© 2024 Program Studi Ilmu Lingkungan Sekolah Pascasarjana UNDIP
JURNAL ILMU LINGKUNGAN
Volume 22 Issue 6 (2024) : 1670-1678 ISSN 1829-8907

NH4 Modelling with ARIMA and LSTM

Hanna Arini Parhusip[1]*, Suryasatriya Trihandaru[1], dan Johanes Dian Kurniawan[1]

[1]Magister Sains Data, Fakultas Sains dan Matematika, Universitas Kristen Satya Wacana, Indonesia; email: hanna.parhusip@uksw.edu

ABSTRAK

AI-Mining adalah prototipe yang dirancang untuk mendeteksi berbagai gas lingkungan, termasuk CO2, NH3, NH4, dan hidrogen, bersama dengan suhu, tekanan, dan kelembaban. Studi ini menekankan pentingnya pemodelan data seri waktu NH4 karena peran kritisnya dalam pemantauan lingkungan dan kesehatan. Prediksi NH4 yang akurat memfasilitasi deteksi awal polusi dan intervensi kontrol gas rumah kaca yang tepat waktu. Studi ini menyelidiki efektivitas AI-Mining dalam mendeteksi dan memprediksi tingkat gas, dengan fokus pada pengumpulan dan analisis data. Analisis data awal menggunakan model Autoregressive Moving Average (ARIMA), khususnya ARIMA (1,1,1), yang dijelaskan dengan persamaan $y_t = 0.0311 - 0.0750y_{t-1} + 0.3842\varepsilon_{t-1}$. Meskipun digunakan, kinerja ARIMA Root Mean Square Error (RMSE) ditemukan kurang dibandingkan dengan metode yang lebih maju. Mengingat klasifikasi data yang diperoleh sebagai data besar dan seri waktu, metode Long Short Term Memory (LSTM) juga diterapkan. Model LSTM awalnya menggunakan dua lapisan dengan fungsi aktivasi tanh dan relu, dan kinerjanya dieksplorasi lebih lanjut dengan menambahkan lapisan ketiga dan mengubah jumlah neuron. (64, 128, and 256). Optimizer Adam digunakan secara konsisten di semua variasi LSTM. Hasilnya menunjukkan bahwa peningkatan lapisan dan neuron tidak secara signifikan mempengaruhi kinerja LSTM, dengan nilai RMSE sekitar 0.023. Namun, LSTM secara konsisten melampaui ARIMA dalam akurasi prediksi, menekankan ketahanan dan keandalan. Akibatnya, penelitian ini merekomendasikan penggunaan LSTM untuk memprediksi data lain yang tercatat dalam AI-Mining, menekankan keunggulan dalam menangani set data lingkungan yang kompleks.

Kata kunci: NH_4, ARIMA, LSTM, RMSE, deteksi gas

ABSTRACT

AI-Mining is a prototype designed to detect various environmental gases, including CO2, NH3, NH4, and hydrogen, alongside temperature, pressure, and humidity. This study emphasizes the importance of modeling NH4 time series data due to its critical role in environmental and health monitoring. Accurate NH4 predictions facilitate early pollution detection and timely greenhouse gas control interventions. The study investigates the effectiveness of AI-Mining in detecting and predicting gas levels, focusing on data collection and analysis. Initial data analysis employed the Autoregressive Moving Average (ARIMA) model, specifically ARIMA (1,1,1), described by the equation yt = 0.0311 - 0.0750yt-1 + 0.3842εt-1. Despite its use, ARIMA's Root Mean Square Error (RMSE) performance was found lacking compared to more advanced methods. Given the classification of the obtained data as big data and time series, the Long Short-Term Memory (LSTM) method was also applied. The LSTM model initially used two layers with tanh and relu activation functions, and its performance was further explored by adding a third layer and varying the number of neurons (64, 128, and 256). The Adam optimizer was consistently used across all LSTM variations. Results indicated that increasing layers and neurons did not significantly impact LSTM's performance, with RMSE values around 0.023. However, LSTM consistently outperformed ARIMA in prediction accuracy, highlighting its robustness and reliability. Consequently, the study recommends using LSTM for predicting other recorded data in AI-Mining, underscoring its superiority in handling complex environmental datasets.

Keywords: NH_4, ARIMA, LSTM, RMSE, gas detection

Citation: Parhusip, H. A., Trihandaru, S., dan Kurniawan, J. D. (2024). NH4 Modelling with ARIMA and LSTM. Jurnal Ilmu Lingkungan, 22(6), 1670-1678, doi:10.14710/jil.22.6.1670-1678

1. INTRODUCTION

Greenhouse gases, such as carbon dioxide, methane, and water vapor, play a crucial role in regulating the Earth's temperature (Feng et al., 2023). They trap heat radiated from the Earth's surface and keep it within the atmosphere, creating what is known as the greenhouse effect. This natural process helps maintain a relatively stable and habitable temperature range on our planet. If all greenhouse gases were suddenly removed, the Earth would indeed lose this warming effect, resulting in a significantly cooler climate. However, accurately predicting the exact temperature decrease is a complex task due to the numerous interconnected factors that influence global climate patterns. Moreover, the removal of all greenhouse gases would have other implications beyond temperature. It would disrupt ecosystems, alter weather patterns, and impact ocean currents, among other effects. These changes would further complicate any precise estimation of the temperature decrease. In summary, while it is generally understood that the removal of greenhouse gases would lead to a cooler climate, stating an exact temperature decrease is challenging due to the complexity of Earth's climate system and the many variables involved.

There are several types classified as GHG, namely carbon dioxide (CO_2), methane gas (CH_4), nitrogen oxide (N_2O), sulfur hexaflouride (SF_6), perfluorocarbons (PFCS), and hydrofluorocarbons (HFCS). Apart from the six types of gases that have been defined in Presidential Regulation (Indonesian government) No. 98 of 2021 concerning the National Greenhouse Gas Inventory, there are gas elements that can also cause the greenhouse gas effect due to burning waste and pollution, namely NH_4 gas. The compound NH_4 refers to ammonium, which consists of an ammonium ion (NH_4^+) and can combine with other compounds to form various substances (Quan et al., 2023). Ammonium itself is not considered a greenhouse gas. However, it can indirectly influence greenhouse gas emissions and the greenhouse effect through processes such as nitrogen cycling. Based on study by (Xu et al., 2022), which measured the concentrations of NH3, PM2.5, and its water-soluble SNA to determine the effect of NH3 on PM2.5 pollution in an urban area in Jeonju, South Korea, from May 2019 to April 2020, it was shown that during PM2.5 pollution episodes (daily PM2.5 average ≥ 25 µg/m3), there was a remarkable increase in the fraction of NH4+ and NO3− in PM2.5. Meanwhile the other study performed by (Park et al., 2021), showed that the ratio of aerosol NH4+ to total NHx (NH3 + NH4+) concentrations, denoted as e(NH4+), showed increases of about 28% ($P > 0.05$) during the COVID lockdown relative to pre-COVID. Both studies show that modeling NH4 (ammonium) is crucial since accurate NH4 modeling helps predict changes in emissions, which are significant due to their high global warming potential.

In the nitrogen cycle there is a conversion of various forms of nitrogen as the role of ammonium(Stein & Martin, 2016). The release of excess nitrogen compounds, for example, occurs in the use of fertilizers in agricultural processes so that nitrous oxide is released into the atmosphere. This leads to a higher heating potential than carbon dioxide. So even if ammonium is not a greenhouse gas, its potential on the nitrogen cycle can affect greenhouse gas emissions. So efforts are needed in the agricultural process to reduce the use of ammonium in fertilizers on farms(Malhi et al., 2021).

In this research, a prototype tool called AI-Mining has been produced to be able to monitor the presence of ammonia in the environment where AI-Mining is placed where the gas detected is not only ammonia but also greenhouse gases. However, this article shows how NH4 gas is modeled using ARIMA and LSTM as algorithms that have been used by several researchers in studying (Xayasouk et al., 2020); (Erman et al., 2022). in studying time series data.

ARIMA was known before during the research, and we compare the results by implementing LSTM as the newer method for analyzing time series data. We expect to have better results using LSTM than ARIMA.

The choice of ARIMA and LSTM methods for modeling NH4 gas in time series data is justified by their complementary strengths. ARIMA (AutoRegressive Integrated Moving Average) is a classical statistical approach known for its effectiveness in capturing linear patterns and temporal dependencies in time series data, making it suitable for short-term forecasting(Albeladi et al., 2023; Sandhya Arora, 2024)

On the other hand, LSTM (long short-term memory) networks, a type of recurrent neural network, excel at handling non-linear relationships and long-term dependencies(Alotaibi et al., 2023; Saini & Sharma, 2024), which are often present in environmental data. By employing both ARIMA and LSTM, researchers can leverage the strengths of each method to achieve more accurate and robust predictions of NH4 gas levels, accounting for both linear trends and complex, non-linear patterns in the data.

The research contribution is an inventory of gases that have never been done by many parties in areas that have the potential to contribute to gas changes in the environment, so that it is well monitored by AI-Mining as the novelty of this research. Although still in prototype, this prototype can be easily used in various places that have the potential to cause harmful gases that have an impact on climate change, including radiation that can be caused, such as in mining areas. This article is limited to the discussion of the NH_4 gas analysis that has been inventoried in 2022.

2. METHODS

2.1. ARIMA Revisited

Autoregressive (AR) is one technique in regression. As the name suggests, regression is carried out between the variables themselves. In a standard regression, there is a regression that expresses the dependent variable as a function of the independent variable. There is only one variable in an AR if a conventional regression has a regression that represents the dependent variable as a function of the independent variable. The link between the variable and its lag values is modeled by the AR component. The number of lag values that are highly linked with the current value-a hint that a pattern or trend is present in the data is found through autocorrelation. The most common method is to compute the autocorrelation function (ACF) or the partial autocorrelation function (PACF) of the time series (Weiß et al., 2023). The ACF measures the correlation between the variable and its lagged values while accounting for the correlation contributed by the intermediate lags. The PACF, on the other hand, measures the correlation between the variable and its lagged values, but only after removing the correlation contributed by the intermediate lags.

The correlation structures of the time series are plotted using the ACF and PACF. The most important case to consider is the order of the ARIMA Model which is determined by the combined of ACF and PACF. One of the examples is the ARIMA Model of the Covid-19 in Brazil (Ospina et al., 2023). The decays of PCAF which follow the ACF at a particular lag is leading to the AR model. The opposite will give MA model. Finally, one describes that ARIMA model provides the pattern of the time series.

The ARIMA model is typically denoted as ARIMA (*p, d, q*). The parameters *p,d,q* are indicating the order of autoregression, denoted by "*p*," represents the number of lagged values considered in the model; the order of differencing represents the number of differencing operations required to make the series stationary and the order of the moving average, denoted "*q*", represents the number of lagged residuals considered in the model. Once the appropriate values for *p, d*, and *q* are determined based on the analysis of autocorrelation functions (ACF) and partial autocorrelation functions (PACF) (Flores et al., 2012), The model can be estimated using various methods such as maximum likelihood estimation. Finally, the errors can be used to evaluate the model prediction results by the Root Mean Square Error (RMSE)) (Elshewey et al., 2023).

Modeling NH4 time series data is urgent due to its crucial role in environmental monitoring and its indirect effects on greenhouse gas emissions. Understanding NH4 emission patterns aids in comprehending the nitrogen cycle and developing strategies to mitigate environmental impacts. Accurate modeling is essential for predicting trends and implementing effective agricultural and industrial practices to reduce harmful emissions. This research aims to enhance NH4 monitoring accuracy and reliability using ARIMA and LSTM algorithms, improving management of its climate change effects.

The objectives of this research are to: develop and validate a prototype tool, AI-Mining, for monitoring ammonia and greenhouse gases; model NH4 gas emissions using ARIMA and LSTM algorithms and compare their effectiveness; provide an inventory of NH4 emissions in environmentally impactful areas; enhance understanding of the temporal patterns and indirect effects of NH4 emissions on greenhouse gases and the nitrogen cycle; and propose strategies to reduce ammonium use in agriculture to mitigate its environmental impact.

2.2. Long Short Term Memory Revisited

Data science provides many methods to deal of data time series where Long Short-Term Memory (LSTM) is one of the most preferred methods due to its reliability in various applications. This method is one of the methods of artificial neural networks developed, especially on data related to data at a certain period of time (Ilya et al., 2014) (Y. Kumar et al., 2023). From the name, there are short term and long-term memory. This is related to how the zero gradient process in the neural network in the training process and optimization carried out in back propagation(S. Zhang et al., 2020). Some examples of ARIMA being used compared to LSTM can also be found in other literatures (J. Kumar et al., 2018);(Rhanoui et al., 2019);(Paviglianiti et al., 2022).

Because it is one of the artificial intelligence methods, the components of this method are neurons, as well as the connections between neurons called gates, namely the forget gate, the input gate, and the output gate. Initially, neurons will store information that comes in a data set where neurons determine what data will be retained for further processing. Through the input gate, there are two stages, namely the use of the activation function and the sigmoid function, which will determine the information to be updated. Furthermore, the use of the tanh activation function to calculate candidate values to be added to the cell. Then the output gate will process the information on the next layer. We will study the strengths and weaknesses of used methods here according to the data studied in this research to observe how well the AI-Mining works.

2.3. The Used Data

In this study, the data from NH4 was recorded by AI-Mining from May 23, 2022, to July 1, 2022, with AI-Mining located at the university. The data was collected at minute intervals based on the sensor's technical specifications. The total number of data points is 4,094, which is divided into 20% test data (819 data points) and 80% training data (3,275 data points). This division follows recommendations from the general literature and journals that suggest that the sharing of test data and training data use a composition of 20% and 80%. In this study for the LSTM method, two architectural configurations were used, namely, using 2 hidden layers and 3 hidden layers with a combination of the number of neurons from 2 layers, as shown in Table 1. Similarly, the architecture of LSTM with three hidden layers is shown in Table 2.

The selection of the number of neurons in each layer for the LSTM architecture in this study is grounded in established neural network design practices and time series analysis literature. Starting with 64 neurons allows for simpler, computationally efficient models and serves as a performance baseline. Increasing to 128 and 256 neurons enhances the model's capacity to learn complex patterns, preventing overfitting by gradually increasing complexity. Empirical evidence supports these configurations, with studies demonstrating their effectiveness in capturing time series patterns without significant overfitting. Using two and three hidden layers helps capture hierarchical data representations, balancing model complexity and computational efficiency. These configurations align with best practices in recurrent neural networks, ensuring robust and reliable model performance.

Table 1. Combination Configuration Table Number of Neurons (2 layers)

Number of Hidden Layers	Number of Layer 1 Neurons	Number of Layer 2 Neurons
2	64	64
	128	64
	256	64

Table 2. Combination Configuration Table Number of Neurons (3 layers)

Number of Hidden Layers	Number of Layer 1 Neurons	Number of Layer 2 Neurons	Number of Layer 3 Neurons
3	64	64	64
	128	128	64
	256	256	64

3. RESULTS AND DISCUSSION

3.1. Results of ARIMA Model

The simulation results using the ARIMA algorithm, stationery tests, and collinearity were first carried out to build the ARIMA model (Contents, 2018); (Nkongolo, 2023). There are four important indicators for stationery tests and collinearity. These are the statistics: ADF (Augmented Dickey-Fuller) (Harris, 1992), *p*-value, critical values, autocorrelation of residuals. The ADF (Augmented Dickey-Fuller) statistic used in stationary tests. This statistic is used to test the null hypothesis that data has a root (non-stationary) unit. The more negative the ADF statistic value, the stronger the evidence that the data are stationary. The *p*-value is the probability associated with the ADF statistic. It is used to determine whether the null hypothesis (data has root units) can be rejected. Critical values are threshold values used to compare ADF statistics. If the ADF statistic value is more negative than a certain critical value, we can confidently conclude that the data are stationary. According to autocorrelation, autocorrelation is a measure of the correlation between values in a time series with values in the previous time (lag). Using the code, a stationary test is produced, and the output list is shown in Table 3 for the values of these 4 indicators.

Thus, we have the statistics ADF: -17.87305544394411. In this case, a very negative value (-17.87) indicates that the residual data of the ARIMA model under test is stationary. Additionally, the *p*-value= 3.0224283967001075e-30. In this case, the *p*-value is very small (smaller than the significance level of 0.05), so we can reject the null hypothesis and conclude that the residual data is stationary. The critical values at 1%, 5%, and 10% confidence levels are -3.432, -2.862, and -2.567, respectively. Since the ADF statistic (-17.87) is much more negative than all these critical values, we can conclude that the residual data is stationary. In this case, we see autocorrelation of ARIMA model residues at some lag. Lag 0 has a value of 1.0 because autocorrelation at lag 0 is always 1 (correlation between the value and itself). The autocorrelation values that follow show residual correlation at certain lags. A value close to 0 indicates that residues generally do not have a significant correlation with previous lags. Thus, these results show that the residue of the ARIMA model tested is stationary and has no significant correlation with previous lags. The *p*-value of the simulation results is much less than the confidence level of 0.05, and the ADF statistic value is less than the critical value, so that H_0 is rejected, which means that the time series NH_4 value is stationary. To build the ARIMA model, it is necessary to find the values of *p, d*, and *q* from NH_4 value data using the ACF and PACF plot functions to obtain autocorrelation that will be used to determine the value of the *d* and differencing results 1 & 2 from NH_4 value data to determine the values of the *p* & *q.*

Table 3. Indicators for Testing Stationary Data

Indicators	Value
ADF Statistic	-17.87305544394411
p-value	3.0224283967001075e-30
Critical Values:	1%: -3.4323544253126586
	5%: -2.8624256356937496
	10%: -2.5672414426759285
Autocorrelation of Residuals:	Lag 0: 1.0
	Lag 1: -0.0035863767330169957
	Lag 2: -0.06640419225891948
	Lag 3: -0.0010935217885844964
	Lag 4: 0.007962724257967688
	Lag 5: 0.0346850807964653
	Lag 6: 0.009623604679542173
	Lag 7: -0.0026854913216730275
	Lag 8: 0.016063360391485995
	Lag 9: -0.03586069491871277
	Lag 10: 0.05944626754065233

The results of the collinearity test simulation with ACF plots show that the value of the coefficient is stable from lag 1 to lag 11 with all coefficient values above the shading area, so that all maximum coefficients are significant with a value of 1 and have a high autocorrelation as depicted in Figure 1 (right). From the results of the PACF plot simulation as shown in Figures 2 (right) and 3(right) for the first and second differencing processes, the results for the negative coefficient value in the second differencing are smaller than the first differencing results; this shows an overdifference, so the value d = 1 is taken. For the *p* value of the PACF graph for the first differentiation, it appears that there is one highest coefficient that exceeds the significant area or shadow area for lag 0 and lag 1, so that the *p* value is taken as p = 1. The lag value after 1 has a value that begins to decrease so that the *q* value is taken 1. Through ACF and PACF graph analysis, it can be concluded that the model for ARIMA for NH_4 value data is (1,1,1). From the predetermined ARIMA model (1,1,1), the good value of the AIC model is -2586,236. Through ACF and PACF graph analysis, the model for ARIMA for NH_4 value data is (1,1,1). Summary of those results is shown in Figure 4. The ARIMA *(*1,1,1) model equation can be written, i.e.:

$$y_t = 0{,}0311 - 0{,}0750\,y_{t-1} - 0{,}3842\,\varepsilon_{t-1}$$

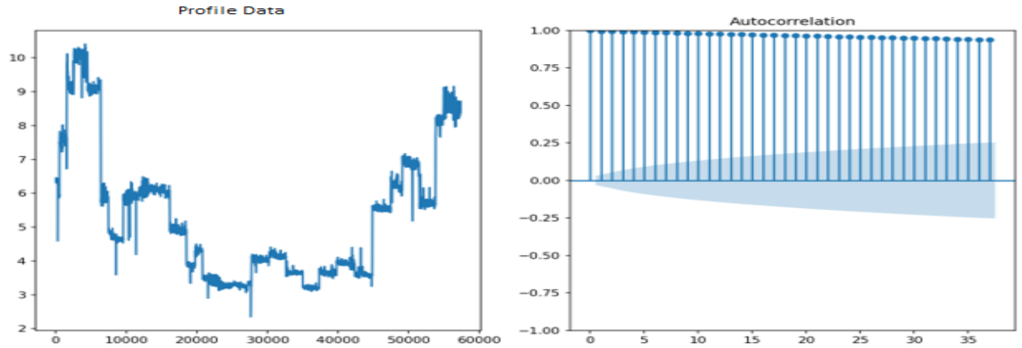


Figure 1. The Profile Data of NH4 (Left) and its Autocorrelation (Right)

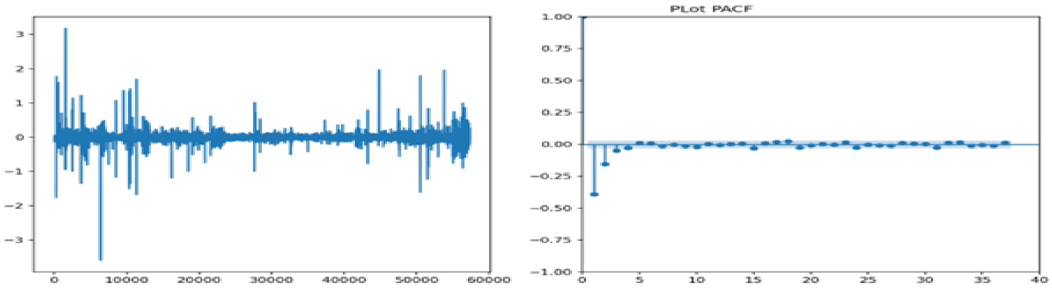


Figure 2. First Differencing Plot (Right)

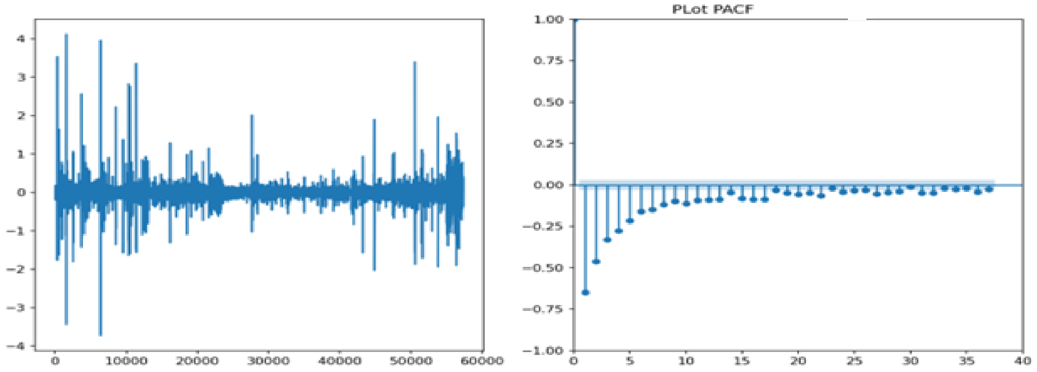
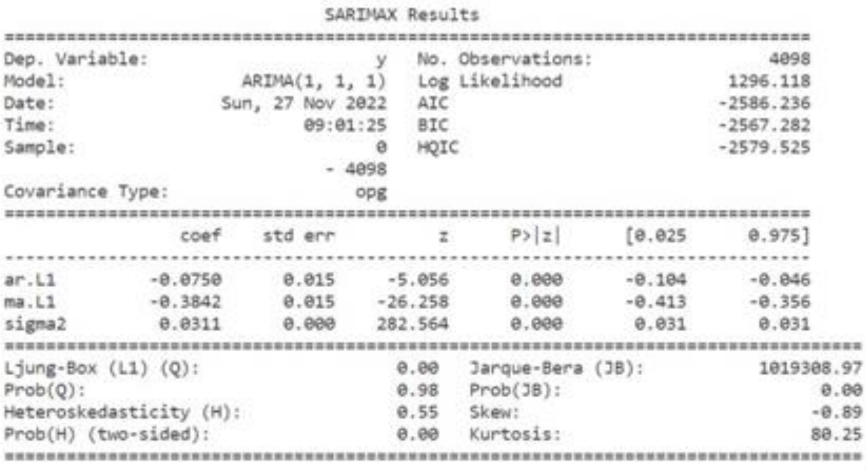


Figure 3. Second Differencing Plot Image



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 4098
Model:                 ARIMA(1, 1, 1)   Log Likelihood                1296.118
Date:                Sun, 27 Nov 2022   AIC                          -2586.236
Time:                        09:01:25   BIC                          -2567.282
Sample:                             0   HQIC                         -2579.525
                               - 4098
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0750      0.015     -5.056      0.000      -0.104      -0.046
ma.L1         -0.3842      0.015    -26.258      0.000      -0.413      -0.356
sigma2         0.0311      0.000    282.564      0.000       0.031       0.031
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):            1019308.97
Prob(Q):                              0.98   Prob(JB):                          0.00
Heteroskedasticity (H):               0.55   Skew:                             -0.89
Prob(H) (two-sided):                  0.00   Kurtosis:                         80.25
===================================================================================
```

Figure 4. SARIMAX Results of NH_4 Data

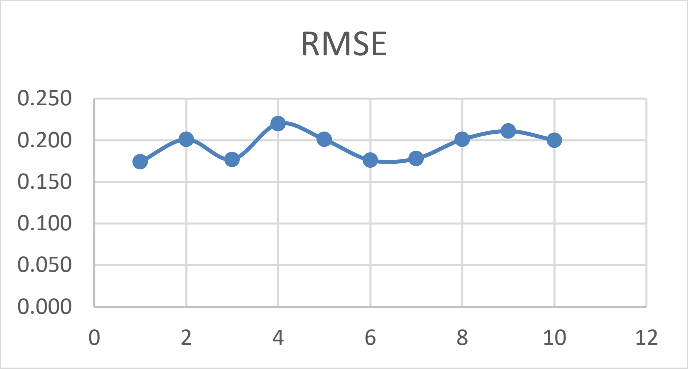


Figure 5. RMSE results

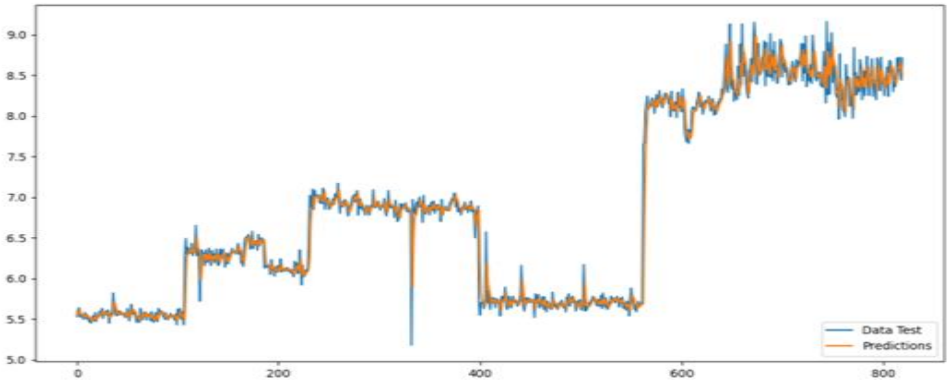


Figure 6. ARIMA Prediction Results

Table 4. RMSE LSTM Value Results 2 Layers using Adam Optimizer

Activation Function	Number of Hidden Layer	Neuron Number	Neuron Number	Average RMSE
Tanh	2	Layer 1	Layer 2	
		64	64	0.022
		64	128	0.023
		64	256	0.024
Relu	2	64	64	0.025
		64	128	0.024
		64	256	0.021

Table 5. Results RMSE LSTM Values using 3 Layers

Activation Function	Number of Hidden Layers	Neuron Number	Neuron Number	Neuron Number	Average RMSE
Tanh	3	Layer 1	Layer 2	Layer 3	
		64	64	64	0.200
		128	128	64	0.027
		256	128	64	0.026
Relu	3	64	64	64	0.021
		128	128	64	0.019
		256	128	64	0.023

The results of the simulation for the RMSE value showed an average RMSE value of 0.1939 from 10 experiments, as depicted in Figure 5. Futhermore, the ARIMA model using 819 test data can be depicted in Figure 6.

3.2. Results on LSTM

Using the LSTM algorithm, 6 configurations were used, and each configuration was carried out 10 times using the number of neurons 64, 128 and 256, the number of hidden layers 2 and 3 layers and the number of epochs 100 times to get the RMSE value. The Adam optimizer is used, while tanh and relu are the activation functions. The summary results of the average experiments conducted on the LSTM method are shown on Tables 5 and 6. The average RMSE values in RMSE LSTM in 2 layer and 3 layers configurations are not significantly different. Furthermore, the addition of the number of neurons in the LSTM method have no significant effect on the value of RMSE. It is depicted from the Tables that the average RMSE value for each neuron configuration tends to be the same value of 0.023 for 2 layers and 0.021 for 3 layers. Additionally, the 2 different activations functions have been employed, i.e. tanh or relu activations. The results of the lost function for experiments with 2 layers and 3 layers using the relu activation function using epoch 100 times show a significant decrease in RMSE value so that looping with epoch 100 times can represent the value of lost function data. This is depicted in Figures 7 and 8. Using 20% of the data or a total of 819 data, the plot prediction results were obtained and depicted in Figures 9 and 10.

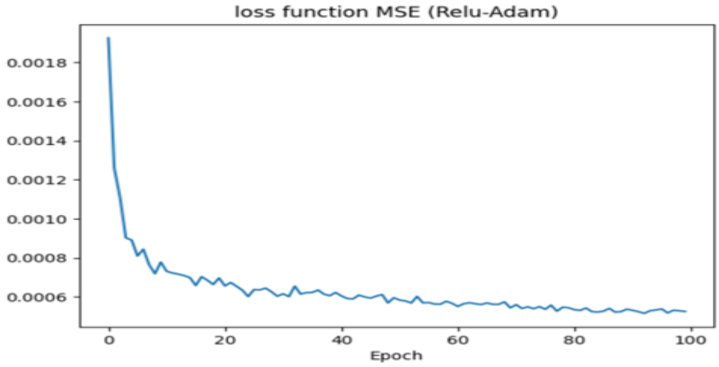


Figure 7. Lost Function Profile using 2 Layers in LSTM with 100 epochs

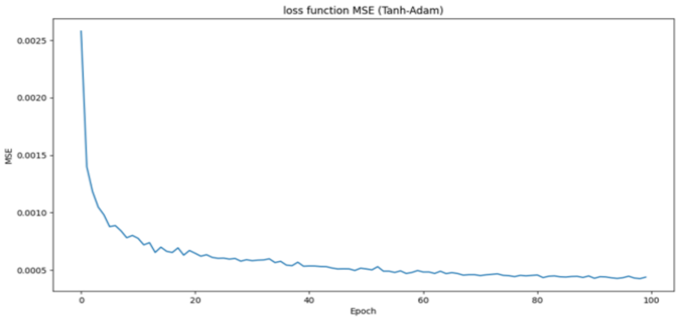


Figure 8. Lost Function Profile using 3 Layers in LSTM with 100 Epochs

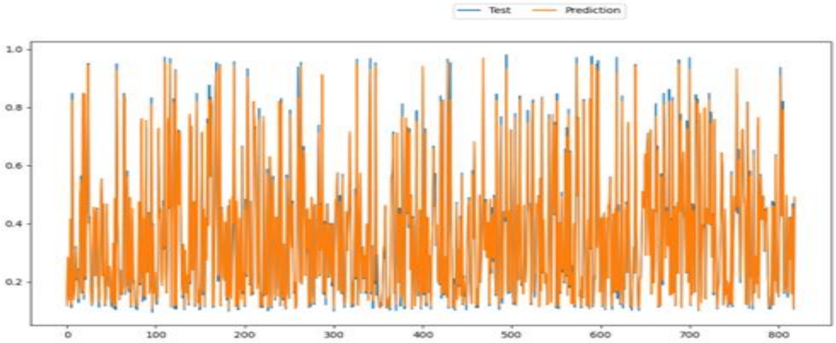


Figure 9. Plot Prediction Data and Test Using Model of LSTM with 2 *Layers*

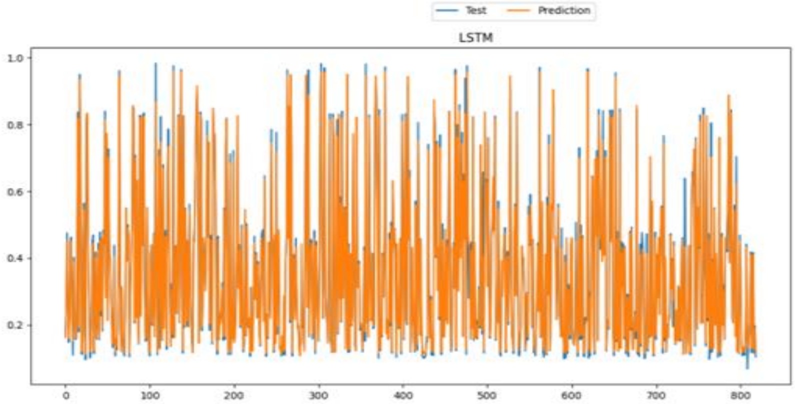


Figure 10. Plot Prediction Data and Test Using Model of LSTM with 3 *Layers*

From the results shown in Figures 10 the prediction result has almost the same profiles between the prediction result (orange color) and the original data (blue color) if the LSTM is 2 layers and 3 layers. Similarly, we observe that the prediction result (denoted by the orange curve) has almost the same shape as the test data (blue curve). This reinforces the result of the average RMSE value, which shows a relatively low value of 0.023, and this means that the LSTM method can predict using 2 layers or 3 layers.

The stable RMSE value is influenced by the use of Adam's optimizer function based on stochastic gradient descent (R. Zhang et al., 2022) which is able to minimize the RMSE value by minimizing loss when adding layers. It can be seen that the prediction results (orange) from ARIMA are not fully able to predict the data, especially when the data experiences a sharp increase and decrease, so that the prediction results are not as good as from the LSTM algorithm.

The NH4 modeling results have significant environmental implications. Accurate NH4 monitoring and prediction can help identify and mitigate sources of emissions, which indirectly contribute to greenhouse gas effects through processes like nitrification and denitrification. Understanding NH4 emissions patterns enables targeted strategies in agriculture and industry to reduce nitrogen compound release, thereby decreasing the overall greenhouse gas footprint. This research highlights the importance of robust environmental monitoring tools, such as AI- mining, to manage and mitigate the impacts of harmful emissions on climate change. Enhanced predictive models like LSTM can facilitate timely and effective interventions, promote sustainable environmental practices, and contribute to global climate goals.

4. CONCLUSION

This study examines NH4 data acquisition and analysis using the AI-Mining tool, comparing ARIMA and LSTM models. The LSTM model demonstrated superior predictive accuracy, with an average RMSE of 0.023, compared to ARIMA's 0.1939. The ARIMA model was identified as ARIMA (1,1,1) with an AIC value of -2586.236. Despite the LSTM's better performance, layer and neuron configuration variations did not significantly affect the RMSE. The AI-Mining tool effectively recorded and processed NH4 data, but the study did not comprehensively address other gases like CO2, N2O, and H2. Future research should include a broader range of gases, explore additional machine learning models, enhance AI-Mining's data capture capabilities, and investigate the long-term environmental impacts of NH4 emissions to develop robust predictive models and mitigation strategies.

Acknowledgments

This research is the result of research in 2021 and 2022 from the Matching Fund research activity funded by the Ministry of Education and Culture and Technology in UKSW which produced a product called AI-Mining.

REFERENCES

Albeladi, K., Zafar, B., & Mueen, A. (2023). Time Series Forecasting using LSTM and ARIMA. *International Journal of Advanced Computer Science and Applications*, *14*(1), 313–320. https://doi.org/10.14569/IJACSA.2023.0140133

Alotaibi, N. D., Jahanshahi, H., Yao, Q., Mou, J., & Bekiros, S. (2023). An Ensemble of Long Short-Term Memory Networks with an Attention Mechanism for Upper Limb Electromyography Signal Classification. *Mathematics*, *11*(18). https://doi.org/10.3390/math11184004

Contents, T. O. F. (2018). Author Information Pack. *Advances in Accounting*, *42*, I–VIII. https://doi.org/10.1016/s0882-6110(18)30184-6

Elshewey, A. M., Shams, M. Y., Elhady, A. M., Shohieb, S. M., Abdelhamid, A. A., Ibrahim, A., & Tarek, Z. (2023). A Novel WD-SARIMAX Model for Temperature Forecasting Using Daily Delhi Climate Dataset. *Sustainability (Switzerland)*, *15*(1), 1–15. https://doi.org/10.3390/su15010757

Erman, M., Anand, S., Adil, Sahu, A., & Arqim, M. (2022). Comparisons of Autoregressive Integrated Moving Average (ARIMA) and Long Short Term Memory (LSTM) Network Models for Ionospheric Anomalies Detection: a Study on Haiti (Mw = 7.0) earthquake. *Acta Geodaetica et Geophysica*, *57*, 195–213. https://doi.org/https://doi.org/10.1007/s40328-021-00371-3

Feng, Y., Li, L. Z. X., Wu, J., Piao, S., Chen, A., & Zeng, Z. (2023). Earth greening mitigates hot temperature extremes despite the effect being dampened by rising CO 2 •. *One Earth*, *3322*(23), 1–22. https://doi.org/https://doi.org/10.1016/j.oneear.2023.12.003

Flores, J. H. F., Engel, P. M., & Pinto, R. C. (2012). Autocorrelation and partial autocorrelation functions to improve neural networks models on univariate time series forecasting. *Proceedings of the International Joint Conference on Neural Networks*, *June*. https://doi.org/10.1109/IJCNN.2012.6252470

Harris, R. I. D. (1992). Testing for unit roots using the augmented Dickey-Fuller test. Some issues relating to the size, power and the lag structure of the test. *Economics Letters*, *38*(4), 381–386. https://doi.org/10.1016/0165-1765(92)90022-Q

Ilya, S., Oriol, V., & Quoc V, L. A. (2014). Sequence to sequence learning with neural networks. *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 3104–3112. https://doi.org/10.5555/2969033.2969173

Kumar, J., Goomer, R., & Singh, A. K. (2018). Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model for Cloud Datacenters. *Procedia Computer Science*, *125*, 676–682. https://doi.org/10.1016/j.procs.2017.12.087

Kumar, Y., Koul, A., Kaur, S., & Hu, Y. C. (2023). Machine Learning and Deep Learning Based Time Series Prediction and Forecasting of Ten Nations' COVID-19 Pandemic. *SN Computer Science*, *4*(1), 1–27. https://doi.org/10.1007/s42979-022-01493-3

Malhi, G. S., Kaur, M., & Kaushik, P. (2021). Impact of climate change on agriculture and its mitigation strategies: A review. *Sustainability (Switzerland)*, *13*(3), 1–21. https://doi.org/10.3390/su13031318

Nkongolo, M. (2023). Using ARIMA to Predict the Growth in the Subscriber Data Usage. *Eng*, *4*(1), 92–120. https://doi.org/10.3390/eng4010006

Ospina, R., Gondim, J. A. M., Leiva, V., & Castro, C. (2023). An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil. *Mathematics*, *11*(14), 1–18. https://doi.org/10.3390/math11143069

Park, J., Kim, E., Oh, S., Kim, H., Kim, S., Kim, Y. P., & Song, M. (2021). Contributions of ammonia to high concentrations of pm2.5 in an urban area. *Atmosphere*, *12*(12). https://doi.org/10.3390/atmos12121676

Paviglianiti, A., Randazzo, V., Villata, S., Cirrincione, G., & Pasero, E. (2022). A Comparison of Deep Learning Techniques for Arterial Blood Pressure Prediction. *Cognitive Computation*, *14*, 1689–1710. https://doi.org/10.1007/s12559-021-09910-0

Quan, F., Zhan, G., Zhou, B., Ling, C., Wang, X., Shen, W., Li, J., Jia, F., & Zhang, L. (2023). Electrochemical removal of ammonium nitrogen in high efficiency and N2 selectivity using non-noble single-atomic iron catalyst. *Journal of Environmental Sciences (China)*, *125*(March), 544–552. https://doi.org/10.1016/j.jes.2022.03.004

Rhanoui, M., Yousfi, S., Mikram, M., & Merizak, H. (2019). Forecasting financial budget time series: Arima random walk vs lstm neural network. *IAES International Journal of Artificial Intelligence*, *8*(4), 317–327. https://doi.org/10.11591/ijai.v8.i4.pp317-327

Saini, K. P., & Sharma, A. (2024). *A Comparison Between Long Short-Term Memory And Prophet For Time Series Analysis And Forecasting Technique*. *30*(4), 8760–8765. https://doi.org/10.53555/kuey.v30i4.2816

Sandhya Arora, M. K. (2024). Forecasting the Future: A Comprehensive Review of Time Series Prediction Techniques. *Journal of Electrical Systems*, *20*(2s), 575–586. https://doi.org/10.52783/jes.1478

Stein, L. Y., & Martin, G. K. (2016). The Nitrogen Cycle. *Current Biology*, *26*(3), R94–R98. https://doi.org/https://doi.org/10.1016/j.cub.2015.12.021

Weiß, C. H., Aleksandrov, B., Faymonville, M., & Jentsch, C. (2023). Partial Autocorrelation Diagnostics for Count Time Series. *Entropy*, *25*(1), 1–21. https://doi.org/10.3390/e25010105

Xayasouk, T., Lee, H. M., & Lee, G. (2020). Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. *Sustainability (Switzerland)*, *12*(6). https://doi.org/10.3390/su12062570

Xu, W., Zhao, Y., Wen, Z., Chang, Y., Pan, Y., Sun, Y., Ma, X., Sha, Z., Li, Z., Kang, J., Liu, L., Tang, A., Wang, K., Zhang, Y., Guo, Y., Zhang, L., Sheng, L., Zhang, X., Gu, B., … Liu, X. (2022). Increasing importance of ammonia emission abatement in PM2.5 pollution control. *Science Bulletin*, *67*(17), 1745–1749. https://doi.org/10.1016/j.scib.2022.07.021

Zhang, R., Song, H., Chen, Q., Wang, Y., Wang, S., & Li, Y. (2022). Comparison of ARIMA and LSTM for prediction of hemorrhagic fever at different time scales in China. *PLoS ONE*, *17*(1 January 2022). https://doi.org/10.1371/journal.pone.0262009

Zhang, S., Lin, M., Zou, X., Su, S., Zhang, W., Zhang, X., & Guo, Z. (2020). LSTM-based air quality predicted model for large cities in China. *Nature Environment and Pollution Technology*, *19*(1), 229–236.