AI AND MACHINE LEARNING TO EXTEND METEO-MARINE STATION OBSERVATIONS INTO THE FUTURE

Joel Azzopardi

Abstract: The real-time availability of data from coastal meteo-marine stations is crucial for various stakeholders, including port authorities, government agencies, researchers, and the general public. While observation data is fundamental, short-term forecasts can significantly enhance planning and decision-making processes. This study explores the application of Machine Learning (ML) techniques to predict hourly values of air temperature, wind speed, atmospheric pressure, and humidity for the next 24 hours. We evaluate three ML models: Long Short-Term Memory Network (LSTM), Random Forest (RF), and Multivariate Linear Regression (LR). The models were trained using Python libraries and Optuna for hyperparameter tuning on datasets of varying lengths from stations in the Malta-Sicily channel. Additionally, we investigated transfer learning with the ERA5 dataset, which provides hourly values over an 83-year period, to address the challenge of limited data availability. The results show that models trained on longer datasets generally achieve better performance. Furthermore, the models demonstrated considerable generalizability, particularly across nearby stations, allowing models trained at one station to be effectively used for predictions at other proximate stations. To support further research and practical application, we have made our models and tools publicly available.

Keywords: Machine Learning, Artificial Intelligence, Transfer Learning, Meteorology, Prediction

Joel Azzopardi, University of Malta, Malta, joel.azzopardi@um.edu.mt, 0000-0001-6709-8530

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Joel Azzopardi, *AI and Machine Learning to extend Meteo-Marine Station Observations into the Future*, pp. 846-857, © 2024 Author(s), CC BY-NC-SA 4.0, DOI: 10.36253/979-12-215-0556-6.73

Introduction

The real-time availability of data from coastal meteo-marine stations is becoming increasingly important. This data is indispensable for stakeholders such as port authorities, government agencies, researchers, and the public. While real-time (now-cast) data is essential, short-term forecasts for the upcoming hours would provide significant additional benefits. The integration of Artificial Intelligence (AI) and Machine Learning (ML) techniques is pivotal in generating these forecasts. Recent advancements in technology and reductions in costs have led to a proliferation of coastal station installations. Notably, a number of stations have been established in the Malta-Sicily channel as part of the i-waveNet project [3]. Real-time observations from these stations are accessible through the i-waveNET Decision Support System developed by the University of Malta [4].

While nowcasts (near-real-time observations) are vital, their utility would be greatly enhanced by incorporating short-term forecasts based on these observations. This paper explores our research into using ML to project coastal stations' observational data into the future. A significant challenge we face is the limited amount of data available for training, as most stations have only become operational in recent months. The literature suggests that deep learning models, especially Long Short-Term Memory Networks (LSTMs), are highly effective but require extensive datasets; even a three-year dataset of hourly observations is often insufficient to train an LSTM effectively.

We evaluate the performance of three ML architectures—Long Short-Term Memory Network (LSTM), Random Forest (RF), and Multivariate Linear Regression (LR)—to predict hourly values for air temperature, wind speed, atmospheric pressure, and humidity for the next 24 hours. To address the issue of limited available data, we conducted experiments with different training sets. We used a 32-month dataset from the Cirkewwa station (October 2020 - May 2023) and assessed how models trained on this dataset predict values for the Cirkewwa station and three other stations (two in Malta and one in southern Sicily) for November 2023. Additionally, we employed a one-month dataset from the Cirkewwa station (October 2023) to evaluate how models trained on it generate predictions for this station and the other three stations. Lastly, we trained a model for each station using data from October 2023 for that station and used these models to predict data for November 2023 for the same station.

The rationale behind experimenting with these different training datasets was to determine the extent to which long datasets are necessary for ML predictions across different parameters and to explore whether models can be generalised to apply across various stations. The outcomes help identify potential solutions for scenarios with sparse or missing data.

Furthermore, we explore the potential of using a long-term time series dataset from ERA5, provided by the Copernicus Climate Change Service (C3S), which offers meteorological hourly values from 1940 to 2022 at a coarse spatial resolution (0.5°) [6]. We performed experiments where we trained our models on this dataset and then fine-tuned them using station observational data.

Finally, we are making our model training code, prediction tools, and the best-performing pre-trained models for each parameter publicly available: https://ocean.mt/research/stationDataPredictions.zip

Related Research

Artificial Intelligence (AI) has demonstrated exceptional utility in short-term meteorological forecasting due to its ability to manage the complexities and vast datasets inherent in weather systems. Traditional numerical weather prediction models often struggle with the nonlinearities and high-dimensionality of weather data. In contrast, AI models, such as neural networks, excel in identifying complex patterns within large datasets without needing explicit physical modelling, leading to more accurate and timely forecasts.

AI has been widely applied to predict air temperatures, particularly local temperatures over short-term periods (typically 1 to 3 days in advance) [1, 8, 11]. Some studies have also focused on forecasting seasonal temperature variations [9]. Wind speed prediction, especially at a height of 10 metres, is another frequent application of AI [2, 10, 11, 13, 14]. Accurate wind speed forecasts are crucial due to the growing use of wind power generation and the necessity to predict energy output from wind sources. Other variables forecasted by AI models include humidity [11, 12, 18], atmospheric pressure [12], and rainfall/precipitation [16].

The AI architectures used to predict these meteorological parameters range from deep learning methods (like neural networks and Long Short-Term Memory networks, or LSTMs) to simpler machine learning techniques. LSTMs are particularly prevalent and have been successfully applied to predict wind speed [10] and air temperature [8, 11, 18]. Reports suggest that LSTMs often outperform other models. They are also frequently integrated into hybrid models, combined with other deep learning architectures such as Convolutional Neural Networks (CNNs) [11] and Convolutional Recurrent Neural Networks (CRNNs) [18], which have been reported to yield superior results.

While neural networks and deep learning techniques generally offer enhanced performance, they can struggle with limited training data. Some studies indicate that polynomial regression models outperform artificial neural networks when using a three-year dataset [2].

A popular alternative to deep learning models is Random Forests (RF). RF models are advantageous because they do not require the extensive training data that deep learning approaches do and have proven to be very effective in predicting meteorological variables. The reviewed literature shows that RF models are primarily used for predicting wind speeds [2, 13] and rainfall [15].

Support Vector Machines (SVMs) are another machine learning approach frequently employed in meteorological predictions. SVMs have been successfully used to forecast air temperature [1] and other general weather parameters [15]. Additionally, statistical methods like polynomial regressions have been applied to predict wind speed [2, 14].

Materials and Methods

The data used in this research comes from observations recorded at four coastal meteorological stations—three located in Malta and one in southern Sicily. These stations are depicted in Figure 1. In Malta, the stations are situated at Mgarr Gozo (blue), Cirkewwa (red), and Delimara (green), which are relatively close to each other. Mgarr Gozo is approximately 5 km from Cirkewwa, and Delimara is about 30 km from Cirkewwa. In contrast, the Sicilian station, Marina di Ragusa (yellow), is significantly farther away, approximately 90 km from Mgarr Gozo.

Each of these stations records the following observations at 1-minute intervals:

- Air Temperature
- Atmospheric Pressure
- Relative Humidity
- Wind Speed and Direction

For the period from October 1, 2023, to November 30, 2023, data from all four stations were aggregated into hourly intervals by simple averaging. Additionally, a longer dataset from the Cirkewwa station was available, containing hourly observations from October 5, 2020, to May 11, 2023. This dataset did not require further pre-processing.

Besides these datasets, we also utilised data from the ERA5 Reanalysis dataset for the period from January 1, 1940, to December 31, 2022. This dataset includes hourly observations for the geographical point at 14.50° longitude and 36.00° latitude, which is the closest to the Maltese Islands and most of the described stations. The ERA5 dataset, freely available from the Copernicus Climate Change Service, provides multiple meteorological variables from 1940 onwards at an hourly temporal resolution and a spatial resolution of 0.5°. For our research, we downloaded the following ERA5 parameters:

- 10 m u and v components of wind
- 2 m temperature
- 2 m dew point temperature
- Surface pressure

Pre-processing of the ERA5 data involved several steps: converting temperature from Kelvin to Celsius, calculating wind speed from the wind components, converting pressure from Pascals to millibars (mbar), and calculating relative humidity from the air temperature and dew point temperature. The relative humidity was computed using the formula provided below:

$$RH = 100 \times \frac{\exp\left(\frac{17.625 \times T_d}{243.04 + T_d}\right)}{\exp\left(\frac{17.625 \times T}{243.04 + T}\right)}$$

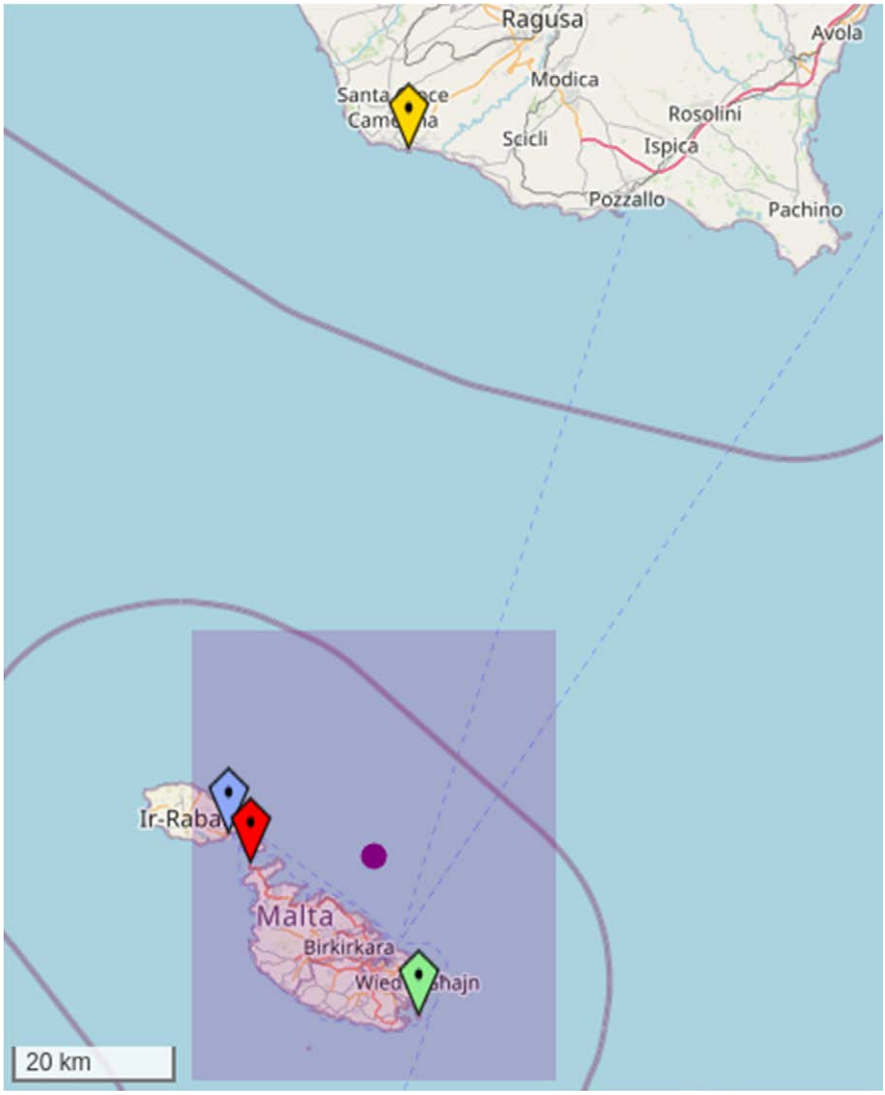

Figure 1 – Map displaying the locations of four stations in the WGS84 coordinate system: Marina di Ragusa (yellow marker, 14.5465°E, 36.7799°N), Mgarr Gozo (blue marker, 14.298°E, 36.024°N), Cirkewwa (red marker, 14.3296°E, 35.9906°N), and Delimara (green marker, 14.5589°E, 35.8217°N). The selected ERA5 reanalysis model cell is marked by a purple square with a central purple dot at 14.5°E, 36.0°N.

We employed three Machine Learning (ML) techniques in our study: Long Short-Term Memory Network (LSTM), Random Forest (RF), and Multivariate

Linear Regression (LR). LSTM and RF were chosen due to their proven success in the literature. LR was selected for its simplicity and its low data requirement for training. Our methodology involved using a lookback period of the past 48 hours of observations to predict the next 24 hours. For simplicity, our predictions always start at 0000 GMT each day. Specifically, each prediction (for the period from *0000* to *2300* of *Day 0*) was based on data from the lookback period *0000* of *Day -2* to *2300* of *Day -1*.

Separate models were trained for each target variable: air temperature, wind speed, atmospheric pressure, and relative humidity. The features used for the predictions included the observed values of air temperature, wind speed, atmospheric pressure, and relative humidity during the lookback period, as well as the hour of the day (0 – 23) and the current day of the year (1 – 365). The hour and day were transformed into sinusoidal signals by multiplying the hour or day ratio by π and then taking the sine of the resulting value. This transformation ensures that 0000 hours is as similar to 2300 as it is to 0100, preserving the cyclical nature of time. Additionally, we scaled all features using the Min-Max Scaler from the sklearn library to ensure they had equivalent ranges.

All models were implemented in Python 3.10 and trained on a Linux Ubuntu system with GPU capabilities. The LSTM model was developed using the Keras library, while the RF and LR models were implemented using the sklearn library. For the LSTM and RF models, we used the Optuna package for hyperparameter tuning. Optuna optimises the search for the best-performing parameters for each ML model.

As previously mentioned, our dataset included observations from the four stations covering the period from October 1, 2023, to November 30, 2023. We reserved the data for November 2023 (November 1, 2023, to November 30, 2023) for testing purposes. All training was conducted on data up to October 31, 2023, allowing us to evaluate the models on a full month of data. However, this approach also meant that training data was limited in some scenarios.

In the initial phase of our research, we focused on evaluating the effectiveness of AI models in scenarios with limited data by using only the observed data from the stations. For this experiment, we trained models for each target variable at each station using the following training sets:

- Data from October 1, 2023, to October 31, 2023, from the same station.
- Data from October 5, 2020, to May 11, 2023, from the Cirkewwa station.
- Data from October 1, 2023, to October 31, 2023, from the Cirkewwa station.

The goal was to evaluate the generalizability of models trained on data from different stations and to explore the feasibility of using longer datasets from other stations to improve model performance.

In the second phase of our research, we incorporated data from the ERA5 reanalysis to train our models. For each target variable at each station, we used the following training datasets:

- Data from January 1, 1940, to December 31, 2022, from ERA5.
- Data from January 1, 1940, to December 31, 2022, from ERA5, followed by fine-tuning with data from October 1, 2023, to October 31, 2023, from the same station.
- Data from January 1, 1940, to December 31, 2022, from ERA5, followed by fine-tuning with data from October 5, 2020, to May 11, 2023, from the Cirkewwa station.
- Data from January 1, 1940, to December 31, 2022, from ERA5, followed by fine-tuning with data from October 1, 2023, to October 31, 2023, from the Cirkewwa station.

This second set of experiments aimed to assess the viability of using a long-term reanalysis dataset to address the challenges associated with the scarcity of training data. By leveraging the extensive historical data from ERA5 and fine-tuning with more recent observations, we sought to improve the accuracy and robustness of the models.

Results

We trained and evaluated a total of 30 models for each target variable, resulting in a combined total of 120 models. Table 1 summarises the results obtained from these models, with Mean Absolute Error (MAE) used as the evaluation metric. To provide a more comprehensive overview, the MAE results for each model configuration were averaged across different stations.

Figure 2 illustrates the air temperature predictions for the Cirkewwa station made by the best-performing model from each of the different machine learning architectures. Figures 3, 4, and 5 follow a similar format: Figure 3 presents the wind speed predictions, Figure 4 the atmospheric pressure predictions, and Figure 5 the relative humidity predictions.

Discussion

Our results indicate that different meteorological parameters exhibit distinct characteristics and therefore require tailored modelling approaches. The simplest model, Multivariate Linear Regression (LR), performed the best for predicting air temperature and relative humidity. Notably, the LR model trained on the two-year Cirkewwa dataset produced the most accurate results for these parameters, even outperforming models trained on data from the station being evaluated. The Random Forest (RF) models trained on the ERA5 dataset and then fine-tuned using the two-year Cirkewwa dataset were the next best performers for both air temperature and relative humidity.

In contrast, for predicting wind speed and atmospheric pressure, the RF model trained exclusively on the ERA5 dataset (without fine-tuning) yielded the best results. This was particularly evident in the case of atmospheric pressure

predictions, where the difference in performance between the ERA5-trained model and models not trained on ERA5 was significant.

Overall, our findings underscore the critical importance of having sufficiently large datasets. Our analysis suggests that models trained on extensive datasets from nearby stations, or using global datasets, tend to perform better than those trained on shorter datasets from the same station.

Another key observation from our results is that deep learning architectures, such as Long Short-Term Memory networks (LSTMs), should not be presumed to provide superior results automatically. We believe the primary reason for this is that the training datasets used in these experiments were smaller than what is typically required for deep learning models to achieve their full potential.

Table 1 – Mean Absolute Error values for each training set and each target variable averaged across all stations.

Training Set	Model	Air Temp. (deg. C)	Wind Speed (m/s)	Atm. Pres. (mbar)	Rel. Hum. (%)
Same Station (Oct 2023)	LSTM	3.994	2.505	5.660	9.603
	RF	4.347	2.375	4.934	8.671
	LR	2.839	8.164	3.587	27.734
Cirkewwa (2 yr)	LSTM	1.710	2.277	6.383	7.529
	RF	1.272	2.540	4.099	6.737
	LR	**1.192**	2.570	7.012	**6.195**
Cirkewwa (Oct 2023)	LSTM	4.253	2.576	5.997	9.572
	RF	5.010	2.989	5.531	8.702
	LR	2.547	5.707	3.447	16.264
ERA5	LSTM	2.238	2.913	5.624	17.680
	RF	1.593	**2.110**	**2.161**	15.614
ERA5 + Same Station (Oct 2023)	LSTM	2.890	2.892	4.946	9.024
	RF	4.339	2.469	4.903	8.450
ERA5 + Cirkewwa (2 year)	LSTM	2.114	3.047	5.042	7.577
	RF	1.242	2.616	4.237	6.691
ERA5 + Cirkewwa (Oct 2023)	LSTM	3.221	3.000	5.080	8.434
	RF	5.053	2.956	5.540	8.609

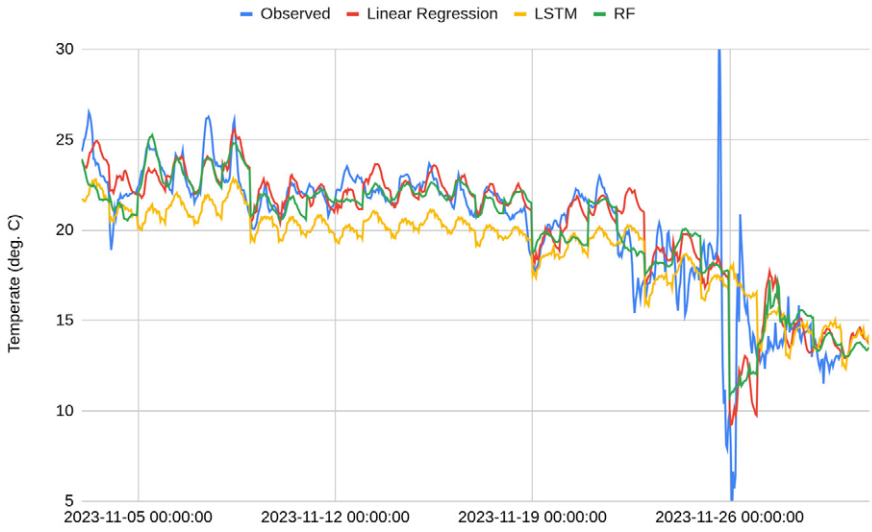

Figure 2 – Air Temperature predictions for the Cirkewwa station.

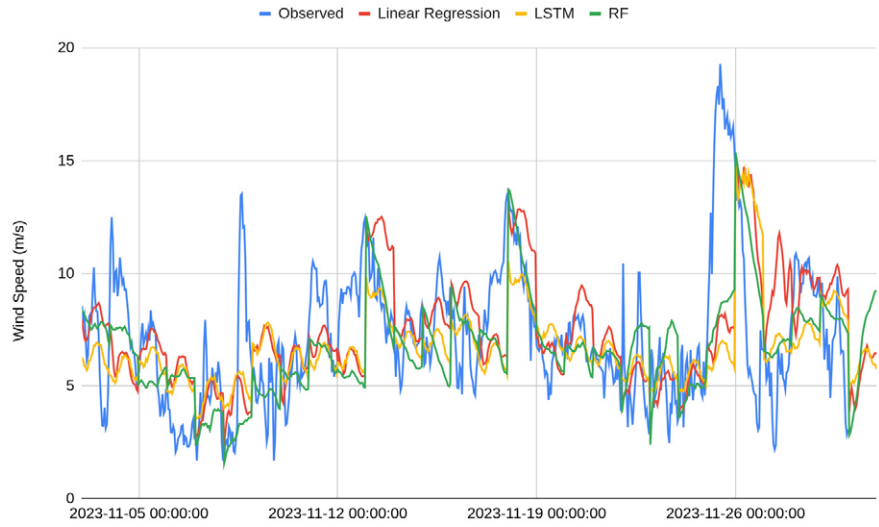

Figure 3 – Wind speed predictions for the Cirkewwa station.

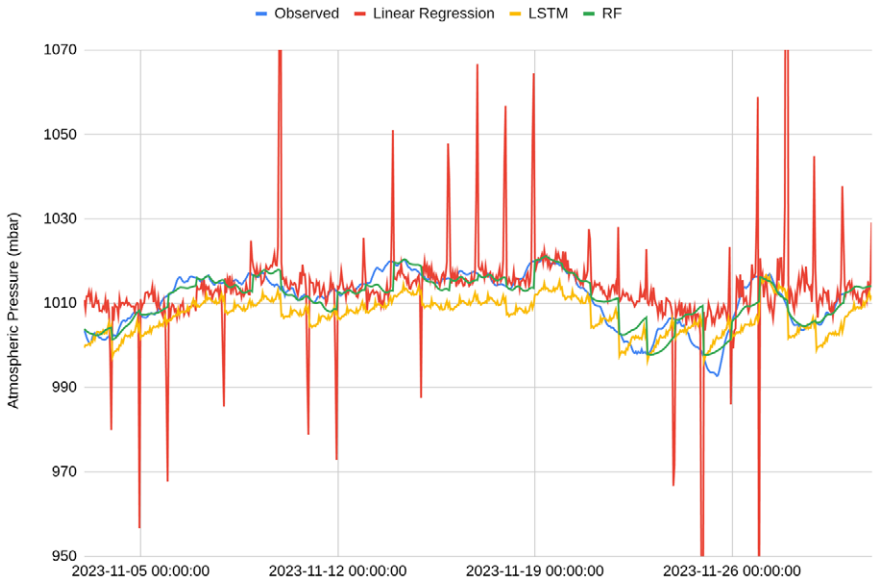

Figure 4 – Atmospheric pressure predictions for the Cirkewwa station.

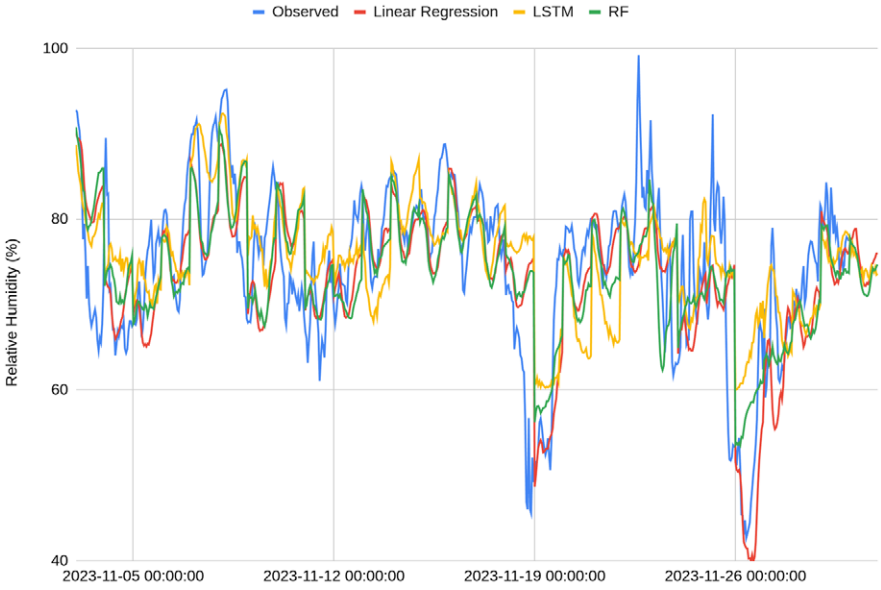

Figure 5 – Relative humidity predictions for the Cirkewwa station.

Conclusion

In this research, we explored the use of LSTMs, RF and LR to predict meteorological parameters based on observations from coastal meteorological stations. Our findings highlighted the critical importance of having sufficiently large training datasets. They suggest that models trained on extensive datasets from nearby stations or global models are preferable to those trained on shorter datasets from the same station.

For future work, we plan to investigate the effectiveness of pre-trained probabilistic forecasting models, such as Lag Llama [18]. Additionally, we intend to experiment with hybrid models that integrate AI with computational physical models, such as the WRF model, to potentially enhance prediction accuracy.

References

[1] Adnan, R.M., Liang, Z., Kuriqi, A., Kisi, O., Malik, A., Li, B. and Mortazavizadeh, F. (2021) - *Air temperature prediction using different machine learning models*, Indonesian Journal of Electrical Engineering and Computer Science 22 (1), 534-541. DOI: 10.11591/ijeecs.v22.i1.pp534-541.

[2] Antor, A.F. and Wollega, E.D. (2020) - *Comparison of machine learning algorithms for wind speed prediction*, Proceedings of the 5th NA International Conference on Industrial Engineering and Operations Management, Detroit, Michigan, USA, August 10 - 14, 2020, pp. 1-8. © IEOM Society International.

[3] Aronica, S., et al. (2022) - *The i-waveNet project and the integrated sea wave measurements in the Mediterranean sea*, 2022 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea), Milazzo, Italy, 2022, pp. 484-487. DOI: 10.1109/MetroSea55331.2022.9950876.

[4] Azzopardi, J., Zammit, A. and Gauci, A. (2024) - *The i-waveNET Decision Support System – user-driven aggregations and analysis of forecasts and observations*, Proceedings of the International Conference on Marine Data and Information Systems: IMDIS 2024, pp. 137-138. DOI: 10.13127/MISC/80.

[5] Bochenek, B. and Ustrnul, Z. (2022) - *Machine learning in weather prediction and climate analyses—Applications and perspectives,* Atmosphere 13 (2), 180. DOI: 10.3390/atmos13020180.

[6] Copernicus Climate Change Service (C3S) (2017) - *ERA5 Reanalysis (Hourly Data on Single Levels) from 1979 to present, Climate Data Store (CDS)*, European Centre for Medium-Range Weather Forecasts (ECMWF). Retrieved from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels.

[7] Frnda, J., Ďurica, M., Nedoma, J., Žabka, S., Martínek, R. and Kostelanský, M. (2019) - *A weather forecast model accuracy analysis and ECMWF enhancement proposal by neural network*, Sensors 19 (23), 5144. DOI: 10.3390/s19235144.

[8] Kreuzer, D., Münz, M. and Schlüter, S. (2020) - S*hort-term temperature forecasts using a convolutional neural network—An application to different weather stations in Germany*, Machine Learning with Applications 1, 100007. DOI: 10.1016/j.mlwa.2020.100007.

[9] Nezhad, E.F., Ghalhari, G.F. and Bayatani, F. (2019) - *Forecasting maximum seasonal temperature using artificial neural networks: Tehran case study*, Asia-

Pacific Journal of Atmospheric Sciences 55 (1), 97-109. DOI: 10.1007/s13143-018-0051-x.

[10] Pati, N., Gourisaria, M.K., Das, H. and Banik, D. (2023) - *Wind speed prediction using machine learning techniques,* Proceedings of the 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, 2023, pp. 1-6. DOI: 10.1109/ICETET-SIP58143.2023.10151597.

[11] Roy, D.S. (2020) - *Forecasting the air temperature at a weather station using deep neural networks,* Procedia Computer Science 170, 392-399. DOI: 10.1016/j.procs.2020.11.005.

[12] Salman, A.G., Kanigoro, B. and Heryadi, Y. (2015) - *Weather forecasting using deep learning techniques*, Proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, pp. 281-285. DOI: 10.1109/icacsis.2015.7415154.

[13] Samuel, G.G., Sankar, P., Samuel, A., Edwin, P. and Manikandan, J. (2021) - *Improved prediction of wind speed using machine learning*, Journal of Physics: Conference Series 1964, 052005. DOI: 10.1088/1742-6596/1964/5/052005.

[14] Şener, U., Kılıç, B.İ., Tokgözlü, A., Aslan, Z. (2023) - *Prediction of Wind Speed by Using Machine Learning.* In: Gervasi, O., et al. - *Computational Science and Its Applications* – ICCSA 2023 Workshops. ICCSA 2023. Lecture Notes in Computer Science, vol 14104. Springer, Cham. DOI 10.1007/978-3-031-37105-9_6

[15] Singh, Nitin, Saurabh Chaturvedi and Shamim Akhter. (2019) - *Weather Forecasting Using Machine Learning Algorithm*. International Conference on Signal Processing and Communication (ICSC) (2019): 171-174.

[16] Singh, S., Kaushik, M., Gupta, A. and Malviya, A.K. (2019) - *Weather forecasting using machine learning techniques*, Proceedings of the 2nd International Conference on Advanced Computing and Software Engineering (ICACSE), 2019. DOI: 10.2139/ssrn.3350281.

[17] Schiller, J., Prokhorenkova, L., Seleznev, S., and Bischofberger, J. (2023) - *Weather forecasting using deep learning methods: A comprehensive review*, arXiv preprint arXiv:2310.08278. DOI: 10.48550/arXiv.2310.08278.

[18] Zhang, Z. and Dong, Y. (2020) - *Temperature forecasting via convolutional recurrent neural networks based on time-series data*, Complexity 2020, 3536572. DOI: 10.1155/2020/3536572.