# *Predicting Global Temperatures: A Comparative Analysis of Prediction Models*

## Introduction

Climate change and its consequences on global temperatures have enhanced the importance of accurate climate modelling, particularly in forecasting surface air temperatures.

This project aims to solve the problem of predicting global surface air temperatures by developing and testing predictive models based on a Linear Basis Function (LBF) approach. The study specifically investigates the predictive performance of three families of basis functions - polynomial, piece-wise constant, and piece-wise linear to discover which is most effective at modelling the relationship between time (year) and global surface air temperature.

The primary goal of this project is to identify the basis function family that minimises prediction error when applied to the ERA5 dataset, which contains monthly surface air temperature records from January 1940 to September 2024. By examining different basis functions, this study aims to develop an effective model for forecasting future temperatures, contributing to climate action initiatives (as part of UN Sustainable Development Goal 13).

The report proceeds as follows: **Section 2** outlines the methodology used to define the LBF model and describes how the three basis function families are implemented and tuned. **Section 3** presents the empirical results, including selecting optimal hyperparameters and the model comparisons based on predictive performance. The **piece-wise linear model** is the best-performing model due to its ability to global trends and local variations, achieving the lowest test MSE. Finally, **Section 4** concludes with an overview of the findings and potential directions for further research, particularly in extending the study to include additional climate variables.

## Methodology

This section provides a comprehensive description of the LBF Model used to predict global surface air temperatures, focusing on three basis function families- Polynomial, Piece-Wise Constant, and Piece-Wise Linear. The methodology covers model definitions, parameter estimation, hyperparameter optimisation, and performance evaluation.

The Linear Basis Function Model is a type of regression that models the target variable, global surface air temperature (y), as a linear combination of features. These features are derived from the input variable (year) using different basis functions. The Equation (1) below represents the general form of the model:

$$y = u^{T}\alpha + \phi(x)^{T}\beta + \varepsilon \tag{1}$$

Where:

- y: Global surface air temperature.
- x: Time variable (year).
- u: Vector of dummy variables representing the month of observation, capturing seasonal variations.
- α: Coefficient vector for month-specific seasonal effects.
- ϕ(x): Vector of basis functions applied to the time variable (year), representing trends over time.
- β: Coefficient vector for the time-dependent basis functions.
- ε: Error term, accounting for randomness or noise not captured by the model.

The choice of the basis function ϕ(x) allows the model to capture complex patterns in the data, ranging from smooth trends to abrupt changes.

Polynomial Basis Functions: The polynomial model is designed to capture smooth and continuous trends in temperature data using the powers of the predictor variable (year). Each additional power adds complexity to the model, allowing it to fit a non-linear curve as mentioned in Equation (2) below:

$$\phi(x) = [1, \phi_1(x), \ldots, \phi_p(x)]^T, \text{ with } \phi_i(x) := x^i \tag{2}$$

Here, Degree of Polynomial (p) is the hyperparameter controlling the model's complexity that determines how flexible the curve is.

Piece-Wise Constant Basis Functions: The piece-wise constant model divides the timeline into discrete segments. Within each segment, the model predicts a constant temperature, if temperature remains the same across that interval as depicted in equation (3):

$$\phi(x) = [1, \gamma_1(x),\ldots,\gamma_k(x)]^T \text{ , with } \gamma_i(x) := I(x>t_i) \tag{3}$$

- $I(x > t_i)$ is an indicator function that activates (equals 1) if x (year) exceeds a specific breakpoint $t_i$ and equals 0 if x is less than or equal to $t_i$.
- k is the number of breakpoints $[\{t_i\}^k{}_{i=1}]$ and hyperparameter that divides the timeline into k +1 segments which are calculated as in Equation (4):

$$t_i := x_{min} + \frac{i(x_{max}-x_{min})}{k+1} \tag{4}$$

Piece-Wise Linear Basis Functions: The piece-wise linear model is an extension of the piece-wise constant approach. It divides the timeline into segments but allows for a linear relationship within each segment. This provides greater flexibility in modelling gradual changes, as showed in Equation (5):

$$\phi(x) = [1, x, \lambda_1(x),\ldots, \lambda_k(x)]^T, \text{ with } \lambda_i(x) := (x-t_i)I(x>t_i) \tag{5}$$

Where:

- $(x - t_i)I(x > t_i)$ creates a linear trend that only activates when the year x surpasses a breakpoint $t_i$.
- The slopes of the linear segments are estimated for each interval and the number of Breakpoints (k) controls the number of segments.

The parameter vectors α and β are estimated using the Ordinary Least Squares (OLS) method, a fundamental technique in linear regression. The goal is to find the coefficients that minimise the Residual Sum of Squares (RSS), which measures the difference between observed and predicted temperatures.

The process begins with a Grid Search which is a systematic method to explore a predefined set of hyperparameter values. Following this, Cross-Validation is conducted by splitting the dataset into training (50%) and validation (25%) and test (25%) sets, where each hyperparameter is evaluated based on its MSE on the validation set. After this step, the Optimal Selection step identifies the hyperparameter value that minimises MSE on the validation set, ensuring the model is neither underfitting nor overfitting. Lastly, the Design Matrix is constructed using the basis functions ϕ(x), which incorporate all polynomial terms or piece-wise indicators relevant to the data, setting up the model to accurately capture temperature trends.

OLS Estimation: The OLS method calculates the optimal values of α and β as shown by solving Equation (6):

$$(\Phi^T\Phi)^{-1}\Phi^T y \tag{6}$$

- Φ: Design matrix with basis functions and dummy variables.

- The matrix inversion step can encounter numerical challenges, especially if the columns of the design matrix are highly correlated (multicollinearity). This issue is particularly relevant in polynomial basis function.

After selecting the optimal hyperparameters, each model's performance is evaluated on a test dataset to ensure it generalises well to new data. The evaluation involves:

Performance Metric: The primary metric used is the Mean Squared Error, which measures the average squared difference between observed and predicted temperatures in validation dataset. A lower MSE indicates better model accuracy.

Model Comparison: The three models were compared based on their ability to capture different patterns in temperature data. The Polynomial Model was tested for its ability to represent smooth, non-linear trends. However, its smooth approximation struggled with abrupt shifts, leading to a relatively higher MSE. The Piece-Wise Constant Model was evaluated for its effectiveness in detecting sudden changes by dividing the data into segments with constant values. While it performed better than the Polynomial Model at identifying sharp shifts, its rigidity in maintaining constant predictions within each segment contributed to a moderately high MSE. The Piece-Wise Linear Model demonstrated the greatest flexibility, successfully capturing both gradual trends and sudden shifts in temperature. This adaptability resulted in the lowest MSE, making it the most accurate model for fitting various patterns in the data. The model with the lowest test MSE (Piece-Wise Linear Model) is selected as the best-performing model, indicating that it best captures the underlying temperature trends without overfitting the training data.

## Empirical Study

This section shares the empirical results from using three different models to predict global surface air temperatures. It covers detail about the dataset, hyperparameter optimisation, model performance evaluation, and temperature predictions for the final months of 2024.

The dataset consists of monthly temperature measurements from January 1940 to September 2024. Each entry includes the Year of the observation, the Month (ranging from 1 for January to 12 for December), and the Temperature, representing the surface air temperature for that month in degrees Celsius.

The hyperparameters for each basis function model were optimised using a validation dataset as depicted in Table 1:

| Model | Hyperparameter Value |
|---|---|
| Polynomial | p = 2 |
| Piece-Wise Constant | k = 24 |
| Piece-Wise Linear | k = 27 |

Table 1: Hyperparameter Value for each basis function model

The Polynomial Model was optimised with a degree of p = 2 (quadratic), effectively capturing the overall trend without overfitting and maintaining a good balance. For the Piece-Wise Constant Model, k = 24 breakpoints were chosen to detect sudden temperature shifts while keeping the model simple. The Piece-Wise Linear Model used k = 27 breakpoints, allowing for localized linear trends within each segment, giving it the flexibility to accurately capture both gradual and sudden changes in the data.

To assess and compare the predicted response values and performance of the three basis function families, it is important to visualise the predicted response values (temperatures) against the actual observed values from the test set. Visualising the predicted versus actual data helps provide a clearer understanding of how well each model captures the underlying patterns in the data, such as trends, fluctuations, and sudden shifts in temperature, as depicted in Figure 1, Figure 2 and Figure 3 below:

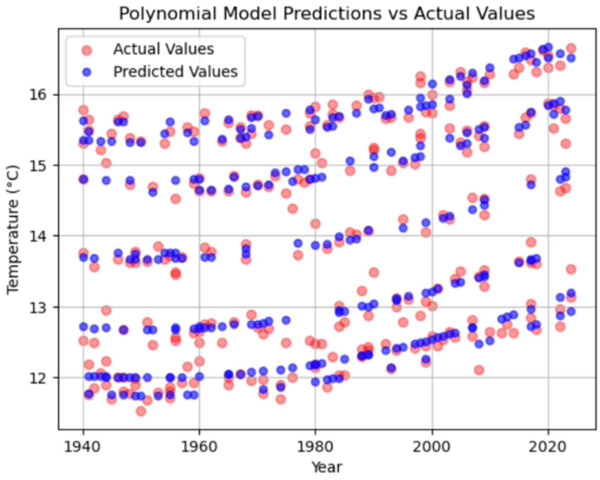

Figure 1: Polynomial Model Prediction

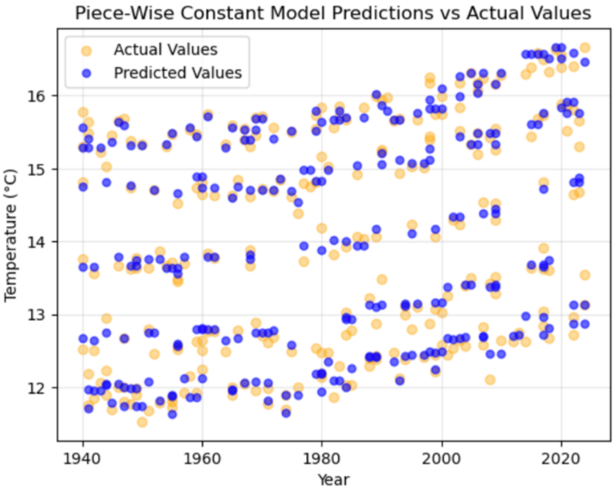

Figure 2: Piece-wise Constant Model Prediction

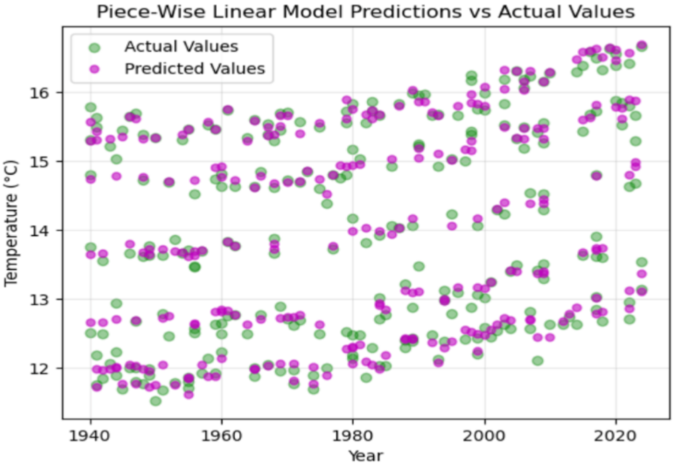

Figure 3: Piece-Wise Linear Prediction

The visualisations of each model reveal their unique strengths and limitations in capturing temperature trends. In Figure 1, the Polynomial Model displays predicted versus actual temperature values, effectively capturing the overall trend and indicating whether temperatures are rising or falling. However, its smooth curve struggles with sudden changes and local variations, resulting in some discrepancies between predictions and actual values. This model is valuable for understanding broad trends but is less accurate with short-term fluctuations, which contributes to a slightly higher error compared to models that adapt to specific time segments. The Figure 2 presents the Piece-Wise Constant Model as horizontal steps, capturing abrupt temperature jumps well but oversimplifying gradual trends. Each segment maintains a constant temperature, making sudden changes easy to identify but lacking flexibility for smoother transitions. This step-like approach highlights stable periods or clear jumps effectively but misses finer details in gradual changes, leading to a slightly higher MSE than the Piece-Wise Linear Model and making it less accurate for gradual shifts, especially toward the end of 2024. In Figure 3, the Piece-Wise Linear Model provides predictions that closely match actual values through a segmented approach that adapts to both gradual and sudden changes. This flexibility allows it to capture both the overall trend and finer variations, making it the most accurate model for seasonal and yearly shifts. Its strong alignment with real data, particularly during steady changes, results in a low test MSE and solidifies its position as the best model for predicting temperature trends.

The overall comparison is that the Piece-Wise Linear Model showed the best fit with segmented lines adapting to localised changes, making it the most accurate choice for both gradual and sudden shifts. The Piece-Wise Constant Model provides a good fit for spotting sudden changes, represented by clear steps, but struggles with gradual temperature variations. The Polynomial Model offers a smooth curve that captures the general trend but lacks the ability to respond to sudden or localised fluctuations effectively.

This comparison highlights the unique strengths and weaknesses of each model. The adaptability of the Piece-Wise Linear Model explains why it performed the best in accuracy metrics like MSE. These observations give a clear picture of how each model behaves when applied to real-world temperature forecasting.

The model evaluation of predictive performance of each model was evaluated using the Mean Squared Error on the test set which measures the average squared difference between observed and predicted temperatures. A lower MSE means better accuracy.

| Model | MSE |
|---|---|
| Polynomial | 0.023243 |
| Piece-Wise Constant | 0.020550 |
| Piece-Wise Linear | 0.017967 |

Table 2: Test MSE values for each basis function model

As mentioned in Table 2, the Polynomial Model with MSE of 0.023243 captured the overall temperature trend striking a balance between complexity and accuracy. However, it did not handle sudden temperature changes well though it shows full rank. The Piece-Wise Constant Model with MSE of 0.020550 effectively detected abrupt changes by splitting the data into segments with constant predictions. It performed better than the polynomial model but struggled with gradual shifts despite having full rank, lacking flexibility for continuous changes. Whereas the Piece-Wise Linear Model with MSE of 0.017967 was the top performer effectively modelling both sudden and gradual shifts in the temperature data. Its flexibility resulted in the lowest test MSE, indicating superior predictive accuracy while also showing full rank.

Based on the best-performing model (Piece-Wise Linear with k = 27), temperature forecasts were generated for the final three months of 2024:

| Month (2024) | Predicted Temperature (°C) |
|---|---|
| October | 13.86 |
| November | 12.77 |
| December | 12.10 |

Table 3: Temperature forecasts for October, November, and December 2024 using the Piece-Wise Linear model

The predictions in Table 3 indicate a gradual decrease in temperature as the year progresses, consistent with expected seasonal cooling patterns. The piece-wise linear model's ability to segment time into linear trends made it especially effective at capturing this seasonal change.

The results showed that the Piece-Wise Linear Model with k = 27 breakpoints performed the best compared to other models. It had the lowest test MSE (0.017967) which means it was most accurate at capturing the complexities in temperature data including both global trends and localised fluctuations. In contrast, the Piece-Wise Constant Model, which uses a step-like pattern, was a full rank model but struggled to handle changes with more gradual shifts which resulted in a higher MSE and lower accuracy than the Piece-Wise Linear Model. Similarly, the Polynomial Model was a full rank model that gave a smooth overview of the data, but missed sharp changes, making it slightly less accurate and resulting in a higher MSE. Overall, the Piece-Wise Linear Model found the right balance between being detailed enough to catch small changes and simple enough to avoid overfitting. This made it the best choice for predicting global temperatures, as it successfully handled both long-term patterns and short-term variations, making it a dependable model for future forecasts.

## Conclusion

This report aimed to predict global surface air temperatures from 1940 to 2024 using three models: Polynomial, Piece-Wise Constant, and Piece-Wise Linear. The results showed that the Piece-Wise Linear Model performed best, achieving the lowest MSE by accurately handling both gradual and sudden temperature changes. The Piece-Wise Constant Model effectively captured sudden shifts with its step-like pattern but struggled with gradual changes, leading to a slightly higher MSE. The Polynomial Model provided a smooth overall trend but missed abrupt jumps, making it less accurate than the piece-wise models. These findings underscore the importance of flexible models for analysing complex data like global temperature trends.

While the study provided valuable insights, there are a few limitations to consider. The study relied on monthly temperature data to represent long-term trends, which may not fully capture short-term weather changes or extreme events, potentially affecting accuracy. Additionally, the models depend heavily on selecting the right settings, incorrect choices can lead to overfitting or underfitting, reducing reliability. Moreover, only year and month were used as predictors, excluding other significant factors like $CO_2$ levels, volcanic activity, and ocean cycles, which could limit the ability of the models to fully explain temperature trends. Finally, as the models were trained on historical data up to 2024, their performance in predicting future conditions remains uncertain considering ongoing climate change.

To enhance the accuracy of future temperature predictions, several improvements could be considered for extending this study. Adding more predictors, such as greenhouse gas levels, sunlight exposure, and ocean temperatures, could help models better understand factors affecting global temperatures. Moreover, exploring advanced models, like those based on deep learning techniques such as Random Forests and Neural Networks, might provide more precise predictions by capturing complexities more effectively. Expanding the study to examine specific regions and seasonal patterns could offer a deeper understanding of how temperatures vary across different areas and times of the year. Additionally, using regularisation techniques like regression could prevent the models from becoming overly complex. In conclusion, while this study has shown the effectiveness of basis function models for temperature prediction and incorporating more sophisticated methods could lead to even more accurate and insightful climate forecasts.

Generative AI (ChatGPT) was used for reframing, editing the final report and taking suggestions for the prediction code.