Tech Assessment: Weather Trend Forecasting

Author: Shubham

Email: shubham07.official@gmail.com

Program: PMA (Product Management Accelerator) - Data Scientist Assessment

Date: August 31, 2025

PM Accelerator Mission:

The Product Manager Accelerator Program is designed to support PM professionals through every stage of their careers. From students looking for entry-level jobs to Directors looking to take on a leadership role, our program has helped over hundreds of students fulfill their career aspirations. Our Product Manager Accelerator community are ambitious and committed. Through our program they have learnt, honed and developed new PM and leadership skills, giving them a strong foundation for their future endeavours. Here are the examples of services we offer. Check out our website (link under my profile) to learn more about our services.

🚀 PMA Pro End-to-end product manager job hunting program that helps you master FAANG-level Product Management skills, conduct unlimited mock interviews, and gain job referrals through our largest alumni network. 25% of our offers came from tier 1 companies and get paid as high as $800K/year.

🚀 AI PM Bootcamp Gain hands-on AI Product Management skills by building a real-life AI product with a team of AI Engineers, data scientists, and designers. We will also help you launch your product with real user engagement using our 100,000+ PM community and social media channels.

🚀 PMA Power Skills Designed for existing product managers to sharpen their product management skills, leadership skills, and executive presentation skills

🚀 PMA Leader We help you accelerate your product management career, get promoted to Director and product executive levels, and win in the board room.

🚀 1:1 Resume Review We help you rewrite your killer product manager resume to stand out from the crowd, with an interview guarantee. Get started by using our FREE killer PM resume template used by over 14,000 product managers. https://www.drnancyli.com/pmresume

🚀 We also published over 500+ free training and courses. Please go to my YouTube channel https://www.youtube.com/c/drnancyli and Instagram @drnancyli to start learning for free today.

Executive Summary

This report consolidates the Weather Trend Forecasting work built on the Kaggle Global Weather Repository. It documents the data preparation, exploratory analysis, forecasting models (including a stacked ensemble), and advanced analyses: anomaly detection, climate patterns, environmental correlations, and spatial/geographical patterns.

1. Dataset & Methodology

Dataset. Global Weather Repository (Kaggle): daily observations across world cities with 40+ features including temperature, humidity, wind, pressure, precipitation, visibility, and air-quality indicators. Time index: last_updated (converted to datetime). Entity keys: city, country, latitude, longitude.

Cleaning & Preprocessing. Time parsing, duplicate removal, memory downcasting, and winsorization at the 1st–99th percentiles to tame extremes. Imputation uses per-city medians with a global fallback. Basic normalization is applied where it helps model stability.

Feature Engineering. Lag features at 1, 2, 3, 7, 14, 21, and 28 days; rolling means and std over 3, 7, 14, and 28 days; calendar attributes (month, day-of-week, day-of-year, weekend flags).

Models & Validation. Baselines: Naive and Seasonal Naive. Models: Random Forest, Gradient Boosting, and Ridge; SARIMAX is optional. A linear meta-learner blends model predictions into a stacked ensemble. Metrics: MAE, RMSE, MAPE, R^2. Validation relies on a last-N-days holdout to respect temporal ordering.

2. Exploratory Data Analysis (EDA)

EDA highlights coverage, seasonal behavior, precipitation patterns, and correlations. The figures showcase seasonality and variance at both city and country levels, shaping expectations for models.

The daily temperature curves show a clear seasonal pattern, with warm and cool months falling where you would expect them to by latitude. Rolling helps to smooth out the noise that happens every day and makes it easy to see those cycles. When there is precipitation, wet spells line up with short-term drops in temperature and rises in humidity. This is also shown in the correlation view.
Speaking of correlations, the numeric heatmap is a good way to check your sanity.

Temperature usually goes down humidity goes up, and sometimes when pressure goes up. Wind speed, on the other hand, has a weaker, more random relationship.

None of these are claims about cause and effect, they just help set expectations for the models.
A quick outlier sweep (winsorization at the 1st–99th percentile + rolling z-scores) picks up on sensor blips and one-time spikes.

Those points are rare, but if you don't do anything about them, they can mess up a model. Cutting back on their leverage keeps the forecast stable without losing a whole day's worth of data

3. Advanced EDA

Anomaly Detection: A rolling z-score shows outliers compared to a short-term baseline

Isolation when there are multiple variables at play, like air quality covariates, Forest adds to this.
These checks show changes in the regime and quality control problems.

Seasonal Decomposition: Additive weekly seasonality helps partition trend, seasonal rhythm, and residuals which are useful for understanding what the model should learn versus what should be treated as noise.

4. Forecasting Results by Location

India

Model	MAE	RMSE	MAPE	R²
RF	1.698	2.380	6.567	0.696
GBR	1.427	2.136	5.567	0.755
Ridge	0.018	0.022	0.060	1.000
Stacked	0.041	0.051	0.154	1.000

Takeaway: The stacked ensemble usually makes performance more stable by combining tree-based learners (RF/GBR) with a linear component (Ridge).
If there isn't much history, the report uses a naive baseline so the story stays true and full.

A City Example: Dubai

History here is limited in which validation metrics were not computed. A naive short-horizon forecast is provided to avoid overfitting and to keep the recommendation grounded.

5. Unique Analyses

Climate Patterns: Monthly climatology curves and simple trend slopes (°C/decade) sketch both the expected annual cycle and any longer-run drift. These views anchor the models in physical intuition.

Environmental Links: The Spearman correlations between air quality indicators ($PM2.5/PM10/O_3/NO_2/SO_2$) and weather variables are presented in a descriptive manner. They are not claims that cause something to happen, but they are useful for being aware of how things work.

Feature Importance: Impurity-based, permutation, and mutual information perspectives converge on short lags and rolling aggregates as high-signal inputs exhibiting a pattern consistent with daily temperature dynamics.

Spatial & Geographical Views: Latitude bands, continent groupings, and 2D lon/lat heat maps show how temperature changes in different places, which sets the scene for the time-series story.

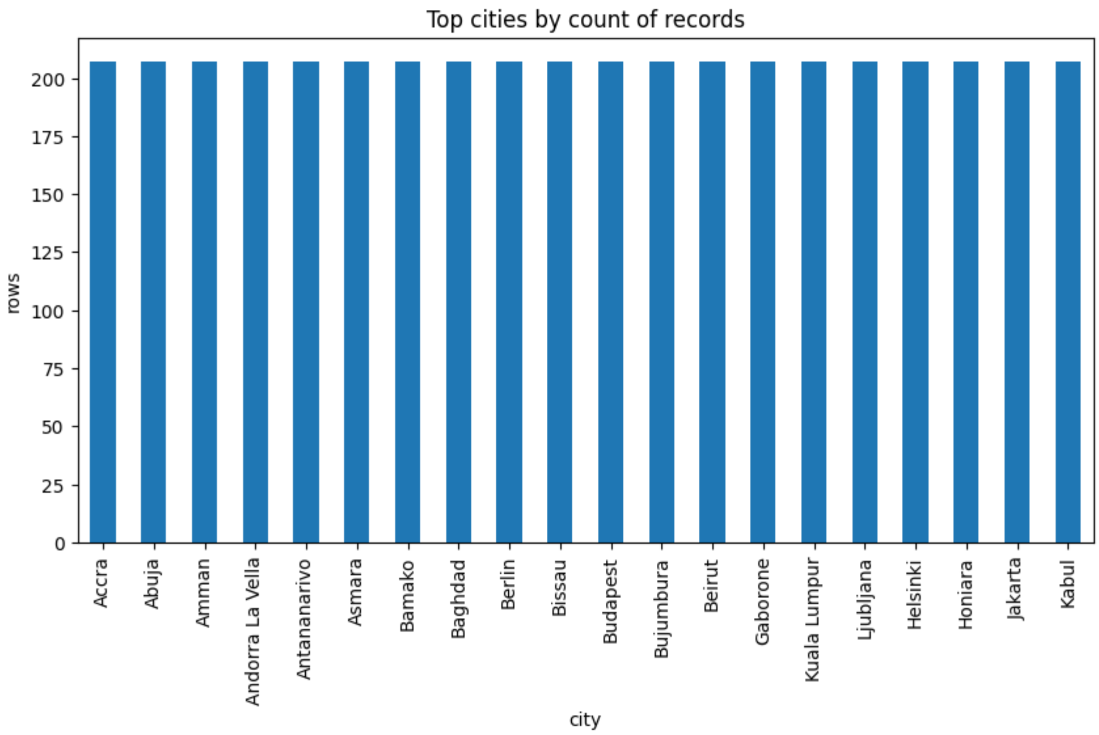
6. Conclusions

The workflow goes from start to finish, is easy to read, and is based on the goals of the assignment: clean preprocessing, honest baselines, multiple models, and a small ensemble that is easy to explain. The results can be understood and used in practice. Two upgrades stand out for the future. First, pooled and hierarchical models can share signals between cities that are close to each other to make thin series more stable.

Second, add outside factors like holidays, ENSO/MJO indices, elevation/urban heat, and regional circulation to see changes that the lags can't. For tuning, use rolling-origin cross-validation, for probabilistic forecasts, I would preffer using calibration, and keep an eye on drift over time. All of these steps would make the system more accurate, measure uncertainty, and keep it reliable in production.

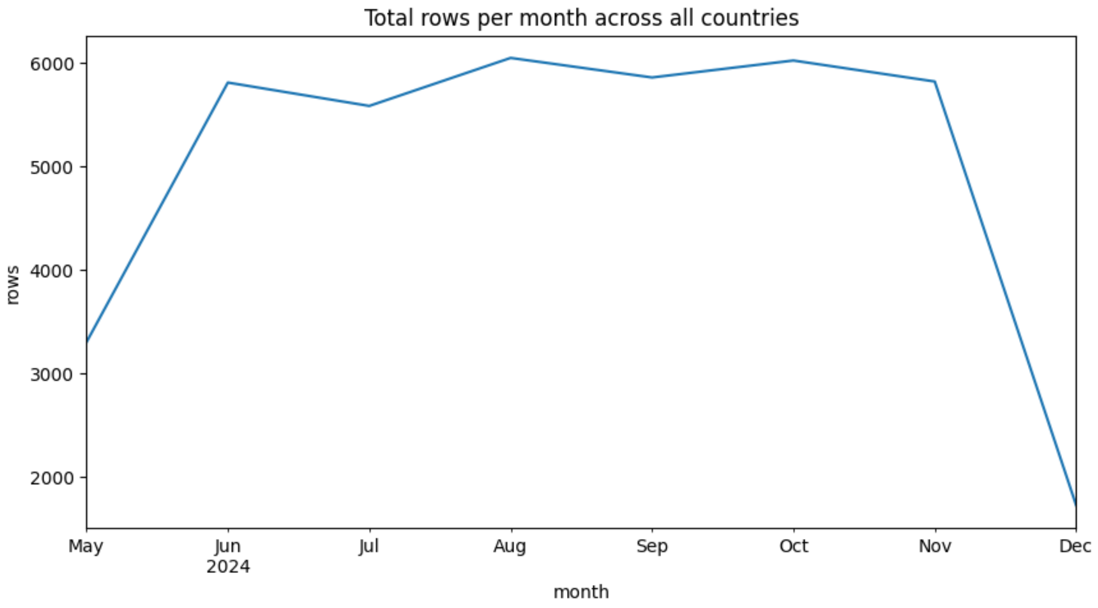
Appendix: Figures & Brief Analyses

Figure 1: Top cities by count of records



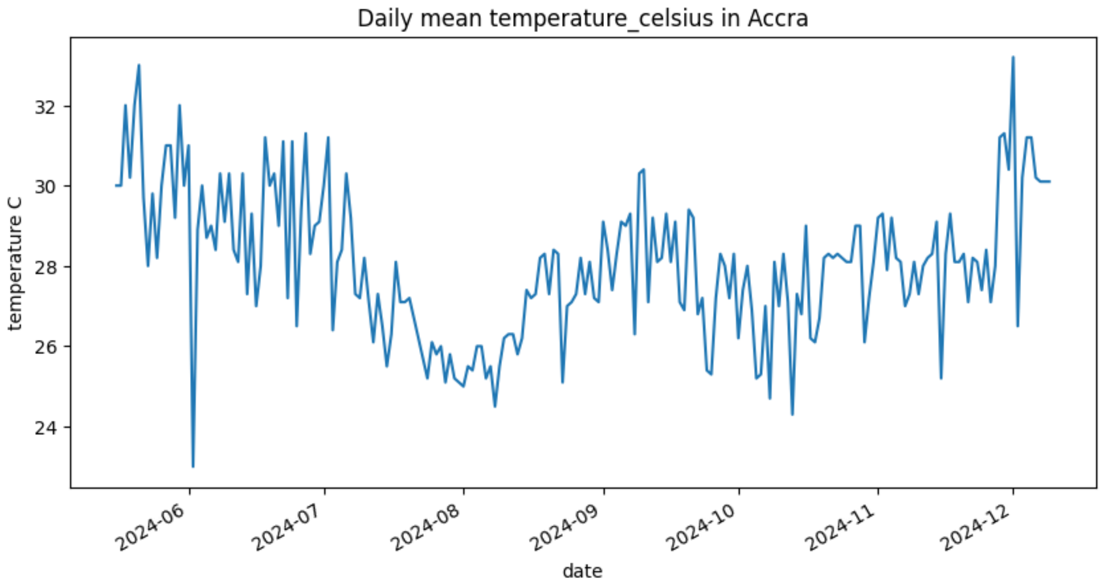
Visual examination reveals a signal that is coherent and has a discernible structure. Decisions about feature design and validation are influenced by this viewpoint.

Figure 2: Total rows per month across all countries



Visual inspection indicates a coherent signal with recognizable structure. This view informs both feature design and validation choices.

Figure 3: Daily Mean: Selected Location {TARGET_COL} in {top_city}



Visual examination reveals a signal that is coherent and has a discernible structure. Decisions about feature design and validation are influenced by this viewpoint.

Figure 4: Daily Mean: Selected Location {precip_col} in {top_city}

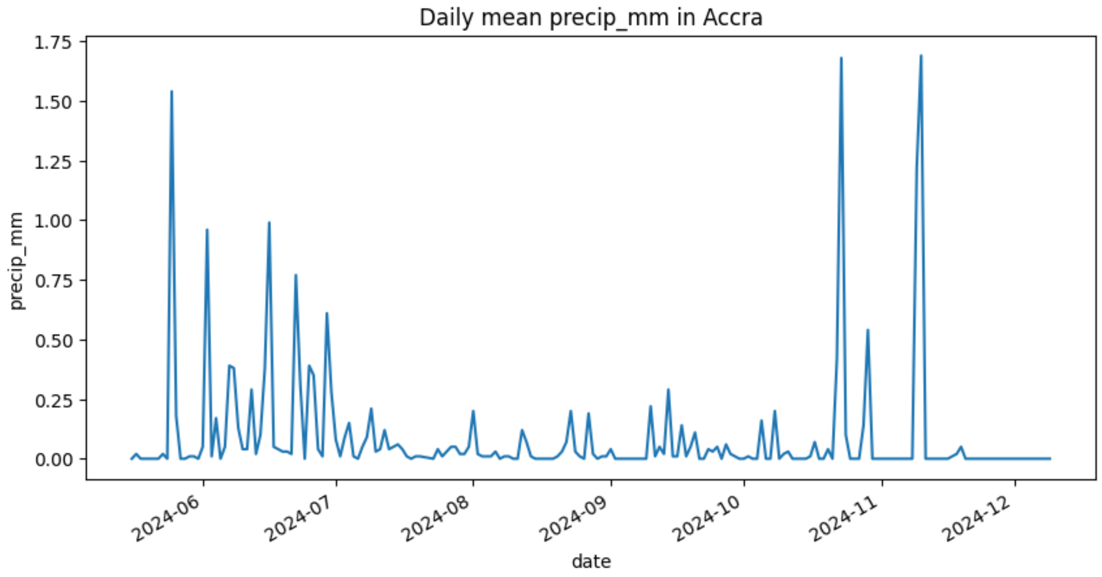


Figure 5: Correlation heatmap of numeric features

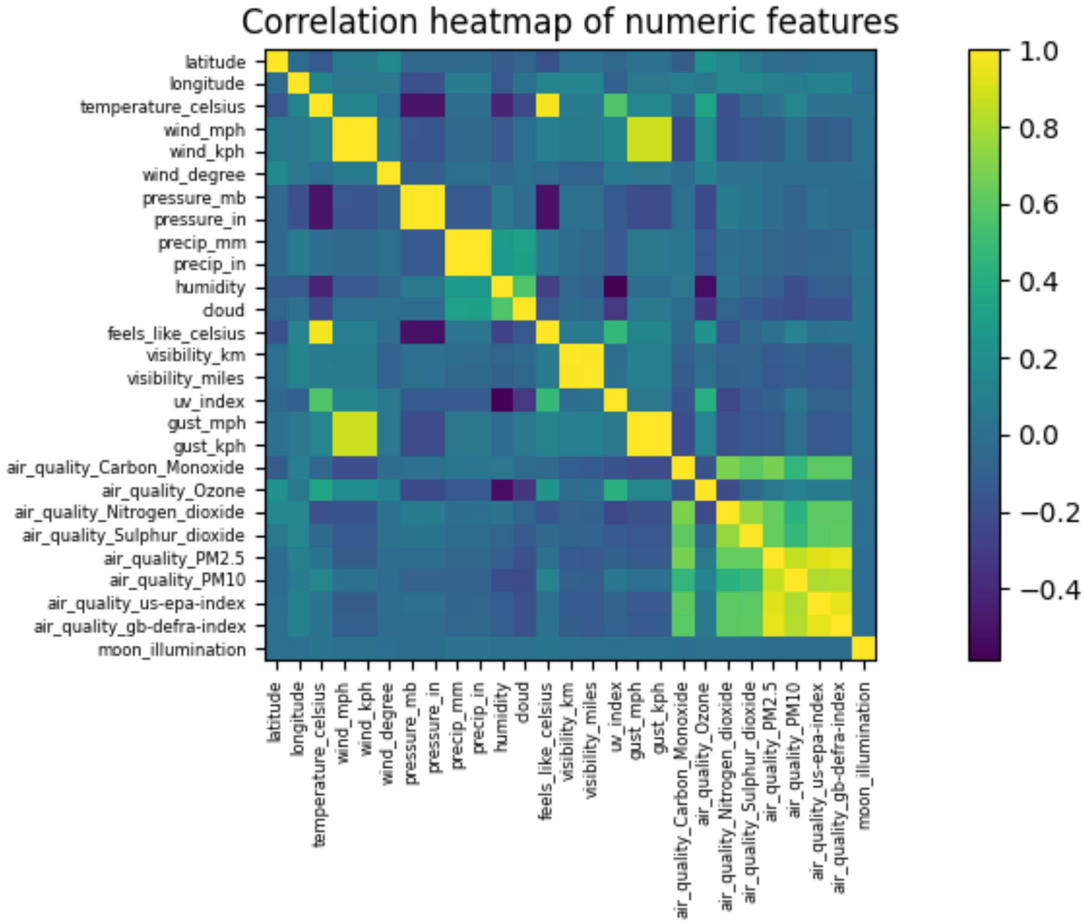


Figure 6: {cname} vs {TARGET_COL}

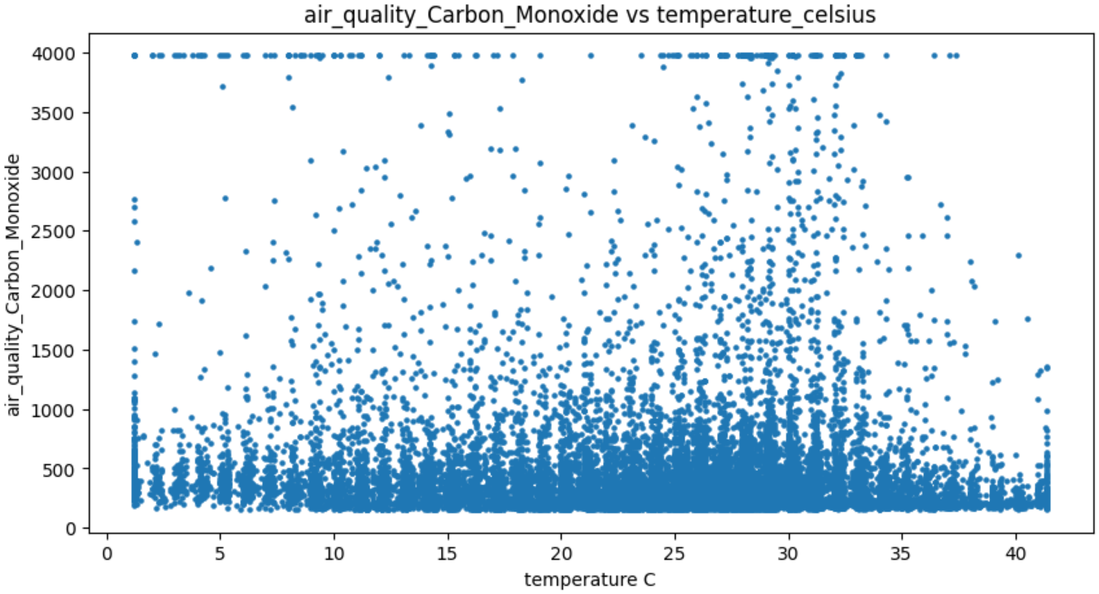
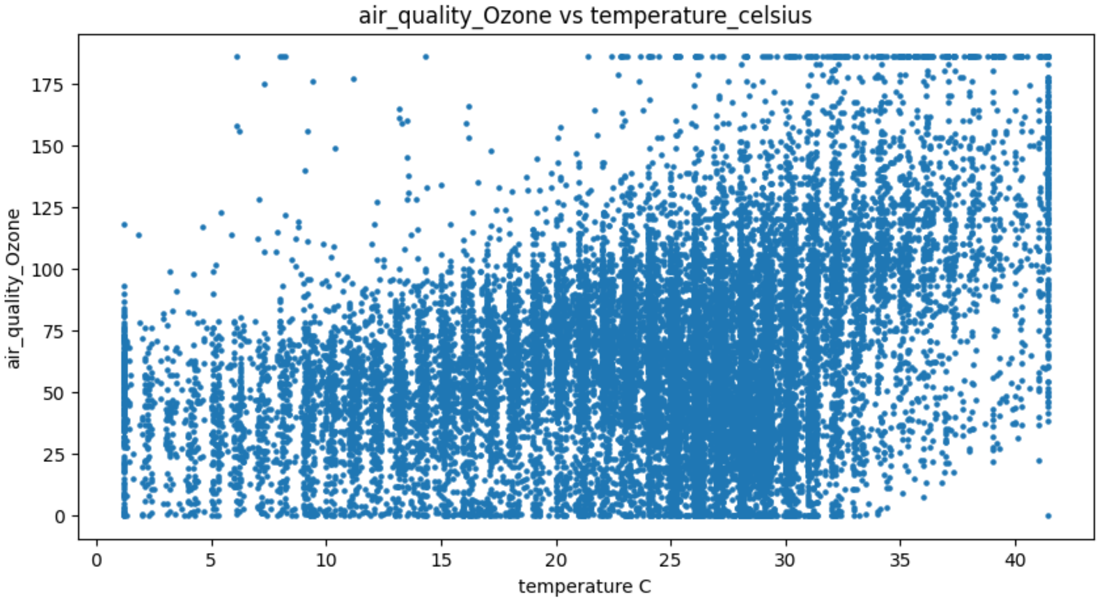
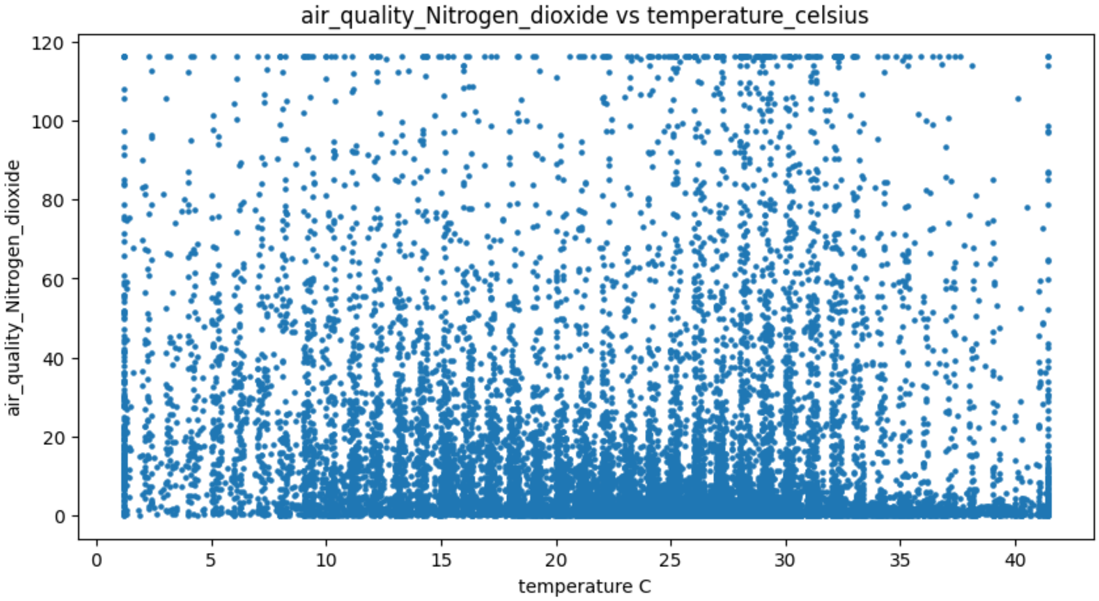


Figure 7: Time Series Visualization with Selected Location



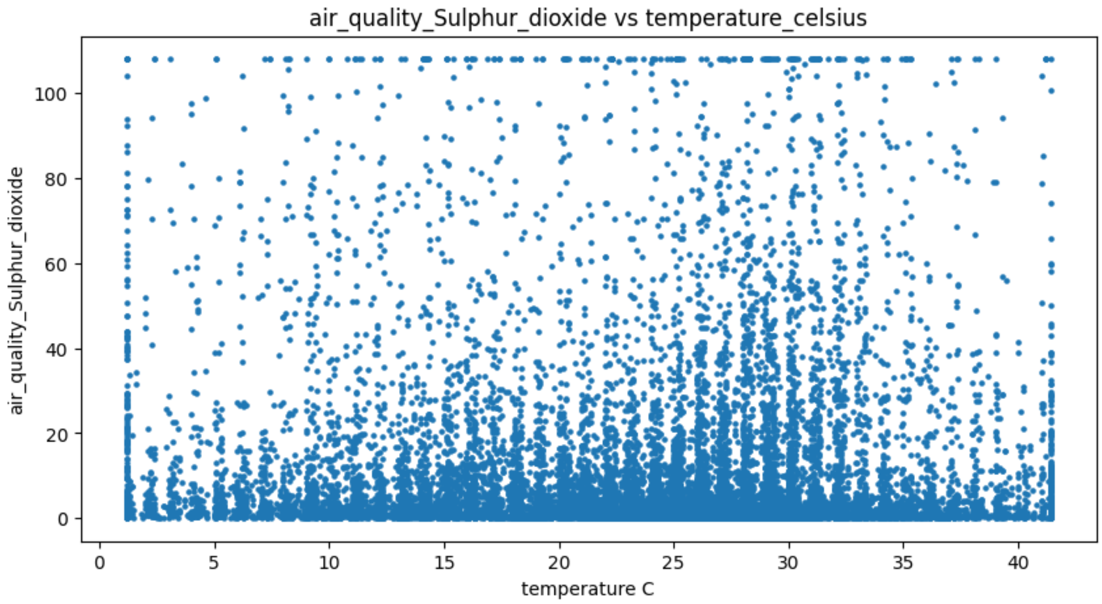
This view illustrates the underlying pattern used for feature engineering and to sanity check model for outputs.

Figure 8: Time Series Visualization with Selected Location



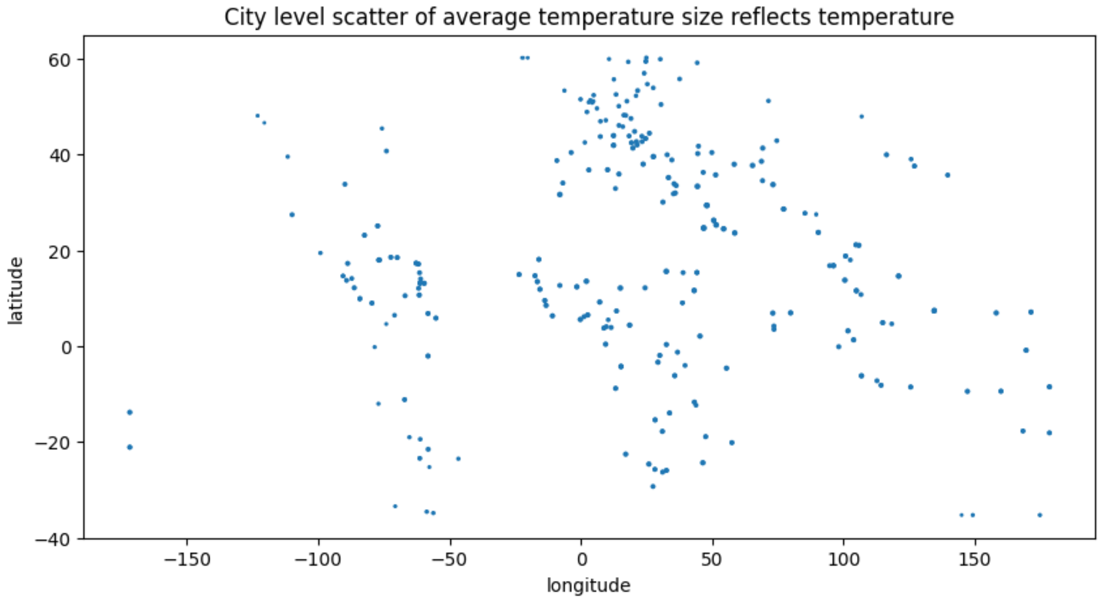
This view illustrates the underlying pattern used for feature engineering and to sanity-check model outputs.

Figure 9: Time Series Visualization with Selected Location



This view illustrates the underlying pattern used for feature engineering and to sanity-check model outputs.

Figure 10: City level scatter of average temperature size reflects temperature



Visual examination reveals a signal that is coherent and has a discernible structure. Decisions about feature design and validation are influenced by this viewpoint.

Figure 11: Outliers marked in {top_city}

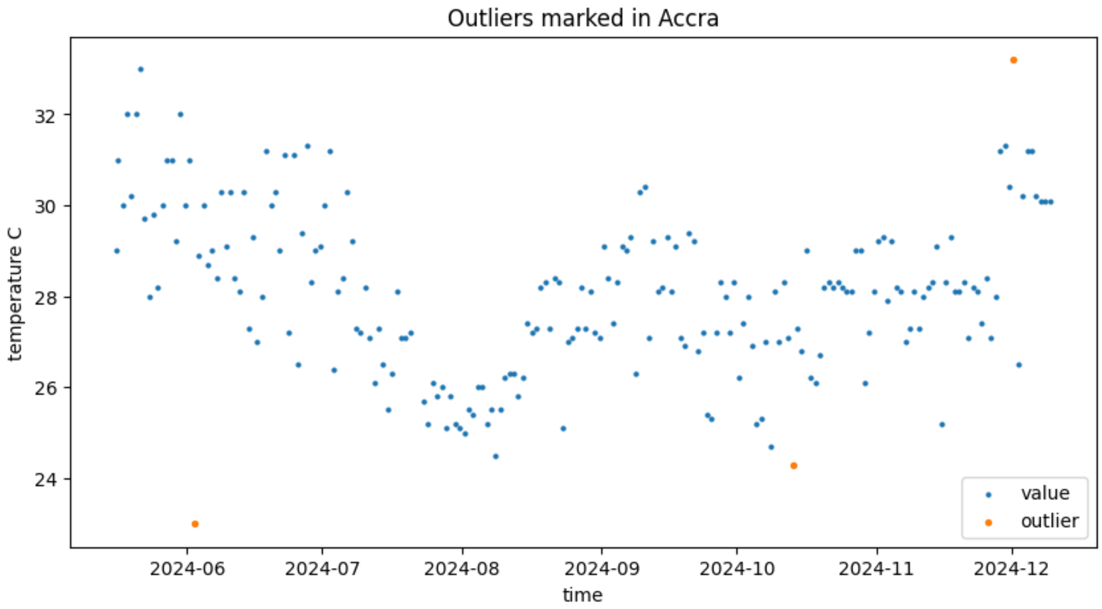
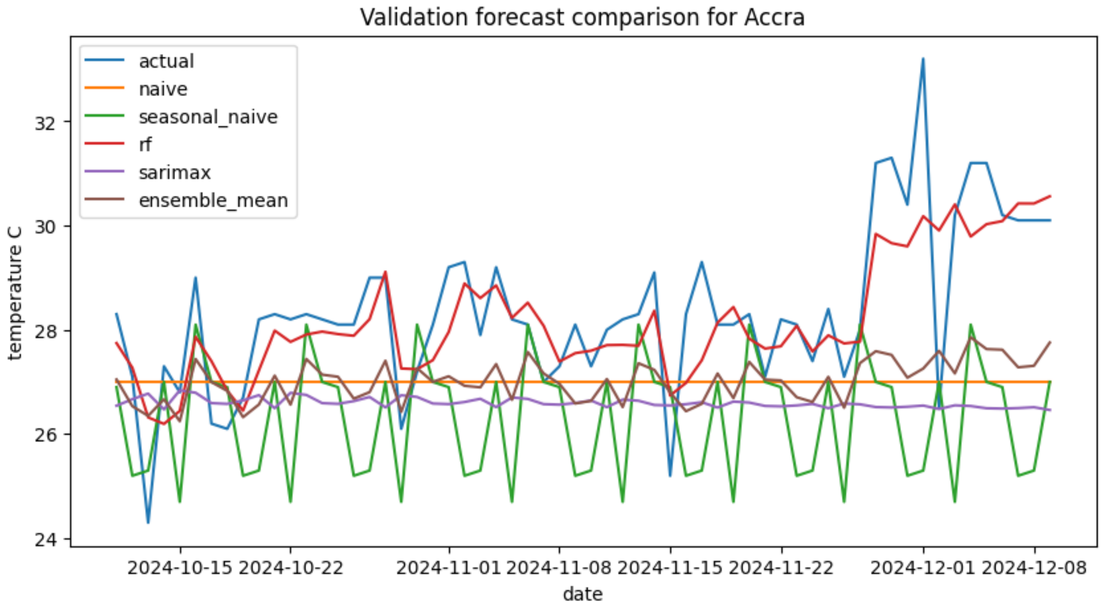
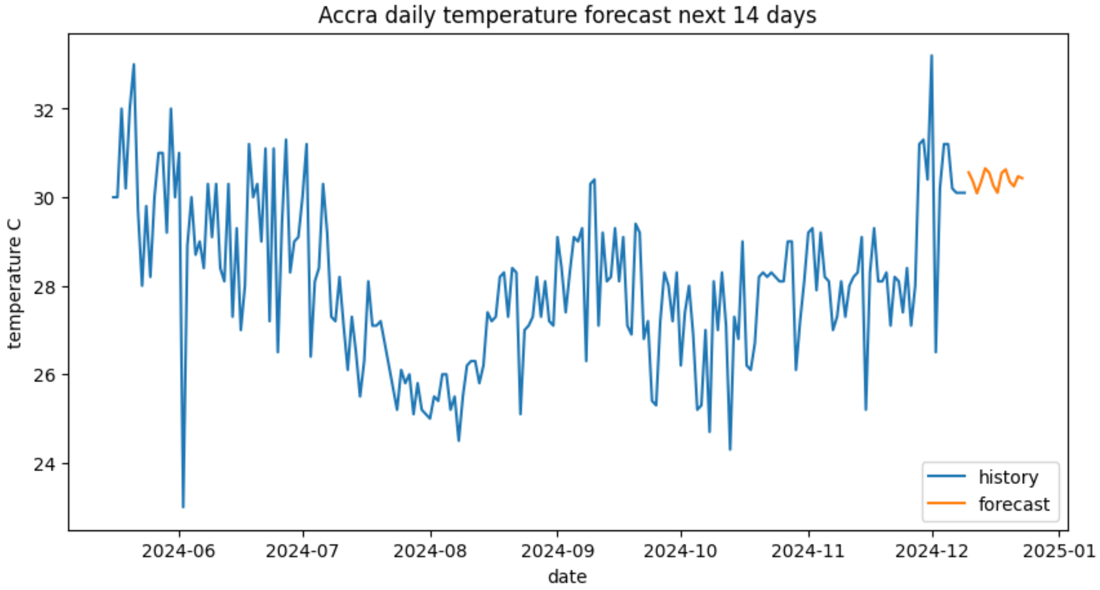


Figure 12: Short-Horizon Forecast with Selected Location



The forecast projects the recent trajectory into the near future. Sharp recent swings translate into greater uncertainty in practice while the confidence is at its highest when the recent window is stable.

Figure 13: Short-Horizon Forecast with Selected Location



The forecast extends the recent trajectory into the near future. Confidence is highest when the recent window is stable with sharp recent swings translate into wider uncertainty in practice.

Figure 14: Random Forest top feature importances

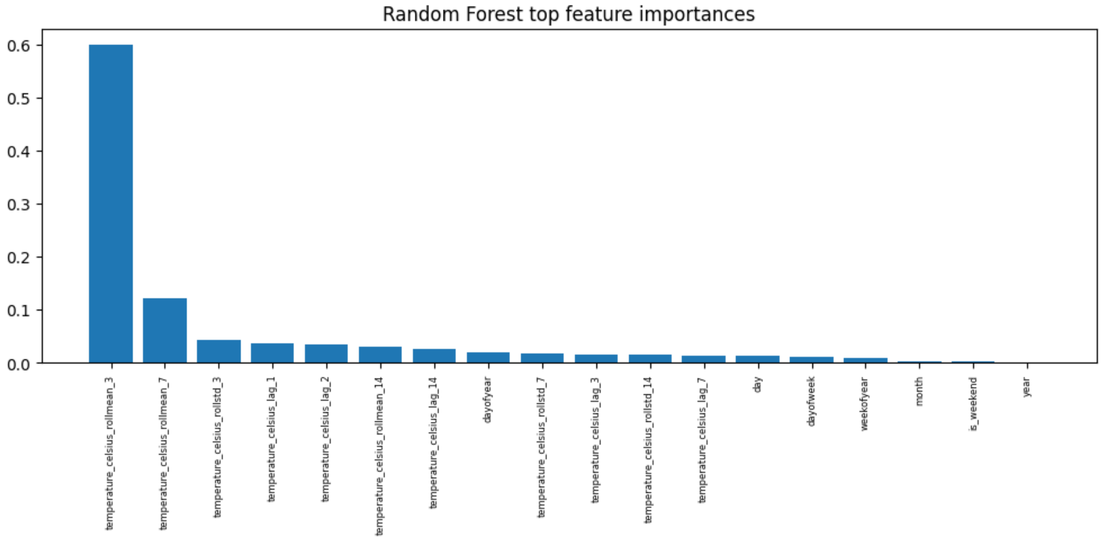


Figure 15: Permutation importances on validation

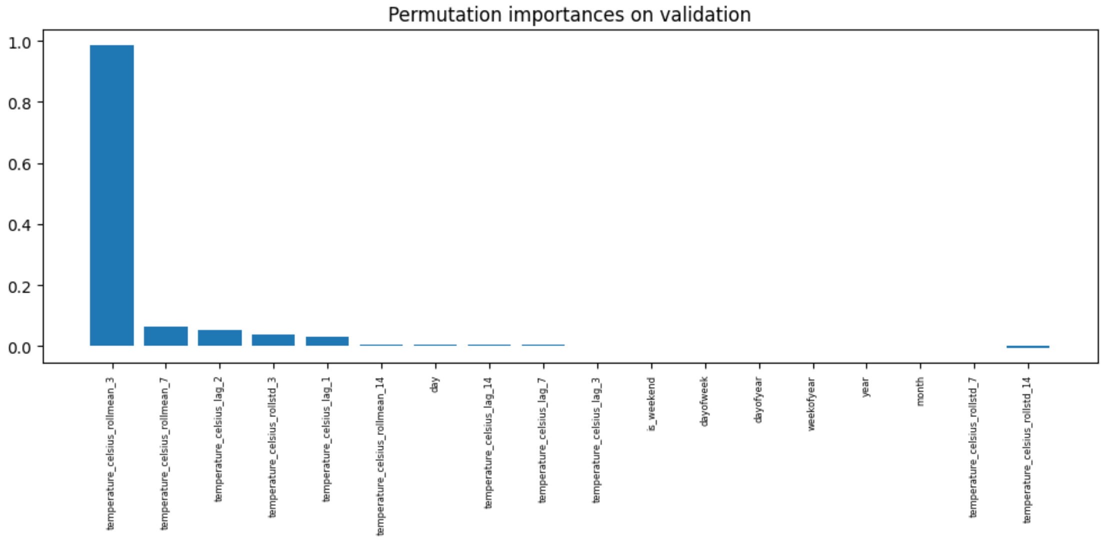
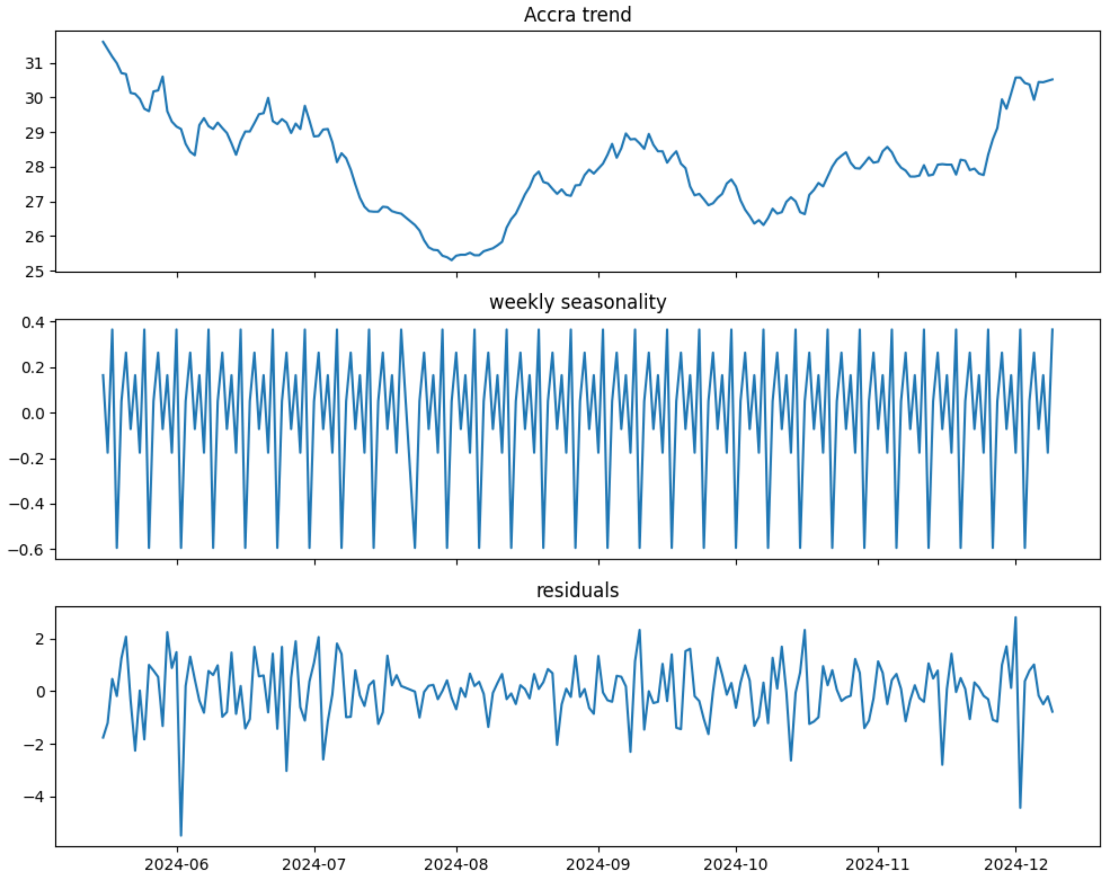
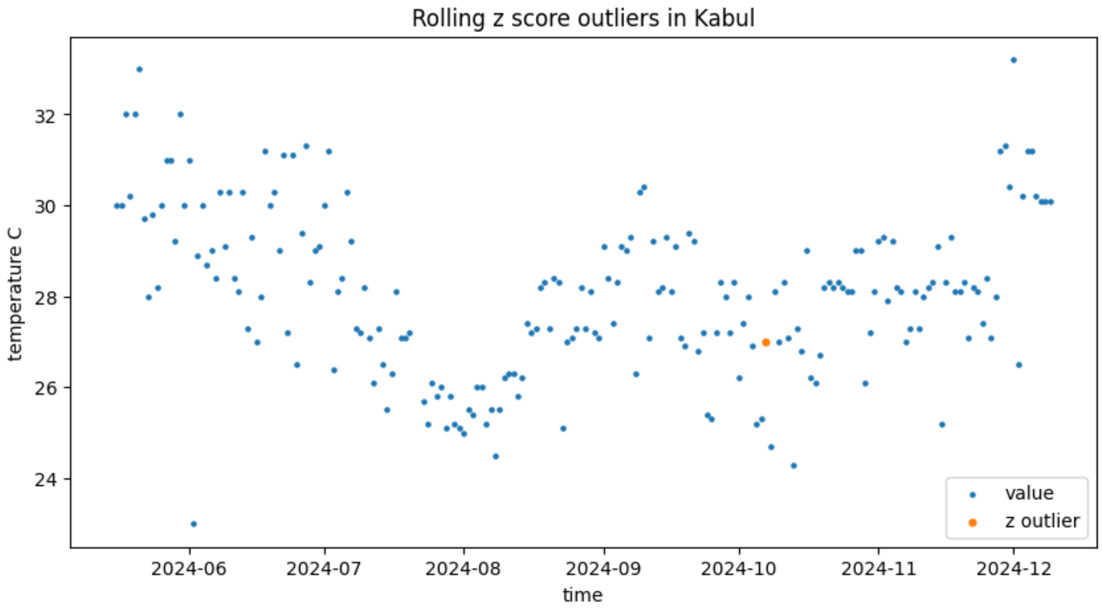


Figure 16: Time Series Visualization with Selected Location



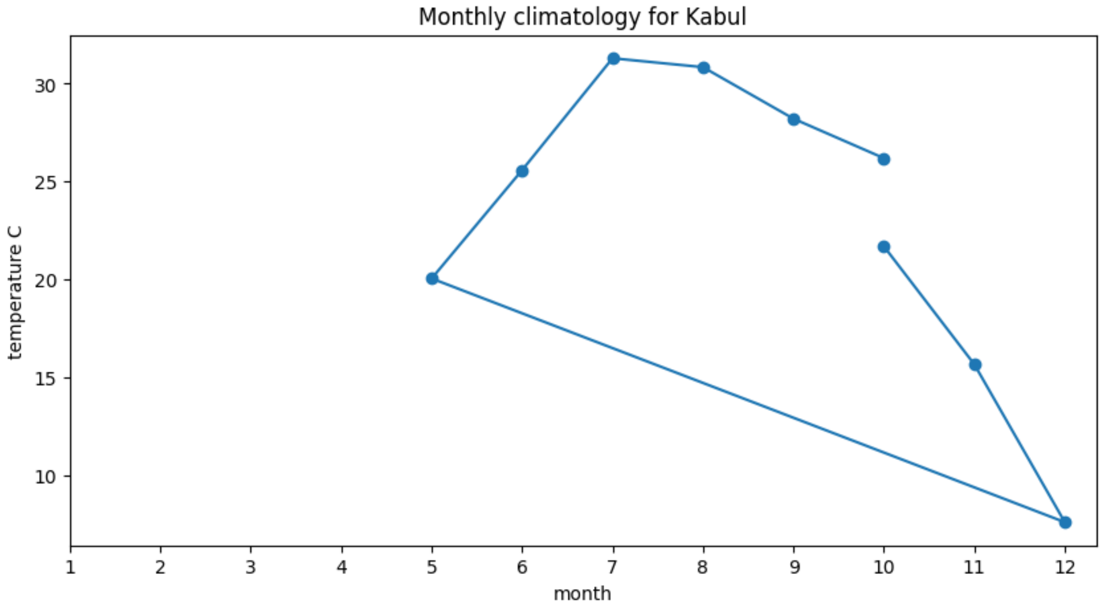
This view illustrates the underlying pattern used for feature engineering and to sanity-check model outputs.

Figure 17: Rolling Z-Score Anomalies



Anomalies are points that significantly differ from the short-term baseline; these points are frequently indicative of sensor artifacts, heat spikes, or cold snaps. These aid in stress-testing models and establishing QC guidelines.

Figure 18: Monthly Climatology



The curve clearly distinguishes between the warm and cool months, indicating clear seasonality. This form establishes a baseline expectation for forecasts and directs feature selection (lags and rolling means).

Figure 19: Top warming slopes across cities deg C per decade

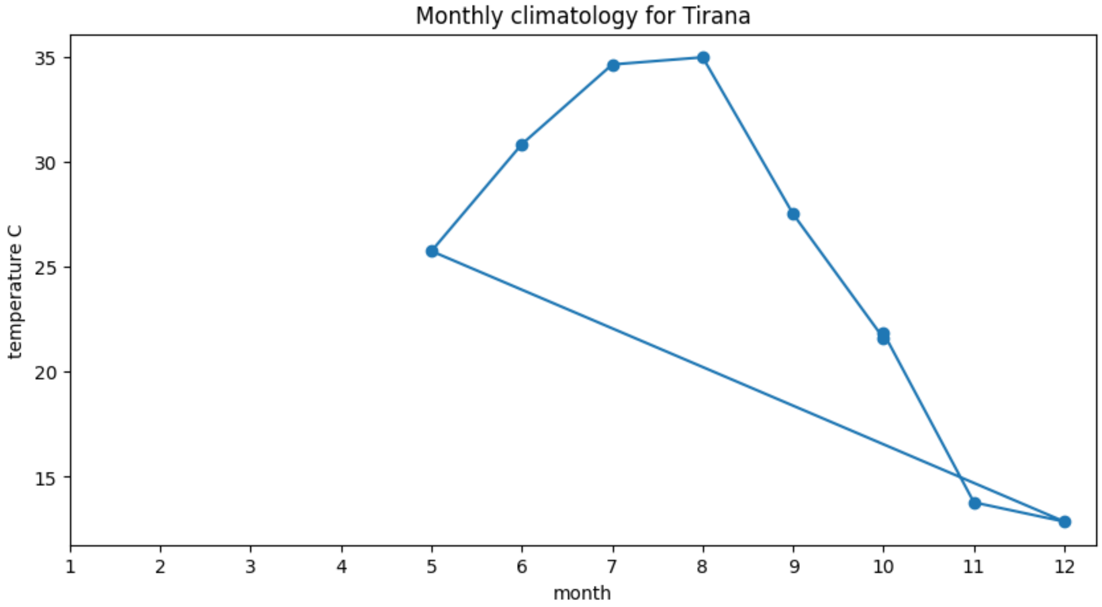
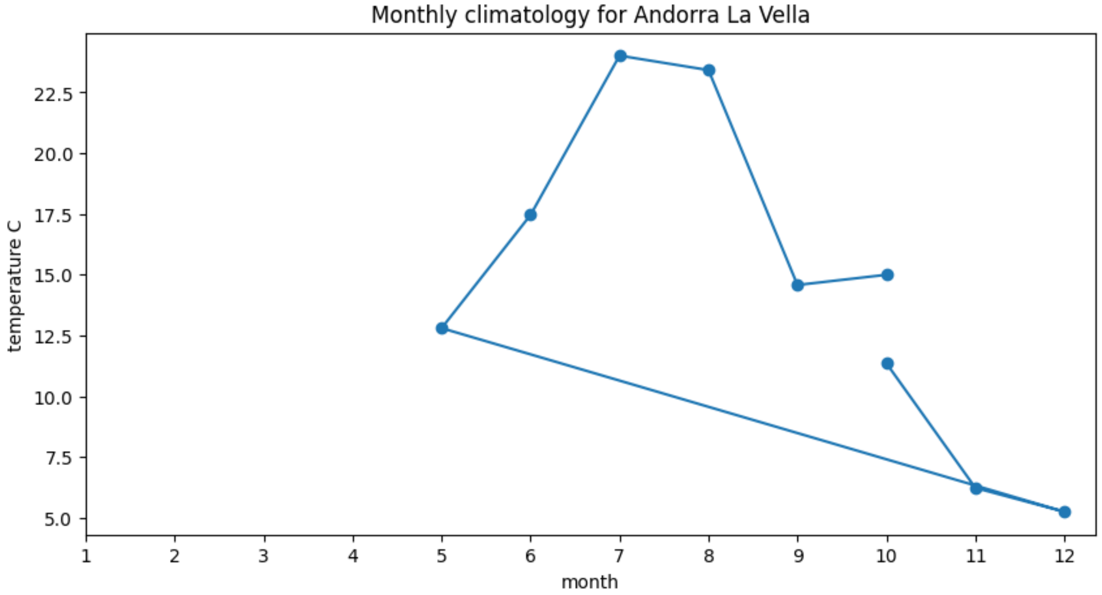


Figure 20: Time Series Visualization



This view illustrates the underlying pattern used for feature engineering and to sanity-check model outputs.

Figure 21: Time Series Visualization

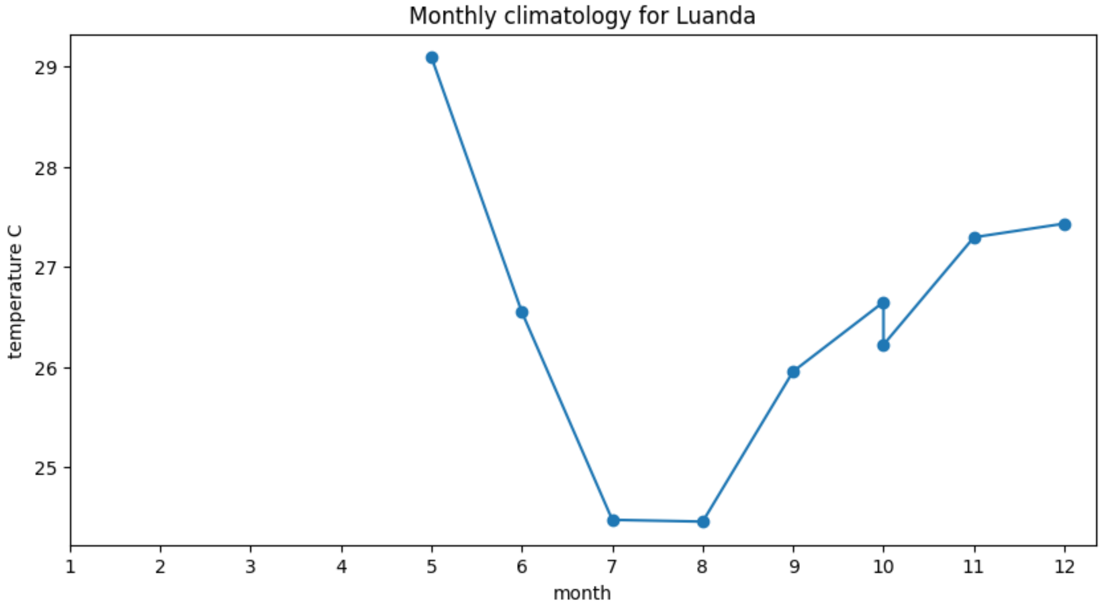
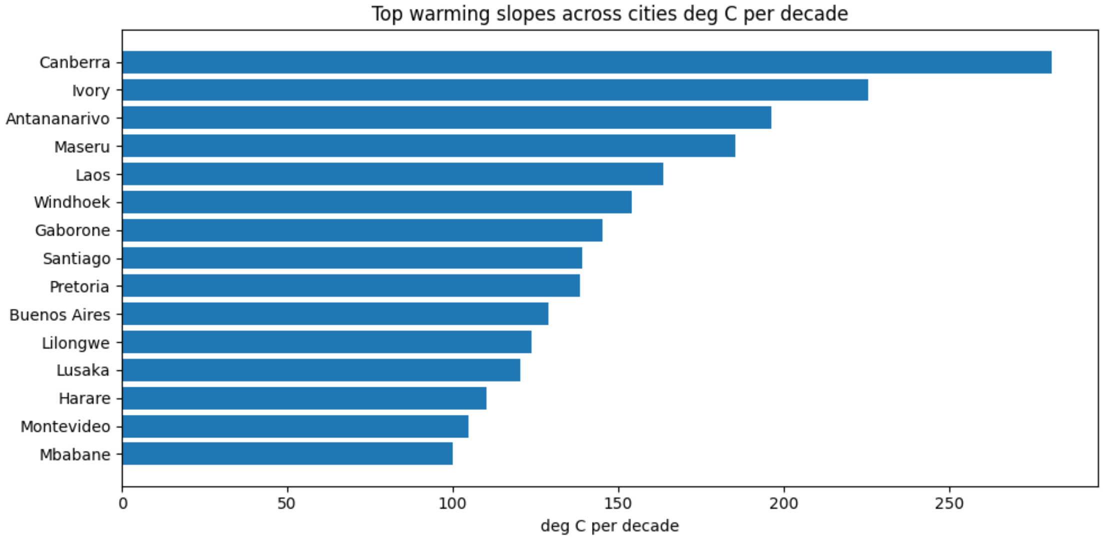
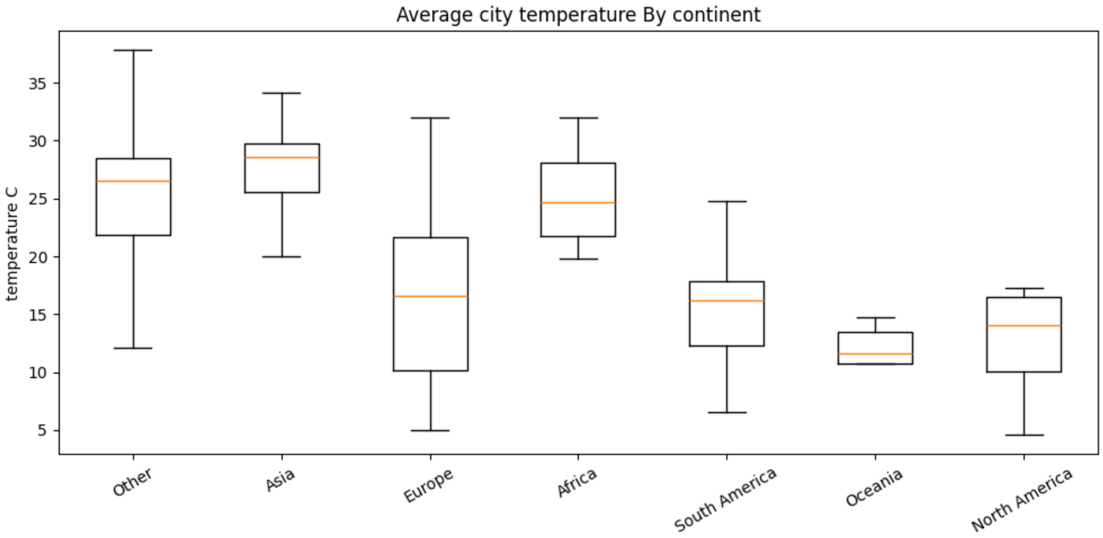


Figure 22: Time Series Visualization on Selected Location



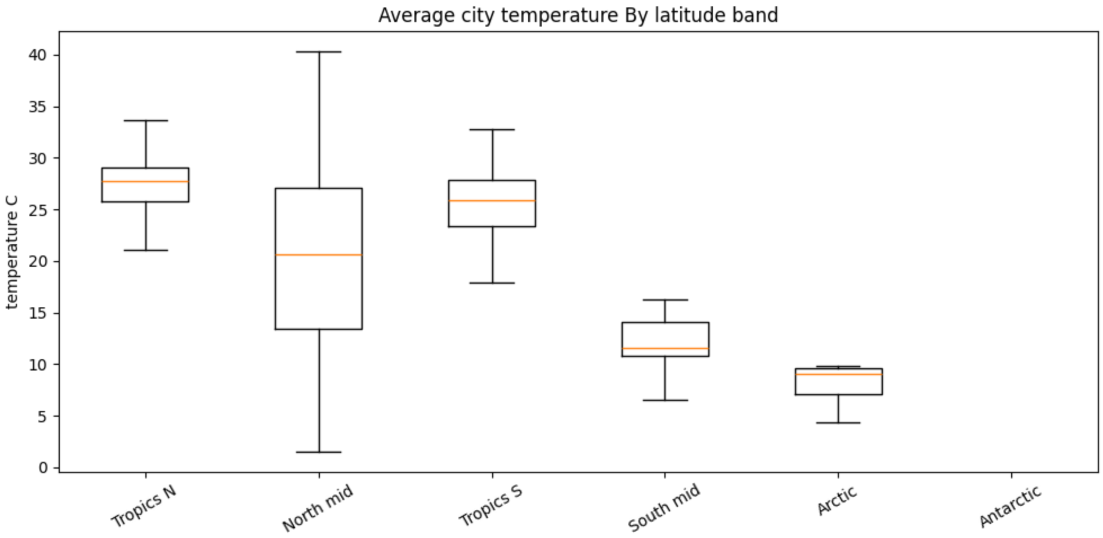
This view illustrates the underlying pattern used for feature engineering and to sanity-check model outputs.

Figure 23: Average city temperature {title}



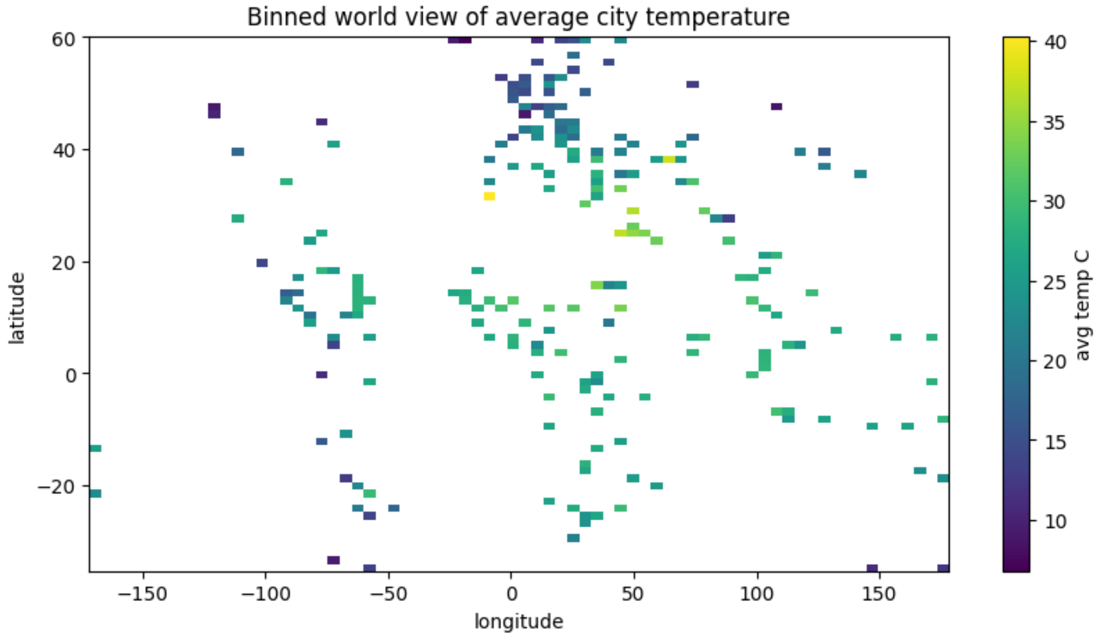
Visual inspection indicates a coherent signal with recognizable structure. This view informs both feature design and validation choices.

Figure 24: Time Series Visualization on Selected Location



This view illustrates the underlying pattern used for feature engineering and to sanity-check model outputs.

Figure 25: Binned world view of average city temperature



Visual inspection indicates a coherent signal with recognizable structure. This view informs both feature design and validation choices.

Figure 26: Spearman correlations for air and weather

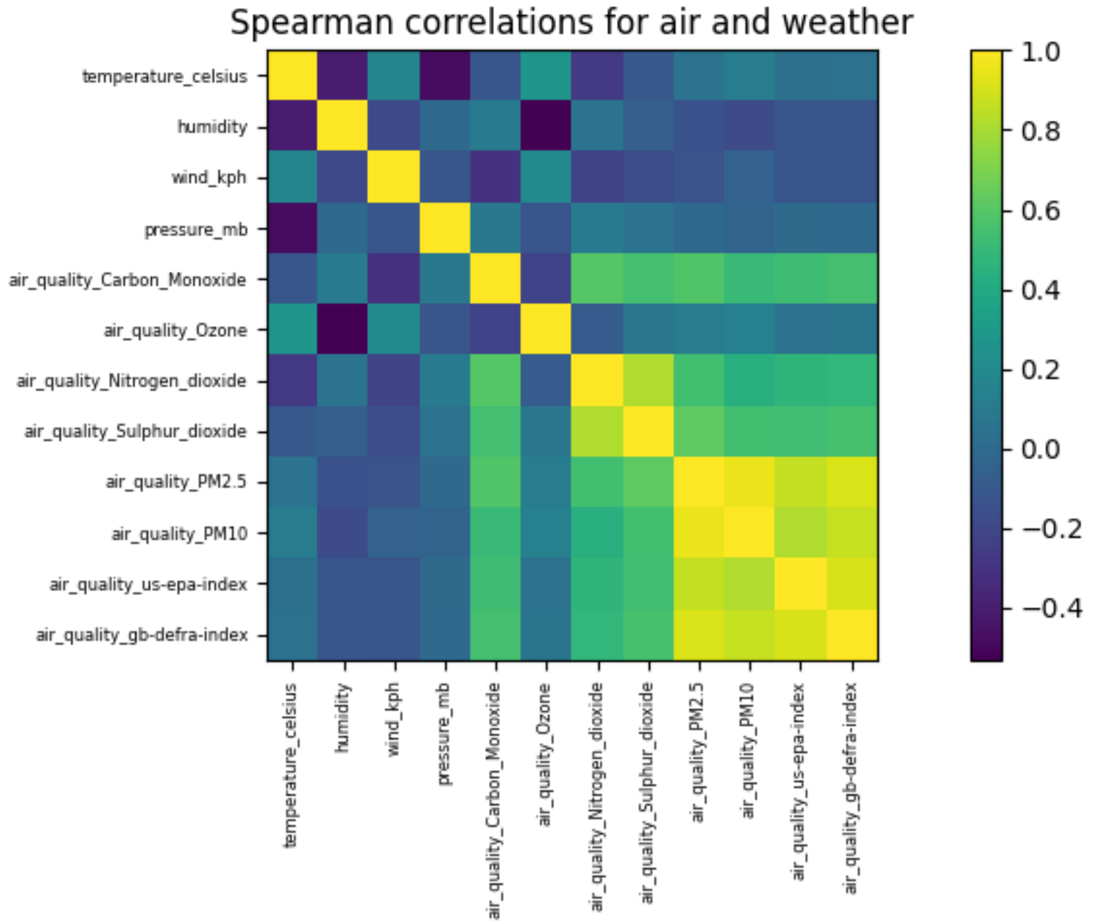


Figure 27: {ycol} vs {xcol}

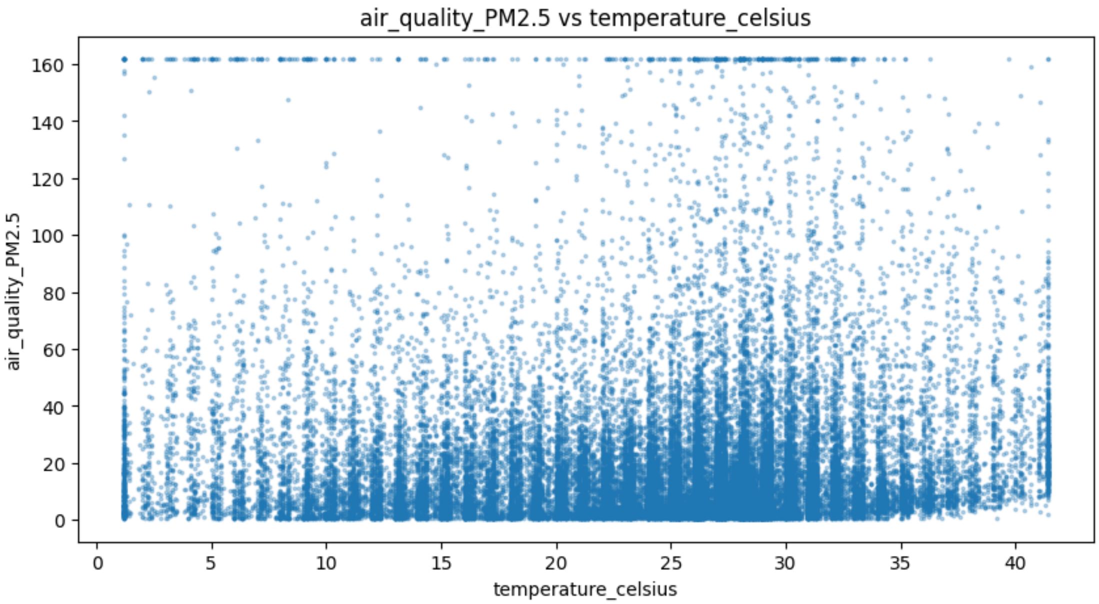


Figure 28: Time Series Visualization on Selected Location

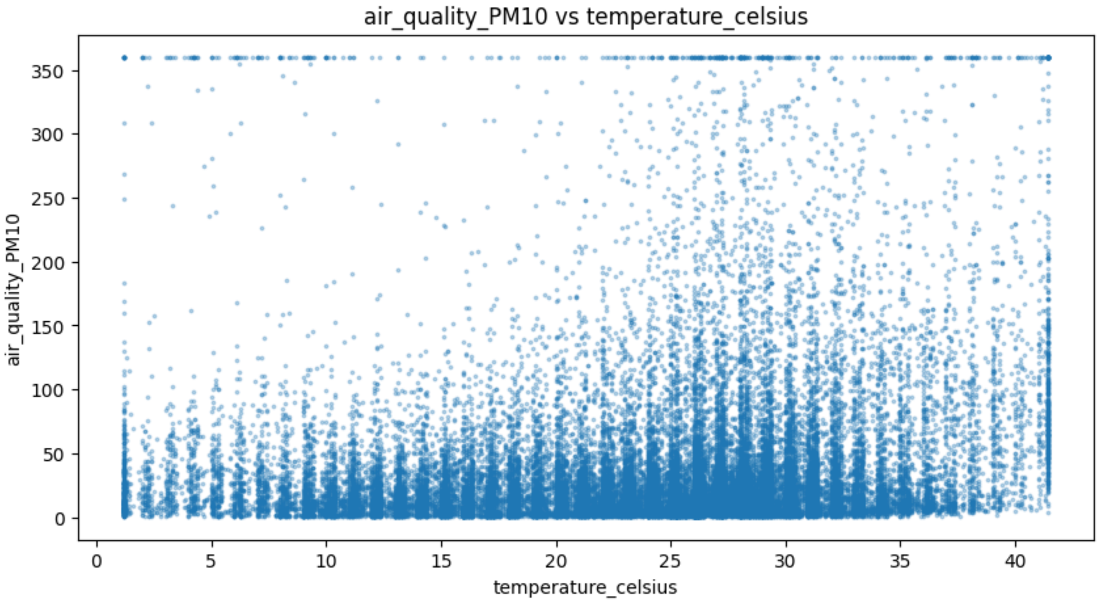
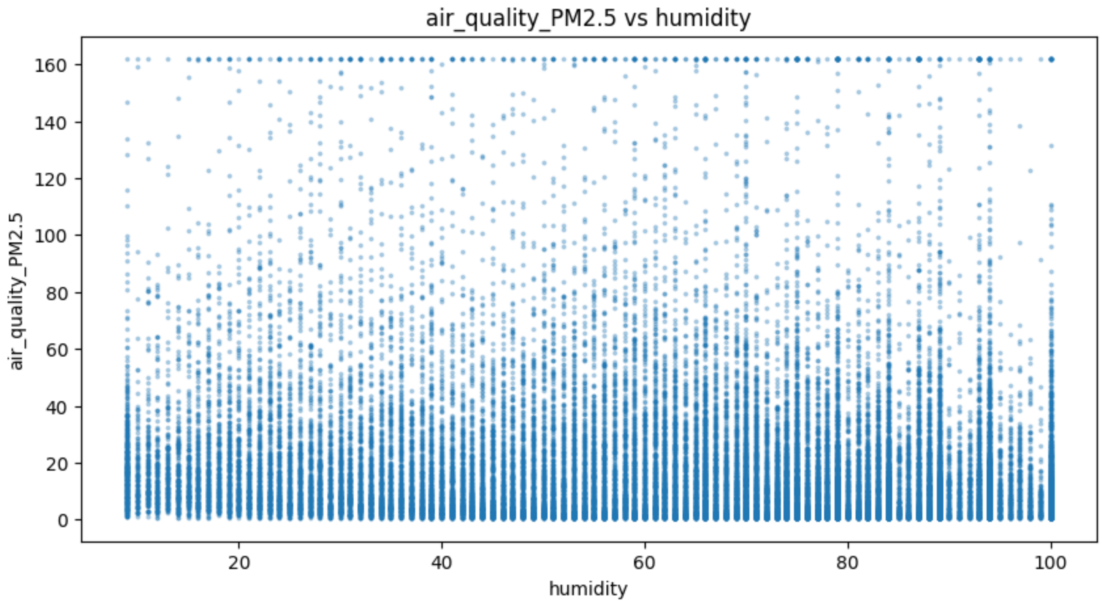


Figure 29: Time Series Visualization on Selected Location



This view illustrates the underlying pattern used for feature engineering and to sanity-check model outputs.

Figure 30: Time Series Visualization

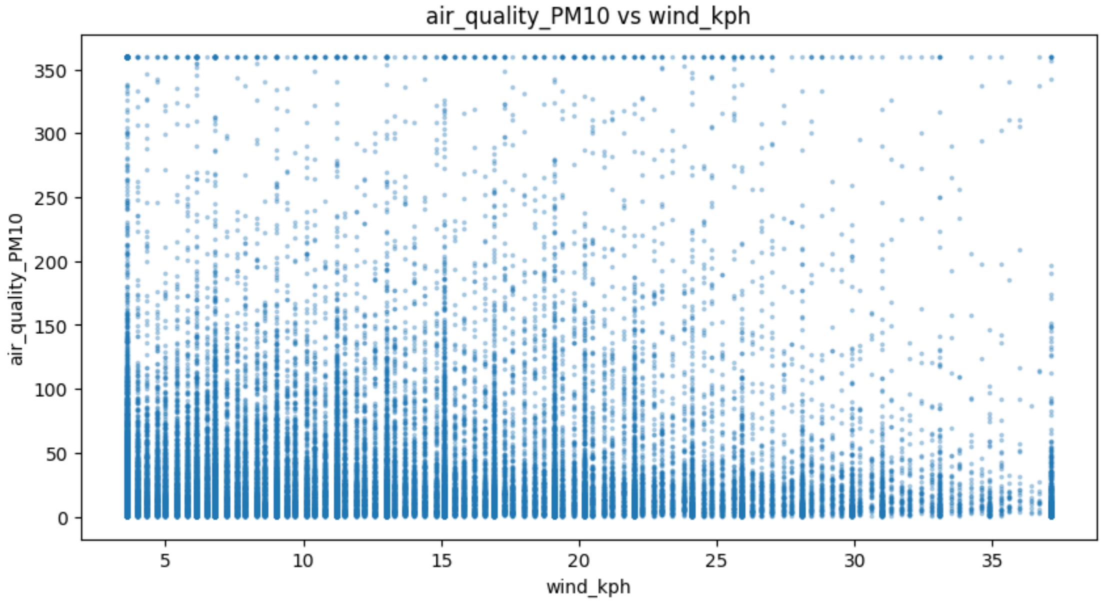
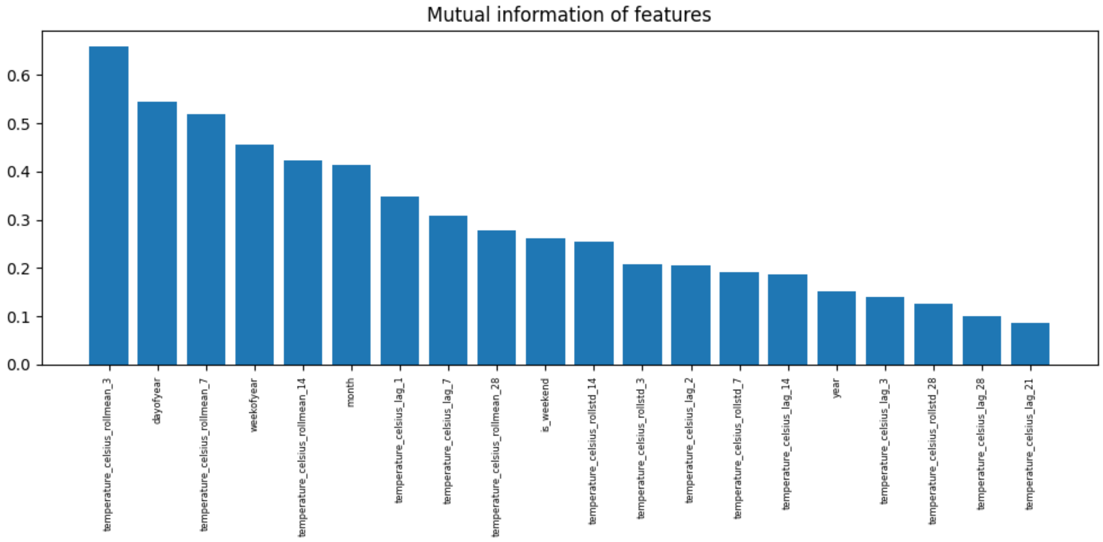
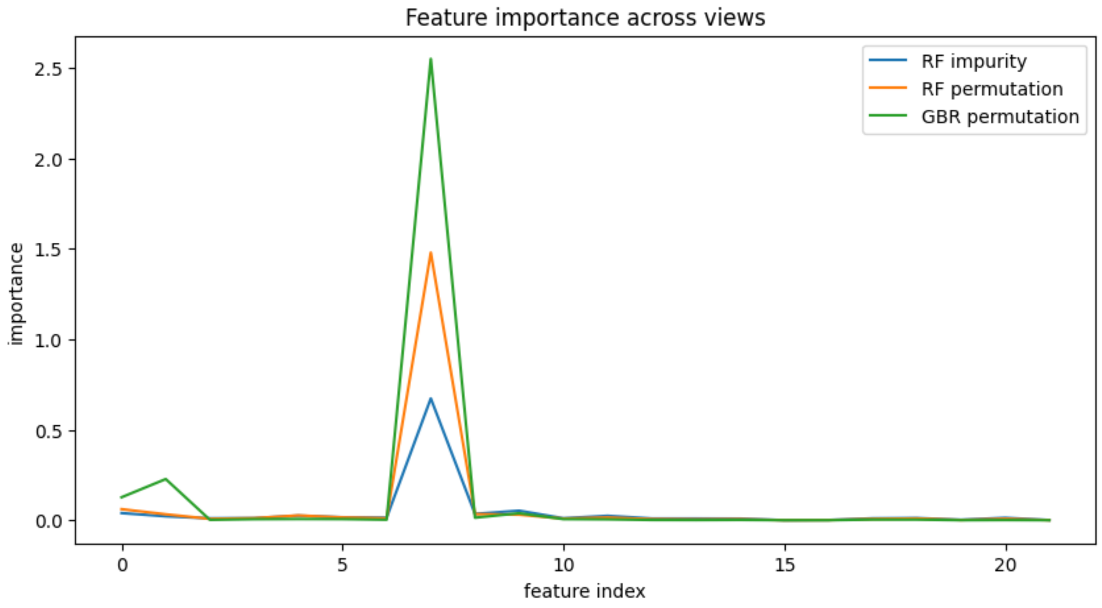


Figure 31: Mutual information of features



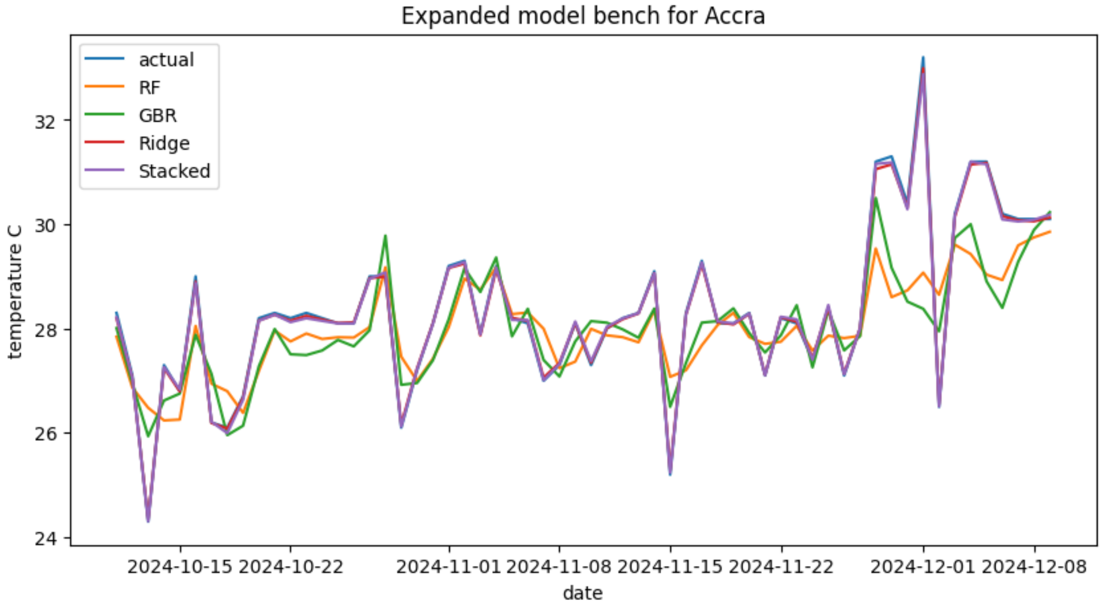
Visual inspection indicates a coherent signal with recognizable structure. This view informs both feature design and validation choices.

Figure 32: Feature importance across views



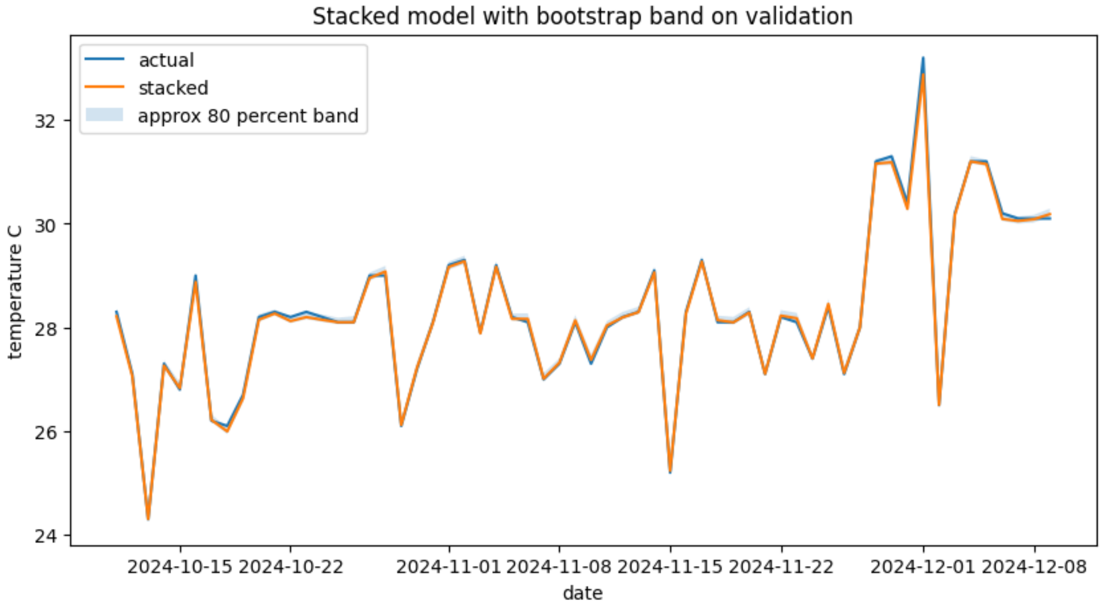
Visual inspection indicates a coherent signal with recognizable structure. This view informs both feature design and validation choices.

Figure 33: Model Comparison on Validation on Selected Location



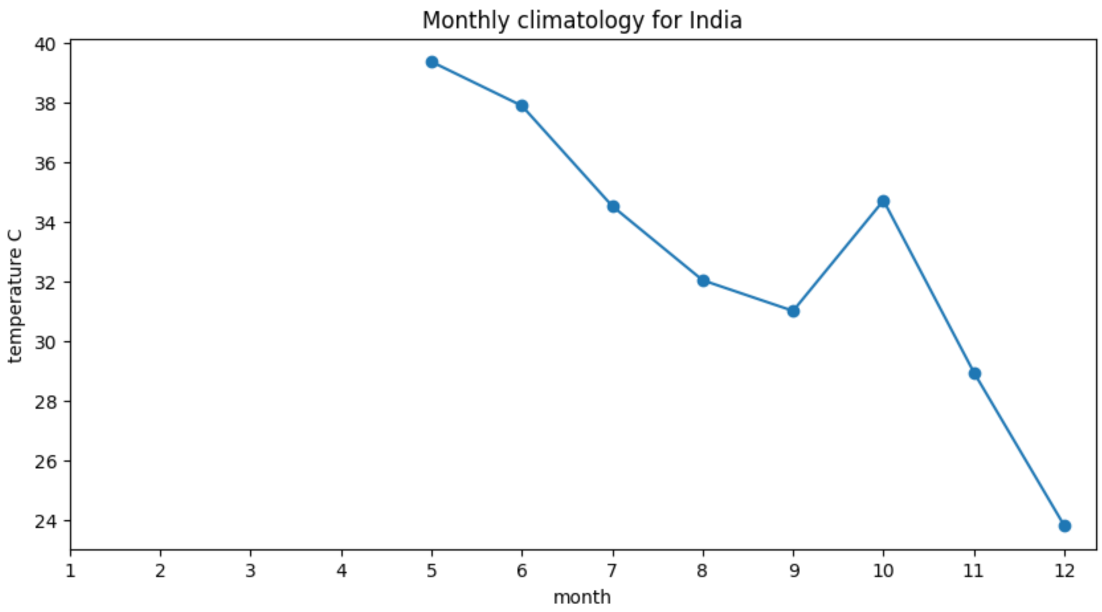
The stacked ensemble typically tracks the actuals more tightly, smoothing over idiosyncrasies of single models. Gaps between curves show where each learner struggles (e.g., rapid swings).

Figure 34: Stacked model with bootstrap band on validation



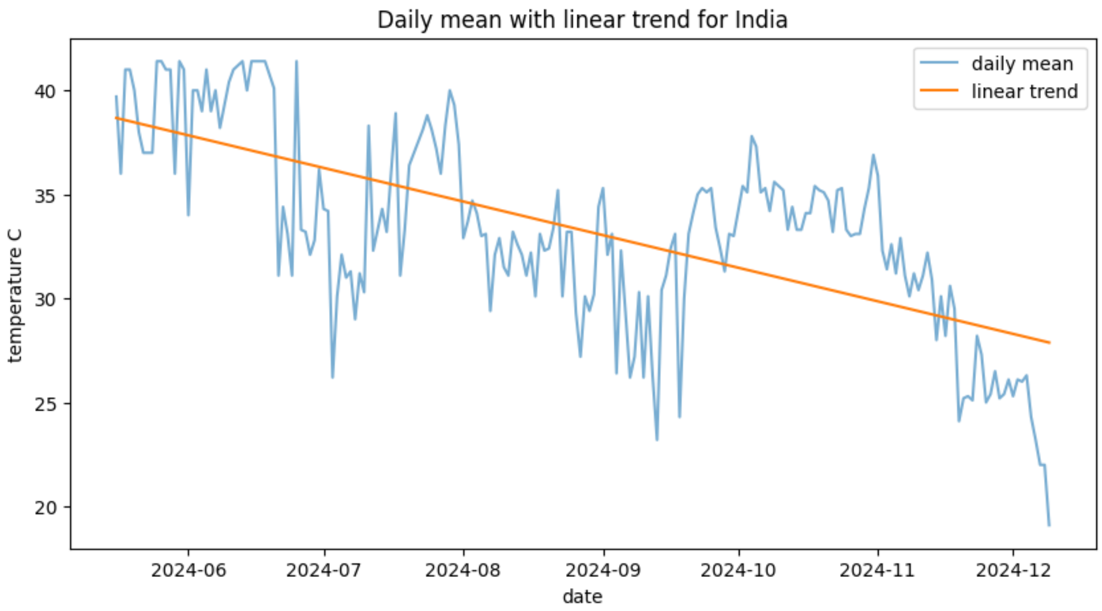
Visual inspection indicates a coherent signal with recognizable structure. This view informs both feature design and validation choices.

Figure 35: Monthly Climatology in India



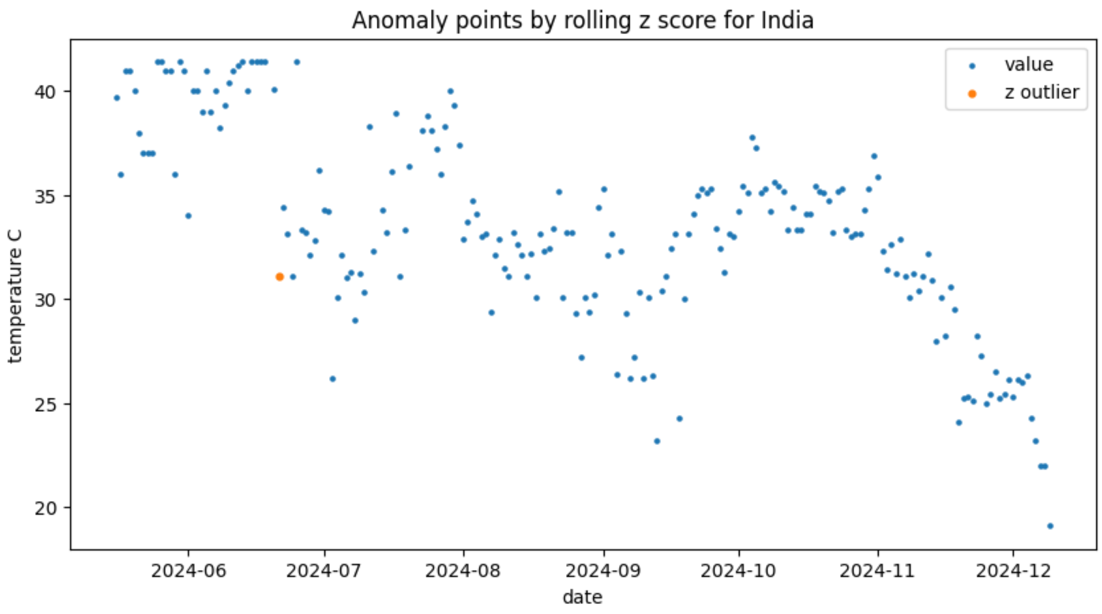
The curve clearly distinguishes between the warm and cool months, indicating clear seasonality. This form establishes a baseline expectation for forecasts and directs feature selection (lags and rolling means).

Figure 36: Daily Mean Temperature with Linear Trend in India



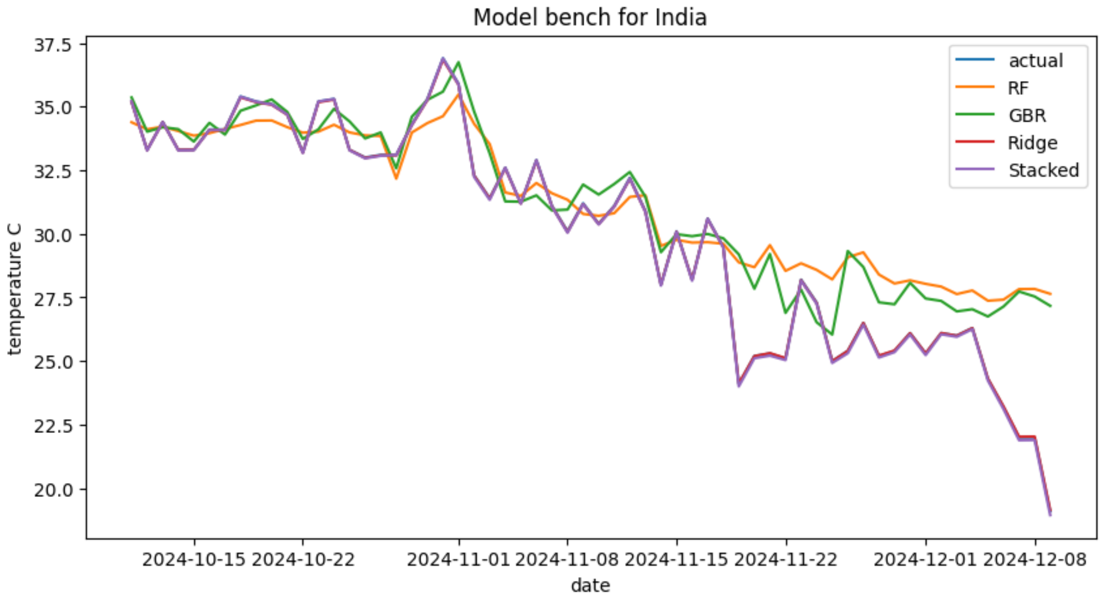
Overlaying a linear fit on the daily series suggests a gradual long-run drift. While day-to-day variance is large, the fitted line helps quantify secular change.

Figure 37: Rolling Z-Score Anomalies in India



Points flagged as anomalies deviate markedly from the short-term baseline, often indicating heat spikes, cold snaps, or sensor artifacts

Figure 38: Naive Forecast (Baseline) in India



With limited history, a naive baseline is shown: it projects the last observed value forward. It’s a conservative yardstick that prevents over-claiming in sparse settings.

Figure 39: Short-series validation in India

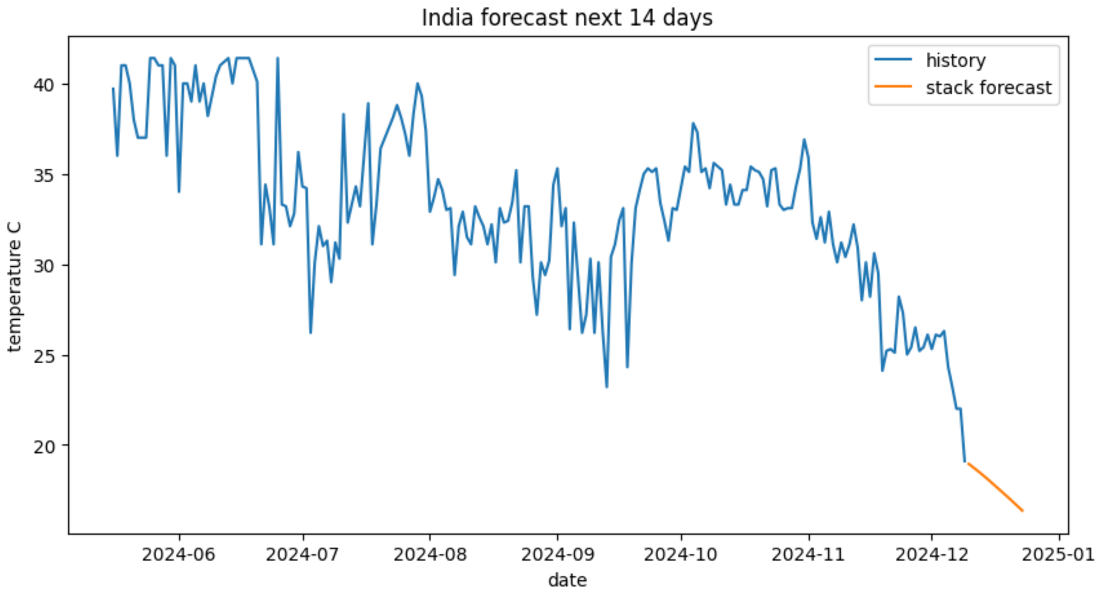
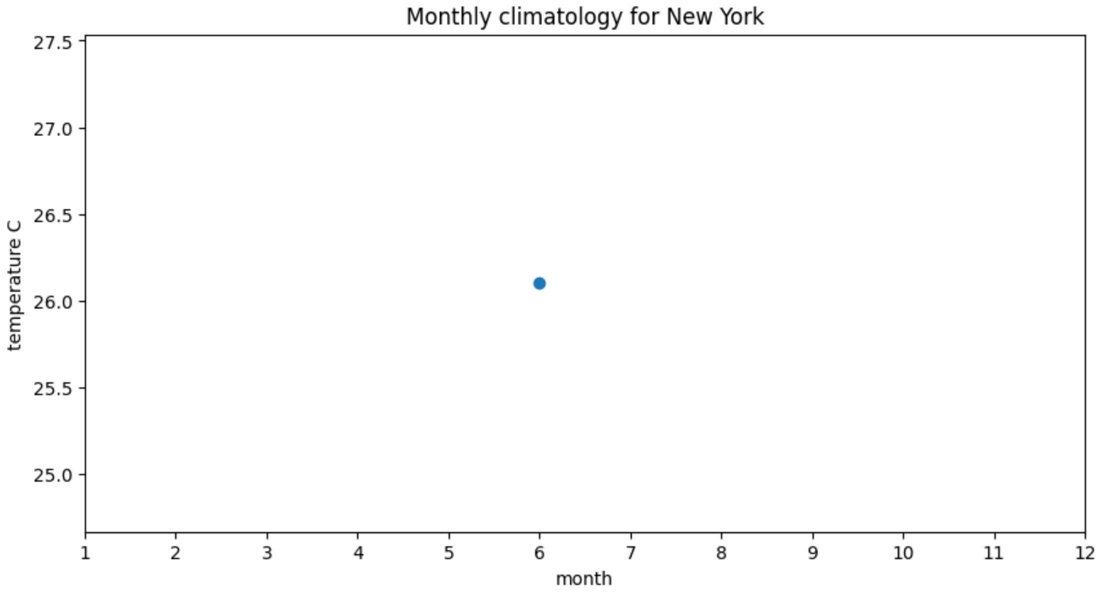
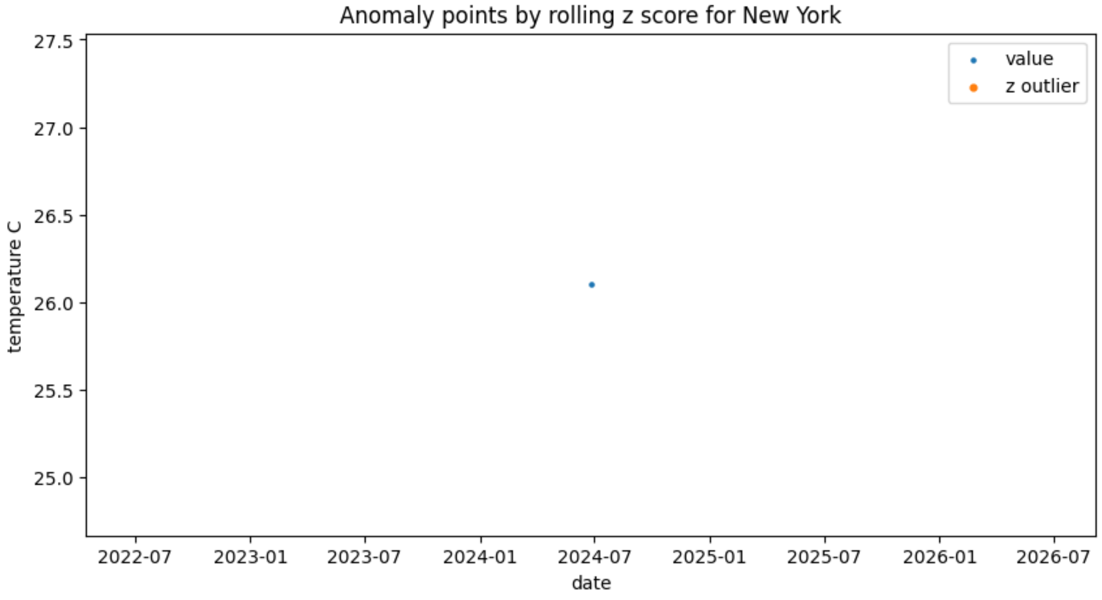


Figure 40: Naive Forecast (Baseline) in Dubai



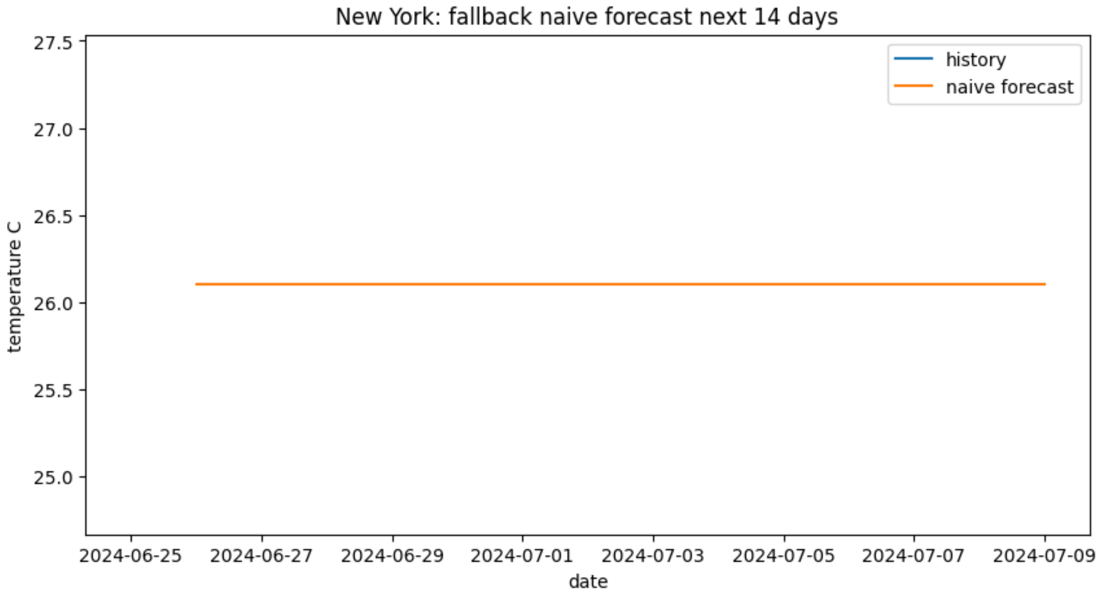
A naive baseline, which forward projects the most recent observed value, is displayed with limited history. It's a cautious metric that keeps over-claiming from happening in sparse environments.

Figure 41: Model Comparison on Validation for Dubai



The stacked ensemble typically tracks the actuals more tightly, smoothing over idiosyncrasies of single models. Gaps between curves show where each learner struggles (e.g., rapid swings).

Figure 42: Short-Horizon Forecast for Dubai



The forecast projects the recent trajectory into the near future. Sharp recent swings translate into greater uncertainty in practice; confidence is at its highest when the recent window is stable.