ANOMALY DETECTION

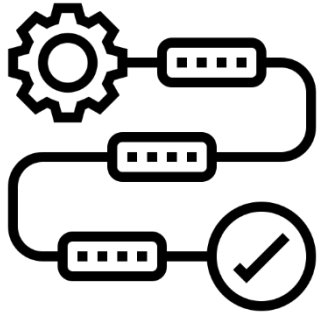
Quality Control using Machine Learning

Gurkanwal Arora, Data Scientist
Office of the Chief Data Officer (OCDO)

Thinesh Sornalingam, Manager
Meteorological Service of Canada (MSC) – Data Management and Operations Section (DMOS)

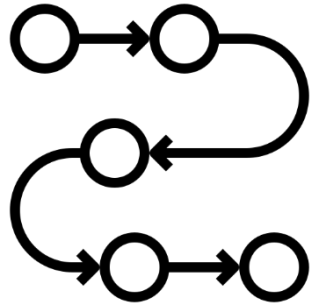
PRESENTATION OUTLINE

Problem & Objective



Methodology

Results



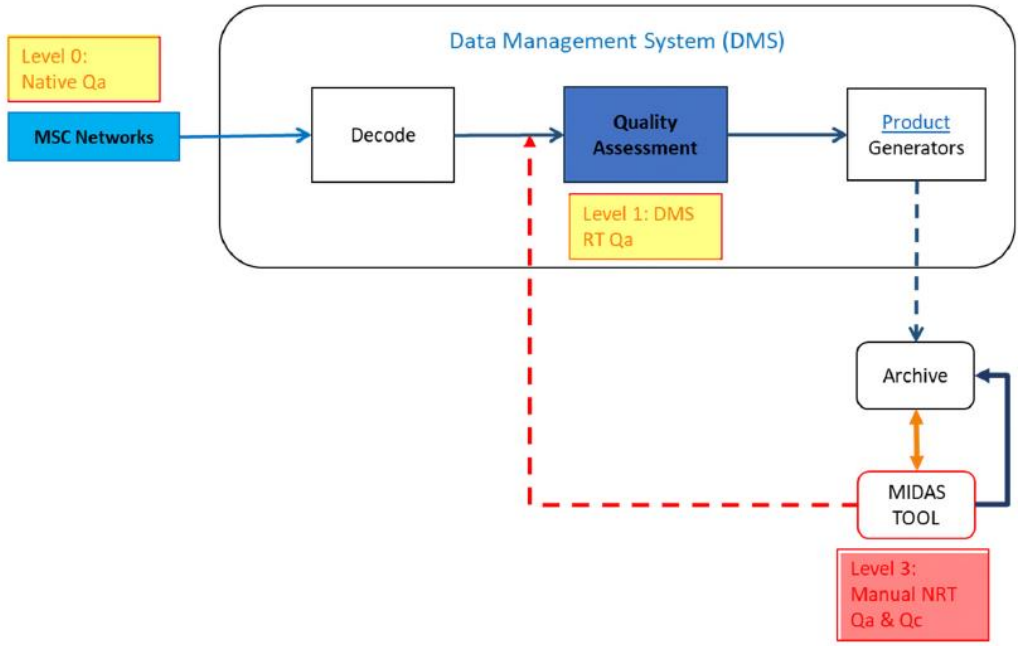
Next Steps

MSC'S QC PROCESS

Table 1 An overview of the structure of the DMS QC System.

QC LEVEL	STATUS	MANAGED BY	PROCESSING METHOD	TIMING
Level 0 (i.e. Native)	Operational	DQU (Support from Data Providers)	Automated	Real-Time
Level 1	Operational Further Enhancements in Planning	DQU	Automated	Real-Time
Level 2	Not Started	DQU	Automated	Non-Real-Time
Level 3	Operational	AOU (Support from DQU)	Manual	Non-Real-Time
Level 4	In Development	AOU (Support from DQU)	Auto/Manual	Non-Real-Time

[Data Quality Unit (DQU) - Wiki (ec.gc.ca)]



PROBLEM & OBJECTIVE

Observations are reviewed manually by QC technicians to improve quality of climate data

- High visibility stations and parameters
- Stations with known historical issues
- Product bound data
- Anomalies reported by users

Thousands of hourly observations, not enough human resources to review all

MSC is looking for a more robust, automated, and sustainable approach to detecting outliers in large data volumes

Objective: Assess the potential of using ML / Anomaly Detection to support the QC process

* The ML model will aid level 3 in the QC process

ETL PROCESS

Arkeon DB

xml-extractor

- 12 copies, 1 per month
- Executed in parallel

xml-to-dict

- Converts xml files into dictionaries
- Chunked by stations

dict-to-tabular

- Converts dictionaries into a single DataFrame

data-clean-up

- Clean up data (i.e., drop columns)
- Prepare target variable

Data preprocessing

- Combines all years
- Further cleaning

Final dataset

- 1 for each element
- Includes all stations and all years
- Ready for machine learning

DATA LABELING

- Rule-based data labeling implemented into the data processing
- Labeling is based on previous manual corrections

Rules for data labeling:

1. Either qa flag value was ‘100’ and changed to a non-null value OR was unequal to ‘100’ and changed to ‘100’ (target = 1)
2. Post mc qa flag value is ‘100’ because of a value override (target = 0)
3. Both flag values are ‘100’ but this is caused by rerun of test after value_override (target = 1)

Table 3 The Data Quality Flags

Flag Value	Flag Name (English/French)	Definition
-10	Suppressed / Réprimer	The data provider has indicated that the value is not to be used or published.
-1	Missing / Manquant	There is no value available.
0	Error / Erreur	The value is erroneous.
10	Doubtful / Douteux	The value may be acceptable but is significantly uncertain.
20	Inconsistent / Incohérent	The value departs significantly from an expected physical relationship with an independently measured, associated variable.
100	Accepted/Passed / Accepté	The value passed all applicable quality assessment test(s) or has been verified as acceptable.

Data Quality Unit (DQU) - Wiki (ec.gc.ca)

DATA PREPROCESSING

Missing Values

- Dropped columns with more than 1% missing instances
- Sklearn’s SimpleImputer (Mean) for remaining missing values

Categorical Features

- Dummy encoding / OHE

MODELING

element	df length	value count			# columns	training data (80%)	test data (20%)
		class 0	class 1	%			
snow_depth_3022	1,718,100	1,705,004	13,096	0.8%	197	1,374,480	343,620
snow_depth_3025	1,702,049	1,688,634	13,415	0.8%	238	1,361,639	340,410
precipitation_amount_12	223,447	199,625	23,822	10.7%	161	178,757	44,690
wind_speed_3003	90,085	82,155	7,930	8.8%	337	72,068	18,017

Classification Algorithms Tested:

- Random Forest (RFC)
- Gradient Boosting (GBC)
- Ada Boost (ABC)
- Extreme Gradient Boosting (XGBC)
- Light Gradient Boosting (LGBC)
- Voting Classifier (VTC)
- Deep Learning (DL)

DEALING WITH THE IMBALANCE

- Class weights
- Over Sampling
  - SMOTE
  - ADASYN
  - Random
- Under Sampling
  - Random

HYPER-PARAMETERS

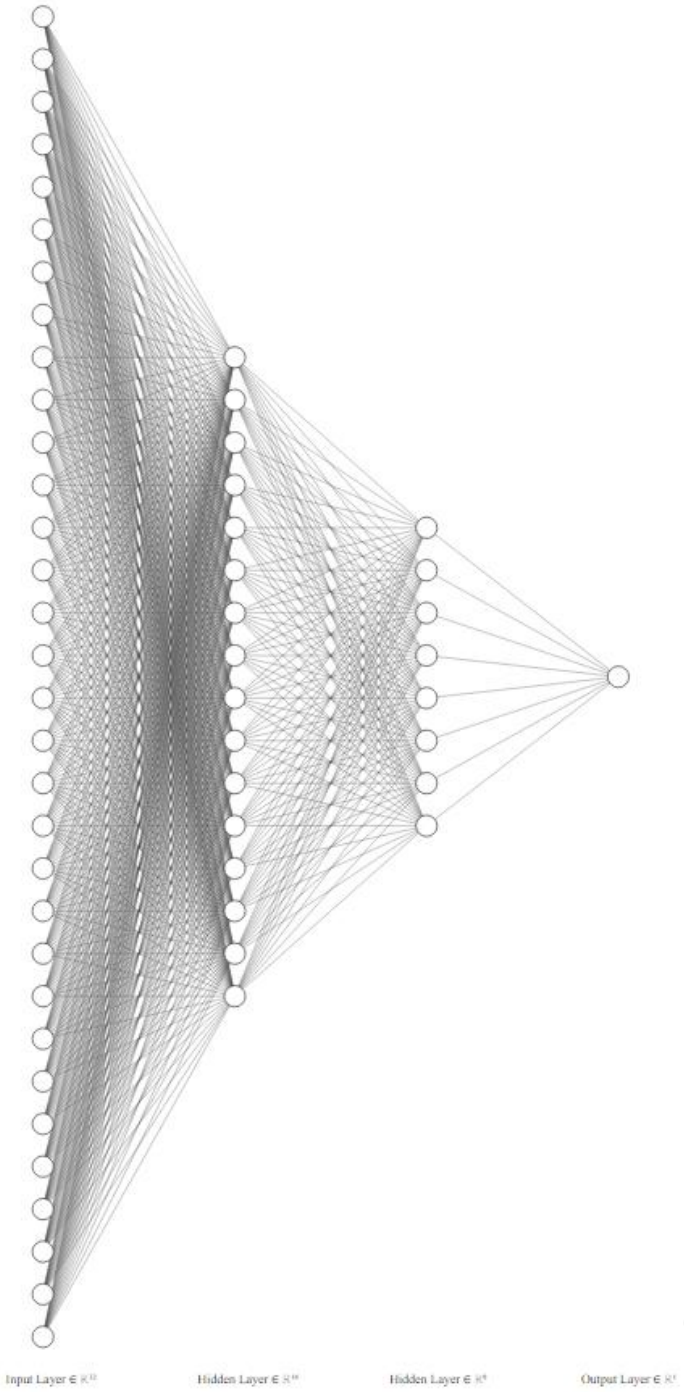
	n_estimators	max_depth	class_weight	learning_rate	num_leaves
RFC	500	None	None	n/a	n/a
GBC	1000	3	n/a	0.1	n/a
ABC	1000	n/a	n/a	0.1	n/a
XGBC	1500	None	None	0.1	n/a
LGBC	500	-1	None	n/a	31

DEEP LEARNING

Model: "sequential"

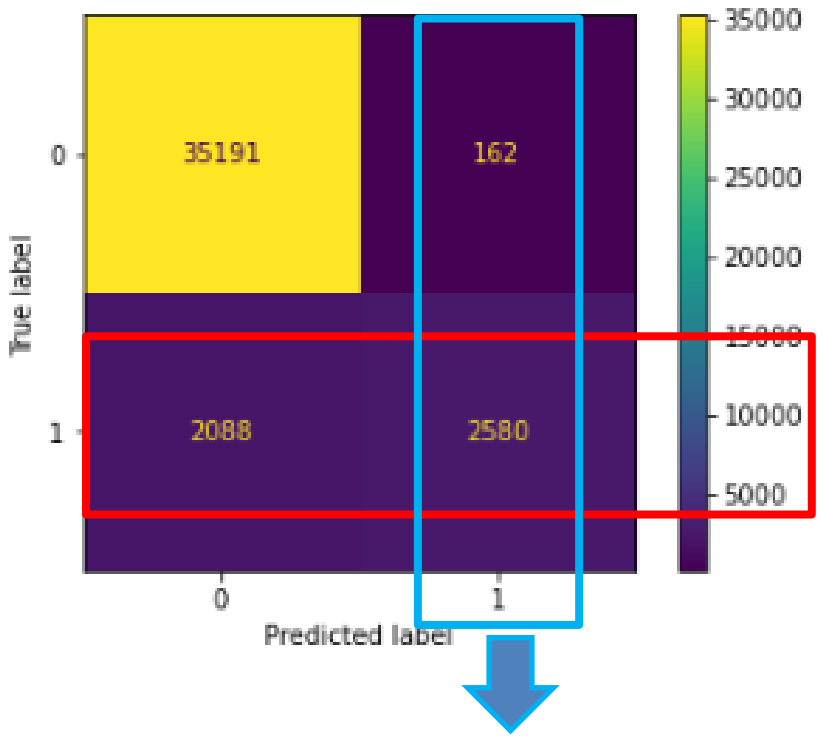
Layer (type)	Output Shape	Param #	activation	Dropout
dense (Dense)	(None, 32)	5152	‘relu’	
dropout (Dropout)	(None, 32)	0		0.2
dense_1 (Dense)	(None, 16)	528	‘relu’	
dropout_1 (Dropout)	(None, 16)	0		0.2
dense_2 (Dense)	(None, 8)	136	‘relu’	
dropout_2 (Dropout)	(None, 8)	0		0.2
dense_3 (Dense)	(None, 1)	9	‘sigmoid’	

Total params: 5,825
Trainable params: 5,825
Non-trainable params: 0



RESULTS

precipitation_amount_12 - RFC



Class	0	1	Macro-Avg
Recall	0.995	0.553	0.774
Precision	0.944	0.941	0.942
F1-Score	0.969	0.696	0.833

$$recall = \frac{2580}{2580 + 2088} = 0.553$$

$$precision = \frac{2580}{2580 + 162} = 0.941$$

Recall : % of actual anomalies the model correctly predicted
Precision : % of all predicted anomalies that are correct

RESULTS

All stations, 2018 – 2021 data
Macro-avg scores

Element	Performance Metric	Baseline	Classical ML				Deep Learning
		DummyClassifier*	RFC	XGBC	LGBC	Voting	DL
snow_depth_3022	Recall	0.499	0.917	0.917	0.918	0.918	0.850
	Precision	0.499	0.995	0.995	0.931	0.931	0.952
	F1-Score	0.499	0.953	0.953	0.925	0.925	0.895
snow_depth_3025	Recall	0.499	0.913	0.951	0.910	0.931	0.817
	Precision	0.499	0.995	0.992	0.947	0.993	0.961
	F1-Score	0.499	0.950	0.970	0.928	0.960	0.876
precipitation_amount_12	Recall	0.502	0.946	0.948	0.951	0.946	0.931
	Precision	0.501	0.959	0.955	0.955	0.959	0.943
	F1-Score	0.501	0.952	0.951	0.953	0.952	0.937
wind_speed_3003	Recall	0.500	0.960	0.973	0.976	0.969	0.909
	Precision	0.500	0.966	0.976	0.979	0.974	0.852
	F1-Score	0.500	0.963	0.975	0.978	0.972	0.878

* DummyClassifier randomly predicts based on distribution

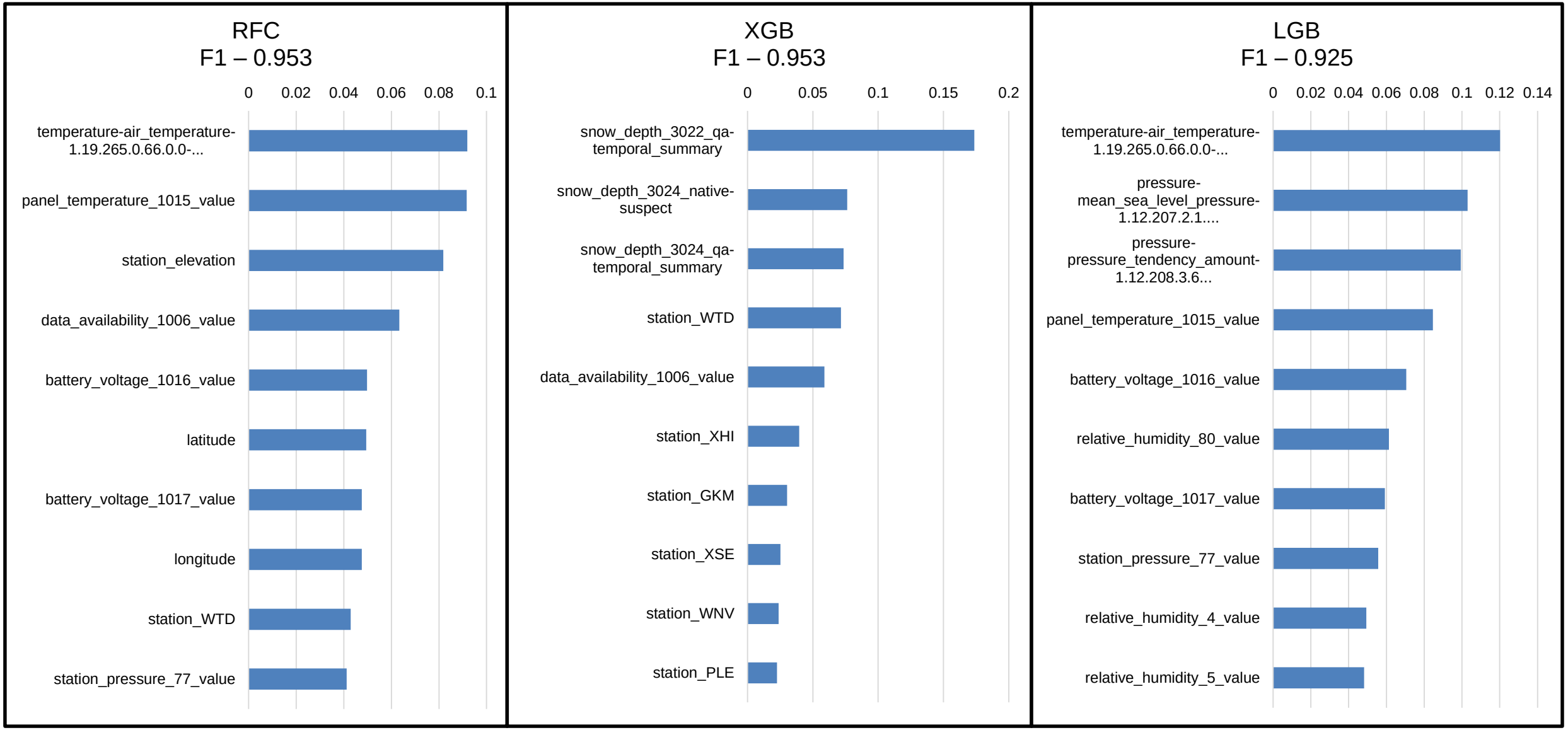
RESULTS

All stations, 2018 – 2021 data
Class 1 scores

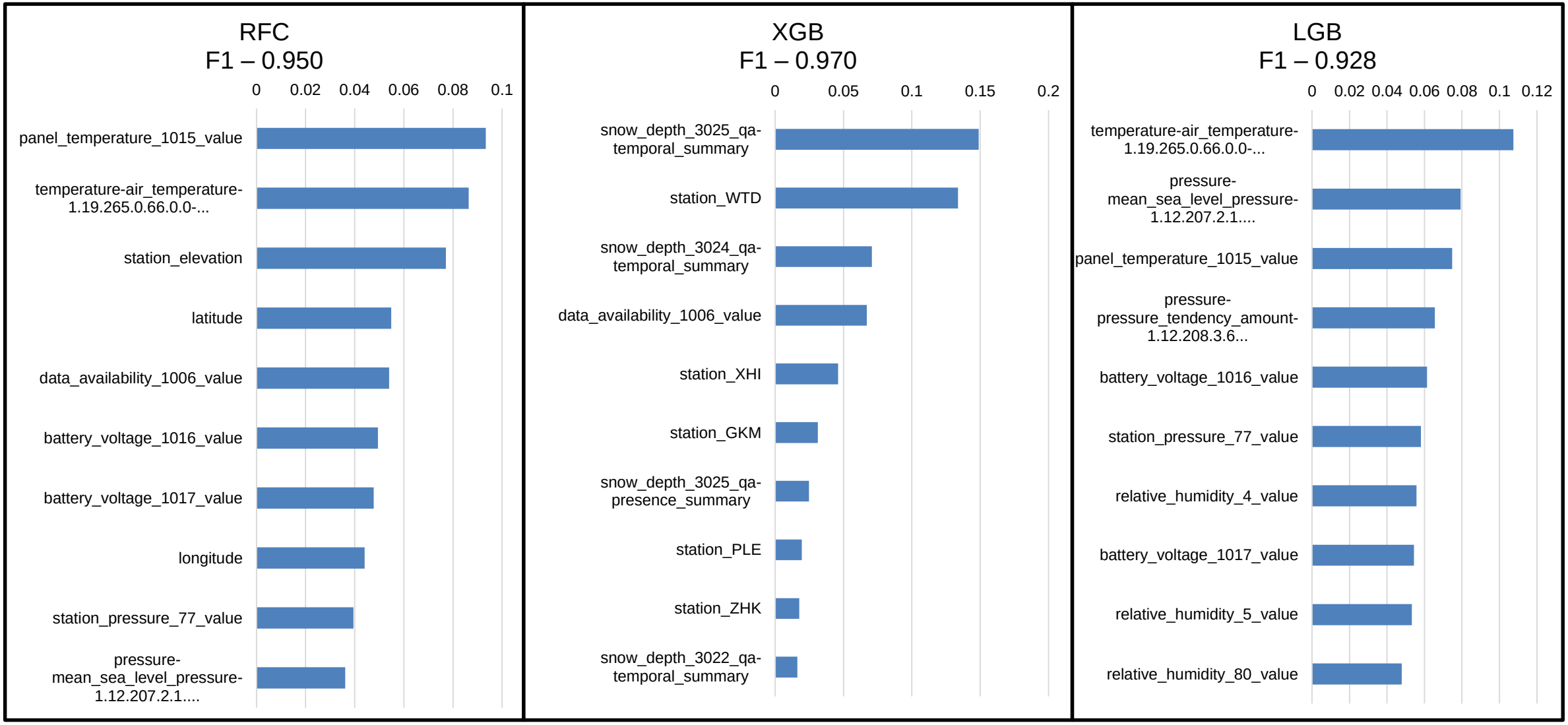
Element	Performance Metric	Baseline	Classical ML				Deep Learning
		DummyClassifier*	RFC	XGBC	LGBC	Voting	DL
snow_depth_3022	Recall	0.01	0.835	0.835	0.838	0.838	0.701
	Precision	0.01	0.992	0.992	0.864	0.864	0.906
	F1-Score	0.01	0.907	0.907	0.851	0.851	0.791
snow_depth_3025	Recall	0.01	0.825	0.901	0.821	0.861	0.634
	Precision	0.01	0.991	0.984	0.896	0.988	0.926
	F1-Score	0.01	0.901	0.941	0.857	0.920	0.753
precipitation_amount_12	Recall	0.11	0.900	0.905	0.911	0.899	0.872
	Precision	0.11	0.930	0.920	0.921	0.929	0.901
	F1-Score	0.11	0.915	0.913	0.916	0.914	0.887
wind_speed_3003	Recall	0.09	0.925	0.950	0.956	0.943	0.851
	Precision	0.09	0.940	0.958	0.962	0.953	0.718
	F1-Score	0.09	0.933	0.954	0.959	0.948	0.779

* DummyClassifier randomly predicts based on distribution

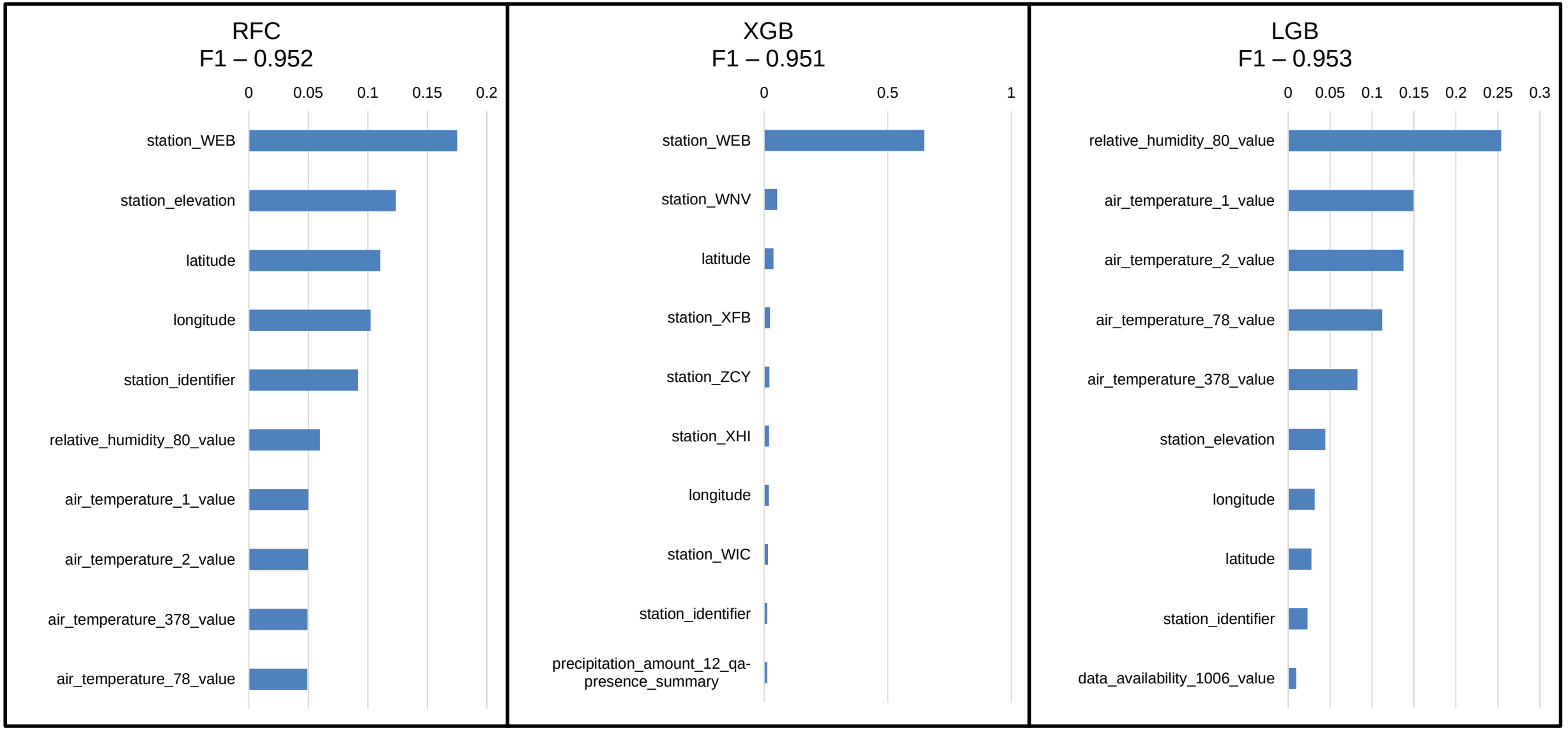
SNOW_DEPTH_3022 FEATURE IMPORTANCES



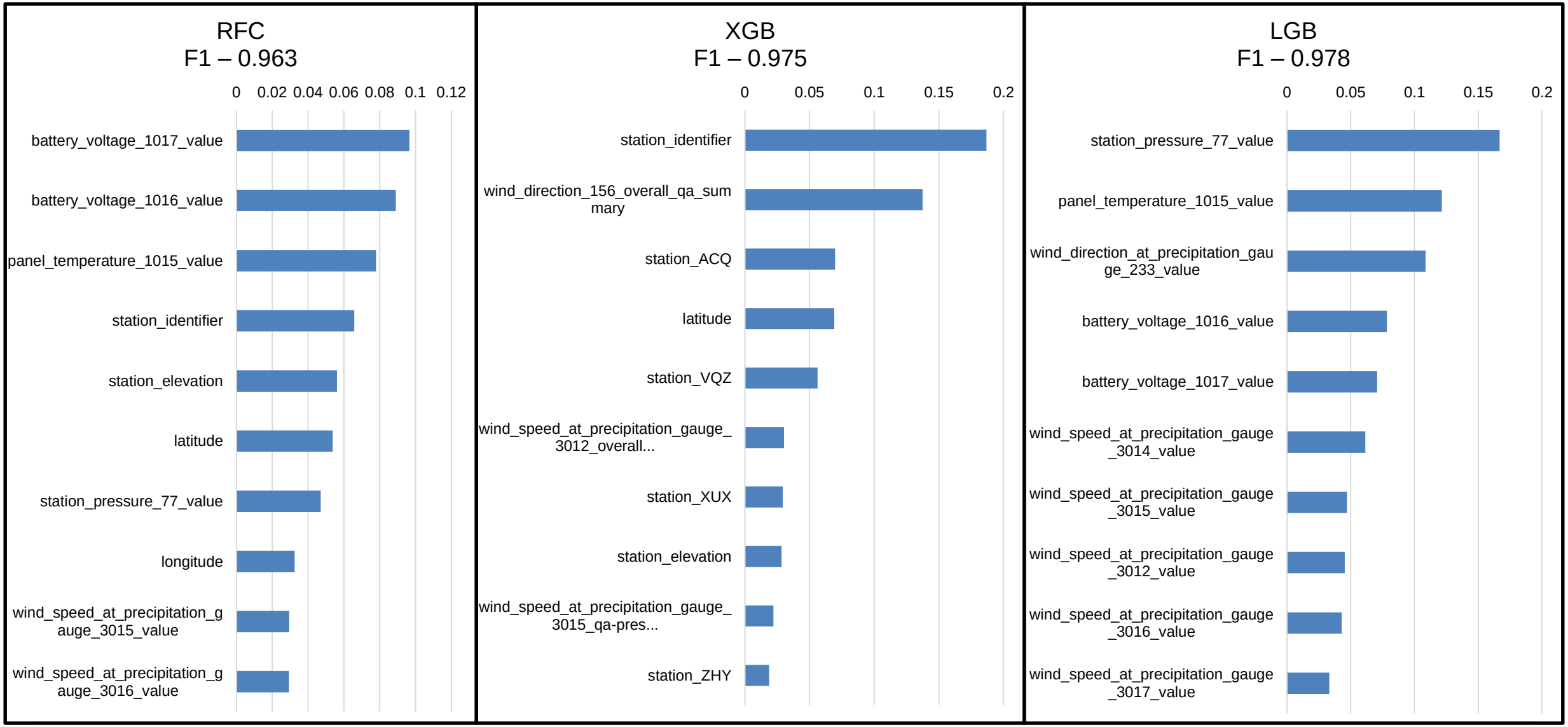
SNOW_DEPTH_3025 FEATURE IMPORTANCES



PRECIPITATION_AMOUNT_12 FEATURE IMPORTANCES



WIND_SPEED_3003 FEATURE IMPORTANCES



NEXT STEPS

- Set up a pilot environment to test the models in production
- Set up real-time detection and alert system (email alerts)
- Test runs with the end-users (QC technicians)