Climate Change Prediction Model

Contents

Introduction

Climate change is one of the most pressing global challenges, with rising temperatures posing significant risks to both natural ecosystems and human society.

This project focuses on forecasting land surface temperature using traditional machine learning techniques and historical climate data. Land surface temperature is a critical climate indicator that directly influences human health, agricultural yields, and water resource availability. Increases in surface temperature can exacerbate water scarcity, reduce crop productivity, and increase the incidence of heat-related illnesses—issues that are especially impactful in regions with already extreme climates.

Problem Definition

The objective of this project is to develop a predictive machine learning model that can forecast future land surface temperature trends in Egypt using historical data to support environmental policy-making by providing data-driven insights into future temperature dynamics.

Literature Review

1. Forecasting Land Surface Temperature

Numerous studies have utilized historical LST data to predict future trends. For instance, [Sobrino et al., 2004] demonstrated the utility of satellite-derived LST for monitoring urban heat islands, while [Li et al., 2013] emphasized the importance of LST in hydrological and agricultural modeling.

2. Machine Learning for Climate Prediction

Support vector machines have been effectively used to predict temperature and rainfall patterns ([Pal and Deswal, 2008]), while random forests have been employed for downscaling climate data ([Bedia et al., 2013]) and predicting extreme weather events. These models can capture non-linear relationships between climatic variables and are particularly effective when combined with feature selection and preprocessing techniques.

3. Feature Engineering from Climate Variables

Commonly used climate predictors include precipitation, atmospheric pressure, wind speed and direction, and humidity. [Chen et al., 2020] showed that combining multiple meteorological variables significantly improves the accuracy of temperature forecasting models. Incorporating spatial and temporal features, such as lag variables and seasonal indicators, further enhances model performance.

4. Climate Data Sources and Preprocessing

ERA5 (provided by the Copernicus Climate Data Store) have become widely used in recent years due to their comprehensive coverage and temporal consistency. Preprocessing GRIB or NetCDF climate files into structured formats like Pandas DataFrames is a crucial step in building efficient ML pipelines. Techniques such as dimensionality reduction, normalization, and handling of missing data are often employed to prepare the data for modeling.

Our work focus on the limited regional focus on Egypt. Also the underuse of traditional machine learning in this interpretable and resource constrained settings.

Dataset

Our dataset was ERA5 hourly time-series data from the Climate Data Store project. The variables we requested are Precipitation, Land Surface Temperature, Atmosphere Surface Pressure, Mean Sea Level Pressure, and Wind Speed. The location was Lat: 29.84178°, Long: 30.65322° in Giza, Egypt, from Jan, 2000 till May, 2025.

Preprocessing & Feature Engineering

We loaded the CSV file into python jupyter notebook, the data has no null values. The time-serise plots shows that all variables are have seasonal patterns. the Mean Sea Level Pressure (msl) and Surface Pressure (sp) were almost the same, so we dropped the msl.

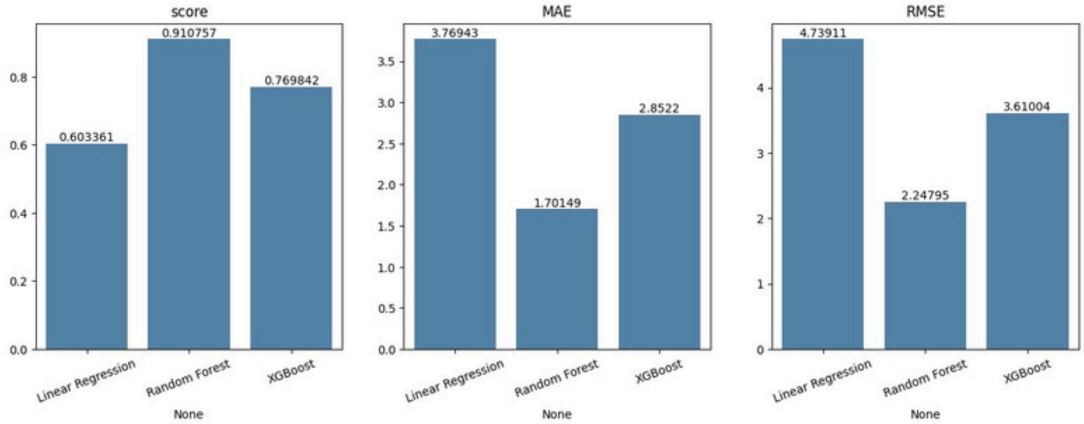
We created hour, day of week, month, quarter, and year as new features from the date. Also, we created wind speed and wind direction from u10 and v10 variables by finding the magnitude between them as the speed as meter/second, and calculating the direction in degree, clockwise, starting from North=0 using this equation:

$$Wind\ Direction = \left(180 + \frac{180}{\pi} \cdot \tan^{-1}\left(\frac{u}{v}\right)\right) \% \ 360$$

Methodology

We selected the landsurface temp as our target variable because rising temperatures affecthealth, crops, water, etc. We tested 3 regression models: linear regression, randomforest, and xgboost.

Split the data intotrain and test sets and used cross-validation to detect overfitting/underfitting or failing to generalize patterns in the data. Trained the models using cross validation with 3 folds. Then, We compared their performance using score, mean absolute error (MAE) and root mean square error (RMSE) and here's the results:

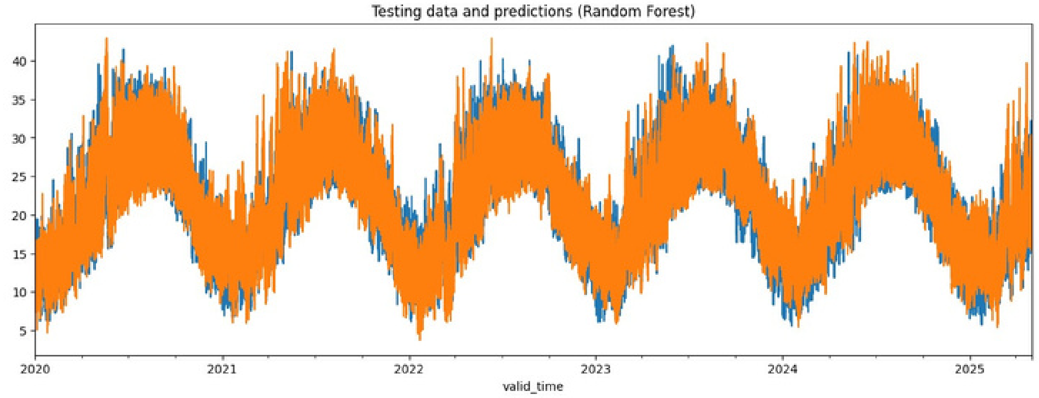


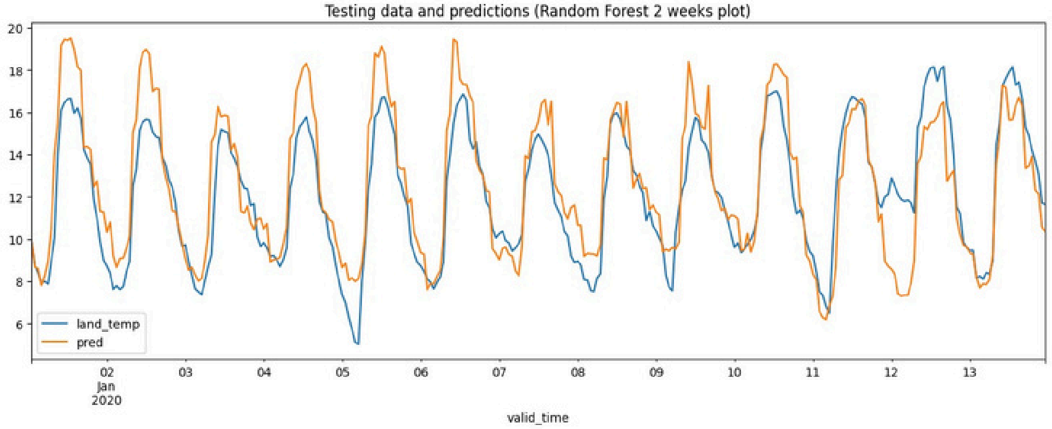
Obviously the best model score is Random Forest with clear overfitting achieving an R squared score of 99% for training and above 91% for testing. Also, Random Forest had the lowest MAE and RMSE compared to the other models.

Further, we tuned the random forest hyper parameters, the parameters were n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features. The hyper parameters tuning didn’t boost the model prediction that much, but it reduced the overfitting achieving accuracy of 92% for the training set and near 91% for the testing set.

Result Analysis

Analyzing the results ofthe model selection process, Random Forest is the most suitable model for forecasting land surface temperature in our current dataset, next are some predictions plots from the test data and its predictions using Random Forest:

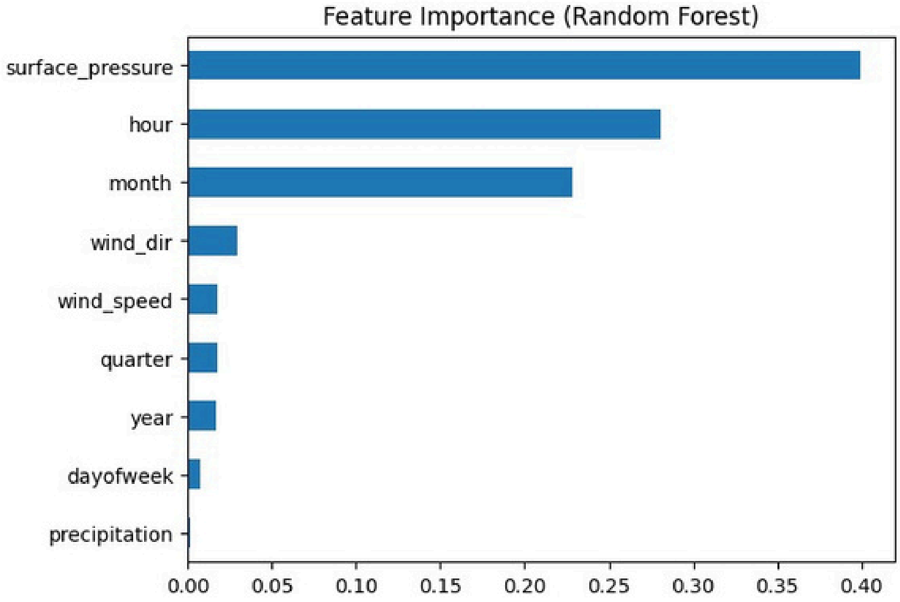




At some points the model missed the pattern but the predictions line suggeststhat the model is able to generalize the data.

After hyper parameters tuning, the Random Forest model had an average cross validation score of 0.904 with standard deviation of 0.006, training score of 0.922 and testing score of 0.907. These results indicate that there are no training problems.

We decided to investigate it to get more insights about the model performance and the dataset. The features importance plot of the Random Forest model shows that the most important features are atmosphere surface pressure and which hour in the day, next to them is the month and the least important feature is total precipitation.



This means that atmosphere surface pressure is highly correlated with temperature changes. Also, daily and seasonal patterns had strong indications of temperature, unlike precipitation, wind speed and direction which had weak influence.