Application of Machine Learning Algorithms to COSMIC-2 QC Metrics

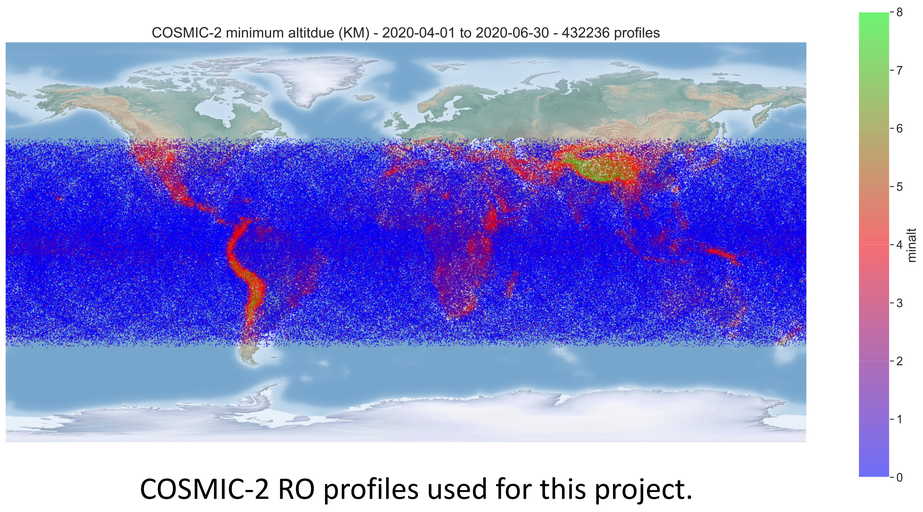
Maggie Sleziak-Sallee, Jan-Peter Weiss, Hanna Huelsing
COSMIC Program, University Corporation for Atmospheric Research, Boulder, CO
Davide Del Vento
Computational & Information Systems Lab, National Center for Atmospheric Research, Boulder, CO

Abstract

We apply data analysis and machine learning to a 3-month FORMOSAT-7/COSMIC-2 (C2) mission neutral atmosphere retrieval dataset comprised of approximately 430K profiles. The dataset includes metrics such as geographic location, local time, signal-to-noise ratio, altitude range, bending angle noise from 60-80km among others. We analyze correlations between all input parameters and with our standard processing quality control (standard-QC) determination of whether a profile is "good" or "bad". We then apply different machine learning classification algorithms to predict good and bad profiles from the set of input parameters. The set profiles determined as bad by standard-QC along with their ML classification results are then compared to numerical weather prediction analysis products. Our goal is to utilize ML to gain a deeper understanding of whether certain metrics can be used to optimize our standard bending angle retrieval quality control procedures.

Introduction

- The COSMIC Data Analysis and Archive Center (CDAAC) is an end-to-end processing and analysis system for ground- and space-based Global Navigation Satellite System (GNSS) data focusing on radio occultation (RO) applications. We process data and publish products from a variety of space missions, including FORMOSAT-7/COSMIC-2 (C2), in near real-time, post-processing, and re-processing modes. Near real-time products are delivered to operational centers for assimilation into weather and space weather analysis and prediction systems
- COSMIC-2 RO retrievals go through a standard processing quality control (standard-QC) process which determines if the profile is "good" or "bad"
- We use data analysis and machine learning (ML) algorithms to gain a deeper understanding of whether certain metrics can be used to optimize our standard bending angle retrieval quality control procedures
- The end goal is to determine whether all profiles currently marked as 'bad' by our standard-QC show the same conditions, or whether some of the flagged bad profiles could potentially be usable



COSMIC-2 RO profiles used for this project.
Color scale represents profile lower altitude.

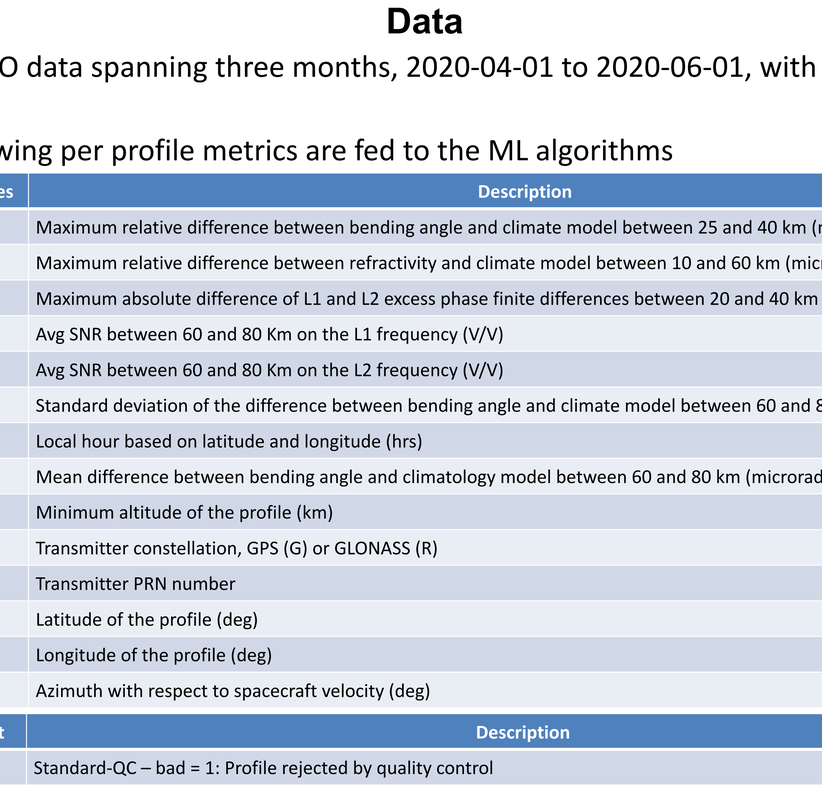
Data

- We use RO data spanning three months, 2020-04-01 to 2020-06-01, with 432236 profiles
- The following per profile metrics are fed to the ML algorithms

Predictors/Features	Description
reldevmax	Maximum relative difference between bending angle and climate model between 25 and 40 km (microrad)
difmaxref	Maximum relative difference between refractivity and climate model between 10 and 60 km (microrad)
qcl2d	Maximum absolute difference of L1 and L2 excess phase finite differences between 20 and 40 km (m/sample)
snr1avg	Avg SNR between 60 and 80 Km on the L1 frequency (V/V)
snr2avg	Avg SNR between 60 and 80 Km on the L2 frequency (V/V)
stdv	Standard deviation of the difference between bending angle and climate model between 60 and 80 km (microrad)
localhour	Local hour based on latitude and longitude (hrs)
smean	Mean difference between bending angle and climatology model between 60 and 80 km (microrad)
minalt	Minimum altitude of the profile (km)
conid	Transmitter constellation, GPS (G) or GLONASS (R)
occsat	Transmitter PRN number
latoccpt	Latitude of the profile (deg)
lonoccpt	Longitude of the profile (deg)
azimuth	Azimuth with respect to spacecraft velocity (deg)

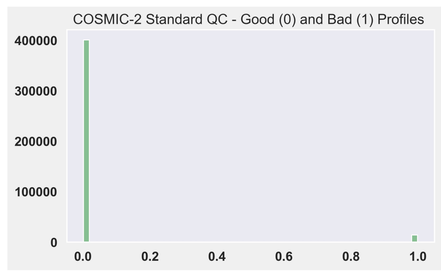
Predictand/Target	Description
bad	Standard-QC – bad = 1: Profile rejected by quality control

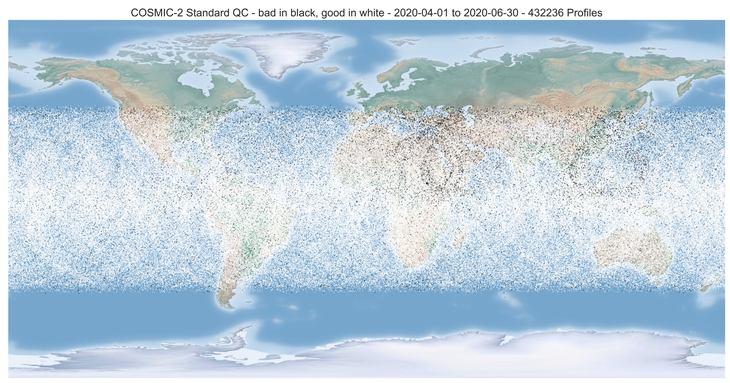
Correlation Between Variables



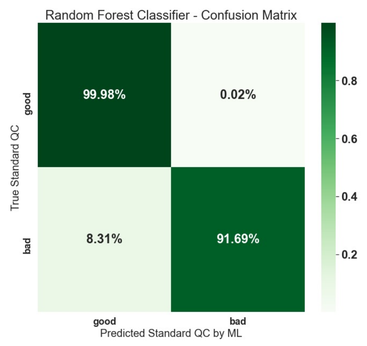
Results

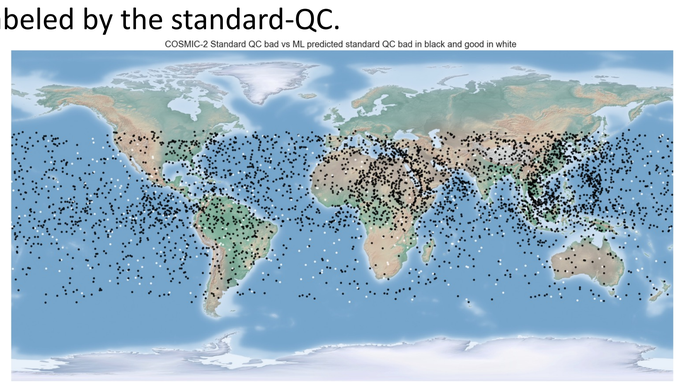
- Profiles marked with bad standard-QC account for 3% of data, making this an imbalanced ML classification problem
- ML algorithms were trained with 77% of the data and tested with 33%



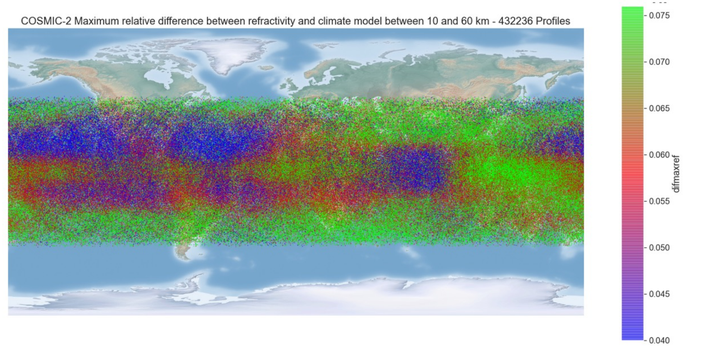


- Random Forest classifier, a machine learning algorithm consisting of decision trees, showed best results in the ROC AUC at 97% as well as the recall at 94%, disagreeing with 8.31% of the bad observations labeled by the standard-QC.





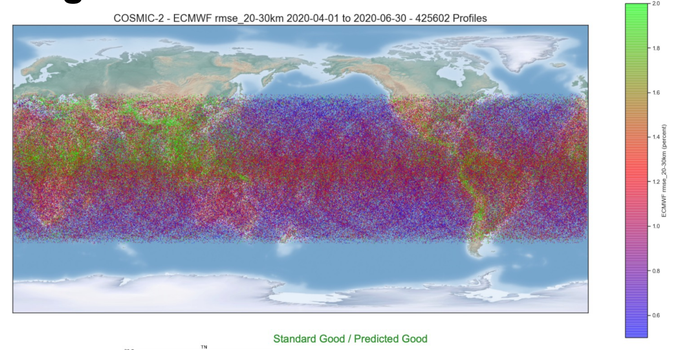
- Model interpretation shows maximum relative difference between bending angle and climatology (reldevmax), maximum relative difference between refractivity and climatology (difmaxref), and maximum absolute difference of L1 and L2 excess phase finite differences as the variables having highest impact on the model.

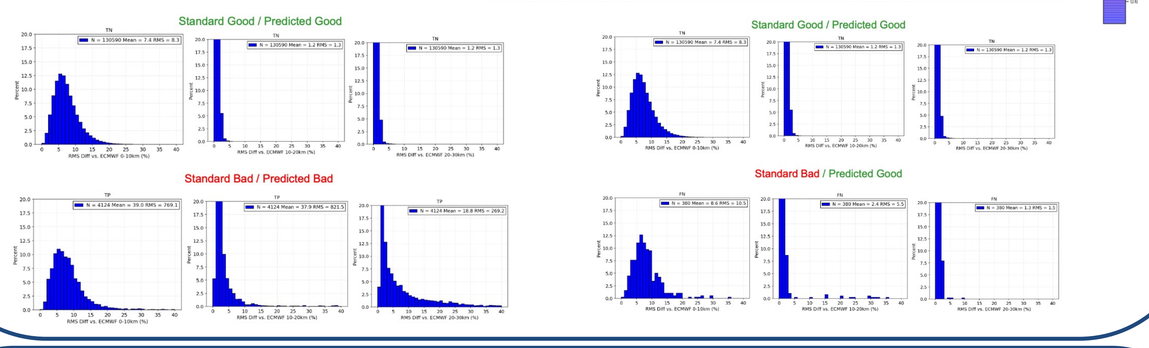


Variable	Importance
difmaxref	0.385514
reldevmax	0.301106
qcl2d	0.084942
stdv	0.071802
snr2avg	0.068505
smean	0.033942
latoccpt	0.009835
snr1avg	0.008940
azimuth	0.008719
minalt	0.008368

Comparing to ECMWF

- Distribution of COSMIC-2 bending angle vs. ECMWF for profiles flagged bad by standard-QC, but predicted by ML as good, is more similar to profiles flagged as good by the standard-QC.





Summary and Future Work

- Random Forest machine learning algorithm was chosen due to best recall and ROC
- Difmaxref, reldevmax, and qcl2d variables are top metrics impacting the model in its prediction of whether a profile is good or bad
- From the test set (137,068 observations), 4655 were marked bad by standard-QC, and the model disagreed with 387 of them (8.31%)
- Distribution of COSMIC-2 bending angle vs. ECMWF for profiles flagged bad by standard-QC, but predicted by ML as good, is more similar to profiles flagged as good than bad by the standard-QC
- Further analysis is currently in progress for the profiles flagged bad by standard-QC but predicted as good by the ML algorithm