Investigating the Use of Machine Learning (ML) to Assess Tropospheric Doppler Radar Wind Profiler (TDRWP) Data Quality

Corey Walker (corey.d.walker@nasa.gov), Stephen Thompson (stephen.c.thompson@nasa.gov) Amentum Space Exploration Division (ASED)

INTRODUCTION

Manual Quality Control (MQC) of Tropospheric Doppler Radar Wind Profiler (TDRWP) data is essential for defining an accurate climatology for downstream aerospace vehicle assessments [1]. MQC traditionally takes around 30.5 hours per year of radar data. The Marshall Space Flight Center Natural Environments Branch (MSFC NE) used machine learning (ML) to test the feasibility of automating the MQC process, showing a potential to reduce labor by 300% [2]. However, analysis of the model showed some false positives. We compared a neural network to the model to validate it and develop a process for assessing comparable solutions in the future.

MANUAL QC PROCESS

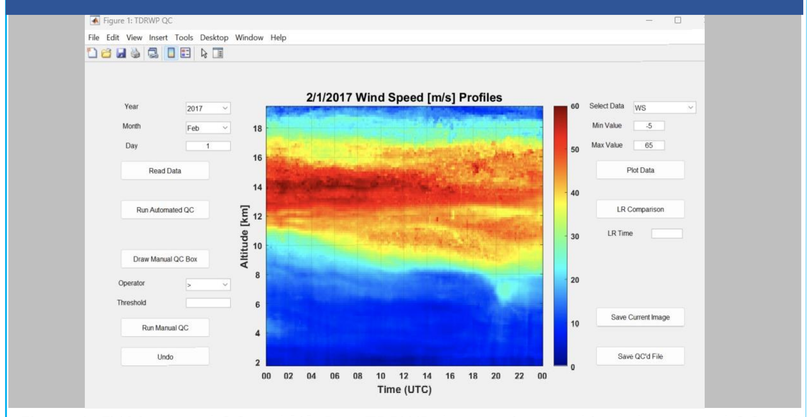


Figure 1: EV44 manual QC tool [1]: Raw TDRWP data are ingested into the manual QC tool and automated flags are generated. A manual QC check is done with an operator and manual QC flags are appended.

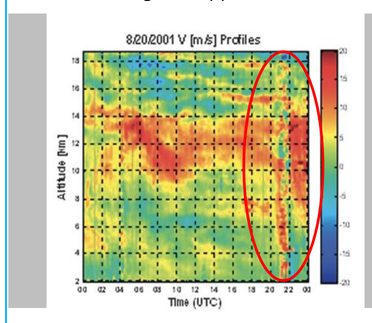


Figure 2: Convection Example [1]

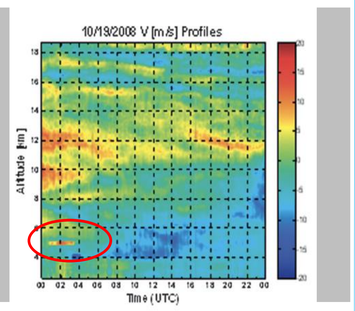


Figure 3: Non-Wind Object [1]

CURRENT ML PROCESS

- The MQC tool takes ~5 minutes per 24-hour image. Therefore, 365 images (for a full year) takes ~30.5 hours of labor.
- A random forest model developed by MSFC NE takes 1 second per 24-hour image, taking 365 seconds or 0.1 hour to QC a year, but has problems with false positives. [2]

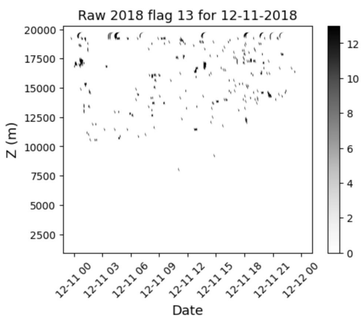


Figure 4: Human Labeled QC Data [2]

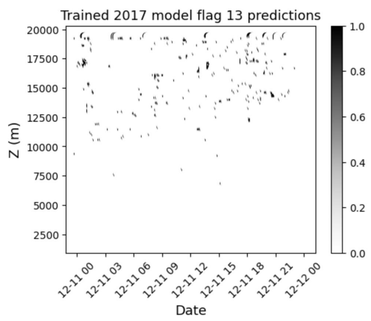


Figure 5: Modeled QC Data [2]

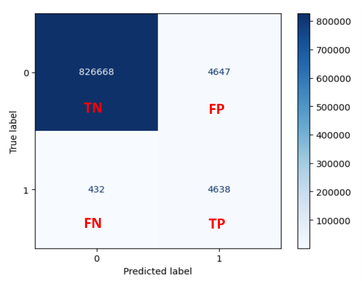


Figure 6: Confusion matrix of the 2017 Random Forest model showing number of True / False Positives (TP / FP) and True / False Negatives (TN / FN) [2]

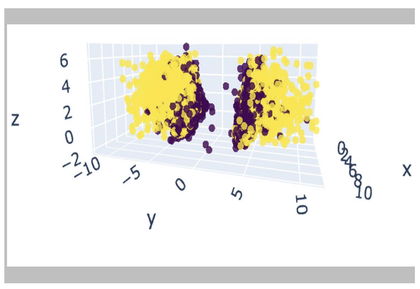


Figure 7: FP (purple) vs TP (yellow). X=wind shear (shr). Y=vertical motion (w), Z= beam-averaged spectral width (sw3). [2]

DATA

MSFC NE trained a random forest model using methods described in [2] and compared it to a trained Long Short Term Memory Network (LSTM) on the MQC task. Both models were trained on 324 radar days in 2018 containing 9,411,697 good (0) vs manual (13) flag examples containing shear (shr), vertical motion (w), and mean vertical spectral width (sw3) components. These models were assessed on the rest of the 2019-2022 data. Automated Quality Control Flags (AQC) were removed from the training dataset.

Flag (Automated)	Reason
0	Good data
0.2	Vertical beam spectral with check failed
0.3	Vertical beam 0-Doppler shift check failed
0.4	Vertical beam noise return failed
0.5	Vutomated convection flag tripped (data not removed)
1	Missing profile
2	Invalid wind speed or direction
3	Exceesive spectral width
4	Excessive internally-computed shear
5	Excessive vertical wind speed
6	FG prop check failed
7	Excessive noise level (Inconsistent data)
8	Excessive horizontal wind shear check (computed from output winds)
9	Small-median check failed
11	Invalid signal return from oblique beam
12	Isolated datum, dodgerblue
Flag (Manual)	**Reason**
10	Data removed manually attributed to ground clutter
13	Data removed manually
14	Data previously flagged for convection (flag == 0.5) removed manually

Table 1: Output quality control flags by automated and manual reason. This project only considers the flags highlighted in yellow.

Year	Total Gates	Total Days	Percent 0.0	Percent 13.0
2018	9411697	324	98.96	1.04
2019	9539479	314	98.66	1.34
2020	9720770	310	98.87	1.13
2021	10373417	340	98.97	1.03

Table 2: Summarized data for each year with percentage by flag

METHODS

1. Trained an LSTM model with the following architecture and information flow:

Input Layer → **Masking Layer** → **LSTM Layer** → **Dense Layer** → **Apply Round** → **Output Flags**

- **Input layer** – Input timestamps of radar gates containing the shr, w and sw3 values. The input size at 1 timestamp is 119 x 3. The batch size is 16. This means the model will generate 16 timestamps of MQC before updating model weights. Larger batch sizes cause faster training.
- **Masking layer** – This is used to preserve input shape without the model being influenced by AQC. This masks all rows containing AQC (shr, w, sw3, and the associated flag) with -999.99.
- **LSTM layer –** This is a collection of memory cells containing long and short term information that can be remembered, or forgotten depending on backpropagation. The LSTM returns the 16 hidden state vectors of size 119 x 16.
- **Dense layer –** This neuron outputs 16 vectors that are size 119 x 1 containing QC probabilities between 0-1. Anything above .5 is likely flag 13.
- **Round Function –** Applied to generate a 119 x 1 binary of 1 vs 0 for each of the 16 output vectors.

2. Retrain the 2017 Random Forest model on 2018 data.

3. Generate FP plots over time for each model.

4. Compared visual results of the Random Forest model vs the Neural Network on sample dates containing non-wind objects.

5. Generate a confusion matrix for all test years for the trained Neural Network and Random Forest (2019, 2020, 2021).

VISUAL RESULTS

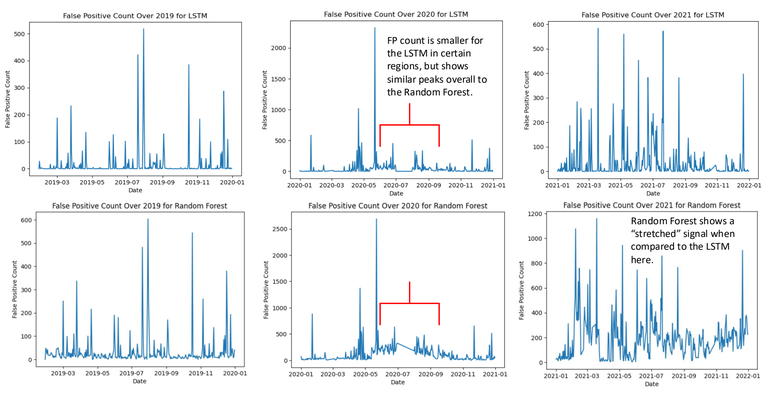


Figure 8: LSTM (top row) vs Random Forest false positive (bottom) count for all years

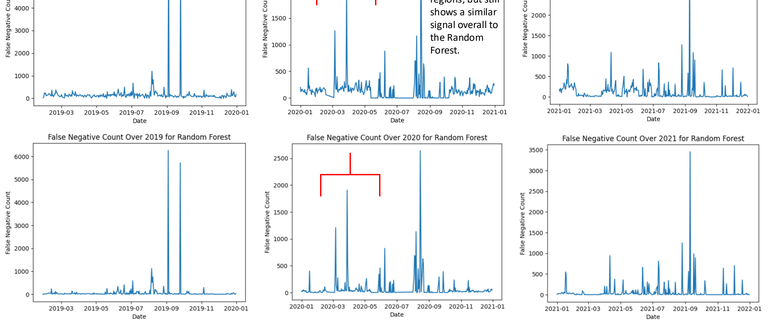


Figure 9: LSTM (top row) vs Random Forest (bottom) false negative count for all years.

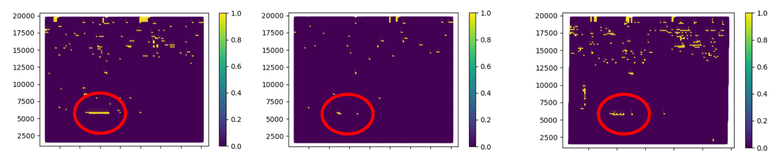


Figure 10: Human Labeled QC data (yellow) on September of 2019 (left), vs Random Forest (middle), vs LSTM (right) on non wind object (circle).

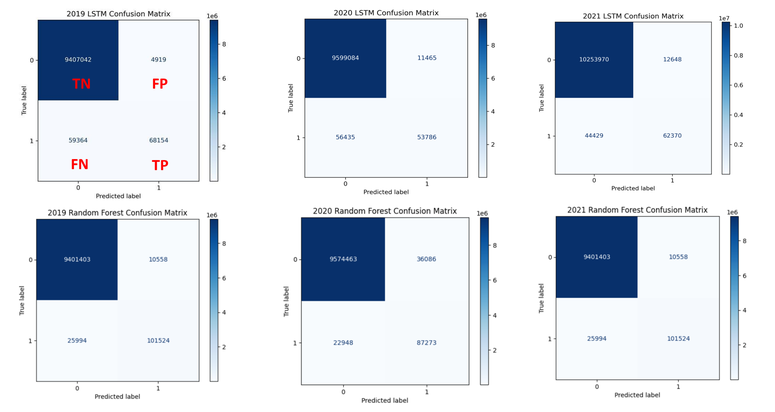


Figure 11: Confusion matrices for LSTM (top row) vs Random Forest (bottom) for test years

CONCLUSION

- Similar FP vs FN counts across observations suggest particular reasons for failure conditions independent of model type. Hypothesis is that FNs may be related to entire profiles being removed when good data may still be there. FPs may be locations where the user missed QC. Quantifying similar failure conditions may help MSFC NE clean the data for future model training.
- LSTM shows better capture of the non-wind object compared to Random Forest. This suggests a need to experiment with different models to understand what may better generalize this feature.
- Confusion Matrix results suggest that the LSTM reduces model FP rate, but increases the FN rate. This may suggest that model tuning is needed, or a different architecture all together. This study has lead MSFC to continue using this ML process to improve the MQC of TDRWP data.

REFERENCES

[1] Barbré , R. E., 2012: Quality Control Algorithms for the Kennedy Space Center 50-MHz Doppler Radar Wind Profiler Winds Database. *J. Atmos. Oceanic Technol.*, 29, 1731–1743, https://doi.org/10.1175/JTECH-D-11-00205.1.

[2] Walker, C., 2024: Automating Manual Quality Control (QC) of Radar Data with Machine Learning (ML). Poster Presented at Marshall Space Flight Center Summer of Artificial Intelligence.

***** This work was done under NASA Contract 80MSFC18C0011