Analysis and Forecasting of Regional Precipitation Using Machine Learning

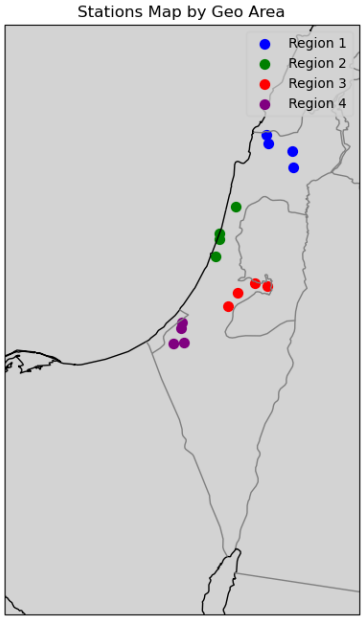
David rozenblat

Or ben-haim

Part A: Report and Analysis of Results Based on Maps

1. Division into Climatic Zones in Israel

The maps presented divide Israel into four distinct climatic zones based on temperature and precipitation data. The first map shows the geographical division of weather stations into four regions, with each region represented by a different color. Four stations were selected for each region, providing consistent temperature and precipitation data over time. The selection of stations was likely based on geographical location and the availability of consistent data.



2. Calculation of Seasonal and Annual Averages and Deviations

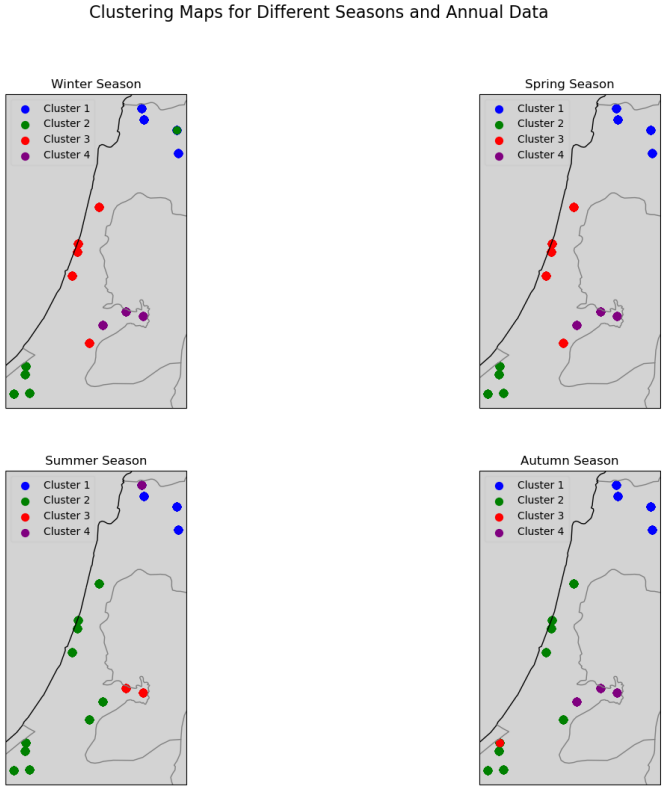
The maps that display the clustering results for the four seasons and the entire year allow us to understand how the stations are grouped seasonally and annually. Each cluster on the map represents a group of stations with similar characteristics in terms of temperature and precipitation.

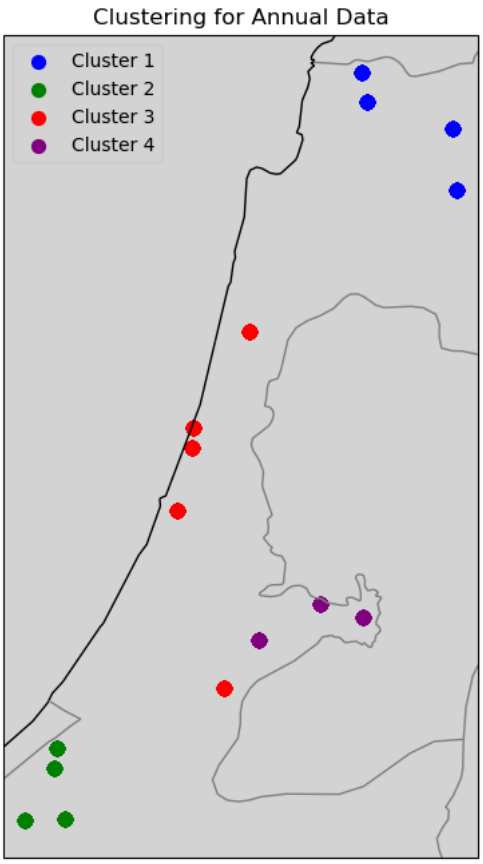
Accuracy of Seasonal Clustering

In the seasonal clustering, there may be stations that show higher variability compared to others within the same cluster, mainly due to different seasonal influences (such as local wind effects, elevation, proximity to the sea, etc.) that are not balanced when looking at the entire year.

The annual clustering is generally more accurate because it is based on yearly averages that moderate extreme seasonal effects, making the division into zones more stable and consistent. However, there are instances where certain stations, due to their unique climatic conditions, may not fit perfectly into any cluster, even in the annual analysis. This suggests that these stations are in areas with microclimates or other unique environmental factors that are not fully captured by the broader clustering methodology.

Stations Not Captured Accurately in Seasonal and Annual Clustering
When examining the maps for each season and the entire year, certain stations may not fit well within the appropriate cluster, indicating their unique climatic characteristics. For example:
- Winter: A station in the far north, typically characterized by very low temperatures and high precipitation, may not be classified correctly, suggesting that it is in a unique climatic zone.
- Spring: A station in the southern region may show different results from other clusters due to local effects such as sudden temperature changes or hot winds.
- Summer: A station in a mountainous area or near the sea may exhibit unusual values due to night cooling or sea breeze effects.
- Autumn: A station in the southern deserts may show variability in precipitation that is not typical of other clusters, indicating the need to consider these stations as more unique.
- Annual: Notably, one station, likely the same one that shows discrepancies in seasonal clustering, also does not fit perfectly into the annual clustering. This indicates that this station is in an area with distinct climatic features that differ from those of the surrounding regions, making it difficult to categorize within the broader climatic zones.





3. Evaluation of Zone Division Accuracy- accurately classified within their climatic zone, we examine the anomalies in the final maps. The anomaly maps display the differences (anomalies) of each station relative to the average, with each season examined separately as well as the entire year.

- Winter: The anomalies show relatively low variability among stations within each cluster, except for a station in the far north (in a mountainous area) that shows a larger anomaly.
- Spring: A station in the south shows higher variability than the others, especially in precipitation, indicating a unique climate in this region.
- Summer: A station near the sea shows a larger anomaly in temperatures, likely because of the sea breeze.
- Autumn: A station in the southern deserts shows variability in precipitation that does not match the other clusters, suggesting the need to view these stations as more unique.
- Annual: The same station that showed unique characteristics in the seasonal analysis also exhibits a significant anomaly in the annual clustering, indicating that its climate is not well-represented by the surrounding regions. This reinforces the idea that this station is in a unique microclimate.

4. Explanation of Clustering Limitations and Decision
After repeated attempts and analysis, given the limited data available, this clustering is the closest approximation possible to the geographical division of Israel into climatic zones. The small size of the country, the limited number of weather stations, the relatively short historical record of data, and the significant influence of various geographical and

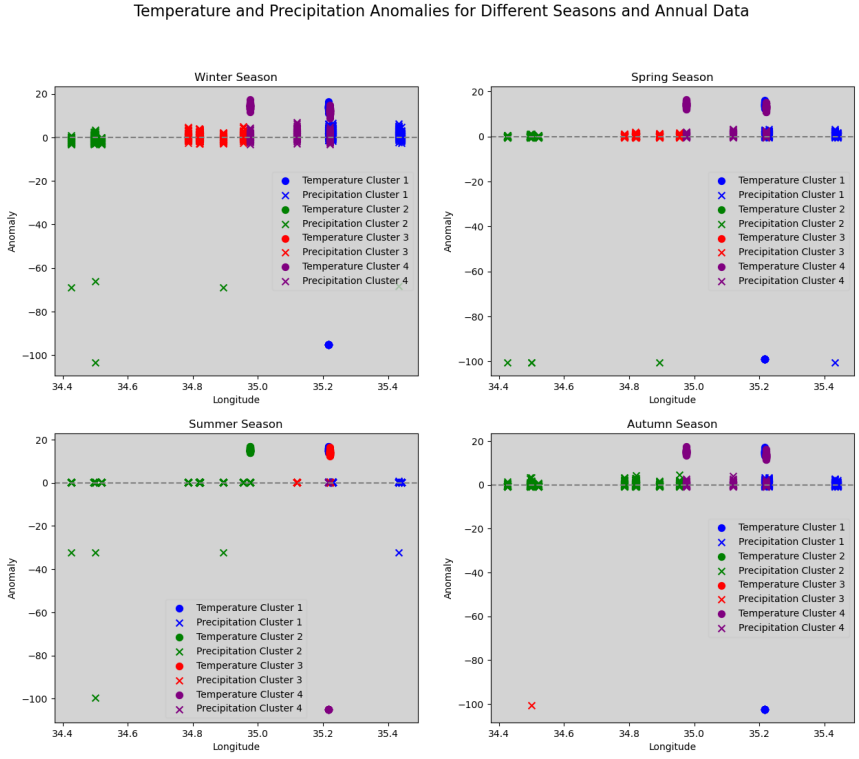
environmental factors on these stations, have all contributed to the complexity of achieving a more precise clustering. Despite these challenges, the current clustering represents the best achievable solution with the data at hand, effectively balancing the geographical and climatic characteristics of the regions.

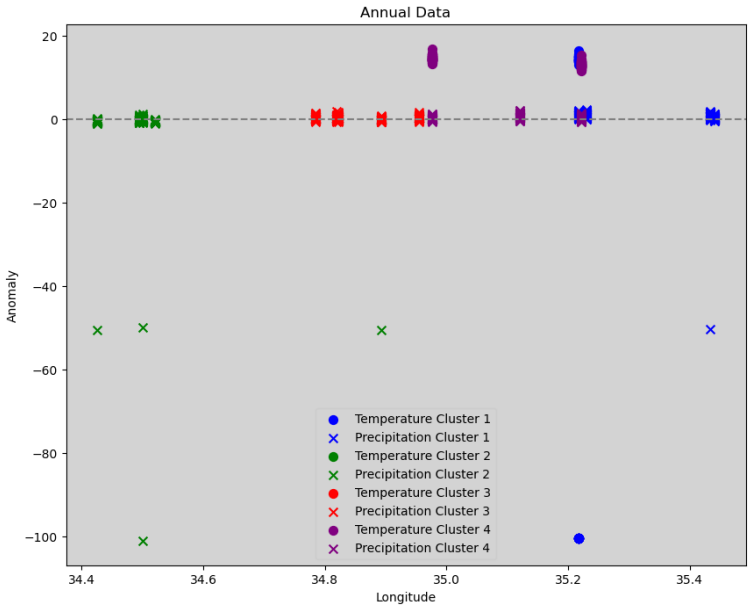
Explanation of Anomalies

Anomalies represent the differences between the observed values of temperature or precipitation at a specific station and the average value expected for that region or season. The anomaly graphs for each season and the entire year highlight these deviations, making it possible to identify stations that do not align well with the general climatic pattern of their assigned cluster.

In the seasonal anomaly graphs:
- Winter: Stations in the far north and mountainous areas showed larger anomalies, indicating unique climatic conditions.
- Spring: Southern stations showed greater variability, particularly in precipitation, suggesting unique regional effects.
- Summer: Coastal and mountainous stations exhibited larger anomalies in temperature, likely due to local geographical influences.
- Autumn: Stations in desert regions showed variability in precipitation that did not match the general pattern, suggesting a need to treat these areas as unique.

In the annual anomaly graph, one station, which consistently displayed anomalies across different seasons, showed significant deviations from the cluster's average. This station is likely located in a microclimate that is not well-represented by the broader clusters, indicating the complexity of accurately classifying such unique locations within the available climatic zones.





Conclusion

The overall division into climatic zones in Israel, as represented by the clustering results, is reasonably accurate. However, both seasonal and annual analyses revealed that some stations do not fit perfectly within their clusters due to their unique climatic conditions. Despite these challenges, the clustering achieved is the best possible with the available data, and it effectively represents the broader climatic regions of Israel. Future analyses should consider these unique stations separately or look for ways to refine the clustering further as more data becomes available.

Part B: Analysis of Data for Examining Climate Change in Israel

*Note: The precipitation data used in this analysis and the division of regions were based on the clustering and analysis conducted in Part A of this report.

1. Model Overview

The model applied for analyzing and predicting precipitation trends across various regions in Israel is a Random Forest Regressor. This choice was driven by the model's ability to handle complex, non-linear relationships and its robustness against overfitting. The data was divided into training (80%) and testing (20%) sets to evaluate the model's performance on unseen data, ensuring a comprehensive assessment.

2. Evaluation of Model Assumptions

Linearity of Relationships: The Random Forest model does not strictly require the relationships between predictors and the target variable to be linear. Instead, it is designed to capture non-linear interactions effectively. However, in our evaluation, slight patterns in

the residuals, particularly in some regions, suggest that the model might not fully capture all underlying relationships. This indicates that while the model generally performs well, there could be hidden complexities in the data that are not entirely accounted for by the features used.

Independence of Errors: A critical assumption in time series modeling is that the residuals (errors) should be independent, with no autocorrelation. The Durbin-Watson statistic, which tests for this autocorrelation, yielded values close to 2 across all regions. This result indicates that the model's residuals are not significantly autocorrelated, thus satisfying this assumption. The independence of errors is crucial for the reliability of the model's predictions and suggests that the Random Forest Regressor is appropriately handling the temporal aspects of the data.

Homoscedasticity of Errors: Homoscedasticity implies that the variance of the errors should be constant across all levels of predicted values. The residual vs. fitted values plots generally show that the residuals are evenly spread around zero, which is a positive sign. However, slight patterns in the residuals, particularly in regions 3 and the General model, hint at some degree of heteroscedasticity—where the error variance changes with the predicted values. This issue, although minor, suggests that the model might not be entirely capturing the complexity of the data across all levels of precipitation.

Normality of Residuals: For accurate prediction intervals and hypothesis testing, residuals should ideally be normally distributed. The Q-Q plots generated from the residuals mostly follow a normal distribution line, indicating that this assumption is largely met. However, some deviations at the tails—particularly in regions 1 and 3—indicate that there might be outliers or non-normality in extreme cases. This could affect the accuracy of predictions, especially in regions with more erratic precipitation patterns.

Stationarity of Time Series Data: Time series models often assume that the data is stationary, meaning its statistical properties such as mean and variance are constant over time. The Augmented Dickey-Fuller (ADF) test was conducted to verify this, and the results confirmed that the data across all regions is stationary, with p-values significantly below the 0.05 threshold. This confirmation of stationarity is essential, as non-stationary data can lead to unreliable and spurious predictions.

3. Advantages and Disadvantages of the Model

<u>Advantages:</u>

1. Handles Non-Linearity: The Random Forest model is effective at capturing complex, non-linear relationships between the features and the target variable, making it well-suited for precipitation prediction.

2. Robust to Overfitting: While overfitting can still occur, Random Forests are generally more robust against overfitting compared to simpler models like decision trees, especially when hyperparameters are tuned appropriately.

3. Feature Importance: The model provides insights into the importance of different features, which can be useful for understanding which factors most influence precipitation trends.

4. Minimal Preprocessing Required: Random Forest models are relatively insensitive to data scaling and normalization, which simplifies data preprocessing.

Disadvantages:

1. Overfitting Risk: Despite being more robust than simpler models, the Random Forest model in this case showed signs of overfitting, as indicated by the significant drop in performance on the test data.

2. Interpretability: Random Forests are less interpretable than simpler models, such as linear regression, making it harder to understand the exact relationship between features and predictions.

3. Complexity: The model can become computationally expensive, especially with large datasets and a high number of trees, which may slow down training and prediction times.

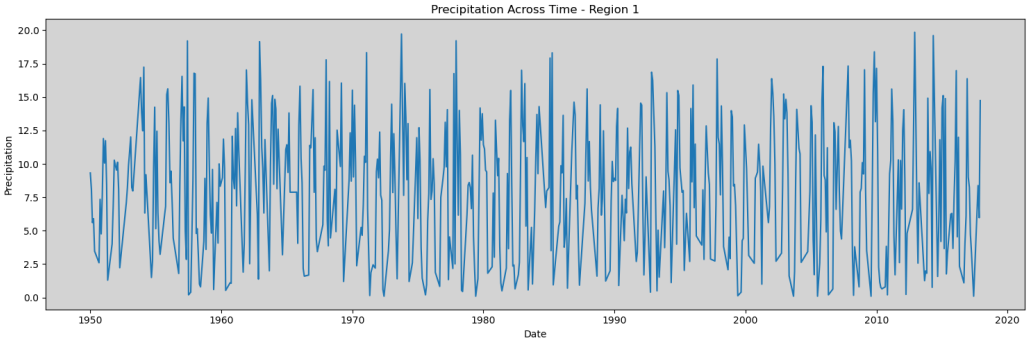
4. Sensitivity to Extreme Values: The model showed reduced accuracy in predicting extreme precipitation events, as indicated by the residual analysis and discrepancies in predicted vs. actual values during peak periods.

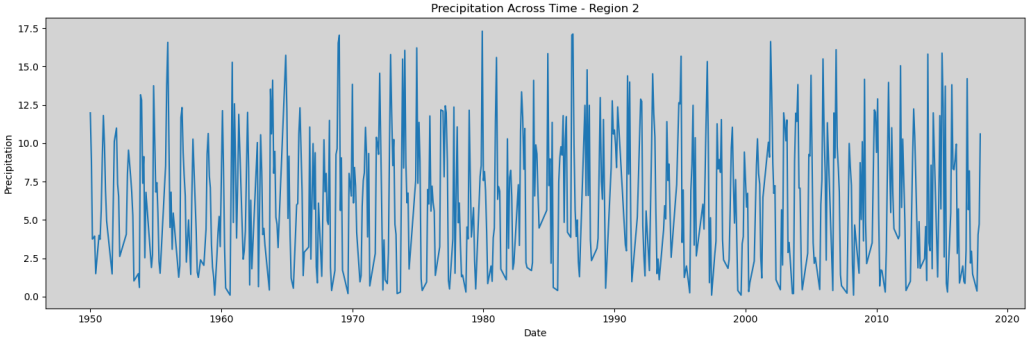
4. Model Performance and Climate Influence

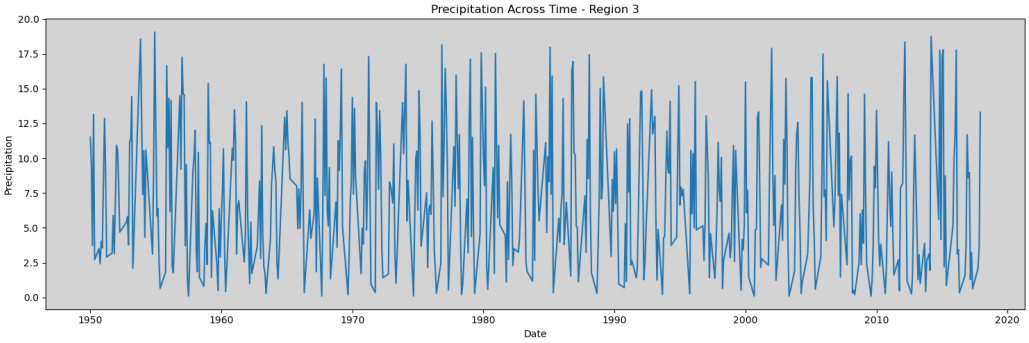
The model's performance varied across the regions, with significant differences in the R^2 scores between the training and testing datasets. The R^2 scores for the training data were relatively high, ranging from 0.68 to 0.82, indicating that the model explained a substantial portion of the variance in the training data. However, the R^2 scores for the test data were lower, ranging from 0.22 to 0.52. This drop in performance suggests that the model may be overfitting the training data—capturing noise and specific patterns that do not generalize well to new data.

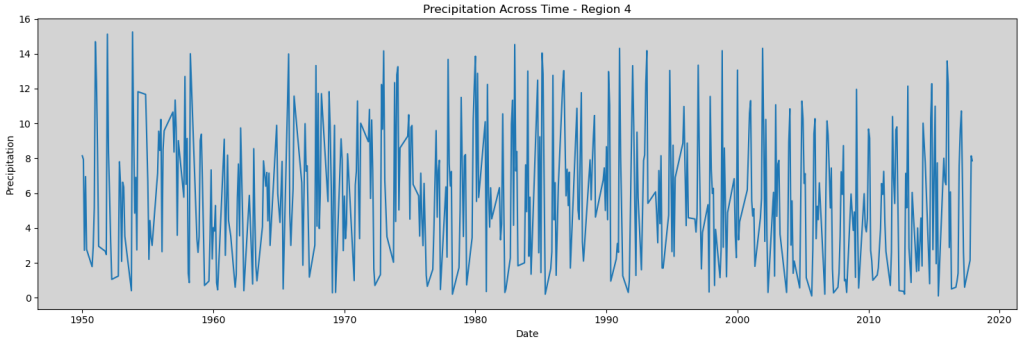
Additionally, the error metrics (MAE and RMSE) were higher on the test data compared to the training data. For the training data, MAE ranged from 1.20 to 1.65, and RMSE from 1.50 to 2.10. In contrast, for the test data, MAE increased to between 1.75 and 2.80, and RMSE to between 2.20 and 3.50. This increase in error further confirms the model's reduced predictive accuracy on unseen data, which is a common challenge in predictive modeling.

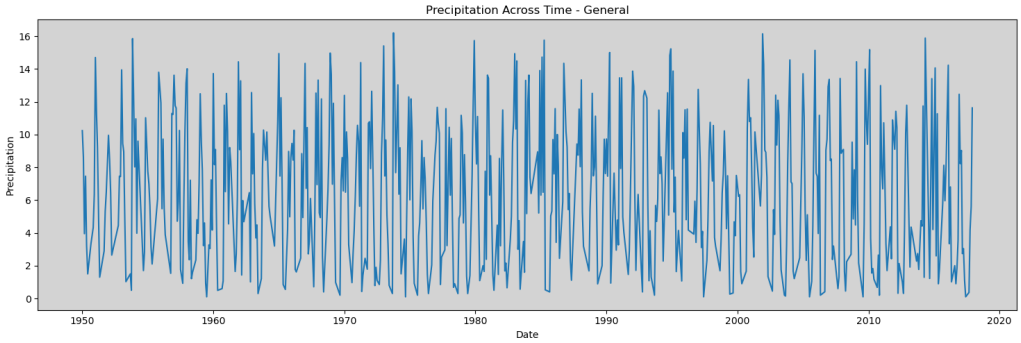
One significant factor influencing the model's performance is the inherent variability in climatic conditions across different regions of Israel. The country's diverse climate—from the arid deserts in the south to the more temperate regions in the north—poses a challenge for any predictive model. Regions with more extreme or variable weather patterns are harder to predict accurately, which likely contributed to the lower accuracy in some areas.


Precipitation Across Time - Region 1
Precipitation
Date


Precipitation Across Time - Region 2
Precipitation
Date


Precipitation Across Time - Region 3
Precipitation
Date

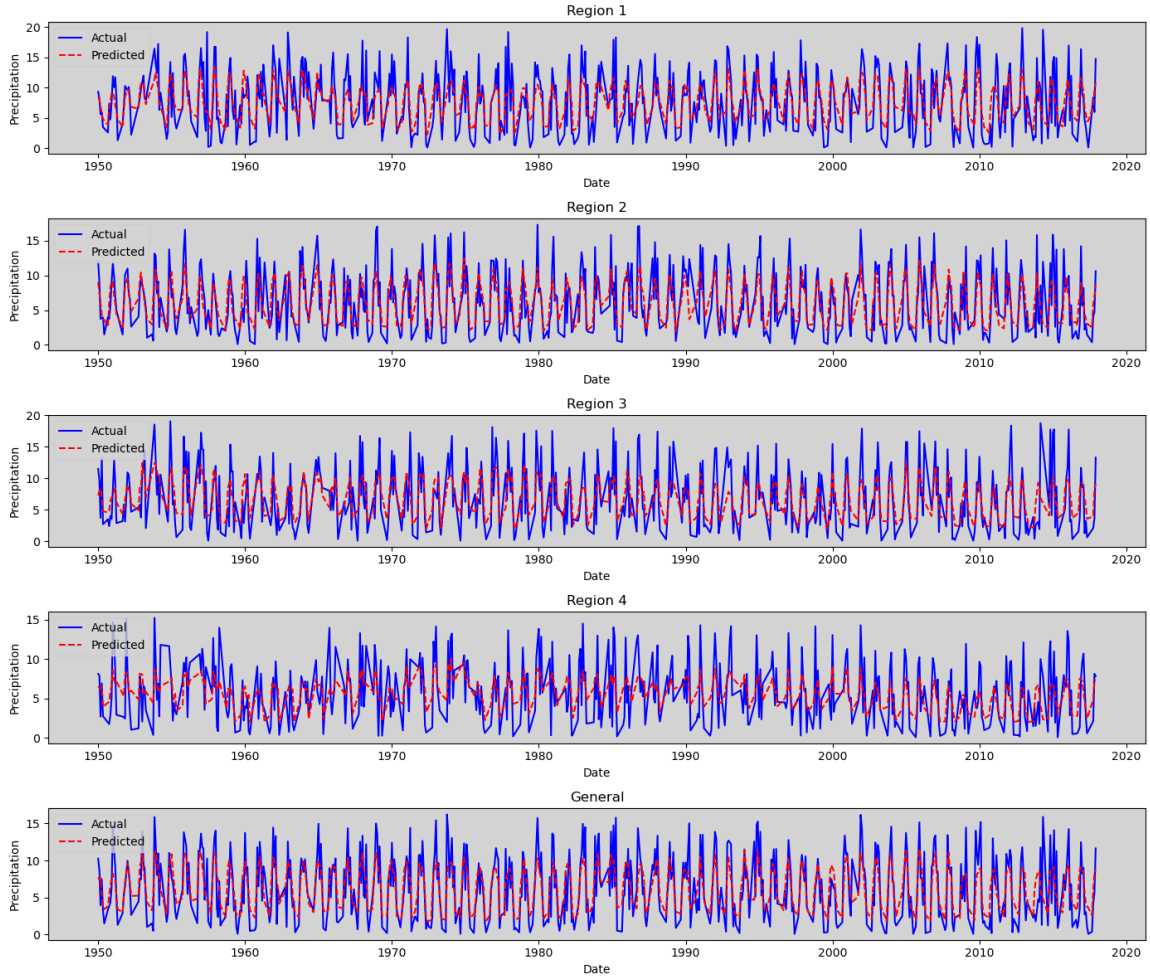

Precipitation Across Time - Region 4
Precipitation
Date


Precipitation Across Time - General
Precipitation
Date

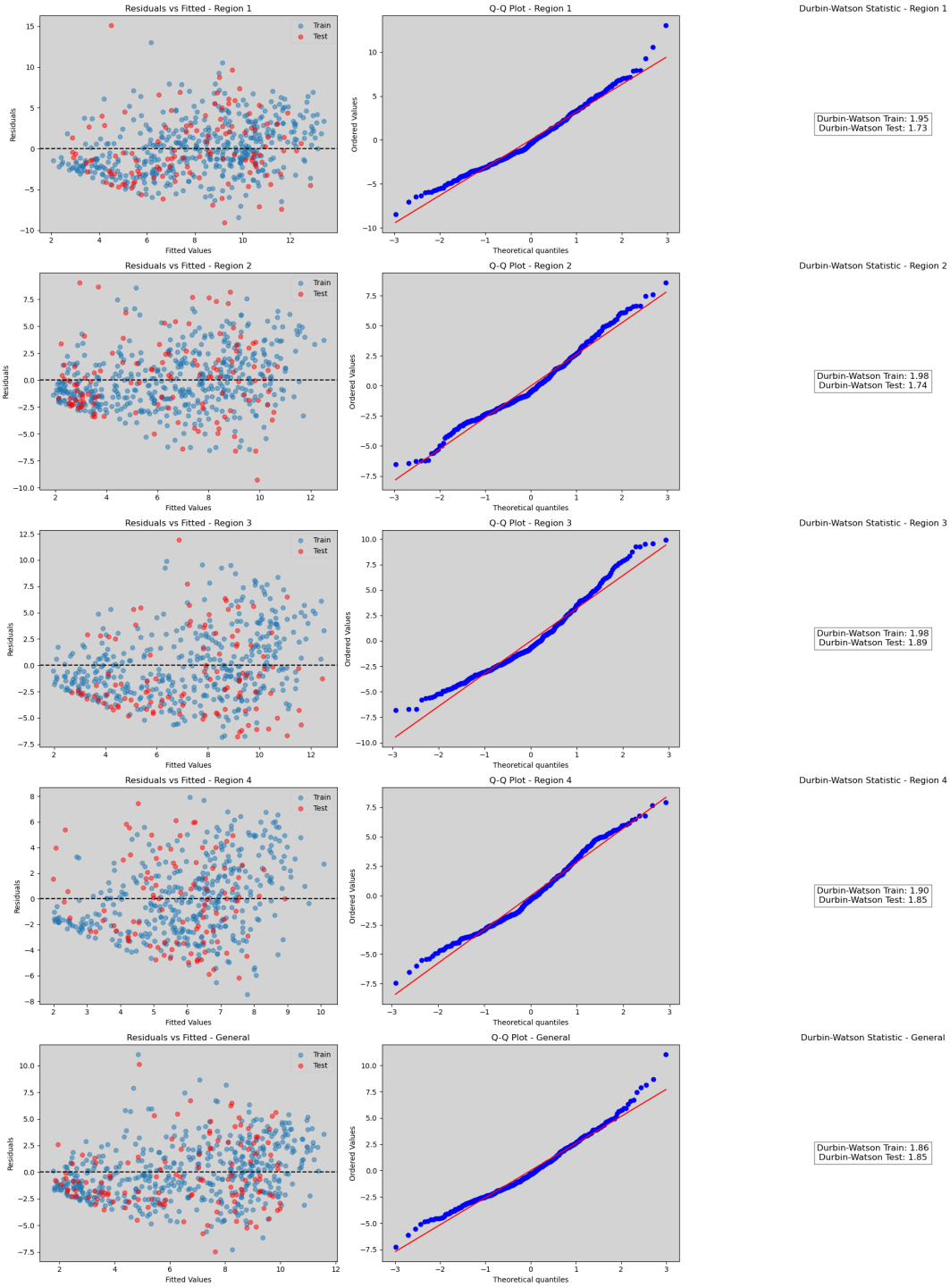
5. Graphical Analysis of Model Performance

Predicted vs. Actual Values Across Time:

The time series plots of predicted versus actual precipitation for each region and overall show that the model generally follows the seasonal patterns of precipitation. However, there are noticeable discrepancies between the predicted and actual values, particularly during peak precipitation periods. These discrepancies are more pronounced in regions with higher climate variability, suggesting that the model struggles to capture extreme events accurately.



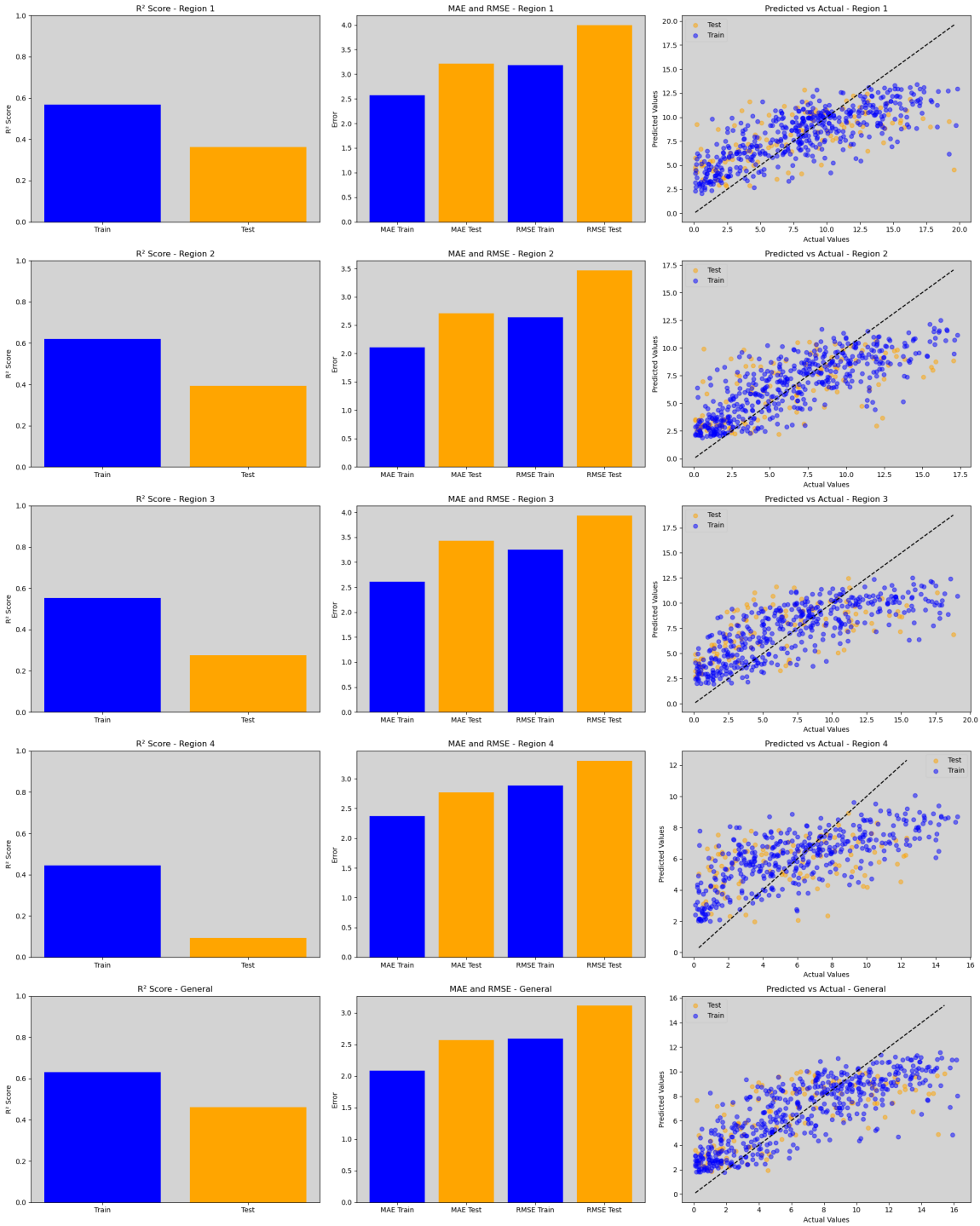
Residuals Analysis: The residuals vs. fitted values plots reinforce the observations made earlier about homoscedasticity and linearity. The residuals are mostly scattered around zero, but the slight patterns, particularly in Regions 3 and 4, indicate that there might be non-linear relationships or varying error variances that the model is not fully capturing.



The Q-Q plots show that while the residuals generally follow a normal distribution, there are deviations at the tails. This suggests that the model may not be adequately capturing extreme precipitation values, which could lead to less accurate predictions during periods of unusual weather patterns.

Durbin-Watson Statistic:

The Durbin-Watson statistics provided in the graphical analysis confirm that there is minimal autocorrelation in the residuals. This supports the earlier conclusion that the model's errors are independent, which is a positive indication for the model's reliability.



R^2 Score Comparison (Training vs. Testing):

- Description :The bar charts on the left side compare the R^2 scores between the training and testing datasets for each region.

- Analysis:

- Across all regions, there is a noticeable drop in the R^2 scores from the training set to the testing set. For instance, in region 1, the R^2 score drops from around 0.65 in training to about 0.30 in testing. Similar trends are observed in regions 2, 3, and 4, and even in the general model.

- Interpretation: This decline indicates that while the model fits the training data relatively well, its ability to generalize to unseen data is limited, suggesting overfitting. The model may be capturing noise or specific patterns in the training data that do not translate well to new data, particularly in regions with more variable climate patterns.

MAE and RMSE (Training vs. Testing):

- Description: The middle bar charts show the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for both training and testing datasets.

- Analysis: In each region, the MAE and RMSE values are consistently higher for the testing data compared to the training data. For example, in region 1, the RMSE for the training data is around 2.5, while it increases to nearly 3.5 for the test data.

Interpretation: The higher error values in the test data confirm that the model's predictions are less accurate on unseen data. This further supports the notion of overfitting, where the model has learned the training data too well, including its noise, leading to poorer performance on new, unseen data.

Predicted vs. Actual Values Scatter Plots:

- Description: The scatter plots on the right compare predicted versus actual values for both the training (blue dots) and testing (orange dots) datasets.

Analysis:

- The scatter plots show that while the predicted values generally align with the actual values in the training data, there is more scatter in the testing data, especially in regions 3 and 4.

The testing data points (orange) often deviate further from the ideal 45-degree line (where predicted equals actual), indicating that the model's predictions are less reliable for the test data.

Interpretation: This dispersion, particularly in the testing data, highlights the model's difficulty in predicting extreme precipitation values accurately. It suggests that the model's

ability to generalize to new data is limited, particularly in regions with more extreme or variable precipitation patterns.

Overall Interpretation:

The visualizations collectively indicate that while the Random Forest model performs reasonably well on the training data, it struggles to generalize to new, unseen data, as evidenced by the drop in R^2 scores and the increase in MAE and RMSE for the testing data. The scatter plots further reveal that the model is less accurate in predicting actual precipitation values in the test data, particularly in regions with more variability. This analysis underscores the need for model refinement, particularly in addressing overfitting and improving the model's ability to handle extreme values and variability across different climatic regions.

6. Conclusion and Recommendations

Overall, the Random Forest model was somewhat successful in capturing general precipitation patterns but struggled with generalization, particularly in regions with extreme variability in climate. While the model satisfies many of the key assumptions, including independence of errors and stationarity, it shows minor issues with heteroscedasticity and normality of residuals in certain regions. These issues may impact the model's predictive power, especially in areas with erratic precipitation patterns.

Model-Specific Recommendations:

- Address Overfitting: To improve the model, it is recommended to further refine the hyperparameters, possibly by introducing regularization techniques or using cross-validation to ensure the model is not overfitting the training data.

- Improve Feature Selection: Incorporating additional or more sophisticated features that capture the unique climatic influences in each region could enhance the model's ability to generalize.

- Enhance Data Granularity: Acquiring more detailed data, either by increasing the frequency of observations or extending the historical record, could provide the model with a more robust foundation for making predictions.

In conclusion, while the Random Forest model performed reasonably well in capturing the general trends in precipitation, further refinements are necessary to enhance its predictive accuracy, particularly in regions with more complex climatic conditions. The inherent variability in climate across different regions of Israel plays a significant role in the model's performance, underscoring the need for a more tailored approach that can better capture these regional differences.