Using Machine Learning Models to Predict Summer Maximum Temperatures in Taiwan through Global Sea Surface Temperature

Cian-Wei Chiang[1], Wan-Ling Tseng[2], Yi-Chi Wang[3]

Department of Environmental Engineering, National Cheng Kung University, Tainan, Taiwan[1]

International Degree Program in Climate Change And Sustainable Development National Taiwan University, Taipei, Taiwan[2]

Academia Sinica, Taipei, Taiwan[3]

Acknowledgments: Cheng-Wei Lin[3], Shih-How Lo[3]

Abstract: This study employs Python scikit-learn machine learning models, including linear regression and random forest, to predict the average summer (June, July, August) maximum temperatures (predictand) for regions in Taiwan below one thousand meters in elevation. The predictions are based on detrended and anomaly May global sea surface temperature (predictor). A comparison is made between the predictions using global sea surface temperature with and without performing empirical orthogonal function analysis (EOF). In addition, we conducted predictions of categorical types of the following summer (i.e. hot anomaly/normal/cool anomaly). The Python SHAP module is utilized to analyze key features employed by the models as primary predictors. Finally, the overall performance of the prediction models is calculated using K-Fold cross-validation.

Motivation and Research Problem

- High temperatures can potentially have adverse effects on humans, animals, and plants. Additionally, they might lead to issues such as increased electricity demand. However, through intervention and prevention, we have the capacity to mitigate the challenges posed by high temperatures. Therefore, our subject is predicting the maximum temperatures in the summer seasons in Taiwan.
- Here, we utilize machine learning models to explore the prediction problem of summer temperature, taking a data-driven approach. The research uses global sea surface temperatures in May (SST; predictor) to forecast the average of June, July, and August (JJA) maximum temperatures (Tmax; predictand) in Taiwan.

Methodology

Correlation map of average JJA Tmax and SST in May

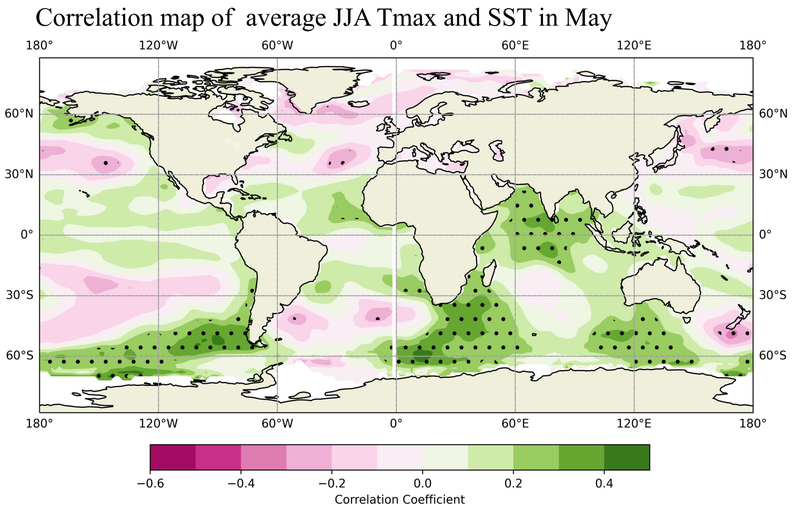


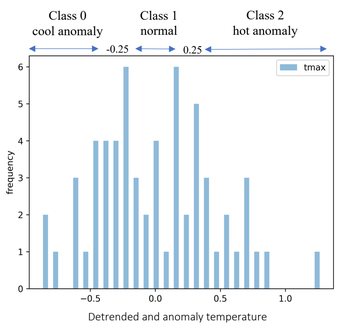
Figure 1 Correlation map of global sea surface temperature in May and average JJA maximum temperature in Taiwan. The dots indicate significance at the 0.05 level. Data source:

- TCCIP grid 5km monthly maximum temperature (JJA data averaged) (1960-2020)
- NOAA 2.0 degree extended reconstructed monthly sea surface temperature V5 (1960-2020)

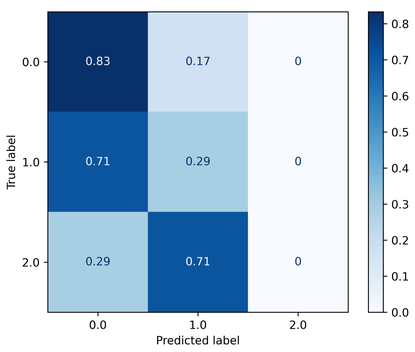
- Data Processing:
  Before model training, global sea surface temperature (SST) is weighted along latitude. Detrend, anomaly, and empirical orthogonal function analysis (EOF) are applied to SST. 128 components are retained after EOF, explained variance ratio is 0.9691.JJA maximum temperature in Taiwan excludes areas with elevation higher than one thousand meters.
- Machine Learning Models:
  sklearn.ensemble.RandomForestClassifier, sklearn.linear_model.LinearRegression, and sklearn.ensemble.RandomForestRegressor are used. (Pedregosa et al., 2011.)
- Model Analysis and Inspection:
  1. LinearRegression: sklearn.linear_model.LinearRegression.coef_
  2. RandomForestClassifier and RandomForestRegressor: SHAP Module. (Lundberg, 2017.)
  3. K-Fold cross-validation

Results: season forecast (hot anomaly/normal/cool anomaly) Random Forest Classifier

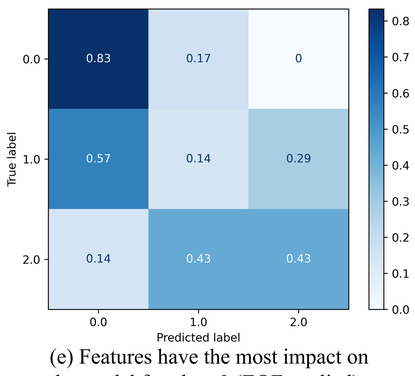
(a) Histogram of detrended and anomaly average JJA Tmax



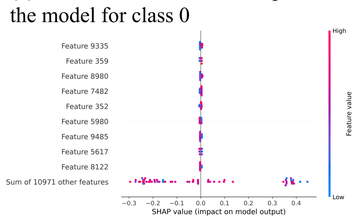
(b) Confusion matrix of RandomForestClassifier



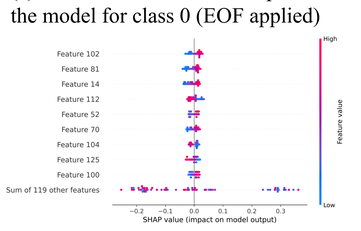
(c) Confusion matrix of RandomForestClassifier (EOF applied)



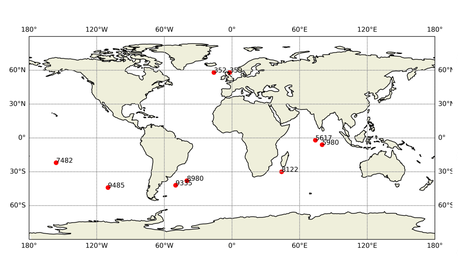
(d) Features have the most impact on the model for class 0



(e) Features have the most impact on the model for class 0 (EOF applied)



(f) Visualization of features



(g) Visualization of features (EOF applied)

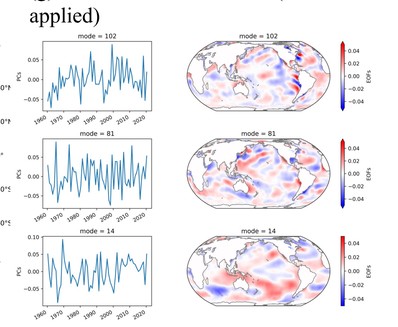
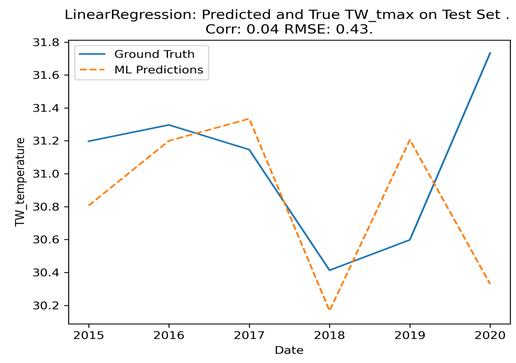


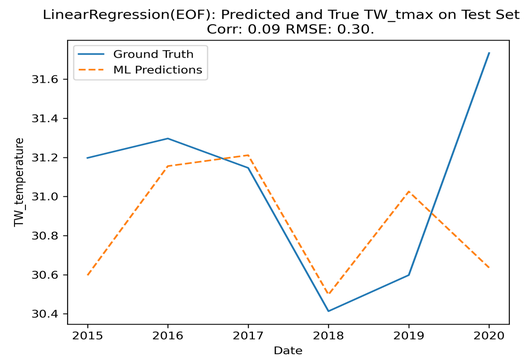
Figure 2 (a) The class predicted is defined by the distribution of detrended and anomaly average JJA Tmax, class 0 (first 34%) is defined as cool anomaly, class 1 is normal, and class 2 (last 33%) is hot anomaly.
(b) (c) performance measurement for model on season predictions. (d) (e) (f) (g) shows the nine most important features used by the model to predict class 0.

Results: JJA temperature forecast Linear Regression

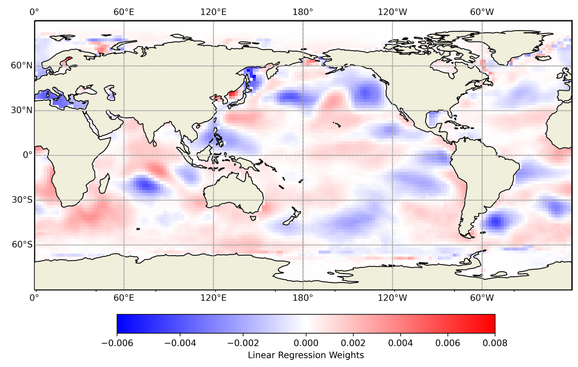
(a) Correlation coefficient and root-mean-square error (RMSE) of predictions and true values



(b) Correlation coefficient and root-mean-square error (RMSE) of predictions and true values (EOF applied)



(c) Estimated coefficients for the linear regression model



(d) PCs and EOFs modes of the four largest estimated coefficients

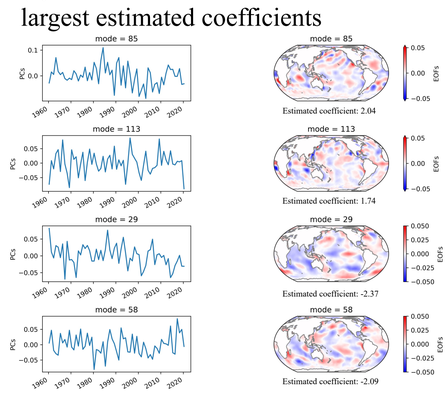
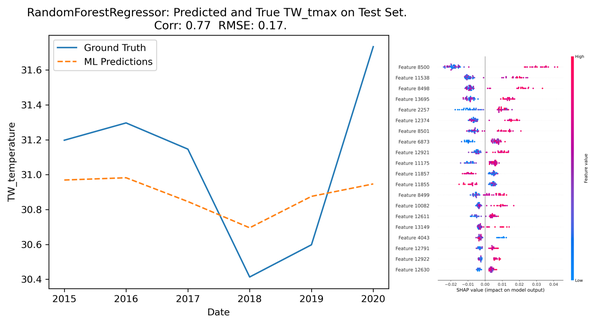


Figure 3 (a) (b) Correlation coefficient and root-mean-square error of the linear regression model. 90% of the data is used in training and validation, and 10% is used in testing. (c) shows the estimated coefficients of the linear regression model. (d) shows the PCs and EOFs modes of the four largest estimated coefficients.

Results: JJA temperature forecast Random Forest Regressor

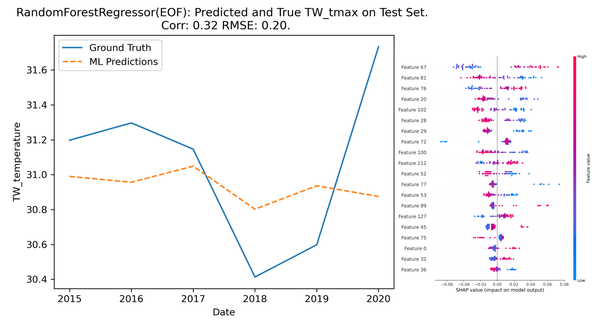
(a) Correlation coefficient and root-mean-square error (RMSE) of predictions and true values

(b) Features with the twenty-largest SHAP value

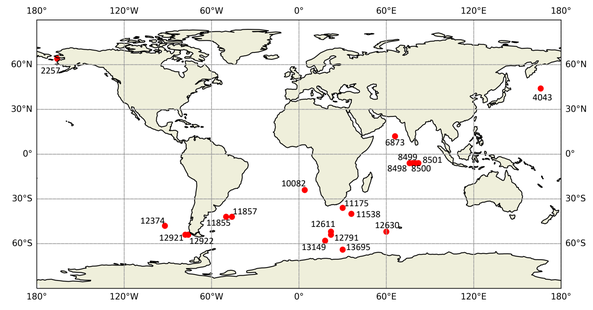


(c) Correlation coefficient and root-mean-square error (RMSE) of predictions and true (EOF applied)

(d) Features with the twenty-largest SHAP value



(e) Visualization of features



(f) Visualization of features (EOF applied)

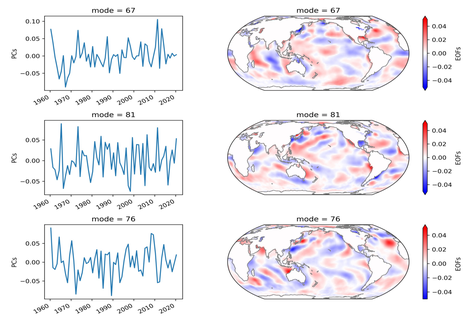


Figure 4 (a) (c) Correlation coefficient and root-mean-square error of the random forest regression model. 90% of the data is used in training and validation, and 10% is used in testing. (b) (d) shows the twenty most important features used by the model. (e) presents the visualization of features. (f) shows the PCs and EOFs modes of the first three features.

Conclusion and Future Work

- **We experimented a data-driven framework with physical explainability for predicting summer-averaged Tmax from SST in May. It can provide a prototype for similar applications.**
- **K-Fold cross-validation for summer classification:**
  - RandomForestClassifier: mean accuracies are 0.3615 and 0.3285 (EOF).
  - ➢ RandomForestClassifier model exhibits better predictive ability for forecasting cool anomaly compared to predicting other categories from the confusion matrices.
- **K-fold cross-validation for summer Tmax:**
  - LinearRegression: mean correlation coefficients are 0.089 and 0.139 (EOF).
  - RandomForestRegressor: mean correlation coefficients are 0.349 and 0.139 (EOF).
  - ➢ Overall, the performance of RandomForestRegressor model is superior to the linear one.
  - ➢ Performing empirical orthogonal function analysis on the predictor variables might not provide significant enhancement to the predictions. Other methods for selecting features may be needed.

- **References:**
  - Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
  - A unified approach to interpreting model predictions. Available at: https://github.com/shap/shap. Lundberg, S. M. 2017.
  - Xarray: N-D labeled Arrays and Datasets in Python. Hoyer, S. & Hamman, J., Journal of Open Research Software. 5(1), p.10. DOI: https://doi.org/10.5334/jors.148, 2017.