Short-term Korea East-sea Temperature Forecasting Approach based on Seq2Seq Model using Multi Parameters

1st Daeseung Park
KAIST Convergence Research Center
for College of Engineering
Daejeon, Republic of Korea
dspark@kaist.ac.kr

2nd A-Ryoung Kim
KAIST Convergence Research Center
for College of Engineering
Daejeon, Republic of Korea
aryoung9622@kaist.ac.kr

3rd Chae-Seok Lee*
KAIST Convergence Research Center
for College of Engineering
Daejeon, Republic of Korea
quarry@kaist.ac.kr

4th Ho-jong Chang*
KAIST Convergence Research Center
for College of Engineering
Daejeon, Republic of Korea
hojoungc@kaist.ac.kr

***Abstract*—Recently, sea surface temperatures worldwide have been recording the highest levels in the historical records of observations. Scholars from various fields are expressing great concern over the heightened sea surface temperatures and the potential aftermath they might bring. Sea surface temperature affects global climate change, marine ecosystems and marine disasters[1]–[3]. Therefore, researching sea surface temperature prediction models is a crucial endeavor. Generally, RNN models such as LSTM and GRU are used for time series forecasting. While Seq2Seq models are primarily used in Natural Language Processing, there are notable cases where they have shown remarkable results in time series forecasting[4]–[6]. In this study, utilizes a Seq2Seq model as the foundation for sea temperature predictions. Additionally, analyzes sea surface temperature data from the East Sea over the past decade to identify external factors with high correlations. Furthermore, examines and proposes methods to create a better prediction model by comparing and analyzing the inputs of the encoder and decoder.**

***Index Terms*—AI, Deep-Learning, Model, Seq2Seq, Forecasting, Time-series, Prediction**

I. Introduction

Recently, sea surface temperatures worldwide have been reaching historically unprecedented levels in the records of observations. Scholars from various fields are deeply concerned about the elevated sea surface temperature and the potential aftermath of heightened sea surface temperature. Sea surface temperature affects global climate change, marine ecosystems and marine disasters. Furthermore, global climate change is raising sea surface temperatures. This interplay could lead to a mutually reinforcing process of global warming. Ultimately resulting in an irreversible climate change. And it is already becoming a reality[1]–[3].

As a result, the academic community is actively researching various approaches to predict sea surface temperatures. In addition, sea surface temperature prediction research is becoming increasingly important and urgent.

Generally, RNN models such as LSTM and GRU are used for time series forecasting. On the other hand, Seq2Seq is mainly used in natural language processing because of its model characteristics[7]. Therefore, this architectural advantage can yield high performance when predicting target data based on multiple variables. Also, there are notable cases where they have shown remarkable results in time series forecasting[4]–[6].

In the case of sea surface temperature, there exists a strong correlation between the surface depth and temperature. Particularly in the West Sea of Korea, the correlation coefficient between surface temperature and temperature is over 0.92, indicating a high correlation[8].

This study analyzes sea surface temperature data from the East Sea over the past decade to identify external factors with high correlations. Furthermore, examines and proposes methods to create a better prediction model by comparing and analyzing the inputs of the encoder and decoder. Through experiments, it is demonstrated that the multivariate-based Seq2Seq model shows improved prediction performance than the single-variable based Seq2Seq model.

II. External Factor Variable Correlation Analysis

In this paper, the objective is to predict sea surface temperatures, especially along the coast of Busan and Pohang in the East Sea of Korea. While there are studies that have explored the correlation between sea surface temperature and temperature in the West Sea[8], the research on correlation in the East Sea remains limited. Therefore, prior to constructing the prediction model, an analysis of Spearman correlation and Random Forest variable importance was conducted on a total of 12 variables, including sea surface temperature, temperature, pressure, humidity, wind speed, wind direction,

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

GUST, maximum wave height, significant wave height, mean wave height, wave period, and wave direction.

$$r_s = \frac{\sum_{i=1}^{n}(i - \frac{n+1}{2})(R_i - \overline{R})}{\sqrt{\sum_{i=1}^{n}(i - \frac{n+1}{2})^2}\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2}} \tag{1}$$

Spearman correlation is utilized as a measure to assess the statistical dependence between two variables[9]. Following equation (1), it quantifies the simple relationship between the two variables.

$$I_j = \left| \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x_j})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \right| \tag{2}$$

The analysis of correlation in Random Forest variable importance is proposed by leveraging the fact that more important variables tend to have larger correlations with the response variable[10]. Following equation (2), it quantifies the importance relationship between the two variables.

A. Spearman Correlation Analysis

First, in the analysis of correlations among external factors, as shown in Fig. 1, temperature, humidity, and air pressure exhibited correlations with sea surface temperature in that order. In particular, temperature showed a high correlation value of 0.87.

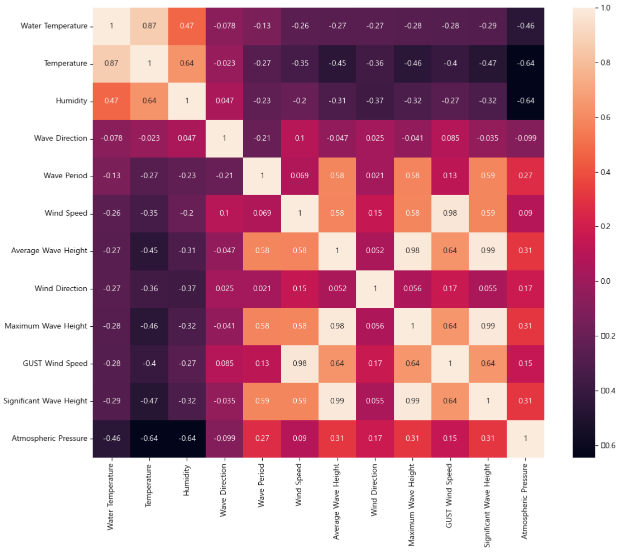

Fig. 1. Spearman Correlation Analysis.

The reason air pressure was ranked fourth in Fig. 1 is that it had a negative value of -0.46, indicating an inverse relationship with sea surface temperature. In correlation analysis, inverse relationships also hold significance, so the analysis was conducted based on absolute values. Consequently, air pressure was ranked fourth in terms of correlation.

B. Random Forest Variable Importance Correlation Analysis

Similarly, in the analysis of Random Forest variable importance, as shown in Fig. 2, temperature exhibited the highest correlation with sea surface temperature. However, the 2nd to 4th ranked correlations differed from those of Spearman Correlation. In the Random Forest variable importance analysis, the order of importance for the 2nd to 4th ranked correlations was industry, wind direction, and humidity.

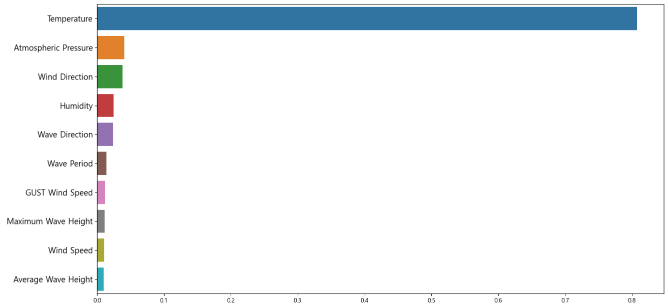

Fig. 2. Random Forest Analysis.

Based on the analyses of Spearman Correlation and Random Forest variable importance, it can be anticipated that including temperature, which shows common correlations, along with sea surface temperature in the learning process could lead to improved prediction performance.

III. Implementation Details

To implement the predictive model, the hardware configuration was structured as shown in Table I. The development environment and library composition for conducting experiments with the prediction model are outlined in Table II.

TABLE I
HARDWARE SPECIFICATIONS

	Specifications
CPU	AMD Ryzen 9 5950X 16-Core Processor 3.40 GHz
RAM	DDR4 64GB 3200MHz
GPU	NVIDIA GeForce RTX 3060 GDDR6 12GB VRAM
OS	Windows 10 Pro 22H2

TABLE II
DEVELOPMENT ENVIRONMENTS

	Environments
IDE	PyCharm 2023.1.2
Interpreter	Anaconda 3
Base Interpreter	Python 3.8
Tensorflow-GPU	2.10.0
Scikit-learn	1.2.2
Pandas	1.5.3
Numpy	1.24.3

The conceptual diagram of sea surface temperature Seq2Seq model used in the experiments is presented in Fig. 3. A total of 9 years' worth of hourly data from 2012 to 2021 were utilized for training. In time series learning, for short-term forecasting, it is effective to learn from a short period of the past and predict a short period into the future. Therefore, the Sequence Length was set to 9 and the Label Length was set to 3. In other words, the model was designed to learn from the past 9 hours to predict the next 3 hours.

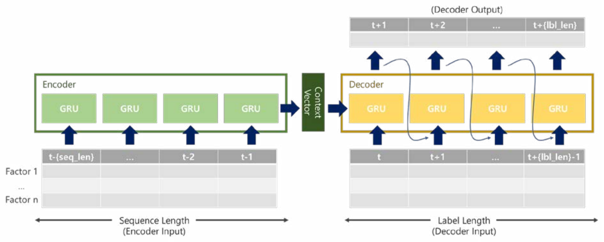

Fig. 3. Seq2Seq Concept Design.

Based on Fig. 3, the actual implemented model structure is depicted in Fig. 4. ”input_1” serves as the Encoder Input, which receives historical data. Upon input to ”input_1,” the ”rnn” generates the Encoder Status. ”input_2” functions as the Decoder Input, allowing additional input of future forecast values. In ”rnn_1,” the Decoder generates prediction values based on the Encoder Status and Decoder Input.

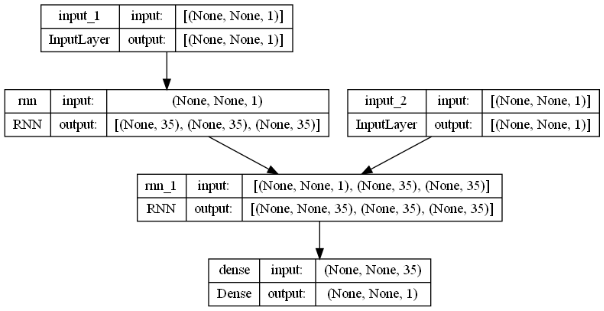

Fig. 4. Seq2Seq Model Structure.

IV. Experimental Results

In Fig. 5, seawater temperature and air temperature were Min-Max Scaled to values between 0 and 1 for training. As a result, it was observed that there is a consistent pattern and cycle in temperature variations each year, indicating a high correlation.

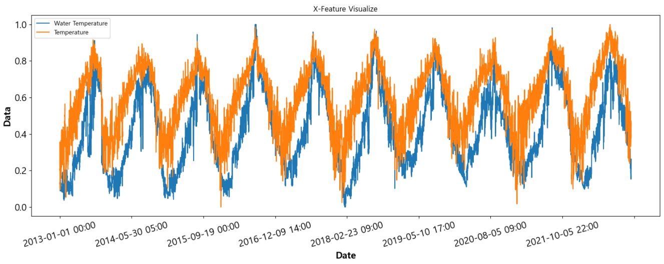

Fig. 5. Min-Max Scaled Graph of Temperature and Water Temperature.

Based on the Min-Max Scaled Data, training was conducted with 25 epochs and a batch size of 32. The results are shown for the 2022 one-year prediction period in Fig. 6 and Fig. 7. In Fig. 6, the prediction results were obtained from training using only the Water Temperature single variable, resulting in an RMSE of 0.2389. In Fig. 7, the results were obtained from training using both Water Temperature and Temperature variables, resulting in an RMSE of 0.2376, indicating an improvement in prediction performance. This improvement is attributed to the inclusion of temperature, which has a high correlation with seawater temperature. Notably, temperature changes more rapidly than seawater temperature, serving as an indicator for early changes and leading to enhanced performance.

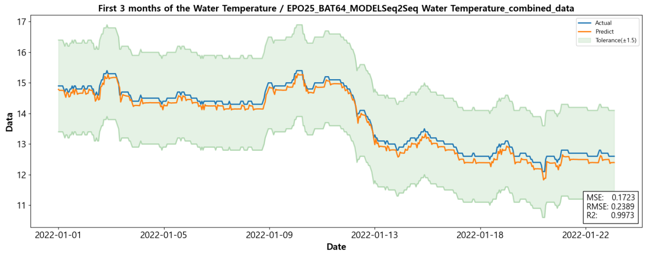

Fig. 6. 1 Month Prediction Result of Water Temperature Input Parameter.

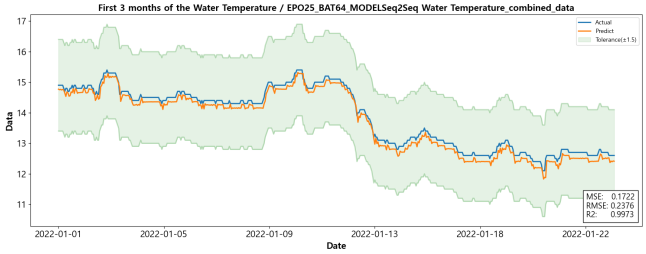

Fig. 7. 1 Month Prediction Result of Water Temperature and Temperature Input Parameter.

Fig. 8 demonstrates excellent prediction results for the one-year period, including during periods of significant temperature changes in the summer. This improved performance is achieved by using a shorter Sequence Length to reduce the training period, enhancing responsiveness to larger temperature variations.

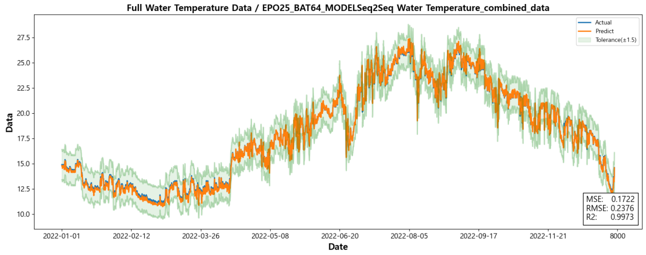

Fig. 8. 1 Year Prediction Result of Water Temperature and Temperature Input Parameter.

Fig. 9 represents the distribution of predicted values against actual values. The red represents actual values, blue represents predicted values, and green represents the error range. The majority of prediction results are distributed within the error range of +-1.5.

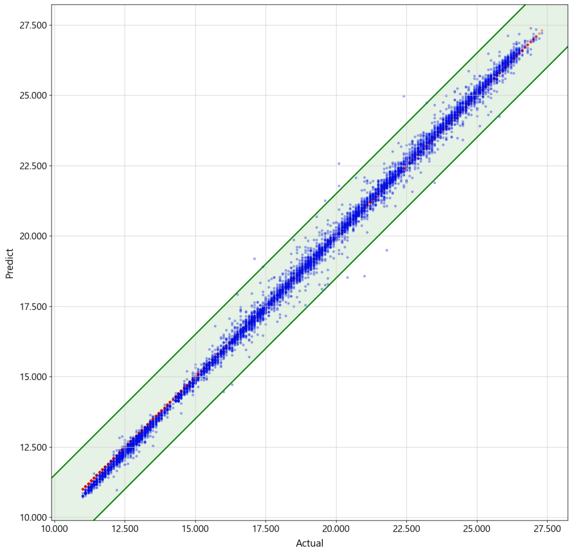

Fig. 9. 1 Year Prediction Result Parity Graph.

Table III provides an overview of the entire experimental results. When seawater temperature and air temperature were used as Encoder Input, the model achieved the best result with an RMSE of 0.2376. However, in models that utilized Decoder Input, which assumed forecasted data for seawater temperature and air temperature, meaningful results were not obtained. This can be attributed to the gradient vanishing problem due to long-term data in the training set over 9 years[11].

Consequently, through additional experiments, when sea surface temperature was used as Encoder Input and air temperature as Decoder Input, results showed an RMSE of 0.5795 for the 9-year training scenario. Conversely, when the training data was limited to 3 years, the result improved to an RMSE of 0.1928, confirming that the issue was related to gradient vanishing problem[11].

TABLE III
TEST RESULT

Test Case		**Result Score**		
Encoder-I	**Decoder-I**	***MAE***	***RMSE***	***R2***
Water-Temp	Empty	0.1722	0.2389	0.9973
Water-Temp	Water-Temp	7.9676	8.8647	-2.7078
Water-Temp, Temp	Empty	0.1721	0.2376	0.9973
Water-Temp, Temp	Water-Temp, Temp	0.6070	0.8187	0.9683

V. Conclusion

In this paper, we conducted research and experiments on a seawater temperature prediction model for the East Sea of the Korean Sea, based on a Seq2Seq model using multiple variables. The experimental results showed that the single-variable prediction model had an RMSE of 0.2389, while the multiple-variable prediction model exhibited an RMSE of 0.2376, indicating an improvement of approximately 0.55% in performance. Furthermore, for the forecast data-based prediction model using Decoder Input, meaningful results could not be obtained in this study due to the gradient vanishing problem caused by long-term training data. However, additional experiments demonstrated that by changing to short-term data, the issue of gradient vanishing was resolved, and normal performance was achieved[11]. Moreover, compared to previous studies[4]–[6], this research confirmed an improvement of approximately 43% in prediction performance over the demonstrated RMSE of 0.34. Through this study, we anticipate that the predictive model developed can contribute to predicting temperature changes in the East Sea of the Korean Sea and forecasting various weather anomalies resulting from climate change and global warming. Future research will need to focus on researching prediction models that utilize short-term data to address the performance degradation caused by the gradient vanishing problem.

Acknowledgment

This research was supported by the Government-wide R&D Fund for Infections Disease Research (GFID), funded by the Ministry of the Interior and Safety, Republic of Korea. (grant number: 20016180).

References

[1] E. J. Rohling, G. L. Foster, K. M. Grant, G. Marino, A. P. Roberts, M. E. Tamisiea, and F. Williams, “Sea-level and deep-sea-temperature variability over the past 5.3 million years,” *Nature*, vol. 508, no. 7497, pp. 477–482, Apr 2014. [Online]. Available: https://doi.org/10.1038/nature13230

[2] T. Geng, F. Jia, W. Cai, L. Wu, B. Gan, Z. Jing, S. Li, and M. J. McPhaden, “Increased occurrences of consecutive la niña events under global warming,” *Nature*, vol. 619, no. 7971, pp. 774–781, Jul 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06236-9

[3] J. Song, G. Tong, J. Chao, J. Chung, M. Zhang, W. Lin, T. Zhang, P. M. Bentler, and W. Zhu, “Data driven pathway analysis and forecast of global warming and sea level rise,” *Scientific Reports*, vol. 13, no. 1, p. 5536, Apr 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-30789-4

[4] Q. He, W. Li, Z. Hao, G. Liu, D. Huang, W. Song, H. Xu, F. Alqahtani, and J.-U. Kim, “A tma-seq2seq network for multi-factor time series sea surface temperature prediction,” *Computers, Materials & Continua*, vol. 73, no. 1, pp. 51–67, 2022. [Online]. Available: http://www.techscience.com/cmc/v73n1/47792

[5] S. F. Stefenon, L. O. Seman, L. S. Aquino, and L. dos Santos Coelho, “Wavelet-seq2seq-lstm with attention for time series forecasting of level of dams in hydroelectric power plants,” *Energy*, vol. 274, p. 127350, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544223007442

[6] Z. Masood, R. Gantassi, Ardiansyah, and Y. Choi, “A multi-step time-series clustering-based seq2seq lstm learning for a single household electricity load forecasting,” *Energies*, vol. 15, no. 7, 2022. [Online]. Available: https://www.mdpi.com/1996-1073/15/7/2623

[7] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 3104–3112.

[8] J. Kim and D. Kim, "A study on correlations between sea surface temperature and air-temperature of the yellow sea of korea," *The Journal of Korean Island*, vol. 30, no. 1, pp. 151–172, 2018.

[9] C. Spearman, "The proof and measurement of association between two things," *International Journal of Epidemiology*, vol. 39, no. 5, pp. 1137–1150, 10 2010. [Online]. Available: https://doi.org/10.1093/ije/dyq191

[10] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, vol. 27, no. 3, pp. 659–678, May 2017. [Online]. Available: https://doi.org/10.1007/s11222-016-9646-1

[11] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, vol. 6, pp. 107–116, 1998. [Online]. Available: https://api.semanticscholar.org/CorpusID:18452318