Anna Binoy
Sumegha M T

Final Presentation

APPLICATION OF MACHINE LEARNING IN PREDICTING GASEOUS PROPERTIES OF EARTH ATMOSPHERE

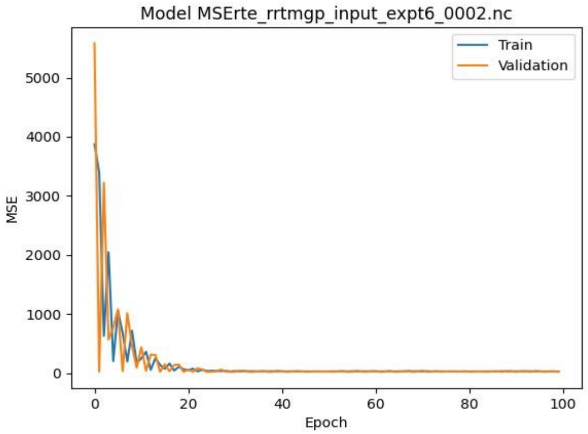
Under the guidance of Dr. Subhankar Mishra and Dr. Jayesh M Goyal

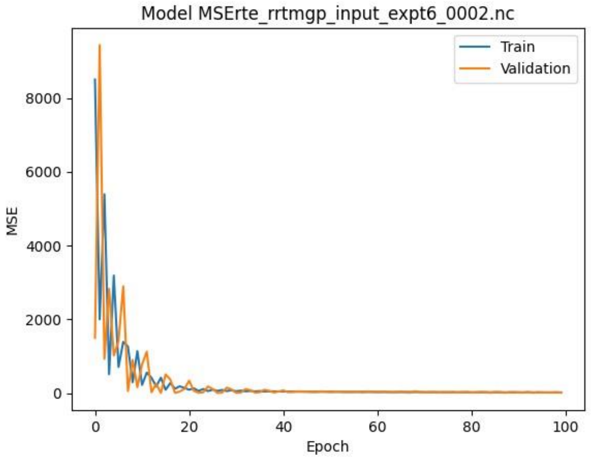
CONTRIBUTIONS

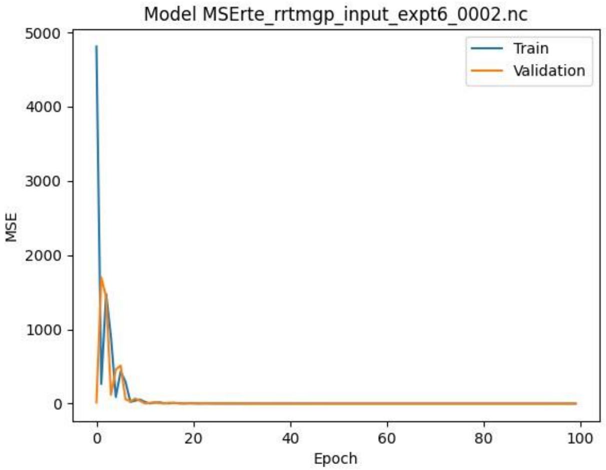
- Trained different datasets on Random Forest and XG Boost
- Obtained good accuracy in the prediction of Single Scattering Albedo which increased with the increase in the size of the training data
- Average accuracy for the prediction of Planck Source Function.
- When the layers were spliced into upper and lower atmosphere, further accuracy was achieved for SSA prediction in the lower atmosphere.

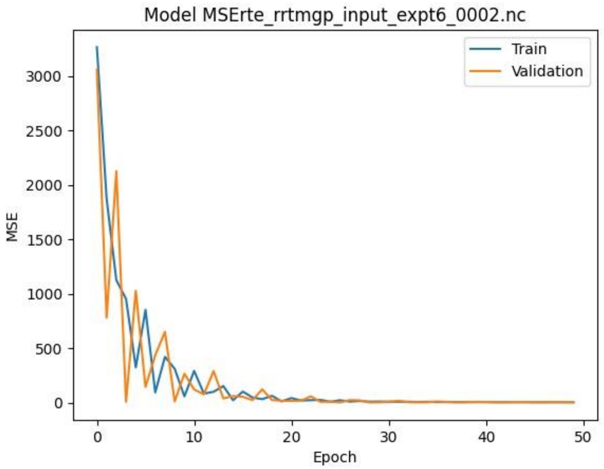
MIDWAY RECAP

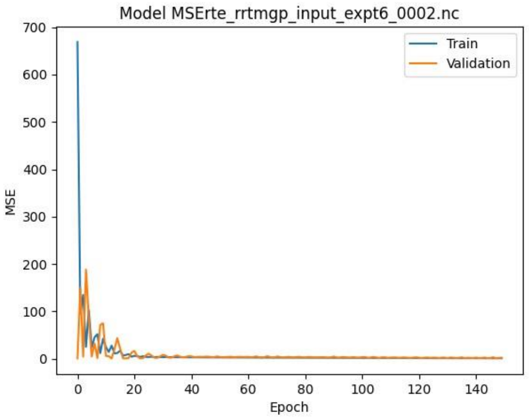
- Generated perturbed input atmospheric profiles, each containing the data of 60 layers (10 kms) and 100 different sites.
- Developed a feed-forward neural network to emulate the look-up tables and trained it using 3600 atmospheric profiles.
- The layer temperature was used to predict the water vapor concentration. Here, temperature was the parameter while the number of epochs and learning rate were the hyperparameters.
- Faced problems while labelling the profiles; unable to label them.

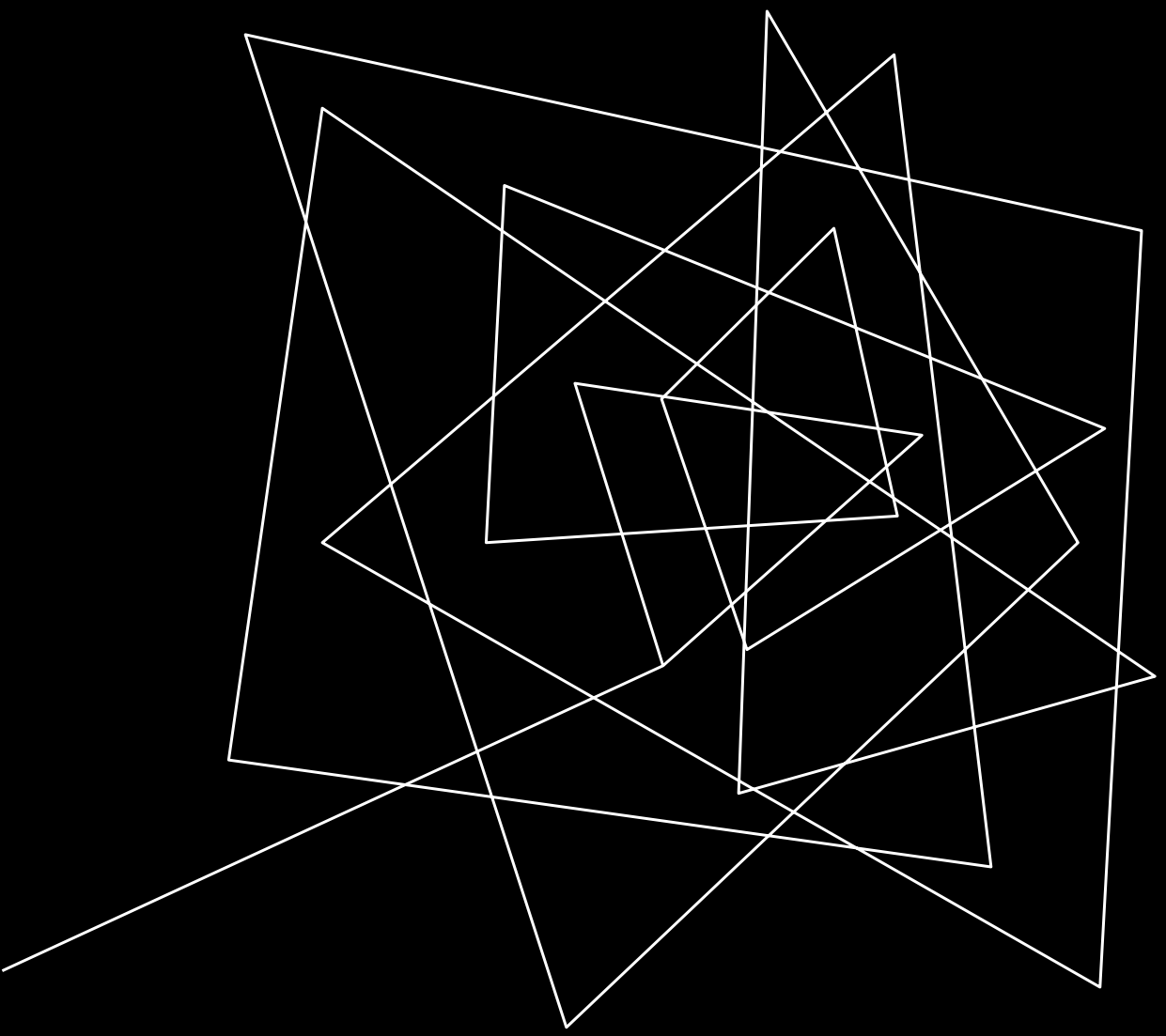












FURTHER PROGRESS…

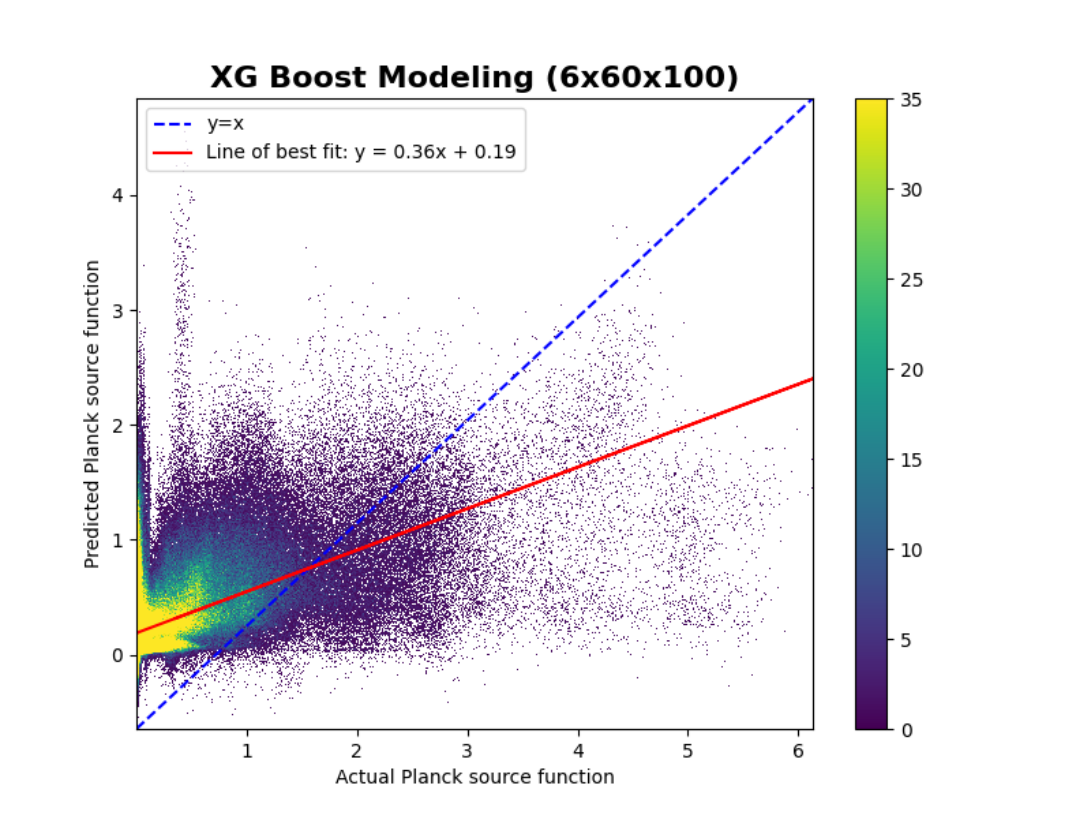
- Successfully labelled the input atmospheric profiles.
- Concatenated 9000 such input atmospheric profiles to a single netCDF file , as for the corresponding output files and divided them into training, testing and validation in 3:1:1 ratio .
- Tried training a neural network model but failed due to GPU Memory saturation.
- Smaller number of concatenated input files were created and divided according to the above ratio.

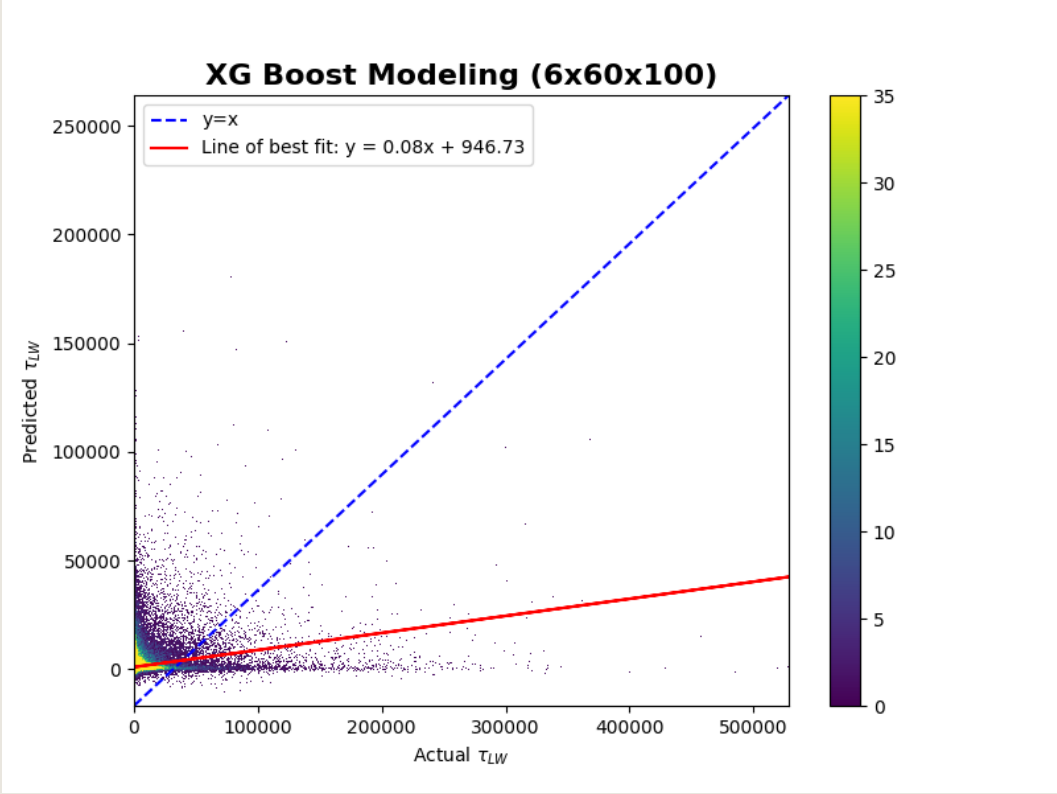
FURTHER PROGRESS…

- Training neural networks over the smaller dataset didn’t work out, again due to GPU memory saturation.
- Parallelly, we were working on SVR as well as Random Forest.
- Due to the difference in the dimensions of the input parameters (nx60x100) and output (nx256x60x100 or nx224x60x100), we realized SVR doesn’t work. Random Forest worked!
- Predictions were evaluated according to MSE error and model accuracy.
- Random Forest predicted Planck Source Function to an average accuracy, SSA to a good accuracy but the short wavelength and long wavelength optical depth showed a very bad accuracy.

FURTHERMORE PROGRESS…

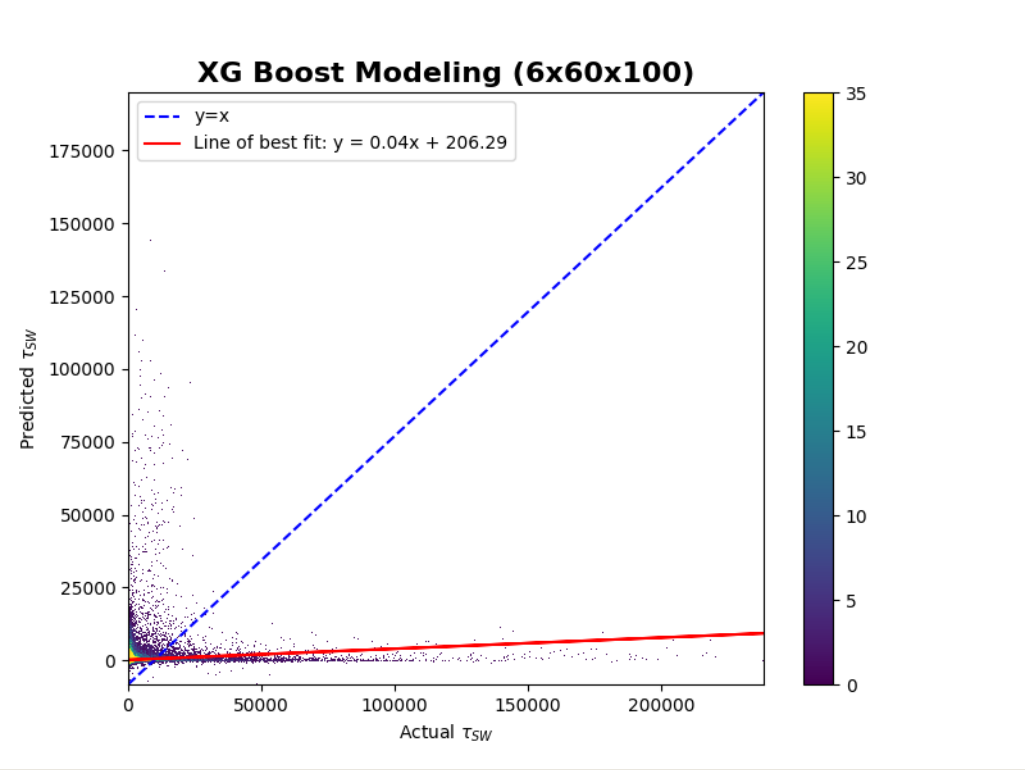
- Furthermore, we trained the data in XG Boost model.

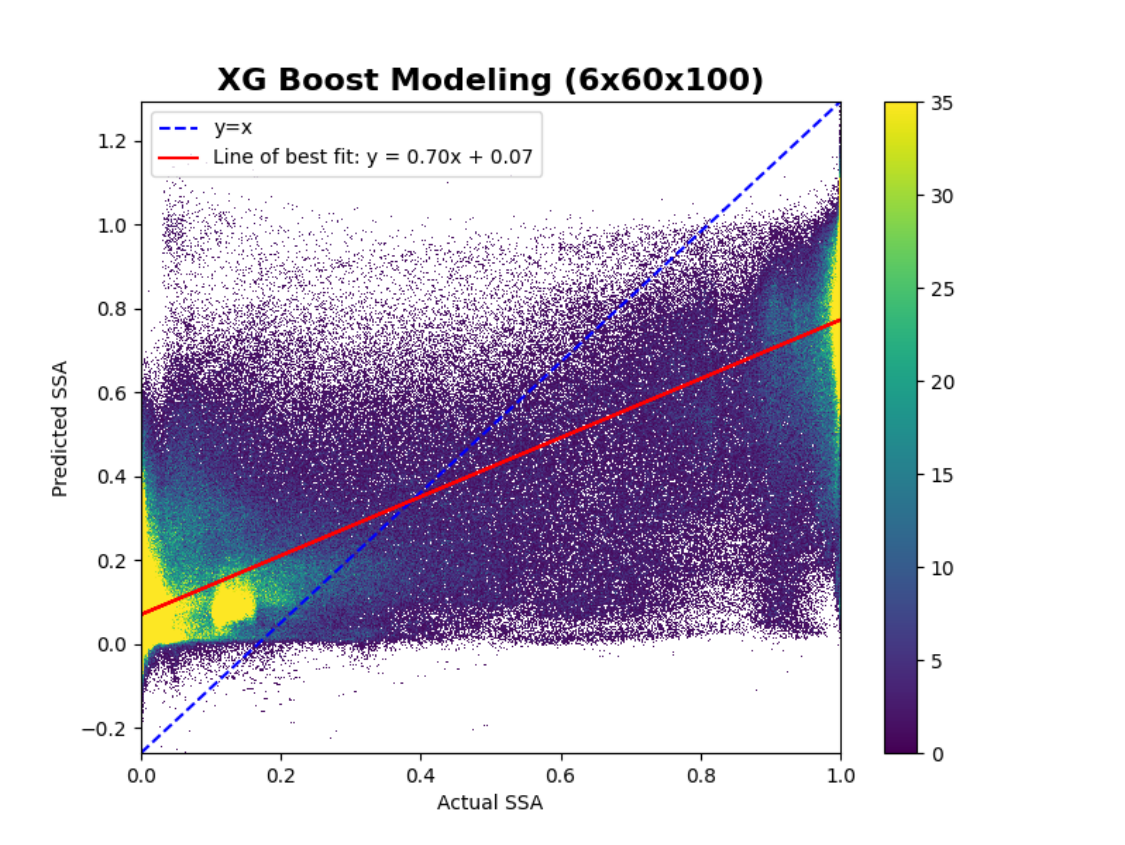




FURTHERMORE PROGRESS…

- Furthermore, we trained the data in XG Boost model.





MODEL COMPARISON

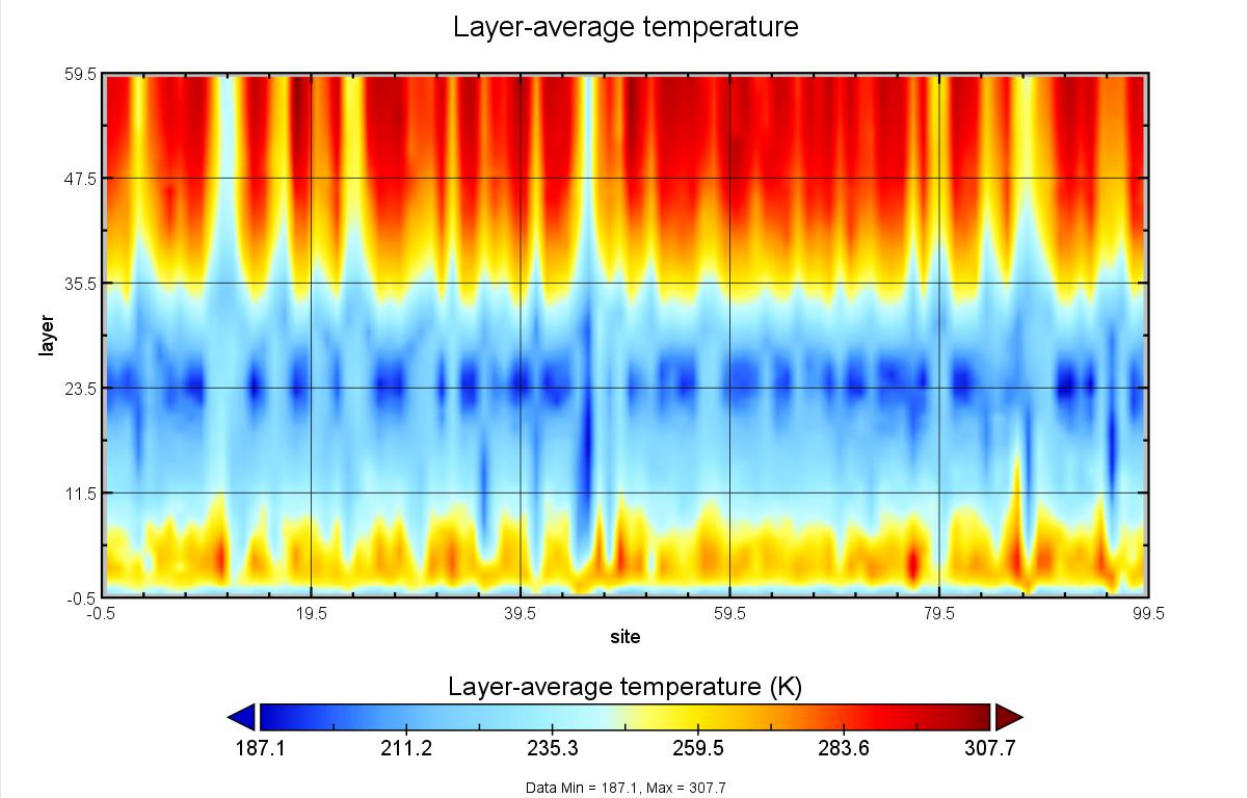
Output	Dataset(x60x100)	Model	MSE	Model Accuracy
Planck Source Function	6	Random Forest	0.169488	0.550532
		XG Boost	0.668303	0.125143
	10	Random Forest	0.225189	0.385423
	20	Random Forest	0.173947	0.504639
	60	Random Forest	0.206073	0.418546
	100	Random Forest	0.184279	0.459319
Optical Depth(Short wavelength)	6	Random Forest	547682.3	0.544293
		XG Boost	295694.5	0.754909
	10	Random Forest	1485905	0.12008
	20	Random Forest	6409936	0.373953
	60	Random Forest	2769729	0.118147
	100	Random Forest	4832451	0.165518

MODEL COMPARISON

Output	Dataset(x60x100	Model	MSE	Model Accuracy
Single Scattering Albedo	6	Random Forest	0.040752	0.702703
		XG Boost	0.034746	0.746511
	10	Random Forest	0.050555	0.630096
	20	Random Forest	0.042998	0.679089
	60	Random Forest	0.050555	0.631237
	100	Random Forest	0.048486	0.649238
Optical Depth (Long wavelength)	6	Random Forest	0.169488	0.550532
		XG Boost	0.668303	0.125143
	10	Random Forest	0.225189	0.385423
	20	Random Forest	0.173947	0.504639
	60	Random Forest	0.206073	0.418546
	100	Random Forest	0.184279	0.459319

FURTHERMORE PROGRESS…

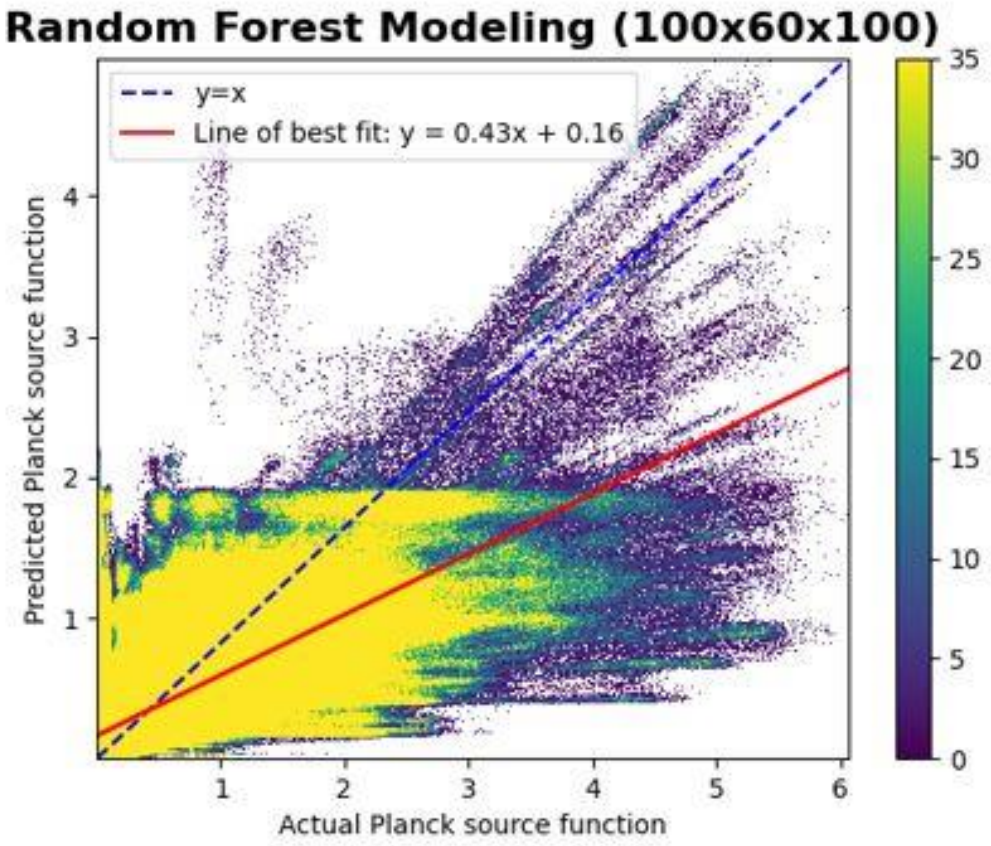
- The memory problem was still persistent in some cases and the datasets were layer wise sliced.
- Furthermore, after examining the temperature v/s layer and pressure v/s layer plots, we sliced the atmospheric profiles into upper and lower layers.
- Random Forest was run on sliced input data points of 60x60x100 and 100x60x100 along with their corresponding output files.
- RF showed excellent accuracy in predicting SSA for the lower atmosphere profiles of 100 input files.
- Hyperparameter tuning for RF was done using grid search

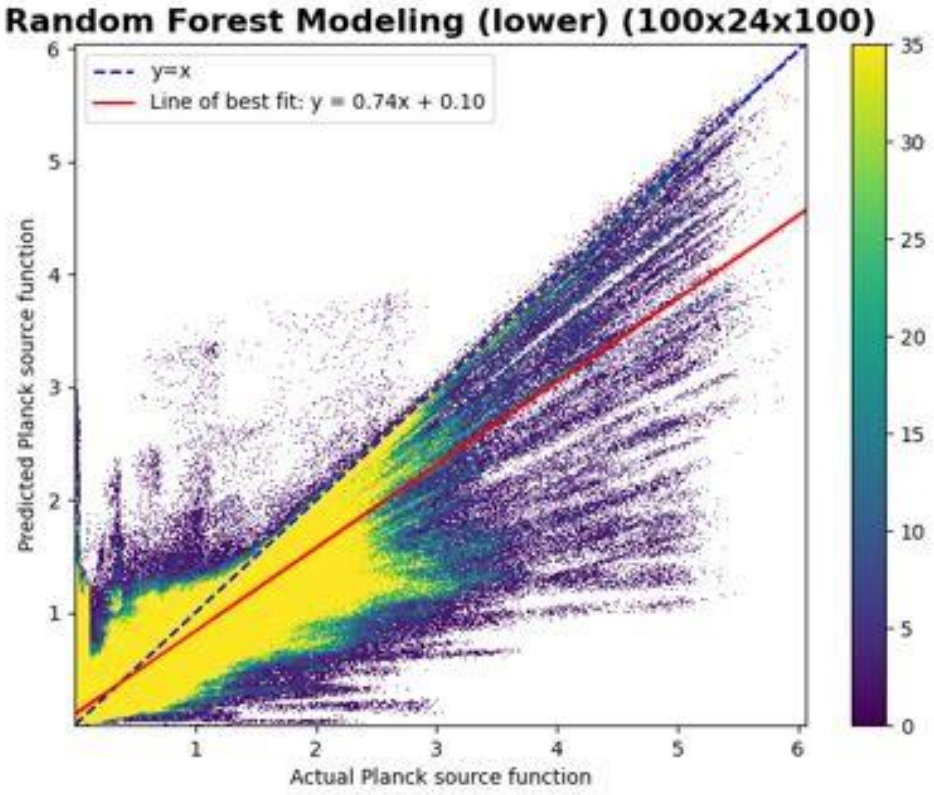


RANDOM FOREST ANALYSIS

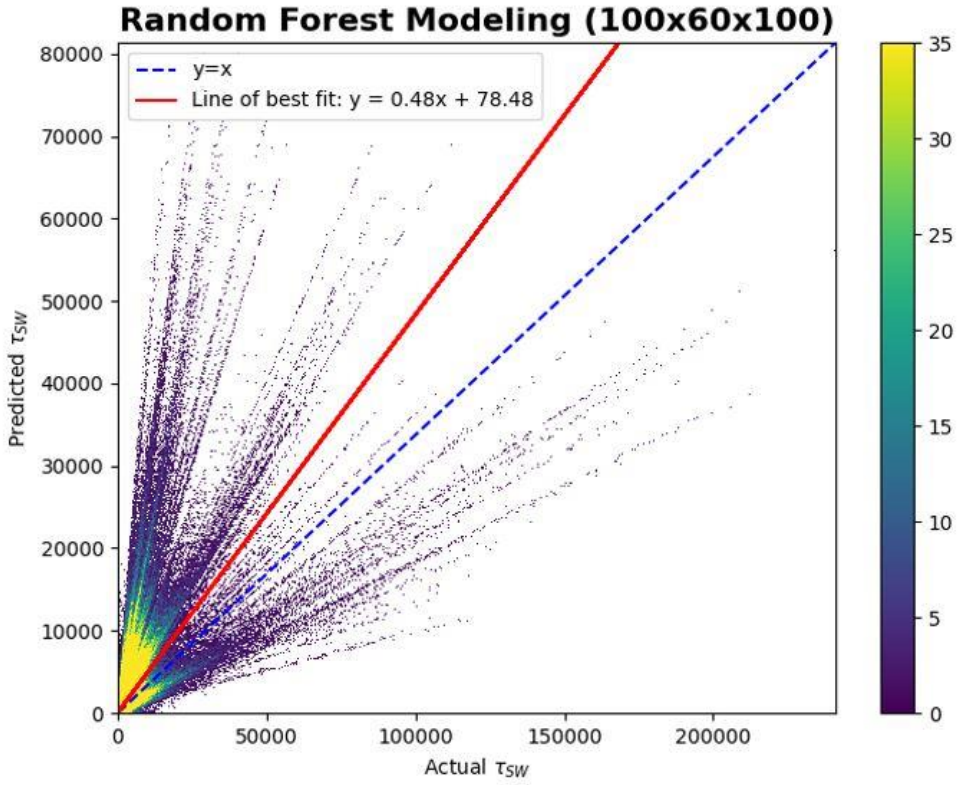
Output	Atmosphere	Dataset	MSE	Accuracy	Maximum Depth	No. of Decision trees
Plank Function	Lower	60	0.156803	0.687026	20	500
		100	0.084088	0.823788	20	200
	Upper	60	0.044074	0.779488	20	500
Opacity (Short wavelength)	Lower	60	1315943	0.257394	20	500
		100	1767329	0.470325	20	500
	Upper	60	3467961	0.230475	20	100
Opacity (Long wavelength)	Lower	60	29990350	0.596966	20	200
		100	15090078	0.748676	20	500
	Upper	60	6121689	0.187367	20	500
Single Scattering Albedo	Lower	60	0.016479	0.877228	20	500
		100	0.006626	0.951047	20	200
	Upper	60	0.034433	0.752852	20	500

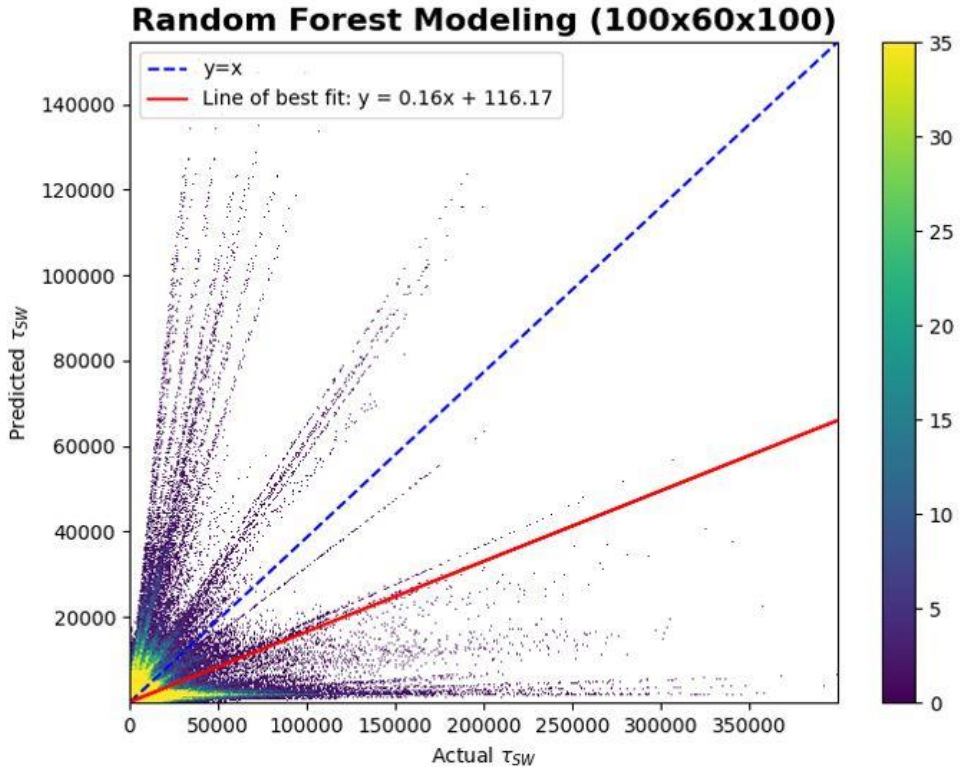
Density Scatter Plots of Predicted v/s Actual Planck Source Function



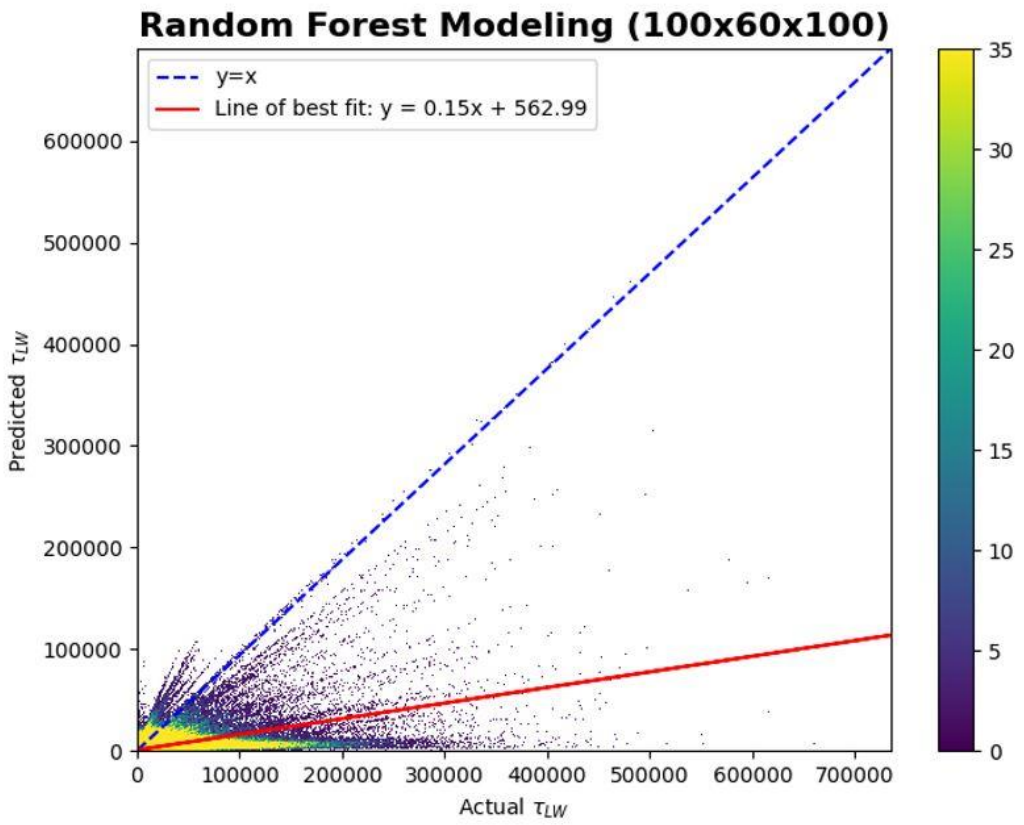


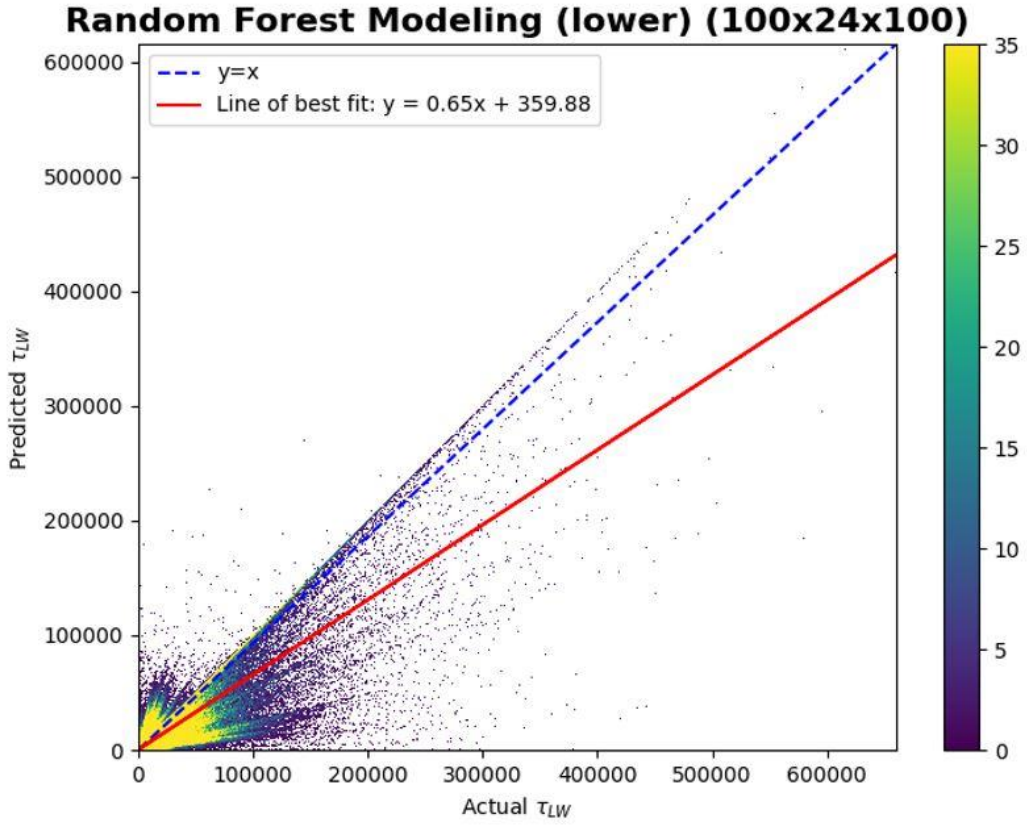
Density Scatter Plots of Predicted v/s Actual Optical depth (Shortwave)



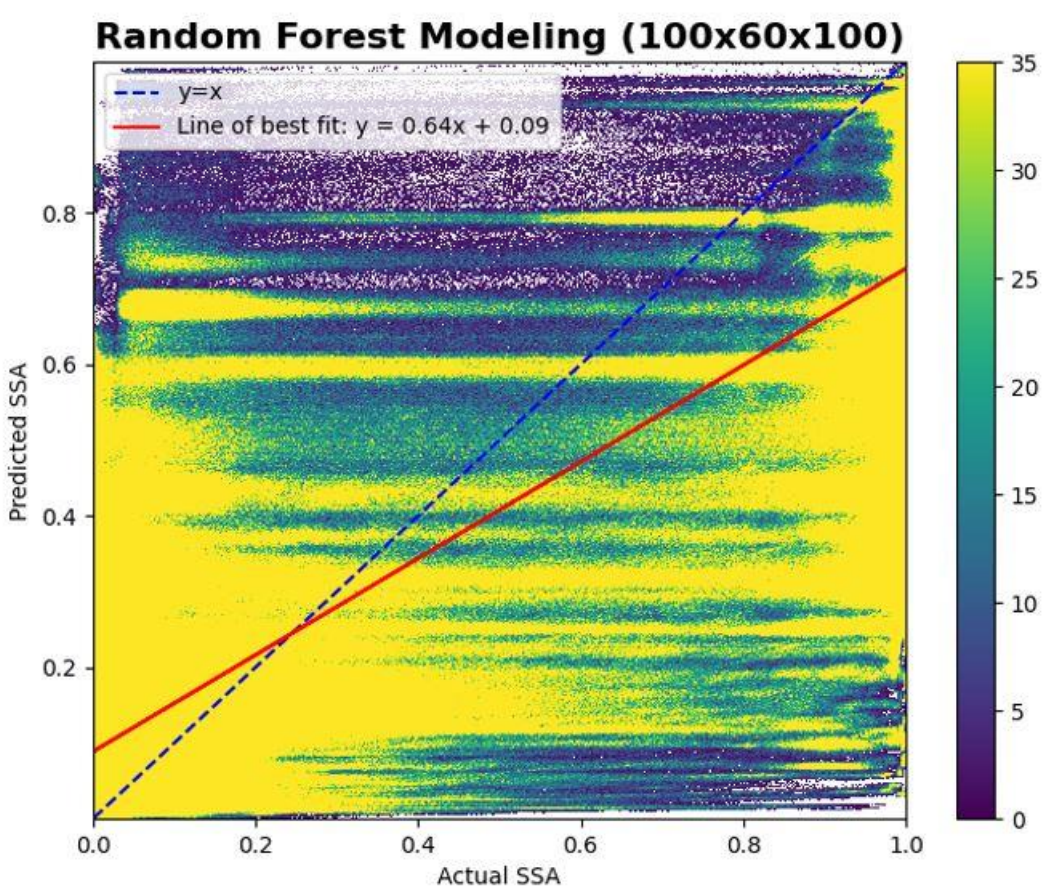


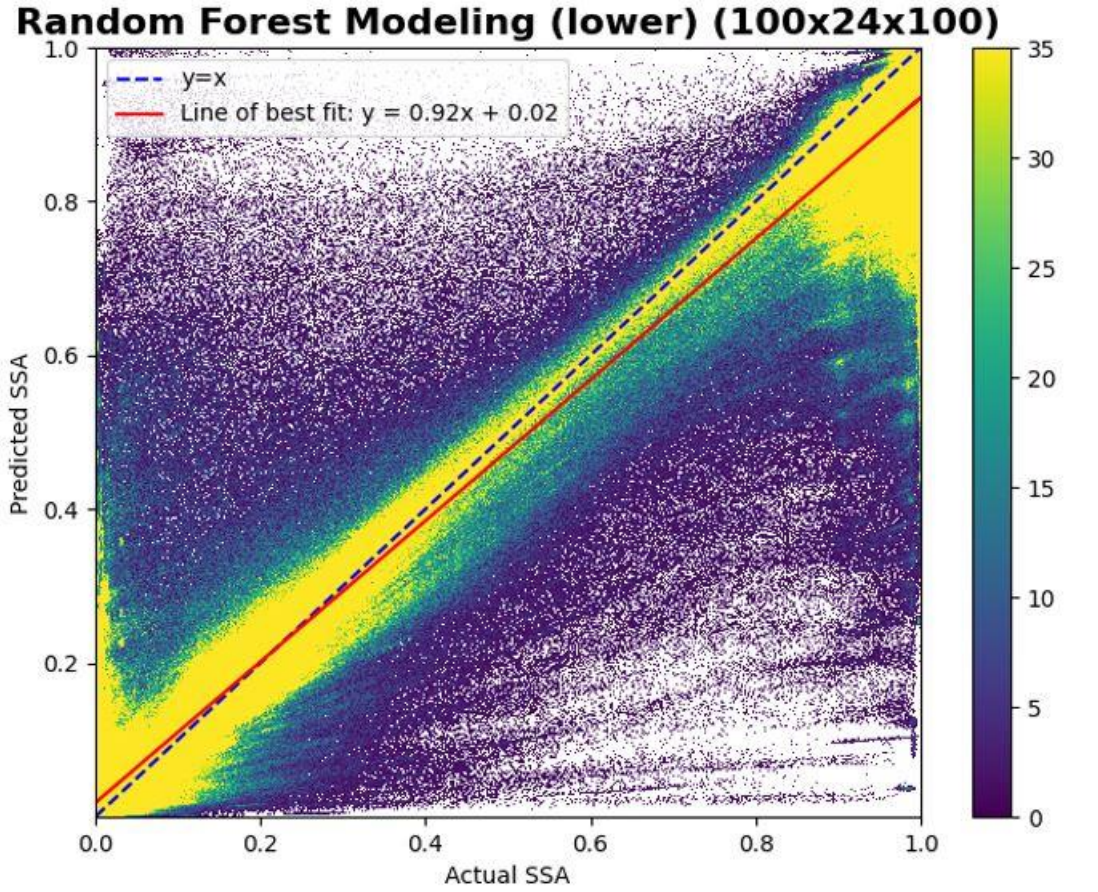
Density Scatter Plots of Predicted v/s Actual Optical depth (Longwave)





Density Scatter Plots of Predicted v/s Actual SSA





CONCLUSIONS

The Random Forest model we developed shows great accuracy for SSA prediction. The accuracy was enhanced when the layers were slice into upper and lower atmosphere.

Hyperparameter tuning by grid search on RF showed that a maximum depth of 20 is the best hyperparameter

Pipelines for RF, XG Boost and Neural Networks are ready to train bigger datasets for better accuracy, provided memory limit issues are resolved.

FURTHER PLANS

To implement the pipelines on other models and datasets.

Once the results are obtained, we'll move forward for publications.