Supplementary Information

Final Report of Time Series Analysis

Chia-Ruei Liu, Htet Lin Naing, Junkang Xu, Zhiwen Huang

This file includes:

1. Supplementary Methodology
1.1. Data
1.2. Daily Temperature Data Analysis
1.2.1 Trend Component
1.2.2 Seasonal Component
1.2.3 Error Component
1.2.4 Forecasting and Model Performance Evaluation

2.Supplementary Tables

3.Supplementary Figures

4.Contriubution

1.0 Supplementary Methodology

1.1 Data

The ERA-Interim dataset is a global atmospheric reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). Reanalysis is a systematic approach to producing spatially and temporally gridded datasets through the integration of observational data with numerical weather models via data assimilation techniques. This methodology plays a pivotal role in advancing climate monitoring and analytical research. The ERA-Interim dataset is widely used for climate research, weather analysis, and many other atmospheric studies.

We choose the place of the sea of Cortez (109.5°W, 25.5°N), also known as the Gulf of California, to analysis its daily temperature changes from the beginning of 1979 to the end of 2017 for 14245 daily temperature data. The region is predominantly characterized by a desert or semi-arid climate, with summers marked by extreme heat, where daytime temperatures frequently exceed 40°C (104°F), particularly in inland areas. Winters are comparatively mild, with daytime temperatures ranging between 20°C and 25°C (68–77°F) and significantly cooler nocturnal conditions. We divide the data into 90% training set and 10% test set for further analysis. (Fig.S1).

1.2 Daily Temperature Data Analysis

1.2.1 Trend Component

To analyze the trend component of the time series, we first fit the polynomial regression models of varying degrees to the data. The trend represents the long-term movement of the data, often caused by external factors or structural changes over time. In our analysis, polynomial models from degree 0 to degree 4 are evaluated to determine which one best captures the trend. From the analysis, we are able to choose one most fitful model to do the de-trending.

The performance of each model is assessed using a series of statistical metrics. The Residual Sum of Squares (RSS) quantifies the total squared differences between the observed and predicted values, while the Residual Standard Error (RSE) provides a measure of the residual variability, accounting for the number of parameters in the model. The Sum of Squares for Regression (SSR) is used to evaluate how much variation in the data is explained by the model, and the R^2 statistic indicates the proportion of variance accounted for by the model. Additionally, model selection is guided by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which penalize models with higher complexity to avoid overfitting. Finally, an F-statistic is calculated to compare nested models and test whether adding more parameters significantly improves the fit.

1.2.2 Seasonal Component

In analyzing the seasonal component of time series data, we choose to use two distinct methods: OLS regression for seasonal mean and harmonic regression. The objective is to compare these methods and determine which is better suited for capturing the recurring patterns in the data.

The first method, OLS regression for seasonal mean, involves representing each time period (e.g., days or months) as a category using one-hot encoding. Ordinary Least Squares (OLS) regression is then applied to estimate the average seasonal effect for each time period. This approach is highly flexible, as it does not assume smooth or periodic behavior, making it particularly effective for modeling irregular seasonal patterns. However, this flexibility comes at the cost of complexity, as it requires estimating many parameters, which can lead to overfitting or computational challenges, especially for high-frequency data.

The second method, harmonic regression, models the seasonal component using sinusoidal functions, such as sine and cosine, to represent smooth and periodic cycles. This approach assumes that the seasonality is regular and predictable, which works well for data with consistent seasonal patterns, such as annual or monthly cycles. Harmonic regression is computationally efficient and parsimonious, relying on fewer parameters to capture the seasonal behavior. However, it may struggle to accurately represent irregular or abrupt changes in seasonality, as it inherently assumes smooth and periodic variations.

To determine the most appropriate method, the two approaches are compared based on flexibility, simplicity, accuracy, and residual analysis. Flexibility is key for irregular seasonal patterns, favoring OLS regression, while harmonic regression excels in simplicity and parsimony for data with smooth, periodic cycles. Accuracy is evaluated using statistical metrics such as residual sum of squares (RSS), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC), which balance model fit and complexity. Residual analysis is performed to confirm that the seasonal component is effectively captured, leaving no significant patterns in the residuals.

1.2.3 Error Component

The error component in time series analysis represents the unexplained variation remaining after removing the trend and seasonal components. To ensure this component behaves as random noise, its structure is analyzed through the autocorrelation function (ACF) and partial autocorrelation function (PACF). These tools identify potential correlations or dependencies within the residuals, which could suggest lingering patterns. Use residual diagnostics to validate the results. Residual scatter plots are examined for randomness, and statistical tests like the Ljung-Box test are used to confirm the absence

of autocorrelation. If the residuals are independent and show no patterns, the error component can be considered successfully isolated. This process ensures that the time series decomposition has effectively captured the trend and seasonal components, leaving only random, unstructured variation.

1.2.4 Forecasting and Model Performance Evaluation

Forecasting in time series analysis involves predicting future values by combining the trend, seasonal, and residual components identified earlier. The trend captures long-term patterns and is extrapolated into the future, while the seasonal component models recurring cycles for daily temperature data. The residual component, representing random fluctuations, is forecasted using a time series model to account for dependencies in the noise. Confidence intervals are calculated to quantify uncertainty, providing a range within which future values are likely to fall. To evaluate the forecast's performance, metrics such as root mean square error (RMSE), mean relative prediction error (MRPE), and empirical coverage rate (ECR) are used. These metrics assess the accuracy of the predictions and the reliability of the uncertainty estimates, ensuring the model effectively captures the dynamics of the time series and produces robust forecasts.

2.0 Supplementary Tables

The Residual Sum of Squares (RSS) was calculated for polynomial models of different orders (1 to 4) (Table.1). While the RSS values decreased as the polynomial order increased, the reduction became marginal at higher orders. For instance, the RSS for the 3rd-order model was 217138.69, only slightly lower than the 4th-order model's RSS of 217134.68. This diminishing improvement indicates that increasing the model's complexity beyond the 3rd order yields limited benefits. To evaluate the significance of the improvement between models, an F-test was conducted. Comparing the 2nd-order and 3rd-order models, the F-statistic was 9.27, and the P-value was 0.00232. The low P-value (< 0.05) indicates that the 3rd-order model provides a statistically significant improvement over the 2nd-order model. Based on these findings, the 3rd-order polynomial model is selected as the optimal model. It offers a good balance between model fitness and complexity, significantly outperforming the 2nd-order model while avoiding the unnecessary complexity of the 4th-order model.

We represent the third-order polynomial model to capture the trend based on the original data (Fig.1a) and plot the residuals over time after de-trending (Fig.1b). It can be observed that, after de-trending, the residuals consistently fluctuate around zero, indicating that the de-trending process is effective.

Table.S1| Polynomial model evaluation. Different numbers in the Poly line represent different orders of the polynomial model, Rss is the residual sum of squares, AIC is Akaike Information Criterion (AIC), and BIC is Bayesian Information Criterion.

Poly	1	2	3	4
RSS	217566.62	217293.64	217138.69	217134.68
AIC	73524.6	73510.3	73503.0	73504.8
BIC	73539.6	73532.7	73532.9	73542.2

Table.S2a| Comparison of seasonal mean and harmonic regression of the training data.

Model	RSS	AIC	BIC
Seasonal Mean	15300.66	39742.71	42477.72
Harmonic Regression	18477.04	41466.94	41481.88

Table.S2b| Comparison of seasonal mean and harmonic regression of the test data for forecasting.

Model	RMSE	MRPE	ECR
Harmonic Regression	1.238	0.038	0.936
Seasonal Mean	1.107	0.035	0.965

Table.S3| AIC and BIC for ARIMA(x,0,0) from x=0 to 14. From the comparison of AIC and BIC to different values of x, we can see that ARIMA(13,0,0) has the lowest AIC and BIC among these 14 models, indicating that ARIMA(13,0,0) is the most optimal model for error component.

x	1	2	3	4	5	6	7
AIC	23461	23460	23242	23178	23114	23087	23072
BIC	23483	23489	23280	23223	23167	23147	23139
x	8	9	10	11	12	13	14
AIC	23050	23025	23003	23002	23000	22967	22954
BIC	23125	23107	23093	23099	23104	23073	23074

3.0 Supplementary Figures

Fig.S1| Origin data and data set partition. The left side of the red vertical bar represents 90% training set and the right side represents 10% test set.

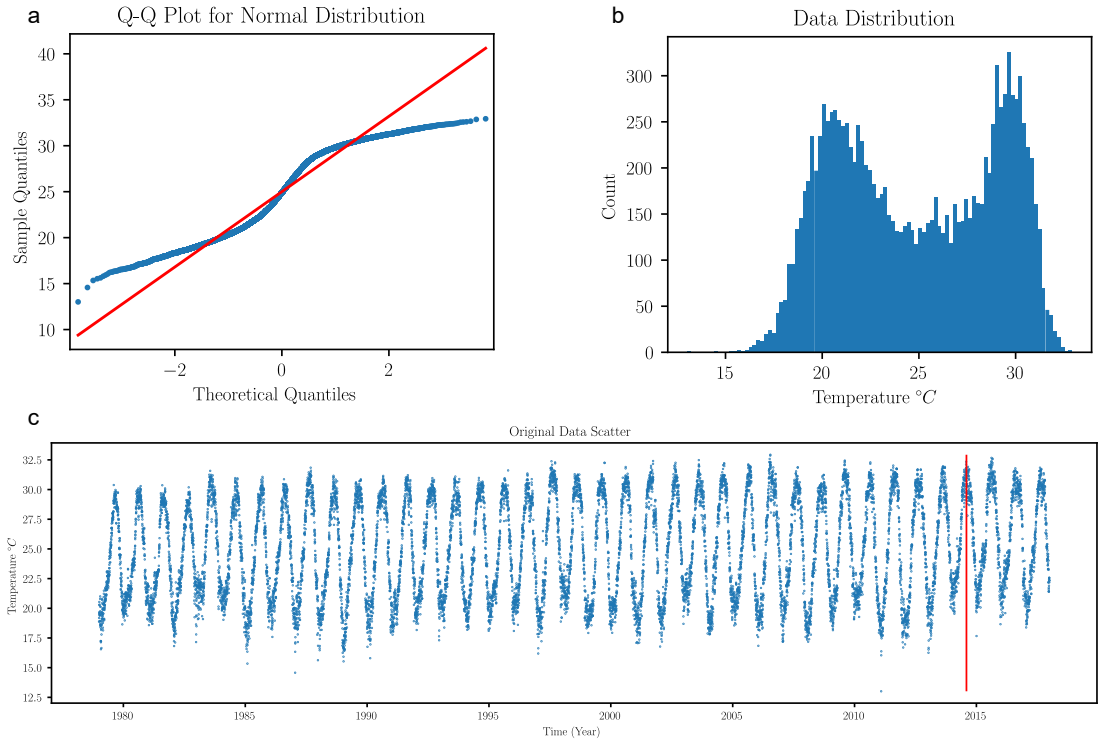


Fig.S2| Details of the Harmonic Regression Model. a compares the detrended data (blue points) with the predicted seasonal component (red curve) from the Seasonal Mean model, **Residual Scatter Plot b** shows the residuals after removing both trend and seasonal components, **c** is the Q-Q plot.

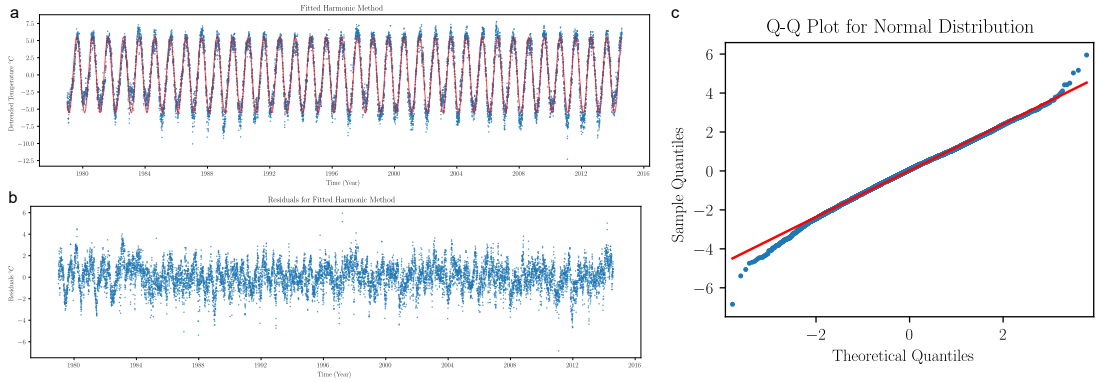
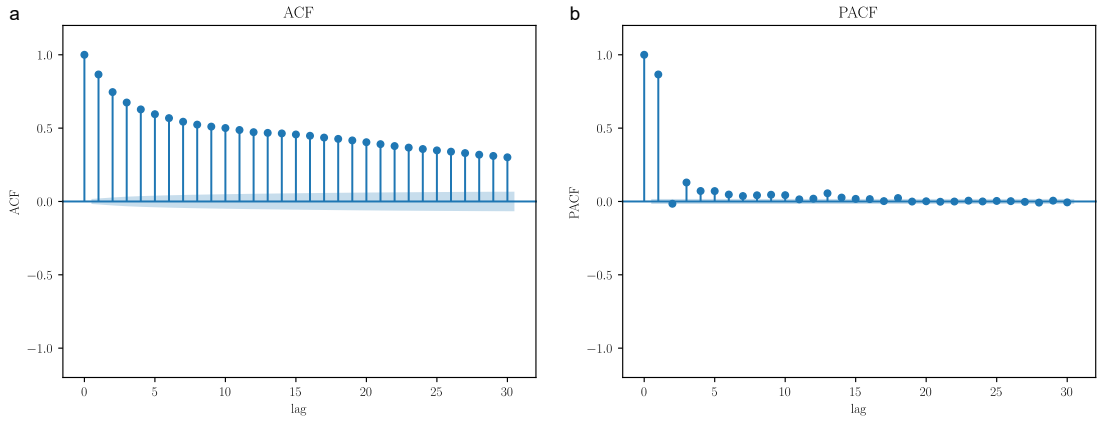


Fig.S3| ACF and PACF plot of the error component. In the ACF plot (left), the autocorrelations gradually decrease in a geometric pattern, indicating the data has a single lag dependency, typical of an AR(1) process. In the PACF plot (right), only the first lag has a significant partial autocorrelation, while the rest are close to zero, confirming that the process can be adequately described using a single autoregressive term. These characteristics are consistent with an AR(1) model.



4.0 Contribution

Chia-Ruei Liu primarily did the analysis part.

Zhiwen Huang prepared the PowerPoint slides.

Htet Lin Naing gave the presentation.

Jungkang Xu wrote the final report.

We all provided our own ideas across all tasks although there are specific assignments to contribute.