Filling Gaps in Daily Temperature Data with a CNN-LSTM Model

Thea Camille-Maxime
Department of Computer Science and Applied Mathematics
National University of Water and Environmental Engineering
Rivne, Ukraine
ORCID: 0009-0002-6709-7364

Olena Belozerova
Department of Computer Science and Applied Mathematics
National University of Water and Environmental Engineering
Rivne, Ukraine
ORCID: 0000-0001-9934-1013

***Abstract*—The challenge of missing data poses a significant difficulty in the practical use of recorded meteorological data, a concern that has become apparent in Ukraine in recent years. This paper presents a deep learning methodology for imputing consecutive missing data within weather station records. By integrating convolutional neural network and Long Short-Term Memory (LSTM) layers, both recognized as prominent techniques in weather data forecasting and imputation, this approach provides reliable results for filling gaps in daily air temperature data.**

Keywords—meteorology data, time series imputation, missing weather data, convolutional neural network, LSTM.

V. Introduction

Meteorological measurement stations are essential tools for monitoring weather conditions, yet they are susceptible to technical failures, resulting in gaps in measured data. In recent years, the lack of measured meteorological data has become a noticeable problem in Ukraine. According to the U.S. National Climatic Data Center (NCDC) [1], 32 out of 123 meteorological stations in Ukraine have been closed during the pandemic, and 57 more stations have stopped operating after the Russian invasion in 2022, leaving only 34 working stations by the end of 2023. Moreover, in the years 2022-2023, the remaining stations failed to record data in 16.8% of days on average. In contrast, over the five years preceding 2022, this value did not exceed 5.2% over all Ukrainian stations.

Thus, the application of weather station data often calls for gap-filling procedures. Common approaches to address missing data involve analyzing previously recorded data and employing statistical methods. These methods typically draw upon temporal data from the station under consideration [2, 3], spatial data from neighboring stations, or a combination of both [4]. Imputation models range in complexity from interpolation techniques and linear regression to more sophisticated models such as AutoRegressive Integrated Moving Average (ARIMA) [3] models and deep learning.

Among the latter, recurrent neural networks (RNN), Long Short-Term Memory (LSTM) models [5] and their variations [2, 6] have demonstrated effectiveness in handling missing data. Additionally, techniques like Kalman smoothers have been proposed to address discrepancies between imputed and recorded data following data gaps [2].

Moreover, insights from weather forecasting research address similar challenges and can offer valuable contributions. For instance, some studies propose an addition of convolutional layers preceding recurrent layers in neural networks [7]. Convolutions aid in summarizing input data, thereby enhancing the model's ability to discern patterns within the data [8].

This study introduces the application of a convolutional LSTM neural network to the gap-filling problem within meteorological data analysis. Currently we narrowed our focus to imputing missing values within daily air temperature time series data due to its predictable dynamics.

VI. Methods

A. Convolutional LSTM model structure

LSTM (Long Short-Term Memory) models make use of LSTM cells, an advancement over traditional RNN cells. These specialized cells retain memory of past input and output values, making them particularly well-suited for analyzing time series data. Notably, LSTM models excel in identifying long-term patterns within input sequences. Some studies suggest that bidirectional LSTM cells (BiLSTM) provide superior predictive results for meteorological data [2].

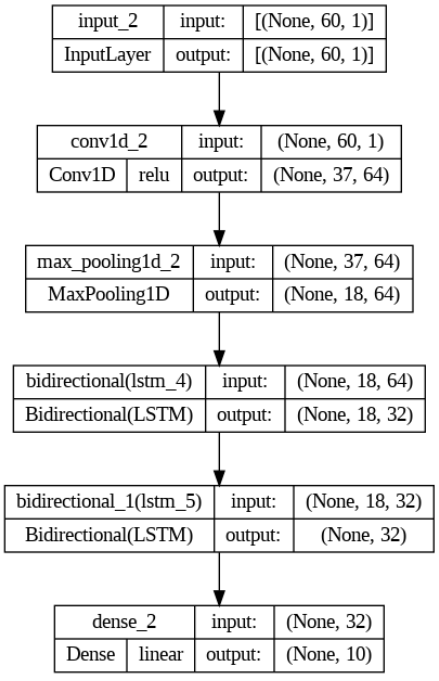

Рис. 1.Structure of the CNN-BiLSTM neural network for predicting missing values in 10-day intervals.

Before the main LSTM layer, input data are processed through a convolutional neural network (CNN) layer. Initially

developed for image processing, CNNs have found utility in various domains requiring feature extraction and data preprocessing. In the context of time series data, convolutions facilitate pattern recognition, alleviating the computational burden on LSTM layers.

The architecture of the proposed neural network is illustrated in Fig. 1. Input data comprise the 60 consecutive temperature measurements preceding the data gap. These data are initially processed through a 1D convolutional layer with 3-size filters, followed by a max-pooling layer. Subsequently, two BiLSTM layers with rectified linear unit (ReLU) activation functions perform core model computations. The final dense layer generates a sequence of 10 values representing the imputed data for a 10-day missing period. This duration was chosen based on the observation that data gaps typically did not exceed this timeframe in the studied weather stations.

Additionally, we propose incorporating measurements following the gap as initial data, aiming to enhance imputation accuracy and achieve smoother transitions between pre- and post-gap data. The modified model maintains the same structure but incorporates an additional 10 values following the gap, resulting in a total of 70 input values.

B. Experimental Setting and Data

We test the proposed model on three meteorological stations in various parts of Ukraine that provide data with the least percent of missing values: Rivne, Vinnytsia, and Sumy.

We use daily weather records provided by the NCDC Climate Data Online service. Data records from 2000 to 2021 are selected from the target stations, with records preceding 2020 allocated for model training. A separate model is trained for each station.

Data for the years 2020 and 2021 are used for testing the model. This data span is segmented into 10-day intervals, and the model is tasked with predicting values within these artificially created gaps using truthful input data. The predicted data are combined into a single time series and evaluated against the actual data. Imputation accuracy is assessed through the following metrics: mean absolute error (MAE), root mean squared error (RMSE), correlation (R), and coefficient of determination (R^2).

Two distinct numerical experiments are conducted within this study. The first model (Model A) is a convolutional BiLSTM neural network that is given the 60 days of data preceding the gap. The second model (Model B) is trained on both 60 preceding values and 10 values following the gap.

VII. Results and Discussion

Table 1 presents the results of accuracy assessment conducted on the base Model A. The network is trained in batches of 128 records with early stopping after 100 epochs of non-improvement on validation data. On average, convergence required approximately 300 epochs across the test stations.

ТАБЛИЦЯ II. Evaluation results for model A

Station	Metric			
	MSE	***RMSE***	***R***	***R^2***
Rivne	3.04	3.84	0.957	0.817
Vinnytsia	3.00	3.88	0.906	0.817
Sumy	3.44	4.34	0.899	0.807

Across the three target stations, Model A demonstrates an MAE of 3.16°C and an RMSE of 4.09°C. The correlation and coefficient of determination (R^2) are 0.92 and 0.81, respectively. Note that R and R^2 scores are extremely forgiving in this application due to seasonality of the data.

In contrast, Model B, which trains on both preceding and subsequent values surrounding the data gap, yields improved average scores of 2.86°C for MAE and 3.72°C for RMSE, as detailed in Table 2. Thus, the proposed CNN-BiLSTM model exhibits reasonable accuracy in temporal imputation of missing air temperature values. The inclusion of actual measurement data after the gap enhances imputation accuracy by 9.5% in MAE and 7.4% in RMSE.

ТАБЛИЦЯ III. Evaluation results for model B

Station	Metric			
	MAE	***RMSE***	***R***	***R^2***
Rivne	2.84	3.62	0.916	0.963
Vinnytsia	2.64	3.45	0.925	0.856
Sumy	3.10	4.10	0.910	0.828

This study represents a fundamental application of the proposed approach. Future research could explore the integration of multiple weather parameters and additional measurement stations to develop a more robust and reliable solution for addressing missing data challenges in meteorological datasets. Further development of state-of-the-art deep learning models will be able to address the challenges within meteorological data analysis.

References

[1] National Centers for Environmental Information, National Oceanic and Atmospheric Administration. “Climate Data Online (CDO) provides free access to NCDC's archive of global historical weather and climate data in addition to station history information”. NOAA.gov. https://www.ncei.noaa.gov/cdo-web/ (retrieved May 4, 2024).

[2] C. Xie, C. Huang, D. Zhang, and W. He, “BiLSTM-I: A Deep Learning-Based Long Interval Gap-Filling Method for Meteorological Observation Data,” Int. J. Environ. Res. Public Health, vol 18, no.19, 2021, Art. no. 10321.

[3] E. Afrifa-Yamoah, U. A. Mueller, S. M. Taylor, and A. J. Fisher, “Missing data imputation of high-resolution temporal climate time series data,” Meteorol Appl., vol. 27, 2020, Art. no. e1873.

[4] B. Henn, M. S. Raleigh, A. Fisher, and J. D. Lundquist, “A Comparison of Methods for Filling Gaps in Hourly Near-Surface Air Temperature Data,” J. Hydrometeor., vol. 14, pp. 929–945, 2013.

[5] S. Liang, L. Nguyen, and F. Jin, "A Multi-variable Stacked Long-Short Term Memory Network for Wind Speed Forecasting," presented at the IEEE International Conference on Big Data, Seattle, WA, USA, Dec. 10-13, 2018, pp. 4561-4564.

[6] Y. Wang, K. Liu, Y. He, Q. Fu, W. Luo, W. Li, “Research on Missing Value Imputation to Improve the Validity of Air Quality Data Evaluation on the Qinghai-Tibetan Plateau,” Atmosphere, vol. 14, no. 12, 2023, Art. no. 1821.

[7] D. Kreuzer, M. Munz, and S. Schlüter, “Short-term temperature forecasts using a convolutional neural network — An application to different weather stations in Germany,” Machine Learning with Applications, vol. 2, 2020. Art. no. 100007.

[8] A. P. Wibawa, A. B. P. Utama, H. Elmunsyah, Utomo Pujianto, Felix Andika Dwiyanto, and Leonel Hernandez, “Time-series analysis with smoothed Convolutional Neural Network,” J. Big Data, vol. 9, 2022, Art. no. 44.