Machine Learning-based Global Lightning Prediction from Convective Parameters

Dominique Brunet (Environment Canada)
Mateusz Taszarek (AMU, Poland)
Jerry Su (U of Waterloo, Canada)
John Hanesiak (U of Manitoba, Canada)

Motivation

Prediction of lightning

- In the past: Global lightning climatology (before the 90s)
- In the coming days and weeks: Lightning forecast from NWP or AIWP storm environment
- In the future: Lightning projection in different climate scenarios

Problem

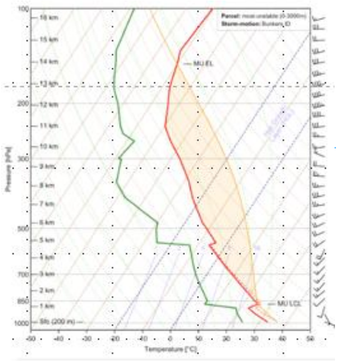
Lightning prediction over land in mid-latitudes is well established, but it is less so in the tropics or over the oceans.

Proposed solution

Interpretable AI (ensemble decision trees) applied on a massive dataset.

Training:

ERA5



thundeR
rawinsonde package

compute

convective parameters

	MIXR [g/kg]	CAPE [J/kg]	CAPE03 [J/kg]	CAPEHGL [J/kg]	CIN [J/kg]	LI [K]	LCL [m]	LFC [m]	EL [m]	WMAXSHEAR [m2/s2]
SB	19.1	5995	181	1522	-10	-14	1155	1480	13900	1644 (E 1914)
MU	19.1	5995	181	1522	-10	-14	1155	1480	13900	1644 (E 1914)
ML	17.7	5030	122	1356	-30	-13	1240	1760	13530	1506 (E 1721)

Bulk wind shear [m/s]	
Sfc - 1 km:	7.1
Sfc - 3 km:	11.8
Sfc - 6 km:	15.0
Sfc - 8 km:	19.0
Sfc - HGL:	13.7
Effec. (SB):	17.5
Effec. (MU):	17.5
Effec. (ML):	17.2

	SRH RM [m2/s2]	SRH LM [m2/s2]
Sfc - 100 m:	21	16
Sfc - 500 m:	35	12
Sfc - 1 km:	73	-15
Sfc - 3 km:	94	-63

Mean wind [m/s]	
Sfc - 1 km:	4.8
Sfc - 2 km:	4.9
1 - 3 km:	8.4
Sfc - 6 km:	8.0

Lapse rate [K/km]	
Sfc - 1 km:	9.1
Sfc - 3 km:	7.9
3 - 6 km:	7.6
500700 hPa:	7.4

Precip. water [mm]:	34
2 - 5 km RH [%]:	42
Sfc - 2 km RH [%]:	50

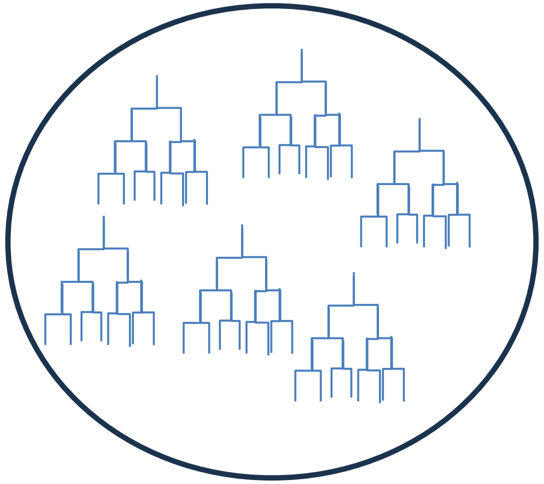
Moisture flux [g/s/m2]:	57
4 km DCAPE [J/kg]:	1359
4 km delta theta-e [K]:	40

SHIP:	2.4
SCP:	9.8
STP:	1.0

DMLC XGBoost

X: features

Ensemble Decision Trees (EDT)

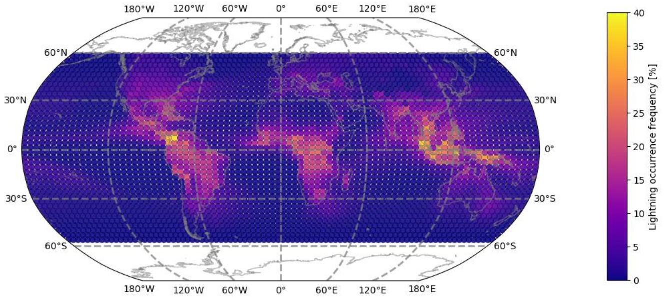


ENGLN

EARTH NETWORKS

detect

lightning occurrence

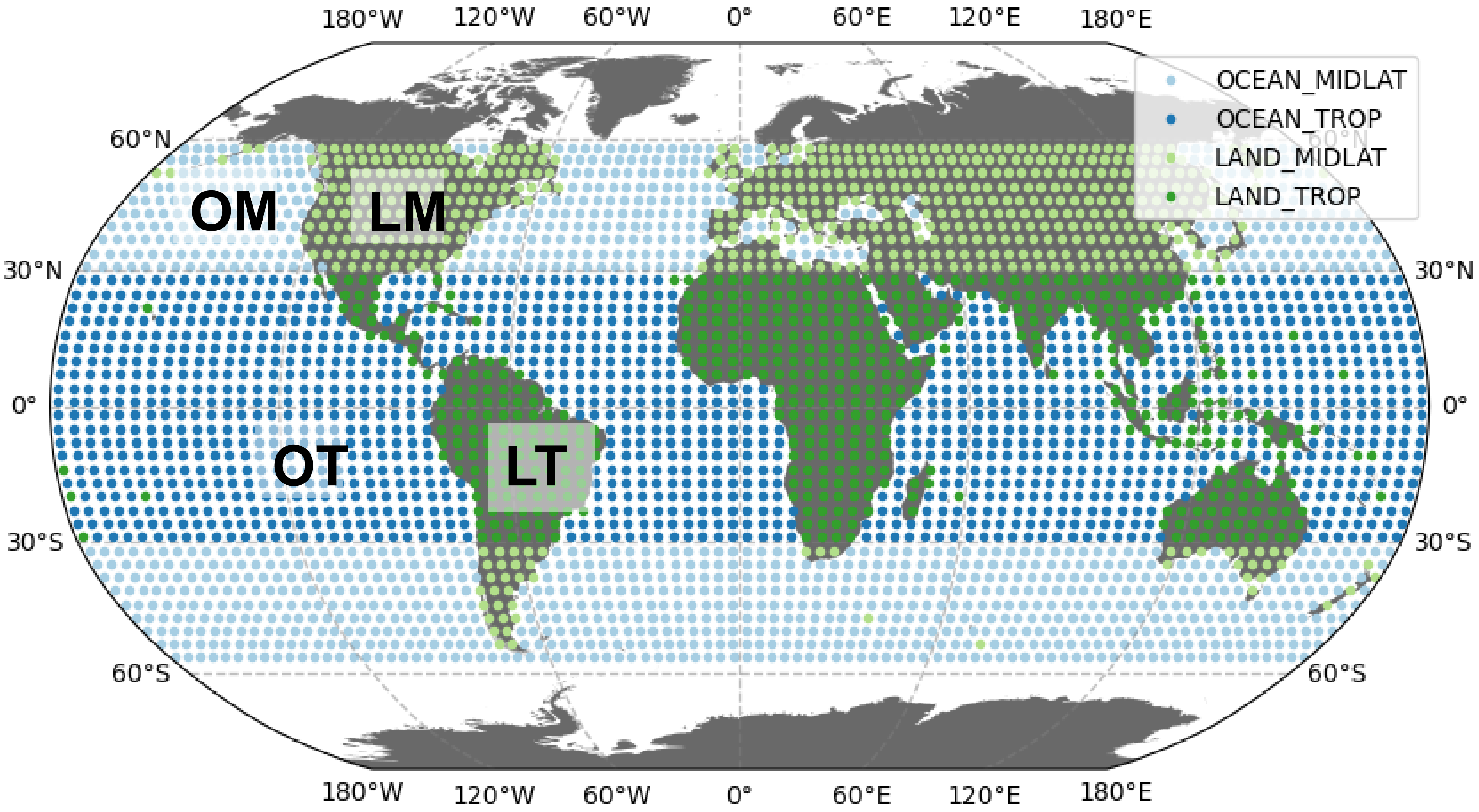


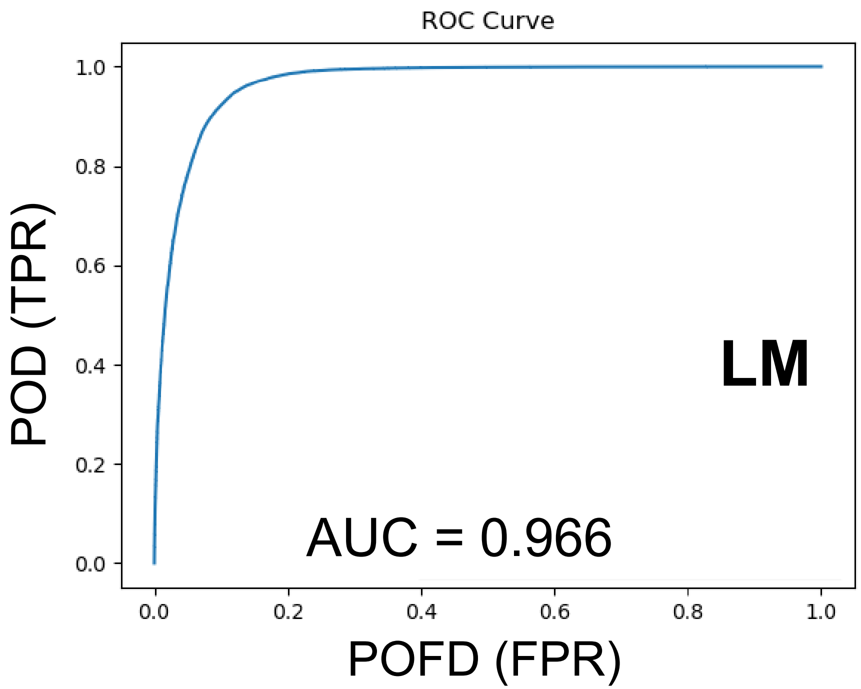
y: predicted classes

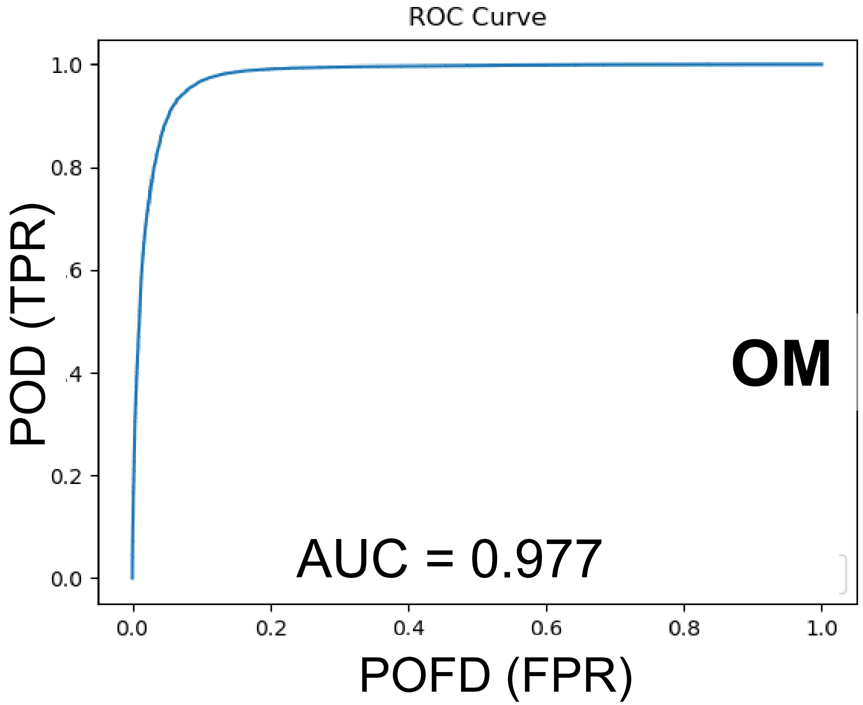
Inference: Apply EDT on ERA5-like data to predict lightning occurrence

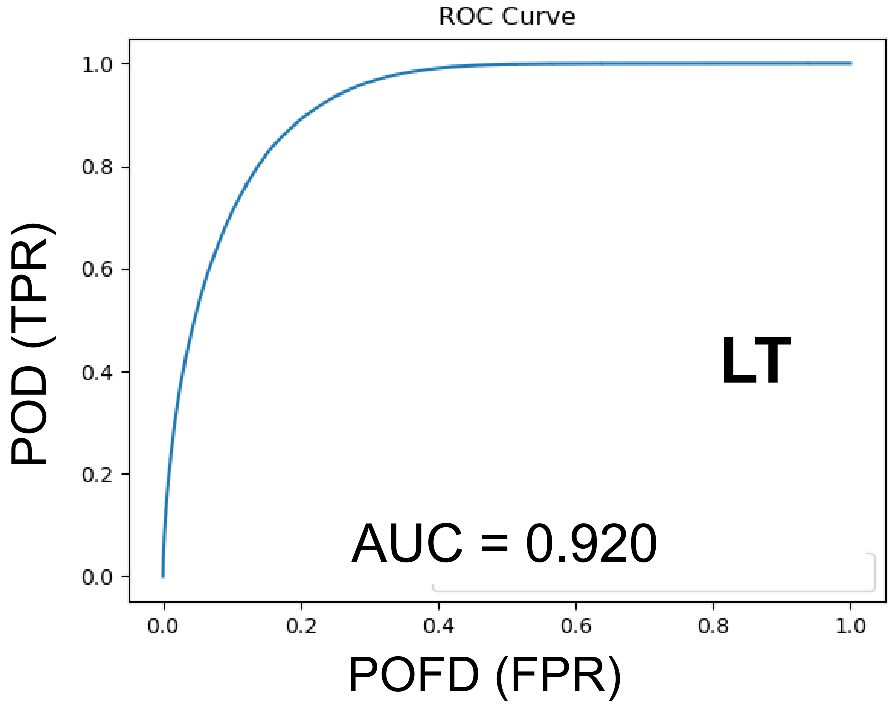
Experiment

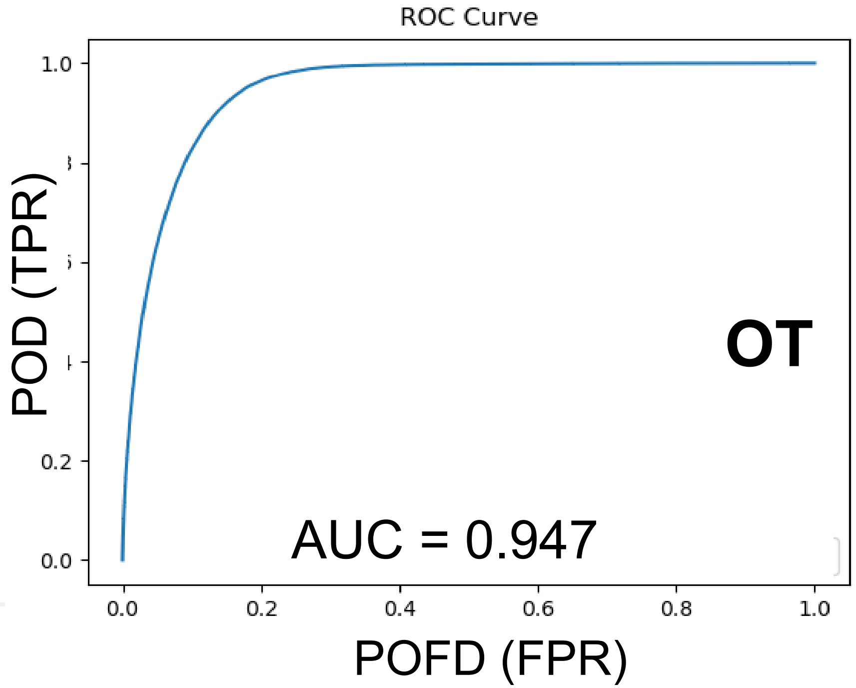
- 153 convective parameters tested
- 51,000,000 cases
- 3,500,000 cases after QC and data balancing
- Split:
  - Train: 2019-2021
  - Validation: 2022
  - Test: 2023







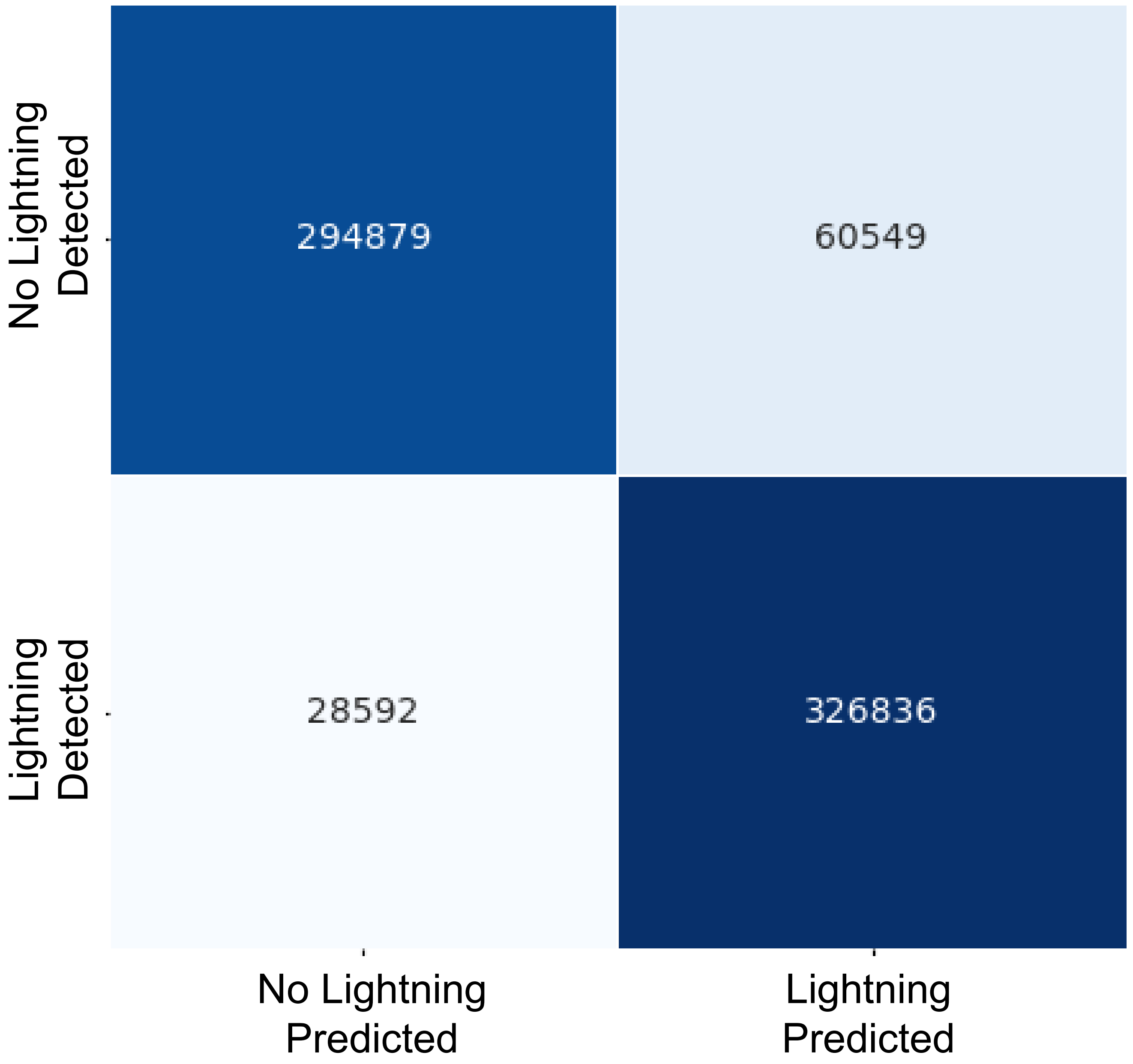




Results

Trained globally

Global ROC-AUC = 0.938



Test dataset (2023)

Global domain

Threshold: 50% (default)

POD = 92%

POFD = 17%

SR = 84%

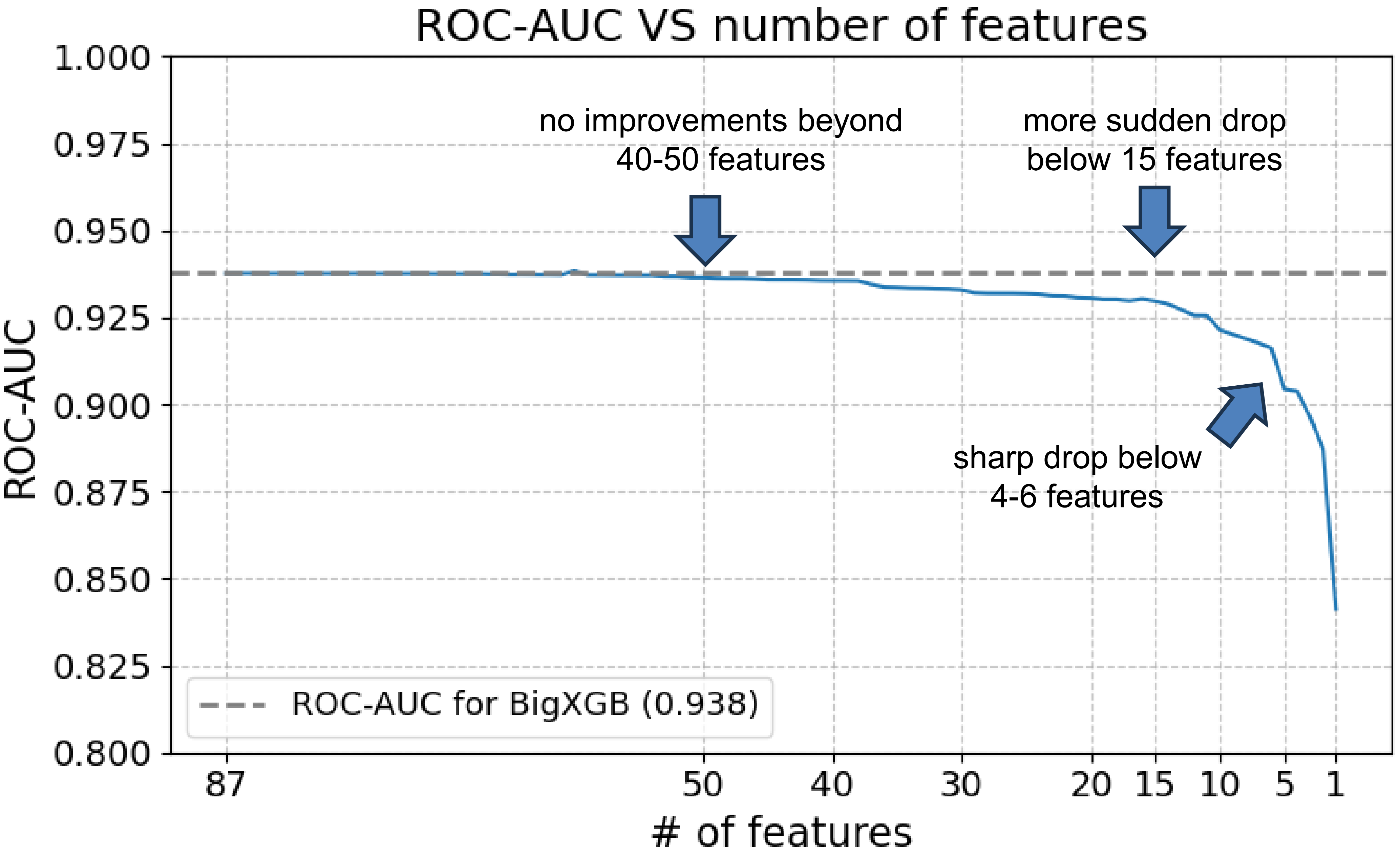
CSI = 79%

(balanced 50-50)

SR = 15%

CSI = 15%

(true climatology of lightning: 3.2%)



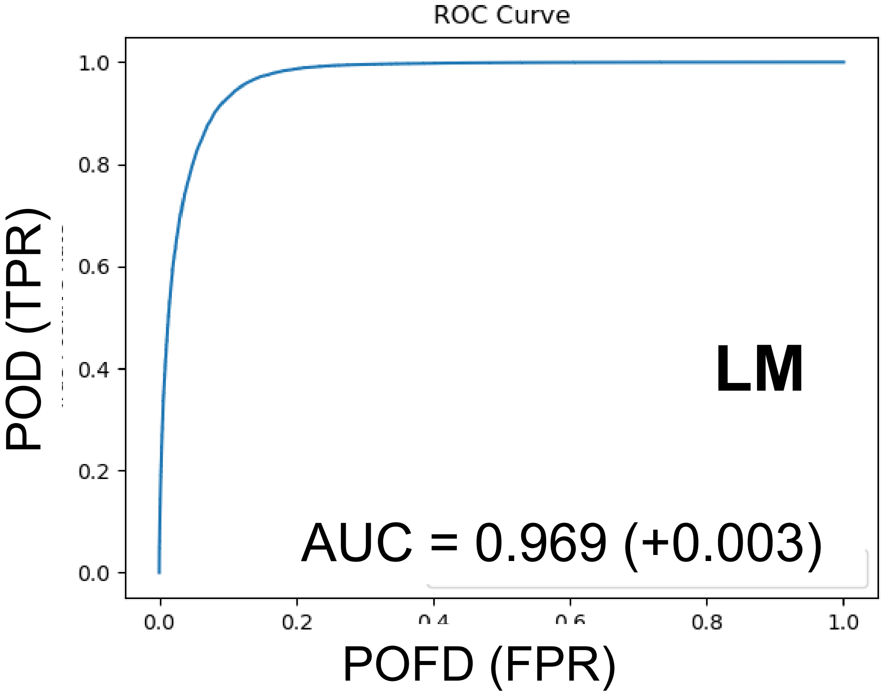
(convective parameters)

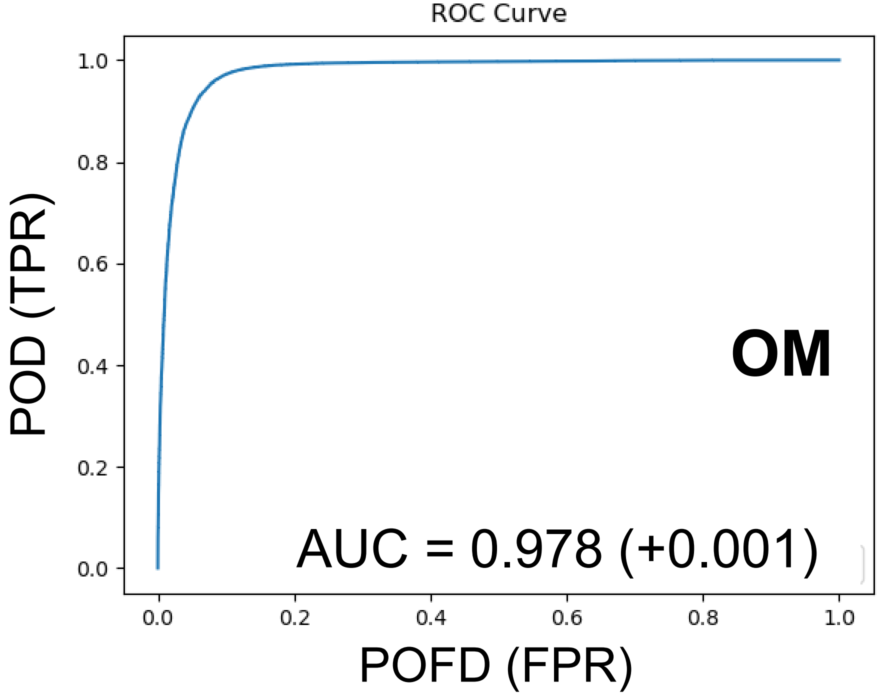
Experiment: Training by regions

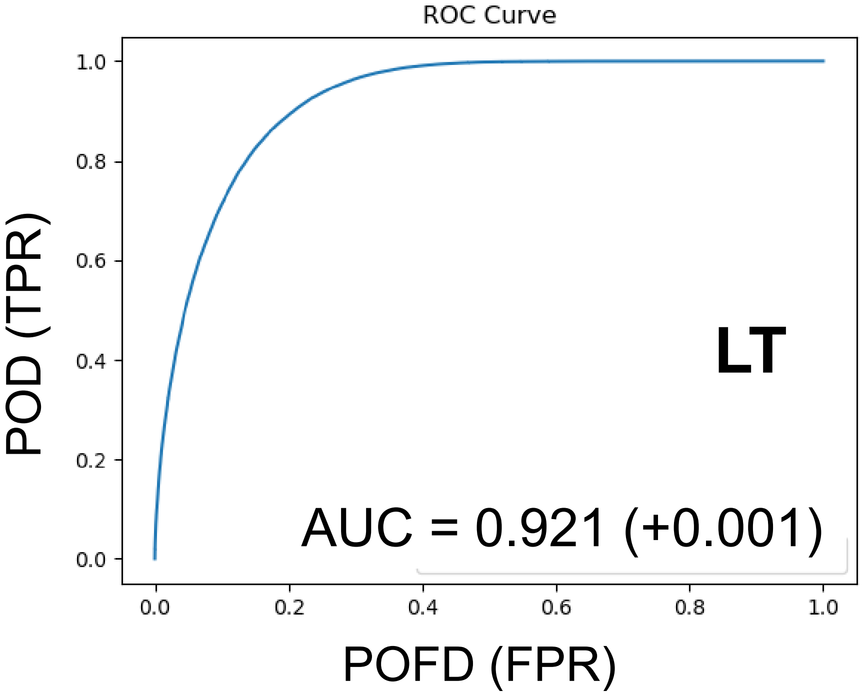
- Are the scores the same for universal VS specialized training?
- What are the best convective parameters for each region?

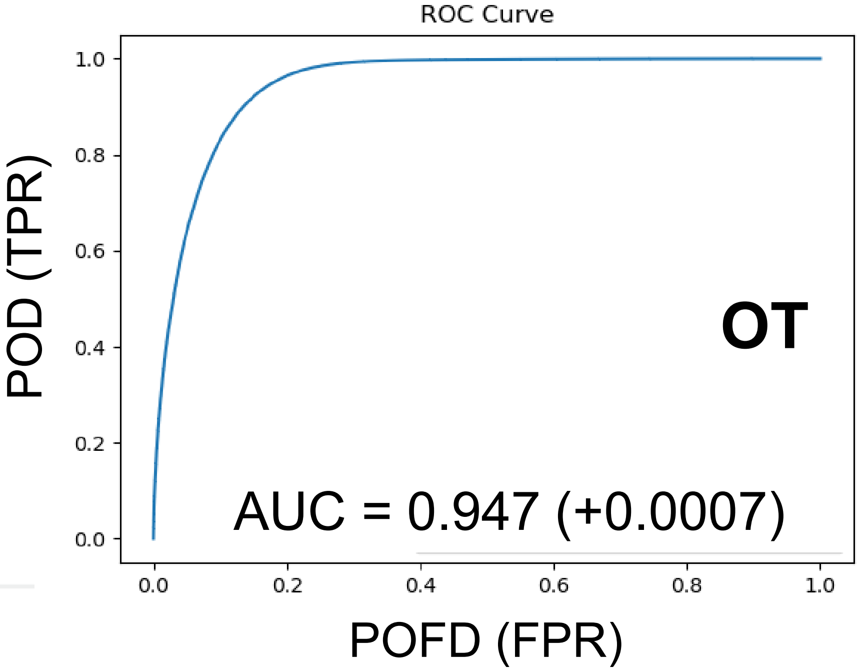
Results

Trained by regions









	Global	LM	OM	LT	OT
1.	**MU5_CIN_4km**	MUML_LI_eff	MUML_LI_eff	**MU5_CIN_4km**	RH_500850
2.	MU5_LI_eff	MU_LI_eff	MU5_LI_eff	MUML_LI	MU5_LI
3.	tciw	MUML_equal_layer	MUML_CAPE_HGL	MU5_LI	tciw
4.	tcslw	MUML_CAPE_HGL	MU_LI_eff	tciw	MUML_LI
5.	MUML_LI_eff	tciw	MU_CAPE_HGL	MUML_LI_eff	tcslw
6.	MU5_CAPE_HGL	MU5_LI_eff	**MU_CIN_4km**	tcslw	MU5_CAPE_HGL
7.	MUML_CAPE_HGL	**MU5_CIN_4km**	tciw	THETAE_02km	RH_14km
8.	MUML_LI				ML_CAPE_HGL
9.	LR_06km	tcslw	ML_LCL_hgt	ML_LI_eff	MU5_LI_eff
10.	MU_LI_eff	Moisture_Flux_SR	vigd	LR_06km	RH_ML_LCL_3km

MU5_CIN_4km: Convective inhibition for lifted parcel (MU5: most-unstable above 500 m AGL) in 4 km layer with the bottom of that layer at the parcel lifting height.

	Global	**LM**	**OM**	**LT**	**OT**
1.	MU5_CIN_4km	**MUML_LI_eff**	**MUML_LI_eff**	MU5_CIN_4km	RH_500850
2.	**MU5_LI_eff**	**MU_LI_eff**	**MU5_LI_eff**	**MUML_LI**	**MU5_LI**
3.	tciw	MUML_equal_layer	MUML_CAPE_HGL	**MU5_LI**	tciw
4.	tcslw	MUML_CAPE_HGL	**MU_LI_eff**	tciw	**MUML_LI**
5.	**MUML_LI_eff**	tciw	MU_CAPE_HGL	**MUML_LI_eff**	tcslw
6.	MU5_CAPE_HGL	**MU5_LI_eff**	MU_CIN_4km	tcslw	MU5_CAPE_HGL
7.	MUML_CAPE				km
8.	**MUML_LI**	**MU_LI**	BS_eff_MU	MU5_CAPE_HGL	MUML_CAPE_HGL
9.	LR_06km	tcslw	ML_LCL_hgt	**ML_LI_eff**	**MU5_LI_eff**
10.	**MU_LI_eff**	Moisture_Flux_SR	vigd	LR_06km	RH_ML_LCL_3km

MU5_LI_eff: The largest difference *between* lifted parcel (MU5) *and* ambient temperature found in the entire profile assuming environmental temperature is below -10 °C and the height is at least 2 km above lifted condensation level.

	Global	LM	OM	LT	OT
1.	MU5_CIN_4km	MUML_LI_eff	MUML_LI_eff	MU5_CIN_4km	RH_500850
2.	MU5_LI_eff	MU_LI_eff	MU5_LI_eff	MUML_LI	MU5_LI
3.	**tciw**	MUML_equal_layer	MUML_CAPE_HGL	MU5_LI	**tciw**
4.	**tcslw**	MUML_CAPE_HGL	MU_LI_eff	**tciw**	MUML_LI
5.	MUML_LI_eff	**tciw**	MU_CAPE_HGL	MUML_LI_eff	**tcslw**
6.	MU5_CAPE_HGL	MU5_LI_eff	MU_CIN_4km	**tcslw**	MU5_CAPE_HGL
7.	MUML_CAPE_HGL	MU5_CIN_4km	**tciw**	THETAE_02km	RH_14km
8.	MUML_LI	[illegible]	[illegible]	[illegible]	_CAPE_HGL
9.	LR_06km	**tcslw**	ML_LCL_hgt	ML_LI_eff	MU5_LI_eff
10.	MU_LI_eff	Moisture_Flux_SR	vigd	LR_06km	RH_ML_LCL_3km

tciw: Total column cloud ice water (ERA5).
tcslw: Total column supercooled liquid water (ERA5).

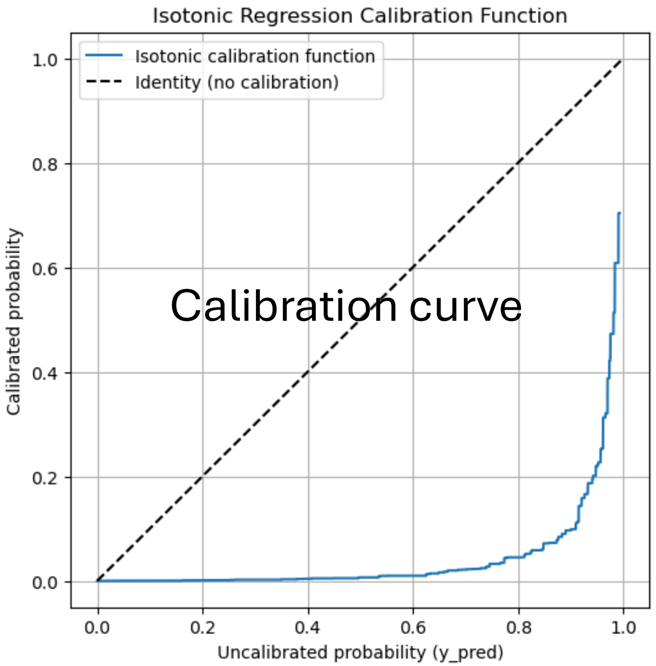
	Global	LM	OM	LT	OT
1.	MU5_CIN_4km	MUML_LI_eff	MUML_LI_eff	MU5_CIN_4km	RH_500850
2.	MU5_LI_eff	MU_LI_eff	MU5_LI_eff	MUML_LI	MU5_LI
3.	tciw	MUML_equal_layer	**MUML_CAPE_HGL**	MU5_LI	tciw
4.	tcslw	**MUML_CAPE_HGL**	MU_LI_eff	tciw	MUML_LI
5.	MUML_LI_eff	tciw	**MU_CAPE_HGL**	MUML_LI_eff	tcslw
6.	**MU5_CAPE_HGL**	MU5_LI_eff	MU_CIN_4km	tcslw	**MU5_CAPE_HGL**
7.	**MUML_CAPE_HGL**	MU5_CIN_4km	tciw	THETAE_02km	RH_14km
8.	MUML_LI	MU_LI	BS_eff_MU	**MU5_CAPE_HGL**	**MUML_CAPE_HGL**
9.	LR_06km	[illegible]	[illegible]	[illegible]	MU5_LI_eff
10.	MU_LI_eff	[illegible]	[illegible]	[illegible]	RH_ML_LCL_3km

MU5_CAPE_HGL : Convective Available Potential Energy *for* lifted parcel (MU5) in the Hail Growth Layer (-10 °C to -40 °C).

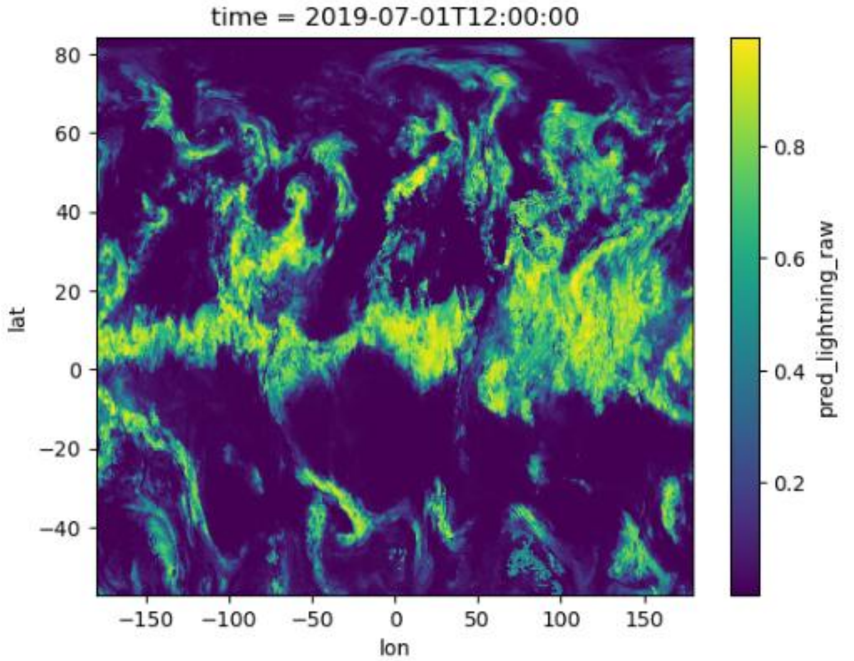
	Global	LM	OM	LT	OT
1.	MU5_CIN_4km	MUML_LI_eff	MUML_LI_eff	MU5_CIN_4km	**RH_500850**
2.	MU5_LI_eff	MU_LI_eff	MU5_LI_eff	MUML_LI	MU5_LI
3.	tciw	MUML_equal_layer	MUML_CAPE_HGL	MU5_LI	tciw
4.	tcslw	MUML_CAPE_HGL	MU_LI_eff	tciw	MUML_LI
5.	MUML_LI_eff	tciw	MU_CAPE_HGL	MUML_LI_eff	tcslw
6.	MU5_CAPE_HGL	MU5_LI_eff	MU_CIN_4km	tcslw	MU5_CAPE_HGL
7.	MUML_CAPE_HGL	MU5_CIN_4km	tciw	THETAE_02km	**RH_14km**
8.	MUML_LI	MU_LI	[illegible]	[illegible]E_HGL	MUML_CAPE_HGL
9.	LR_06km	tcslw	ML_LCL_hgt	ML_LI_eff	MU5_LI_eff
10.	MU_LI_eff	Moisture_Flux_SR	vigd	LR_06km	**RH_ML_LCL_3km**

RH_500850: Mean Relative Humidity *between* 500 hPa and 850 hPa.

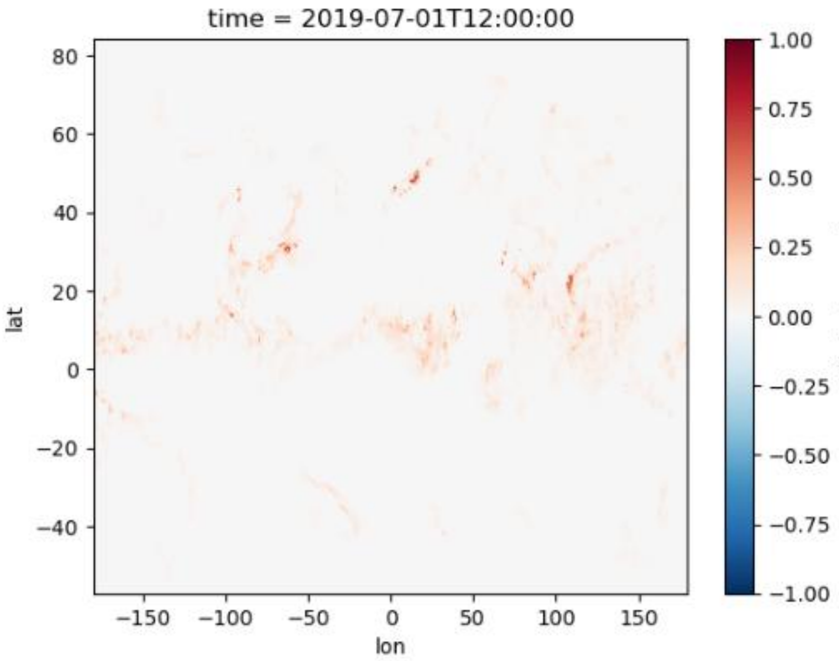
Calibrated lightning probability



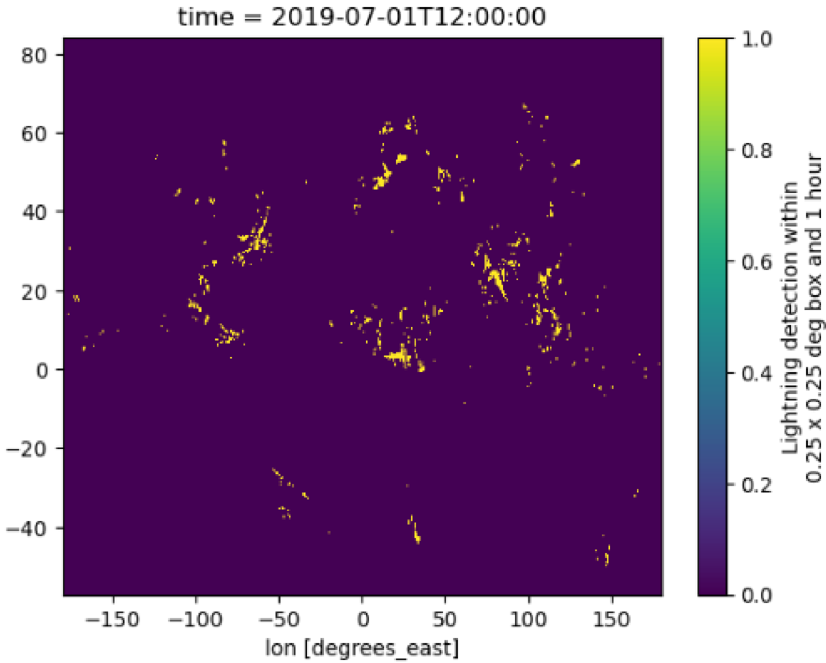
Raw XGBoost output



Calibrated lightning probability



Detected lightning



Concluding remarks

- *State-of-the-art prediction of lightning occurrence*; recommended number of convective parameters to use depend on desired trade-off between computational cost, interpretability and discriminative power.
- Different flavors of the same convective parameter help to boost performance. Open question: *why is it the case?*
- Small performance gap between regionalized and global model: *could new predictors close that gap?*
- Final calibrated model needs to be verified and benchmarked against baselines on full test year, then potential to deploy the model for climatology, forecasts and climate change studies.

Contact information: Dominique.Brunet@ec.gc.ca

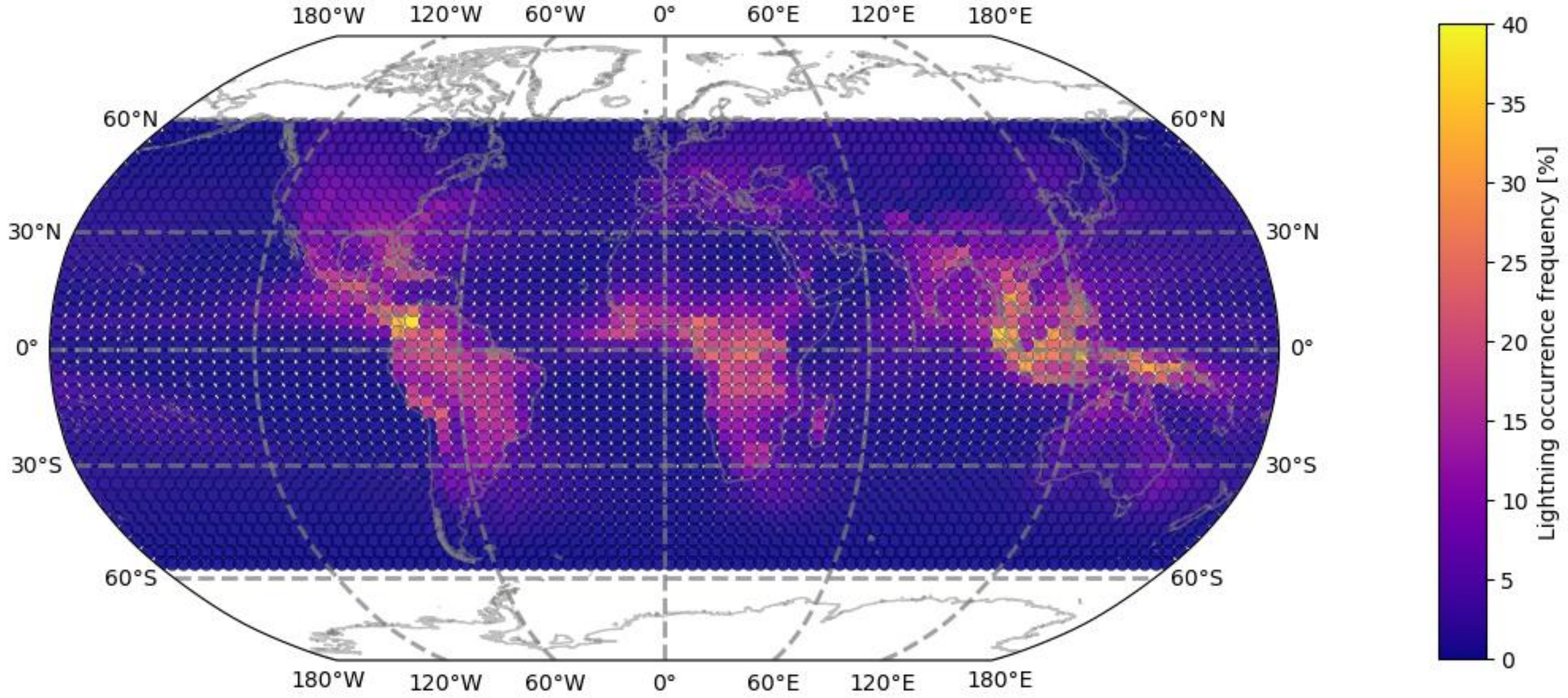
EXTRAS

Convective Parameters Count

Convective parameter type	Prefix	Count
Most-unstable parcel (CAPE, CIN, LI, buoy, etc.)	MU	25
Most-unstable-mean-layer parcel (CAPE, CIN, LI, buoy, etc.)	MUML	25
Most-unstable above 500 m parcel (CAPE, CIN, LI, buoy, etc.)	MU5	16
Mixed-layer parcel (CAPE, CIN, LI, buoy, etc.)	ML	7
Bulk wind shear (from level X to Y)	BS	19
Wind speed (mean or max between level X and Y)	MW, WS	6
Storm-relative winds, storm-motion vectors	SRW, CBV	9
Moisture parameters (mean between level X and Y, flux)	RH, Moisture, PRCP	21
Temperature lapse rates (between level X and Y)	LR	10
Theta-e and downdraft parameters	THETAE, Cold_Pool	9
Height of specific levels (isotherms, parcel)	HGT	6
Direct ERA5 fields (precipitation, total column water)		13
Static features (angle, mask)	Sun, Land	3
Lightning climatology	Climatology	1
Total (no composite or SB parameters)		**170**

Earth Networks Global Lightning Detection Network (ENGLN)

Lightning occurrence: lightning detection +/-1h and +/-0.25deg around ERA5 grid

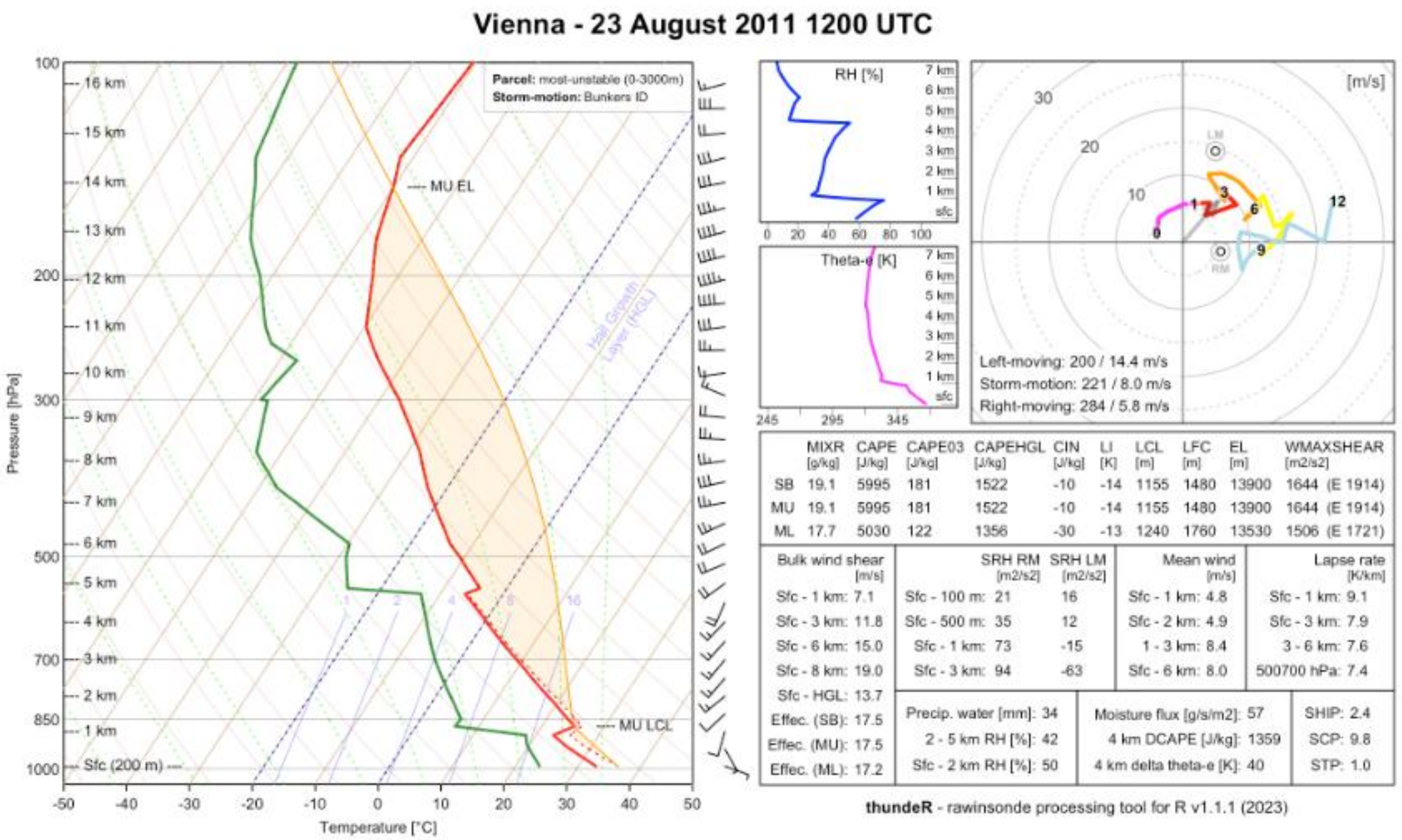


Period: 2019-2023

ThundeR

v.1.5 (pre-release)

```
data("sounding_vienna") # load example dataset (Vienna rawinsonde profile for 23 Aug 2011 12UTC):
pressure = sounding_vienna$pressure # vector of pressure [hPa]
altitude = sounding_vienna$altitude # vector of altitude [meters]
temp = sounding_vienna$temp  # vector of temperature [degree Celsius]
dpt = sounding_vienna$dpt # vector of dew point temperature [degree Celsius]
wd = sounding_vienna$wd # vector of wind direction [azimuth in degrees]
ws = sounding_vienna$ws # vector of wind speed [knots]
sounding_save(filename = "Vienna.png", title = "Vienna - 23 August 2011 1200 UTC", pressure, altitude, tem
p, dpt, wd, ws)
```



- Computation and visualisation of an extensive suite of convective parameters for severe weather forecasting
- Efficient computation in C++
- R interface and Python calls via rpy2 library: https://bczernecki.github.io/thundeR/
- Web interface: www.rawinsonde.com

XGBoost

XGBoost is an optimized distributed gradient boosting library designed to be highly **efficient**, **flexible** and **portable**.

- Type of **Ensemble Decision Trees**
- **Gradient Boosting**: Alternative training method to **Random Forests**
- Can handle **billions** of examples

```python
# Model Parameters
params = {
    'max_depth': 10,
    'learning_rate': 0.003,
    'eval_metric': 'logloss',
    'tree_method': 'hist',
    'objective': 'binary:logistic',
    'verbosity': 1,
}
n_boosting_rounds = 3500
```

Hyperparameters optimized with GridSearchCV

Data pre-processing

- #features: 170 (153 after removing features with # Nans > 1%)
- #samples:
  - LM = 8M
  - OM = 17M
  - LT = 8M
  - OT = 19M

  Total: 51M
- Data split:
  - Train = 2019-2021
  - Validation = 2022
  - Test = 2023
- Data balancing (random under-sampling):
  - LM = 3.1% => train = 0.3M, validation = 0.1M, test = 0.1M
  - OM = 0.8% => train = 0.2M, validation = 0.1M, test = 0.1M
  - LT = 10.7% => train = 1M, validation = 0.3M, test = 0.3M
  - OT = 2.7% => train = 0.6M, validation = 0.2M, test = 0.2M

Rank	Convective parameters	Normalized information gain
1	MU5_CIN_4km	100.0%
2	MU5_LI_eff	95.9%
3	tciw	32.8%
4	tcslw	16.9%
5	MUML_LI_eff	10.0%
6	MU5_CAPE_HGL	8.2%
7	MUML_CAPE_HGL	7.7%
8	MUML_LI	7.3%
9	LR_06km	6.6%
10	MU_LI_eff	4.8%
11	LR_16km	4.4%
12	vigd	3.4%
13	RH_500850	3.0%
14	MU5_buoy_HGL	3.2%
15	MU_CAPE_HGL	3.0%
16	LR_03km	2.8%
17	HGT_ISO_M10	2.8%
18	RH_14km	2.5%
19	THETAE_02km	2.5%
20	MU_LI	2.4%
21	MU5_LI	2.4%
22	RH_ML_LCL_3km	2.4%
23	RH_25km	2.3%
24	MU5_CAPE_M10	2.2%
25	THETAE_01km	2.1%
26	tcrw	2.0%
27	LR_04km	1.8%
28	cp	1.8%
29	vimd	1.7%
30	tp	1.7%
31	MU_CIN_4km	1.7%
32	MU5_CAPE	1.6%
33	MU_cold_cloud	1.6%
34	SRW_0500m_MW	1.6%
35	tclw	1.5%
36	LR_26km	1.5%
37	BS_18km	1.5%
38	Cold_Pool_Strength	1.4%
39	MUML_CAPE	1.4%
40	HGT_ISO_0	1.4%

Explanation of “features”

1. **MU5_CIN_4km** : Convective inhibition for lifted parcel (MU5) in 4 km layer with the bottom of that layer at the parcel lifting height.
2. **MU5_LI_eff** : The largest difference *between* lifted parcel (MU5) *and* ambient temperature found in the entire profile assuming environmental temperature is below -10 °C and the height is at least 2 km above lifted condensation level.
3. **tciw** : Total column cloud ice water (ERA5).
4. **tcslw** : Total column supercooled liquid water (ERA5).
5. **MUML_LI_eff** : Same as 2., but the lifted parcel (MUML) is the most-unstable parcel after mixing 500 m layers (moving window), up to 3 km AGL.
6. **MU5_CAPE_HGL** : Convective Available Potential Energy *for* the most-unstable parcel after excluding lowest 500m from the profile *and* in the Hail Growth Layer (-10 °C to -40 °C).

Interpretation of “features”

- **CIN** - highlights convective initiation failure modes when there is too much inhibition
- **CAPE_HGL** - measure of buoyancy in a lightning-favorable charge separation zone
- **LI** - measure of atmospheric instability, supplements CAPE in marginal situations when CAPE is zero, but lightning can still form
- **tciw / tcslw** - proxy for lift, indicative of presence of hail/graupel
- Both **MU5** and **MUML** parcels are worth to use simultaneously and they supplements each other and account for different atmospheric scenarios