agriculture

MDPI

Article

Prediction of Carbon Dioxide Concentrations in Strawberry Greenhouse by Using Time Series Models

Seung Hyun Shin [1], Nibas Chandra Deb [2], Elanchezhian Arulmozhi [2], Niraj Tamrakar [2], Oluwasegun Moses Ogundele [2], Junghoo Kook [1], Dae Hyun Kim [3] and Hyeon Tae Kim [2,*]

[1] Department of Smart Farm, Institute of Smart Farm, Gyeongsang National University, Jinju 52828, Republic of Korea
[2] Department of Biosystems Engineering, Institute of Smart Farm, Gyeongsang National University, Jinju 52828, Republic of Korea
[3] Department of Biosystems Engineering, Kangwon National University, Chuncheon 24341, Republic of Korea
* Correspondence: bioani@gnu.ac.kr; Tel.: +82-55-772-1896

Abstract: Carbon dioxide (CO_2) concentrations play an important role in plant production, as they have a direct impact on both plant growth and yield. Therefore, the objectives of this study were to predict CO_2 concentrations in the greenhouse by applying time series models using five datasets. To estimate the CO_2 concentrations, this study was conducted over a four-month period from 1 December 2023 to 31 March 2024, in a strawberry-cultivating greenhouse. Fifteen sensors (MCH-383SD, Lutron, Taiwan) were installed inside the greenhouse to measure CO_2 concentration at 1-min intervals. Finally, the dataset was transformed into intervals of 1, 5, 10, 30, and 60 min. The time-series data were analyzed using the autoregressive integrated moving average (ARIMA) and the Prophet Forecasting Model (PFM), with performance assessed through root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). The evaluation indicated that the best model performance was achieved with data collected at 1-min intervals, while model performance declined with longer intervals, with the lowest performance observed at 60-min intervals. Specifically, the ARIMA model outperformed across all data collection intervals while comparing with the PFM. The ARIMA model, with data collected at 1-min intervals, achieved an R^2 of 0.928, RMSE of 7.359, and MAE of 2.832. However, both ARIMA and PFM exhibited poorer performances as the interval of data collection increased, with the lowest performance at 60-min intervals where ARIMA had an R^2 of 0.762, RMSE of 19.469, and MAE of 11.48. This research underscores the importance of frequent data collection for precise environmental control in greenhouse agriculture, emphasizing the critical role of short-interval data collection for accurate predictive modeling.

Keywords: ARIMA model; carbon dioxide; Prophet Forecasting Model; strawberry

Citation: Shin, S.H.; Deb, N.C.; Arulmozhi, E.; Tamrakar, N.; Ogundele, O.M.; Kook, J.; Kim, D.H.; Kim, H.T. Prediction of Carbon Dioxide Concentrations in Strawberry Greenhouse by Using Time Series Models. *Agriculture* **2024**, *14*, 1895. https://doi.org/10.3390/agriculture14111895

Academic Editors: Weiwei Chen, Qiuju Xie and Li Guo

Received: 12 September 2024
Revised: 21 October 2024
Accepted: 24 October 2024
Published: 25 October 2024

1. Introduction

Strawberries are widely cultivated around the world for their rich vitamin content, antioxidant properties, and health benefits [1]. They are also highly valued by consumers for their attractive red color, distinct aroma, and sweet taste [2]. Strawberries contain high levels of vitamin C, minerals, flavonoids, and phytochemicals, and are recognized as a functional food with numerous health-enhancing properties [3]. Additionally, the phytochemical profile of strawberries includes phenolic compounds and vitamins, contributing to their antioxidant capacity and health benefits, which has led to a significant increase in consumer demand over the past 20 years [4]. Research has shown that strawberries, rich in antioxidants and bioactive compounds, can significantly lower fasting blood glucose levels in individuals diagnosed with type 2 diabetes [5]. For these reasons, strawberries are one of the most extensively studied fruits from agronomic, genetic, and nutritional perspectives [6].

Monitoring and managing CO_2 concentration are therefore vital for ensuring optimal conditions for strawberry growth and productivity in greenhouse environments [7]. The intervals for collecting CO_2 concentration data vary significantly from study to study, ranging from weekly, hourly, to even minute intervals [8]. Sokolov, S.V. [9] have emphasized the importance of long-term measurement of CO_2 concentrations in greenhouses, highlighting the critical role of controlling these levels. This body of research underlines the importance of accurately collecting and analyzing environmental data, such as CO_2 concentration. The diversity of sampling intervals and methodologies emphasizes the complexity and variability in atmospheric CO_2 concentration monitoring. Accurate and timely environmental data are essential for smart farm managers to make informed decisions regarding crop management [10].

Despite extensive research on CO_2, studies aimed at determining the optimal intervals for collecting CO_2 concentration data remain scarce. Despite technological advancements, the data collection intervals in existing smart farm systems still require significant improvement. If data are collected too infrequently, critical environmental changes may not be detected in time, preventing timely responses. Conversely, collecting data too frequently can lead to an accumulation of unnecessary data, consuming substantial resources in storage and processing and leading to inefficient use of resources. Therefore, setting appropriate data collection intervals is crucial for efficient resource use and accurate detection of environmental changes. However, the current data collection cycles in existing smart farm systems often have various issues. Existing smart farm systems face challenges with data collection cycles ranging from infrequent to excessively frequent data collection, leading to resource wastage or overlooking significant environmental changes. Small- to medium-sized farms often employ customized data collection methods, focusing on key observations rather than adopting formal systems, which hinders the adoption of Farm Management Information Systems (FMISs) [11]. The decision on data collection intervals plays a crucial role in optimizing data collection strategies. For example, the time of day when temperatures change the most typically occurs during the transition between day and night [12,13]. Furthermore, elucidating the relationship between temperature and CO_2 levels is imperative for optimal farm management. Fluctuations in temperature can markedly affect atmospheric CO_2 concentrations, which, in turn, influence plant growth and soil health [14]. Elevated temperatures during daylight hours can enhance the rate of CO_2 uptake by plants via photosynthesis. In contrast, reduced temperatures at night decelerate this process [15]. By synchronizing data collection with these pivotal environmental interactions, farmers can more precisely evaluate the effects of climatic conditions on crop productivity and make more informed decisions regarding resource allocation. This necessitates studies to decide the intervals at which to collect data efficiently. Therefore, predicting CO_2 concentrations at different collection intervals is crucial for guaranteeing the optimal growth of plants.

Additionally, in this study, the ARIMA model and PFM were utilized to predict the concentration of CO_2 within the greenhouse, and the predictive performance of each model was derived to determine the optimal data collection interval. Time series model analyses such as the ARIMA model and PFM were used to find the optimal data collection intervals [16]. Considering their excellent ability to analyze time series data, these models have become increasingly indispensable in the agricultural sector over the past 20 years. They are primarily used to meet the critical need for precise forecasting of crop yields, market prices, and environmental conditions, which are essential for effective farm management and planning [17]. The PFM was proposed by Desai and Shingala [18] as a forecasting model for wheat yield predictions, achieving high accuracy through the use of the FB PFM algorithm. These models predict CO_2 concentration data to identify patterns, trends, and seasonality. Previous studies [19] have evaluated the performance of ARIMA and PFM in comparison with other time series forecasting models, demonstrating that ARIMA can achieve higher accuracy in short-term predictions for agricultural data when compared to deep learning models such as LSTM. Similarly, M'barek et al. [20] compared PFM with

LSTM-based models and found that PFM performs better on shorter time series data. The results of such predictions provide crucial criteria for determining when to schedule subsequent data collections. For instance, during periods with strong patterns or seasonality, data collection can be increased to capture fluctuations more accurately. Conversely, when trends are stable, the collection intervals can be extended to use resources more efficiently. Furthermore, the ARIMA model or PFM can be applied to the data to accurately predict future CO_2 concentrations in the greenhouse, providing essential information for greenhouse management. Moreover, the results of time series analysis can play a significant role in establishing such strategies. It is ideal to perform time series analysis on collected data to continuously adjust and optimize the sampling strategy based on the analysis results and engage in an iterative process.

According to the existing literature, there was a knowledge gap in utilizing ML models to predict CO_2 concentrations at various collection intervals [21–23]. Consequently, the current study aims to construct an optimal model that is suitable for short-term predictions and has low complexity. This research uses the ARIMA model and PFM to forecast CO_2 concentrations. Statistical time-series forecasting models increase efficiency in data processing and variable selection through the use of appropriate input parameters, which is crucial for enhancing prediction accuracy. Moreover, selecting suitable parameters maximizes the model performance and ensures an efficient learning process. The principal aims of this research are as follows:

(1) By building time series forecasting models for CO_2 concentration predictions, optimal hyperparameters with reduced complexity are selected to accommodate various collection intervals (1-min, 5-min, 10-min, 30-min, and 60-min intervals);
(2) Comparing the accuracy of all models using 5 different datasets: (1) dataset collected at 1-min intervals, (2) dataset collected at 5-min intervals, (3) dataset collected at 10-min intervals, (4) dataset collected at 30-min intervals, and (5) dataset collected at 60-min intervals;
(3) Comparing the performance of two-time series forecasting models for predicting CO_2 concentrations.

2. Materials and Methods

The overall flowchart of the research study method is shown in Figure 1. This flowchart illustrates the process of collecting CO_2 in a greenhouse and predicting future concentrations using time series models. Initially, sensors are installed to measure and record CO_2 concentrations, and the collected data undergo preprocessing to ensure suitability for analysis. Subsequently, time series models are applied based on historical data to forecast future CO_2 concentrations, involving steps like feature selection and model training. Finally, the predicted data are analyzed to identify trends, and environmental management strategies are formulated based on these insights.

2.1. Experimental Greenhouse Features

The research was conducted in a greenhouse, located at the smart farm research center of Gyeongsang National University, Gyeongsangnam-do, Republic of Korea. A flat arch-shaped UV-resistant greenhouse, featuring a 2-layer polyethylene covering with thicknesses of 0.1 mm and 0.075 mm, was used for this research [23]. The greenhouse dimensions are 7.7 m in width, 19.7 m in length, and 3.6 m in height. The Gyeongsangnam-do province's climatic condition is more suitable for strawberry cultivation and production. Specifically, the province encompasses 2280 hectares of cultivated land managed by 4739 farming households, producing a total of 67,762 tons of strawberries annually.

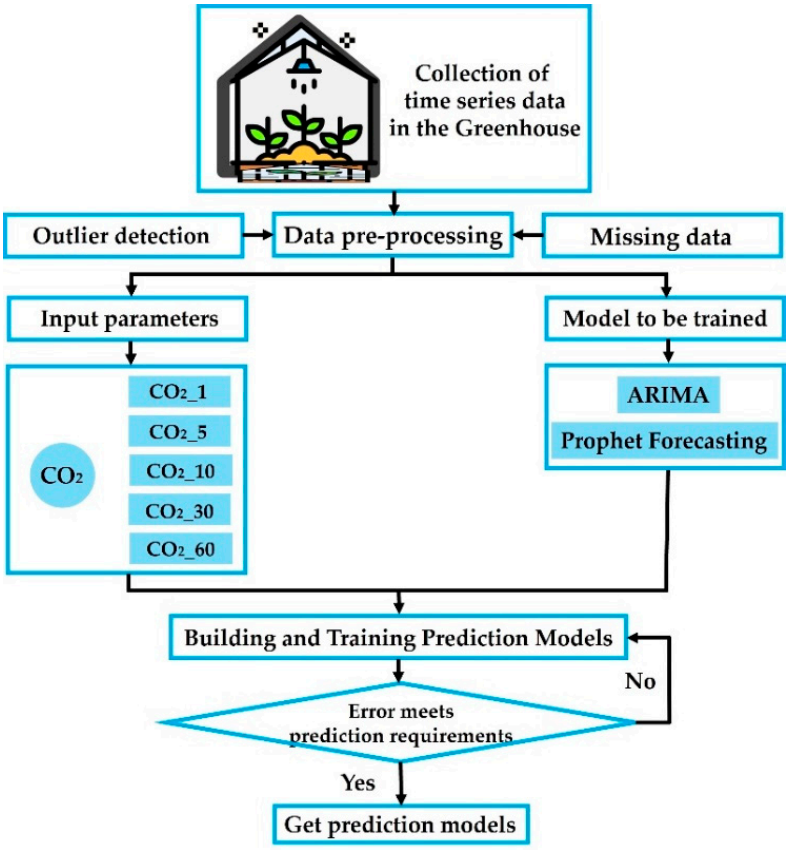


Figure 1. The flowchart of the study.

2.2. Experimental Design and Data Collection

The experiment was conducted between 1 December 2023 to 31 March 2024 inside an experimental greenhouse. A total of 15 sensors (MCH-383SD, Taipei, Taiwan) were installed for this study. For each bed, three sensors were placed—one at each end and one in the middle to monitor the environmental data every day. According to the information provided by the sensor company, the sensor has an accuracy of 95 ± 5%. The experimental greenhouse and CO_2 sensors are shown in Figure 2b. This sensor is capable of collecting data on temperature, humidity, and CO_2 concentration. However, in this study, only the CO_2 concentration data were utilized. The time-series dataset was collected every minute, comprising records of CO_2. All data loggers and electronic devices were adjusted to reduce any potential instrument errors prior to the experiment. The crops grown inside the greenhouse were the Seolhyang strawberries (*Fragaria* × *annanassa Duch.*), which are popular and commonly cultivated in South Korea. Thirty-day-old Seolhyang strawberry seedlings, uniform in size and vigor, were planted. A total of 500 plants were planted in rows, 100 per row, as shown in Figure 2a. The spacing between the plants was maintained at 0.2 m, ensuring sufficient sunlight and air for the plants. Additionally, a mixture of BioPlus compost (BIOPLUS CO., Ltd., Seoul, Republic of Korea) which includes coconut waste and other biodegradable materials, along with Hoagland solution, was applied to each row. Specifically, the electrical conductivity (EC) of the Hoagland solution was maintained at a stable level of 1.5 mS/cm for all treatments. Moreover, the BioPlus compost soil comprised cocopeat (68.86%), perlite (11.00%), peat moss (11.00%), and zeolite (9.00%) [24]. During the initial stages of growth, 20–30 mL of irrigation water was applied daily to each plant, and during the overall ripening stage, 30–50 mL was applied. In this study, we employed a simple drip irrigation system that was positioned well away from the sensors and covered with plastic, ensuring that irrigation did not affect the sensor readings.

(a) (b)

Figure 2. (**a**) The strawberry experiment in the vinyl greenhouse (VGH), (**b**) carbon dioxide sensor (MCH-383SD, Taipei, Taiwan) used in this study.

2.3. Preprocessing of the Dataset for Prediction Models

The time-series dataset contained 2,613,600 records. The raw data could include outliers, missing values, or inconsistent data [25]. Prior to applying the time-series data to a single model, data preprocessing was conducted. Initially, missing data were addressed through linear interpolation [26]. Outliers were identified and removed using an Isolation Forest [27]. After data preprocessing, the collection intervals of the originally collected data, which were at 1-min intervals, were transformed into 10-, 30-, and 60-min intervals for time-series analysis of each interval (Table 1).

Table 1. Recorded CO_2 data counts based on collection intervals.

Interval (min)	Dataset	Number of Records	Number of Outliers
1	CO_2_1	2,613,600	81,021
5	CO_2_5	522,720	15,158
10	CO_2_10	261,360	6534
30	CO_2_30	87,120	2003
60	CO_2_60	43,560	958

2.4. Model Implementation

2.4.1. Modeling of ARIMA

The ARIMA model was developed in 1970 by George Box and Gwilyn Jenkins, which is commonly known as the Box–Jenkins method [28]. ARIMA is a forecasting technique well-suited for analyzing data that are interrelated over time, excluding the influence of independent variables [29]. This model specifically leverages features of time series data, such as autocorrelation, trends, and seasonality, to achieve high accuracy in short-term predictions. The flowchart of the ARIMA model is shown in Figure 3a. The ARIMA model operates under the assumption that the time series data are in a stationary state, meaning that the data's mean and variance are constant over time [30]. If the data are non-stationary, they must be converted into a stationary state through a differencing process. The Augmented Dickey Fuller (ADF) test is effective in identifying the presence of a unit root in the series, assisting in determining whether the series is stationary [31].

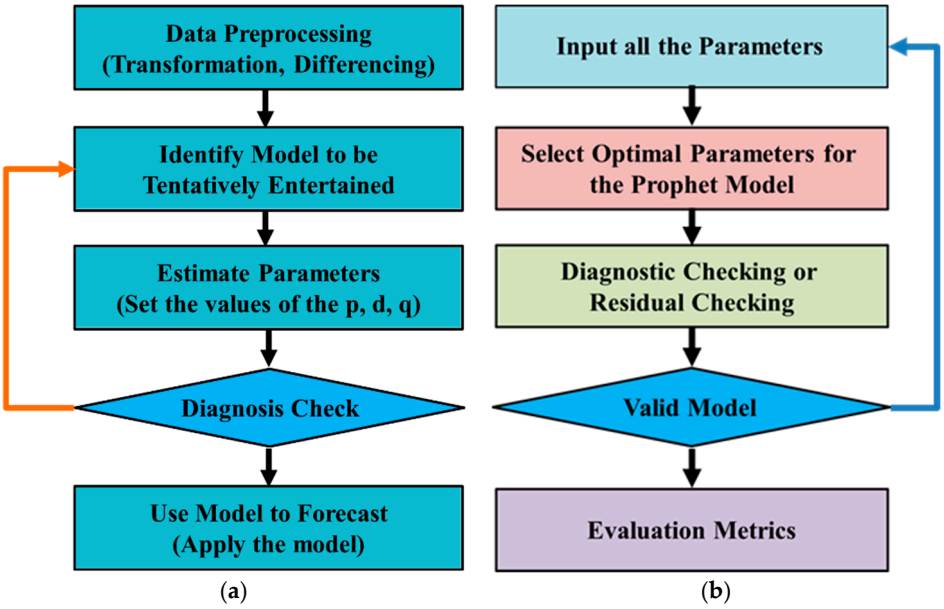


Figure 3. Flowchart: (**a**) Box–Jenkins Methodology; (**b**) PFM.

In the autoregressive integrated (AR) part of the model, the influence of previous values on the present value is modeled. Here, the tilde $\varnothing_1$, $\varnothing_2$, …, $\varnothing_p$ represents the coefficients indicating the influence of these past values, and the at symbol c denotes the model's constant term [32]. This section captures the autocorrelation in time series data, playing a crucial role in predicting current values based on past data patterns. The moving average (MA) part of the model explains how the prediction errors from the time series affect the current value [33]. The theta coefficients θ_1, θ_2, … , θ_q represent the coefficients that indicate the influence of these past prediction errors. This component models the impact of random shocks in time series data and is useful for capturing the 'noise' aspect of the data. The backshift operator B is used in time series data to shift observations back in time [34]. This operator integrates the autoregressive part, differencing, and moving average components of the ARIMA model, allowing it to handle non-stationarities in the time series and model the effects of autocorrelation and random shocks simultaneously. The variable d represents the number of differencing operations required to stabilize the data, helping the time series achieve a constant mean and variance. The terms AR, MA, and the backshift operator can be represented by the following equations (Equations (1)–(3)) [35], as follows:

$$y_t = c + \varnothing_1 y_{t-1} + \varnothing_2 y_{t-2} + \ldots + \varnothing_p y_{t-p} + \epsilon_t \quad (1)$$

$$y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (2)$$

$$\left(1 - \varnothing_1 B - \varnothing_2 B^2 - \ldots - \varnothing_p B^p\right)(1 - B)^d y_t = C + \left(1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q\right)\epsilon_t \quad (3)$$

To build a model, it is essential to accurately set parameters that match the characteristics of the dataset. This defines the type of ARIMA model, and each parameter is adjusted to reflect specific patterns in the time series. To implement an ARIMA model, specific parameters of the dataset (p, d, q) are required.

These parameters were determined by analyzing the data's Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). ACF measures the correlation between different time points within the time series data, indicating the linear relationships between data points at specific lags. In contrast, PACF shows the pure correlation between these time points, excluding the influence of earlier lags.

After determining the model's parameters, two criteria, AIC and BIC, can be used to select the most suitable model. AIC measures the balance between model complexity and goodness of fit to the data, aiming to minimize overall information loss while including

a penalty for model complexity. BIC uses a similar approach but tends to favor simpler models by imposing a greater penalty based on sample size. The AIC and BIC values for each ARIMA model can be plotted graphically, with model complexity (e.g., order of AR and MA) on the *x*-axis and AIC or BIC values on the *y*-axis, to see how these values change with model complexity. These two criteria are used competitively to enhance the predictive performance of the model.

By utilizing these statistical tools and criteria, the most suitable ARIMA model for the collected data was selected, enabling more accurate predictions. Accurate parameter setting and appropriate model selection are crucial for effectively modeling complex time series patterns and predicting future data points. To select the model, we first conducted an analysis of the ACF and the PACF to understand the data's autocorrelation. Based on these results, we determined the order (p, d, q) of the ARIMA model. Among several candidate models, we chose the one with the lowest Akaike Information Criterion (AIC) value, thereby enhancing the accuracy of our predictions.

2.4.2. Modeling of the Prophet

PFM is a decomposable time series forecasting model created by Facebook, which relies on an additive approach that incorporates trend, seasonal, and holiday components [35]. The PFM is powerful yet user-friendly, providing fast and accurate forecasts for data that change over time [36,37]. In this study, the default settings of the model were employed, which is obtained from the study of Toharudin et al. [38]. The flowchart of PFM is presented in Figure 3b. The Prophet algorithm's time series is broken down as illustrated in the Formula (4) [39].

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (4)$$

In this formula, g(t) is used to represent the long-term trend of the time series. s(t) indicates the periodic changes in the time series. h(t) models some of the irregular variations in the time series. ϵ_t represents the residual volatility that the model fails to predict, encompassing not only forecasting errors but also other types of noise and patterns not captured by the model [40]. This includes unmodeled influences, measurement errors, and any intrinsic data variations that are not accounted for within the model's parameters.

In the PFM, g(t), s(t), and h(t) are implemented using the following principles:

- Principle of g(t): The Prophet implements trend components in two ways—using either the Saturating Growth Model or the Piecewise Linear Model;
- Principle of s(t): The Prophet applies Fourier series to capture periodicity, allowing the model to flexibly account for recurring patterns;
- Principle of h(t): In the Prophet model, each holiday is treated as an independent component, with separate dummy variables assigned for each holiday.

2.5. Model Performance Metrics

Various performance metrics were used to evaluate the performance of the model based on the data collection intervals. The coefficient of determination (R^2) measures the closeness between actual data and predicted values. The Root Mean Square Error (RMSE) is calculated as the square root of the average of the squared differences between predicted values and actual values. The Mean Absolute Error (MAE) calculates the average of the absolute differences between predicted values and actual values. Also, the MAE provides an intuitive understanding of the average magnitude of errors and has the advantage of being less affected by outliers. For a comprehensive evaluation, the model's results were analyzed with the coefficients of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) [23], which are defined by equations (Equations (5)–(7)) [41].

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} |y_i - p_i|^2}{n}} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - p_i)}{\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\right)\sum_{i=1}^{n} y_i} \quad (6)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - p_i|}{n} \quad (7)$$

2.6. Modeling Analysis Software

The data collected in this study were analyzed using the Jupyter Notebook environment (Python version 3.8.0). Python is a versatile programming language that plays a critical role in data analysis across numerous research disciplines due to its robust libraries and frameworks [42]. Additionally, the data were visualized using the OriginPro software package (version 9.5.5, OriginLab, Northampton, MA, USA).

3. Results

3.1. Microclimate of Experimental Greenhouse

The variations in temperature, relative humidity, and CO_2 concentration inside the greenhouse were analyzed during the experimental period. The ranges of temperature, relative humidity, and CO_2 were 4.29–35.50 °C, 7.6–98.1%, and 356.92–596 ppm, respectively, the in experimental period. It was observed that the relation between temperature, CO_2 concentration, and humidity displayed distinct patterns. Specifically, during the experiment period, a negative correlation was observed between temperature and humidity ($r = -0.550$, $p < 0.01$), indicating that as temperature increased, humidity tended to decrease. Additionally, a weak negative correlation was observed between temperature and CO_2 concentration ($r = -0.169$, $p < 0.01$), suggesting a slight decrease in CO_2 concentration with rising temperature. However, the humidity made a weak positive relation with CO_2 concentration ($r = 0.006$, $p = 0.049$).

3.2. The Results of the ARIMA Model

In this study, the concentration of CO_2 was predicted using the ARIMA model. While analyzing time series data with an ARIMA model, it is essential to verify whether the data exhibit stationarity [43]. The stationarity of the time series data was verified through the Augmented Dickey-Fuller (ADF) test. The ADF test statistic for the CO_2 concentration data was −9.225884, which is below the critical values at significance levels of 1%, 5%, and 10% (−3.958454, −3.410526, and −3.127071, respectively). Additionally, the *p*-value was extremely low at 1.126898×10^{-13}, providing strong evidence to reject the null hypothesis of it being non-stationary. Therefore, the data meet the conditions of being stationary, indicating that no further differencing is necessary. This suggests that the data are suitable for applying the ARIMA model in time series analysis. Figure 4 illustrates that the CO_2 concentration dataset is stationary. To sum up, the data are in a stationary state and are ready for processing with the ARIMA model.

To obtain reliable results, it is crucial to determine the appropriate parameters for the ARIMA model. The parameters (*p*, *d*, *q*) are defined as follows:

- p: Order of the AR term;
- d: Number of differencing required to make the time series stationary;
- q: Order of the MV term.

In this study, the autocorrelation structure of five datasets with varying collection intervals was analyzed using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. These analyses were essential for identifying the autocorrelation patterns of each dataset and developing suitable predictive models. The correlation coefficient is displayed on the *x*-axis, while the number of lags is plotted on the *y*-axis [35]. The ACF plot for the CO_2_1 dataset shows a positive correlation beyond the first lag, suggesting the presence of a Moving Average (MA) component. The MA(1) model, which is the first order of the moving average model, explains how the error term from a previous point in time predicts the current value in time series data. This model is represented by Equa-

tion (8) [44]. It is used to analyze the impact of random shocks on future values in time series data. The slow decline in the correlation coefficients suggests that the MA model is necessary, particularly an MA(1) model, since the coefficients remain relatively high after the first lag. Additionally, the dataset exhibits a strong autoregressive effect in the first two lags, which requires an AR(2) model. The AR(2) model, or the second order autoregressive model, illustrates the relationship of the current value with the values of the two preceding points in time series data. This model is represented by Equation (9) [44]. It is useful for analyzing patterns and trends in time series data and for predicting future values. Additionally, the PACF plot shows relatively high partial autocorrelation coefficients at the first two lags, followed by a sharp decline to near-zero values, suggesting that the order of the AR model should be two. The sharp decrease after the first two lags indicates that additional AR effects are not significant. Thus, the MA(1) model can effectively capture the high correlation coefficient at the first lag. Therefore, an ARIMA (2,0,1) model was selected for the CO_2_1 dataset.

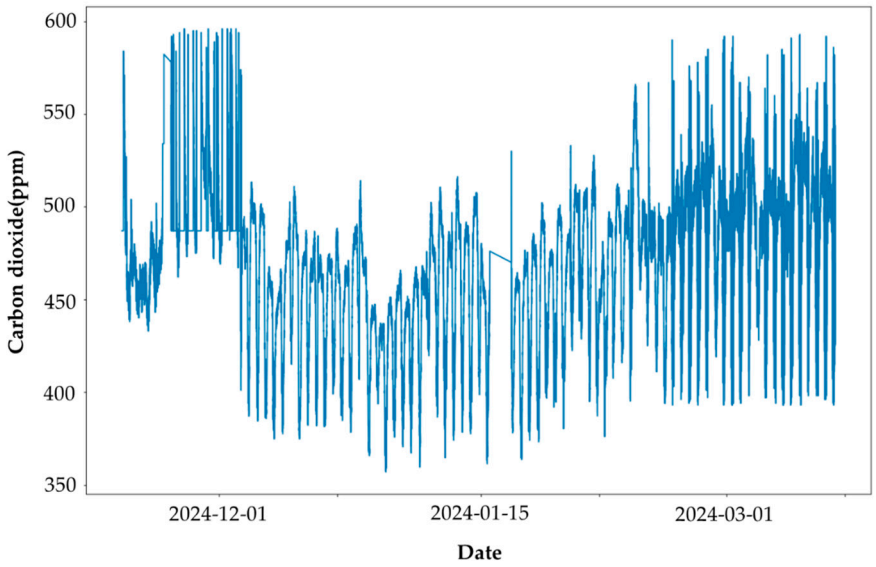


Figure 4. Stationary test results for the ARIMA model.

$$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} \quad (8)$$

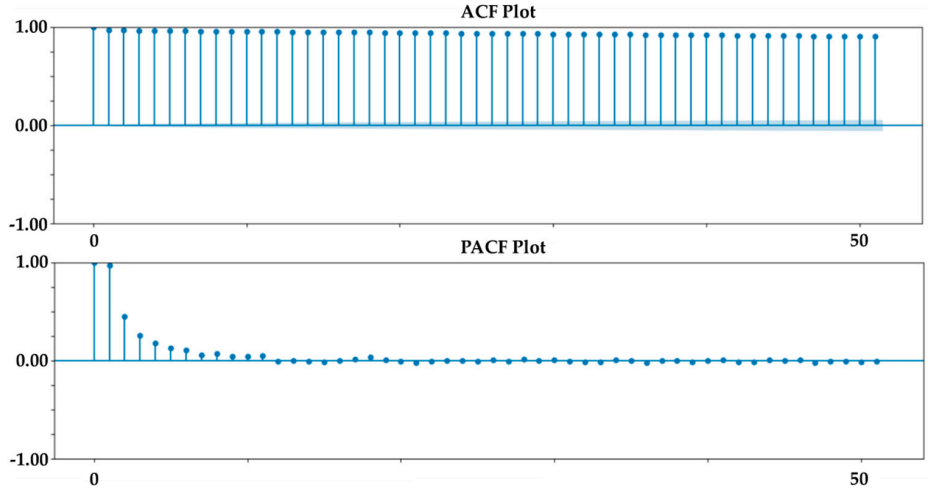
$$X_t = c + \varnothing_1 X_{t-1} + \varnothing_2 X_{t-2} + \epsilon_t \quad (9)$$

In the datasets with CO_2_5, CO_2_10, CO_2_30, and CO_2_60, the Autocorrelation Function (ACF) plots exhibit a high positive correlation coefficients at the first lag, which gradually decrease. This trend suggests the possibility of presence Moving Average (MA) components, particularly indicated by the sustained positive correlations at the first two lags, which makes the MA(2) model suitable for explaining the data's correlation pattern. The persistence of positive values beyond the first two lags provides a basis for setting the moving average order at q = 2. Additionally, the PACF plots show a sharp truncation of the correlation coefficients after the initial two lags, with subsequent lags converging significantly toward zero. This indicates that an AR(2) model can adequately explain the autoregressive structure of the data. The significant partial autocorrelation at the first two lags strongly supports the necessity of two AR terms to model the autoregressive characteristics of the time series data, particularly as the high values at the first and second lags justify selecting an autoregressive order p = 2. The PACF plots generally show a sharp decline in correlation coefficients after the first two lags, approaching zero, indicating that the order of the AR component is two, and further AR effects are not significant beyond these lags. This means that two AR terms can sufficiently model the autoregressive characteristics of the time series data. Based on the common patterns observed in the ACF and

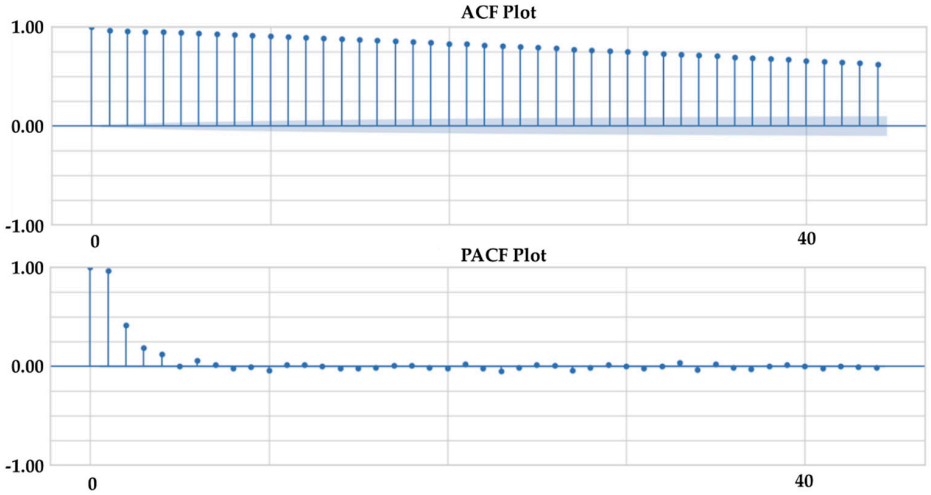
PACF, an ARIMA (2,0,2) model was chosen for the datasets with CO_2_5, CO_2_10, CO_2_30, and CO_2_60. This model captures both the autoregressive and moving average properties of each dataset, effectively reflecting the effects up to the second lag. The AR order of two terms explains the autoregressive dynamics found in the initial two lags, while the MA order of two terms is necessary to model the moving average effects observed at the initial lags. The ACF and PACF plots are depicted in Figure 5.

In this study, we considered both the model fit and complexity using the AIC and Bayesian Information Criterion (BIC). Thus, these two indices were crucial in selecting the optimal ARIMA models, particularly advantageous for interpretation and prediction.

For the CO_2_1 dataset, the ARIMA (2,0,1) model displayed the lowest AIC and BIC values, calculated at −123.45 and −118.90, respectively. Likewise, for the CO_2_5, CO_2_10, CO_2_30, and CO_2_60 datasets, the ARIMA (2,0,2) model showed the lowest AIC and BIC values, both recorded at −123.45 and −118.90, respectively. The minimum values indicate that the model that provides the best balance, between fitting the data and maintaining model simplicity compared to other parameter combinations. The (p, d, q) values of the ARIMA model were determined following a thorough trial-and-error process, identifying the model parameters that maximize efficiency and effectiveness (Figure 6).

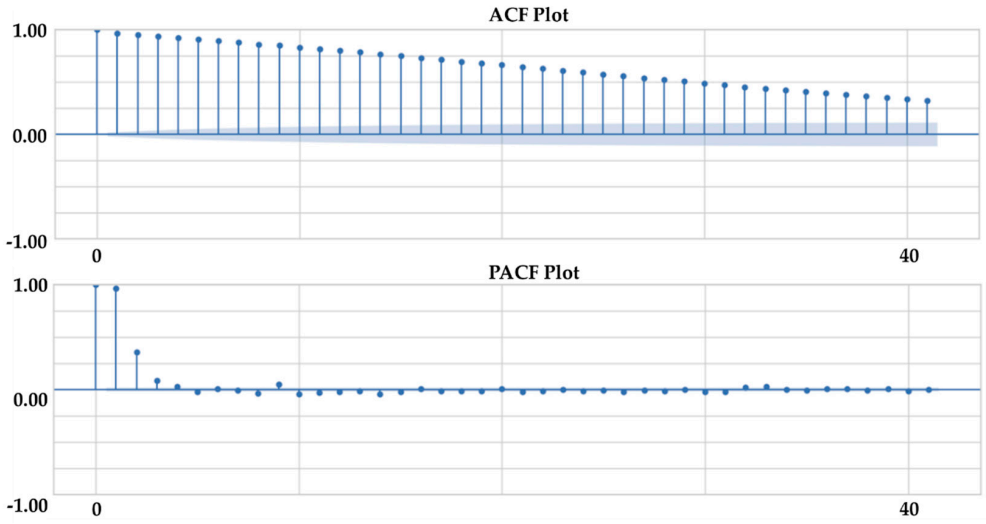


(**a**)

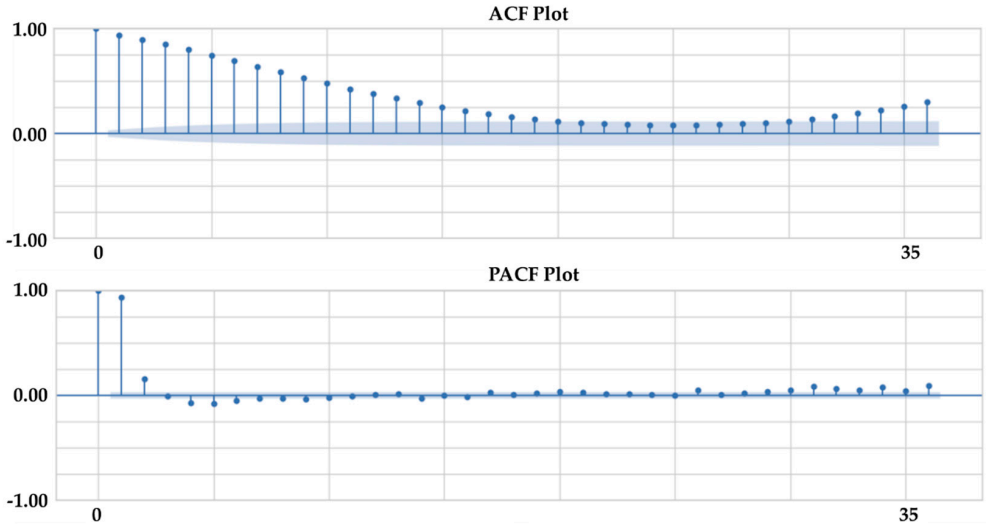


(**b**)

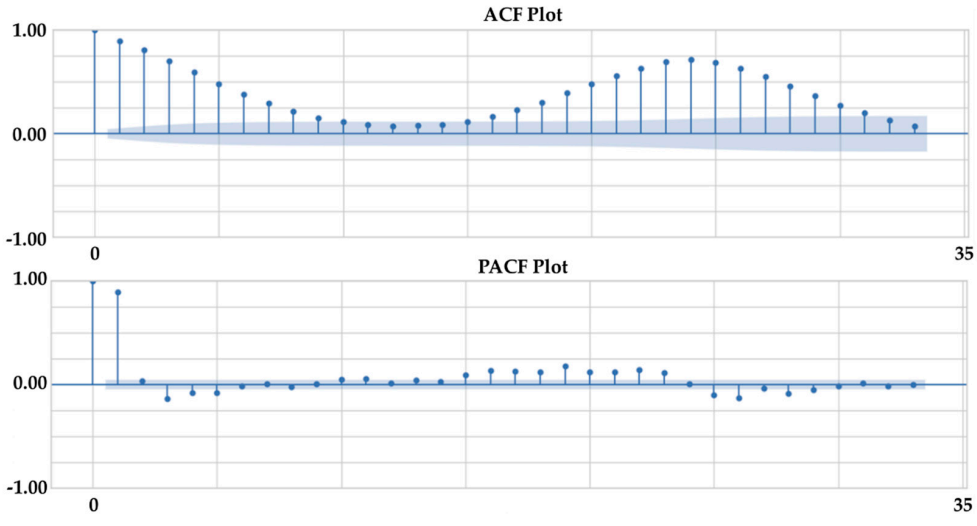
Figure 5. *Cont.*



(**c**)



(**d**)



(**e**)

Figure 5. Selecting p, d, and q parameters for the ARIMA model: (**a**) 1 min ACF and PACF diagram; (**b**) 5 min ACF and PACF diagram; (**c**) 10 min ACF and PACF diagram; (**d**) 30 min ACF and PACF diagram; (**e**) 60 min ACF and PACF diagram.

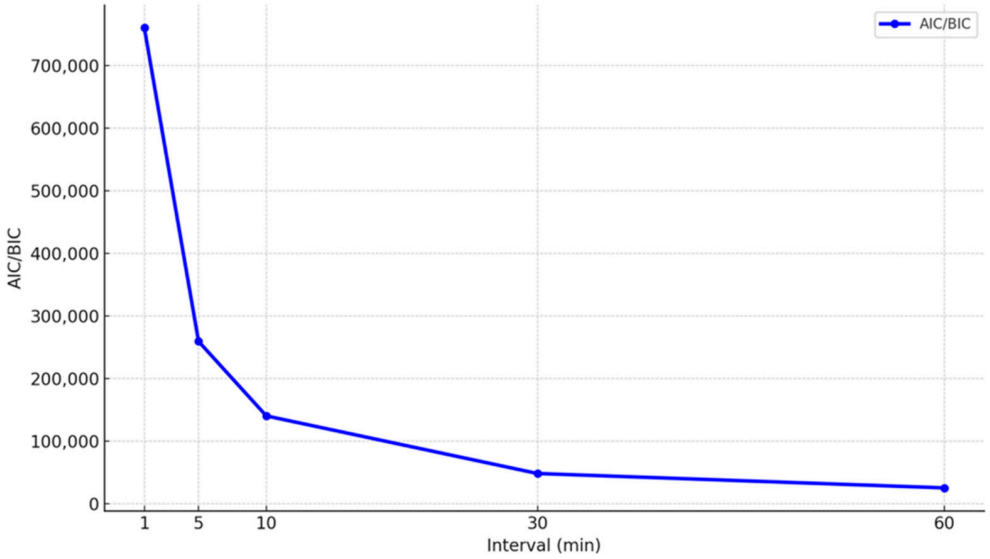


Figure 6. Selecting p, d, and q values for the ARIMA model: AIC and BIC.

3.3. Dataset Performance

In this section, the results of the CO_2_1, CO_2_5, CO_2_10, CO_2_30, and CO_2_60 datasets are compared to conduct the performance analysis. The highest performing datasets testing data were obtained from the ARIMA model in CO_2_1 prediction (MAE = 2.832, RMSE = 7.359, R^2 = 0.928). For CO_2 concentration prediction, all the models performed better when using the CO_2_1 dataset compared to the other datasets. The lowest performing datasets testing data were obtained from the PFM in CO_2_60 prediction (MAE = 19.158, RMSE = 25.04, R^2 = 0.753). The performance outcomes of all models for predicting CO_2 concentrations are presented in Table 2. Overall, the results suggest that the ARIMA model outperformed the PFM across the five datasets.

Table 2. Model performance under five datasets.

Dataset	Models	Training			Testing		
		RMSE	MAE	R^2	RMSE	MAE	R^2
CO_2_1	ARIMA	7.160	2.556	0.965	7.359	2.832	0.928
	PFM	21.417	17.601	0.981	22.388	18.645	0.951
CO_2_5	ARIMA	8.691	3.798	0.945	9.21	4.119	0.844
	PFM	22.424	17.944	0.966	22.483	18.914	0.921
CO_2_10	ARIMA	9.799	4.568	0.931	10.369	7.327	0.817
	PFM	22.607	18.085	0.857	23.493	18.96	0.883
CO_2_30	ARIMA	12.735	6.957	0.877	16.614	7.375	0.815
	PFM	23.145	18.114	0.846	24.848	19.05	0.879
CO_2_60	ARIMA	16.034	9.766	0.804	19.469	11.48	0.762
	PFM	23.198	18.451	0.743	25.04	19.158	0.753

RMSE and MAE are measured in ppm, while R^2 is unitless.

In the CO_2_1 dataset, the ARIMA model exhibits a 2.78% higher RMSE in the test data compared to the training data, with a 10.80% increase in MAE and a 3.83% decrease in R^2. This indicates that the model demonstrates slightly higher errors while maintaining consistent predictive capabilities in the test data, and generally, the ARIMA model has shown high performance with this dataset, though a slight performance degradation is observed in the test data. In the PFM, the RMSE in the test data is 4.54% higher than in the

training data, the MAE has increased by 5.93%, and the R^2 has decreased by 3.06%. This shows a drop in prediction accuracy in the test data.

In the CO_2_5 dataset, the ARIMA model shows a 5.98% higher RMSE in the test data compared to the training data, an 8.45% increase in MAE, and a 10.69% decrease in R^2. This indicates a decline in model consistency, particularly evident in the significant drop in R^2, highlighting more pronounced performance degradation in the test data compared to the training data. In the PFM, the RMSE in the test data is 0.26% higher than in the training data, with MAE increasing by 5.40% and R^2 decreasing by 4.66%. The prediction accuracy for the test data was not consistently maintained. In the CO_2_10 dataset, the ARIMA model shows a 5.82% higher RMSE in the test data, a 60.25% increase in MAE, and a 12.24% decrease in R^2. This indicates a significant degradation in model performance. As the dataset intervals increase, the decline in performance becomes more apparent, with the increase in MAE in the test data being particularly noteworthy. The PFM exhibits a 3.92% higher RMSE in the test data compared to the training data, a 4.84% increase in MAE, and a 3.03% decrease in R^2. This suggests that the model struggles to maintain consistent performance compared to the training data.

In the CO_2_30 dataset, the ARIMA model exhibits a 30.44% higher RMSE in the test data compared to the training data, a 6.01% increase in MAE, and a 7.07% decrease in R^2. This highlights a pronounced performance degradation as the data intervals lengthen, indicating a decline in the model's generalization ability, particularly evident in the significant difference in RMSE. The PFM also shows deterioration in consistent predictive capabilities, with a 7.35% higher RMSE in the test data, a 5.17% increase in MAE, and a 3.89% decrease in R^2, further confirming the challenges in maintaining model performance with longer data intervals.

In the CO_2_60 dataset, the ARIMA model shows a significant degradation in performance at the longest data interval, with RMSE in the test data being 21.41% higher than in the training data, MAE increasing by 17.60%, and R^2 decreasing by 5.22%. This indicates the largest decline in generalization ability at the longest data interval. The PFM also demonstrates reduced generalization capability in this dataset, with a 7.93% higher RMSE in the test data, a 3.84% increase in MAE, and a 1.34% decrease in R^2, confirming that the model struggles to maintain performance relative to the training data in the face of extended data intervals.

3.4. Model Performance

Overall, the ARIMA model demonstrates satisfactory prediction results during both training and testing periods. The performance of the two models for CO_2_1, CO_2_5, CO_2_10, CO_2_30, and CO_2_60 is shown in Figures 7 and 8.

For the CO_2_1 prediction training time, the ARIMA model results are comparatively higher (MAE = 2.556, RMSE = 7.160, and R^2 = 0.965) followed by the PFM (MAE = 17.601, RMSE = 21.417, and R^2 = 0.981). Since validation assessment is crucial, as previously mentioned, this study considered the testing outcomes as an indicator of the model's performance. Consequently, during the testing phase, the ARIMA model outperformed other models (MAE = 2.832, RMSE = 7.359, and R^2 = 0.928). The PFM struggled to produce a more competitive outcome than the ARIMA model for the CO_2_1 predictions as reflected in the results (MAE = 558% high, RMSE = 204% high, and R^2 = 2.477% low).

By CO_2_1 prediction results, the CO_2_5 exhibited a similar performance pattern. The ARIMA model was outperformed during the training and testing time compared to PFM. The top performance was achieved from the ARIMA model for the CO_2_5 predictions during the training time (MAE = 3.798, RMSE = 8.691, and R^2 = 0.945) and testing time as well (MAE = 4.119, RMSE = 9.21, and R^2 = 0.844). According to the testing results, PFM was attained second (358.5% higher MAE, 144.5% higher RMSE, and 9.116% lesser R^2) (refer to Table 2). As noted earlier, the ARIMA model demonstrated better performance in the D3 predictions. The results of CO_2_10 prediction training and testing time were comprehensively explained in Table 2. According to this, the ARIMA model performance surpassed

the PFM performance during the training time. The difference observed when compared to the ARIMA model performance was PFM (158% higher MAE, 126.5% higher RMSE, and 8.080% lesser R^2).

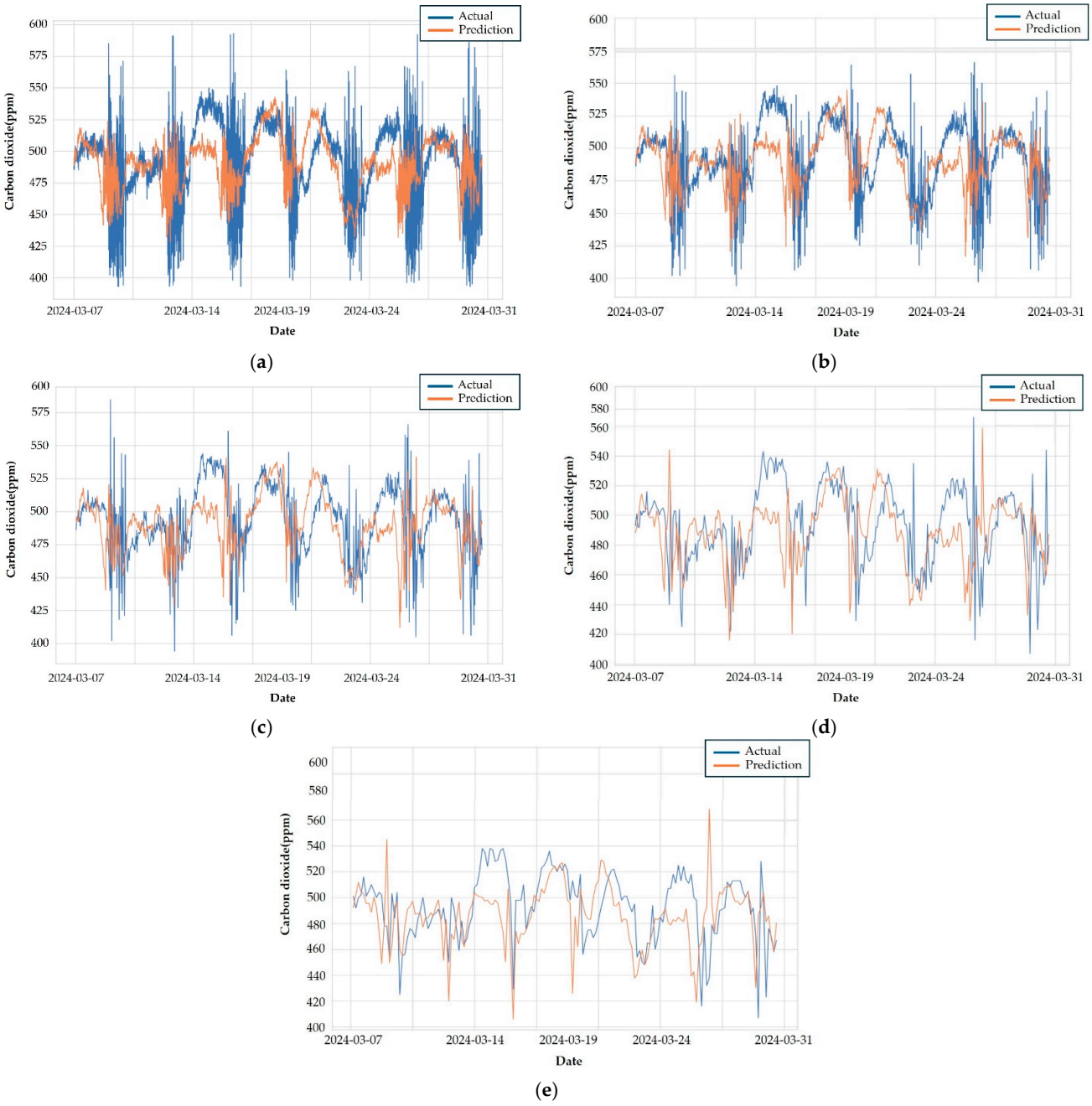


Figure 7. Comparison of observed and predicted values using ARIMA for CO_2 concentration predictions: (**a**) 1 min; (**b**) 5 min; (**c**) 10 min; (**d**) 30 min; (**e**) 60 min.

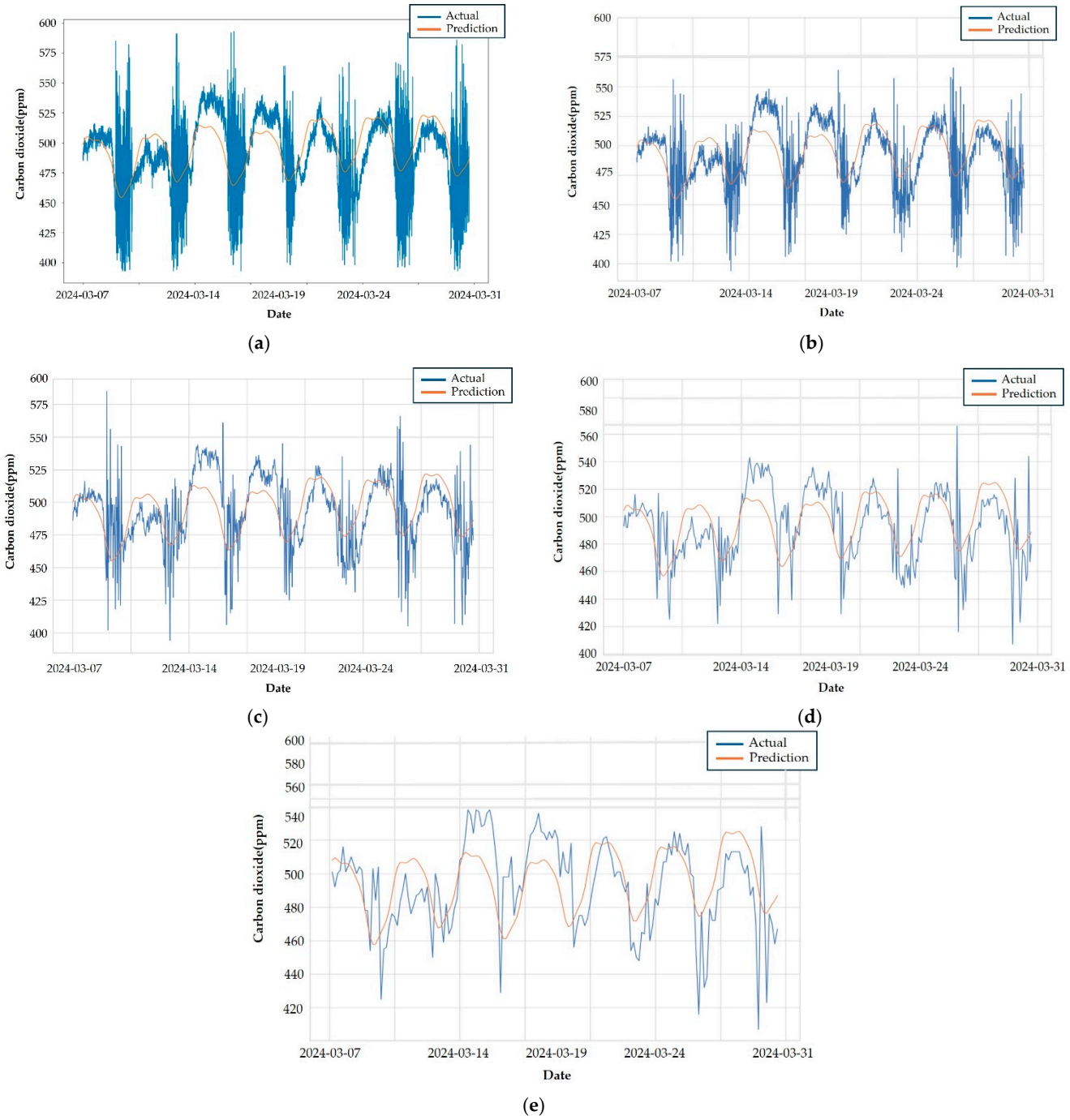


Figure 8. Comparison of observed and predicted values using PFM for CO_2 concentration predictions: (**a**) 1 min; (**b**) 5 min; (**c**) 10 min; (**d**) 30 min; (**e**) 60 min.

When evaluating the outcomes of CO_2_30 predictions, the ARIMA model results of training time (MAE = 6.957, RMSE = 12.735, and R^2 = 0.877) and testing time (MAE = 7.375, RMSE = 16.614, and R^2 = 0.815) were superior, whereas PFM carried out the second. During the testing time, PFM (158.960% higher MAE, 49.540% higher RMSE, and 7.840% lesser R^2) was placed in the second position.

Similar to the other findings, the prediction results from the CO_2_60 dataset also demonstrate that the models behave in a similar manner. The ARIMA model executes better results than PFM (88.870% higher MAE, 44.690% higher RMSE, and 7.580% lesser R^2) during the training time. Furthermore, the comparison of the evaluation metrics' results between the training and testing periods is presented in Figure 9.

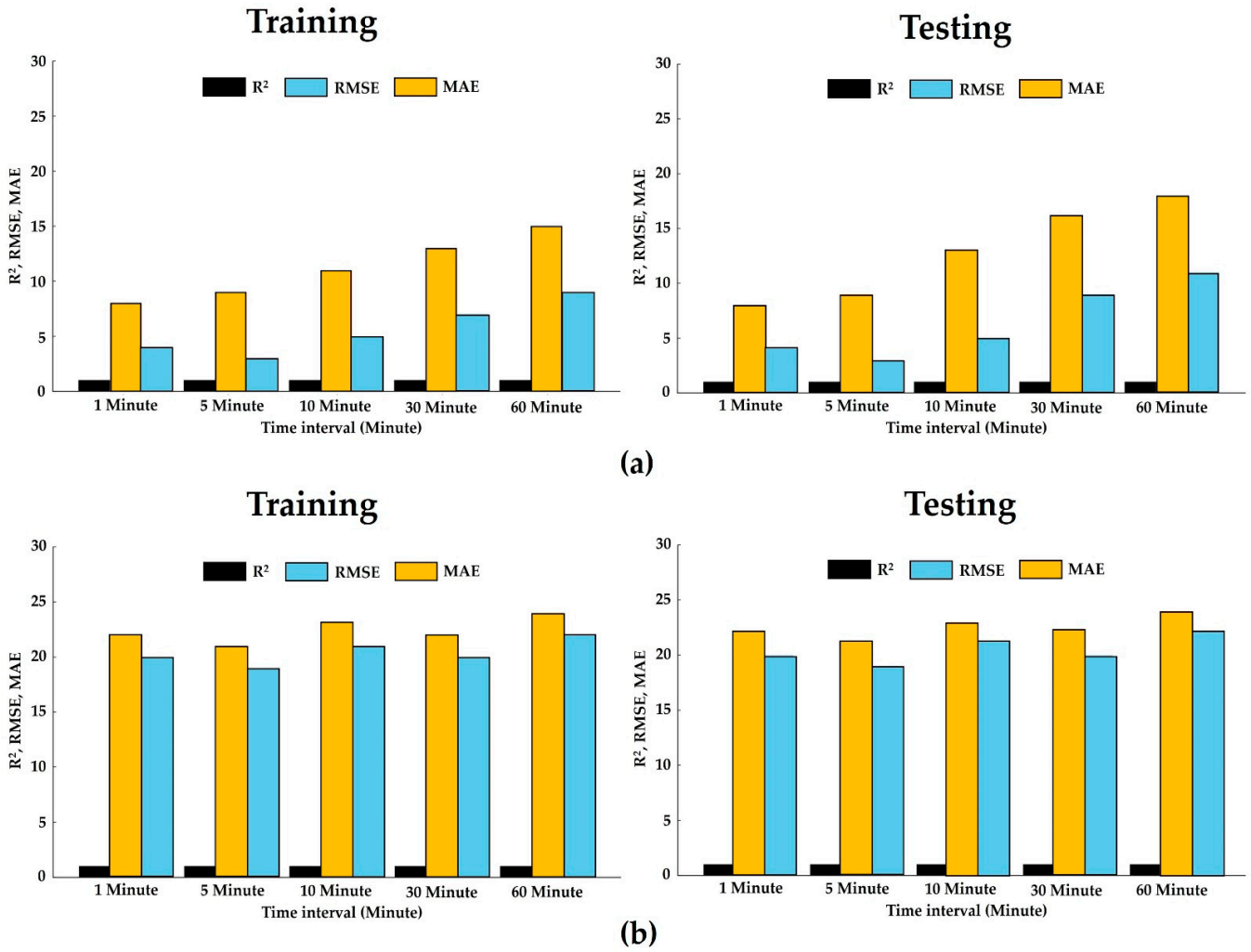


Figure 9. (**a**) The evaluation metrics results of the ARIMA model for CO_2 concentration predictions; (**b**) The evaluation metrics results of PFM for CO_2 concentration predictions.

3.4.1. Performance of the ARIMA Model

The performance of the ARIMA model was evaluated using the RMSE, MAE, and R^2 metrics. When comparing the performance between CO_2_1 and CO_2_60 datasets, the CO_2_1 dataset exhibited the least error and maximum performance, showing a 62.20% lower RMSE, 75.33% lower MAE, and a 21.78% higher R^2 compared to the CO_2_60 dataset. When comparing the performance between CO_2_5 and CO_2_60 datasets, the ARIMA training results for RMSE were 20.10% lower, MAE was 31.25% lower, and R^2 was 9.95% higher. Additionally, when comparing the performance between the CO_2_10 and CO_2_60 datasets, the ARIMA training for RMSE was 29.03% lower, MAE was 61.35% lower, and R^2 was 13.59% higher. Finally, when comparing the performance between the CO_2_30 and CO_2_60 datasets, the ARIMA training results showed a 55.71% reduction in RMSE, a 61.6% reduction in MAE, and a 13.87% increase in R^2. In summary, data collection at a 1-min interval demonstrated the highest model performance with an R^2 of 0.928, RMSE of 7.359, and MAE of 2.832. The performance of the model in predicting CO_2 concentration decreased as the data collection interval increased, thus showing the lowest performance at the CO_2_60 dataset (R^2 = 0.762, RMSE = 19.469, MAE = 11.48).

The comparison results between the actual values and predicted values are presented in Figure 7. According to the ARIMA prediction outcomes, the datasets with intervals of CO_2_1, CO_2_5, CO_2_10, CO_2_30, and CO_2_60 followed a similar performance pattern. The

training and testing prediction performance demonstrated that the CO_2_1 dataset achieved the best performance in predicting CO_2 concentrations compared to other datasets.

3.4.2. Performance of PFM

The results of the PFM are presented in Table 2. It was observed that the CO_2_1 dataset provided the best performance compared to other datasets (RMSE = 22.388, MAE = 18.645, R^2 = 0.951). Conversely, the lowest performance was observed in the CO_2_60 dataset (RMSE = 25.04, MAE = 19.158, R^2 = 0.753). While comparing the performance between these two datasets, the CO_2_1 dataset exhibited the least error and maximum performance, showing a 10.59% lower RMSE, 2.68% lower MAE, and a 26.29% higher R^2 compared to the PFM testing results for the 60-min dataset. When comparing the performance between the CO_2_5 and CO_2_60 datasets, the PFM testing results for RMSE were 0.42% lower, MAE was 1.42% lower, and R^2 was 3.26% higher. Additionally, when comparing the performance between CO_2_10 and CO_2_60 datasets, the PFM testing for RMSE was 4.70% lower, MAE was 1.66% lower, and R^2 was 7.70% higher. Finally, when comparing the performance between the CO_2_30 and CO_2_60 datasets, the PFM testing results were 0.77% lower in RMSE, 0.56% lower in MAE, and 16.73% higher in R^2. Consequently, data collection at a 1-min interval demonstrated the highest model performance compared to other datasets. Similar to the ARIMA model predictive performance results, the performance of the PFM in predicting CO_2 concentration decreased as the data collection interval increased, thus showing the lowest performance at the 60-min interval.

The comparison results between the actual values and predicted values are presented in Figure 8. According to the PFM prediction outcomes, the datasets with intervals of CO_2_1, CO_2_5, CO_2_10, CO_2_30, and CO_2_60 followed a similar performance pattern. The training and testing prediction performance demonstrated that the 1 CO_2_1 dataset achieved the best performance in predicting CO_2 concentrations.

3.5. Model's Performance Comparison and the Proposed Model

In the dataset with CO_2_1, the ARIMA model presents a 67.13% lower RMSE and an 84.81% lower MAE compared to the PFM, indicating greater accuracy in terms of error metrics. Conversely, the R^2, which indicates the proportion of variance the model explains, is 2.48% higher in the PFM, suggesting that it may slightly better reflect the variability in the data. In the dataset with CO_2_60, the ARIMA model continues to outperform with a 22.25% lower RMSE and a 40.08% lower MAE, while also achieving a 1.20% higher R^2. This consistency suggests that the ARIMA model is generally superior to the PFM, making it a preferable choice for predicting CO_2 concentrations overall.

4. Discussion

4.1. Comparative Analysis of Models in CO_2 Concentration Prediction

CO_2 and plant growth are closely linked due to the precise influence of CO_2 on photosynthesis, nutrient uptake, biomass, and chloroplast diversity [45–47]. However, key factors influencing the final quality and quantity of production, including seed germination, growth of roots and shoots, stem length, flower growth, and leaf development, primarily depend on CO_2 concentrations [48,49]. Plant growth is significantly affected by CO_2 concentrations, but direct measurement of CO_2 can be time-consuming, costly, and labor-intensive. Consequently, this study evaluated the accuracy of CO_2 concentration predictions within a greenhouse using the ARIMA model and the PFM developed by Facebook, across various data collection intervals. The research analyzed the performance of the prediction models using five different datasets. The results indicate that the 'CO_2_1' dataset was more effective in accurately modeling CO_2 concentrations than the 'CO_2_5', 'CO_2_10', 'CO_2_30', and 'CO_2_60' datasets. This comparison of the two models' predictive performance with the five datasets helped identify an appropriate method for predicting greenhouse CO_2 concentrations.

Another previous study [50] conducted research to predict greenhouse environmental variables over a period of 45 days. This research utilized the ARIMA model to predict temperature and humidity, achieving a minimum error rate of 0.4% and a forecasting accuracy of 95%. In the current study, the same ARIMA model was applied to predict CO_2 concentrations in a greenhouse over a period of 121 days, with the results achieving a predictive accuracy of 92.8% on test data. These results are consistent with the previous findings and further reaffirm the reliability of the ARIMA model. PFM effectively processes daily, weekly, and annual seasonal data and accurately identifies complex patterns of CO_2 concentration that reflect changes in both internal and external greenhouse conditions [51].

This study reveals that PFM is somewhat limited compared to the ARIMA model. The ARIMA model excels at capturing rapid changes in CO_2 concentrations in data collected at 1-min intervals, whereas PFM effectively identifies major trends in CO_2 concentrations from data gathered every 30 min. While ARIMA requires careful tuning of its p, d, and q parameters to prevent overfitting, PFM offers a more flexible approach with less intensive parameter adjustments. Such optimization strikes an efficient balance between computational load and predictive accuracy, which is essential for real-time greenhouse management systems. Considering the overall predictive performance and minimal error metrics, the research concludes that the ARIMA model is more suitable for predicting CO_2 concentrations.

4.2. Model Accomplishment

The performance of the ARIMA model is negatively affected as the data collection interval increases. Firstly, expanding the data collection interval from 1 min to 60 min leads to significant information loss. Data exhibiting high volatility, such as carbon dioxide concentrations, can change rapidly over time. Shorter intervals are more effective at capturing these fine variations. Secondly, as intervals widen, the responsiveness and temporal resolution of the model decrease, which challenges the model's ability to capture recent changes and natural patterns, including periodicities. This directly undermines the accuracy of predictions. Since ARIMA models rely heavily on the autocorrelation of time series data, a reduced temporal resolution significantly hampers the model's ability to learn data autocorrelation and periodicity. Furthermore, a decrease in the number of data points diminishes the amount of information available for the model to learn, particularly exacerbating performance degradation in sparse data conditions. Therefore, for effective prediction, it is ideal to collect data at as short an interval as possible, though this comes with increased costs and effort in data processing and storage. Considering these factors, the performance of the ARIMA model is superior when data collection intervals are shorter and deteriorate as intervals lengthen.

In contrast, the performance decline of the PFM is also notable as the interval widens. Specialized in analyzing various temporal elements such as trends, seasonality, and holiday effects in time series data, the PFM is particularly affected by changes in the data collection interval [37]. Firstly, increasing the interval leads to a loss of detailed data and a decrease in the accuracy of estimating seasonality and trends [38]. Data collected at 1-min intervals can capture subtle changes in CO_2 concentrations, aiding the model in learning more accurate trends and patterns. In contrast, data at 60-min intervals may miss important fluctuations or trends. Secondly, a wider interval can lead to less accurate identification of fine-grained seasonal patterns, such as hourly or daily variations, which diminishes the model's predictive power in environments where short-term changes are critical [40]. Thirdly, a wider gap between data points lowers the resolution of the time series, thereby increasing the variability in statistical estimates. Generally, predictive models tend to perform better when trained on a larger number of data points [36]. Therefore, as the interval widens and the number of available data points decreases, it negatively impacts model performance. Finally, the resolution of the data also affects the model's propensity for overfitting and its ability to generalize predictions. High-resolution data can increase the risk of overfitting, but this risk can be managed through proper data handling and model

tuning. Conversely, low-resolution data may lead to underfitting, degrading the model's generalization capabilities in making predictions.

Consequently, while the ARIMA model enables more precise predictions with high-resolution data collected at short intervals, suggesting its suitability for rapidly changing environmental conditions, the PFM, while useful for analyzing seasonal variability and trends in time series data, sees its ability to capture fine changes diminish as the data collection interval widens. Considering the impact of data collection intervals, the choice of model and data collection strategy should be carefully determined based on the research objectives and available resources. For instance, the ARIMA model may be more appropriate in situations where real-time detection of environmental changes is required, whereas the PFM might be advantageous for long-term seasonal variability. Such decisions will vary depending on costs, data storage and processing capabilities, and the required accuracy of predictions.

4.3. Influence of Input Variables and Models on CO_2 Concentration Prediction

In this study, the data collection interval had a substantial impact on the prediction accuracy of the ARIMA model and PFM for estimating CO_2 concentration. Considering the various input combinations, the dataset with CO_2_1 produced the most accurate results across all models. This suggests that shorter data collection intervals offer better accuracy compared to datasets gathered over longer intervals. Generally, the R^2 values increased as the data collection intervals decreased, while RMSE and MAE values showed a downward trend. As more input parameters were incorporated, there was a noticeable improvement in model accuracy. However, incorporating multiple input variables raised computational costs and added complexity to the model, potentially limiting its practical application [52]. As shown in numerous studies, the amount and relevance of input parameters have a substantial impact on prediction accuracy [53].

Comparing the two-time series forecasting models, the ARIMA model demonstrated relatively higher stability and less sensitivity to input data variations in prediction accuracy [54]. Considering the increased rates of RMSE and MAE between the training and testing phases, the PFM also exhibited high stability, but ARIMA generally outperformed in overall performance. This stability may be attributed to ARIMA's ability to effectively capture complex relationships and patterns in time series data and generalize well with unseen data [55]. The stability of the two models varied with different combinations of input data, with ARIMA showing less sensitivity to these variations. Primandari et al. [56] utilized the PFM to predict CO_2 concentrations. The PFM is noted for its high predictive accuracy and low error values, effectively managing the seasonality and change points in CO_2 levels, which showed a continuing upward trend without any reduction in recent levels. However, the performance and stability of a model can differ based on the specific datasets used and the challenges faced, making it essential to evaluate model performance across diverse datasets [57]. This remains a valuable consideration when selecting time series forecasting models.

The adaptability of ARIMA to various data intervals and its robust performance suggest its suitability for environments where rapid real-time data processing is crucial. For example, in urban air quality monitoring or industrial environmental management systems, the ARIMA model's quick response to environmental changes is highly beneficial. On the other hand, the PFM's capability to analyze long-term seasonal variations makes it more appropriate for applications like agricultural planning where long-term trends are more relevant.

The impact of different data collection intervals on data file size was analyzed as input variables for a CO_2 concentration prediction model. The collection interval affects not only the performance of the prediction model but also data management, storage costs, and processing times. The data file from CO_2_1 is the largest, at 59.82 MB, reflecting the highest frequency of data collection. This high granularity captures hourly variations in detail, enabling more accurate predictions but also requiring significant resources for data

processing and storage. The data file at CO_2_5 is significantly reduced in size to 11.96 MB, approximately an 80% decrease compared to the CO_2_1 data file. This reduction results in some loss of detailed information, but it eases data processing and management. The size of the data file collected at CO_2_10 is 5.98 MB, which is a 50% reduction from the CO_2_5 data file. This further decreases the need for storage space and speeds up processing, but it also increases the potential for loss in prediction accuracy. The data file collected, CO_2_30, is much smaller at 1.99 MB, significantly reducing data frequency and risking missing important environmental changes. However, this interval significantly cuts costs related to data management and processing. The data file collected at CO_2_60 is the smallest at 0.997 MB, offering minimal data and a high likelihood of missing critical time periods. However, they provide benefits in terms of minimal storage space usage and reduced processing time. The variation in data collection intervals and file sizes clearly illustrates the trade-offs between data quality and quantity and processing costs. High-resolution data enable more accurate analyses but justify increased costs and resource usage, as larger file sizes demand more time and money for data processing and storage. Thus, assessing the model's effectiveness and the suitability of data intervals, considering budget and infrastructure constraints, is crucial in choosing the most efficient data collection strategy. Especially for large facilities or those with limited budgets, selecting an appropriate model and data collection frequency plays a vital role in maximizing resource optimization and efficiency.

This study confirmed that both the ARIMA model and PFM exhibit higher prediction accuracy at shorter data collection intervals. This suggests their utility in applications where real-time or high-frequency data monitoring is crucial. For instance, in urban air quality monitoring or industrial environmental management systems, rapid responses through real-time data are essential when using short data collection intervals. However, the complexity of the models and computational costs tend to increase as the data collection intervals decrease. This can be particularly challenging when dealing with large-scale data, necessitating a cost-effectiveness analysis.

Overall, this research proposes a methodology to determine the optimal data collection frequency for regulating optimal CO_2 concentrations in greenhouse crops and enhancing the efficiency of smart farm operations. This suggests an appropriate collection frequency for efficiently utilizing vast amounts of data in the agricultural sector. These findings underscore the importance of high-frequency data collection in accurately monitoring and controlling CO_2 concentration within greenhouses.

5. Conclusions

In this study, we utilized two-time series models to predict CO_2 concentrations in strawberry greenhouses. The primary objective was to evaluate the optimal data collection intervals needed to achieve high-accuracy predictions of CO_2 concentrations within the greenhouse environment. The results of the study demonstrated that the ARIMA model outperforms the Prophet model (PFM) in predicting CO_2 concentrations across all data collection intervals. Moreover, among the five datasets (CO_2_1, CO_2_5, CO_2_10, CO_2_30, CO_2_60), the ARIMA model and PFM demonstrated the best performance on the CO_2_1 dataset. Overall, the performance of the ARIMA model and PFM improved as the data collection interval shortened. Specifically, the ARIMA model with the CO_2_1 dataset showed that R^2 increased by 21.78%, and RMSE and MAE reduced by 62.20% and 75.33%, respectively, compared to the CO_2_60 dataset. Additionally, the PFM with the CO_2_1 dataset showed that R^2 increased by 26.29% and RMSE and MAE reduced by 10.59% and 2.68%, respectively, compared to the CO_2_60 dataset. This research clearly highlighted the effectiveness of time series models, especially the ARIMA model, in forecasting CO_2 concentrations in a greenhouse. The results offer valuable insights into CO_2 concentration patterns, supporting data-driven decision-making in plant production and environmental management through real-time CO_2 monitoring. However, modeling CO_2 concentrations has some limitations because it depends on other variables such as ventilation conditions, temperature, humidity, and seasonal variations. Therefore, future studies may focus on

developing predictive models that consider ventilation conditions, temperature, humidity, and seasonal changes to predict CO_2 concentrations in plant production.

Author Contributions: Conceptualization, S.H.S.; methodology, S.H.S.; software, S.H.S.; validation, N.C.D., E.A. and N.T.; formal analysis, S.H.S.; investigation, S.H.S.; resources, H.T.K.; data curation, S.H.S., J.K. and O.M.O.; writing—original draft preparation, S.H.S.; writing—review and editing, N.C.D., D.H.K., E.A., N.T. and O.M.O.; visualization, S.H.S. and N.C.D.; supervision H.T.K.; project administration, H.T.K.; funding acquisition, H.T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been financially supported by the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through Technology Commercialization Support Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA) (1545026476).

Data Availability Statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through the Technology Commercialization Support Program, funded by the Ministry of Agriculture, Food and Rural Affairs (MAFRA) (421040-04) and (1545026476) for financial support to conduct this experiment. Moreover, this work was also supported by the Glocal University 30 Project Fund of Gyeongsang National University in 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Prasad, R.; Lisiecka, J.; Raj, K. Strawberry–More than a Popular Summer Fruit: A Mini-Review. *Adv. Nutr. Food Sci.* **2022**, *2*, 1–11. [CrossRef]
2. Balasooriya, H.N.; Dassanayake, K.B.; Tomkins, B.; Seneweera, S.; Ajlouni, S. Impacts of Elevated Carbon Dioxide and Temperature on Physicochemical and Nutrient Properties in Strawberries. *J. Hortic. Sci. Res.* **2017**, *1*, 19–29.
3. Garza-Alonso, C.A.; Olivares-Sáenz, E.; González-Morales, S.; Cabrera-De la Fuente, M.; Juárez-Maldonado, A.; González-Fuentes, J.A.; Tortella, G.; Valdés-Caballero, M.V.; Benavides-Mendoza, A. Strawberry Biostimulation: From Mechanisms of Action to Plant Growth and Fruit Quality. *Plants* **2022**, *11*, 3463. [CrossRef]
4. Rapuru, R.; Bathula, S.; Kaliappan, I. Phytochemical Constituents and Pharmacological Activities of Strawberry. In *Recent Studies on Strawberries*; IntechOpen: London, UK, 2022; ISBN 1803551992.
5. Yuliwati, N.; Nugroho, R.F. The Potential of Strawberry, Rome Beauty Apple, and New Combination on Fasting Blood as Supporting Diet Therapy in Patients with Type II Diabetes Mellitus. *Glob. Med. Health Commun.* **2021**, *9*, 69–75. [CrossRef]
6. Liu, Z.; Liang, T.; Kang, C. Molecular Bases of Strawberry Fruit Quality Traits: Advances, Challenges, and Opportunities. *Plant Physiol.* **2023**, *193*, 900–914. [CrossRef] [PubMed]
7. Nonaka, A.; Hamada, T. Practical Utility Assessment of a Remote System for Monitoring CO_2 in Greenhouses by Using a Farmer-Built IoT System, and Usefulness of the System Together with a State-Space Model in Detecting Anomalous Values. *Agric. Inf. Res.* **2023**, *31*, 95–110.
8. Wang, Y.; Ma, B.; Shen, S.; Zhang, Y.; Ye, C.; Jiang, H.; Li, S. Diel Variability of Carbon Dioxide Concentrations and Emissions in a Largest Urban Lake, Central China: Insights from Continuous Measurements. *Sci. Total Environ.* **2024**, *912*, 168987. [CrossRef] [PubMed]
9. Sokolov, S.V. Optimization of Greenhouse Microclimate Parameters Considering the Impact of CO_2 and Light. *J. Eng. Sci.* **2023**, *10*, G14–G21. [CrossRef]
10. Cappelli, I.; Parri, L.; Tani, M.; Mugnaini, M.; Vignoli, V.; Fort, A. Pervasive IoT Monitoring of CO_2 for Smart Agriculture. In Proceedings of the 2024 IEEE Sensors Applications Symposium (SAS), Naples, Italy, 23–25 July 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
11. Taji, K.; Ghanimi, F. Enhancing Plant Disease Classification through Manual CNN Hyperparameter Tuning. *Data Metadata* **2023**, *2*, 112. [CrossRef]
12. Blay Carreras, E. Transitional Periods of the Atmospheric Boundary Layer. Ph.D. Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 2014.
13. Cox, D.T.C.; Maclean, I.M.D.; Gardner, A.S.; Gaston, K.J. Global Variation in Diurnal Asymmetry in Temperature, Cloud Cover, Specific Humidity and Precipitation and Its Association with Leaf Area Index. *Glob. Chang. Biol.* **2020**, *26*, 7099–7111. [CrossRef]
14. Veni, V.G.; Srinivasarao, C.; Reddy, K.S.; Sharma, K.L.; Rai, A. Soil Health and Climate Change. In *Climate Change and Soil Interactions*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 751–767.

15. Pimentel, C. Plant Responses to High-Temperature Stress. *Arch. Agric. Res. Technol. (AART)* **2022**, *3*, 1–2. [CrossRef]
16. Majidnia, M.; Ahmadabadi, Z.; Zolfaghari, P.; Khosravi, A. Time Series Analysis of Cutaneous Leishmaniasis Incidence in Shahroud Based on ARIMA Model. *BMC Public Health* **2023**, *23*, 1190. [CrossRef]
17. Jdi, H.; Falih, N. Comparison of Time Series Temperature Prediction with Auto-Regressive Integrated Moving Average and Recurrent Neural Network. *Int. J. Electr. Comput. Eng. (2088-8708)* **2024**, *14*, 1770. [CrossRef]
18. Desai, M.; Shingala, A. Time Series Prediction of Wheat Crop Based on FB Prophet Forecast Framework. In *ITM Web of Conferences, Proceedings of the 2nd International Conference on Data Science and Intelligent Applications (ICDSIA-2023), Gandhinagar, India, 28–29 April 2023*; EDP Sciences: Les Ulis, France, 2023; Volume 53, p. 02014.
19. Xia, C. Comparative Analysis of ARIMA and LSTM Models for Agricultural Product Price Forecasting. *Highlights Sci. Eng. Technol.* **2024**, *85*, 1032–1040. [CrossRef]
20. Iaousse, M.; Jouilil, Y.; Bouincha, M.; Mentagui, D. A Comparative Simulation Study of Classical and Machine Learning Techniques for Forecasting Time Series Data. *iJOE* **2023**, *19*, 57. [CrossRef]
21. Basak, J.K.; Paudel, B.; Kim, N.E.; Deb, N.C.; Kaushalya Madhavi, B.G.; Kim, H.T. Non-Destructive Estimation of Fruit Weight of Strawberry Using Machine Learning Models. *Agronomy* **2022**, *12*, 2487. [CrossRef]
22. Karki, S.; Basak, J.K.; Paudel, B.; Deb, N.C.; Kim, N.-E.; Kook, J.; Kang, M.Y.; Kim, H.T. Classification of Strawberry Ripeness Stages Using Machine Learning Algorithms and Colour Spaces. *Hortic. Environ. Biotechnol.* **2024**, *65*, 337–354. [CrossRef]
23. Elanchezhian, A.; Basak, J.K.; Park, J.; Khan, F.; Okyere, F.G.; Lee, Y.; Bhujel, A.; Lee, D.; Sihalath, T.; Kim, H.T. Evaluating Different Models Used for Predicting the Indoor Microclimatic Parameters of a Greenhouse. *Appl. Ecol. Environ. Res.* **2020**, *18*, 2141. [CrossRef]
24. Madhavi, B.G.K.; Basak, J.K.; Paudel, B.; Kim, N.E.; Choi, G.M.; Kim, H.T. Prediction of Strawberry Leaf Color Using RGB Mean Values Based on Soil Physicochemical Parameters Using Machine Learning Models. *Agronomy* **2022**, *12*, 981. [CrossRef]
25. Jaihuni, M.; Basak, J.K.; Khan, F.; Okyere, F.G.; Arulmozhi, E.; Bhujel, A.; Park, J.; Hyun, L.D.; Kim, H.T. A Partially Amended Hybrid Bi-GRU—ARIMA Model (PAHM) for Predicting Solar Irradiance in Short and Very-Short Terms. *Energies* **2020**, *13*, 435. [CrossRef]
26. Larson, D.M.; Bungula, W.; Lee, A.; Stockdill, A.; McKean, C.; Miller, F.F.; Davis, K.; Erickson, R.A.; Hlavacek, E. Reconstructing Missing Data by Comparing Interpolation Techniques: Applications for Long-term Water Quality Data. *Limnol. Oceanogr. Methods* **2023**, *21*, 435–449. [CrossRef]
27. Goad, P.M.; Deore, P.J.; Patil, V.B. A Novel Approach for Detecting Outliers by Using Isolation Forest with Reducing under Fitting Issue. *Preprint* **2022**. [CrossRef]
28. Rodhan, M.; Jaaz, A. Box-Jenkins Modelling and Forecasting of Wti Crude Oil Price. In Proceedings of the 2nd International Multi-Disciplinary Conference Theme: Integrated Sciences and Technologies, IMDC-IST 2021, Sakarya, Turkey, 7–9 September 2021.
29. Gui, J.; Sun, H.; Jia, D.; Yan, T.; Cui, X.; Zhao, L. Forecasting Calling Activity Based on ARIMA Model. In Proceedings of the 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 26–28 May 2023; IEEE: Piscataway, NJ, USA, 2023; Volume 3, pp. 1247–1250.
30. Siami-Namini, S.; Tavakoli, N.; Namin, A.S. A Comparison of ARIMA and LSTM in Forecasting Time Series. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1394–1401.
31. Fowler, C.; Cai, X.; Baker, J.T.; Onnela, J.-P.; Valeri, L. Testing Unit Root Non-Stationarity in the Presence of Missing Data in Univariate Time Series of Mobile Health Studies. *J. R. Stat. Soc. Ser. C Appl. Stat.* **2024**, *73*, 755–773. [CrossRef] [PubMed]
32. Wei, X.; Liu, X.; Fan, Y.; Tan, L.; Liu, Q. A Unified Test for the AR Error Structure of an Autoregressive Model. *Axioms* **2022**, *11*, 690. [CrossRef]
33. Chen, Y.; Wang, K. Prediction of Satellite Time Series Data Based on Long Short Term Memory-Autoregressive Integrated Moving Average Model (LSTM-ARIMA). In Proceedings of the 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 19–21 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 308–312.
34. Nau, R. The Mathematical Structure of Arima Models. *Duke Univ. Online Artic.* **2014**, *1*, 1–8.
35. Kumari, S.; Kumar, N.; Rana, P.S. A Big Data Approach for Demand Response Management in Smart Grid Using the Prophet Model. *Electronics* **2022**, *11*, 2179. [CrossRef]
36. Anand, P.; Sharma, M.; Saroliya, A. A Comparative Analysis of Artificial Neural Networks in Time Series Forecasting Using Arima Vs Prophet. In Proceedings of the 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), Gautam Buddha Nagar, India, 9–11 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 527–533.
37. Mwafulirwa, D. Forecasting Population Demographics in Lilongwe City: Leveraging Prophet and Time Series Analysis Techniques. *Int. J. Emerg. Trends Sci. Technol.* **2024**. [CrossRef]
38. Toharudin, T.; Pontoh, R.S.; Caraka, R.E.; Zahroh, S.; Lee, Y.; Chen, R.C. Employing Long Short-Term Memory and Facebook Prophet Model in Air Temperature Forecasting. *Commun. Stat.-Simul. Comput.* **2023**, *52*, 279–290. [CrossRef]
39. Almazrouee, A.I.; Almeshal, A.M.; Almutairi, A.S.; Alenezi, M.R.; Alhajeri, S.N. Long-Term Forecasting of Electrical Loads in Kuwait Using Prophet and Holt–Winters Models. *Appl. Sci.* **2020**, *10*, 5627. [CrossRef]
40. Riady, S.R. Stock Price Prediction Using Prophet Facebook Algorithm for BBCA and TLKM. *Int. J. Adv. Data Inf. Syst.* **2023**, *4*, 1–8. [CrossRef]

41. Yang, J.-W.; Dashdondov, K. In-Depth Examination of Machine Learning Models for the Prediction of Ground Temperature at Various Depths. *Atmosphere* **2022**, *14*, 68. [CrossRef]
42. Ye, C.; Shen, Z.; Wu, Y.; Loskot, P. Reconsidering Python Syntax to Enhance Programming Productivity. *Int. J. Res. Appl. Sci. Eng. Technol.* **2024**, *12*, 776–785. [CrossRef]
43. Taslim, D.G.; Murwantara, I.M. Comparative Analysis of ARIMA and LSTM for Predicting Fluctuating Time Series Data. *Bull. Electr. Eng. Inform.* **2024**, *13*, 1943–1951. [CrossRef]
44. Kaur, S.; Rakshit, M. Seasonal and Periodic Autoregressive Time Series Models Used for Forecasting Analysis of Rainfall Data. *Int. J. Adv. Res. Eng. Technol.* **2020**, *10*, 2019. [CrossRef]
45. Mehrotra, S.; Tripathi, K.P. Enhancement of Carbon Assimilates and Macronutrients in Legumes under Elevated CO_2 Concentration. *Int. J. Plant Environ.* **2022**, *8*, 52–63. [CrossRef]
46. Bouain, N.; Cho, H.; Sandhu, J.; Tuiwong, P.; Zheng, L.; Shahzad, Z.; Rouached, H. Plant Growth Stimulation by High CO_2 Depends on Phosphorus Homeostasis in Chloroplasts. *Curr. Biol.* **2022**, *32*, 4493–4500. [CrossRef] [PubMed]
47. Kaur, H.; Kumar, A.; Choudhary, A.; Sharma, S.; Choudhary, D.R.; Mehta, S. Effect of Elevated CO_2 on Plant Growth, Active Constituents, and Production. In *Plants and Their Interaction to Environmental Pollution*; Elsevier: London, UK, 2023; pp. 61–77.
48. Kutschera, U.; Ehnes, I. World Climate Declaration: Exhaled Carbon Dioxide Promotes Plant Development. *Eur. J. Environ. Earth Sci.* **2023**, *4*, 1–4. [CrossRef]
49. Madhu, M.; Hatfield, J.L. Dynamics of Plant Root Growth under Increased Atmospheric Carbon Dioxide. *Agron. J.* **2013**, *105*, 657–669. [CrossRef]
50. Parra, J.A.P.; Cruz, O.A.T.; Méndez, Y.L.A. Dispositivo Basado En Modelo Arima Para Predicción de Variables Ambientales (Temperatura, Humedad, Velocidad Del Aire) En El Área Agrícola Del Departamento Del Meta. *Rev. GEON (Gestión Organ. Neg.)* **2020**, *7*, 1–12.
51. Junsuk, K.; Tae, J.K. Application of Facebook's Prophet Model for Forecasting Meteorological Data. *J. Korean Soc. Hazard Mitig.* **2021**, *21*, 53–58.
52. Wang, L.; Yan, Y.; Wang, X.; Wang, T. Input Variable Selection for Data-Driven Models of Coriolis Flowmeters for Two-Phase Flow Measurement. *Meas. Sci. Technol.* **2017**, *28*, 035305. [CrossRef]
53. Gao, J. Time-Series Prediction Research Based on Combined Prophet-LSTM Models. In Proceedings of the 2022 18th International Conference on Computational Intelligence and Security (CIS), Chengdu, China, 16–18 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 143–147.
54. AlOmar, M.K.; Hameed, M.M.; Al-Ansari, N.; Razali, S.F.M.; AlSaadi, M.A. Short-, Medium-, and Long-Term Prediction of Carbon Dioxide Emissions Using Wavelet-Enhanced Extreme Learning Machine. *Civ. Eng. J.* **2023**, *9*, 815–834. [CrossRef]
55. Zhang, Y.; Meng, G. Simulation of an Adaptive Model Based on AIC and BIC ARIMA Predictions. *J. Phys. Conf. Ser.* **2023**, *2449*, 012027. [CrossRef]
56. Primandari, A.H.; Thalib, A.K.; Kesumawati, A. Analysis of Changes in Atmospheric CO_2 Emissions Using Prophet Facebook. *Enthusiastic Int. J. Appl. Stat. Data Sci.* **2022**, 1–9. [CrossRef]
57. Gupta, V.; Pandya, P.; Kataria, T.; Gupta, V.; Roth, D. Multi-Set Inoculation: Assessing Model Robustness Across Multiple Challenge Sets. *arXiv* **2023**, arXiv:2311.08662.