From Temperature Anomalies to El Niño – Using Machine Learning to Predict El Niño Index

Michaela Robinson, Data Science, Fisk University **Chloe Banks,** Computer Science, Alabama State University **Mentor**: **Dr. Lei Qian**, Meharry Medical College

BACKGROUND

Temperature anomalies are deviations from long-term average values that show how current temperatures differ from historical norms. The **Southern Oscillation Index (SOI)**, which measures air pressure differences between Tahiti and Darwin, links these anomalies to El Niño events: negative SOI values are typically associated with warmer sea surface temperatures and disrupted weather patterns worldwide. Climate variability, especially **El Niño–Southern Oscillation (ENSO)** events, strongly influences weather, agriculture, and ecosystems. Accurate prediction of ENSO and its associated global temperature anomalies is critical for climate research, disaster preparedness, and policy planning.

Traditional statistical forecasting methods for ENSO events often struggle to capture the complex, non-linear interactions between oceanic and atmospheric variables, resulting in limited prediction accuracy. Recent advances in machine learning offer promising opportunities to improve ENSO prediction by identifying intricate patterns in large climate datasets that conventional approaches may miss.

In this study, we analyzed NOAA historical climate datasets—including global temperature anomalies, SOI, and sea surface temperatures—to explore patterns of global temperature anomalies and develop forecasting models for El Niño events. We examined the global temperature anomaly trends from 1850–2024 by identifying the 10 hottest and 10 coolest years within this 175-year period and applying multiple smoothing algorithms, including moving averages, Gaussian Smoothing, and **LO**cally **WE**ighted **S**catterplot **S**moothing (LOWESS), to reduce noise and better understand underlying trends.

We implemented multiple **machine learning frameworks** that apply **Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs)** to capture non-linear relationships between oceanic and atmospheric indicators. By integrating multiple data sources, advanced preprocessing, and deep learning models, our approach demonstrates the forecasting. This work highlights the potential of machine learning for more accurate ENSO prediction and provides a scalable blueprint for combining large climate datasets with data-driven models, contributing to better understanding and preparedness in a changing climate.

TOOLS and MATERIALS

Core datasets are derived from the **National Oceanic and Atmospheric Administration (NOAA)**, including monthly **global land and ocean temperature anomaly records** and **Southern Oscillation Index (SOI) measurements, and the COBE-SST 2 Sea and Ice measurements** dating from 1950 to 2025. ENSO phase indicators (El Niño, La Niña) are incorporated to contextualize anomaly trends. Supplementary climate drivers, such as long-term CO_2 concentration and sea surface temperature indices, are prepared for future model extensions.

The predictive framework is built heavily in Python, leveraging Jupyter Notebooks for exploratory analysis and prototyping. Google Colab provides a cloud-based environment with GPU acceleration for deep learning training runs, ensuring computational efficiency without requiring local high-performance hardware. Model development leveraged **PyTorch** for building and training both CNN and LSTM architectures, taking advantage of its dynamic computation graph and GPU acceleration. **Scikit-learn** supported preprocessing, feature scaling, train-test splitting, and baseline regression models for comparison.

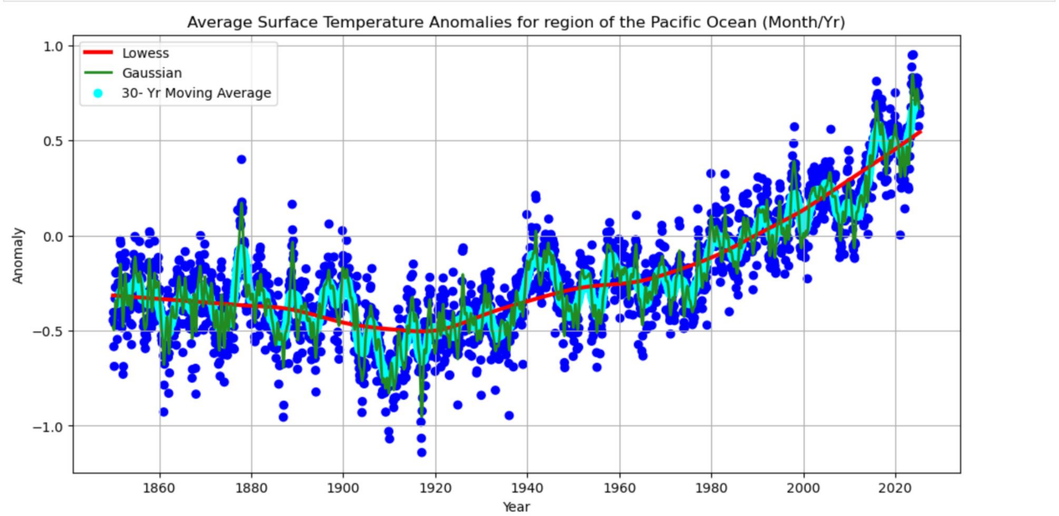
The predictive modeling framework was developed using **Python** as the core programming language, with the primary development environment being **Jupyter Notebook** for constructing and iterating on the LSTM, CNN, and GRU models.. This architecture was implemented using **PyTorch** to leverage GPU-accelerated deep learning capabilities, while **scikit-learn** supported data preprocessing, feature scaling, and baseline regression models for performance comparison. Additional libraries such as **xarray** and **netCDF4** were employed for handling time-series climate data in CSV and NetCDF formats, and **matplotlib**, **seaborn** were used to visualize anomaly trends, ENSO phases, and model predictions.

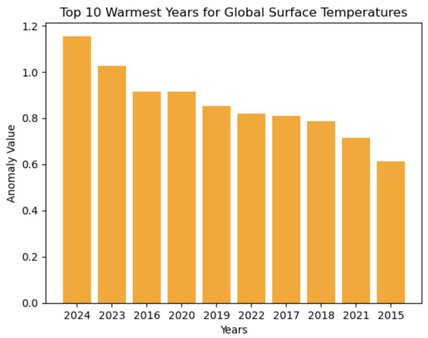
METHODOLOGY

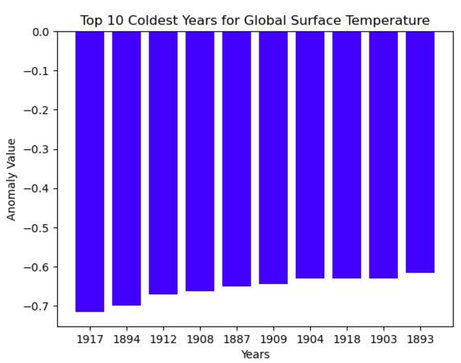
To predict the El Niño/La Niña Index, we collected key datasets, including ENSO indices (SOI, Niño 3.4 anomalies), global temperature, and additional predictors like sea surface temperatures and atmospheric variables. The data was preprocessed by aligning temporal resolution, normalizing features, handling missing values, and engineering lagged variables to capture delayed ENSO effects. To aid with visualization of global temperature anomaly trends, we applied smoothing techniques including Gaussian filtering, LOWESS, and moving averages. We then trained and compared multiple models, including deep learning approaches (LSTM, CNN, GRU) to leverage both spatial and temporal patterns. The models were validated using time-series cross-validation and evaluated using mean squared error (MSE) across three inputs: SST-only, global anomaly-only, and a combined dataset. The results were visualized with bar charts to compare model performance for each forecast horizon (1 month, 3 months, 6 months).

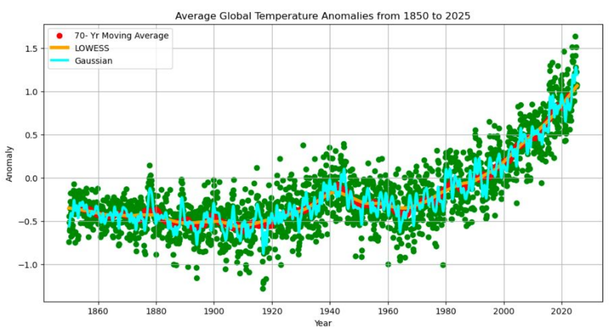
The charts below indicate that there has been an increase in global temperatures since 1960. The Pacific Ocean (El Niño region) shows an accelerated increase as well. Additionally, the warmest years for global surface temperatures mostly occur within the last 10 years, with 2024 being the highest.

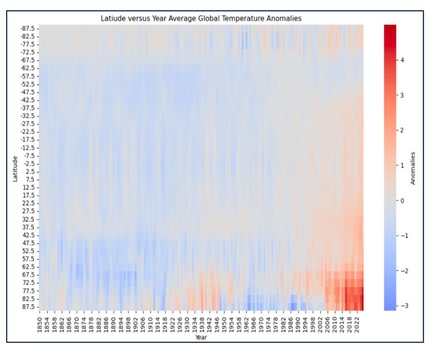
Software and Platforms





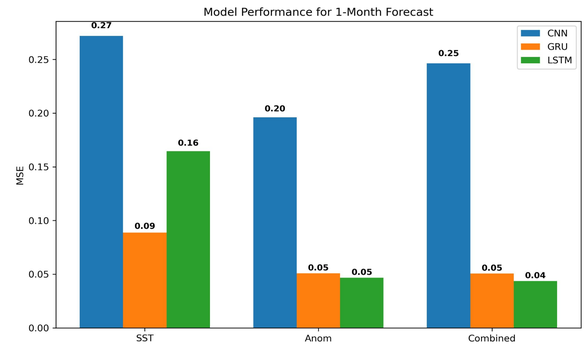


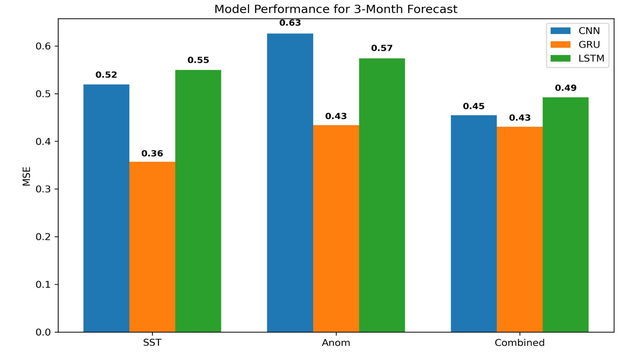


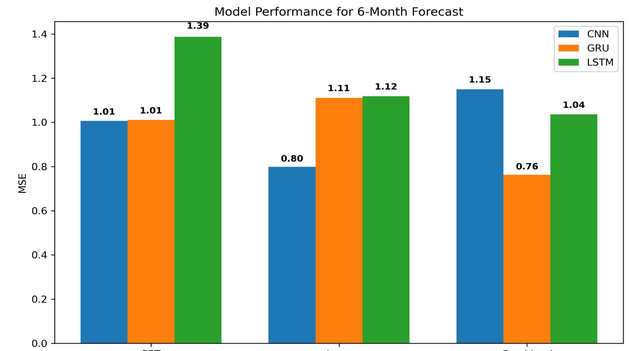


RESULTS

Model Performance







When performing **1 month predictions,** the combined dataset achieved the lowest errors with the LSTM model showing the best performance (MSE=0.04). For **3 month forecasts**, all models began declining in performance. The combined data set also shows that the GRU outperformed other model accuracies (MSE=0.43). **Forecasts 6 months** ahead shows increased errors, but the combined dataset with the LSTM and and GRU models produced the best results.

ACKNOWLEDGEMENTS

This research was supported by MS-CC, Meharry Medical College, and Fisk University. We also thank Dr. Hussain, Dr. Muallem, and other faculty members for their valuable support and assistance

REFERENCES

- Huang, B.; Yin, X.; Menne, M. J.; Vose, R. S.; Zhang, H. -M. (2024-02-13). *NOAA Global Surface Temperature Dataset (NOAAGlobalTemp), Version 6.0*
- Japan Meteorological Agency (JMA). (2014). *Centennial in situ Observation-Based Estimates of Sea Surface Temperature (COBE-SST2)* [Data set]. NOAA Physical Sciences Laboratory. https://psl.noaa.gov/data/gridded/data.cobe2.html
- National Centers for Environmental Information. (2025, August). *El Niño / Southern Oscillation (ENSO) monitoring*. NOAA Climate Monitoring. Retrieved August 1, 2025, from NOAA NCEI website: https://www.ncei.noaa.gov/access/monitoring/enso/