Modeling Global Urban Air Temperature Trends Using Machine Learning on Satellite Land Surface Data

***Taseen Islam*[1], *Kip Nielsen*[2,3], *Shaunak Sharma*[2,4], *Ashley Grey*[2,5], *Audrey Lofthouse*[2,6], *Hamidreza Norouzi*[7], *Reginald Blake*[7]**

[1] *CUNY Macaulay Honors College,* [2] *NASA Goddard Institute for Space Studies,* [3] *University of Kansas,* [4] *South Brunswick High School,* [5] *Baruch College Campus High School,* [6] *Brigham Young University,* [7] *CUNY New York City College of Technology*

Introduction

Accurate predictions of air temperature are essential for understanding atmospheric phenomena; yet, the scarcity of available air temperature data poses challenges for comprehensive analysis and forecasting. Land surface temperature (LST) and air temperature are often shown to be correlated, but air temperature data is not as widely available as LST data is, making coupled analysis and predictions difficult. NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) satellite provides continuous global coverage of LST data. However, Automated Surface Observing Systems (ASOS) weather stations that capture air temperature data are only available at specific points, creating a lack of data between ASOS stations where only MODIS data is available, especially in urban areas. A supervised machine learning model was developed using the K-Nearest-Neighbors (KNN) regression algorithm to model air temperature trends and predict air temperature values when given land surface data from the MODIS satellite.

Objectives

- Create a machine learning model that can accurately predict air temperature values when taking in land surface data as inputs.
- Develop a model that can fill in gaps in data when air temperature data is missing.
- Develop a model that can extrapolate air temperature values into the past before air temperature data in a region was recorded and into the future where data has not been recorded yet.

Materials

A supervised machine learning model was developed using the K-Nearest-Neighbors (KNN) regression algorithm. KNN is an algorithm that takes a hyperparameter value k and when given a sample datapoint, looks for the k points closest to that point in the training set. The target feature's values (in this case air temperature) for the k points are then averaged to provide a prediction for the value for the sample point.

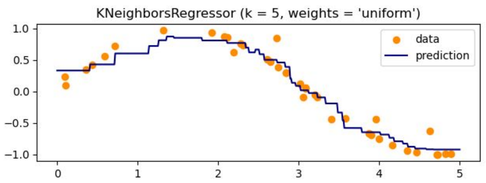
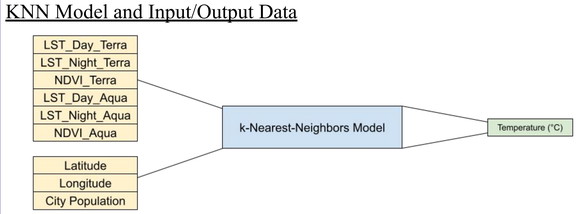


Figure 1: An example of the KNN algorithm used for regression from scikit-learn User Guide Chapter 1.6. Nearest Neighbors

227 ASOS weather stations spanning multiple cities across 125 countries around the world were randomly selected. LST data was obtained from MODIS observations at 1 km resolution at the weather station locations while air temperature data was taken from the stations themselves.

KNN Model and Input/Output Data



Data and Results

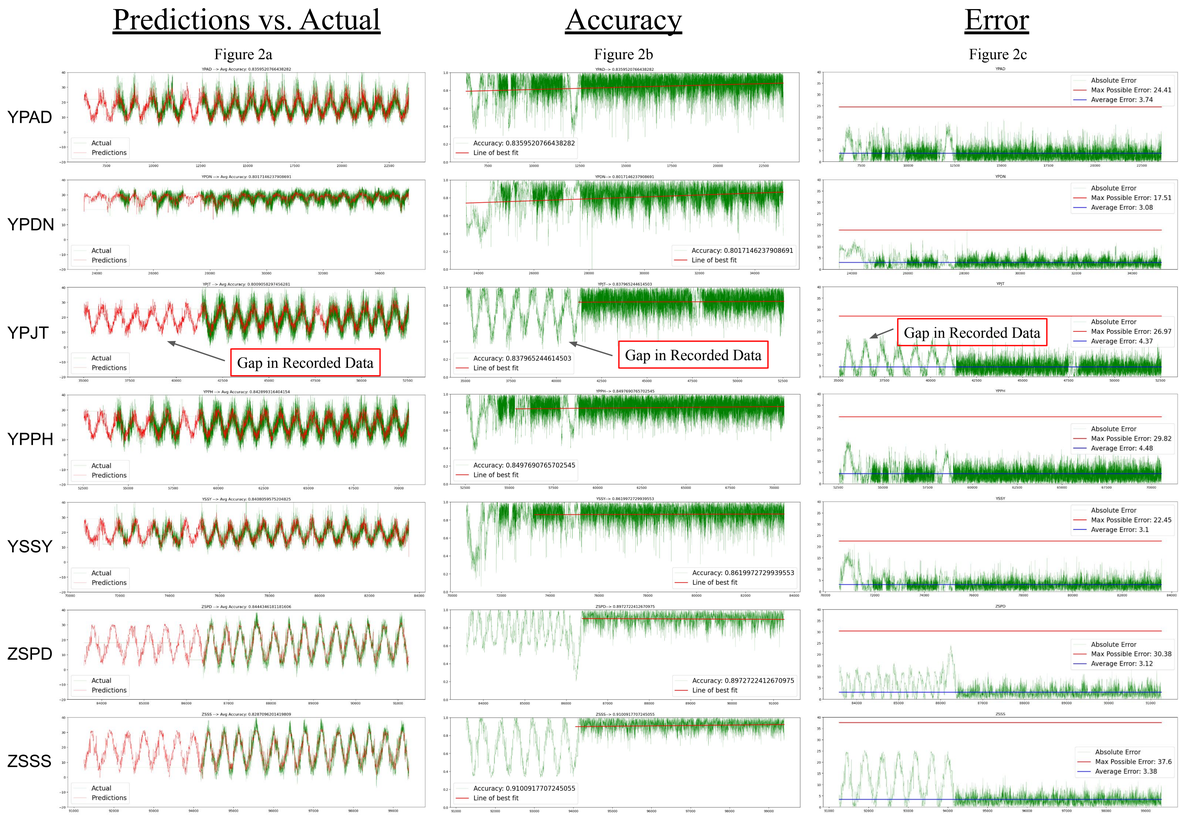


Figure 2: These plots show the model's predictions vs actual values (2a) for 8 stations across the world. The model is performing with a high degree of accuracy as is shown by the relative accuracy (2b) and absolute error (2c) graphs. It is even accurately extrapolating and predicting trends during times where no actual air temperature recordings are available. This is seen in Figure 1a in places where there is a red line but no green line. The y-axis for 2a and 2c is in °C while the x-axis for all three subfigures is the indices associated with each datapoint.

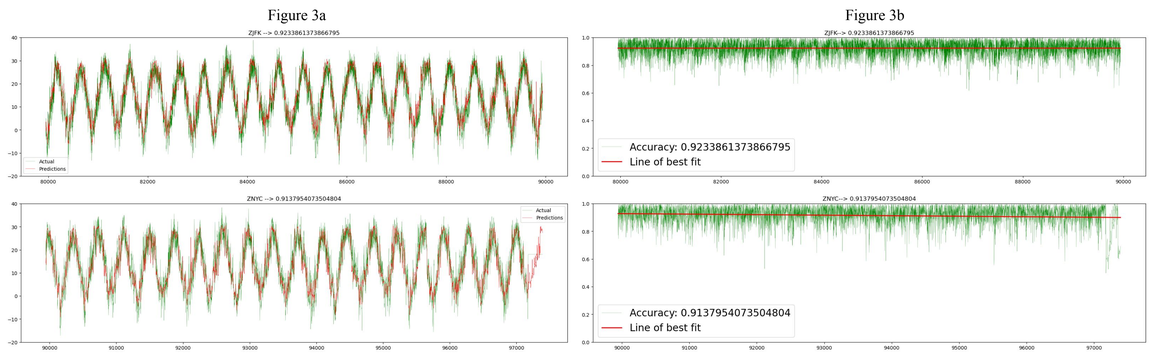


Figure 3: The model is performing very well in New York City. Two stations, ZJFK (JFK Airport) and ZNYC (Midtown Manhattan), are shown here with their predictions vs actual air temperatures graphs (3a) and their accuracy graphs (3b). ZJFK has an accuracy of 92.3% while ZNYC has an accuracy of 91.4%.

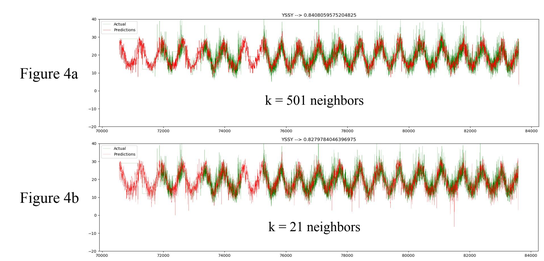


Figure 4: Raising the number of neighbors in the model brings greater general accuracy at the cost of losing some variance. This is called the Bias-Variance Tradeoff. When k-neighbors = 21, the model is 82.8% accurate, but when k-neighbors = 501, the accuracy goes up to 84.1%. However, the model predicted more outliers and greater extremes when a lower k-neighbors hyperparameter was used. This is shown by more frequent and larger red spikes on Figure 4b than 4a.

Methods

Since air temperature is taken hourly but LST is taken more sparsely for a single location, data points would have missing LST values. Stations were first individually cleaned by removing any data points where all four LST values were missing. Cubic spline interpolation was then applied to the LST values to fill in any remaining missing values. The stations were combined into training and testing sets and then the training set was scaled using a Robust Scaler. A function was created to apply the same scaling function to the unscaled testing set values before they were used by the model.

Conclusions

The KNN model performs with 82% of predictions falling within ±10% from the actual value, an RMSE of 4.9 °C, and a median absolute error of 3.2 °C. The model performed much better on stations that had larger testing dataset sizes, with the model often performing with an accuracy over 85% and sometimes even over 90%. The model is able to accurately predict and model air temperature data for the US and other countries and is even able to model air temperature during times where ASOS air temperature data is not available, filling in data gaps in regions without recorded air temperature readings. This capability presents a unique opportunity to generate comprehensive air temperature predictions for urban regions with heterogeneous land cover types. The model has the potential to continue predicting future air temperature trends and values across the world, fill in gaps in data, and provide spatial heat maps of air temperature estimates to create a more comprehensive understanding of air temperature in urban settings.

Future Work

While the model is already performing very well, further additions can be made to improve it. More input parameters can be added, like land cover type, building height, and proximity to water. Introducing more weather stations from currently excluded cities can create an even more robust dataset for the model that would allow it to better model natural air temperature variance without sacrificing accuracy even when the number of neighbors is lowered, combating the Bias-Variance tradeoff and introducing a wider range of locations where the model can be accurately applied. Heatmaps can also be created of regions in between ASOS stations to create a more comprehensive understanding of global air temperature dynamics.

References

1. Brownlee, J. (2020, August 17). kNN Imputation for Missing Values in Machine Learning [web log]. Retrieved from https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/.
2. Ding, X., Y. Zhao, Y. Fan, Y. Li, and J. Ge, 2023: Machine learning-assisted mapping of city-scale air temperature: Using sparse meteorological data for urban climate modeling and adaptation. *Build. Environ.*, **234**, 110211, https://doi.org/10.1016/j.buildenv.2023.110211.
3. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Acknowledgements

This summer research project was supported by NOAA CESSRST Cooperative Agreement Grant #NA16SEC4810008 (Summer Bridge), NSF Grant AGS-2150432 (REU), NSF Grant ICER-2023174 (IUSE), and Pinkerton Foundation (HIRES). The authors would like to express their gratitude to their faculty advisors Dr. Hamidreza Norouzi and Dr. Reginald Blake for their assistance and guidance. The statements contained within the poster are not the opinions of the funding agency or the U.S. government but reflect the author's opinions.