On the use of CERES SSF dataset to understand different machine-learning algorithms for clear-sky detections in infrared hyperspectral observations

Chongxing Fan[1], Xianglei Huang[1], Xiuhong Chen[1], Qing Yue[2]

[1]University of Michigan, Ann Arbor, Michigan, USA

[2]JPL/Caltech

2021 Fall CERES Science Team Meeting

Oct 13, 2021

Hyperspectral Remote Sensing

Clear-sky detection from rich spectral information can be beneficial

- Application:
  - Data assimilation
  - Trace gas retrieval
  - Clear-sky flux estimation
- Imager like MODIS
  - Fine spatial resolution but limited spectral channels
- Hyperspectral sounder like AIRS
  - Thousands of spectral channels but coarse spatial resolution

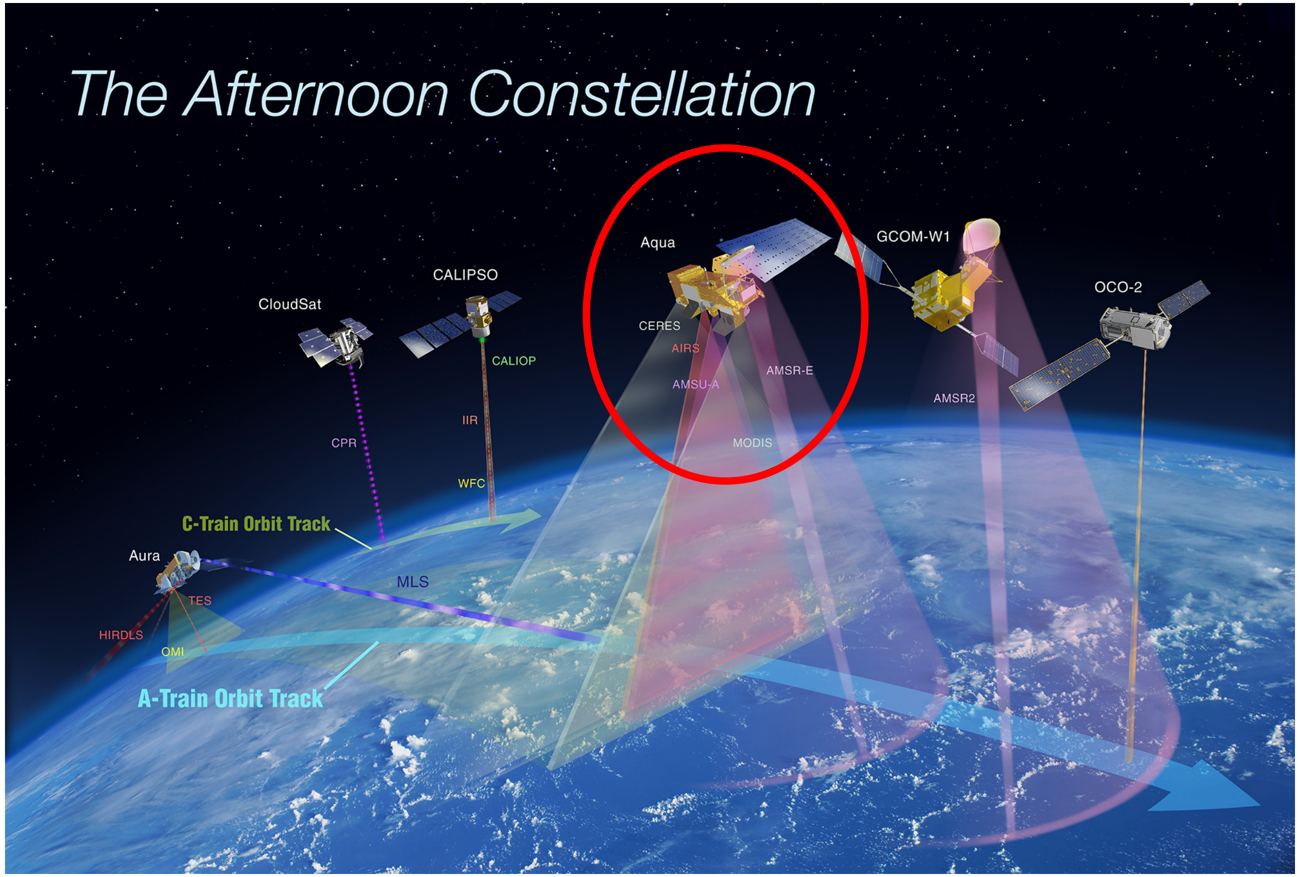


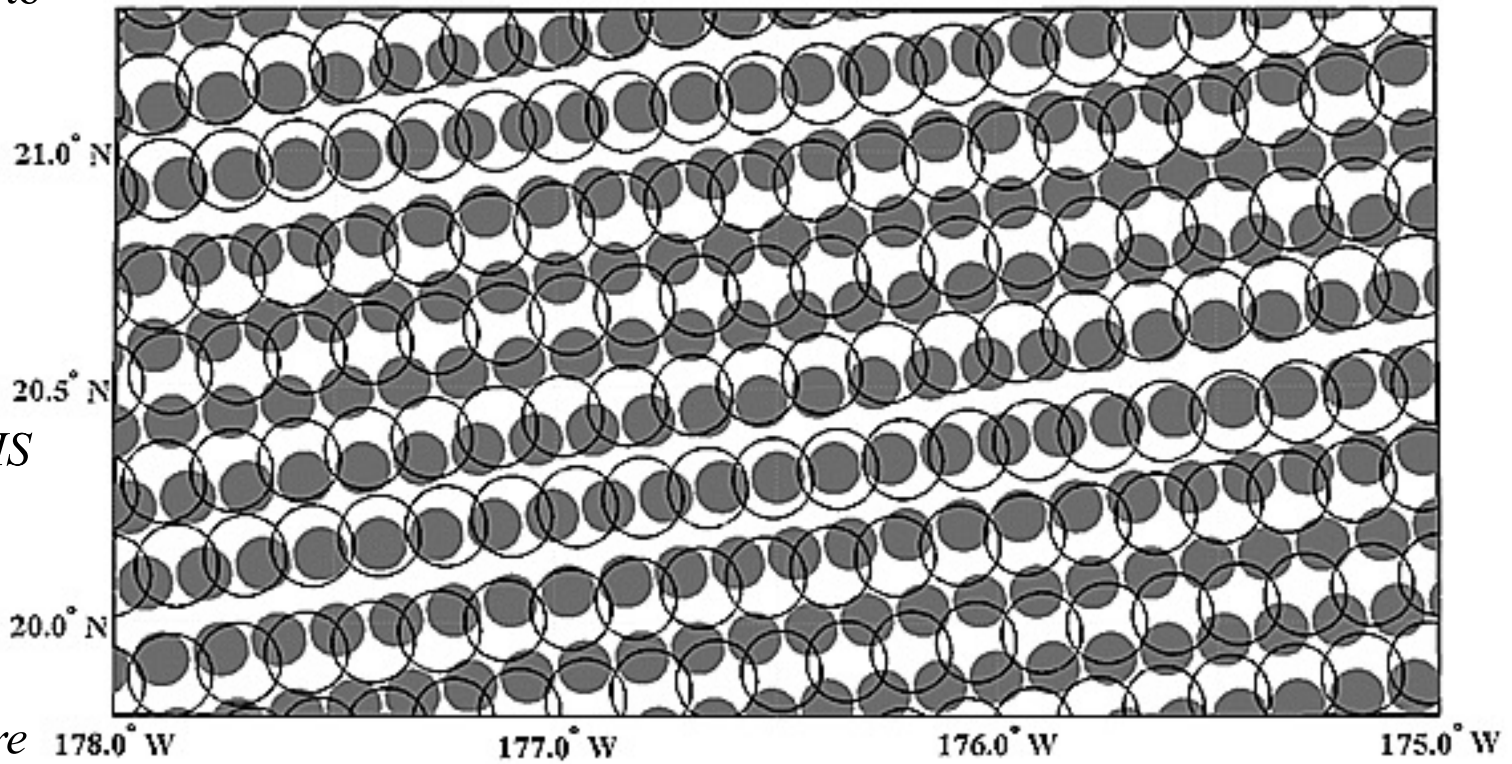
Image from NASA (https://atrain.nasa.gov/)

Collocations of CERES and AIRS footprints

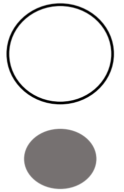
We have used such collocation advantage to
1. derive spectral flux from AIRS
2. assess the stability of FM3 over the years
3. retrieve CFC-11 from AIRS and CrIS (Chen et al., 2020)
4. evaluate GEOS-5 T/q profiles
5. and here…

- *Can we use pixel-based CERES-MODIS clear-sky detection result to train a clear-sky detection method for the infrared sounder like AIRS or CrIS?*

- *How does such trained method compare to those physical-based methods?*



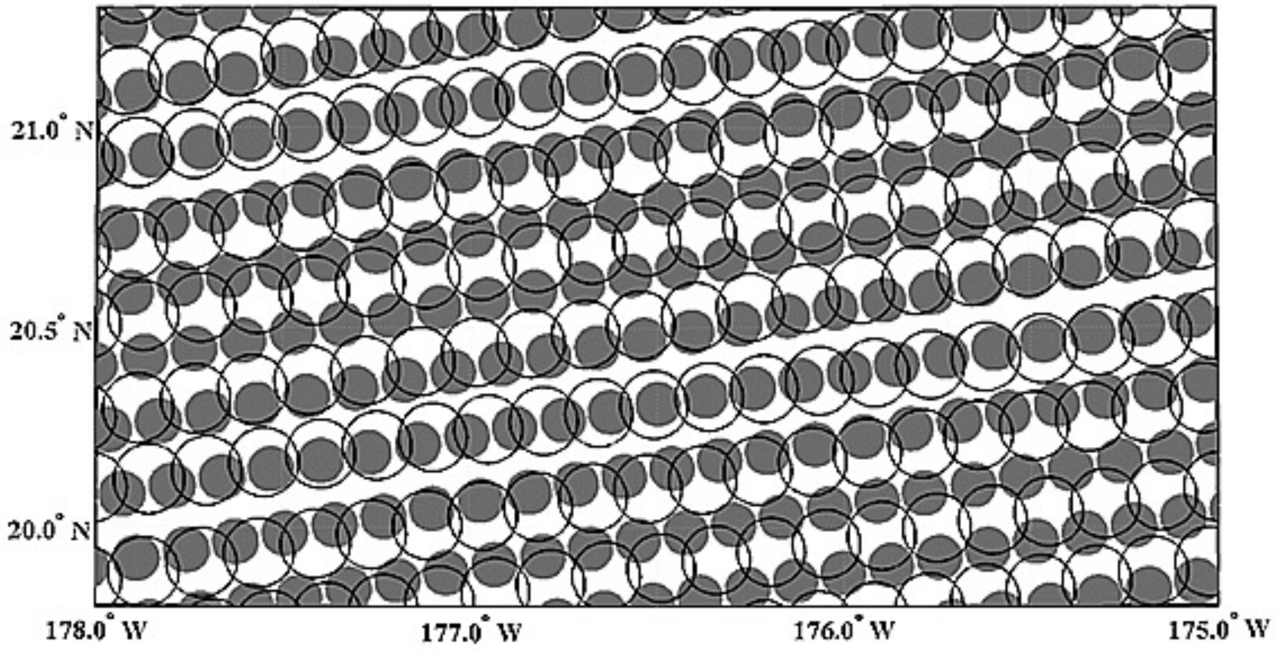
Huang et al. (2008)



Data & Methodology

Clear-sky detection from AIRS infrared hyperspectral observations

- Data:
  - AIRS brightness temperature (BT) from 1,598 thermal infrared channels
  - HadCRUT sea surface temperature (SST)
  - CERES-MODIS Ed4 cloud flag (Minnis et al., 2021)
  - MODIS Cloud Product Collection 6/NASA EOS WorldView
- Data Selection:
  - Nadir view
  - Tropical ocean
- Data Processing:
  - Collocation of CERES and AIRS footprint
  - Equally sampled: cloudy samples = clear samples
  - Normalization $\boldsymbol{X}' = \frac{\boldsymbol{X}-\overline{\boldsymbol{X}}}{\sqrt{S(\boldsymbol{X})}}$
  - Grid hyperparameter tuning



Huang et al. (2008)

Goal of Our Study

Clear-sky detection from AIRS infrared hyperspectral observations

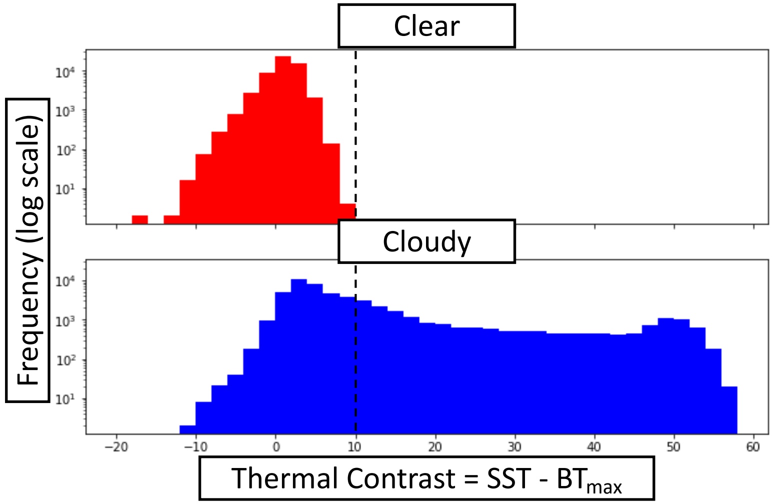
- Machine learning models:
  - Linear-kernel support vector classifier (LinearSVC)
  - Random forest classifier
  - Gradient boosting classifier
  - Fully-connected artificial neural network (FNN)
  - 1D convolutional neural network (CNN)
- Compared to common physically-based algorithms:
  - Bispectral method (cloudy if $BT_{8\mu m} - BT_{11\mu m} > 0$)
  - Thermal contrast threshold method (cloudy if $SST - BT_{max} > \sigma$)
- Spectral classification, without spatial and temporal information
- No visible and near-infrared channels, assessing the potential of exploiting abundant spectral information in the thermal radiation spectrum for clear-sky detection

Evaluation

Training on year 2004 (106,558 spectra), evaluation on year 2008 (1,159,955spectra)

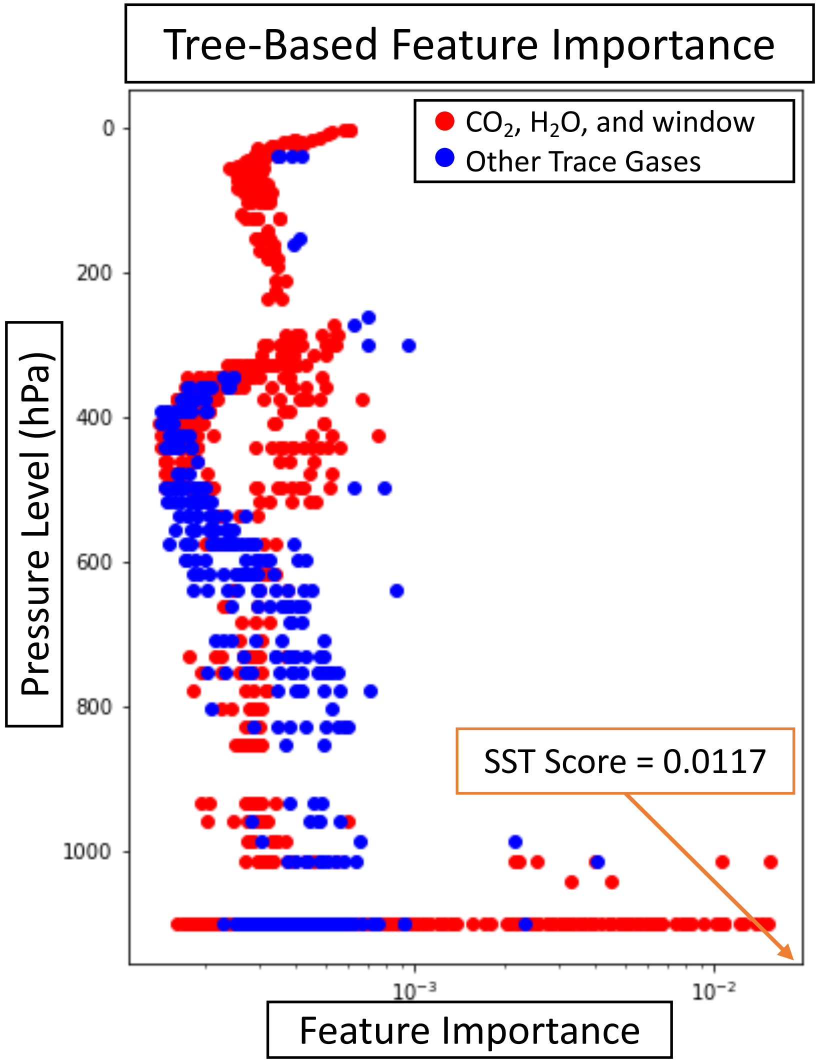
Model Name	True Positive	False Negative	True Negative	False Positive	Accuracy
LinearSVC	78.96% (2)	15.57% (2)	4.97% (4)	0.50% (4)	83.93% (2)
RF	74.45% (5)	20.08% (5)	5.03% (2)	0.45% (2)	79.48% (5)
GB	75.46% (4)	19.07% (4)	4.98% (3)	0.49% (3)	80.44% (4)
FNN	79.06% (1)	15.46% (1)	4.92% (5)	0.55% (5)	83.98% (1)
1D-CNN	77.61% (3)	16.91% (3)	5.09% (1)	0.39% (1)	82.70% (3)
Thermal Contrast	34.14%	60.39%	5.47%	0.00%	39.61%
Bispectral Algorithm	21.56%	72.97%	5.47%	0.00%	27.03%

- ✓ Traditional algorithms perform poorer than all machine learning models
- ✓ LinearSVC and 1D-CNN are slightly superior
- ✓ Relatively balanced cloudy-sky detection rate (~81%) and clear-sky rate (~91%) despite of severely imbalanced dataset



Feature Importance Analysis

ML models can exploit refined structure of the contrasts between BT and SST



- ✓ Bands with weighting function peak on the ground have relatively greater importance
- ✓ High thermal contrast samples correctly predicted by all models
- ✓ SST is a significant predictor in all models. Nevertheless, eliminating SST variable has little impact on 1D-CNN

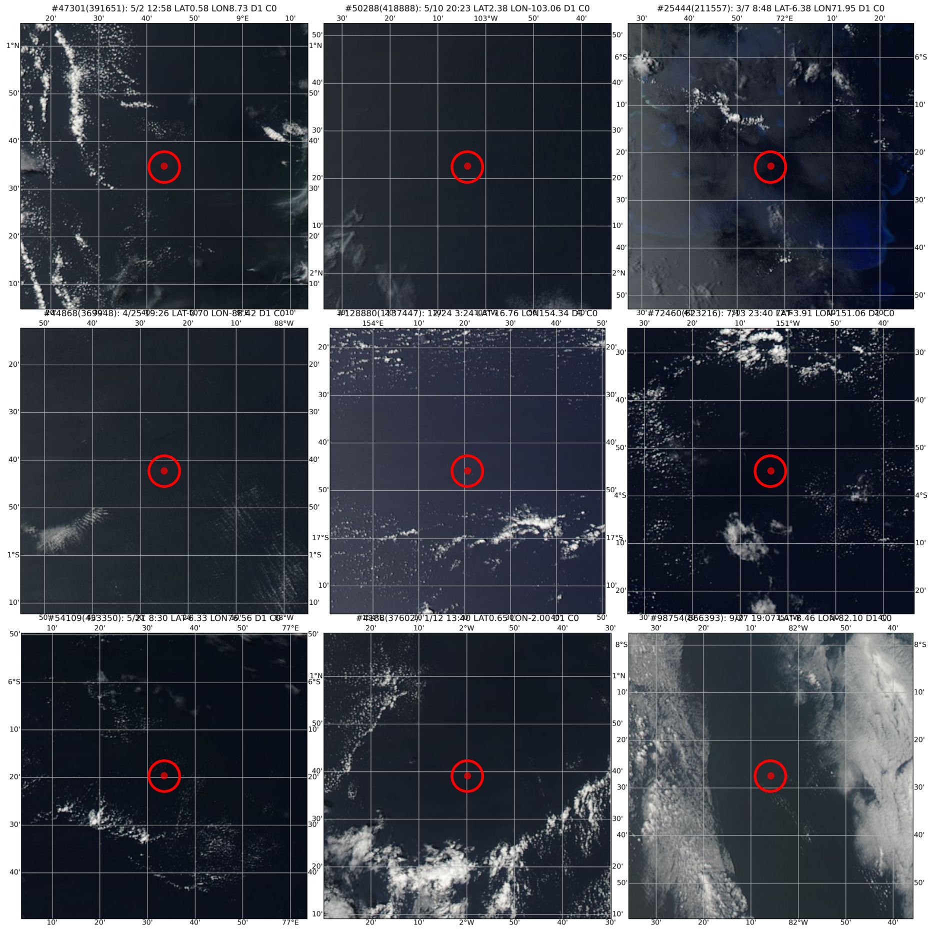
Model Name	Accuracy (All)	Accuracy ($SST-BT_{max}$>10K)
LinearSVC	83.93%	100.00%
RF	79.48%	100.00%
GB	80.44%	100.00%
FNN	83.98%	100.00%
1D-CNN	82.70%	100.00%

UNIVERSITY OF MICHIGAN

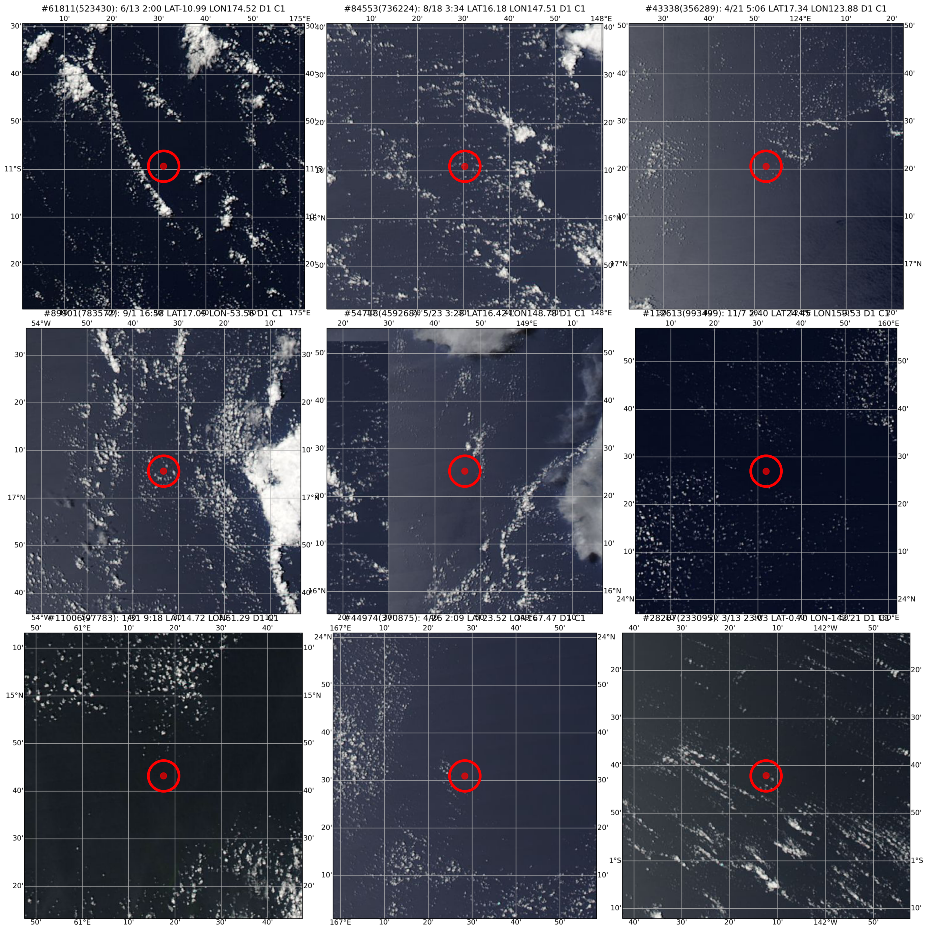
Case Study

Enumerate the cases where models succeed or fail in prediction

Clear-sky Scenes
All models fail



Cloudy-sky Scenes
All models fail

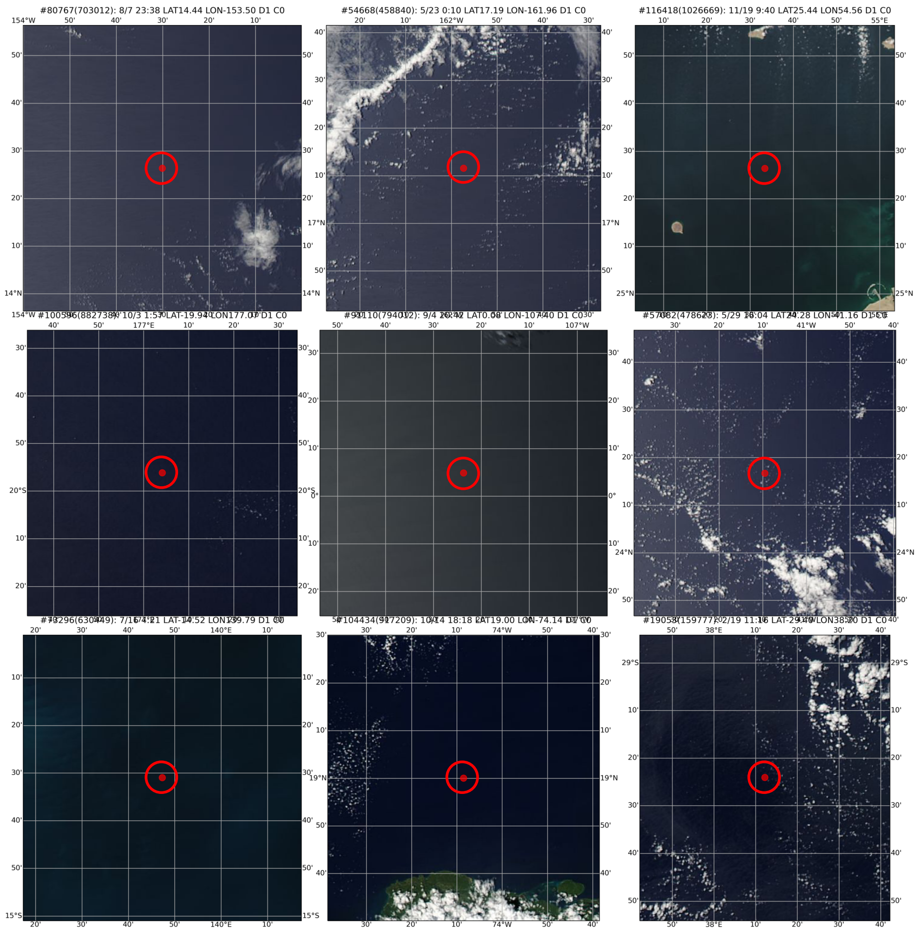


MODIS Corrected Reflectance (True Color) – EOS WorldView

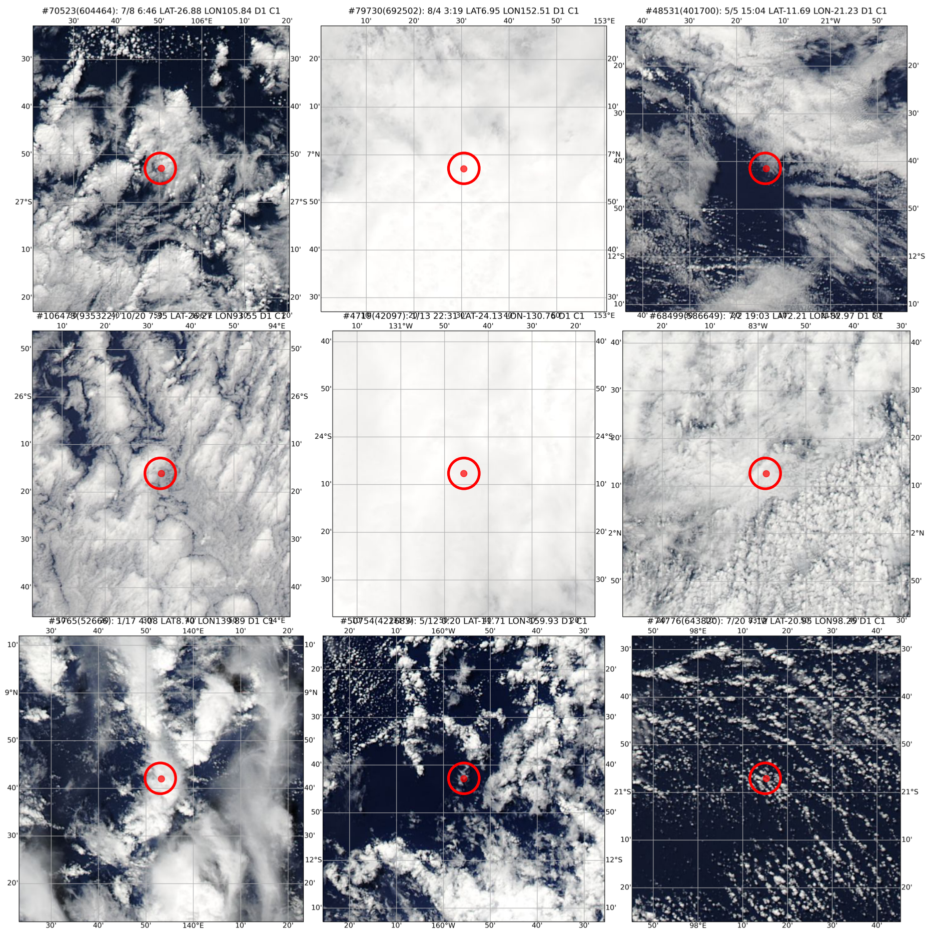
Case Study

Enumerate the cases where models succeed or fail in prediction

Clear-sky Scenes
All models succeed



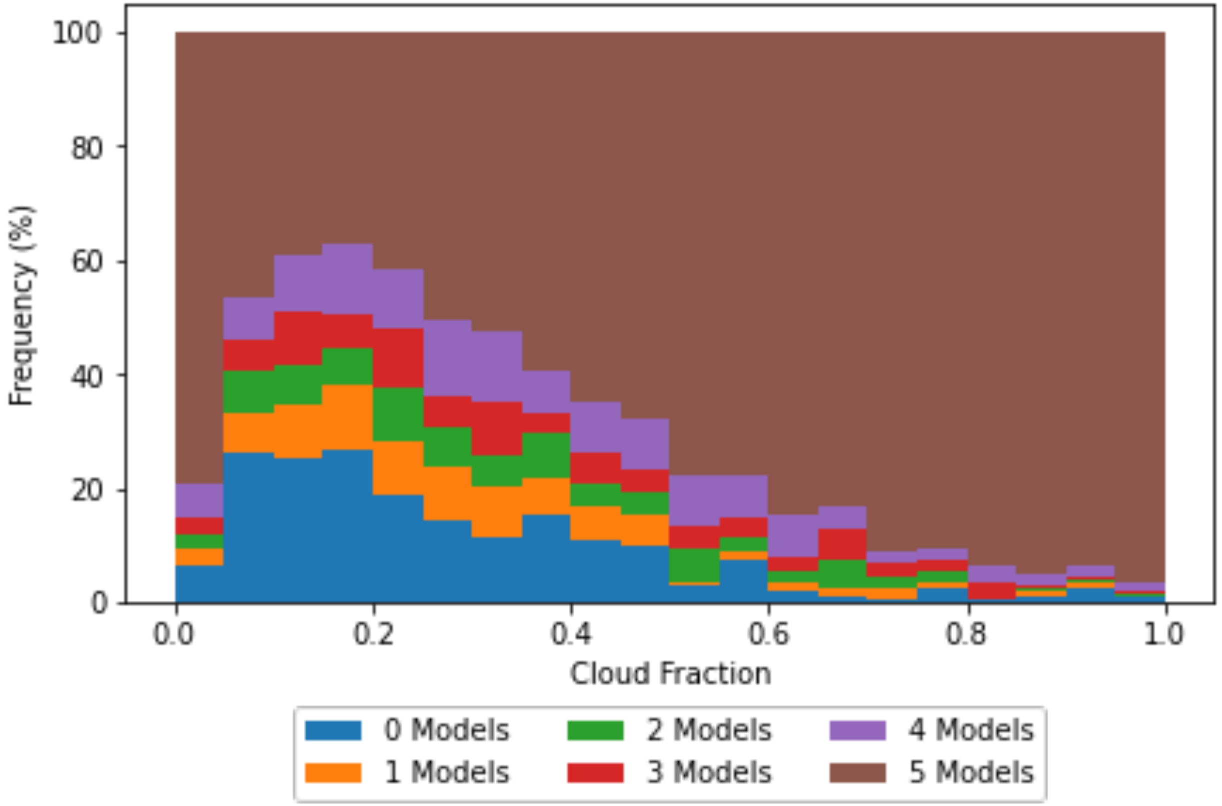
Cloudy-sky Scenes
All models succeed



MODIS Corrected Reflectance (True Color) – EOS WorldView

Error Analysis

Broken clouds are one of the error sources

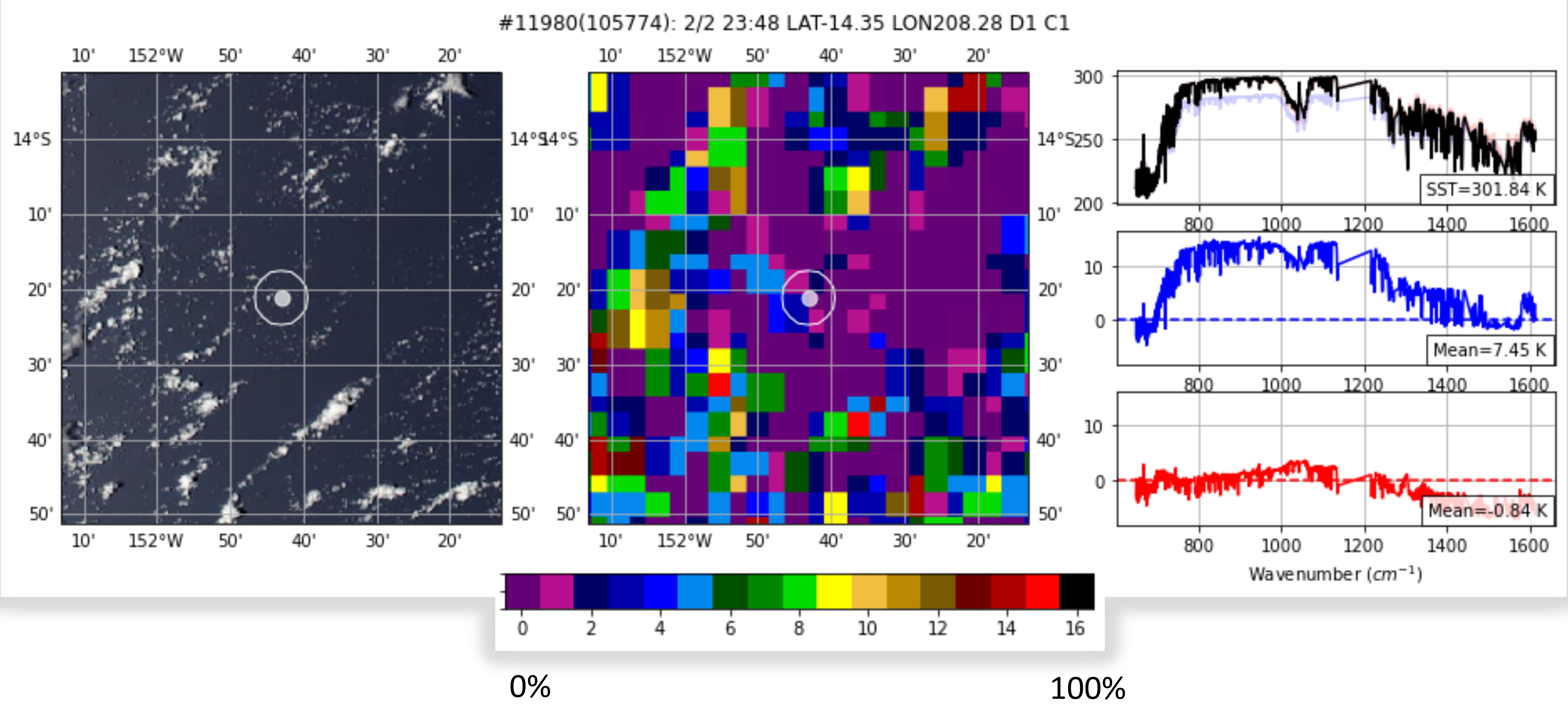


- ✓ Increasing error rate when the cloud fraction decreases from 1 to ~0.2.
- ✓ Slight decreasing error rate when the cloud fraction decreases from 0.2 to 0.05.

Cloud fraction from MOD06/NASA EOS WorldView

Error Analysis

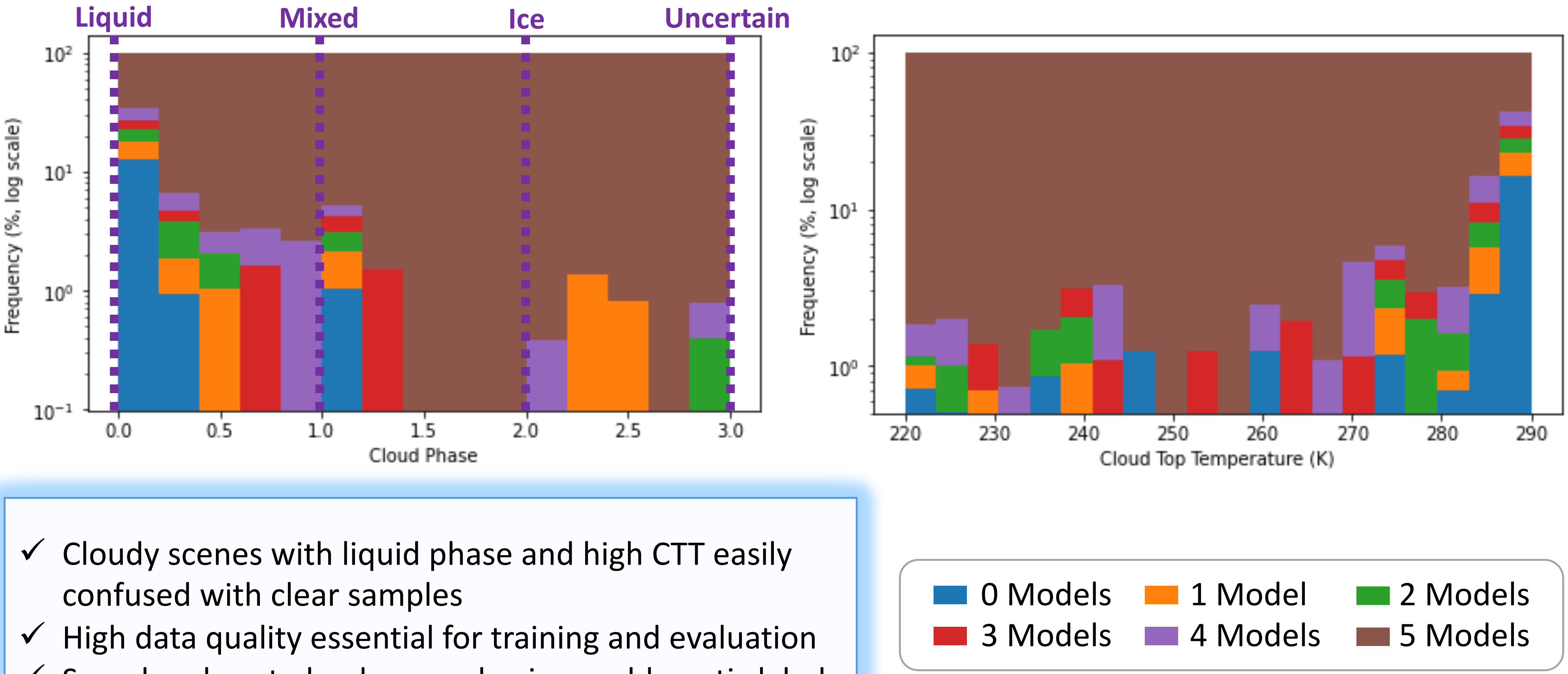
Broken clouds are one of the error sources



UNIVERSITY OF MICHIGAN

Error Analysis (true cloud scenes)

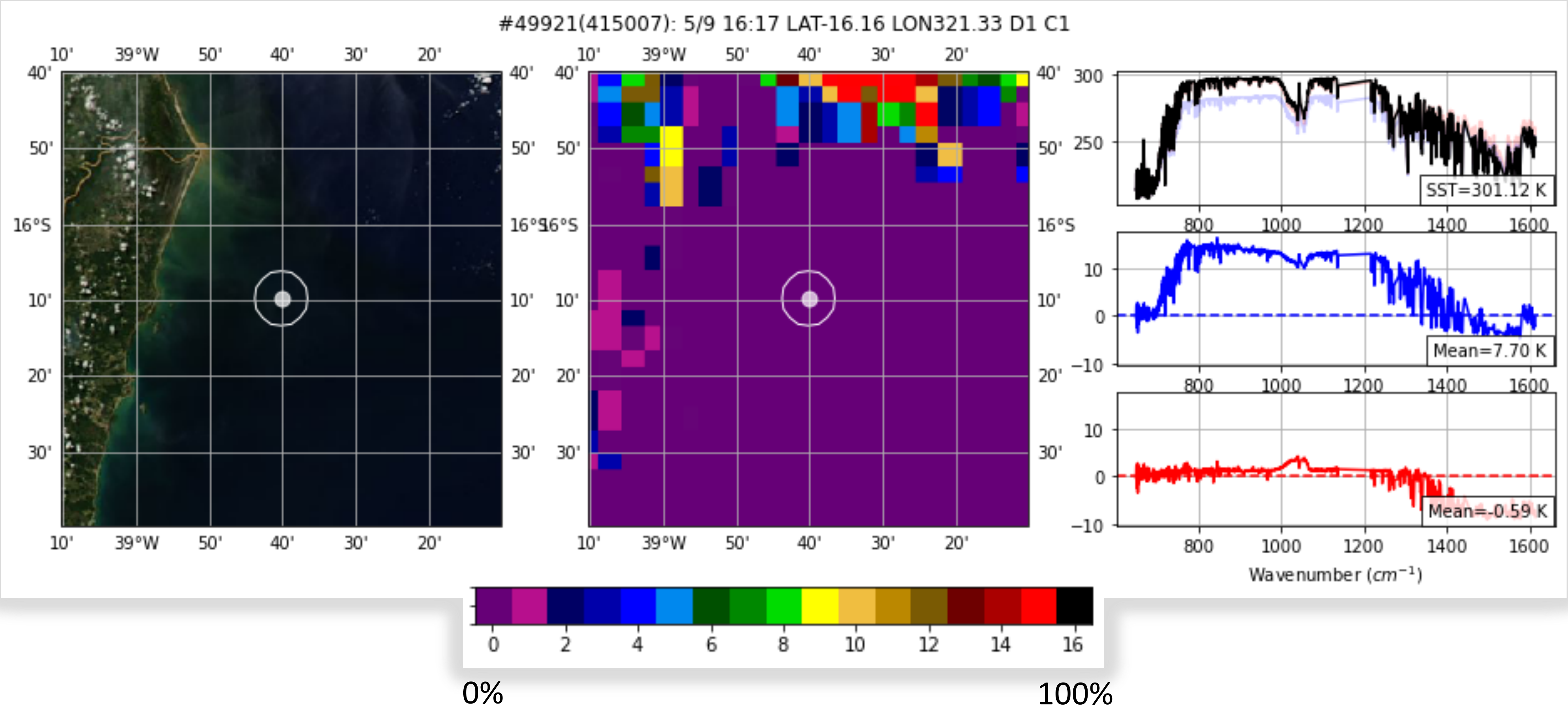
Indistinguishable low clouds and data quality issue are likewise critical



- ✓ Cloudy scenes with liquid phase and high CTT easily confused with clear samples
- ✓ High data quality essential for training and evaluation
- ✓ Samples close to landmasses having problematic labels

Error Analysis

Indistinguishable low clouds and data quality issue are likewise critical



Take-home Messages

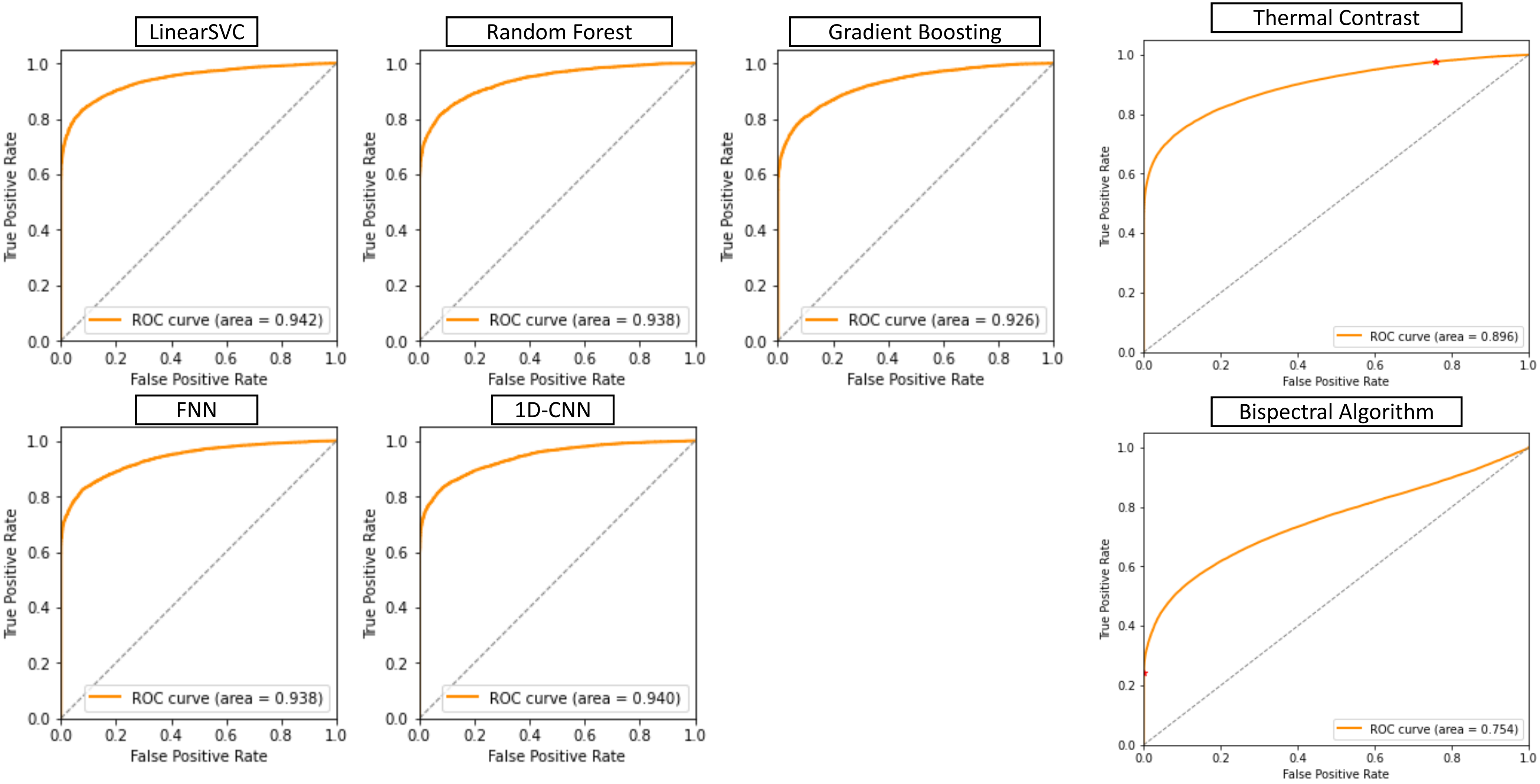
From feature importance attribution and error analyses to physical interpretability

- In terms of clear-sky detection, ML models performs well
  - Linear SVC and 1D-CNN slightly better than others;
  - ML feature importance can be related to the physics.
- Broken clouds and low clouds are responsible for most errors.
- Training data quality is also critical
- As a preliminary study, we might not fully unleash the power of ML learning techniques yet
  - Spatial information is not exploited yet
  - Impose *a priori* channel correlations in the detection

Thank You!

References:

1. Huang, X., Yang, W., Loeb, N. G., & Ramaswamy, V. (2008). Spectrally resolved fluxes derived from collocated AIRS and CERES measurements and their application in model evaluation: Clear sky over the tropical oceans. *Journal of Geophysical Research: Atmospheres*, *113*(D9).
2. Minnis, P., Sun-Mack, S., and Co-authors (2021). CERES MODIS Cloud Product Retrievals for Edition 4—Part I: Algorithm Changes. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(4), 2744–2780.
3. Chen, X. H., X. L. Huang, and L. L. Strow (2020), Near-global CFC-11 Trends as Observed by Atmospheric Infrared Sounder from 2003 to 2018, JGR-Atmospheres, 125(22), e2020JD033051.


LinearSVC
ROC curve (area = 0.942)
Random Forest
ROC curve (area = 0.938)
Gradient Boosting
ROC curve (area = 0.926)
Thermal Contrast
ROC curve (area = 0.896)
FNN
ROC curve (area = 0.938)
1D-CNN
ROC curve (area = 0.940)
Bispectral Algorithm
ROC curve (area = 0.754)
True Positive Rate
False Positive Rate