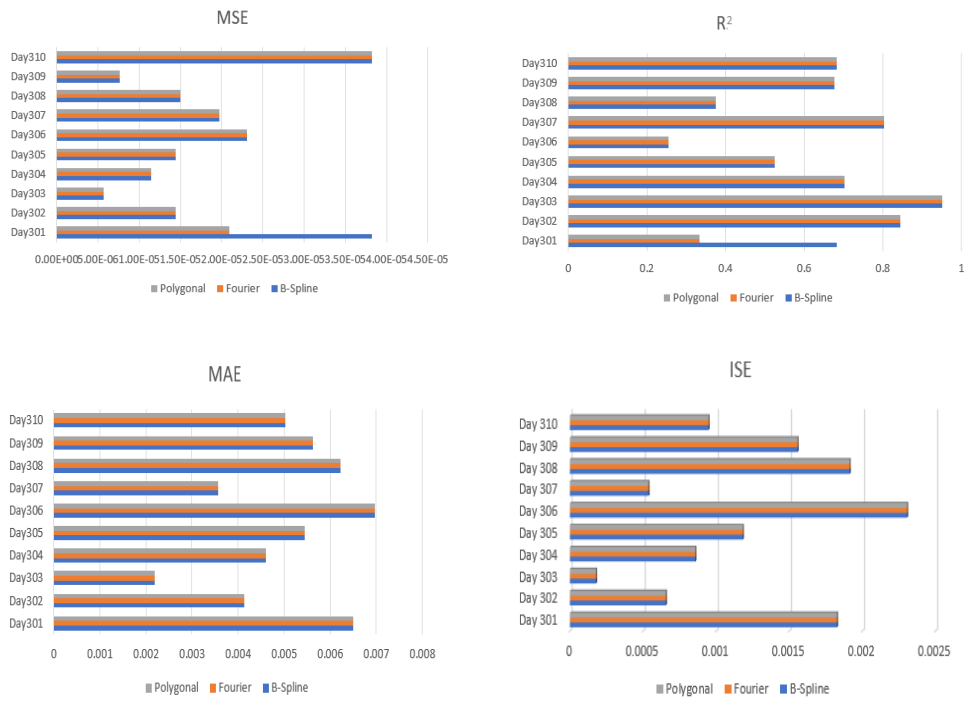


Figure 3: The stacked horizontal-bar-line chart of MSE, MAE and R^2 between the actual versus forecasted curve (with different basis- spline, fourier and polygonal) in Shah Alam using the Multi-step ahead forecast

Larger variation in the values of MSE, MAE and R^2 is observed for Shah Alam (Figure 3) data set for the future Day 301-Day 310 forecast, compared to the data set in Klang and Batu Muda (Figure 4). The Multi-step ahead model of Shah Alam was affected the most by the Polygonal function, but there was a consistent effect of other basis functions on the forecasting performance (MSE, MAE and R^2). The forecasting performance was not affected by the different basis functions for the dataset of other locations. However, there was an abrupt effect of the B-spline for the future Day 301 forecast for the Klang data set.



(a) Batu Muda



(b) Klang

Figure 4: The stacked horizontal-bar-line chart of MSE, MAE and R^2 between the actual versus forecasted curve (with different basis- spline, Fourier, and polygonal) in (a) Batu Muda and (b) Klang using the Multi-step ahead forecast.

3.2.2. Iterative one-step ahead forecasting method

The following Figures 5 and 6 illustrate the results of MSE, MAE and R^2 . The same pattern of basis function effect is observed with larger variation in the values of MSE and MAE. ISE and R^2 is seen for the Shah Alam (Figure 5) data set for the Day 301-Day 310 forecast, compared to the data sets in Klang and Batu Muda (Figure 6). The results indicate that the Iterative One-step ahead method of Shah Alam was affected the most by the Polygonal function. Still, there was a consistent effect of other basis functions on the forecasting performance (MSE, MAE, ISE and R^2) for the datasets at Batu Muda and Klang.

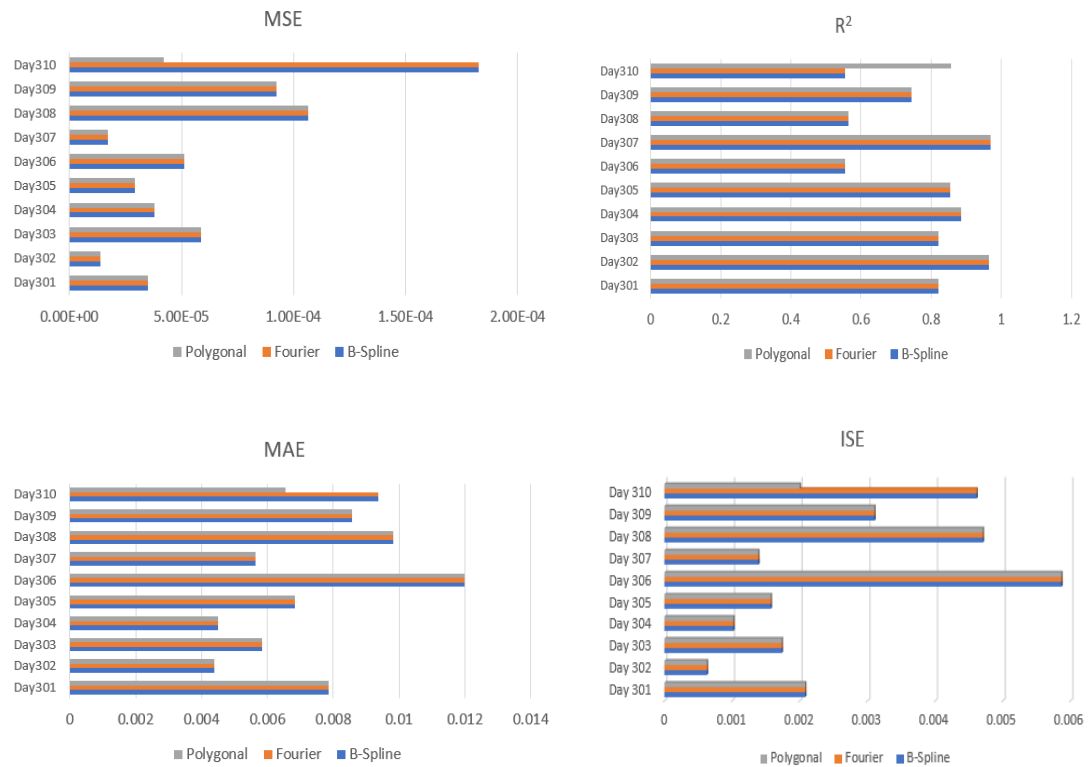
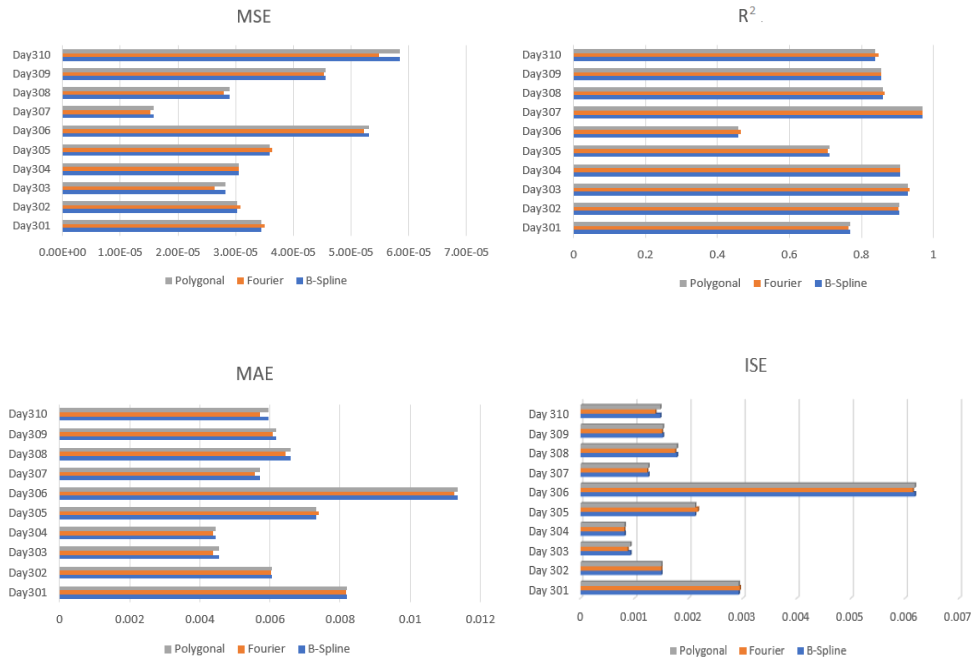


Figure 5: The stacked horizontal-bar-line chart of MSE, MAE and R^2 between the actual versus forecasted curve (with different basis- spline, fourier and polygonal) in Shah Alam using the iterative one-step ahead forecast

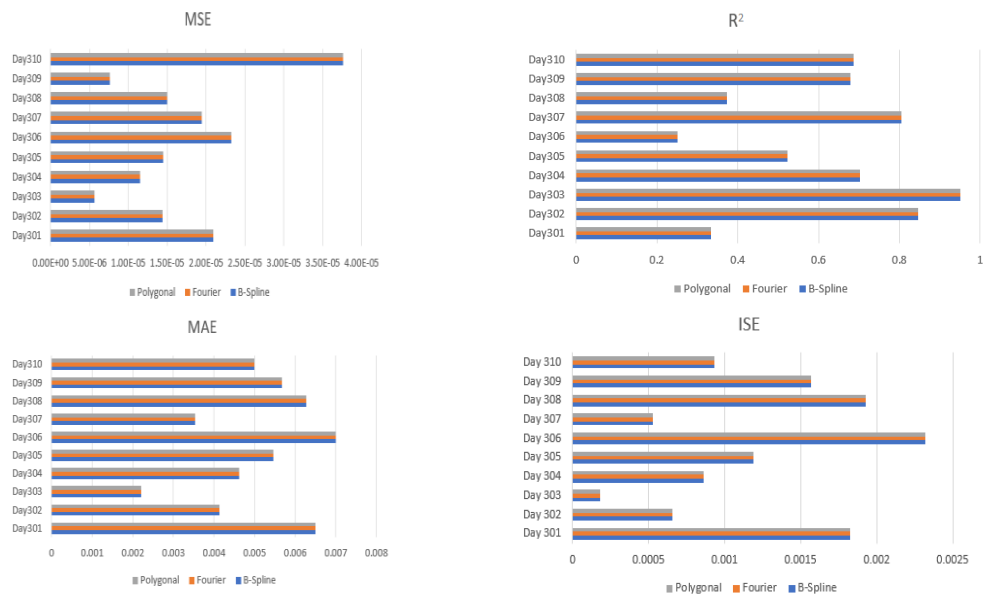
3.2.3. Comparison of Overall Performance Between Multi-step versus Iterative-step Methods

Table 5 below summarizes the overall performance of the Multi-step ahead and Iterative One-step ahead forecasts for the future 10-day curves, Day 301-Day 310 for the all-locations data set. The results provide a comprehensive comparison of forecasting performance across methods, basis functions, and locations using four evaluation metrics, which are MSE, MAE, ISE, and R^2 . The analysis results indicate that variations in the types of basis functions influence forecasting performance across stations. Within the station, data in Shah Alam is shown to be affected by the different basis functions (B-spline, Fourier and Polygonal), with Fourier receiving the lowest MSE and ISE. Meanwhile, the data at other locations only show a marginal effect. Additionally, the results of the coefficient of determination (R^2) show that the model of the Multi-step ahead method yielded a better performance compared to the Iterative One-step ahead method. This is shown by the overall R^2 values of the forecasting performance

using the Multi-step ahead method ($R^2 = 0.8017, 0.8214, 0.6269$) at each respective location. The values are higher than the Iterative One-step ahead method ($R^2 = 0.7831, 0.8208, 0.6154$).



(a) Batu Muda



(b) Klang

Figure 6: The stacked horizontal-bar-line chart of MSE, MAE and R^2 between the actual versus forecasted curve (with different basis- spline, fourier and polygonal) in (a) Batu Muda and (b) Klang, using the iterative one-step ahead forecast

Table 5: Assessing the overall performance of multi-step ahead and Iterative one-step ahead methods

Method	Performance Indicator	Location	B-Spline	Fourier	Polygonal	Overall mean
Multi-step ahead	MSE	Shah Alam	6.24E-05	7.78E-05	4.82E-05	6.28E-05
		Batu Muda	3.58E-05	3.58E-05	3.58E-05	3.58E-05
		Klang	1.88E-05	1.71E-05	1.71E-05	1.76E-05
	MAE	Shah Alam	7.48E-03	7.06E-03	7.19E-03	7.24E-03
		Batu Muda	6.63E-03	6.63E-03	6.63E-03	6.63E-03
		Klang	5.03E-03	5.03E-03	5.03E-03	5.03E-03
	ISE	Shah Alam	2.66E-02	2.93E-02	2.40E-02	2.66E-02
		Batu Muda	2.04E-02	2.04E-02	2.04E-02	2.04E-02
		Klang	1.20E-02	1.20E-02	1.20E-02	1.20E-02
	R-squared	Shah Alam	0.7733	0.8278	0.8040	0.8017
		Batu Muda	0.8214	0.8214	0.8214	0.8214
		Klang	0.6500	0.6153	0.6153	0.6269
Iterative One-step ahead	MSE	Shah Alam	6.25E-05	6.25E-05	4.84E-05	5.78E-05
		Batu Muda	3.61E-05	3.55E-05	3.61E-05	3.59E-05
		Klang	1.70E-05	1.70E-05	1.70E-05	1.70E-05
	MAE	Shah Alam	7.48E-03	7.48E-03	7.20E-03	7.39E-03
		Batu Muda	6.63E-03	6.54E-03	6.63E-03	6.60E-03
		Klang	5.04E-03	5.04E-03	5.04E-03	5.04E-03
	ISE	Shah Alam	2.67E-02	2.67E-02	2.41E-02	2.58E-02
		Batu Muda	2.05E-02	2.03E-02	2.05E-02	2.04E-02
		Klang	1.20E-02	1.20E-02	1.20E-02	1.20E-02
	R-squared	Shah Alam	0.7731	0.7731	0.8032	0.7831
		Batu Muda	0.8202	0.8220	0.8202	0.8208
		Klang	0.6154	0.6154	0.6154	0.6154

Figure 7 illustrates the physical form of the actual versus predicted curves for the first nine days next curves at Shah Alam station (i.e., day 301-309). This visual comparison allows assessment of how well the model reproduces the overall diurnal shape, including peak timing and magnitude. A close alignment between the forecasted and observed curves indicates good model performance, while discrepancies highlight periods where the model underestimates or overestimates Ozone levels. While quantitative metrics such as MSE, MAE, ISE and R^2 summarize forecasting accuracy, curve comparisons provide additional insight into pattern consistency and structural fidelity. Highlight is given to the agreement between observed and forecasted peak Ozone levels, as accurate peak representation is critical for air quality assessment.

The results show that the method accurately forecasts the overall curve and captures the Ozone peak around 3:00 p.m. for most days, except on days with substantially shifted peak timing. This finding opens a gap for future enhancements that may focus on improving adaptability to atypical peak timing by incorporating peak-related features, flexible basis functions, or additional explanatory variables.

4. Conclusions

This study applied the FDA approach to analyze Ozone concentration levels and visualize, evaluate, and forecast daily (diurnal) Ozone levels for the future 10 days ahead using the FTS model. Specifically, the study aims to compare the effect of different basis functions, including B-spline, Fourier, and Polygonal, on the forecasting performance of two existing forecasting model methods: the Multi-step and the Iterative One-step ahead. The predictive performance of Ozone curves depends on the data variation of different data sets while maintaining the forecasted curve pattern. Based on the R^2 , the Multi-step ahead approach showed modestly

better performance than the Iterative One-step ahead method for this dataset; however, this finding should be interpreted cautiously and does not imply universal superiority across functional time series applications. Several limitations should be acknowledged, including the relatively short 10-day test horizon, the absence of explicit modelling of seasonal or meteorological factors, the relatively high basis dimension compared to the 24-hourly observations per curve, and the absence of formal statistical testing to determine the significance of performance differences. Future research may consider longer evaluation periods, the inclusion of seasonal and meteorological covariate effects, more parsimonious basis representations, and statistical significance testing, as well as extending the framework to other air pollutants. These findings are supported and in line with the research of Din and Shaadan (2022), whereby in their study, it was found that the Multi-step ahead method outperformed the Iterative One-step ahead method in forecasting monthly diurnal maximum API curves. For future research, it is recommended to explore additional types of basis functions and extend the application to other air pollutants while also addressing seasonal effects that were not explicitly modeled in this study with a larger dataset.

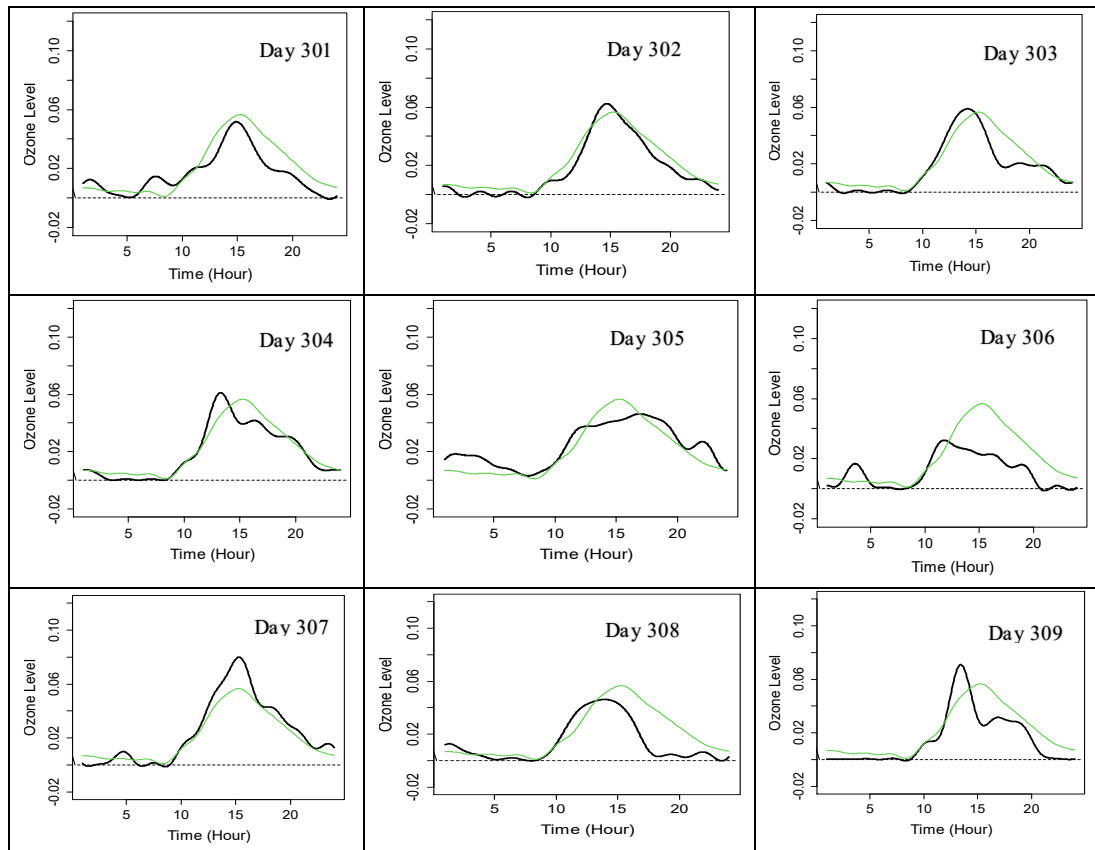


Figure 7: The comparison between the actual (black) and forecasted curve (green) of the Multi-step ahead method at Shah Alam station for an illustration

Acknowledgement

The authors wish to thank the Department of Environment (DOE), Putrajaya, Malaysia, for providing the data utilized in this study, Universiti Teknologi MARA and the Institute of Statistics Malaysia for the support. Not to forget, big thanks to all the reviewers.

References

- Acuna E. & Rodriguez C. 2004. The treatment of missing values and its effect in the classifier accuracy. In, Banks D., House L., McMorris F.R. Arabie P. & Gaul W. (eds.). *Classification, Clustering and Data Mining Applications*: 639-647. Berlin Heidelberg: Springer.
- Ahmad A.N., Abdullah S., Mansor A.A., Dom N.C., Ahmed A.N., Ismail N.A. & Ismail M. 2023. Assessment of daytime and nighttime ground level ozone pollution in Malaysian urban areas. *Malaysian Journal of Medicine and Health Sciences* **19**(6): 242-248.
- Bosq D. 2000. *Linear Processes in Function Spaces: Theory and Applications*. New York: Springer.
- Cao J., Liu J., Cheng Y., Ai S., Li F., Xue T., Zhang Q. & Zhu T. 2024. Impacts of different vehicle emissions on ozone levels in Beijing: Insights into source contributions and formation processes. *Environment International* **191**: 1–11.
- De Marco A., Garcia-Gomez H., Collalti A., Khaniabadi Y.O., Feng Z., Proietti C., Sicard P., Vitale M., Anav A. & Paoletti E. 2022. Ozone modelling and mapping for risk assessment: an overview of different approaches for human and ecosystems health. *Environmental Research* **211**: 1–11.
- Din W.W. & Shaadan N. 2022. Application of functional time series model in forecasting monthly diurnal maximum API curves: A comparison between multi-step ahead and iterative one-step ahead approach. *Malaysian Journal of Fundamental and Applied Sciences* **18**: 124-137.
- Hyndman R.J. & Shang, H.L. 2009. Forecasting functional time series. *Journal of the Korean Statistical Society* **38**(3): 199-211.
- Hyndman R.J. & Ullah M.S. 2007. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis* **51**(10): 4942-4956.
- Merz B., Kuhlcke C., Kunz M., Pittore M., Babeyko A., Bresch D.N., Domeisen D.I.V, Feser F., Koszalka I. & Kreibich H. (2020). Impact forecasting to support emergency management of natural hazards. *Reviews of Geophysics* **58**(4): 1–52.
- Mohd Ali, M.F., Syed Mohamad, S.F., Mohammad Yusof, A.S., Ismail, S. & Ibrahim, N.A. 2022. A GARCH study on exchange rate determinants: A case of Malaysia. *Journal of Statistical Modeling and Analytics (JOSMA)* **4**(1): 72-84.
- Naji A., Ismail A., Kamrani E. & Sohrabi T. 2014. Correlation of MT levels in livers and gills with heavy metals in wild tilapia (*Oreochromis mossambicus*) from the Klang River, Malaysia. *Bulletin of Environmental Contamination and Toxicology* **92**(6):674-679.
- Ramsay J.O. & Silverman B.W. 2005. *Functional data analysis*. New York: Springer New York.
- Schuster N.A., Rijnhart J.J.M., Twisk J.W.R. & Heymans M.W. 2022. Modeling non-linear relationships in epidemiological data: The application and interpretation of spline models. *Frontiers in Epidemiology* **2**: 1-10.
- Shang H.L., Smith P.W., Bijak J. & Wisniowski A. 2021. A multilevel functional data method for forecasting population, with an application to the United Kingdom. *International Journal of Forecasting* **32**(3) :629-649.
- Tabachnick B.G. & Fidell L.S. 2007. *Using multivariate statistics*. 5th Ed. Boston: Pearson Education, Inc.
- Xie J., Tang X., Zheng F., Wang X., Ding N., Song Y., Luo H., Wu H., Kong L. & Zhu J. 2024. Improvement of the ozone forecast over Beijing through combining the chemical transport model with multiple machine learning methods. *Atmospheric Pollution Research* **15**(8): 1–9.

Center of Mathematical Sciences

Faculty of Computer and Mathematical Sciences

Universiti Teknologi MARA

40450, Shah Alam

Selangor, MALAYSIA

E-mail: rosrosslan@gmail.com, shahida@tmsk.uitm.edu.my*, sayang@tmsk.uitm.edu.my

Received: 27 July 2025

Accepted: 13 June 2026

*Corresponding author