Date: January 31st, 2025

Brief Summary

Data Processing:

I downloaded Sea Surface Temperature(SST) data from the NOAA OISST dataset for the period May-August from 2020-2024 (Link). I extracted the SST variable from NetCDF files and combined them into a single NetCDF file which contains SST data for 615 days [May-August, 2020-2024, hour 12.00pm], cropped/subset the dataset to focus/store on a specific geographical region with Latitude: 0° to 35° and Longitude: 260° to 360°. The selected region covers the Gulf of Mexico, The Caribbean Sea, and the Atlantic Ocean up to the Saharan Dessert. After processing, I converted the NetCDF file into CSV format and split the dataset into Training set: 2020-2022 (May-August), Validation Set: 2023 (May-August) and Test set: 2024 (May-August).

I also downloaded DUST data from NASA's MERRA-2 dataset (Link) for the same period. I extracted the DUCMASS variable from NetCDF files and combined them into a single NetCDF file which contains dust data for 615 days [May-August, 2020-2024, hour 12.300pm]. The DUCMASS variable has 24 hourly time indices per day, so I selected the 12.30PM timestamp to align with SST variable at 12.00PM. After regridding the DUCMASS latitude and longitude to match the SST grid, cropped/subset the dataset to focus/store on a specific geographical region with Latitude: 0° to 35° and Longitude: 260° to 360° (same as SST region). After processing, I converted the NetCDF file into CSV format and merged the SST dataset and Dust datasets into a single CSV file containing [Time (from SST), Latitude (from SST), Longitude (from SST), SST, DUCMASS]. I split this combined dataset into training, validation and test datasets.

Model Comparison:

I wrote Python scripts to test and compare various Machine Learning (ML) and Deep Learning(DL) models for SST prediction with dust data and without dust data. I initially experimented with traditional ML models such as Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbors. After evaluating their performance, I experimented with Boosting models like XGBoost, Gradient Boost, Adaboost, which showed promising results. To further improve accuracy, I experimented with advanced deep learning models like Deep learning model such as LSTM, CNN and hybrid CNN-LSTM model.
I evaluated each model based on Mean Absolute Error (MAE), Mean Squared Error(MSE), and Root Mean Squared Error (RMSE). During training each model was trained on training set. The Validation set helped to monitor overfitting issues, and both the validation and test set was used to evaluate final model performance on unseen data.

Without adding dust data as a feature, XGBoost, Neural Network and LSTM performed the best. To examine whether dust data improved SST predictions, I re-trained the models using the combined SST-dust dataset, adding DUCMASS as an input feature. Now, the input features include [day, month, year, latitude, longitude, and ducmass], while the target variable remains SST. After adding dust data, I observed significant improvements in MAE, MSE, and RMSE scores, particularly in advanced DL model LSTM, CNN and hybrid CNN-LSTM. These models showed notable performance improvements after integrating dust data.

Date: January 31st, 2025

Here I have added some examples of True SST, CNN SST prediction, CNN SST prediction with dust:

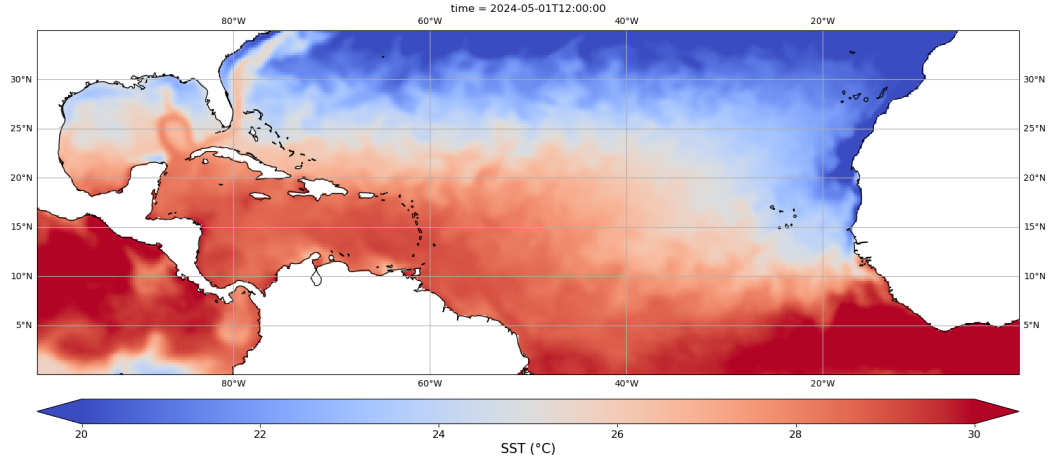


Fig: Actual SST of Day (2024-05-01)

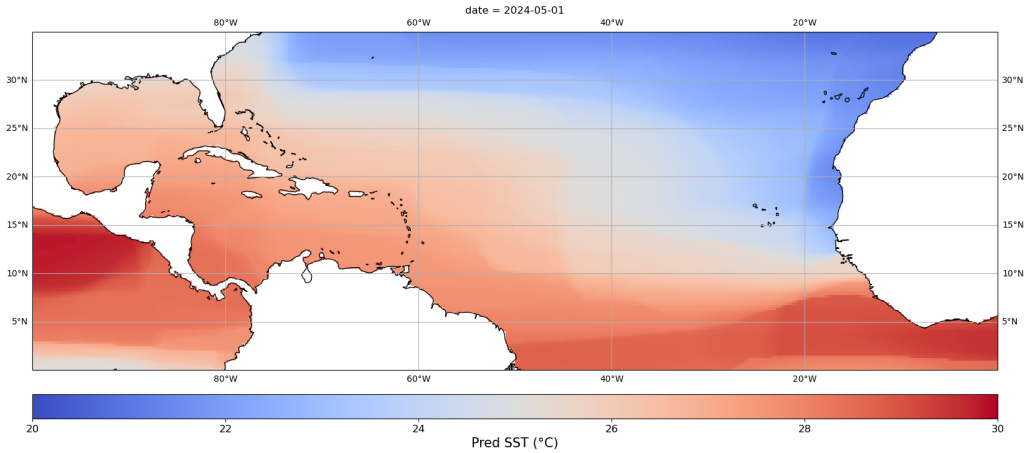


Fig: SST prediction of CNN Day (2024-05-01)

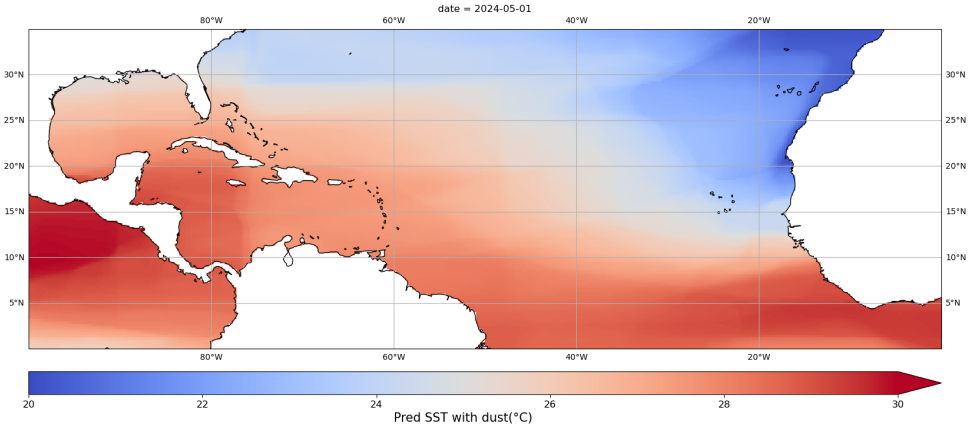


Fig: SST prediction with Dust of CNN Day (2024-05-01)

Model	Validation MSE	Validation MAE	Validation RMSE	Test MSE	Test MAE	Test RMSE
Linear Regression	3.19	1.52	1.79	4.29	1.74	2.07
Gradient Boosting	1.33	0.96	1.15	1.12	0.90	1.06
AdaBoost	2.76	1.45	1.66	2.53	1.38	1.59
K-Nearest Neighbors	1.47	0.98	1.21	1.16	0.90	1.08
Decision Tree	1.83	1.09	1.35	1.82	1.10	1.35
Random Forest	1.82	1.09	1.35	1.81	1.10	1.35
XGBoost	1.33	0.95	1.15	1.07	0.87	1.03
Neural Network	0.94	0.73	0.97	1.86	1.05	1.36
CNN	1.25	0.85	1.12	2.27	1.09	1.51
LSTM	1.20	0.88	1.09	1.35	0.90	1.16
LSTM-CNN	1.57	0.97	1.25	1.90	1.07	1.38

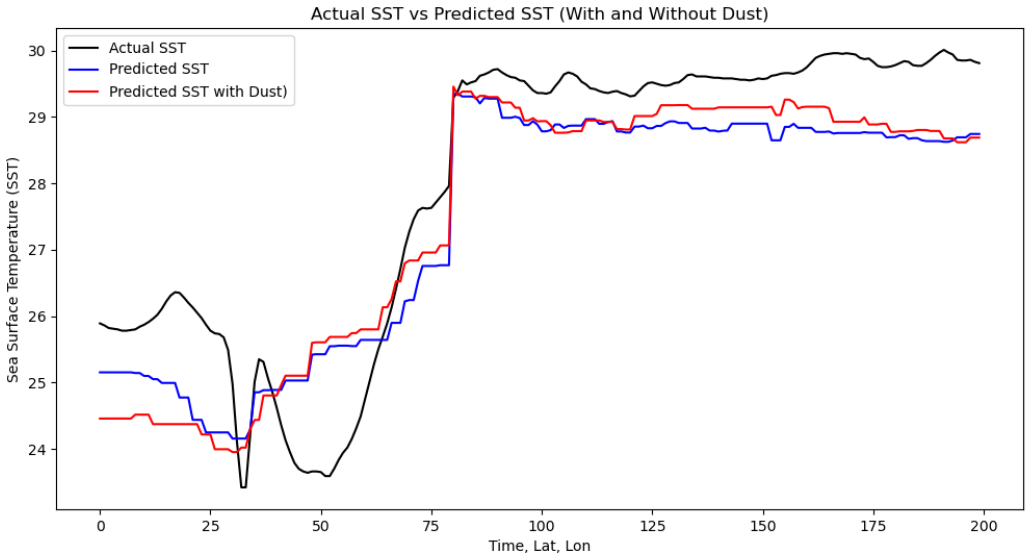
Fig: SST model evaluation results

Model	Validation MSE	Validation MAE	Validation RMSE	Test MSE	Test MAE	Test RMSE
Linear Regression	3.18	1.52	1.78	4.33	1.76	2.08
Gradient Boosting	1.32	0.96	1.15	1.13	0.91	1.06
AdaBoost	2.69	1.43	1.64	2.51	1.37	1.58
K-Nearest Neighbors	1.47	0.98	1.21	1.15	0.90	1.07
Decision Tree	1.83	1.10	1.35	1.82	1.10	1.35
Random Forest	1.82	1.09	1.35	1.81	1.10	1.35
XGBoost	1.29	0.93	1.14	1.07	0.88	1.03
Neural Network	0.98	0.75	0.99	2.11	1.10	1.45
CNN	1.08	0.80	1.04	1.07	0.80	1.04
LSTM	1.17	0.88	1.08	1.12	0.85	1.06
LSTM-CNN	1.52	0.98	1.23	1.62	1.03	1.27

Fig: SST model (added dust) evaluation results

Date: January 31st, 2025

Visualization of first 200 lat and lon points from test set:



The model struggled to accurately capture the underlying patterns of SST, as evidenced by the discrepancies between the predicted and actual SST distributions. However, after adding the dust feature (DUCMASS) into the model, a noticeable improvement was observed in the evaluation results. The addition of dust as a feature provided the model with more relevant environmental context, which helped in better understanding and predicting SST variations. This improvement highlights the significance of including atmospheric variables that influence SST patterns.

It is important to note that the model operates on a gridded dataset, predicting SST values across a mesh of latitude and longitude points. Unlike models that predict SST at specific station locations, this approach provides a comprehensive spatial representation of SST across the region of interest.