LADPS Next-day Maximum Temperature Prediction using Machine Learning

Zelin Wang

Department of Atmospheric and Oceanic Science, University of California Los Angeles

AOS C204: Introduction to Machine Learning for Physical Sciences

Dr. Alexander Lozinski

1. Introduction

In recent weather forecasting systems, AI and machine learning are playing a more and more important role in improving the accuracy of predictions, and even enabling pure AI-driven new models, some of which even have a higher accuracy than traditional numerical prediction models controlled by physical equations. This highlights the promising future of applying AI techniques in conventional weather forecasting.

The Local Data Assimilation and Prediction System (LDAPS) is an important numerical prediction model operated by the Korean Meteorological Administration. It is a high-resolution weather model that aims to quickly predict small-scale, local weather events across the Korean Peninsula. Like all other numerical weather forecasting models, the LDAPS is not able to accurately predict the temperature for every single day, which is partly because of the nonlinear and complex characteristics of weather systems. As Figure 1 shows, the histogram of the distribution ranges from –7°C to 7°C, and it generally exhibits a normal distribution.

In this report, I will try to apply several classic machine learning techniques to independent variables, such as the temperature of the current day, geographic auxiliary variables, and LDAPS-predicted variables, to predict the next-day maximum temperature. In the end, I will compare the prediction based on the testing data with the observed next-day temperature and the LDAPS-predicted next-day temperature, in order to evaluate the performance of each machine learning method in this context. While we don’t necessarily expect that the prediction based on these ML methods can exceed the accuracy of LDAPS, it is meaningful to evaluate which machine learning method comes closer to the LDAPS prediction or to the observed next-day temperature.

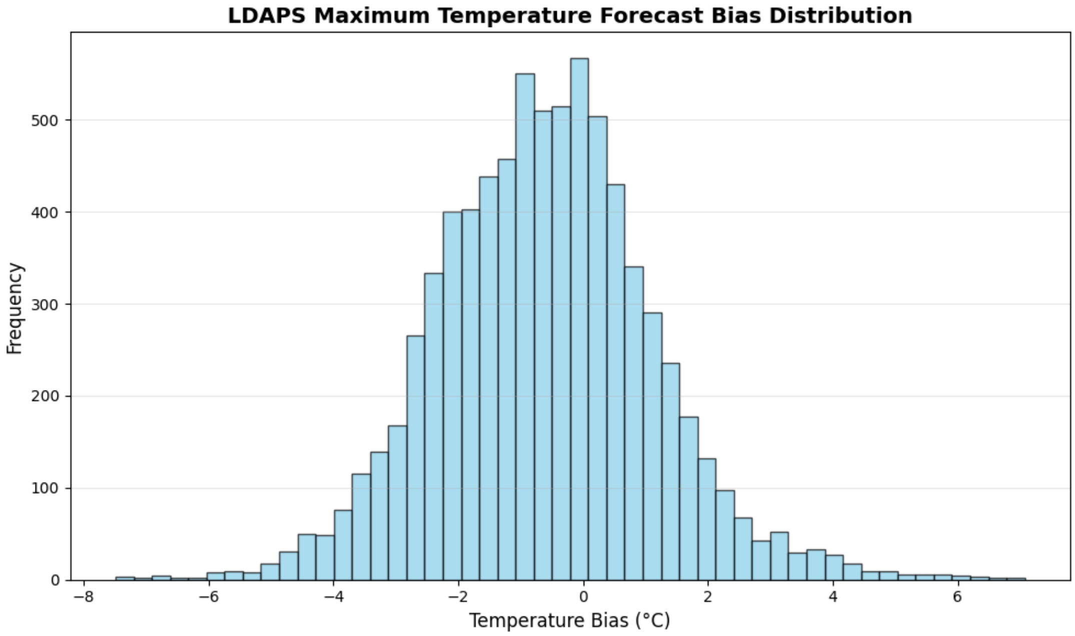


Fig. 1. Histogram of LDAPS Maximum Temperature Bias

2. Data

For this project, the LDAPS data is obtained from a publicly available dataset in UC Irvine Machine Learning Repository. The data is available in summers from 2013-2017. The input data contains the LDAPS model's next-day forecast data, in-situ maximum and minimum temperatures of present-day, and geographic auxiliary variables over Seoul, South Korea in the summer. There are two outputs (i.e. next-day maximum and minimum air temperatures) in this data, while in this project, I will be mainly predicting the maximum air temperature.

LDAPS model's next-day forecast data as input data contains: next-day LDAPS-predicted temperature, next-day average wind speed, next-day max/min relative humidity; next-day average latent heat flux; next-day average cloud cover over 4*6-hour period; next-day split average precipitation over 4*6-hour period. The five geographic auxiliary variables contain latitude, longitude, elevation, slope and solar radiation. These data will be used to predict the next-day maximum temperature.

I designed a parallel experiment, that is, in order to avoid having the final prediction overly rely on the next-day LDAPS-predicted temperature, I removed the next-day LDAPS-predicted temperature from the input data in the second experiment, to see whether the performance shows a significant change compared to the first experiment.

3. Preprocessing

First, to make sure our models work, it’s necessary to clean the data by removing the NANs. And for both experiments, we choose the target variable: next-day maximum temperature, and features.

I noticed that the dataset contains time series for all 25 stations, in case the 25 stations show significantly different characteristics, I plotted the temporal evolution of the everyday maximum temperature for every station in Fig.2, and an average of these stations. As the time is separated by each year, there are five figures representing the summer seasons from 2013 to 2017. From these figures we can observe that the station spread is generally limited, and it seems in most cases these stations follow the same increasing or decreasing trend.

And to better understand the relationships of the variables within the data, I first applied a multi-correlation between the next-day maximum temperature and different variables one by one. As Fig. 3 shows, the variables are ordered according to their correlation with next-day maximum temperature (regardless of positive or negative). It turns out that the top six variables which have the highest correlation with next-day maximum temperature are LDAPS_LH, LDAPS_Tmax_lapse, LDAPS_RHmin, Present_Tmax, LDAPS_RHmax, and LDAPS_WS respectively. And for those variables, Fig. 4 shows the specific correlation distribution and trend of these six variables with next-day maximum temperature.

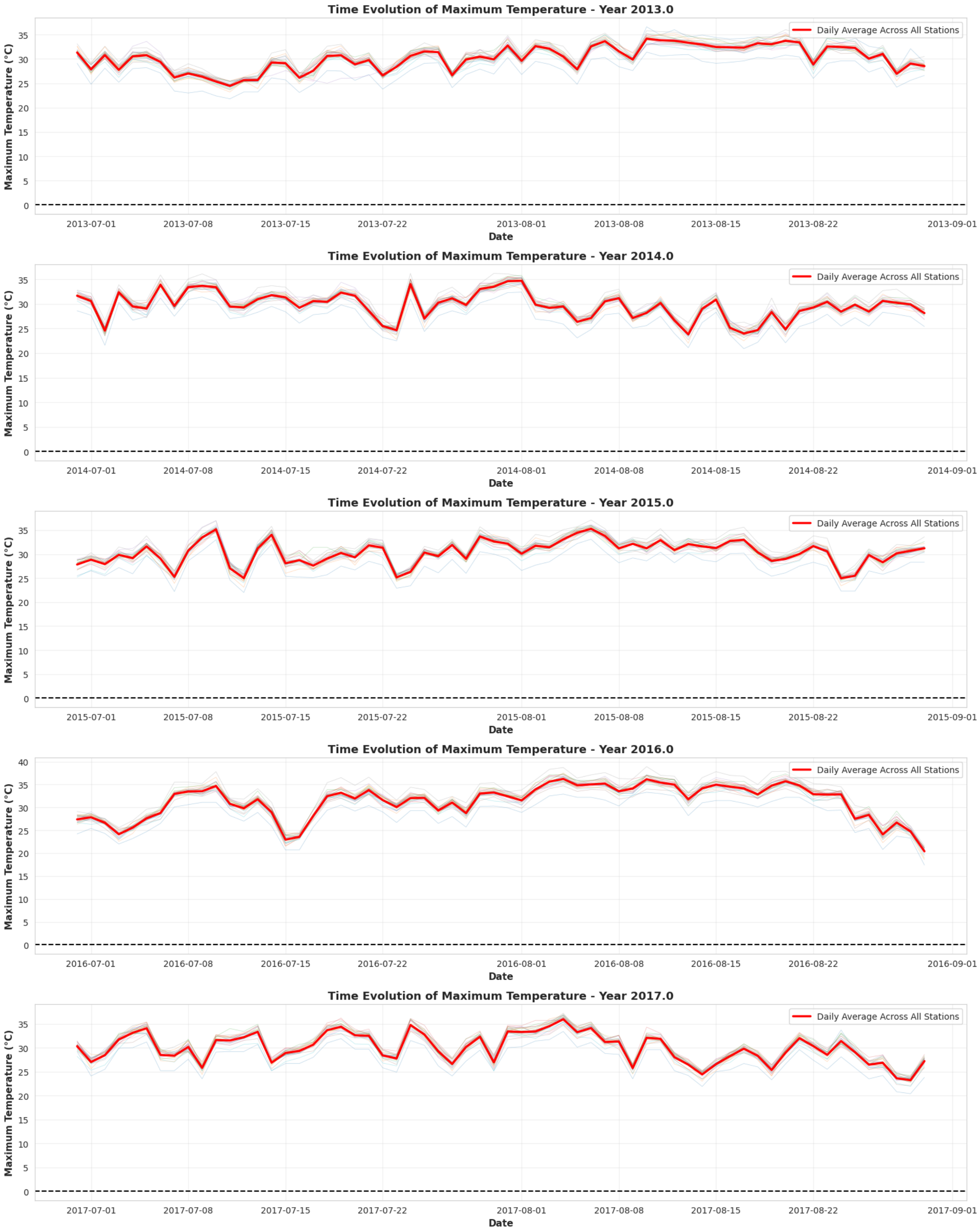


Fig. 2. The temporal evolution of the daily Maximum temperature for the 25 stations, for summers of different years from 2013-2017, respectively.

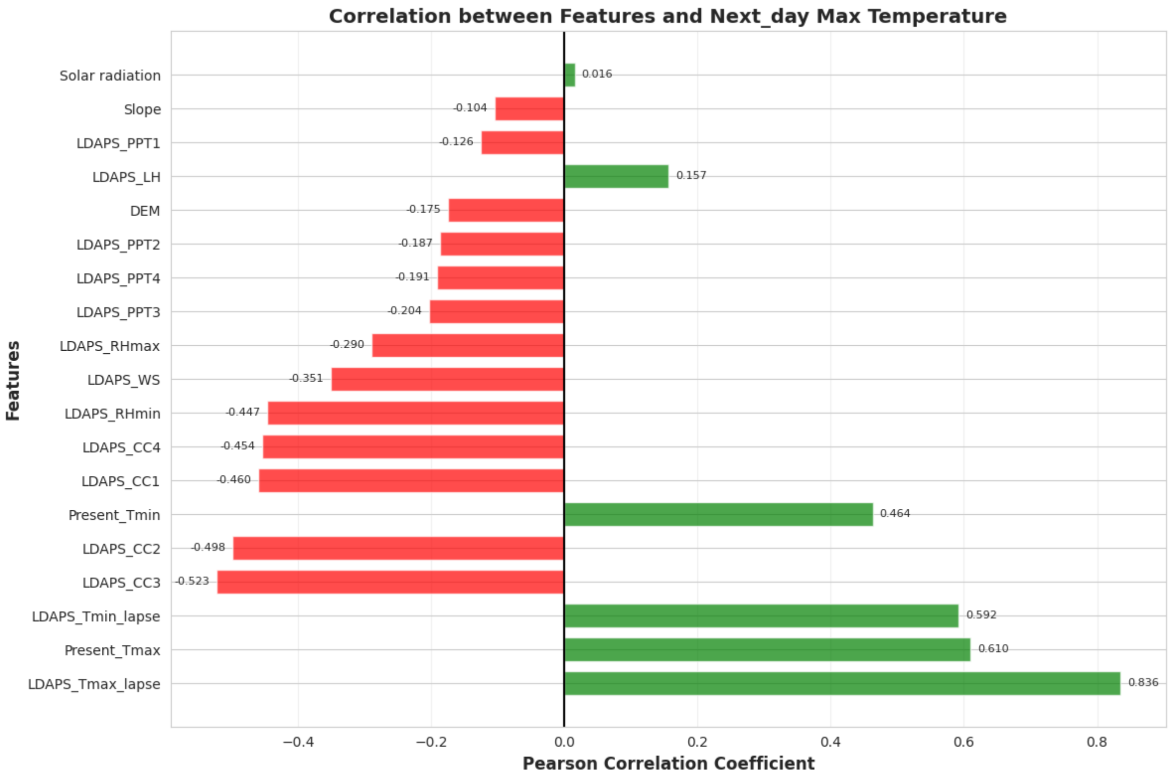


Fig. 3. Correlations between different variables and next-day maximum temperature.

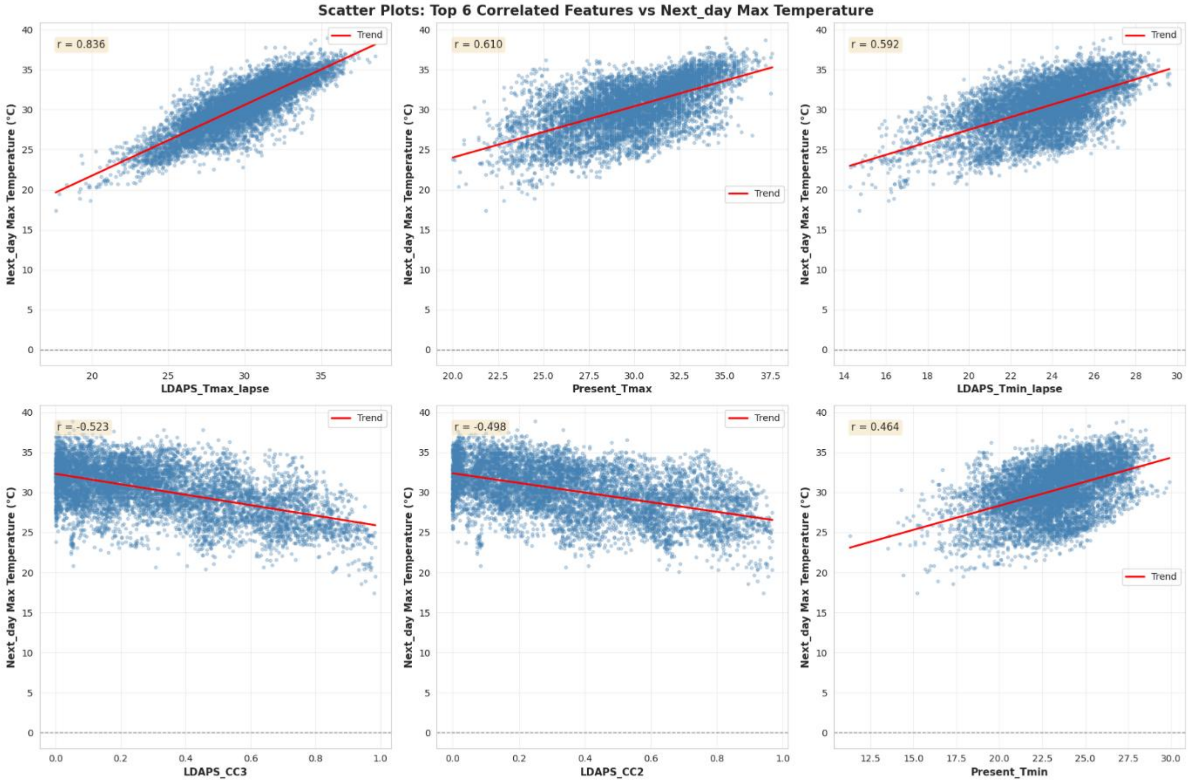


Fig. 4. Top 6 Correlated features' correlation map with Next-day Temperature.

4. Modelling

After learning about the features and their correlation with next-day max temperature in the data, we know that this is a regression task. I plan to include 4 machine learning techniques: Linear Regression, Decision Tree, and Random Forest, as well as MLP Regressor, to do the regression and compare their performance based on their Root Mean Square Error (RMSE), and then evaluate its difference from the LDAPS-predicted temperature.

Linear Regression serves as a baseline model because it is simple and easy to interpret, which helps us understand the basic linear relationships between input features and the next-day maximum temperature. Decision Tree is chosen because it can capture non-linear patterns in the data without requiring data normalization, and it can also show us which features are most important for prediction through its tree structure. Random Forest, as an ensemble method that combines multiple decision trees, usually provides better accuracy and is more robust against overfitting compared to a single decision tree. Finally, MLP Regressor (Multi-Layer Perceptron), a type of neural network, is included to test whether a more complex model can learn hidden patterns in the data that simpler models might miss. By comparing these four methods, we can evaluate which approach works best for predicting temperature.

For both experiments, the training set and testing set are split by a 70% to 30% ratio, and they are ordered based on time, meaning that earlier temperature records are used as the training set to test the most recent data, which mostly fall in 2017 and 2016.

5. Results

First, for Experiment 1, which includes the LDAPS-predicted next-day Tmax in the features, we conducted the 4 machine learning models mentioned above, and their behavior is shown in Fig. 5. Their performance based on mean absolute error (MAE), rooted mean squared error (RMSE), and R^2 is displayed in Table 1. Here, MAE and RMSE measure the average size of the prediction errors (MAE uses absolute values, RMSE penalizes larger errors more strongly). Therefore, lower MAE and RMSE indicate better predictions. R^2 measures how much of the variance in observations is explained by the model, so higher R^2 indicates better performance.

Based on these results, we find that Linear Regression appears to be the best model that has the predictions closest to the observations, given its lowest MAE and RMSE, and highest R^2. Following that are Random Forest and MLP Regressor, and the worst modeling performance is from the Decision Tree.

And given by Fig. 6, the tree structure of the decision tree and the first tree in the random forest, as well as Fig. 7, the top 10 important features in these models to predict the next-day max temperature, we can find out an obvious characteristic: both the decision tree and random forest highly rely on the LDAPS-predicted next-day Tmax, because this feature serves as the first-level split for both models and leads the other features very strongly in Fig. 7.

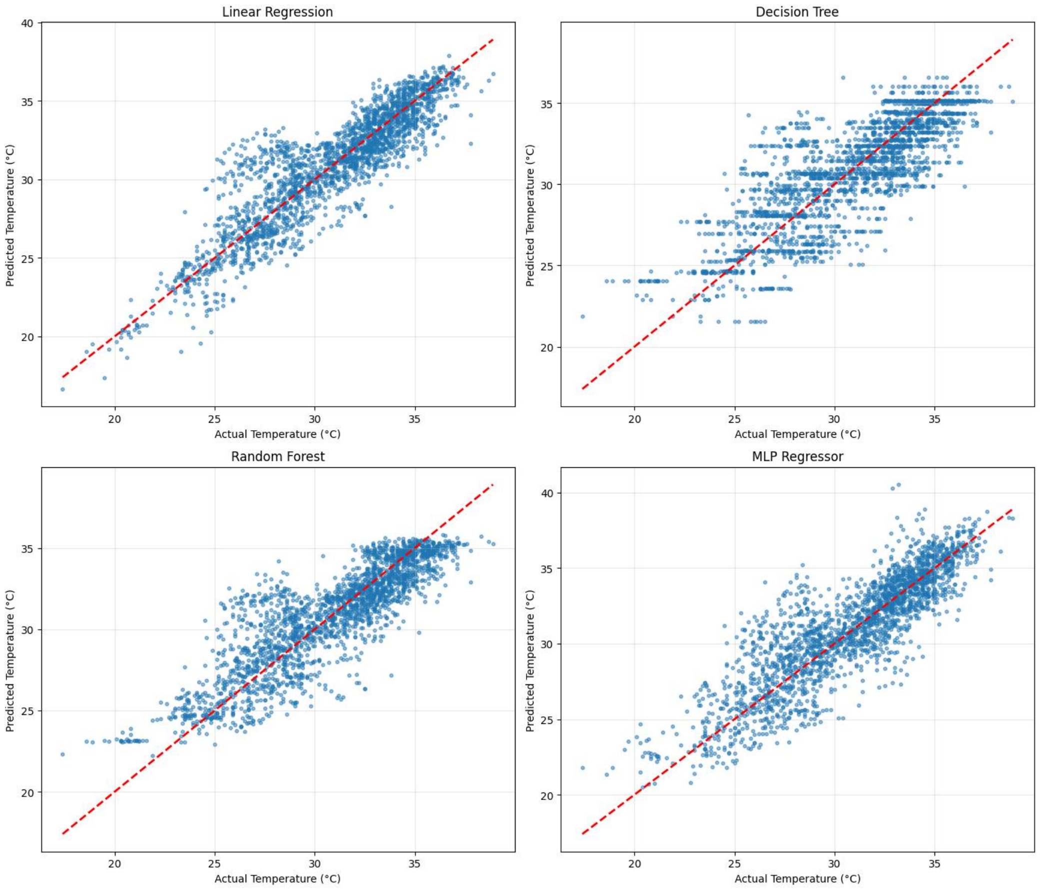


Fig. 5. Prediction vs Actual Temperature for the four training models in Experiment1.

Models	Linear Regression	Decision Tree	Random Forest	MLP Regressor
MAE	1.2167°C	1.6408°C	1.3574°C	1.3837°C
RMSE	1.6437°C	2.1243°C	1.7928°C	1.8567°C
R^2	0.7962	0.6597	0.7576	0.7400

Table 1: Model Performance display based on their MAEs, RMSEs, and R^2 in Experiment1.

Models	Linear Regression	Decision Tree	Random Forest	MLP Regressor
MAE	1.5182°C	2.0258°C	1.7095°C	1.9677°C
RMSE	1.9061°C	2.5950°C	2.2058°C	2.6050°C
R^2	0.7260	0.4921	0.6330	0.4882

Table 2: Model Performance display based on their MAEs, RMSEs, and R^2 in Experiment2.

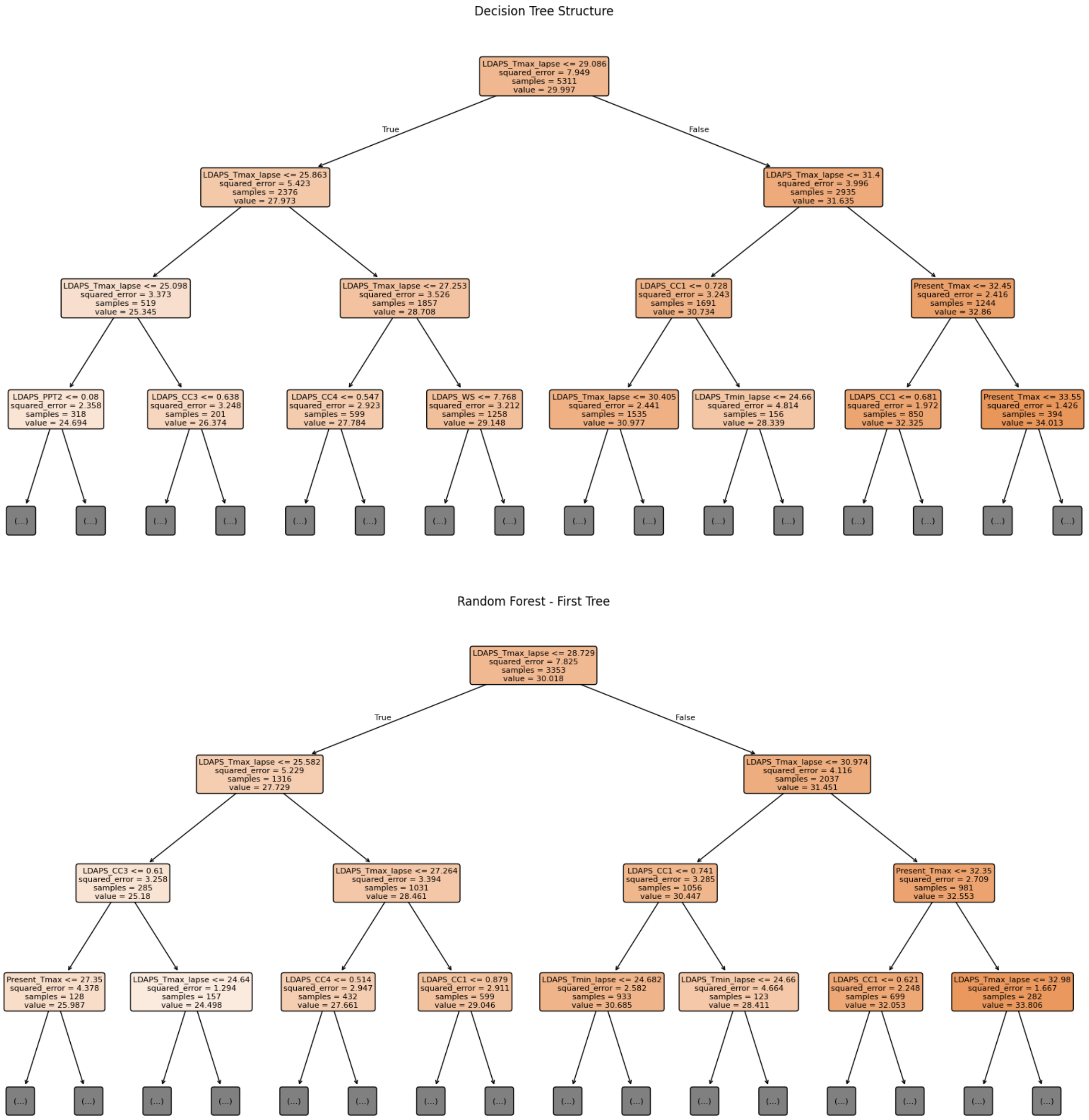


Fig. 6. The structure of the decision tree, and the first tree in the random forest.

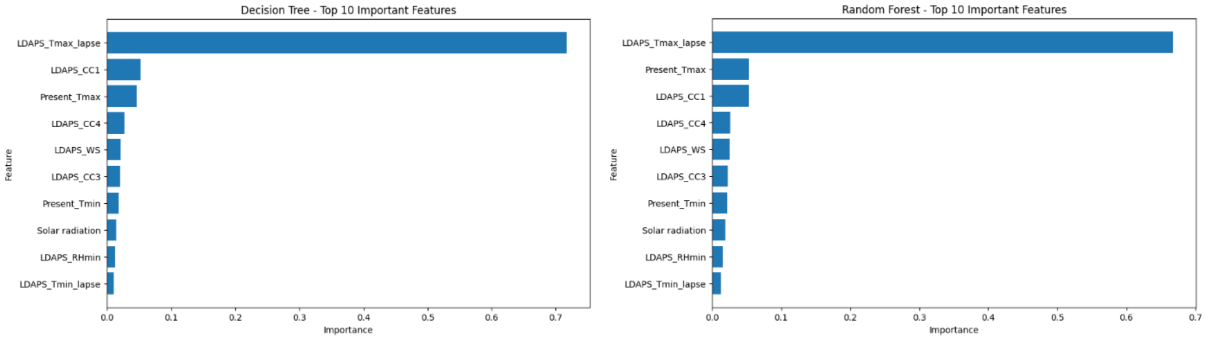


Fig. 7. The relative importance features for the decision tree, and random forest modelling.

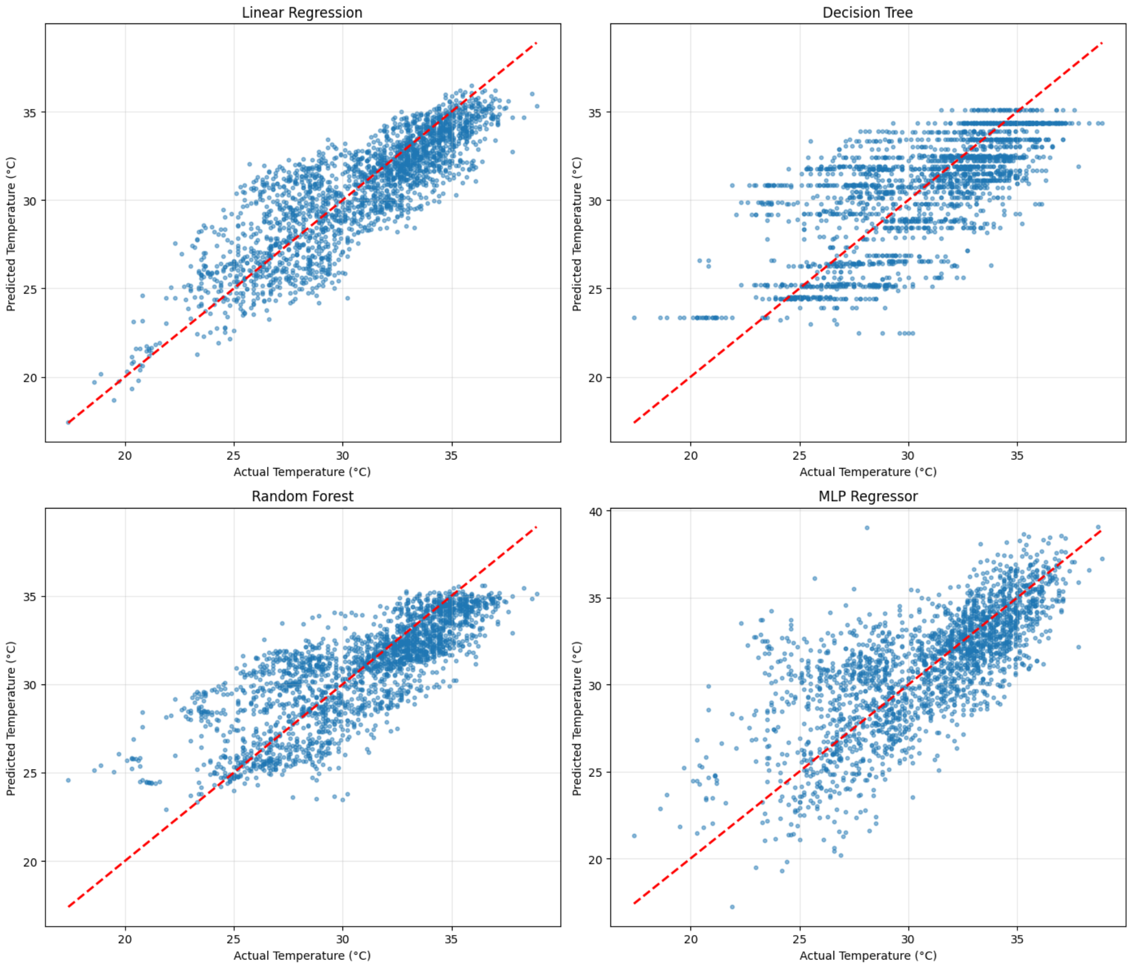
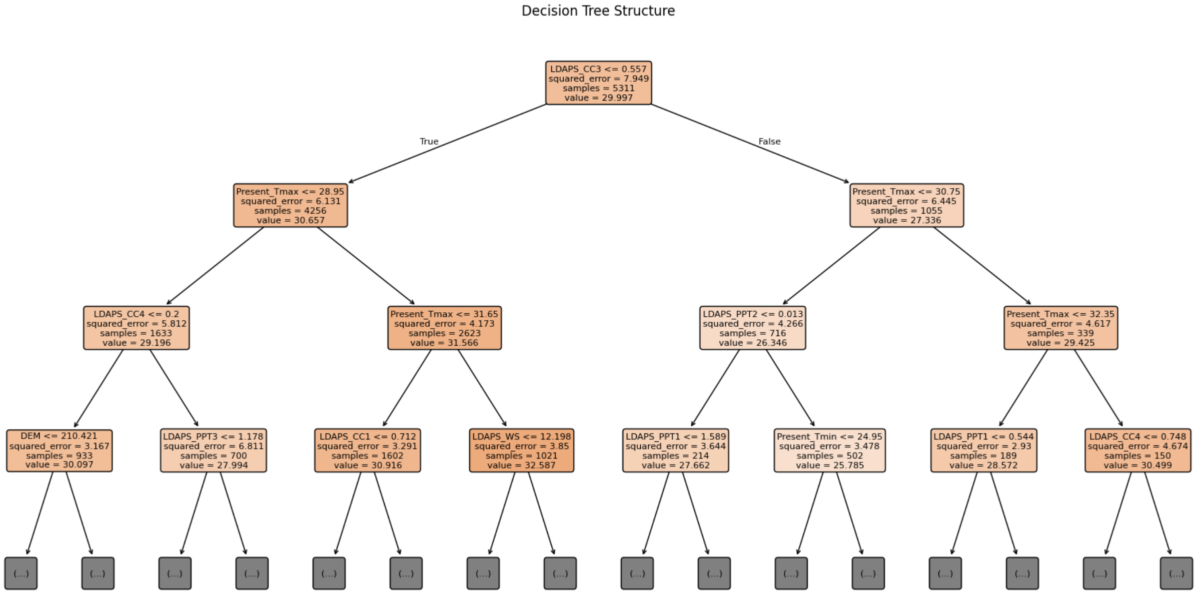


Fig. 8. Same as Fig. 5, but in Experiment2.



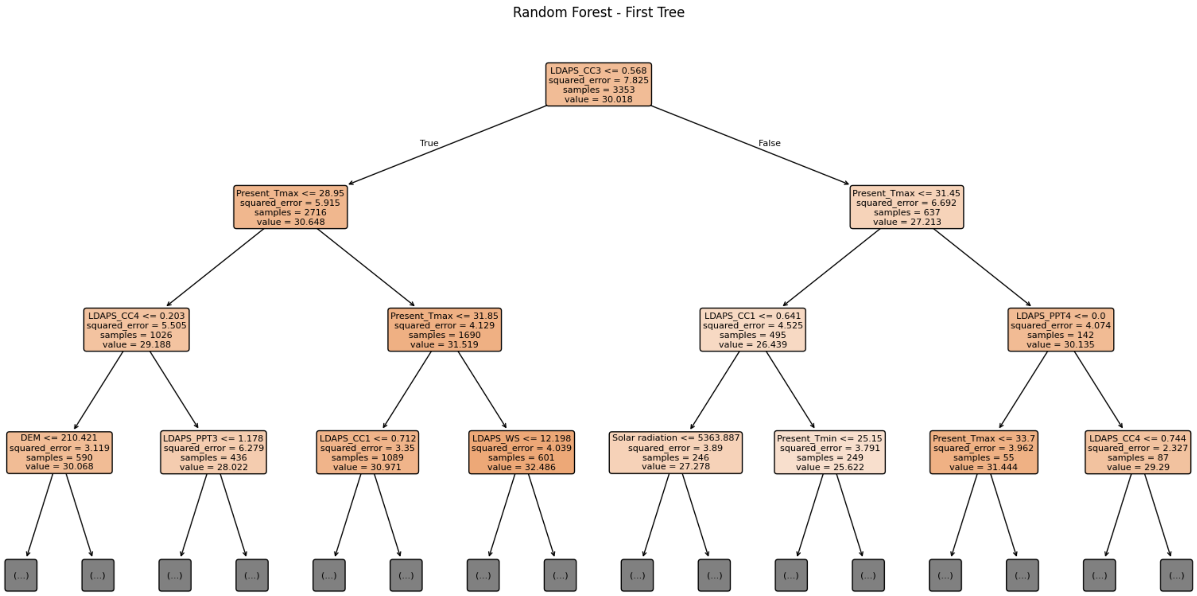


Fig. 9. Same as Fig. 6, but in Experiment2.

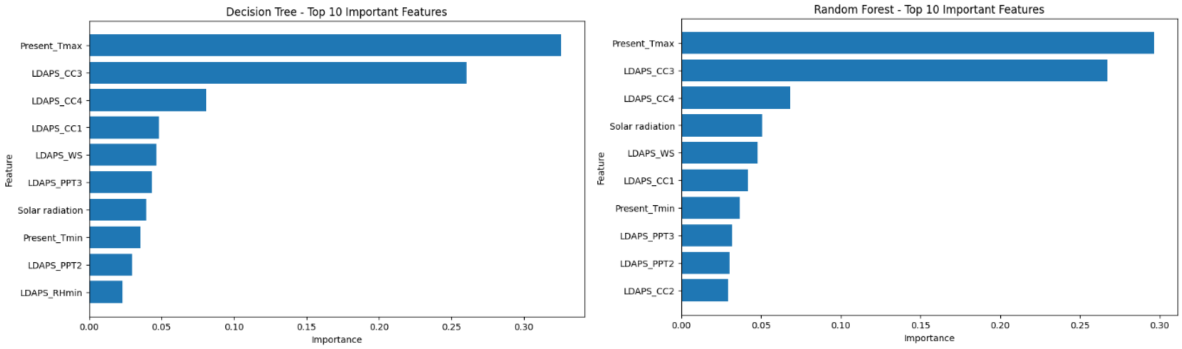


Fig. 10. Same as Fig. 7, but in Experiment2.

Consequently, it is interesting to look at the experiment where we exclude the LDAPS-predicted Tmax from the features, which manually removes the model's reliance on this variable, and then see which variables become the most important, and how the models' performance based on MAE, RMSE, and R^2 changes under this new experiment.

For the same modelling methods, we removed the LDAPS-predicted Tmax from the features, and the results are shown in Table 2 and Fig. 8–10. Compared to Table 1, the MAE and RMSE of each modelling method both increase significantly, while R^2 decreases significantly, which means that without the LDAPS-predicted Tmax serving as a feature, the prediction performance for next-day maximum temperature drops substantially. And in Experiment 2, Linear Regression is still the model with the smallest error. After removing the LDAPS-predicted Tmax as the main dependent variable, Present-Tmax and LDAPS_CC3/4 become the most important variables for Decision Tree and Random Forest. This makes sense because the temperature of the next day often inherits information from the previous day's temperature, and cloud (especially after noon) cover also has a very significant impact on maximum temperature.

6. Discussion

Based on our results, especially from the comparison between Experiment 1 and Experiment 2, we find that the LDAPS-predicted next-day Tmax serves as an important feature for models to learn in order to predict the next-day observed Tmax.

If the LDAPS-predicted next-day Tmax is not included in the features, only Linear Regression shows a slight improvement in accuracy; when the LDAPS-predicted next-day Tmax is included as a feature, Linear Regression, Random Forest, and MLP Regressor can improve the LDAPS forecast accuracy. This finding indicates that these simple regression models do not truly learn the physics in numerical weather forecasting, because they still rely on the LDAPS-predicted data. And the models that can genuinely improve traditional weather forecasts would need to be more complex and coupled.

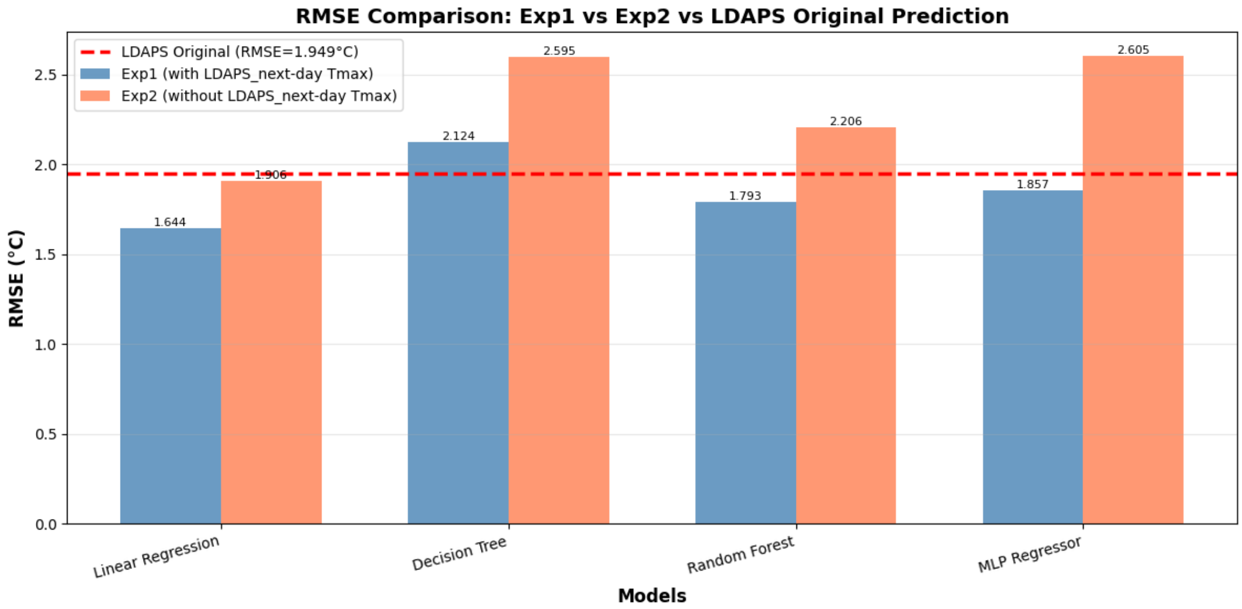


Fig. 11. Comparison of RMSEs between LDAPS-predicted next-day Tmax and machine learning models predicted next-day Tmax for 4 methods and 2 experiments.