Machine Learning based investigation of the variables affecting summertime lightning over South Great Plain and Southeastern South America

[1]Siyu Shan, [1]Dale Allen, [1,2]Zhanqing Li, [1]Kenneth Pickering

Email: syshan@umd.edu

[1]Department of Atmospheric and Oceanic Science

[2]Earth System Science Interdisciplinary Center (ESSIC)
University of Maryland, College Park

August 10th, 2023 ASR Meeting

Background

- Lightning is affected by many factors, many of which are not routinely measured, well understood, or accounted for in physical models.

- CAPE
- Rain rate
- Cloud thickness: warm/cold cloud depth
- Wind shear
- Aerosol
- ……

Scientific Questions

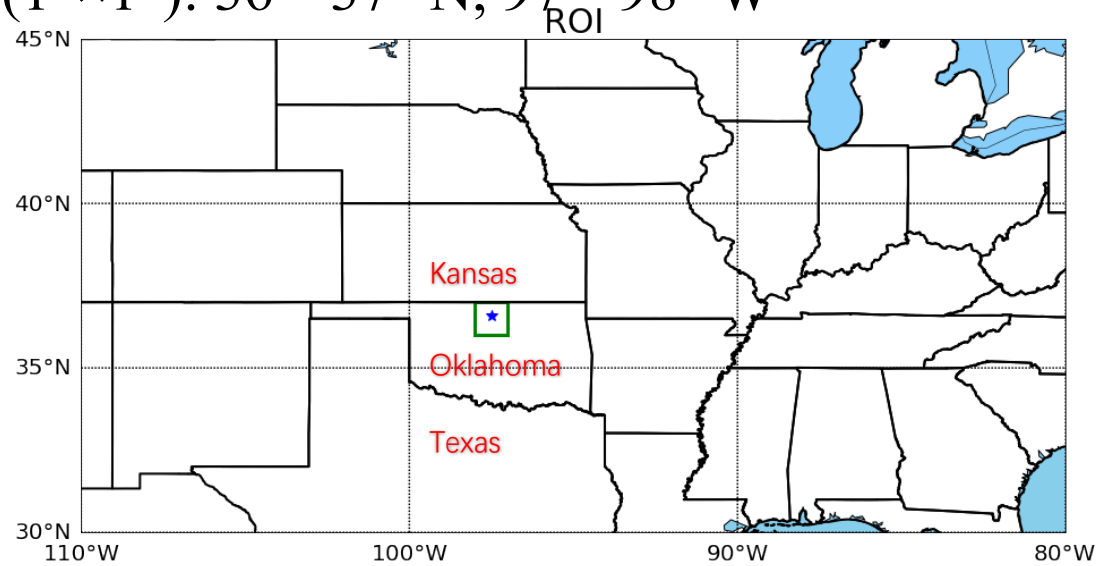
1. What are the meteorological variables most strongly associated with lightning at the Southern Great Plains (SGP) site?

2. What is the best predictor of Intra-Cloud (IC) flash fraction?

3. What are the meteorological variables most strongly associated with lightning at the Southeastern South America in clean and polluted cases?

Region of Interest

- Green Box (1°×1°): 36°~37° N, 97°~98° W



- Atmospheric Radiation Measurement (ARM)
- Southern Great Plains (SGP) Site (blue star): 36.6° N, 97.5° W

Time Period and Data Sources

Time Period

- 2012-2020 Summer Months: June, July, August and September (JJAS)

Data Sources

- Lightning Flashes: ENTLN (Earth Networks Total Lightning Network)
- Most meteorological variables: (i.e., CAPE, Cloud Thickness, Rain Rate, > 10 dBz Radar Reflectivity Volume & Centroid, CCN concentration) obtained or calculated from ARM SGP site
- Wind: ERA-5 Reanalysis
- $PM_{2.5}$ Concentration: US EPA (Environmental Protection Agency) Measurements

ML is used to predict whether lightning will occur in a 1°×1° grid box centered at SGP during hours that convective clouds are detected at ARM site.

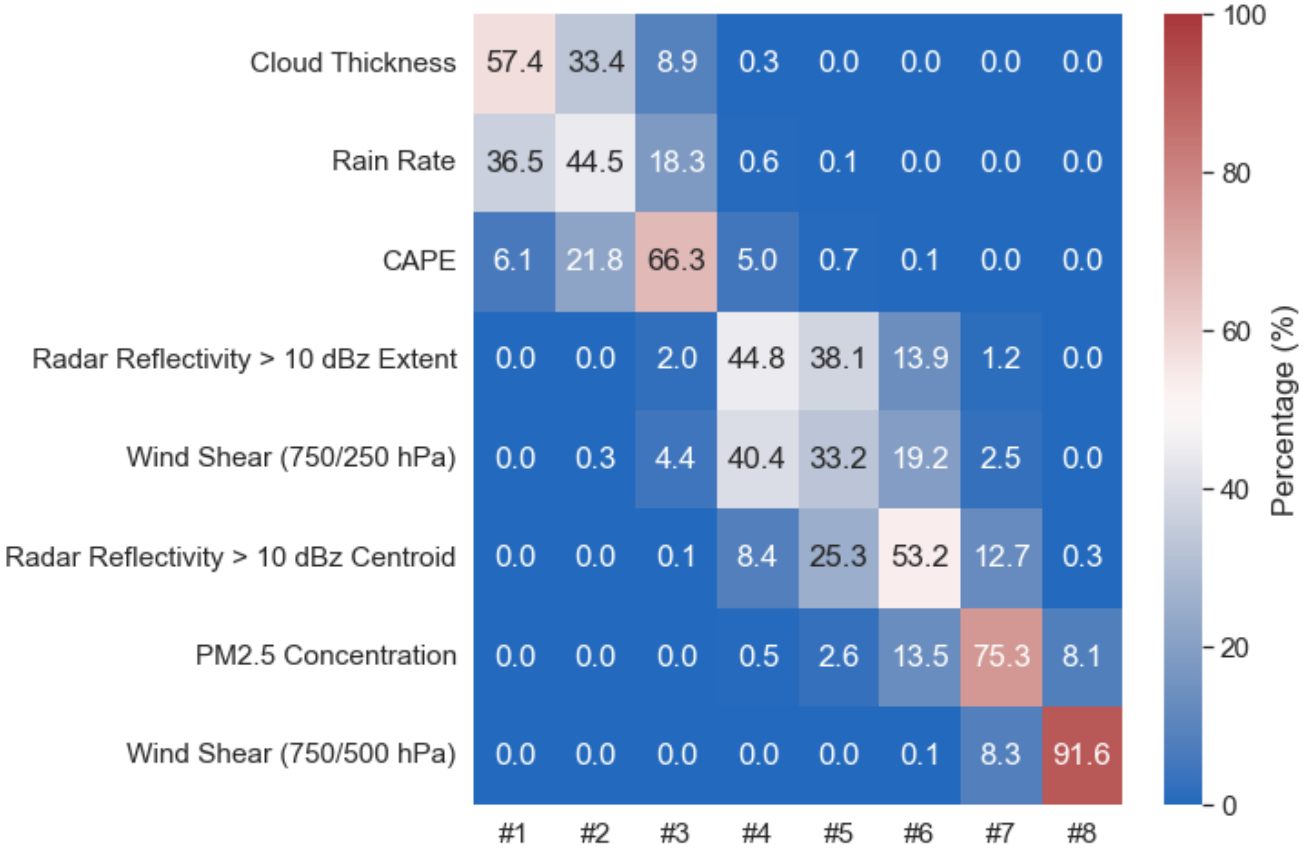
Error Matrix

	Prediction: No Lightning	Prediction: Lightning
Truth: No Lightning	24.5 % ± 2.5 %	13.0 % ± 2.6 %
Truth: Lightning	10.4 % ± 2.5 %	52.1 % ± 2.9 %

ML Model Results

- Random Forest with K-Fold (10 repeats, 10 splits)
- Performance:
  - Accuracy: ~77%
  - Area Under Curve (AUC): 0.85 ± 0.04

CAPE, Cloud Thickness and Rain Rate are the Most Important Variables determining lightning occurrence.



- How do we know this?
- This plot shows the chance of variables being the different most importance in the ML model.
- For example, in the ML model, Cloud Thickness is the most important (#1) variable among 57.4% of the cases, in 33.4% of the cases it is the second important (#2) and in 8.9% of the cases it is third important (#3).

Intra-Cloud (IC) Flash Fraction is positively correlated with $\sqrt{CAPE}$

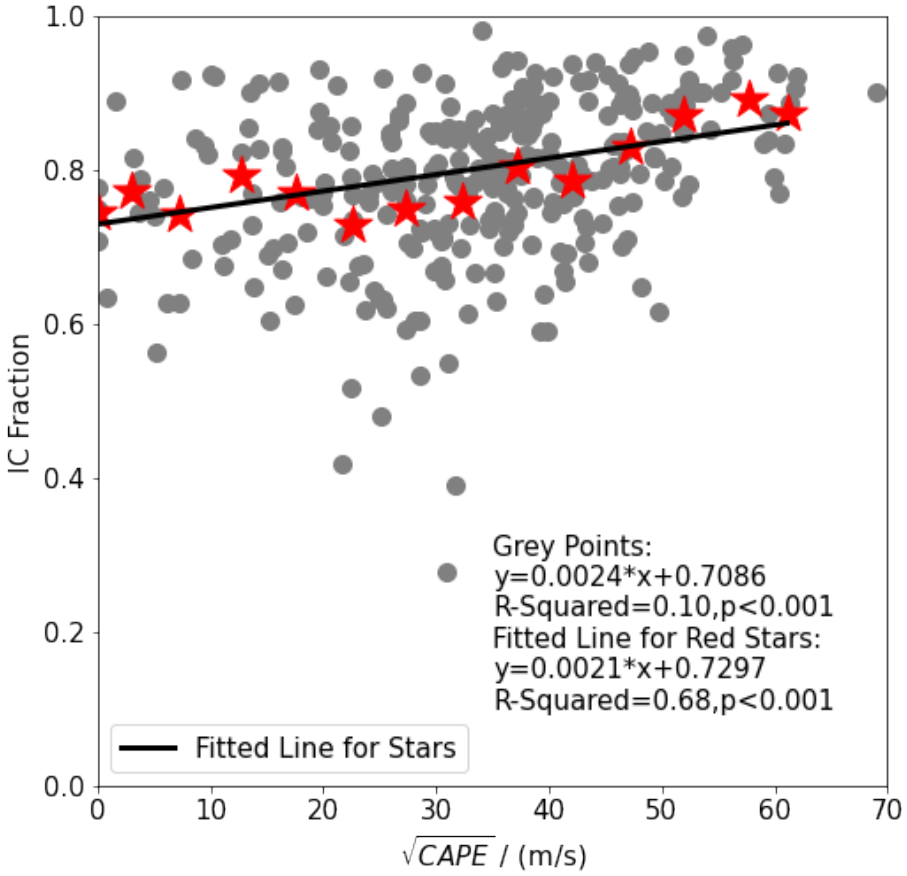
Convective Hours Identification

- Analysis restricted to 256 hours with convective clouds detected at ARM SGP site and plentiful flashes (Hourly Flash Count > Flash Median = 162.5), to ensure the statistics are meaningful.

Intra-Cloud (IC) Flash Fraction

- IC Flash Fraction = IC Flashes / Total Flashes
- $\sqrt{CAPE}$ has the most positive linear coefficient value with IC flash fraction.

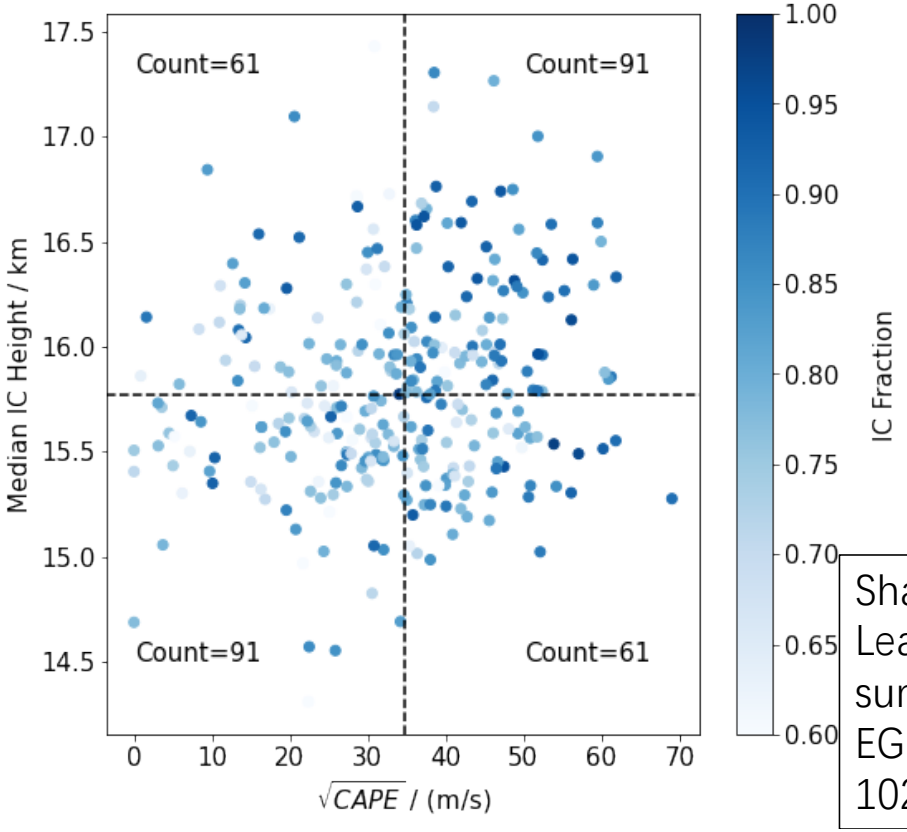
Intra-Cloud (IC) Flash Fraction is positively correlated with $\sqrt{CAPE}$



Intra-Cloud (IC) Flash Fraction

- Scatter plot shows the positive relationship between $\sqrt{CAPE}$ and IC flash fraction.
- Red stars show means for 14 bins.
- As $\sqrt{CAPE}$ increases from 0 to 60 m/s, IC flash fraction increases from 0.7 to ~0.9.

Intra-Cloud (IC) Flash Fraction is positively correlated with $\sqrt{CAPE}$



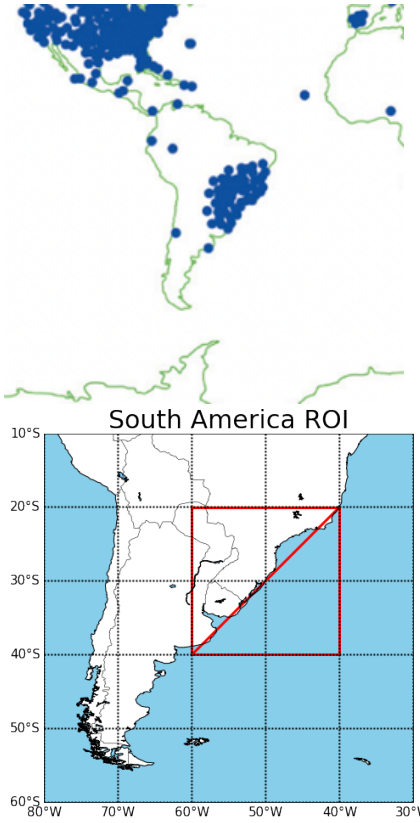
- Hypothesis: Higher $\sqrt{CAPE}$ represents a stronger convective environment. Stronger updraft brings the electrification zone further above the SFC, resulting in a greater IC flash fraction.
- This hypothesis is supported by the fact that higher IC flash fractions are associated with higher IC heights.
- Chi-square test: $p < 0.001$

Shan, S., Allen, D., Li, Z., Pickering, K., and Lapierre, J.: Machine Learning based investigation of the variables affecting summertime lightning frequency over the Southern Great Plains, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2023-1020, 2023.

SA ROI (20°-40° S, 40°-60° E)

- Pros:
- 1. Plenty of ENTLN sensors and long enough observations
- 2. Biomass burning (polluted/clean contrast)
- 3. Land/ocean contrast
- 4. Few study about lightning over this region

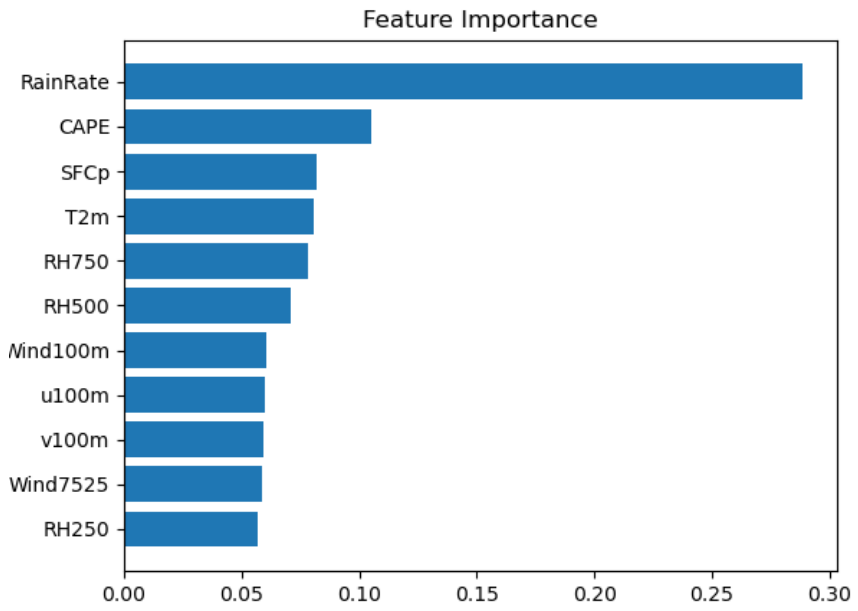
- Data Sources:
- ERA-5
- Rain Rate from IMERG
- Lightning from ENTLN



Meteorological Variables Chosen for ML Investigation of Lightning > Median? (80539 data points > median)

- All (1631088): accuracy 98.0%, AUC 0.970
- Excellent performance!

- All = Clean + Polluted
- VIIRS daily AOT > 75% (polluted) or < 25% (clean)



Meteorological Variables Chosen for ML Investigation of Lightning > Median? (80539 data points > median)

- Ocean (892464): 99.4%, 0.973
- Land (738624): 96.4%, 0.957
- Clean (815544): 99.2%, 0.974
- Polluted (815544): 96.8%, 0.961

- Ocean & Clean (446232): 99.7%, 0.976
- Ocean & Polluted (446232): 99.0%, 0.969
- Land & Clean (369312): 98.6%, 0.970
- Land & Polluted (369312): 94.3%, 0.940

Scientific Questions & Results Review

1. What are the meteorological variables most strongly associated with lightning at the Southern Great Plains (SGP) site?

CAPE (Convective Available Potential Energy), cloud thickness and rain rate are most important, based on ML.

2. What is the best predictor of Intra-Cloud (IC) flash fraction?

The square root of CAPE ($\sqrt{CAPE}$).

3. What are the meteorological variables most strongly associated with lightning at the Southeastern South America in clean and polluted cases?

CAPE and rain rate.

Thanks for your listening!

- Comments and questions are welcome!
- Email: syshan@umd.edu