Can monthly rainfall be predicted using forecasted minimum and maximum temperatures?

Comparing Linear Regression, Random Forest and Support Vector Regression Models

Introduction

This project investigates whether monthly rainfall can be predicted using forecasted minimum and maximum temperatures. Reliable rainfall forecasting is essential for planning in agriculture, flood planning, and water resource allocation across the world. Using data from three Welsh weather stations, containing the monthly records of the minimum and maximum temperatures, and the number of frost days – three supervised regressors were implemented – Linear Regression, Random Forest and Support Vector Regression. This poster presents the modelling process, key results and the overall findings when comparing the three models across different training setups.

Methodology

Data & Features

Monthly records from three Welsh weather stations were used. Features (independent variables) include the month, minimum and maximum temperature and the number of frost days.

Train Test Setups

1. **Cross station generalisation (2 train / 1 test).**
   Train on two stations and test on the third, rotating so each station serves as a test once. This measures the transferability to unseen geographical regions, assessing the impact of unseen microclimates.
2. **Within station performance (single-station 80/20 split).**
   Each station's data is split into 80% for testing and 20% for training. This establishes a station specific baseline when train and test come from the same baseline, isolating model fit from cross site variability.
3. **Pooled dataset (all stations merged, 80/20 split).**
   Combine all station data together, then perform a single 80/20 train test split. This increases sample size, exposing models to a variety of data, reducing over fitting to any single location's climate.

Validation

In setups 2 and 3, the unseen 20% split serves as the validation set. As for setup 1, the entirely unseen third station is the validation set.

Models

Model	Reasoning
Linear Regression	A low complexity, linear baseline - useful for checking whether a simple relationship exists.
Random Forests	Handles non-linear patterns and interaction of features. Often effective at environmental prediction.
Support Vector Regression (SVR)	Kernel based approach can capture more complex relationships if the temperature-rainfall relationship is non-linear.

Hyperparameters

Linear Regression was used with default parameters as a baseline. Random Forest used 300 trees to provide stable performance without excessive computation. SVR used an RBF kernel with $C = 100$, $\gamma = 0.1$ and $\epsilon = 1$, chosen through light exploratory tuning to balance smoothness and model flexibility.

Evaluation

Metric	Description
Mean Squared Error (MSE)	Average squared difference between predicted and actual values. Lower values indicate better performance. Useful for comparing the magnitude in error.
Coefficient of Determination (R^2)	Proportion of variance in the target variable explained by the model. Values closer to 1 imply better fit. Useful metric to capture the goodness of fit.

Results

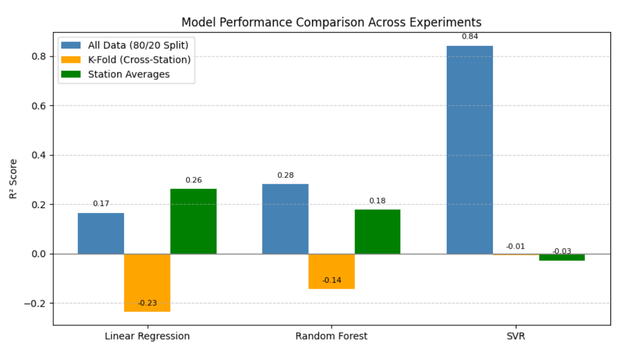

Figure 1: Model performance comparison based on R^2 scores across different experimental setups.

Cross-Station Results (Setup 1)

All models produced low or negative R^2 scores, indicating consistently poor performance when tested on an unseen weather station. This setup yielded the lowest results across all three models, showing that none of them generalised effectively under this condition.

Within-Station Results (Setup 2)

Performance improved on within station 80/20 splits, with each model achieving higher R^2 scores compared to the cross station setup. Support Vector Regression and Random Forests generally obtained stronger station specific results than Linear Regression, which remained the weakest performer across all stations.

Pooled Dataset Results (Setup 3)

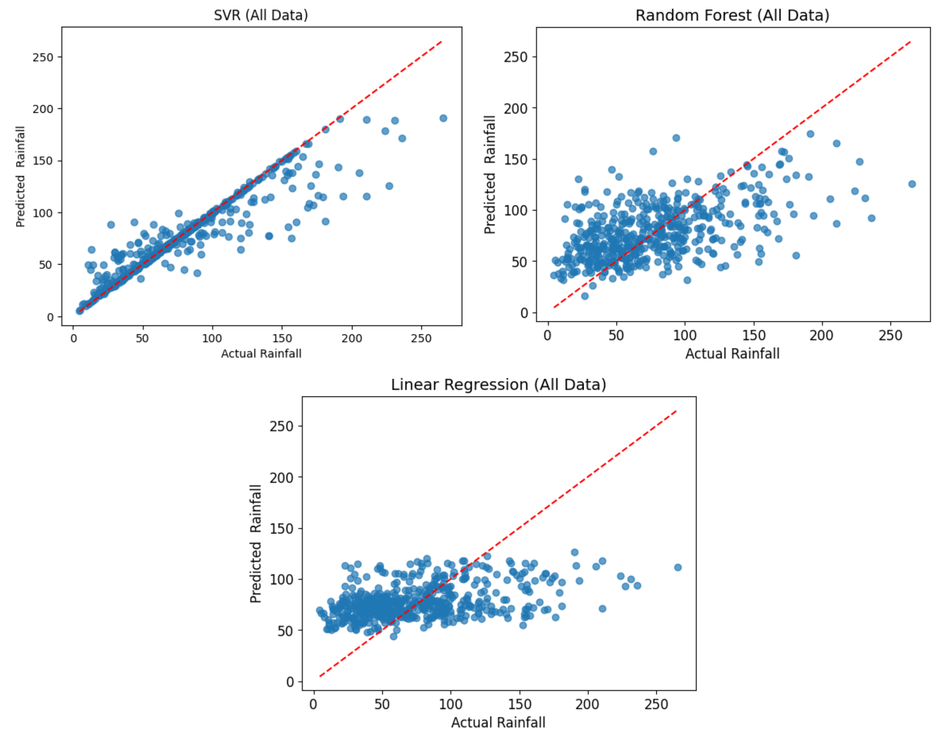

Figure 2: Model predicted values vs actual rainfall values across experimental setup (2)

When trained on the pooled dataset using an 80/20 split, Support Vector Regression achieved the highest R^2 score overall. Random Forests delivered moderate performance, while Linear Regression remained the weakest model. The SVR scatter plot showed predictions clustered closely to the ideal $y = x$ line, with a number of noticeable outliers. Whereas, Linear Regression and Random Forest overall showed a higher variance in their predictions, often straying significantly from the ideal line of perfect prediction.

Error Comparison (MSE Results)

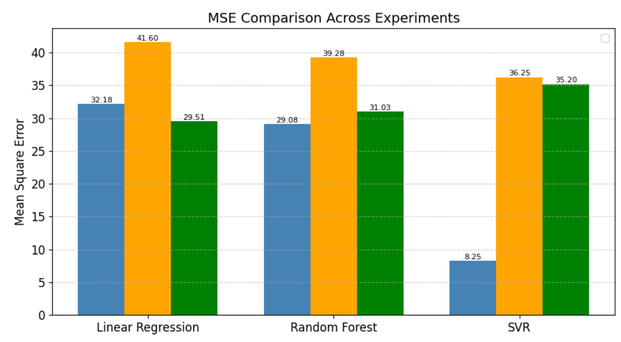

Figure 3: Mean square error comparison between models reflecting magnitude of error across experimental setups.

Across all experiments, SVR produced the lowest mean squared error when trained on the pooled dataset. Random Forests and Linear Regression showed noticeably higher errors, and the cross station setup resulted in the highest MSE values for every model.

Conclusions

All models failed to generalise on unseen weather stations, with consistently low or negative R^2 scores shown in the cross-station setup. This indicates that the rainfall–temperature relationship varies significantly enough across regions, leading to model underperformance, reflecting the influence of local micro-climatic effects and additional meteorological factors not captured in the limited feature set.

Performance improved when models were trained and tested on the same station, which was expected, demonstrating the highest when all stations were combined into a single dataset. Support Vector Regression benefited the most from the wider variety of data, achieving the strongest overall performance, while Random Forests showed only moderate improvement, and Linear Regression remained the weakest model.

Although SVR modelled the general temperature–rainfall pattern effectively, it tended to under predict extreme rainfall values, suggesting that additional features such as humidity, pressure, or wind would be required to improve performance on heavy rainfall events. The overall findings highlight the limitations of using temperature alone for rainfall forecasting, especially across different geographical locations.

Future work could incorporate a wider set of meteorological inputs, larger datasets, or temporal models to better capture seasonal patterns and local region (station) variability.