Improvements in Cloud Detection Using Simple Machine Learning Models

Charles H. White[1], Andrew K. Heidinger[2], Steven A. Ackerman[1]

[1]Department of Atmospheric and Oceanic Sciences, University of Wisconsin – Madison, Madison WI

[2]Center for Satellite Applications and Research (STAR), NOAA/NESDIS, Madison, WI

Summary

Cloud masks are one of the most fundamental cloud products derived from satellite imagers with implications for clear-sky products, cloud-property algorithms, assimilating sounder radiances and other applications. Here, we detail our exploration of gradient boosted methods to predict the presence of clouds from Visible Infrared Imaging Radiometer Suite (VIIRS-SNPP) observations. We use the Clouds from AVHRR Extended (CLAVR-x) cloud mask (Heidinger et al. 2014) as our baseline for comparison, and the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) as our 'truth' dataset.

Overall, this model performs very well compared to CLAVR-x with exceptional improvements over nighttime snow and ice.

Datasets

- One year (2016) of collocations between VIIRS (SNPP) and CALIOP are used to train and evaluate this model
- Collocations are only used under three conditions
  - Time difference between the two platforms < 8 minutes
  - CALIOP cloud optical depth equal to 0 or > 0.01
  - CALIOP 5 km cloud fraction equal 0 or 1.0 (no cloud edges)
- Collocations are split into 3 groups
  - Training set is every other day in 2016 (~50% of all data)
  - Validation and test sets are evenly split from remainder (~25% each)
- 9.92 million globally-distributed clouds are used in training and validation of this model
- Additional information obtained from clear-sky radiative transfer simulations and model reanalysis

Inputs*

Imager Brightness Temperatures: 11μm, 12μm, 8.5μm, 3.75μm

Imager Reflectance: 1.60μm, 1.38μm, 0.65μm, 0.47μm

Clear-Sky Radiative Transfer: 11μm, 12μm, 3.75μm, 0.65μm

Geographic Information: latitude, land/snow/ice cover, coastlines

Other Ancillary Data: $T_{Surface}$, 3.75μm surface emissivity

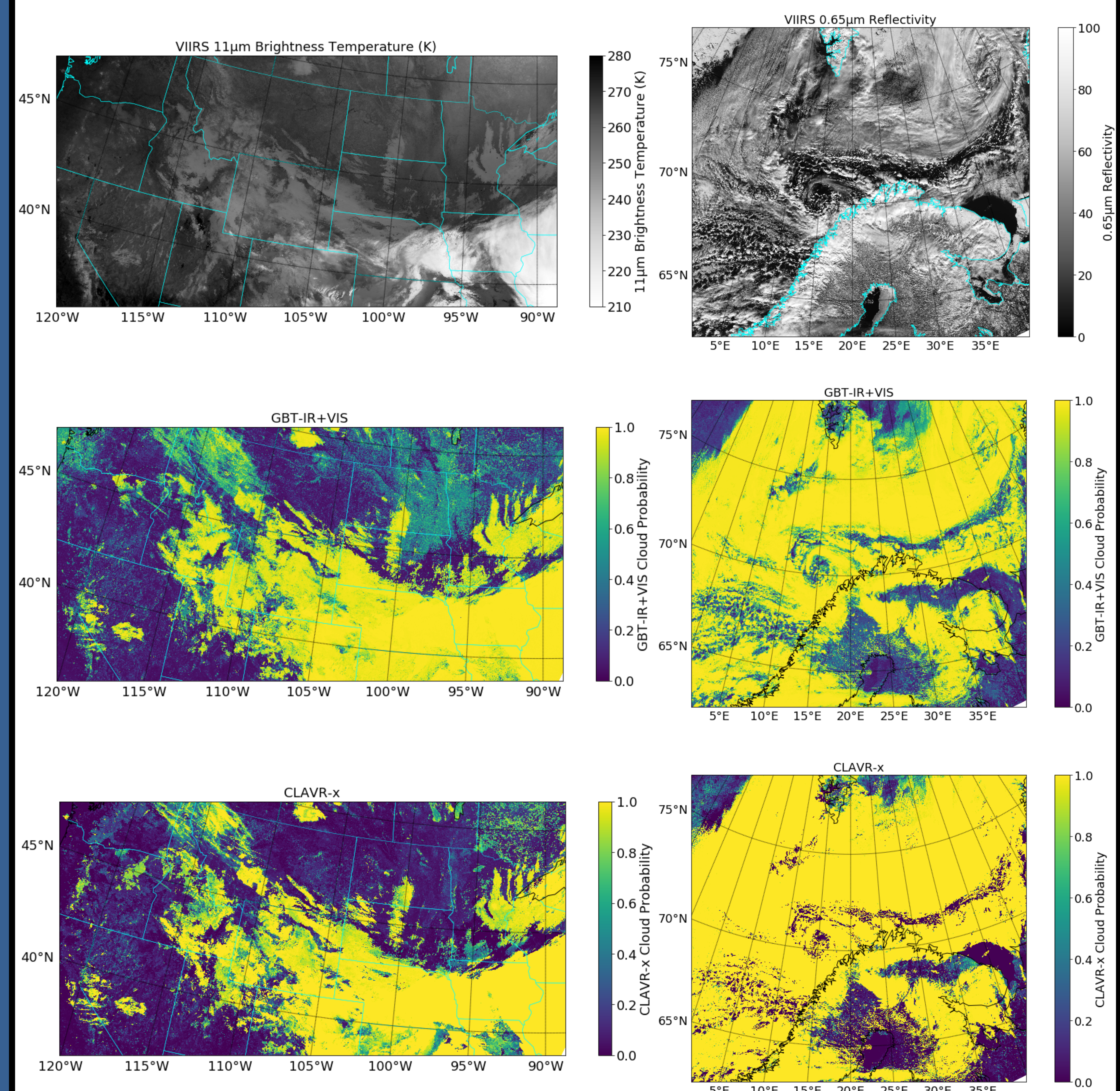
Observed/Clear-Sky Differences: $BT_{11\mu m} - BT_{11\mu m,\ clear\text{-}sky}$

Other Cloud Tests: $BT_{11\mu m} - BT_{12\mu m}$, $BT_{11\mu m} - BT_{3.75\mu m}$

*This list is representative, but not comprehensive. Contact Charles White at cwhite25@wisc.edu for full variable list

Qualitative Comparison

Continental US (Night) Norwegian Sea (Day)



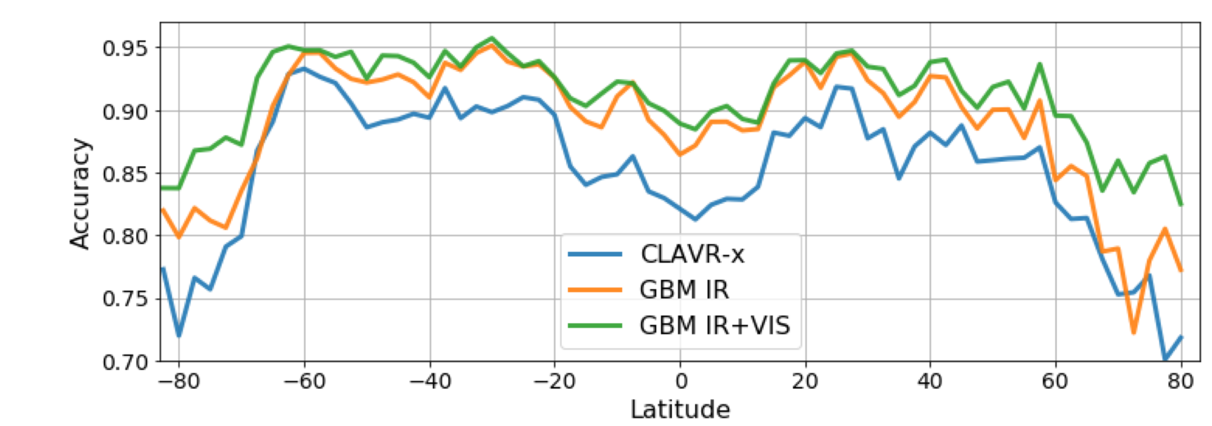
Model Details

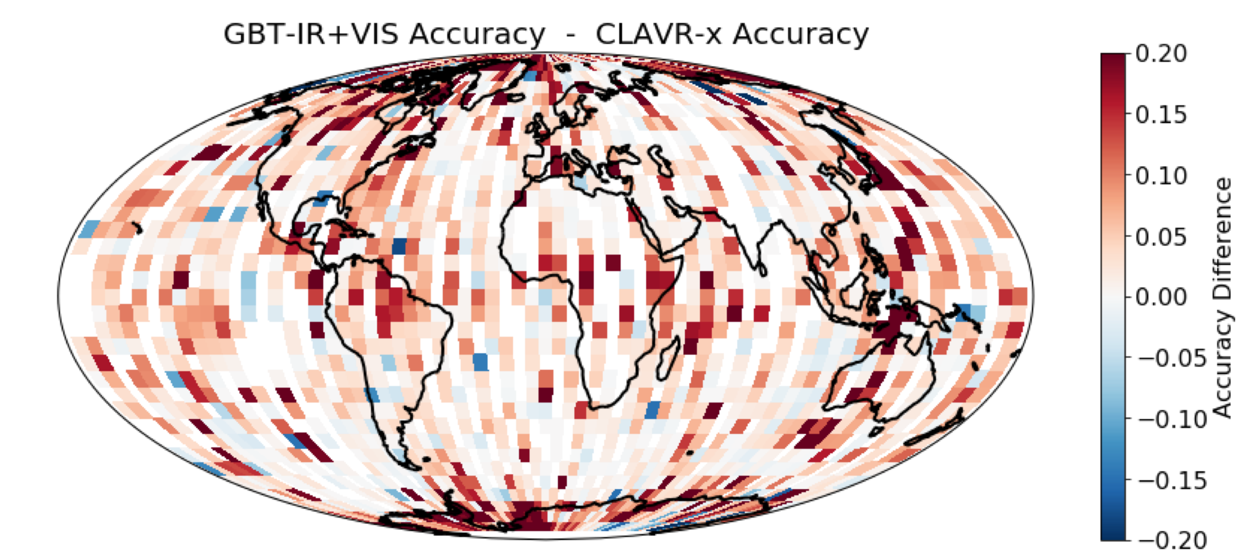
- LightGBM framework (Ke et al. 2017; https://github.com/Microsoft/LightGBM)
- Two gradient boosted decision tree models are made: one with only infrared observations, and another with both infrared and visible observations
- The models are made with a maximum of 150 leaves for each tree, 50% of all features sampled at each split, and a minimum of 1,000 observations at each leaf. The learning rate was set to $0.2(0.98)^{n-1}$ for the n^{th} iteration with early stopping. The IR model resulted in 29 trees, and the IR+VIS model resulted in 65 trees.
- The classification task is binary (0=clear, 1=cloudy)
- The model output is the mean prediction across the ensemble (between 0 and 1)

Quantitative Comparison

Accuracy when Discriminating $\tau > 0.01$ from $\tau = 0$

	Day Land	Night Land	Day Water	Night Water	Day Ice	Night Ice	Day Snow	Night Snow
GBT-IR+VIS	0.94	0.93	0.95	0.91	0.90	0.84	0.89	0.82
GBT-IR	0.91	0.91	0.94	0.91	0.83	0.75	0.85	0.78
CLAVR-x	0.88	0.87	0.90	0.87	0.83	0.68	0.87	0.69





Main Takeaways

Increased accuracy in all scenarios relative to CLAVR-x

Largest improvements are seen at high latitudes and snow/ice covered scenes during the night

These models are more complex than the CLAVR-x naïve bayesian. While not impossible, model interpretation is more difficult.

Acknowledgments

This work was supported by the JPSS CAL/VAL project. The views, opinions, and findings contained in this report are those of the author(s) and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. Government position, policy, or decision.

References

Heidinger, A. K., A. T. Evan, M. J. Foster, and A. Walther, 2012: A naive Bayesian cloud-detection scheme derived from Calipso and applied within PATMOS-x. *J. Appl. Meteorol. Climatol.*, **51**, 1129–1144, doi:10.1175/JAMC-D-11-02.1.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, 2017: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 3146–3154. https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree