MATH6011-Forecasting
2024-25
Professor Alain Zemkoho

Zhe GUAN
Student ID: 36516473
Email: zg2u24@soton.ac.uk

Exponential Smoothing

Data Preparation and Exploratory Data Analysis

This analysis involved forecasting four different time series datasets (MSTA, CH4, GMAF, and ET12) using Exponential Smoothing models.

The data preparation stage included preparing data sets from original Excel files, checking for missing values, and handling non-positive values where required for multiplicative Holt-Winters models. Missing values were managed through interpolation.

A preliminary analysis involved visually inspecting the datasets to assess the presence of trend and seasonality as shown in Fig 1. For the MSTA series, it was observed that the measurements before 1960 contained larger uncertainties as evident from the wider confidence intervals, and exhibited patterns inconsistent with the post-1960 period. As the aim was to forecast values for 2025, only data from 1960 onward were used to ensure better model consistency and forecast accuracy. The other datasets (CH4, GMAF, and ET12) either exhibited smaller measurement errors or did not report uncertainty estimates, and were therefore retained in full without truncation.

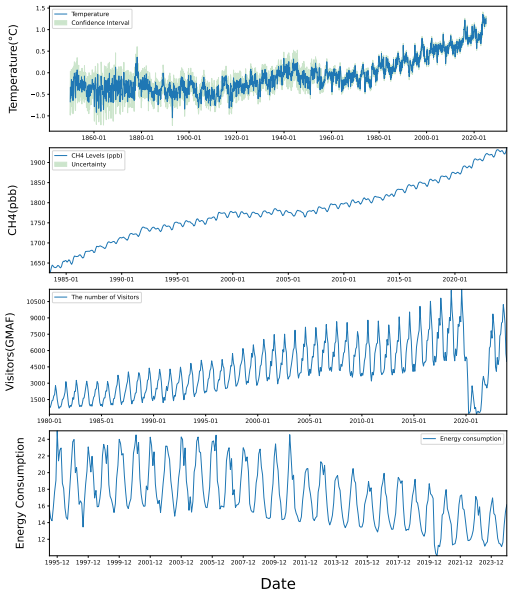


Figure 1: Original Distributions for MSTA, CH4, GMAF and ET12

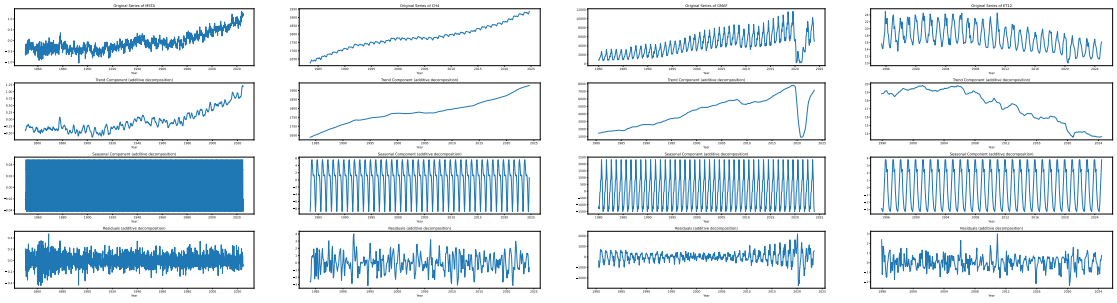
Table 1: Summary of parameters and errors from exponential smoothing models for MSTA

Metrics	**SES Model**	**HES Model**	**HWES Model**
α	0.544	0.542	0.544
β	-	1.58×10^{-17}	2.21×10^{-18}
ϕ	-	-	-
l_0	-0.115	-0.130	1.213
b_0	-	1.73×10^{-3}	1.00×10^{-3}
MSE	0.011	0.011	0.010
MAE	0.081	0.081	0.079
R^2	0.923	0.923	0.925

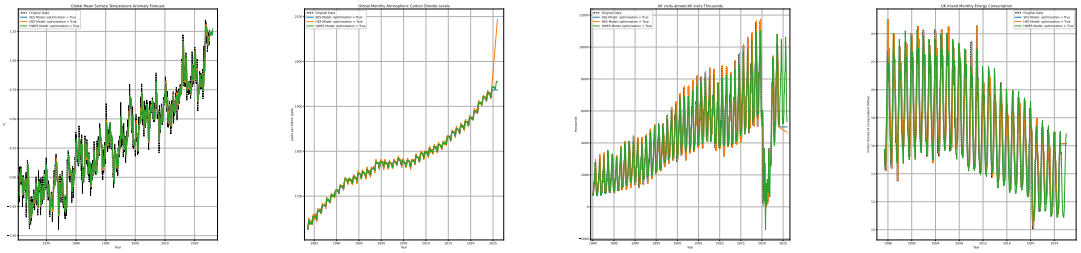
For model selection, three exponential smoothing techniques were evaluated: Simple Exponential Smoothing (SES), Holt's Exponential Smoothing (HES), and Holt-Winters Exponential Smoothing (HWES). Each technique was assessed using both manually specified parameters and automatic optimization within Python. Automatic parameter optimization consistently yielded lower MSEs and was thus adopted for all final forecasts. Specifically, for the MSTA dataset, the performance differences among the SES, HES, and HWES models were relatively minor, as the seasonal component was less pronounced. For the other datasets (CH4, GMAF, ET12), clear seasonality was evident, making Holt-Winters Exponential Smoothing (HWES) the most suitable method for accurately forecasting these seasonal patterns.

Model metrics are presented with tables summarizing key performance metrics (MSE, MAE, R^2) and optimized model parameters for each method as shown in Tab 1. More details for other three datasets are provided in the code framework.

Forecasting prediction with different exponential smoothing methods for four data is shown at Fig 2



(a) MSTA additive decomposition (b) CH4 additive decomposition (c) GMAF additive decomposition (d) ET12 additive decomposition



(e) MSTA multiplicative forecast (f) CH4 multiplicative forecast (g) GMAF multiplicative forecast (h) ET12 multiplicative forecast

Figure 2: Additive and multiplicative decomposition forecasts for the four time series datasets (MSTA, CH4, GMAF, ET12)

Notably, for future predictions, the SES model produced flat forecasts, reflecting its inability to account for trend and seasonality. HES yielded linear forecasts due to its capability of modeling trend but not seasonality. The HWES, however, showed fluctuations consistent with observed seasonal patterns, capturing both trend and seasonal effects effectively.

AutoRegressive Integrated Moving Average

The AutoRegressive Integrated Moving Average(ARIMA) analysis focused on forecasting the MSTA dataset.

Data preparation included checking for and addressing non-positive values using transformations, and applying the Box-Cox transformation to stabilize variance as shown in Fig 3. The effectiveness of these steps was verified through visual comparison of rolling standard deviations.

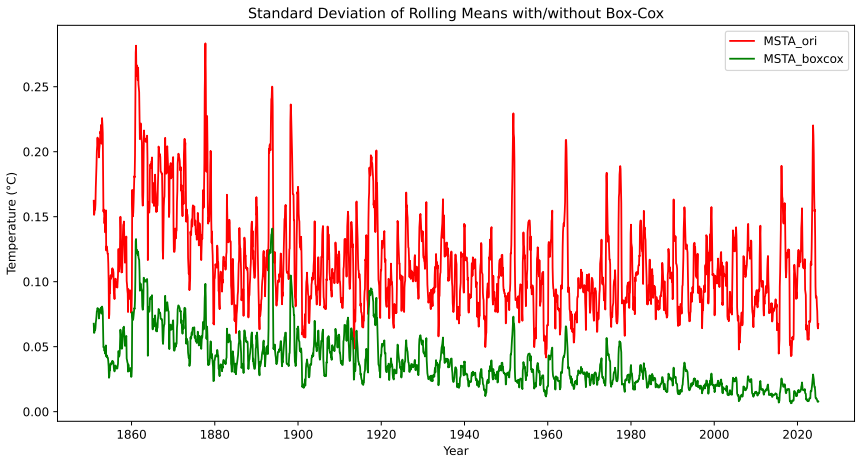


Figure 3: Standard Decviation of Rolling Means with/without Box-cox transformation

The preliminary analysis comprised stationarity checks using Augmented Dickey-Fuller (ADF) tests and visual inspections via Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. Initial findings indicated non-stationarity, necessitating first-order differencing to achieve stationarity as shown in Fig 4.

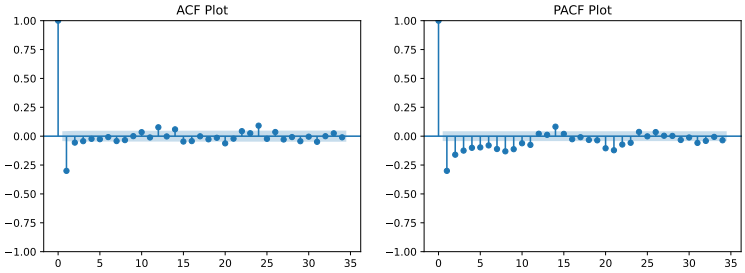


Figure 4: ACF and PACF after first differencing

Model selection was conducted via systematic grid search across ARIMA parameters (p, d, q) employing cross-validation as shown in Fig 5. The optimal ARIMA(7,1,2) model was selected based on the lowest AIC and mean cross-validation MSE. Model performance and accuracy were validated through forecasting and residual analysis.

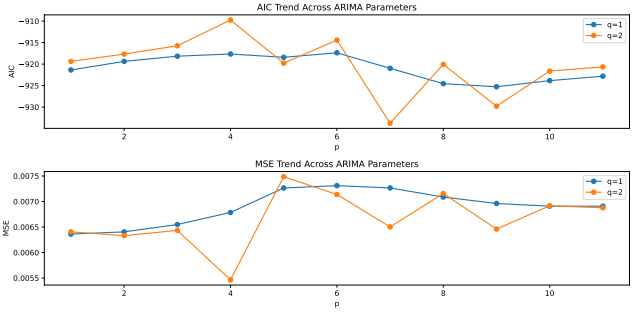


Figure 5: AIC and MSE Trend Across ARIMA Parameters

In comparison to exponential smoothing, ARIMA provided a structured approach to modeling autocorrelations explicitly, while exponential smoothing offered simpler computational implementation. Specifically for MSTA prediction as shown in Fig 6, ARIMA captured complex autocorrelations effectively, while exponential smoothing methods provided simpler but similarly accurate forecasts due to the limited seasonal effects in the dataset.

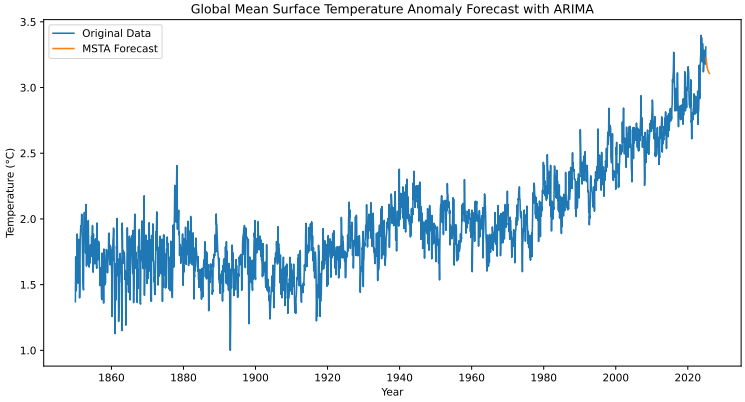


Figure 6: Global Mean Surface Temperature Forecast with ARIMA

Regression prediction

The regression analysis focused on predicting the MSTA dataset using multiple regression techniques with interaction terms and kernel ridge regression. Data preparation involved aligning and standardizing time-series data from multiple variables (CH4, GMAF, ET12), along with generating lagged variables and interaction terms to address potential non-linear relationships and improve predictive accuracy.

Preliminary analysis included identifying optimal lag values for each predictor variable using the AIC to enhance forecasting performance. Additionally, variance inflation factor (VIF) analyses were conducted to ensure multicollinearity remained within acceptable limits, ensuring the stability and interpretability of the regression model.

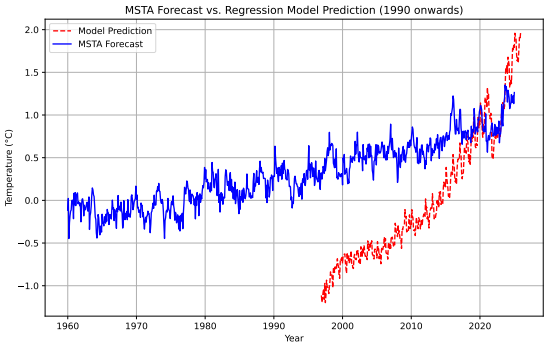


Figure 7: Liner Regression for MSTA with lagged interaction terms

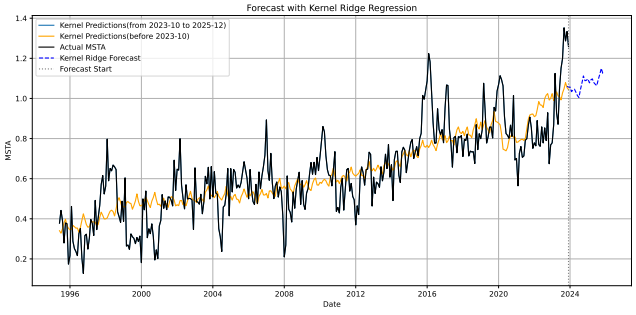


Figure 8: Kernel Regression for MSTA consdering non-linear relationship with CH4, GMAP and ET12

The regression analysis initially considered linear regression with lagged variables and interaction

terms, selected using AIC-based lag optimization. However, since data relationships appeared non-linear, Kernel Ridge Regression was also evaluated. Grid search with cross-validation (GridSearchCV) was conducted over parameters for different kernel functions (RBF, polynomial, sigmoid), varying kernel hyperparameters (alpha, gamma, and degree). The best Kernel Ridge model was selected based on the lowest Mean Squared Error (MSE) obtained via cross-validation. Hence, Kernel Ridge Regression was determined to be suitable and more effective for forecasting MSTA due to its capability to capture non-linear relationships present in the dataset.

- **Linear regression (with and without lagged interaction terms):** yielded poor fit quality (adjusted $R^2 \approx 0.60$ and MSE = 0.359).
- **Kernel Ridge Regression (KRR):** implemented due to its capability in capturing non-linear relationships. Grid search with cross-validation was used to optimize KRR parameters (best found: kernel=`poly`, $\alpha = 0.001$, degree=2). KRR demonstrated superior forecasting performance on MSTA ($R^2 = 0.92$, MSE = 0.0105), clearly outperforming linear models.

Appendix: Code and Data Summary

- **EDA_Preprocess_36516473.py**: Loads and preprocesses raw datasets (MSTA, CH4, GMAF, ET12), standardizes date formats, handles missing values, and exports cleaned data into the `Data_36516473.xlsx`.
- **EDA_Decomposition_36516473.py**: Performs additive and multiplicative seasonal decomposition on all series. Non-positive values are handled when required for multiplicative decomposition. Decomposition plots are saved for each dataset.
- **ExponentialSmoothing_All_36516473.py**: Applies and compares SES, HES, and HWES models on each dataset. Forecasts are visualized, and parameter optimization is optionally enabled.
- **ARIMA_TrainingForecast_36516473.py**: Focuses on the MSTA series. Applies Box-Cox transformation and differencing, verifies stationarity (ADF test), plots ACF/PACF, searches optimal ARIMA(p,d,q), and performs a 12-month forecast.
- **LinearRegression_ModelTraining_36516473.py**: Builds linear regression models (with/without lags) using CH4, GMAF, ET12. Interaction terms and lag optimization (AIC-based) are used. VIF analysis ensures multicollinearity is acceptable.
- **KernelRegression_TrainingEvaluation_36516473.py**: Implements Kernel Ridge Regression (KRR) to model non-linear relationships between MSTA and its predictors. Uses grid search with cross-validation to optimize kernel type and hyperparameters. Forecasts future MSTA based on smoothed HWES inputs.
- **LinearRegression_Forecast_36516473.py**: Applies the trained regression model to forecast MSTA for 2025, using lagged forecasts of CH4, GMAF, and ET12. Results are compared with ARIMA forecasts.

Exported Data Notes: `Data_36516473.xlsx` contains all cleaned datasets in separate sheets. The file `regression_data.xlsx` contains aligned training data over the common overlapping time span across all four variables. Forecast files (e.g., `CH4_forecasts_smooth.xlsx`) store HWES-based forecasts for each respective dataset. All files are located in the `Exported_Data` folder.

Models Notes: All the trained regression model including linear regression `model_with_lags.pkl`, `model_without_lags.pkl` and kernel regression `model_with_kernel.pkl` are located in the `Models` folder.