Columbia University

Climate Change Prediction with Time-Series Modeling

Adeel Arif, Boping Song, Malaikah Khan

APANPS5910

Professor Siddhartha Dalal

December 5, 2025

Abstract

Introduction

Accurate temperature forecasting is essential for understanding climate change impacts and supporting policy and adaptation planning. Deep learning models such as Long Short-Term Memory (LSTM) networks can capture long-term temporal patterns in climate data, yet global and regional climate systems differ in variability, seasonality, and data quality. This study evaluates how LSTM and Bidirectional LSTM (Bi-LSTM) architectures perform in forecasting monthly land temperatures at global and continent-specific scales.

Methods

We used the Berkeley Earth Surface Temperature Dataset, which includes more than 500,000 monthly observations from 1743 to 2013. A standardized preprocessing pipeline was implemented involving geographic harmonization, deduplication, interpolation, and continent assignment. Three neural architectures were trained: a baseline attention LSTM, an enhanced attention LSTM incorporating seasonal and lagged features, and a multi-feature attention Bi-LSTM. Models were trained both globally and for six continents using a 36-month input sequence to predict next-month temperature. Performance was evaluated using RMSE, MAE, and Directional Accuracy.

Results

The Enhanced Attention LSTM achieved the best global performance with an RMSE of 0.0204, an MAE of 0.0133, and a directional accuracy of 0.875. At the continent level, the Bi-LSTM consistently achieved the lowest RMSE and MAE across all six regions, indicating a stronger ability to model localized and cyclical climate dynamics. Directional Accuracy remained similar across architectures, with smaller differences compared to error metrics.

Conclusion

Global forecasts benefit most from enriched feature engineering, while regional forecasting improves with bidirectional temporal encoding. These results provide a reproducible framework for climate time-series modeling and highlight the potential of combining feature-rich LSTMs and Bi-LSTMs for multi-scale climate forecasting.

Introduction

Climate change is one of the most significant global challenges today, with rising temperatures creating escalating risks for ecosystems, infrastructure, and human health. According to NASA's Global Climate Change program, Earth's surface temperature is warming at an unprecedented rate, and human activity is the principal cause (NASA, 2024). These changes are increasing the frequency and intensity of heat extremes, altering precipitation patterns, and threatening global food and water security. The effects of climate change are also not distributed equally. The Intergovernmental Panel on Climate Change (IPCC) reports that communities in Africa face heightened vulnerability due to a combination of high exposure and development constraints that intensify risks to agriculture, strain water resources, and increase climate-related mortality (IPCC, 2023). This uneven burden highlights the need for reliable temperature forecasts that support proactive planning in agriculture, water management, and heat-related health preparedness. Accurate regional forecasts can help policymakers, NGOs, and governments make informed decisions to improve climate resilience.

Machine learning models offer powerful tools for analyzing long-term climate patterns. Deep learning architectures such as Long Short-Term Memory (LSTM) networks are well-suited for capturing complex temporal dynamics, including seasonality, periodic variability, and long-term warming trends (Kong, 2024). However, global temperature forecasting remains difficult. Climate time series are non-stationary, so models must learn both short-term fluctuations and gradual multi-decadal change. Regional climate systems also differ substantially, which means that a single global LSTM model may not perform equally well across all continents, particularly in regions with high variability or limited data, such as Africa. To address these challenges, our project trains both global and continent-specific LSTM models. This approach allows us to account for regional climate structures and evaluate where localized modeling improves performance. We also incorporate Bidirectional LSTMs, which learn temporal patterns by using information from both past and future context (Duy, 2025). By training both global and regional LSTM and Bi-LSTM models, we can directly compare whether one unified model is sufficient or if continent-specific models yield more accurate forecasts.

The goal of this project is to investigate how effectively continent and global LSTM-based models and Bi-LSTM models can forecast future monthly land temperature using historical temperature data. The study examines two key research questions:

1) To what extent can an LSTM-based model reliably forecast future temperatures at a global scale?

2) How does a global LSTM model's performance compare to continent-specific LSTM models in forecasting temperature across different regions (e.g., Europe, Asia, Africa)?

Methods

Dataset

We used the Berkeley Earth Surface Temperature Dataset from Kaggle, which provides monthly average land surface temperatures for all available countries and territories from 1743 to 2013 (Berkeley Earth via Kaggle). Each record includes the timestamp, average temperature, associated measurement uncertainty, and country identifier. After loading the full dataset, the initial corpus contained 577,462 observations across 243 countries and territories.

Preprocessing Pipeline

A standardized ETL pipeline was implemented to prepare the time series for neural forecasting. The procedure consisted of four stages: geographic harmonization, temporal cleaning, gap filtering, and interpolation. A flowchart of the pre-processing pipeline can be seen in Figure 1, presented in the Appendix.

Geographic Normalization

Through the initial review of the dataset, we confirmed that the majority of standard countries were present. However, the dataset also included a considerable number of non-sovereign regions, island territories, and geographic designations that do not map directly to conventional country lists. To enable continent-level aggregation, we required a unified mapping that assigned each entity to a single continental region.

Country labels were mapped to continental regions using the pycountry-convert package, which provides ISO-standard country-to-continent conversions (Koch, n.d.). This automated mapping covered most countries in the dataset. For territories and regions not recognized by the library, including island groups, manual assignment was performed based on geographic proximity and established regional classifications. This step was essential for constructing consistent continent-level datasets and for supporting the training of continent-specific forecasting models.

Deduplication

A small number of countries contained multiple entries for the same year-month timestamp. Duplicate values were aggregated by taking the mean of the reported temperatures. This ensured a single, consistent observation per month.

Gap Removal

Time-series forecasting models rely on input sequences with uniform temporal spacing, as irregular intervals can distort learned temporal dependencies and disrupt predictive performance (Chen, 2023). To enforce this assumption, all country-level series were systematically screened for irregular sampling following established preprocessing guidelines for sequence modeling. Specifically, we computed the time gap between every pair of consecutive observations and flagged sequences containing any interval exceeding three months. Such sequences were removed from the modeling dataset.

Interpolation

Missing temperature values were filled using interpolation. This method estimates values based on adjacent observations and preserves continuity in the underlying signal (Lepot, 2017). Interpolation has been widely used in time-series processing and helps prevent model instability caused by missing temporal inputs. In addition, all temperature values were scaled using min–max normalization before model training.

In addition to handling missing values, we addressed the geographic structure of the dataset before summarizing the final sample counts. The dataset contains numerous Pacific island regions listed as distinct entries, and treating Australia alone as its own continent would have produced an imbalanced distribution where one country dominated an entire region. To maintain consistency with established geographic conventions in climate research, we grouped Australia, New Zealand, and the Pacific island nations under the broader regional category of Oceania (Low, 2023). This ensured that continent-level datasets were more evenly structured and prevented small island territories from forming undersized groups that would reduce the stability and interpretability of continent-specific forecasting models.

After all preprocessing procedures were completed, the final cleaned dataset contained 533,022 monthly observations. The distribution of records across continents was as follows: Europe (157,688), Asia (114,516), Africa (105,230), North America (90,135), South America (29,509), Oceania (32,892), and Antarctica (3,052). Because Antarctica contributed relatively few data points, it was excluded from subsequent analyses. Removing these 3,052 Antarctic observations yielded a final modeling dataset of 529,970 monthly records.

Architecture and Modeling Overview

To evaluate how different neural architectures capture climate dynamics across temporal and geographic scales, we implemented three complementary forecasting models: a baseline attention LSTM, an enhanced multi-feature attention LSTM, and a bidirectional LSTM with attention. Standard LSTMs are well-suited for monthly climate data because they are designed to learn long-term temporal dependencies and can effectively model slow-moving temperature trends (Kong, 2024). The attention mechanism was included to allow the network to learn which historical months within the 36-month input window contribute most to the upcoming temperature, improving interpretability and reducing the risk of the model overemphasizing irrelevant portions of the sequence (Kang, 2023).

The enhanced LSTM extends this baseline approach by incorporating additional seasonal and lagged features that encode multi-year cycles, short-term smoothing, and periodic structure, providing a more robust representation of climate variability. Finally, the bidirectional LSTM was included because climate patterns are cyclical and often depend on relationships that appear both earlier and later within the input window. Unlike a standard LSTM, which processes the sequence in only one direction, a Bi-LSTM learns two sets of hidden states by reading the sequence forward and backward (Duy, 2025). Combining these representations gives the model a more complete view of seasonal structure and repeating transitions, allowing it to capture patterns that a one-directional LSTM may miss.

Each architecture was trained in two configurations: a single global model using all countries combined, and six continent-specific models trained independently on Europe, Asia, Africa, North America, South America, and Oceania. With three architectures and seven models per architecture, the study produced a total of 21 forecasting models. All models were trained to predict the average land surface temperature for the next month (t + 1) using a fixed 36-month historical window.

All preprocessing scripts, architectures, and training code used in this study are available in the project's GitHub repository: https://github.com/adeelarif9/Climate-Change-LSTM-Project.

Baseline Attention LSTM

The baseline model used a single input feature, the monthly average temperature. A single-layer LSTM processed the 36-month window and produced a sequence of hidden representations, one for each month in the input. These hidden states were then passed to an attention layer that learned a set of weights over the 36 time steps and combined them into a single context vector. This allowed the model to identify which past months were most informative for the next-month prediction and to downweight less relevant parts of the sequence. The baseline attention LSTM, therefore, provided a simple but interpretable reference point for evaluating the benefit of additional features and architectural complexity.

Hyperparameters:

The baseline model used one LSTM layer with a hidden size of 64. Training was performed for 30 epochs with a batch size of 64, with a learning rate of 0.001 and mean squared error as the loss function.

Enhanced Attention LSTM (Multi-Feature Model)

The enhanced attention LSTM expanded the input from one feature to six engineered features. In addition to the raw monthly temperature, the model received sine and cosine transformations of the month index, lagged temperatures, and a short-term rolling average. The sine and cosine terms encode the cyclical structure of the annual calendar, which removes the artificial break between December and January and allows the model to learn smooth seasonal variation (Valença, 2024). The lagged features provide explicit information about the temperature one year and two years earlier, which helps the network capture repeating annual and multi-year climate cycles that may not be fully learned from the raw sequence alone (Lazzeri, 2021). The three-month rolling average smooths high-frequency variability and highlights local trends, which can make it easier for the model to identify persistent warming or cooling patterns rather than reacting to noise (Lazzeri, 2021).

Introducing this richer feature set increased the expressive power of the model but also made optimization more challenging. With higher input dimensionality, the gradients flowing through the LSTM can occasionally become very large, which is known as exploding gradients. When this occurs, parameter updates can become extremely large between iterations, causing the loss to fluctuate wildly or diverge instead of converging to a stable minimum (Bajaj, 2025). To address this, we applied gradient clipping during training of the enhanced model. Gradient clipping limits the magnitude of the gradients before the optimizer updates the weights, which keeps training stable while still allowing the network to learn from long sequences and complex feature interactions (Bajaj, 2025).

Bidirectional LSTM with Attention

The bidirectional LSTM with attention was built on the enhanced multi-feature model but differed only in the directionality of the recurrent layer. Instead of using a standard LSTM that processes the input sequence from earlier months to later months, the bidirectional LSTM processes the same 36-month window twice: once in the forward direction and once in reverse. This produces two independent hidden representations for each position in the sequence. These forward and backward representations are combined and then passed to the attention layer, which learns to focus on the most informative time steps across the combined sequence. Using the same six engineered features as the enhanced LSTM, the bidirectional architecture provides a more complete temporal encoding of the input window.

Hyperparameters:

The bidirectional model used the same training settings as the enhanced attention LSTM. The only architectural change was the use of a bidirectional LSTM with 64 hidden units in each direction. Gradient clipping was also applied to ensure stable training due to the increased representational capacity of the bidirectional layer.

Results

Global Metrics

Global performance was evaluated using RMSE, MAE, and Directional Accuracy (DA). DA measures how often the model correctly predicts whether the temperature will rise or fall relative to the previous month. Across all three metrics, the Enhanced Attention LSTM achieved the strongest performance. It obtained an RMSE of 0.0204, compared with 0.0258 for the Baseline Attention LSTM and 0.0211 for the Attention Bi-LSTM. A similar pattern was observed for MAE, where the Enhanced model achieved 0.0133, outperforming the Baseline (0.0155) and Bi-LSTM (0.0155). Directional Accuracy also favored the Enhanced model, which reached 0.875, compared with 0.842 for the Baseline and 0.870 for the Bi-LSTM. Overall, the Enhanced Attention LSTM achieved lower error and slightly higher directional consistency than the other architectures at the global scale. All global metrics can be seen in Table 1, as presented in the appendix.

Continent-Level Metrics

Continent-level evaluation revealed a different performance trend from the global results. For every continent, the Attention Bi-LSTM achieved the lowest RMSE, indicating the strongest numerical performance for regional temperature forecasting. In Europe, the Bi-LSTM achieved an RMSE of 0.04, compared with 3.02 for the Enhanced model and 4.39 for the Baseline. Asia showed a similar structure, with the Bi-LSTM achieving 0.03 versus 1.65 for the Enhanced model and 2.04 for the Baseline. In Africa, the Bi-LSTM reached 0.03, while the Enhanced and Baseline models recorded 1.07 and 1.20, respectively. North America followed the same pattern, with the Bi-LSTM at 0.03, the Enhanced model at 3.08, and the Baseline at 1.41. Results in Oceania and South America were numerically higher overall, but the Bi-LSTM still obtained the lowest RMSE in both regions, recording 0.05 in Oceania and 0.09 in South America.

MAE results mirrored the RMSE findings. Across all six continents, the Bi-LSTM achieved the lowest MAE, with values such as 0.03 in Europe, 0.02 in Asia, 0.02 in Africa, and 0.02 in North America. In Oceania, the Bi-LSTM recorded 0.04, while South America reached 0.07. The Enhanced model consistently recorded intermediate MAE values, while the Baseline model yielded the highest MAE across regions.

Directional Accuracy values remained relatively similar across all architectures. In Europe, DA ranged from 0.87 to 0.88, while in Asia the range was 0.89 to 0.90. Africa displayed values between 0.83 and 0.85, and North America ranged from 0.84 to 0.85. Oceania, which showed weaker performance across RMSE and MAE, exhibited DA values between 0.65 and 0.70. South America displayed values from 0.69 to 0.79. Although the Bi-LSTM often achieved the strongest DA in several regions, differences across architectures were relatively small compared with the error-based metrics.

Overall, the numerical results show that the Enhanced Attention LSTM achieved the best performance at the global level, whereas the Attention Bi-LSTM obtained the lowest RMSE and MAE across all six continents, with relatively modest differences in Directional Accuracy among the three models. All continent-specific results can be seen in Tables 2, 3, and 4 as presented in the appendix.

Discussion

The results show clear differences in how each architecture responds to the structure of global versus regional climate data. At the global scale, the Enhanced Attention LSTM achieved the strongest performance, which indicates that expanded feature engineering plays a central role in forecasting large-scale climate signals. The global temperature series is smoother and more homogeneous than continent-level sequences, and the inclusion of seasonal encodings, lagged features, and a rolling mean provided the model with explicit representations of periodic and long-term dependencies. These engineered features allowed the model to capture repeating annual cycles and gradual multiyear trends more effectively than the baseline or bidirectional architectures.

In contrast, the continent-level experiments revealed consistent advantages for the Attention Bi-LSTM. Climate sequences at the regional scale are far less uniform than the global aggregate and often contain abrupt transitions, localized anomalies, and seasonal asymmetries between the Northern and Southern Hemispheres. The bidirectional architecture is well-suited to this setting because it analyzes the input window in both forward and backward directions, enabling it to detect relationships that depend on how different parts of the seasonal cycle relate to one another within the 36-month context. The superior RMSE and MAE values achieved by the Bi-LSTM across all continents suggest that regional climate dynamics benefit from architectures that capture temporal dependencies in a more flexible and symmetric way.

Directional Accuracy values were more similar across architectures and across continents. This indicates that, although the models varied substantially in their ability to reduce numerical error, they were comparably effective at predicting the direction of month-to-month temperature changes. Direction prediction is a simpler task than estimating absolute values, and even the baseline LSTM was able to capture these directional shifts with reasonable consistency. The smaller differences in DA values, therefore, reflect the relative stability of directional patterns in monthly climate data.

Oceania showed weaker performance across all architectures, which can be attributed to several structural properties of the dataset. Oceania contains fewer total observations than other continents, and many of its regions represent small island climates with highly variable temperature behavior. Additionally, seasonal patterns in Oceania are reversed compared with regions in the Northern Hemisphere, which complicates learning when the dataset contains a mixture of both hemispheric structures. These factors reduce the ability of the models to learn stable temporal relationships, resulting in elevated error values. Antarctic data were excluded entirely because the limited number of observations and irregular sampling patterns did not permit reliable modeling.

Overall, the findings highlight two key conclusions. First, incorporating explicit seasonal and lagged features is essential for improving global climate forecasts, where broad temporal patterns dominate. Second, bidirectional architectures offer clear advantages for regional modeling, where localized climate regimes exhibit more complex temporal behavior. These results suggest that future climate forecasting frameworks may benefit from hybrid global-regional modeling strategies or architectures that adaptively select between unidirectional and bidirectional processing depending on the geographic scale of the input.

Ethical Considerations

To minimize potential bias, differences in data quality and sample size across continents were examined, and all preprocessing steps were standardized to ensure fairness in model comparison. Transparency was prioritized by documenting all modeling procedures, including sequence window selection, normalization strategies, seasonal sine–cosine encoding, and model-specific feature construction. The dataset used in this study is publicly available and contains no individual-level or sensitive information, reducing concerns regarding data privacy. From a societal perspective, machine learning models can support climate science and inform policy discussions, but their outputs should not be interpreted without appropriate domain expertise. Overreliance on model predictions could lead to misinterpretations of climate risks if they are not contextualized within established scientific evidence and policy frameworks. Readers are encouraged to evaluate these results in conjunction with broader climate research rather than as standalone indicators.

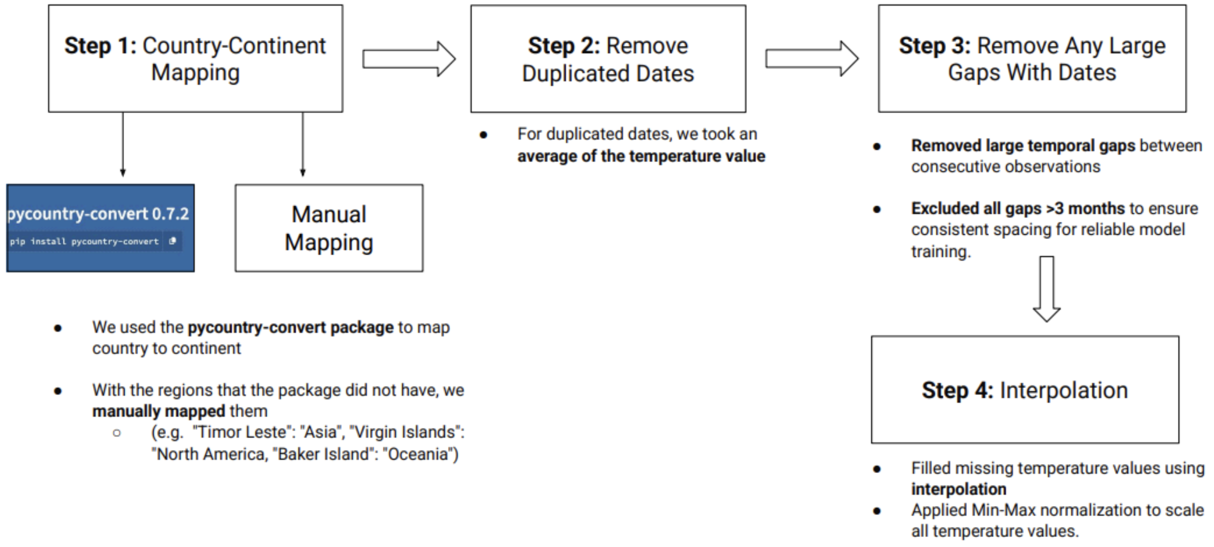
Limitations and Future Work

This study focuses on average temperature as the sole input variable, which limits the ability to capture interactions among other climate factors such as precipitation, atmospheric circulation, or emissions. Some regions, especially Oceania, contain fewer observations and greater variability, which may reduce model stability. Antarctic regions were excluded because of limited data availability. Future work can include additional climate variables to support multivariate forecasting and explore deeper architectures or transformer-based models that may better capture long-range climate patterns. Expanding the dataset and developing hierarchical global and regional models may further improve forecasting accuracy and interpretability.

Conclusion

This study demonstrates that climate forecasting performance depends strongly on both the neural architecture and the geographic scale of analysis. At the global level, the Enhanced Attention LSTM achieved the lowest RMSE and MAE and the highest directional accuracy, highlighting the value of engineered seasonal features, lagged signals, and smoothing operations for modeling large-scale climate trends. At the continent level, the Attention Bi-LSTM consistently produced the lowest error metrics across all six regions, indicating that regional climate sequences benefit from bidirectional temporal encoding and richer contextual information. Directional accuracy remained relatively stable across architectures, suggesting that all models were similarly capable of capturing the direction of month-to-month temperature change. These findings establish a reproducible framework for regional and global climate forecasting and motivate future work on architectures that more effectively capture multi-scale climate variability.

Appendix


Step 1: Country-Continent Mapping
pycountry-convert 0.7.2
pip install pycountry-convert
Manual Mapping
Step 2: Remove Duplicated Dates
Step 3: Remove Any Large Gaps With Dates
Step 4: Interpolation

- We used the **pycountry-convert package** to map country to continent
- With the regions that the package did not have, we **manually mapped** them
  - (e.g. "Timor Leste": "Asia", "Virgin Islands": "North America, "Baker Island": "Oceania")

- For duplicated dates, we took an **average of the temperature value**

- **Removed large temporal gaps** between consecutive observations
- **Excluded all gaps >3 months** to ensure consistent spacing for reliable model training.

- Filled missing temperature values using **interpolation**
- Applied Min-Max normalization to scale all temperature values.

Figure 1: Pre-processing Pipeline Flowchart

Model	RMSE	MAE	Directional Accuracy
Baseline Attention LSTM	0.0258	0.0155	0.842
Enhanced Attention LSTM	0.0204	0.0133	0.875
Attention Bi-LSTM	0.0211	0.0155	0.870

Table 1: Global Model Comparison

Continent	Baseline Attention LSTM	Enhanced Attention LSTM	Attention Bi-LSTM
Europe	4.39117	3.02192	0.041422
Asia	2.036672	1.65024	0.031520
Africa	1.204461	1.07208	0.034632
North America	1.41378	3.075578	0.028855
Oceania	0.6802	0.885159	0.049648
South America	1.082204	2.60925	0.085169

Table 2: RMSE Comparison Across Three Continent-Level Models

Continent	Baseline Attention LSTM	Enhanced Attention LSTM	Attention Bi-LSTM
Europe	3.693739	2.478485	0.033690
Asia	1.249834	1.138645	0.023536
Africa	0.839264	0.723852	0.024564
North America	0.735268	1.732746	0.019073
Oceania	0.489724	0.699873	0.038547
South America	0.722365	2.226599	0.071998

Table 3: MAE Comparison Across Three Continent-Level Models

Continent	Baseline Attention LSTM	Enhanced Attention LSTM	Attention Bi-LSTM
Europe	0.869093	0.881826	0.883119
Asia	0.885417	0.904542	0.881022
Africa	0.830061	0.851309	0.853076
North America	0.84163	0.847692	0.850125
Oceania	0.651606	0.702714	0.699769
South America	0.792405	0.694011	0.687718

Table 4. DA Comparison Across Three Continent-Level Models

Works Cited

Bajaj, Aayush. “Understanding Gradient Clipping (and How It Can Fix Exploding Gradients Problem).” Neptune.ai, 21 July 2022, neptune.ai/blog/understanding-gradient-clipping-and-how-it-can-fix-exploding-gradients-problem.

Chen, Zonglei, et al. “Long Sequence Time-Series Forecasting with Deep Learning: A Survey.” Information Fusion, Apr. 2023, p. 101819, https://doi.org/10.1016/j.inffus.2023.101819.

“Climate Change: Earth Surface Temperature Data.” Www.kaggle.com, Berkeley Earth, www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data.

Duy, Huynh Anh, and Tarapong Srisongkram. “Bidirectional Long Short-Term Memory (BiLSTM) Neural Networks with Conjoint Fingerprints: Application in Predicting Skin-Sensitizing Agents in Natural Compounds.” Journal of Chemical Information and Modeling, 3 Mar. 2025, https://doi.org/10.1021/acs.jcim.5c00032.

IPCC. “Synthesis Report of the IPCC Sixth Assessment Report (AR6) Summary for Policymakers.” Intergovernmental Panel on Climate Change, 2023.

Kang, Qing, et al. “Attention-Based LSTM Predictive Model for the Attitude and Position of Shield Machine in Tunneling.” Underground Space, vol. 13, 1 Dec. 2023, pp. 335–350, https://doi.org/10.1016/j.undsp.2023.05.006.

Koch, Daniel. “Pycountry-Convert: Extension of Python Package Pycountry Providing Conversion Functions.” PyPI, pypi.org/project/pycountry-convert/.

Kong, Yaxuan, et al. “Unlocking the Power of LSTM for Long Term Time Series Forecasting.” ArXiv.org, 2024, arxiv.org/abs/2408.10006.

Lazzeri, Francesca. “Introduction to Feature Engineering for Time Series Forecasting.” Data Science at Microsoft, 2 Nov. 2021, medium.com/data-science-at-microsoft/introduction-to-feature-engineering-for-time-series-forecasting-620aa55fcab0.

Lepot, Mathieu, et al. “Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment.” Water, vol. 9, no. 10, 17 Oct. 2017, p. 796, https://doi.org/10.3390/w9100796.

Low, Remy, and Helen Proctor. “Oceania and the History of Education.” History of Education, vol. 52, no. 2-3, 4 May 2023, pp. 201–219, https://doi.org/10.1080/0046760x.2023.2196512. Accessed 10 Jan. 2024.

Matheus Valença. “Sine and Cosine Transformation and Normalization - Matheus Valença - Medium.” Medium, 20 Aug. 2024, medium.com/@valencamatheus97/sine-and-cosine-transformation-and-normalization-491a6f71c091.

NASA. “Evidence.” Science.nasa.gov, NASA, 23 Oct. 2024, science.nasa.gov/climate-change/evidence/.