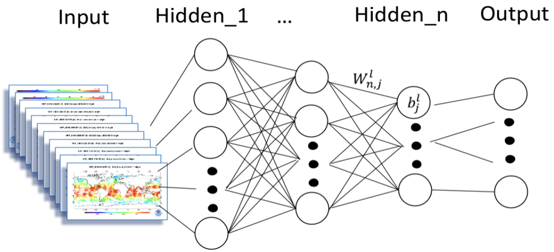
The effect of different surface and atmosphere states on AI_CRTM

Intern: Sungmin Park & Zhuoyu Yang

Mentor: Xingming Liang

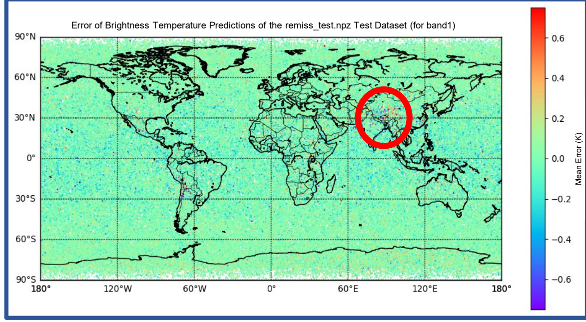
Cooperative Institute for Satellite Earth System Studies

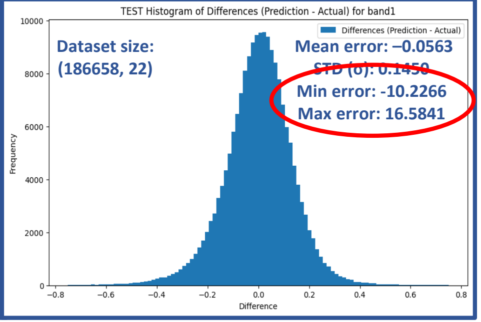
Objective:
Use deep learning to predict ATMS brightness temperatures for 22 bands using CRTM, ECMWF, and ATMS data and also evaluate the impact of various surface and atmosphere states, aiming to improve the model's accuracy across diverse environmental conditions.



Used an 8-hidden-layer residual network consisting of 3 residual blocks with 2 hidden layers each, Batch Normalization, LeakyReLU, Adam optimizer with a clipnorm of 0.5, a custom dropout rate scheduler, and a learning rate scheduler; loss (MSE) started at over 50,000 but reduced to ~0.02 over a span of 2,000 epochs.

	Remiss_test.npz	Separately Preprocessed Data
Training Loss (MSE)	0.0172 K	0.0223 K
Training Data Size	(1493268, 22)	(1297415, 22)





- The input data is extracted from ECMWF and ATMS SDR product. The model reference is a CRTM simulation.
- For the initial test, we used the entire global dataset to train the model. We can see that the results contain more error in the Tibet area. This is due to the data imbalance in the surface pressures—the Tibet area has a surface pressure of ~500, which is much smaller compared to that of the ocean and other land areas.
- To solve this problem, we preprocessed the data to make each bin have ≤50,000 samples, to make it more balanced.

Separately Preprocessed Data by Intern Zhuoyu Yang

Sungmin Park & Zhuoyu Yang || Mentor: Dr. Xingming Liang

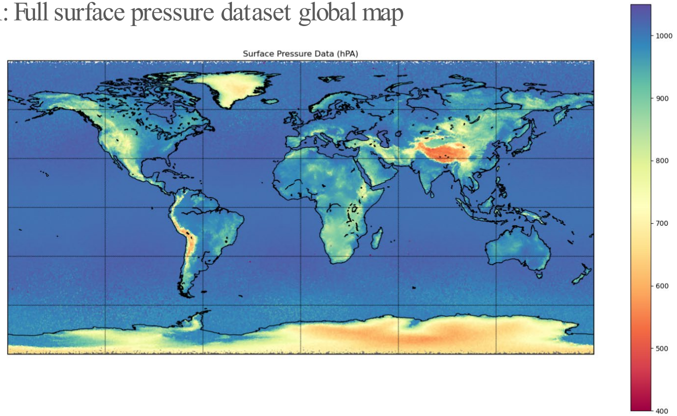
Data Preprocessing is a necessary step in ML to improve data quality and model accuracy.

Input data was extracted from ECMWF and ATMS SDR product and CRTM simulation used as a model reference

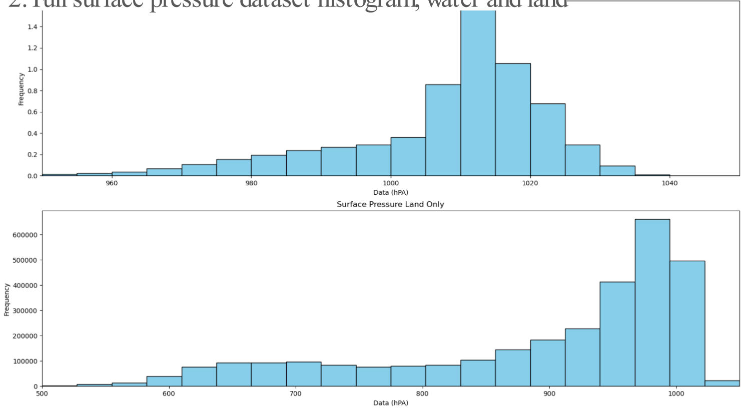
Data was highly concentrated around 1000 hPA and has 3x as many measurements over water than land.

To reduce these biases, we need to control the data distribution.

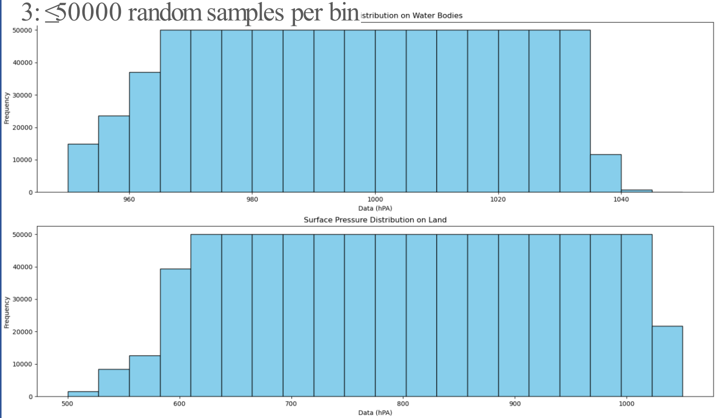
1: Full surface pressure dataset global map



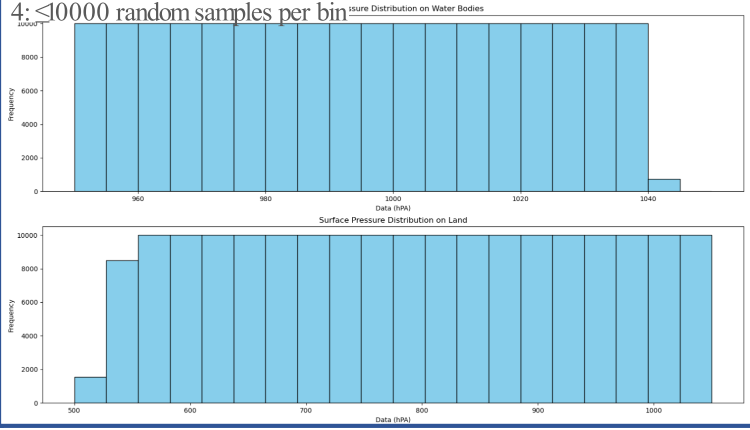
2: Full surface pressure dataset histogram, water and land

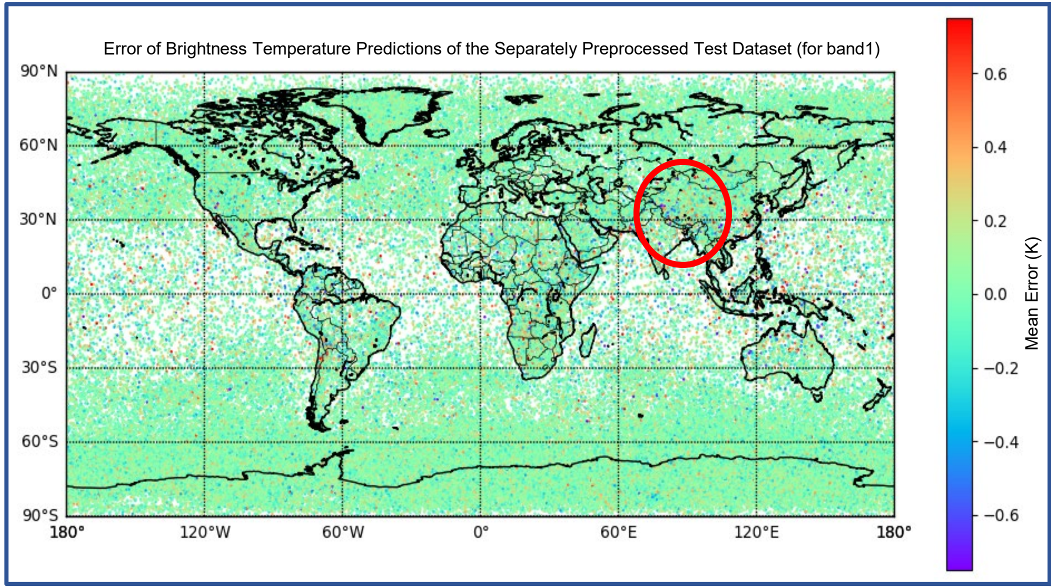


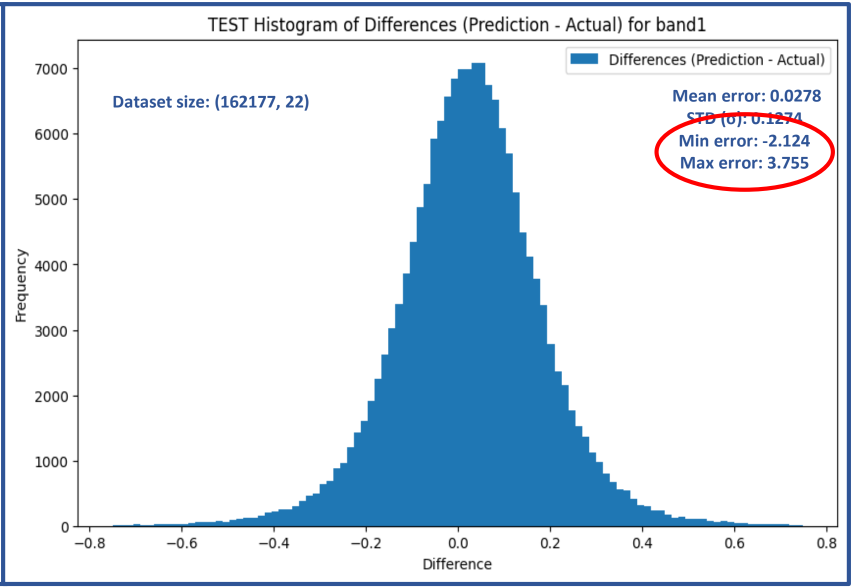
3: ≤50000 random samples per bin



4: ≤10000 random samples per bin







Band #	1	2	3	4	5	6	7	8	9	10	11
Mean error:	0.0278	0.0134	0.0069	-0.0152	-0.0270	-0.0108	-0.0044	-0.0030	-0.0027	-0.0071	-0.0076
STD (σ):	0.1274	0.1253	0.1772	0.1662	0.1805	0.1395	0.1050	0.0971	0.0772	0.0881	0.0811
Min error:	-2.1244	-1.8962	-2.7244	-3.8911	-9.5071	-9.3607	-5.9977	-4.0449	-2.4672	-2.6962	-0.7371
Max error:	3.7552	2.9424	10.335	8.3044	34.197	2.2867	1.6642	1.8112	0.7607	0.7667	0.8107

Band #	12	13	14	15	16	17	18	19	20	21	22
Mean error:	-0.0093	-0.0095	-0.0060	-0.0087	0.0184	-0.0208	-0.0225	-0.0136	0.0043	0.0002	0.0241
STD (σ):	0.0854	0.0932	0.0978	0.1045	0.1854	0.2201	0.2114	0.2085	0.2195	0.2101	0.2279
Min error:	-1.1291	-0.7740	-1.0635	-1.1270	-2.4565	-7.7820	-10.961	-4.7793	-30.321	-6.5577	-2.9279
Max error:	4.2370	7.7358	8.6271	3.5382	3.1490	5.3541	3.1463	9.6498	2.1284	4.8005	4.0736

```
Predictions shape: (162177, 22)
22 BT Predictions for the FIRST data point: [[113.37322 197.04955 134.62224 195.33714 230.7608  234.68584 221.7769
  214.22105 208.9859  207.01625 213.49129 226.56407 241.96156 255.78293
  265.91803 182.82278 219.5138  208.27809 244.66225 245.70848 245.8178
  242.54153]]
Actual BT Values for the FIRST data point: [[113.27304 196.96913 134.96556 195.37862 230.76569 234.6834  221.76035
  214.22044 209.03139 206.95746 213.48311 226.53082 241.94888 255.83594
  265.9458  182.88925 219.789   208.63339 244.5908  245.85696 245.88008
  242.5045 ]]
```

Conclusion:

- After data preprocessing, to balance the data distribution, each bin has ≤50,000 samples.
- Retraining the model with this data, we can see that, in the Tibet area, the error is mitigated. Thus, the minimum and maximum error values both decrease.
- The global statistics for all 22 bands become comparable or better, particularly the minimum and maximum.
- Using balanced data improves the model accuracy, even if the dataset size decreases.