Practical Machine Learning - PHAS 0056

Mini-Project Report: Group C

Data-Driven Forecasting of Climate and Weather Using LSTM and XGBoost

Abstract

Accurate weather and climate forecasting is crucial to many industries such as agriculture and energy, public safety, and environment planning. Whilst traditional, physics-based models like numerical weather prediction (NWP) are widely used, they require significant computational power and are also sensitive to initial conditions. This project aims to explore the use of deep learning methods - specifically, Long Short-Term Memory (LSTM) and traditional machine learning techniques like XGBoost regression, to time series temperature and precipitation predictions. I train and evaluate these models on the Global Historical Climatology Network (GHCN-Daily) dataset, a large archive of meteorological observations. Daily observations are pre-processed, normalized, and converted to input-output sequences suitable for recurrent neural networks. Performance of the model is evaluated based on standard regression metrics like Mean Squared Error (MSE) and Mean Absolute Error (MAE). Our results demonstrate that LSTM models are effective in capturing temporal dependencies, while XGBoost offers strong baseline performance with minimal tuning. This study highlights the strengths and limitations of data-driven approaches and their potential to complement traditional forecasting methods.

Keywords: *climate forecasting, LSTM, XGBoost, time series analysis, weather prediction, GHCN, deep learning, environmental modelling*

1 Introduction

1.1 Scientific Motivation

Weather and climate forecasting accurately is crucial for a wide range of practical applications - from managing power grids and guiding crop cycles to anticipating extreme events and planning urban infrastructure. However, traditional forecast systems, particularly numerical weather prediction (NWP) models, can be somewhat limited. These models rely on solving fluid dynamic equations under initial and boundary conditions, which makes them computationally expensive and sensitive to uncertainties, particularly in regions weakly observed. (Bauer et al., 2015; Benbouzid et al., 2021).

Machine learning (ML) offers an alternative route by learning directly from observational data. Instead of simulating physical processes, ML models learn about the underlying temporal and spatial patterns in the data, enabling efficient and often surprisingly accurate forecasts - even with sparse or noisy inputs (Reichstein et al., 2019). Models specially designed for time series like Long Short-Term Memory (LSTM) networks work particularly well at learning delayed effects and recurring seasonal cycles, whereas tree-based models like XGBoost can learn sharp thresholds and outlier behavior.

In this context, data-driven models hold promise as either standalone tools or as components of hybrid forecast systems, especially over high-latitude regions where physical models are likely to fail.

1.2 Study Region: Arctic Bay, Canada

This study focuses on Arctic Bay CS (GHCN ID: CA002400404) of a well-established weather station located at 73.0°N, –85.02°W in Qikiqtaaluk Region, Canada. Since it falls inside the Arctic Circle, it is among the most developed areas in the realm of climate change. The surface air temperatures above the Arctic have risen nearly four times the global mean between the years 1979, a phenomenon called Arctic amplification Rantanen (2022). This is primarily driven by feedback processes such as ice–albedo loss, ocean–atmosphere heat transfer, and persistence of the greenhouse gases.

High-latitude, long term weather stations like Arctic Bay provide rare and valuable "ground-truth" data to evaluate and validate climate models. The pronounced seasonality of this region as well as its dominant trend signals, and global climatic significance make it an ideal choice for deep modeling.

Figure 1 displays the 365-day rolling average of daily maximum (TMAX) and minimum (TMIN) temperatures from 1999 to 2021. Not only does this graph confirm a persistent warming trend, but it also shows an expanding difference between TMAX and TMIN in recent years—suggesting increased diurnal variability and reduced seasonal cold extremes, both known indicators of Arctic climate change (Screen and Simmonds, 2012).

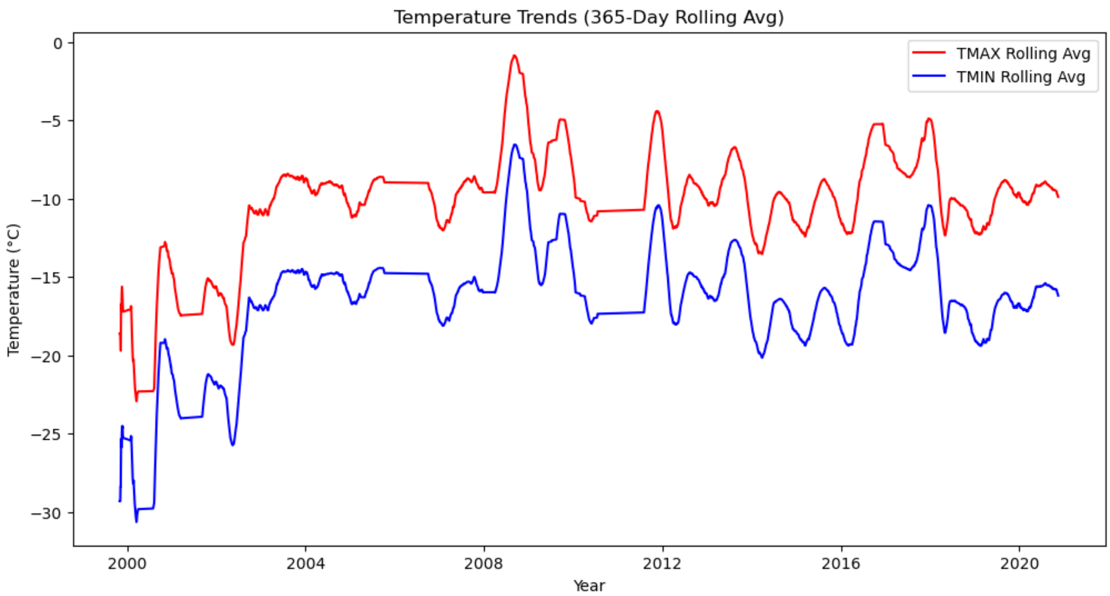

Figure 1: 365-day rolling average of TMAX and TMIN temperatures at Arctic Bay CS (GHCN ID: CA002400404)

1.3 Project Aims and Structure

This project investigates whether machine learning models are effective in predicting temperature and precipitation using the Global Historical Climatology Network Daily (GHCN-D) dataset. The following two tasks are explored:

- **Task 1 — Climate Forecasting:** Predicting the monthly average TMAX, TMIN, and PRCP with seasonal data and LSTM/XGBoost models.
- **Task 2 — Weather Forecasting:** Predicting daily TMAX, TMIN, and PRCP from past daily values using LSTM/XGBoost models.

LSTM and XGBoost models are mainly used and compared in order to evaluate and find the most efficient model for forecasting. Additionally, I explore four extension tasks which investigate long-term temperature forecasts, regional temperatures and spacial correlations between global stations.

2 Data Selection and Pre-processing

2.1 Variable Selection and Pre-processing

The core variables used are daily maximum temperature (TMAX), minimum temperature (TMIN), precipitation (PRCP), and snowfall (SNOW). After merging these series by date, missing or flagged values were addressed through a combination of imputation and logical inference. Specifically:

- **Missing TMAX/TMIN/PRCP values** were filled using forward and backward filling, ensuring temporal continuity for sequence-based models.
- **SNOW values** posed a unique challenge: data were highly sparse, with all zero values after 2007. Due to the lack of informative variability and limited coverage period, the SNOW variable/values were excluded from further analysis for this station

2.2 Outlier Handling and Distribution Analysis

As shown in 2, TMAX and TMIN demonstrate a strong linear relationship with a Pearson correlation coefficient of 0.98. The scatter plot clearly shows a diagonal line of dense points - indicating that warmer days tend to have warmer nights, and vice versa. Such a strong correlation is typical in stable climate regions where the maximum temperatures and daily minimum are influenced by shared atmospheric conditions, such as air mass temperature and cloud cover Dai et al. (1999). However, this does introduce potential multicollinearity concerns in models that attempt to use both as independent features or predict both simultaneously James et al. (2013). Therefore, it is important to ensure that this interdependence is accounted for when selecting features and designing model structures, to avoid any overfitting or redundancy.

TMAX and TMIN exhibited no extreme outliers. PRCP, however, showed a highly skewed distribution with a long positive tail. Logarithmic transformation was applied to PRCP values (after adding a small offset) to improve model learning stability. A total of 894 PRCP values were flagged as statistical outliers, but most were retained to preserve rare event dynamics.

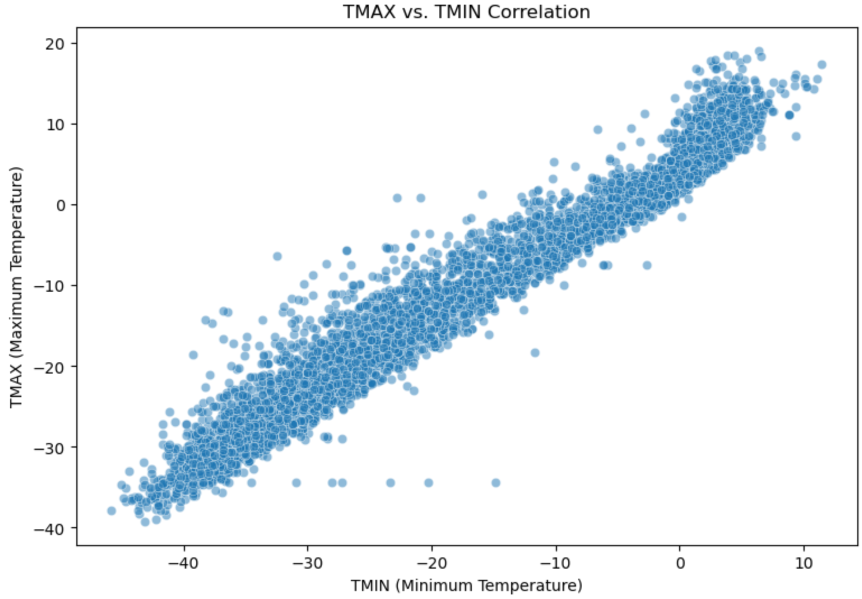

Figure 2: Scatter plot showing correlation between TMAX and TMIN across all daily observations.

Figure 3 illustrates the raw PRCP distributions along with a box plot which shows that most PRCP values are 0. This is expected, given the polar climatology of Arctic Bay, where precipitation occurs infrequently but may be intense when it does occur. Additionally, the box plot allows us to gauge where most of the other values lie - further motivating the use of a log transformation.

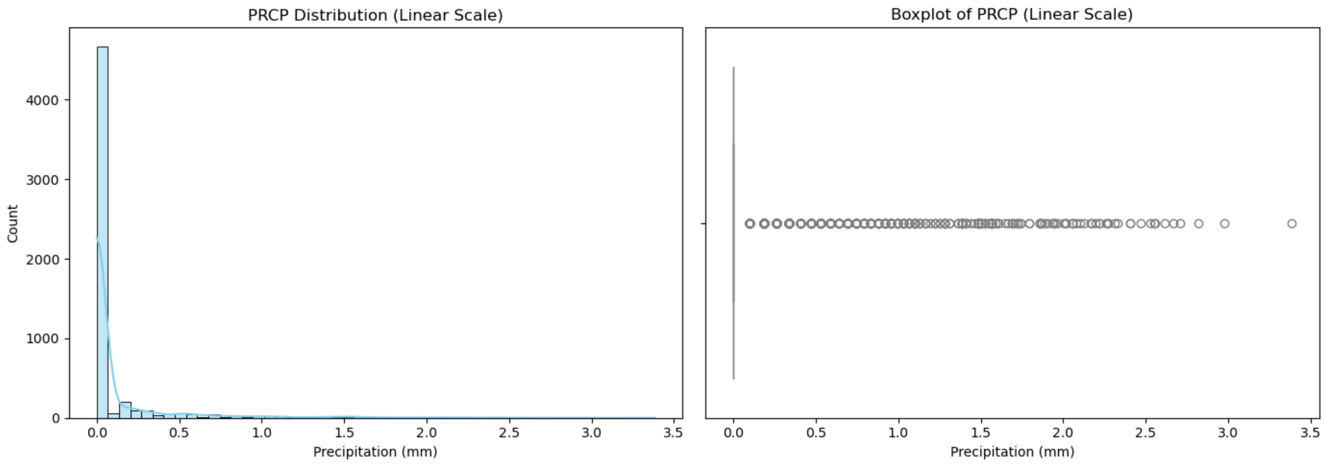

Figure 3: Distribution of raw daily precipitation (PRCP) values at Arctic Bay CS. Values are highly skewed toward zero, with most days showing no rain.

2.3 Feature Engineering

Several additional features were constructed to enrich model input:

- **Lagged variables:** For each core feature, 1-day lag values (e.g., `TMAX_lag1`) were generated. These are essential for models like XGBoost which does not handle sequence data.
- **Calendar features:** Month and day-of-week were included as cyclical proxies for seasonality.

2.4 Scaling and Sequence Preparation

All numerical variables were scaled to the $[0, 1]$ range using Min-Max scaling, fitted only on the training data to avoid leakage.

- For **LSTM models**, input data were reshaped into fixed-length sliding windows (e.g., 30-day sequences) to form supervised learning samples. Each sequence contains past values of TMAX, TMIN, PRCP.
- For **XGBoost models**, flattened lag features were used instead. No temporal memory is retained between samples in this case, so proper lagging and calendar features were essential to encode time implicitly.

2.5 Final Dataset Summary

After preprocessing, the final dataset contained 5,566 rows spanning from November 1999 to November 2021. This timeframe provides over two decades of strong metereological observations, sufficient to capture seasonal, inter-annual and decadal variability - particularly important for Arctic climate analysis.

Variable	**Mean**	**Std**	**Min**	**Max**
TMAX (°C)	–9.66	14.12	–39.2	19.1
TMIN (°C)	–15.92	14.90	–45.8	11.5
PRCP (log)	0.098	0.31	0.00	3.38
SNOW (cm)	0.24	3.08	0.00	80.0

Table 1: Summary statistics of key features after preprocessing and transformation.

Despite gaps in some of the rare data, the station exhibits strong temporal coverage with minimal periods of total data loss. Key descriptive statistics are shown in Table 1. It is clear to see the average daily maximum and minimum temperatures are both well below freezing, reflecting the extreme polar climate at Arctic Bay CS. The maximum TMAX of 19.1 degrees Celsius and minimum TMIN of -45.8 Celsius highlight the wide annual range in surface temperatures - an essential aspect for models to learn.

3 Task 1: Predicting Climate (long-term)

3.1 Methodology

Model setup: Both models were trained on monthly-averaged data from 1999 to 2020 and tested on 2021, using a strict chronological split to prevent data leakage. The training dataset contained approximately 260 samples (12 per year $\times$ 21 years). All features were normalized using Min-Max scaling based on the training set.

LSTM architecture: The LSTM model used a single hidden layer with 50 units and ReLU activation, trained to predict one month ahead using a 12-month sliding window. Inputs included three features: TMAX, TMIN, and PRCP. Dropout (20%) was applied to reduce overfitting, and the model was trained for 200 epochs with early stopping.

XGBoost setup: The XGBoost model was trained using lag features from the previous 12 months for each target variable, along with month encoding. Default hyper-parameters were used with 100 estimators and learning rate 0.1. Unlike LSTM, this model has no temporal memory and learns purely from structured lagged inputs.

Evaluation metrics: Performance was assessed using Mean Absolute Error (MAE) and the coefficient of determination (R^2). LSTM achieved MAE = 1.44 and R^2 = 0.63 on average, while XGBoost achieved MAE = 0.28 and R^2 = 1.00. Although the latter appears near-perfect, its generalisability is questionable given the small dataset size.

3.2 Results and Analysis

. The following plots demonstrate the actual vs predicted TMAX, TMIN and PRCP of the unseen data in 2021, using an LSTM and XGBoost model. A residual plot is also presented side by side.

Figure 4 shows LSTM predictions versus actual monthly values in 2021. General periodicity of the seasonality cycle in the model is captured however, as with other models, the temperature extremes are underestimated. TMAX peaks, during the summer months, are greatly smoothed while TMIN is captured slightly better. PRCP predictions show low variability with missing sharp precipitation spikes - a common challenge when using deep learning to model sparse or zero-inflated series on limited data.

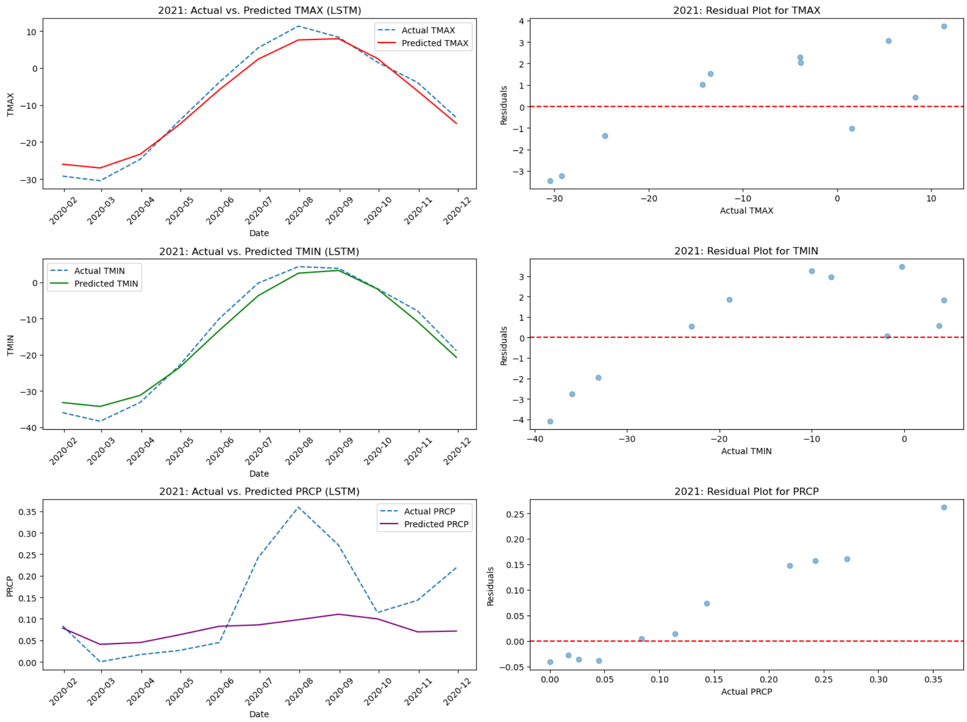

Figure 4: LSTM predictions vs actual monthly values for TMAX, TMIN, and PRCP (2021).

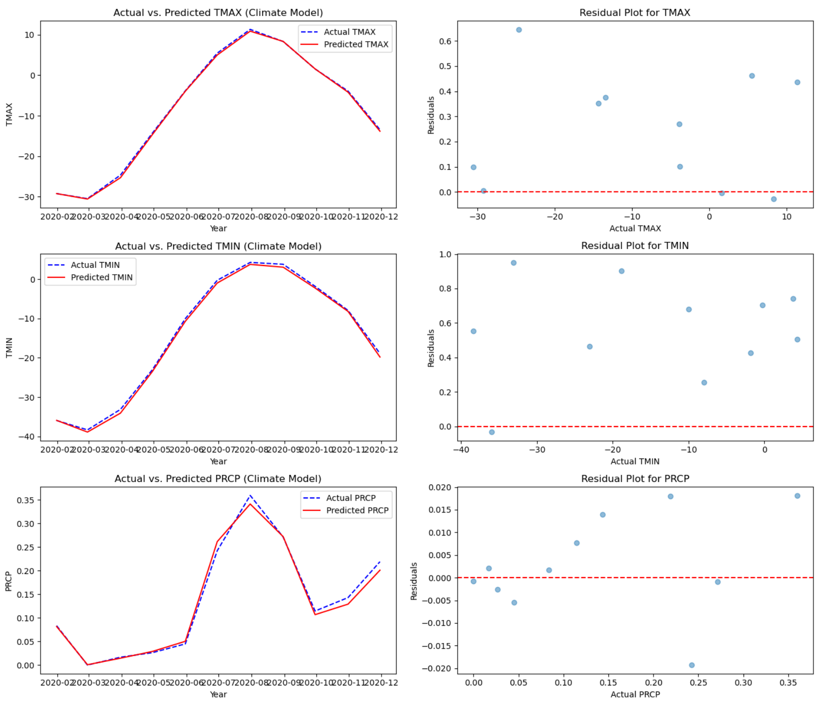

Figure 5: XGBoost predictions vs actual monthly values for TMAX, TMIN, and PRCP (2021).

Figure 5 presents the XGBoost model's predictions. It successfully captures the seasonal cycle with high fidelity, hitting temperature peaks and the general shape of the expected PRCP signal. It even gets the minor fluctuations in precipitation that the LSTM model would otherwise miss. One significant point to consider is that this could be a sign of overfitting rather than genuine predictive robustness, as XGBoost's tree-based structure has the capacity to memorize trends when trained on limited data.

Model	**MAE**	R^2 **Score**
XGBoost	0.28	1.00
LSTM	1.45	0.64

Table 2: Comparison of XGBoost and LSTM performance for climate prediction on 2021 monthly data.

Table 2 directly compares the two model across all variables using Mean Absolute Error (MAE) and the coefficient of determination (R^2). Although XGBoost achieves near-perfect performance (with a 0.28 MAE) on this highly structured, seasonal dataset, its performance are likely influenced by overfitting. LSTM is less precise but offers a more flexible and generalizable modeling approach — a merit that would be more apparent in multivariate or multi-location scenarios.

4 Task 2: Predicting Weather (short-term)

4.1 Objective and Setup

This task focuses on short-term forecasting of daily maximum temperature (TMAX), minimum temperature (TMIN), and precipitation (PRCP). Unlike the previous task, the goal here is to model high-frequency, short-memory signals with strong day-to-day variability. This type of forecasting is more sensitive to noise, autocorrelation, and sharp transitions, and is relevant for operational use cases such as event planning and infrastructure management.

To enable daily prediction, the preprocessed dataset (1999–2021) was structured into a supervised learning format using sliding windows of past daily observations. The models were trained to forecast the next day's values for each variable.

4.2 Model Design and Methodology

LSTM setup: The LSTM architecture was similar to that used in Task 1, but with sequence inputs based on 30-day rolling windows. Each input sample included 30 previous days of TMAX, TMIN, and PRCP. A single-layer LSTM with 64 units was used, followed by a dense output layer. Dropout regularization and early stopping were applied to prevent overfitting.

XGBoost setup: The XGBoost model was trained using lagged features for each variable (e.g., TMAX_lag1, TMAX_lag2, ..., TMAX_lag30), alongside calendar features like day-of-week and month. These additional features help capture periodic effects and improve model discrimination.

Baseline model: A naive baseline was also implemented, where each variable's value for day t is predicted as the value from day $t-1$. This provides a simple benchmark for evaluating model skill.

Training and evaluation: All models were trained on 1999–2020 data and evaluated on 2021. The test year was excluded from model fitting to simulate real-world deployment. Performance metrics include MAE and R^2 on the test set.

4.3 Results and Analysis

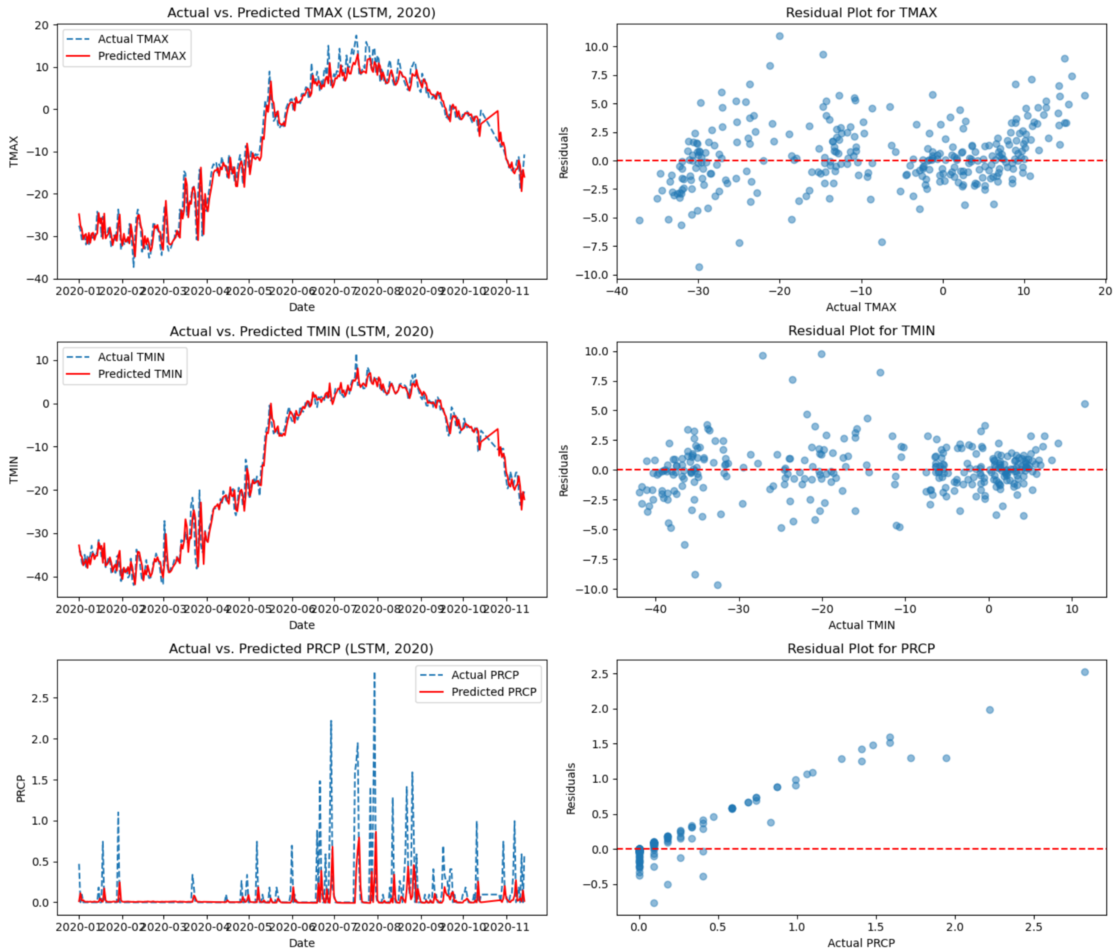

Figure 6: LSTM predictions vs actual daily values for TMAX, TMIN, and PRCP (2021).

LSTM performance: The LSTM model achieved an MAE of 1.44 and an R^2 score of 0.63, similar to its performance in Task 1. The model (presented in Figure 6) captured medium-term fluctuations but struggled with high-frequency transitions. This is likely due to both the data volume (daily forecasting needs much more training data than monthly) and the relatively shallow architecture.

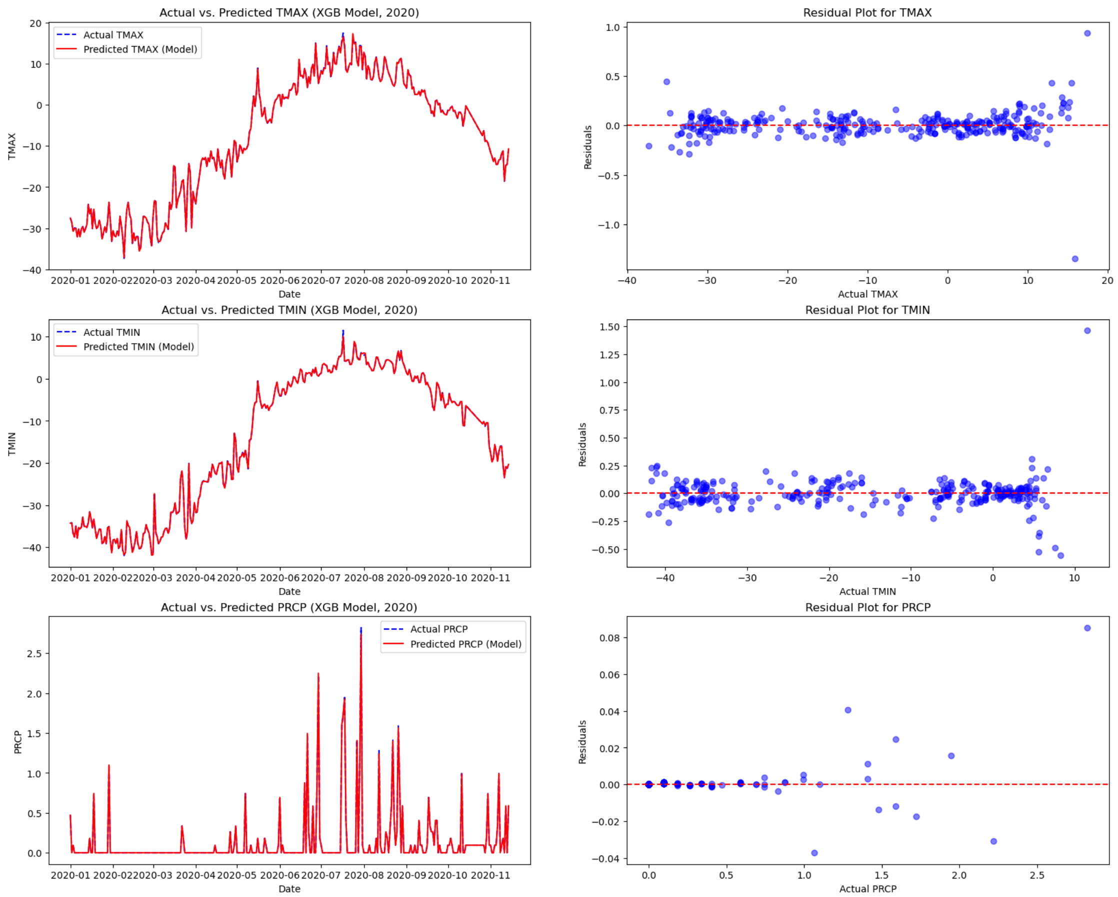

Figure 7: XGBoost predictions vs actual daily values for TMAX, TMIN, and PRCP (2021).

XGBoost performance: XGBoost achieved an MAE of 0.05 and an R^2 score of 1.00, strongly outperforming LSTM and the naive baseline. This highlights its ability to leverage lagged values and calendar patterns in structured time series. The model (seen in Figure 7) successfully replicated spikes and dips in all three variables with minimal error, though generalizability remains a concern due to the risk of overfitting.

Model	MAE	R^2 Score
XGBoost	0.05	1.00
LSTM	1.25	0.61

Table 3: Comparison of XGBoost and LSTM performance for daily weather prediction in 2020.

Also, the baseline comparison tells us that the naive model achieved an MAE of 1.34, emphasizing that while previous-day values are informative, both ML models (especially XGBoost) significantly outperform such simple baselines. In fact you can see in Table 3, the MAE value for XGBoost, for example, is the lowest (0.05).

XGBoost provides robust and highly accurate short-term predictions for Arctic Bay daily weather, far outperforming the LSTM model and naive baseline. These results suggest that tree-based ensemble models are especially effective in structured, high-frequency forecasting tasks, given adequate feature engineering. LSTM, while more generalizable, was heavily limited by training data size and high noise in PRCP.

5 Extensions

5.1 Long-Term Forecasting at Arctic Bay

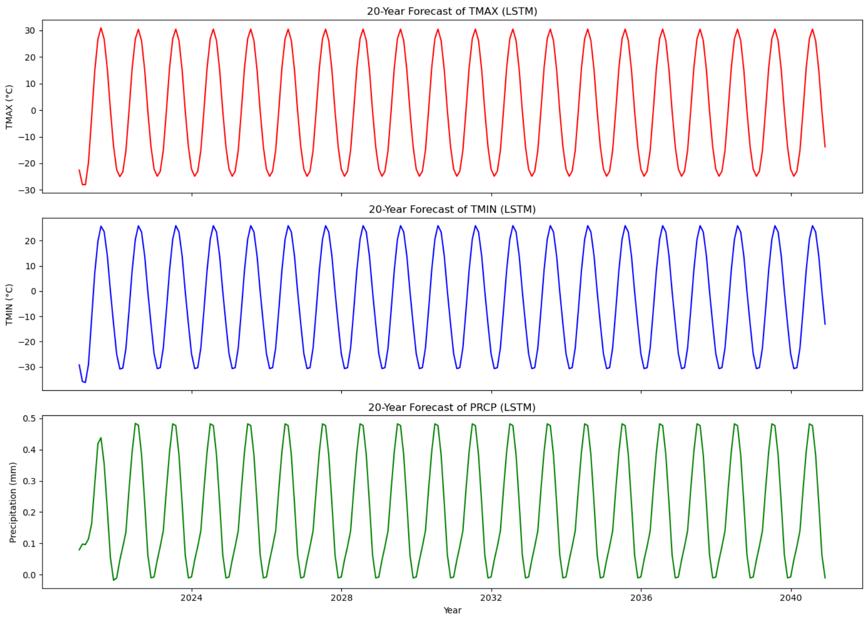

Figure 8: 20-year LSTM forecast for TMAX, TMIN, and PRCP at Arctic Bay CS.

To assess the potential for long-range climate prediction, I used the LSTM model trained on Arctic Bay CS to generate a 20-year forecast of monthly TMAX, TMIN, and PRCP. Starting with the final 12 months of 2021, the model recursively predicted each subsequent month, feeding predictions back as input. This autoregressive setup enables long-term projection but assumes stationarity in underlying climate patterns.

The model successfully preserves annual cycles but shows no trend in warming or rainfall change, reflecting its limitations in extrapolating beyond past dynamics.

5.2 Identifying the Hottest Locations in 2044

To evaluate spatial differences in projected climate, LSTM models were trained on eight global stations with diverse climatic regimes. Each model was used to generate a 20-year forecast, from which the average TMAX in 2044 was extracted.

To assess regional warming trends, LSTM models were trained on eight global stations and used to forecast monthly TMAX through 2044. Figure 9 shows the projected change in average temperature relative to the 2025 baseline.

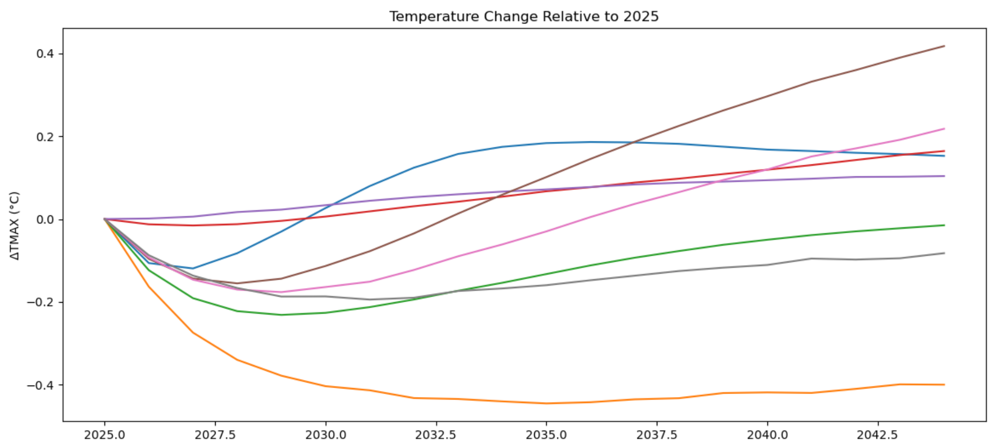

Figure 9: Projected change in TMAX relative to 2025 for eight global stations.

Most stations exhibit gradual warming consistent with global climate trends, with tropical and maritime sites like Montego Bay and Guam maintaining consistently high temperatures (Figure 9). Mid-latitude locations such as Toulouse and Rennes show moderate increases, while Pocatello (USA) unexpectedly displays a slight cooling trend. These variations may reflect model limitations in capturing long-term anthropogenic trends, as the LSTM was trained solely on historical data without incorporating exogenous drivers such as greenhouse gas concentrations. Nonetheless, the forecast preserves broad climatological distinctions and offers insights into relative warming patterns under a data-driven framework.

Montego Bay (Jamaica) and Guam consistently emerge as the hottest stations by 2044, with average TMAX values exceeding 30°C. This aligns with climatological expectations given their low-latitude, maritime locations, where strong year-round solar insolation and minimal seasonal variability sustain persistently high temperatures. In contrast, mid-latitude and continental stations such as Pocatello and Mont-Aigoual exhibit lower projected temperatures and stronger seasonal oscillations. These results demonstrate the model's ability to preserve inter-station climatological distinctions over long forecasting horizons, even when trained only on historical temperature time series. While the LSTM models do not account for long-term climate drivers or emissions scenarios, their projections maintain geographically coherent trends that align well with known climatic regimes.

5.3 Non-Local Predictors: Utah Sun and Global Rainfall

This extension investigates potential teleconnections by examining whether solar input in Utah (TMAX at Salt Lake City) can help predict rainfall (PRCP) at distant stations in Spain, Brazil, and the Philippines. Using detrended monthly data, Pearson correlations were computed over lagged time windows (–12 to +12 months).

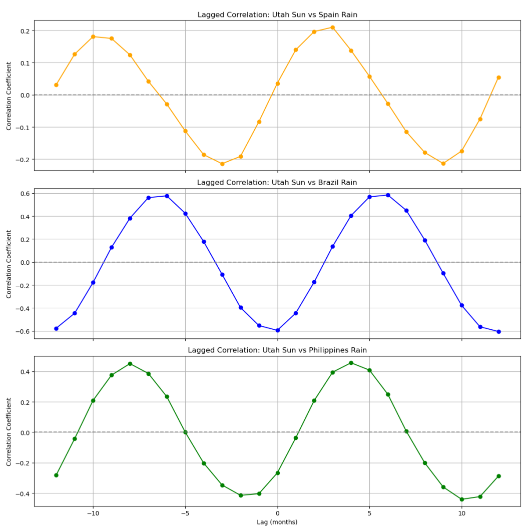

Figure 10: Lagged correlations between Utah TMAX and PRCP at distant stations.

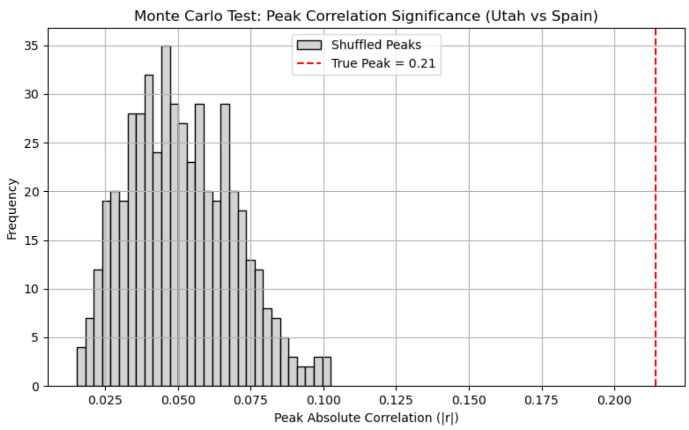

Figure 11: Monte Carlo permutation test for Utah–Spain correlation.

These results suggest that certain apparent teleconnections, such as the weak anti-correlation between Utah temperature and rainfall in Brazil, may be statistically significant, as indicated by peak correlations exceeding shuffled distributions (Figure 11). However, caution is warranted in interpreting these links as causal or physically meaningful. In many cases, the observed correlations are likely driven by common seasonal cycles rather than dynamic atmospheric coupling. For instance, high insolation in Utah during boreal summer coincides with regional dry seasons elsewhere, producing coincidental anti-correlation. These findings highlight the need for rigorous significance testing and climatological context when inferring teleconnections from observational data alone.

5.4 Distance-Based Station Correlation

This task investigated how inter-station correlations of climate variables decay with geographical distance. In addition to maximum temperature (TMAX), I included precipitation (PRCP) to assess how spatial coherence varies between variables.

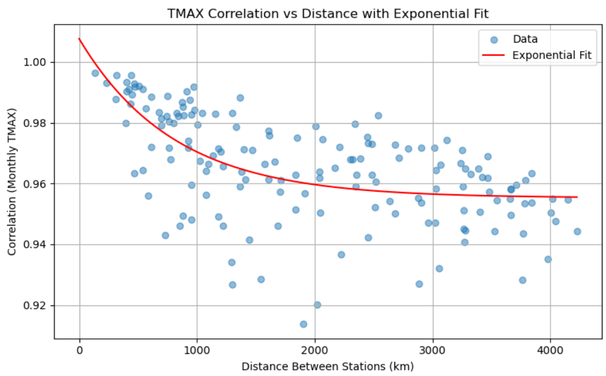

Figure 12: Monthly TMAX correlation vs distance from Arctic Bay.

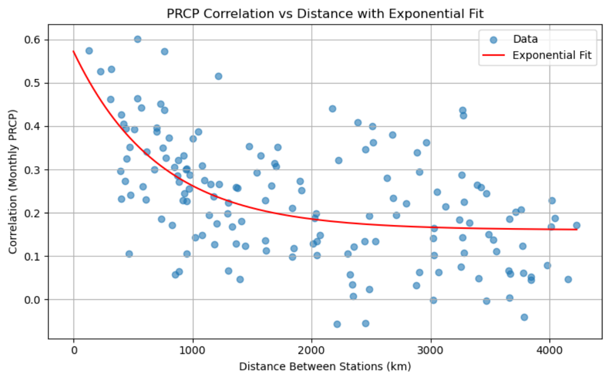

Figure 13: Monthly PRCP correlation vs distance from Arctic Bay.

- **TMAX:** Correlations remained exceptionally high, exceeding 0.95 even at distances above 2000 km. This indicates a strong regional coherence in temperature patterns, likely driven by large-scale atmospheric dynamics.
- **PRCP:** Correlations dropped below 0.3 beyond ∼1000 km, confirming precipitation's highly localized nature due to its sensitivity to mesoscale weather systems and local topography Dai et al. (1999).

These findings highlight the importance of spatial resolution in model design. While temperature data may be reliably extrapolated across large areas, accurate precipitation modeling demands dense station networks or localized predictors.

6 Conclusion

This project applied LSTM and XGBoost models to GHCN-Daily data for Arctic Bay CS to forecast temperature and precipitation at monthly and daily timescales. XGBoost performed better than LSTM in both tasks routinely, demonstrating its ability in tabular time series forecasting. The various extension tasks demonstrated the scalability of the pipeline in forecasting future trends of warming, identification of spatial extremes, teleconnection testing, and spatial correlation decay analysis.

Limitations

Several limitations impacted model performance. The use of uni variate LSTMs limited the ability to learn from multivariate interactions (e.g., temperature–precipitation coupling). The models also lacked exogenous climatic drivers such as ENSO or greenhouse gas trends, which are critical in capturing non-stationary long-term changes in climate. In addition, the Arctic Bay dataset contained gaps and sparsity (especially in snowfall), which impaired training effectiveness and generalizability.

Further Work

Future improvements could include:

- **Multivariate learning**: Addition of more predictors (like pressure, humidity, large-scale indices).
- **Transfer learning**: Utilization of regional/global data to pre-train models and fine-tune locally.
- **Hybrid models**: Blending both physical and ML-based models (e.g., DL-WRF) for physically constrained prediction.
- **Uncertainty estimation**: Application of Bayesian or other ensemble methods to estimate confidence in predictions.

To conclude, this project demonstrates the feasibility of applying traditional machine learning techniques to local-scale climate analysis, while highlighting the importance of integrating larger-scale data and physically based models in future work.

7 Appendix A

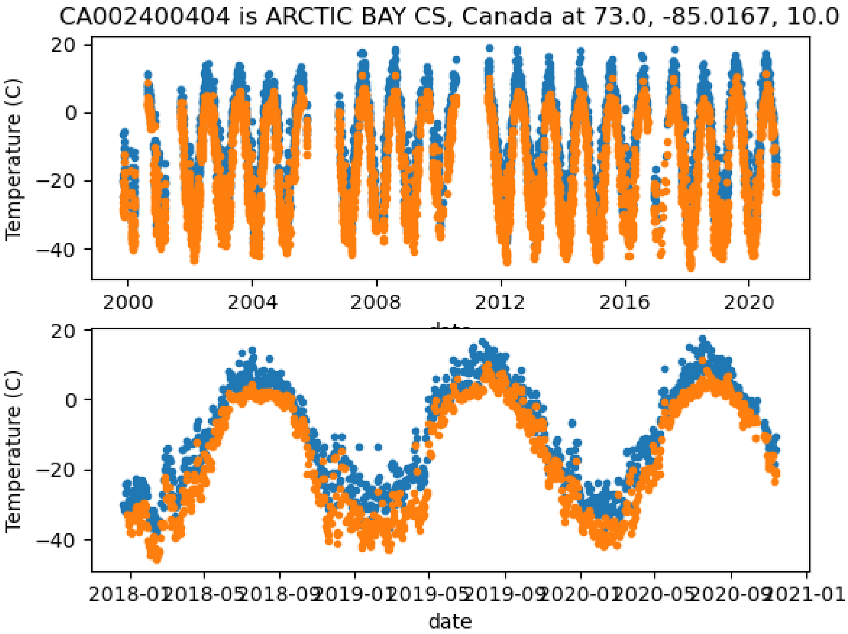

Figure 14: Daily TMAX (blue) and TMIN (orange) recorded at Arctic Bay CS from 1999 to 2021 (top), and zoomed in on recent years (bottom). Strong seasonal cycles are visible.

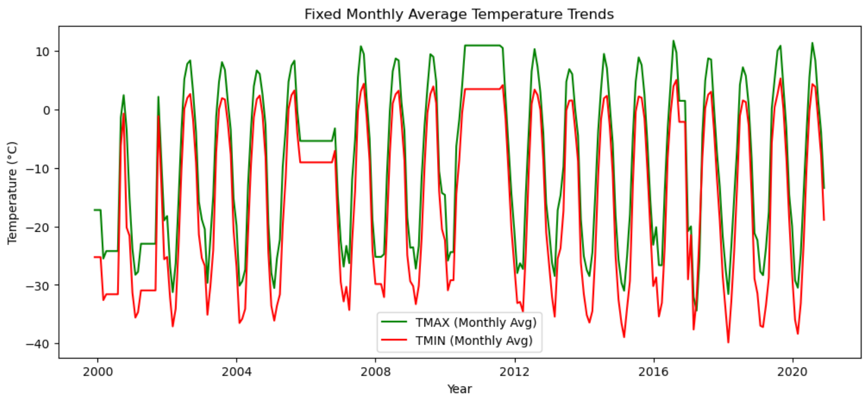

Figure 15: Monthly average TMAX and TMIN at Arctic Bay CS. Peaks in summer and troughs in winter highlight strong annual temperature variation.

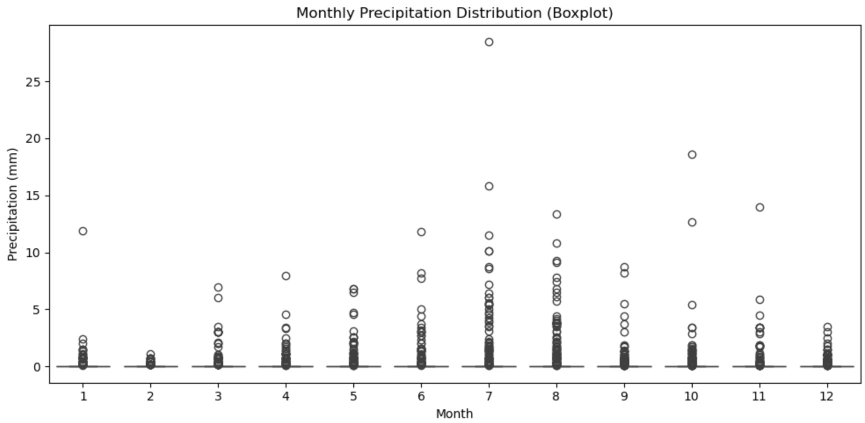

Figure 16: Monthly boxplot of daily precipitation. Most months exhibit low median values and frequent outliers, especially in summer.

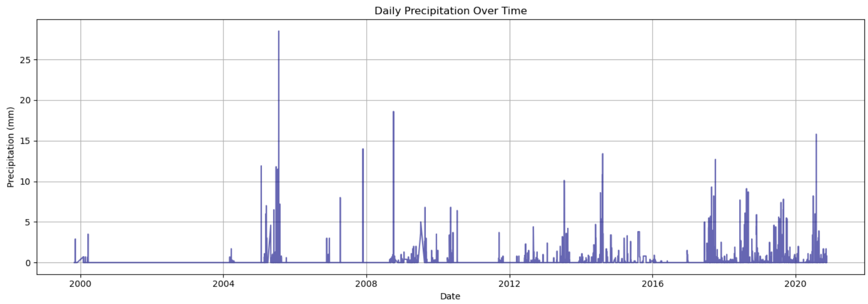

Figure 17: Daily precipitation over time at Arctic Bay CS. Precipitation events are sporadic and concentrated in clusters.

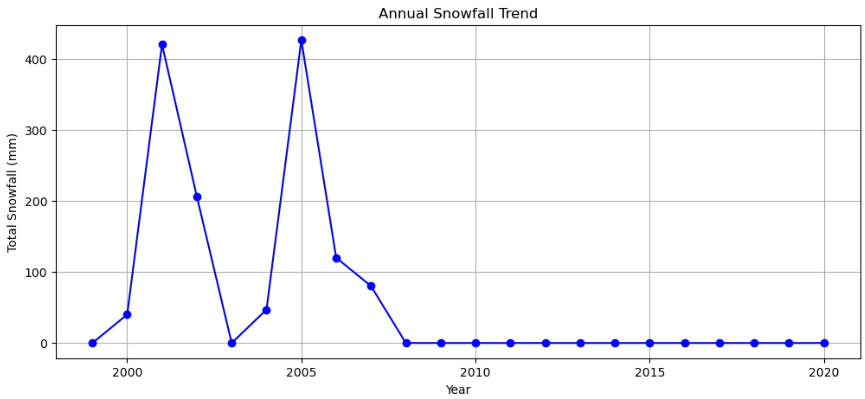

Figure 18: Annual snowfall totals at Arctic Bay CS. Data becomes zero or missing after 2007, motivating the exclusion of SNOW from modelling.

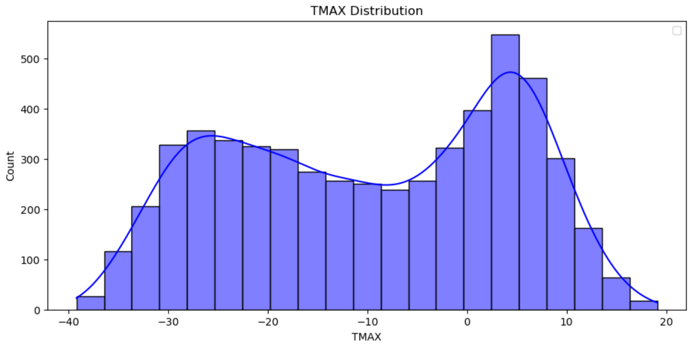

Figure 19: Histogram of daily TMAX values. The distribution is approximately bimodal, reflecting seasonal cold and warm periods.

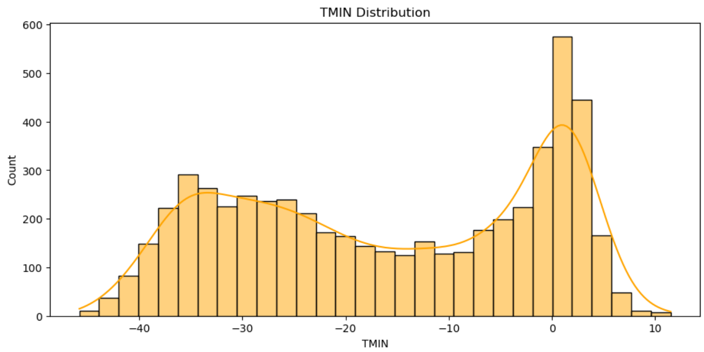

Figure 20: Histogram of daily TMIN values. A sharp peak around 0°C is visible, with a long left tail due to extremely cold days.

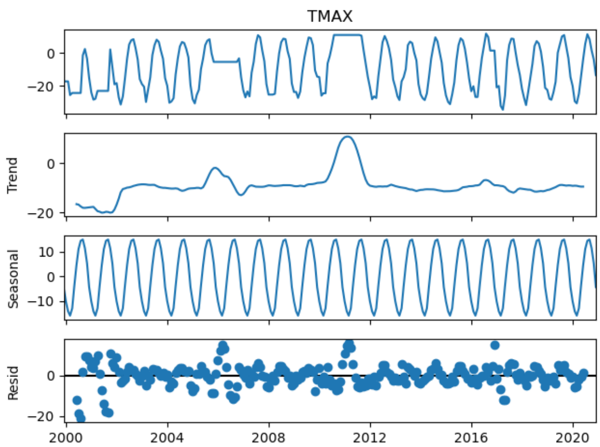

Figure 21: Seasonal decomposition of TMAX into trend, seasonal, and residual components. A clear seasonal signal is present.

References

Bauer, Peter, Alan Thorpe, and Gilbert Brunet (2015), “The quiet revolution of numerical weather prediction.” *Nature*, 525, 47–55.

Benbouzid, D., L. Nicolai, L. Siksnys, J. M. Schneider, and N. Lavayssière (2021), “Machine learning for weather and climate modelling.” *Philosophical Transactions of the Royal Society A*, 379.

Dai, Aiguo, Kevin E. Trenberth, and Thomas R. Karl (1999), “Effects of clouds, soil moisture, precipitation, and water vapor on diurnal temperature range.” *Journal of Climate*, 12, 2451–2473.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013), *An Introduction to Statistical Learning*. Springer, New York. Chapter 6: Linear Model Selection and Regularization.

Rantanen, M. et al. (2022), “The arctic has warmed nearly four times faster than the globe since 1979.” *Communications Earth Environment*, 3, 168.

Reichstein, Markus, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, and Nuno Carvalhais (2019), “Deep learning and process understanding for data-driven earth system science.” *Nature*, 566, 195–204.

Screen, J. A. and I. Simmonds (2012), “The central role of diminishing sea ice in recent arctic temperature amplification.” *Nature*, 464, 1334–1337.