Applying and Evaluating Tree-Based Ensemble Methods to Identify Ice Crystal Chain Aggregates during the IMPACTS Field Campaign

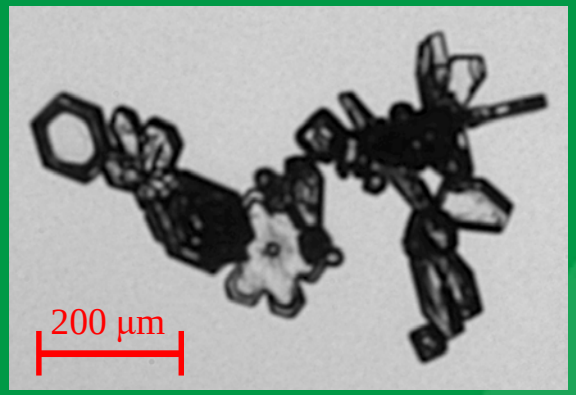
Christian Nairy[1] (christian.nairy@und.edu), David Delene[1], Joseph Finlon[2, 3], John Yorks[2], and Kenneth L. Thornhill[4]

[1]University of North Dakota, Department of Atmospheric Science

[2]National Aeronautics and Space Administration (NASA) - Goddard Space Flight Center

[3]Earth System Science Interdisciplinary Center (ESSIC), University of Maryland

[4]National Aeronautics and Space Administration (NASA), Langley Research Center



105th AMS Annual Meeting – Second Symposium on Cloud Physics: New and Emerging Measurement Methods in Cloud and Precipitation Research III

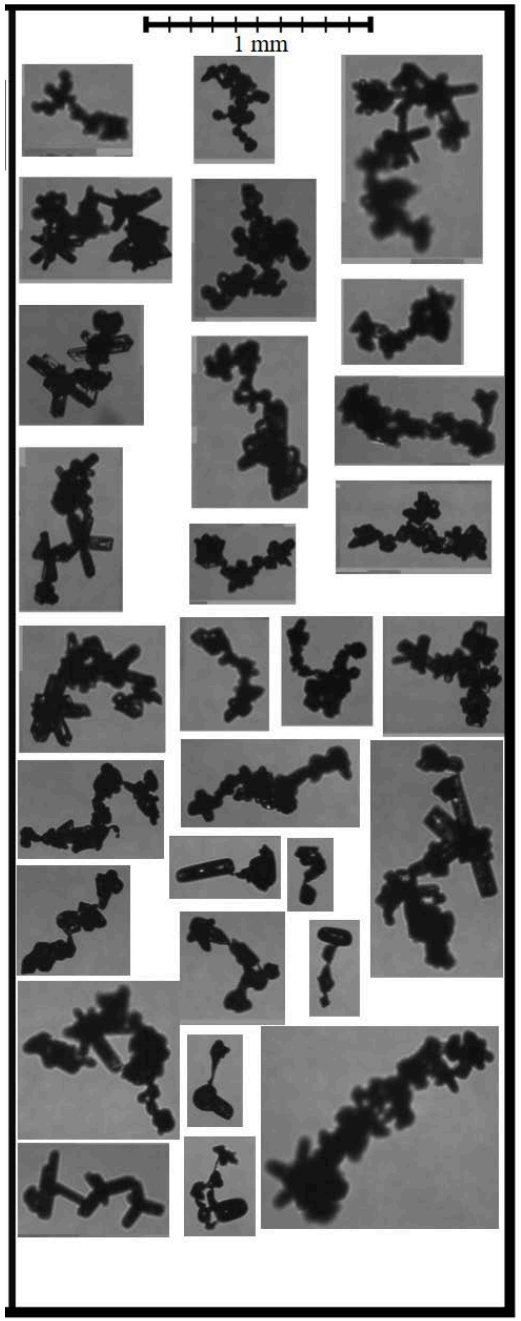
Motivation for Research

- The chain aggregation process is still not well understood.
  - Where/How?
  - Inconsistencies between cloud chamber experiments and aircraft observations.
  - Lack of representation in atmospheric cloud models.

- Influence on cloud radiative transfer properties (Liou 1973; Stephens *et al.* 1990; Baran 2009).

- Contextualizing chain aggregates within storm life-cycles requires analyzing diverse cases, **but manually classifying tens of thousands of particles is highly time-intensive.**

Hawkeye-CPI Images



IMPACTS

Objective

Employ supervised decision tree-based methods to differentiate ice crystal chain aggregates from non-chain aggregates.

Methods:

1. Particle Classifications
   - Classify observed (in-situ) chain and non-chain aggregates from diverse cloud regions from the IMPACTS field campaign.

2. Morphological Particle Attributes and Environmental Parameters
   - Extract attributes and parameters for each classified particle
   - Particle Attributes: D_{max}, Area Ratio, Complexity, Compactness, Curl, Fine Detail, etc.
   - Environmental Parameters: Temperature & Liquid Water Content (LWC)

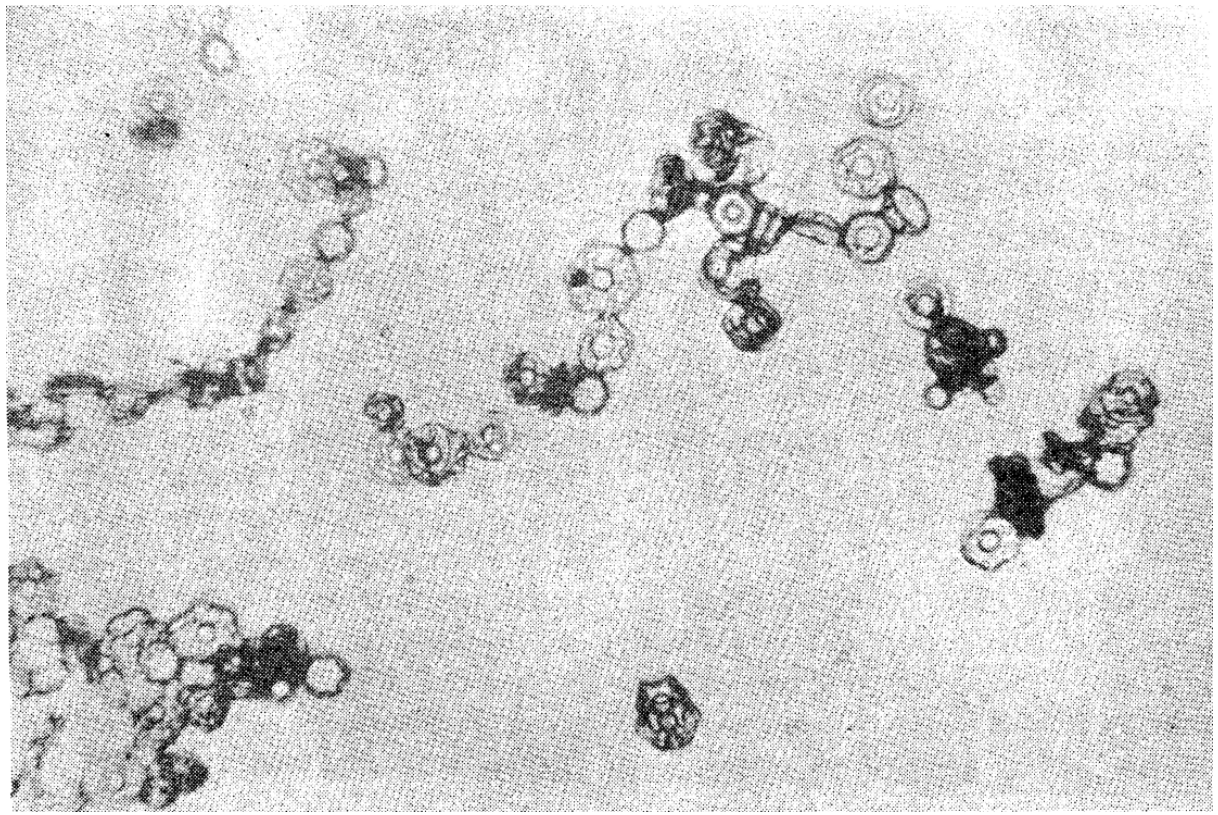
3. Hyper-tune Two Tree-Based Classifiers
   - Train ➢ Test ➢ Validate

4. Comparison of Decision Tree-Based Methods

Previous Chain Aggregate Observations

Laboratory Observations of Chain Aggregates:

- In High Electric Fields (Minimum Threshold: 60 kV m^{-1})
- -5 and -37 °C (Maximum efficiency at -8 °C)
- Ice Crystal Concentrations Between 3 and 4×10^6 m^{-3}
- Aggregation found to be temperature-dependent with electric fields.

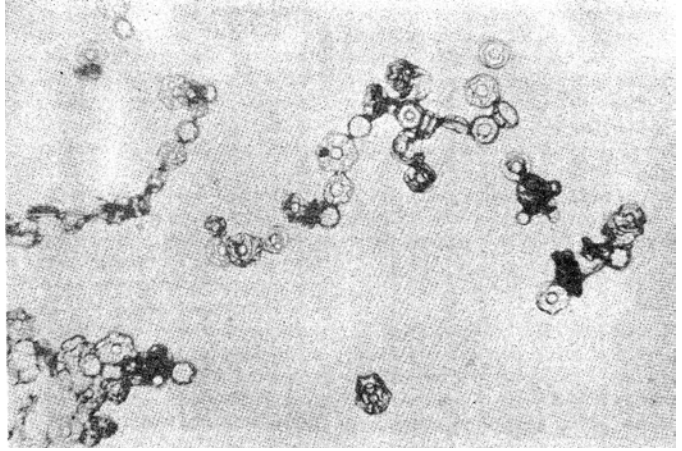


Chain Aggregates on formvar slides that were generated in a **cloud chamber.** Adapted from Saunders and Wahab, 1975.

Previous Chain Aggregate Observations

Laboratory Observations of Chain Aggregates:

- In High Electric Fields (Minimum Threshold: 60 kV m^{-1})
- -5 and -37 °C (Maximum efficiency at -8 °C)
- Ice Crystal Concentrations Between 3 and 4×10^6 m^{-3}
- Aggregation found to be temperature-dependent with electric fields.

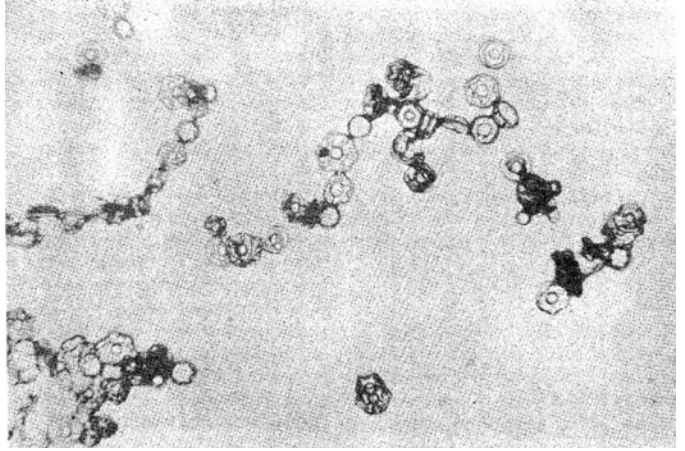


Chain Aggregates on formvar slides that were generated in a **cloud chamber.** Adapted from Saunders and Wahab, 1975.

Previous Chain Aggregate Observations

Laboratory Observations of Chain Aggregates:

- In High Electric Fields (Minimum Threshold: 60 kV m^{-1})
- -5 and -37 °C (Maximum efficiency at -8 °C)
- Ice Crystal Concentrations Between 3 and 4×10^6 m^{-3}
- Aggregation found to be temperature-dependent with electric fields.



Chain Aggregates on formvar slides that were generated in a **cloud chamber.** Adapted from Saunders and Wahab, 1975.

Aircraft Observations of Chain Aggregates:

- Mid- to upper-level clouds produced by **summertime** continental convection in the tropics and sub-tropics.
  - Field Campaigns: CapeEx19, CRYSTAL-FACE, ABFM-II, EMERALD-II
  - -8 to -65 °C (but mainly found < -25 °C)
  - In-situ E-Fields (CapeEx19 & ABFM-II) < 60 kV m^{-1}
- **Wintertime** Continental Convection in the Midlatitudes.
  - Field Campaigns: IMPACTS
  - -13 to -35 °C

Previous Chain Aggregate Observations

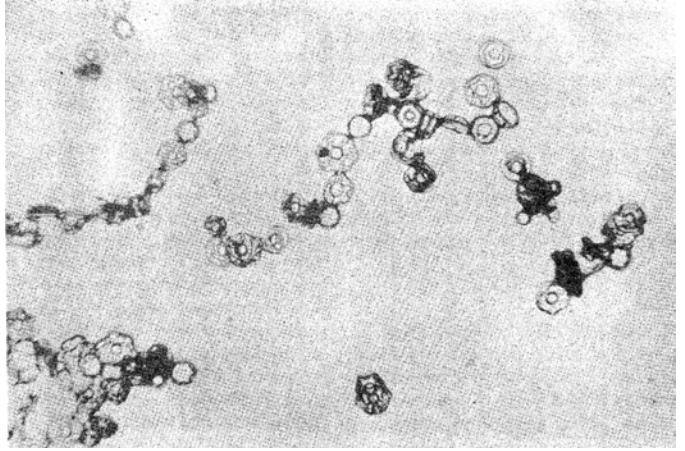
Laboratory Observations of Chain Aggregates:

- In High Electric Fields (Minimum Threshold: 60 kV m^{-1})
- -5 and -37 °C (Maximum efficiency at -8 °C)
- Ice Crystal Concentrations Between 3 and 4×10^6 m^{-3}
- Aggregation found to be temperature-dependent with electric fields.

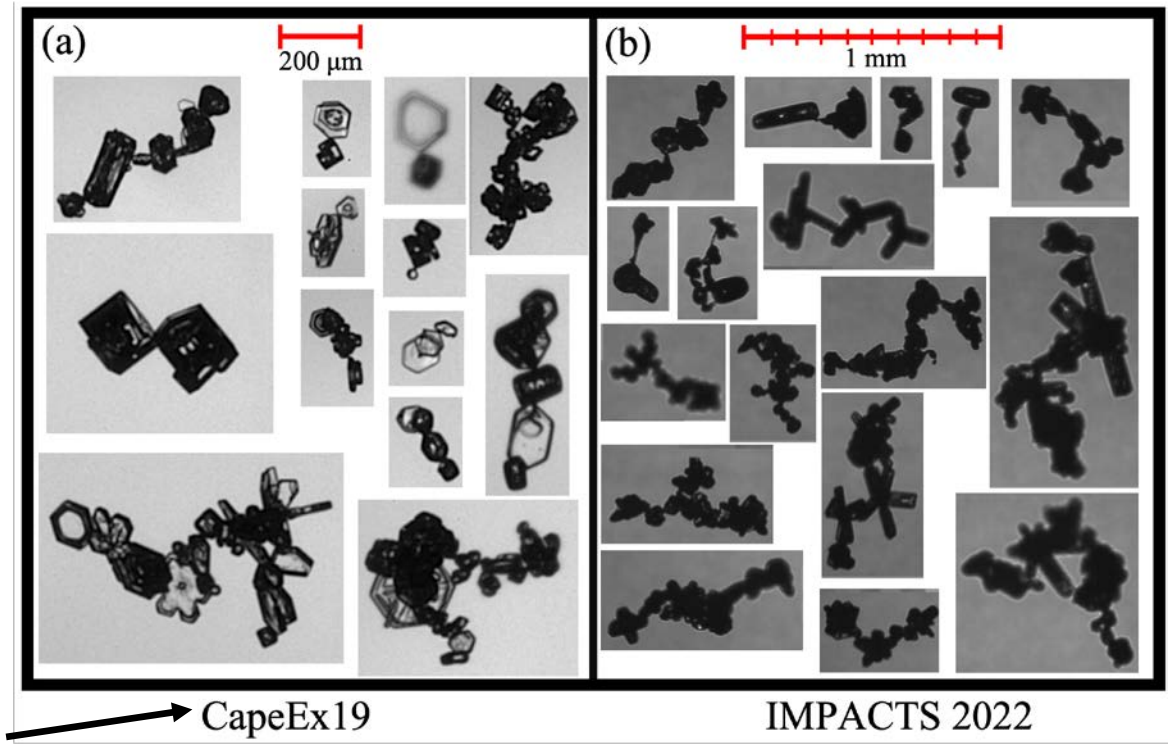
Aircraft Observations of Chain Aggregates:

- Mid- to upper-level clouds produced by **summertime** continental convection in the tropics and sub-tropics.
  - Field Campaigns: CapeEx19, CRYSTAL-FACE, ABFM-II, EMERALD-II
  - -8 to -65 °C (but mainly found < -25 °C)
  - In-situ E-Fields (CapeEx19 & ABFM-II) < 60 kV m^{-1}
- **Wintertime** Continental Convection in the Midlatitudes.
  - Field Campaigns: IMPACTS
  - -13 to -35 °C

Nairy, C. M., D. J. Delene, A. G. Detwiler, J. M. Schmidt, P. R. Harasti, M. Schnaiter, E. Järvinen, T. D. Walker, Ice Crystal Chain Aggregates in Florida Cirrus Cloud Anvils - 3 August 2019 Case Study. Journal of Geophysical Research: Atmospheres. In Review, 2025



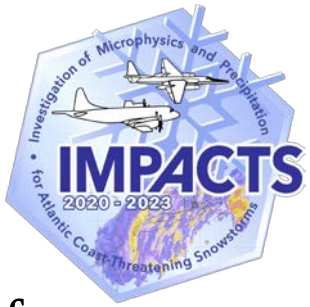
Chain Aggregates on formvar slides that were generated in a **cloud chamber.** Adapted from Saunders and Wahab, 1975.



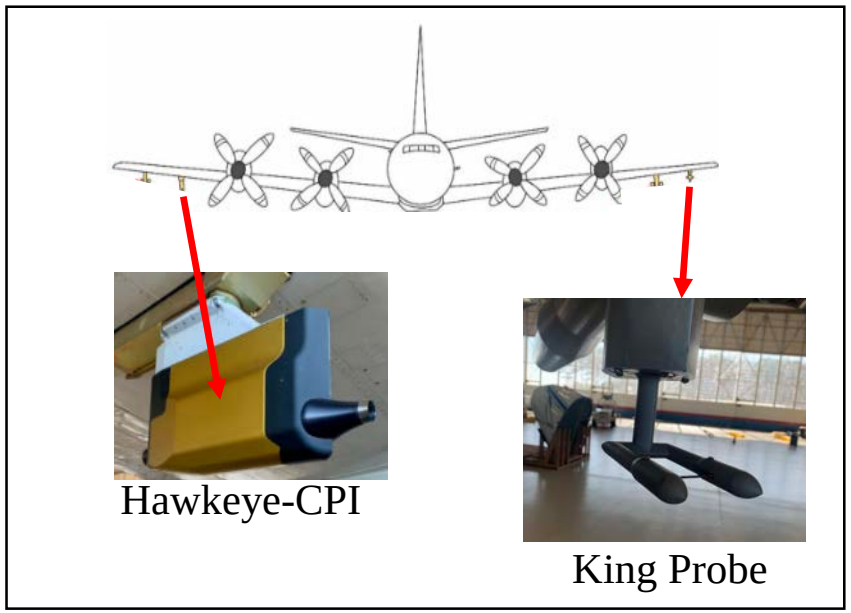
(a) PHIPS images (**in-situ**) of chain aggregates comprised of ice crystals found in cirrus anvil clouds during the CapeEx19 field campaign and (b) CPI images (**in-situ**) of chain aggregates found near the parameter of a convective band associated with a nor'easter off the New England coast.

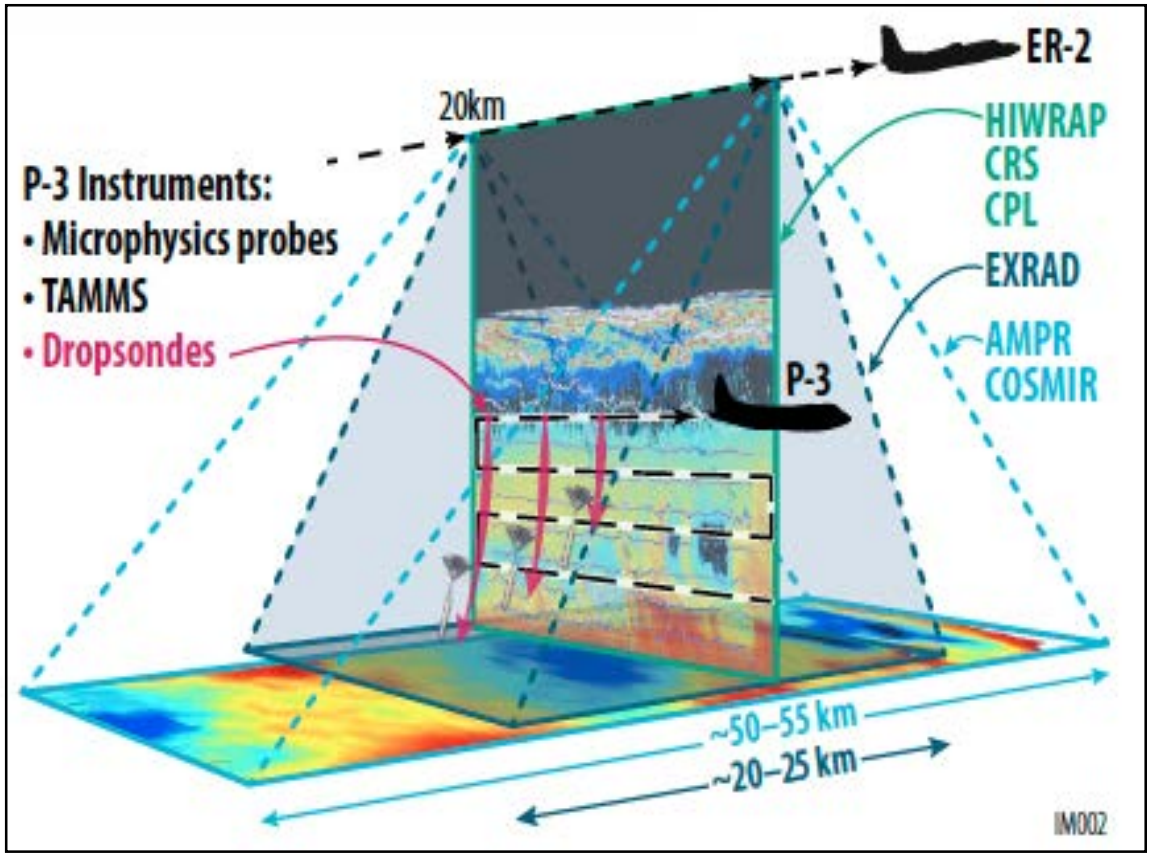
Dataset/Instrumentation

- **I**nvestigation of **M**icrophysics and **P**recipitation for **A**tlantic **C**oast-**T**hreatening **S**nowstorms (IMPACTS).



- NASA P-3b Orion Research Aircraft
  - Hawkeye-Cloud Particle Imaging (CPI) Probe
  - King Probe (Liquid Water Content [LWC])





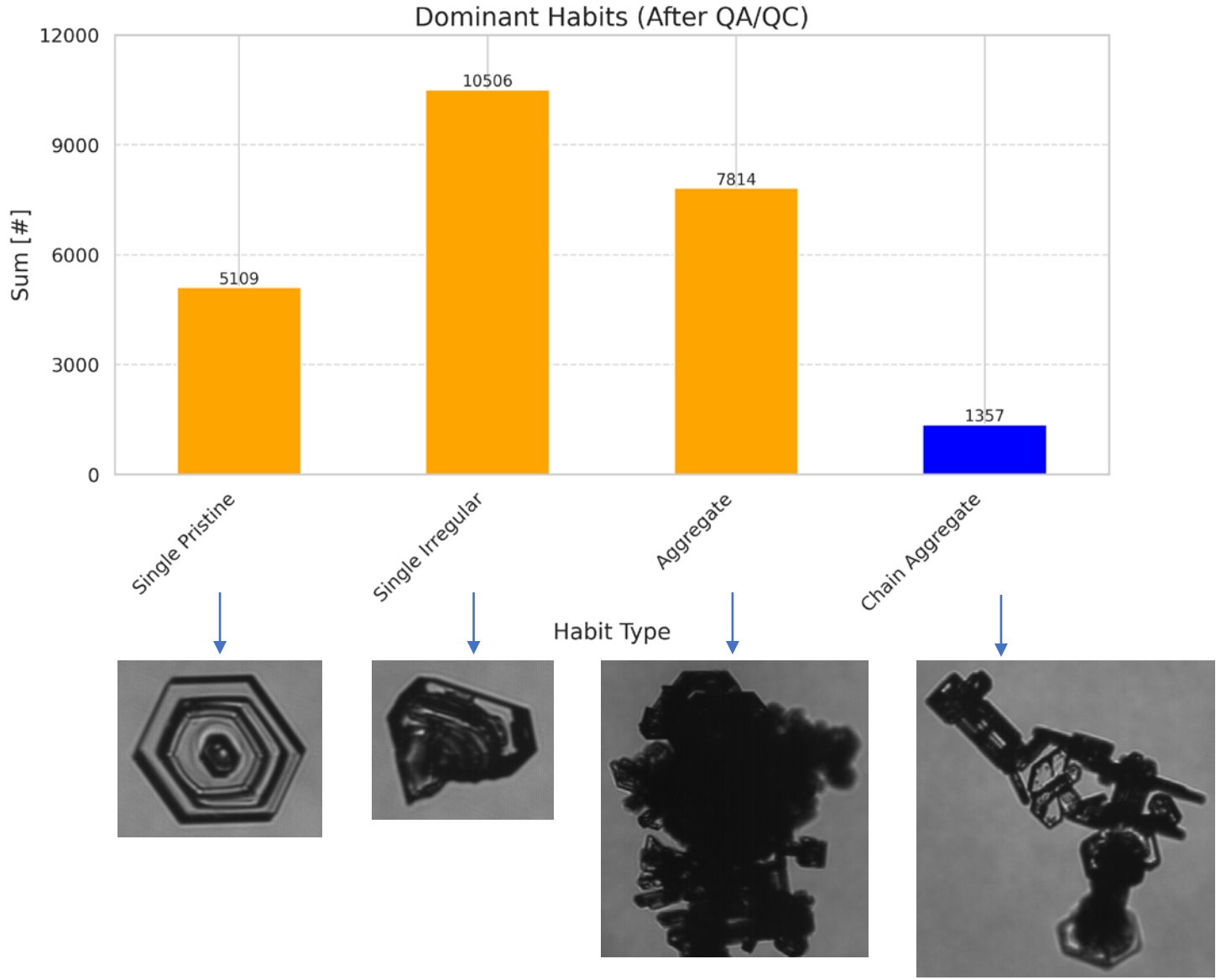
Adapted from the NASA IMPACTS executive summary (https://espo.nasa.gov/impacts/).

Particle Classifications

- 5 Flight Segments Classified:
  Total Time Coverage: 1hr 40m 49s
  Temperature Range: -36 to -5 ºC

- Total # of Particles Classified (after QA/QC):
  - 24,786

- # of Chain Aggregates Classified (after QA/QC):
  - 1,357



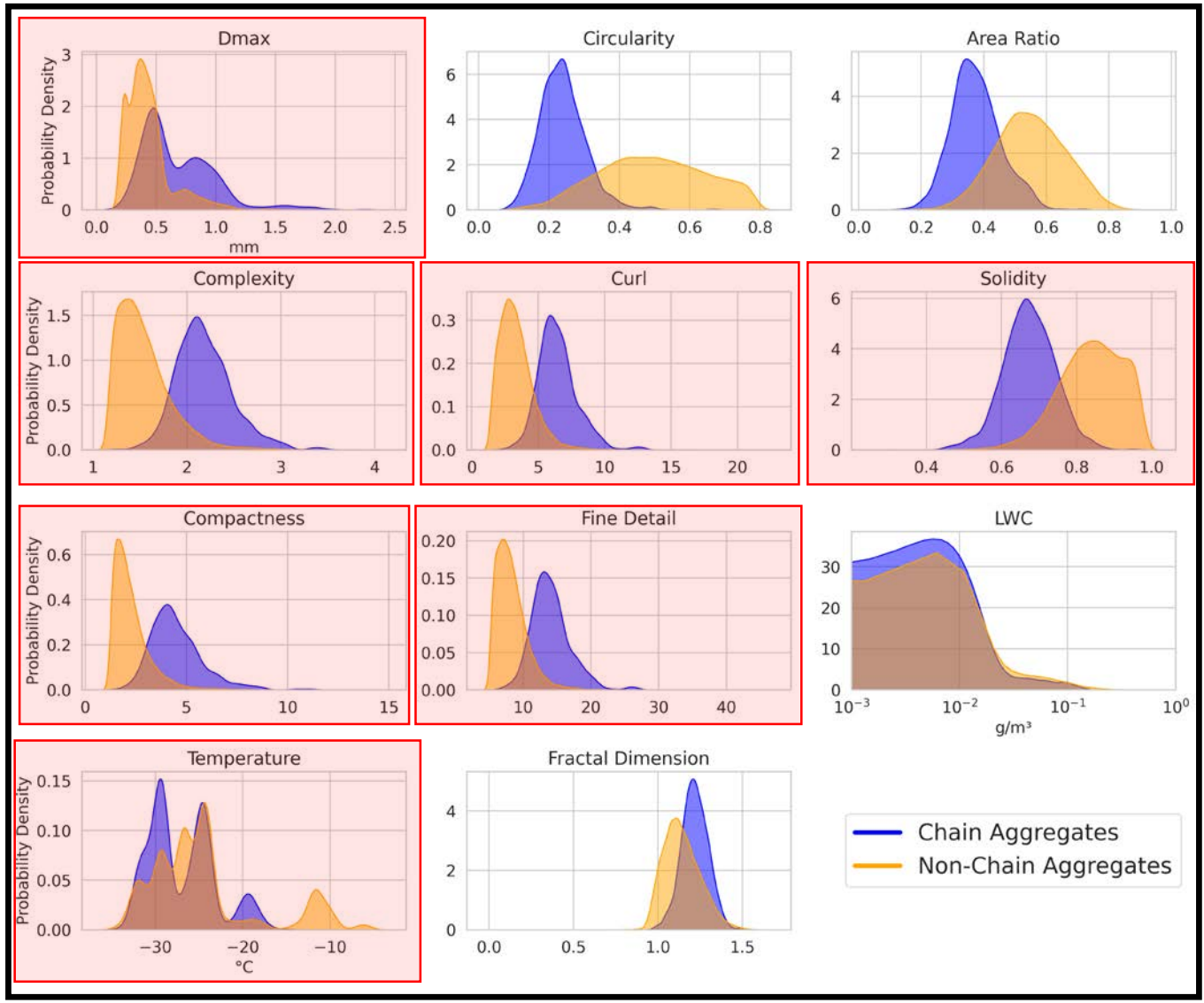
NOT TO SCALE

Morphological Particle Attributes and Environmental Parameters

- Attributes calculated for “all-in” CPI imaged particles

- Key attributes/parameters used found that optimized the decision tree methods:
  - D_{max}, Complexity*, Curl*, Compactness*, Fine Detail*, Solidity*, & Temperature

**Equations in extra slides*



Tree-Based Classifier Setup

- Hyper-tuned for the most optimal result.
  - Key Parameters Used: D_{max}, Complexity*, Curl*, Compactness*, Fine Detail*, Solidity*, & Temperature

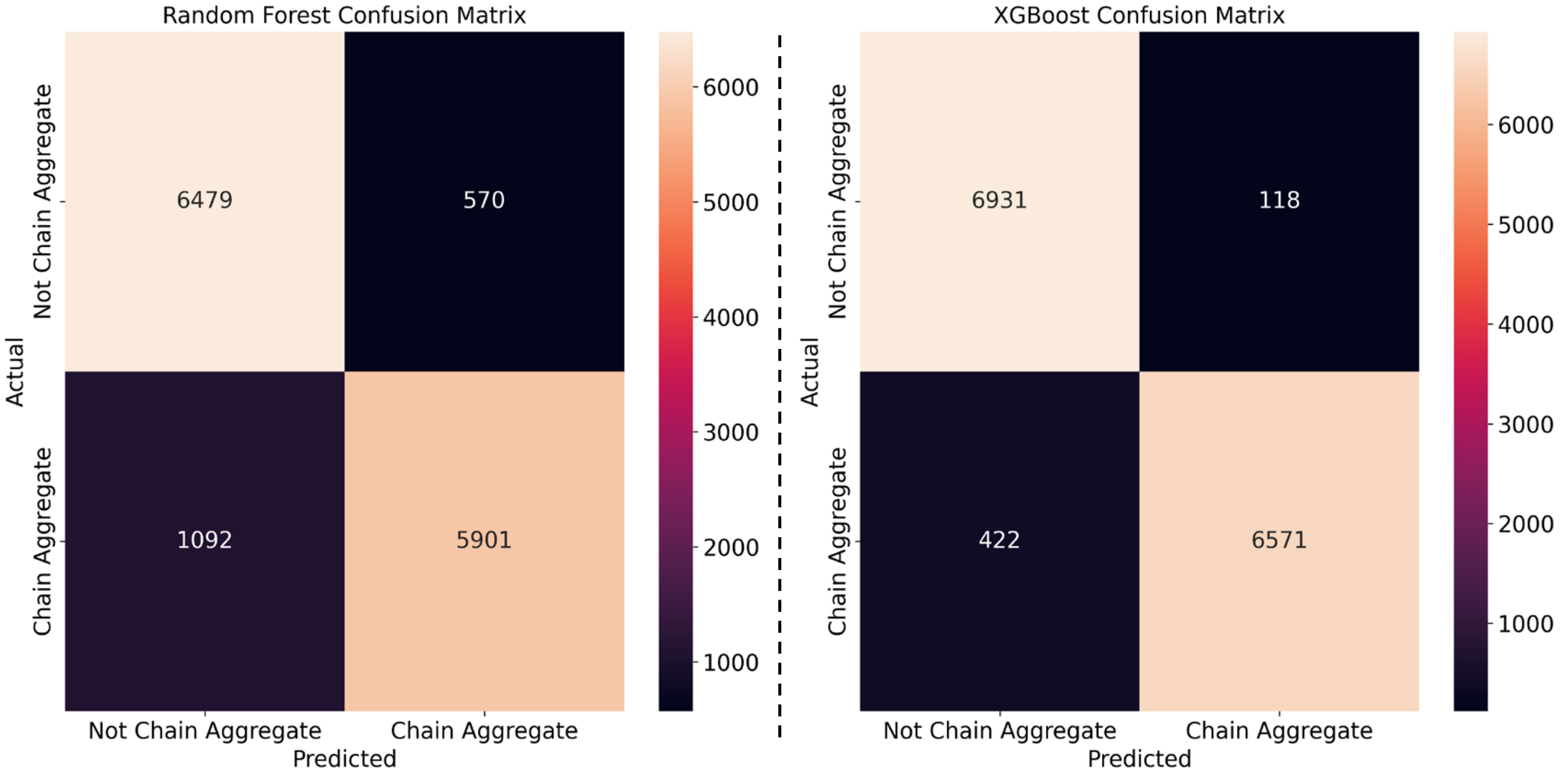
**Equations in extra slides*

- Random OverSampling (ROS) is applied to balance the dataset, improving sensitivity to rare chain aggregate occurrences.
  - Total particles after ROS performed: 46,804 (50% chains, 50% non-chains)

- 70% (36,762) used for training; 30% (14,042) used for testing and validation

1. **Random Forest Classifier**
2. **eXtreme Gradient Boosting (XGBoost) Classifier**

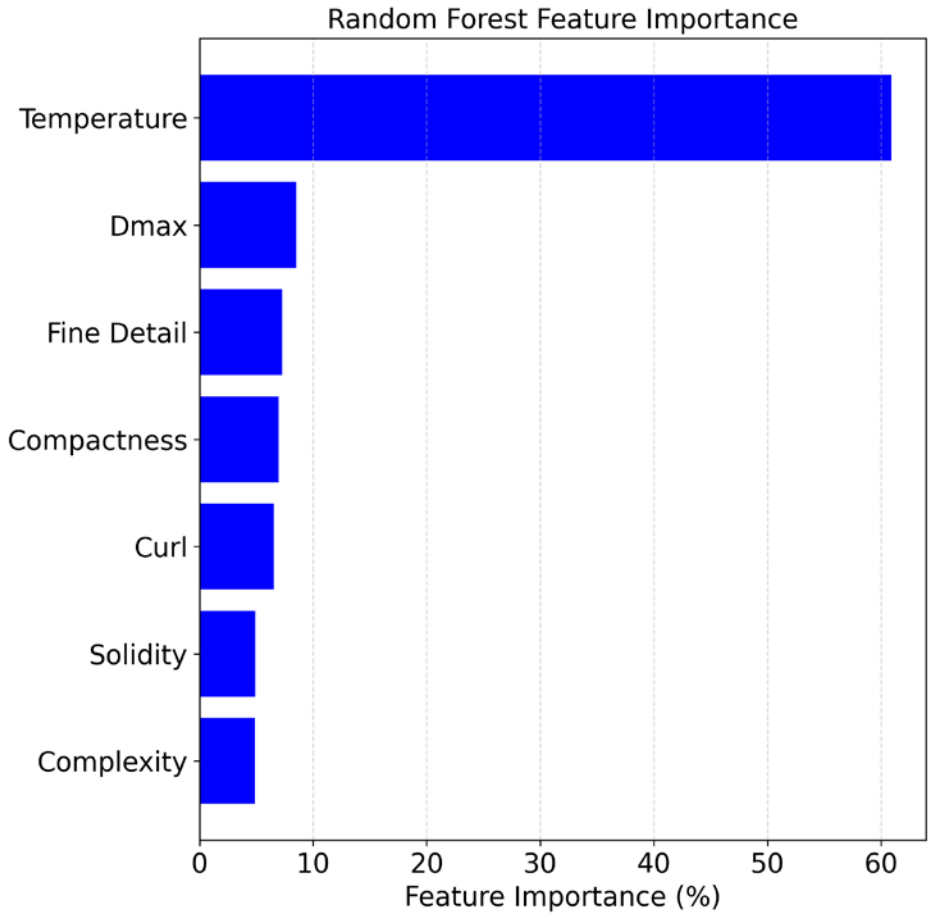
Comparison: Random Forest to eXtreme Gradient Boosting (XGBoost)

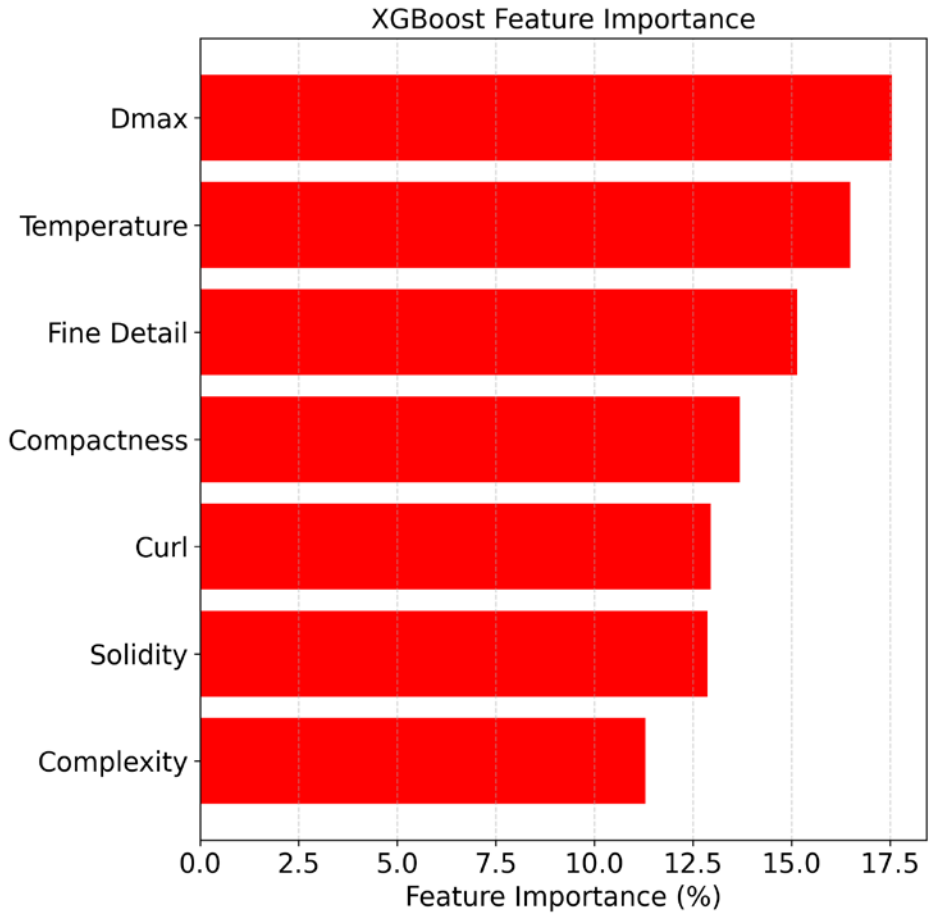


Key Takeaway:

- Less false positives (upper-right) and false negatives (lower left) using the XGBoost Classifier.

Comparison: Random Forest to eXtreme Gradient Boosting (XGBoost)

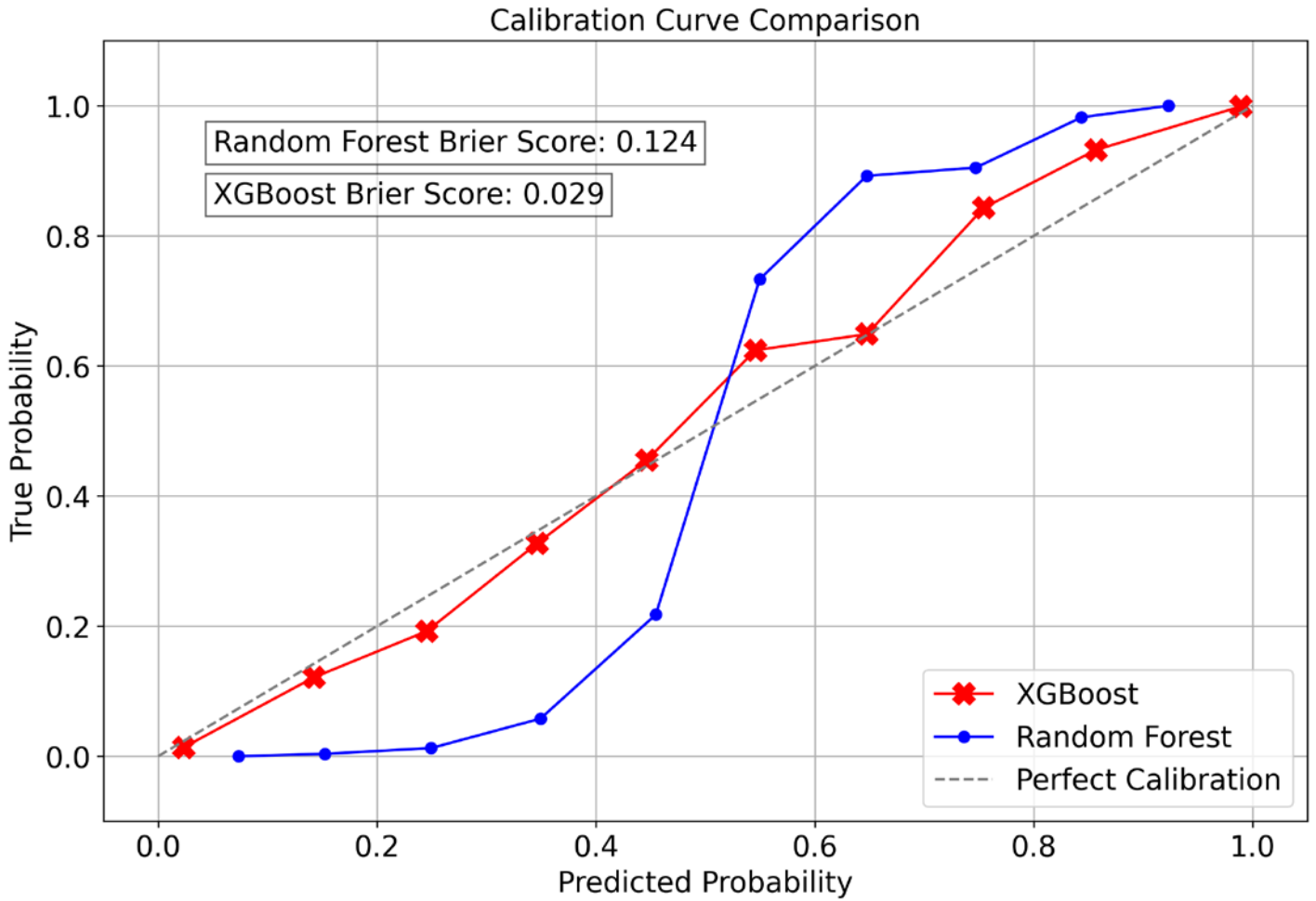




Key Takeaways:

- Random Forest weighs temperature strongly, though, accounts for most of the errors in the model.
- XGBoost recognizes and accounts for those errors, resulting in relatively balanced feature importance weights.

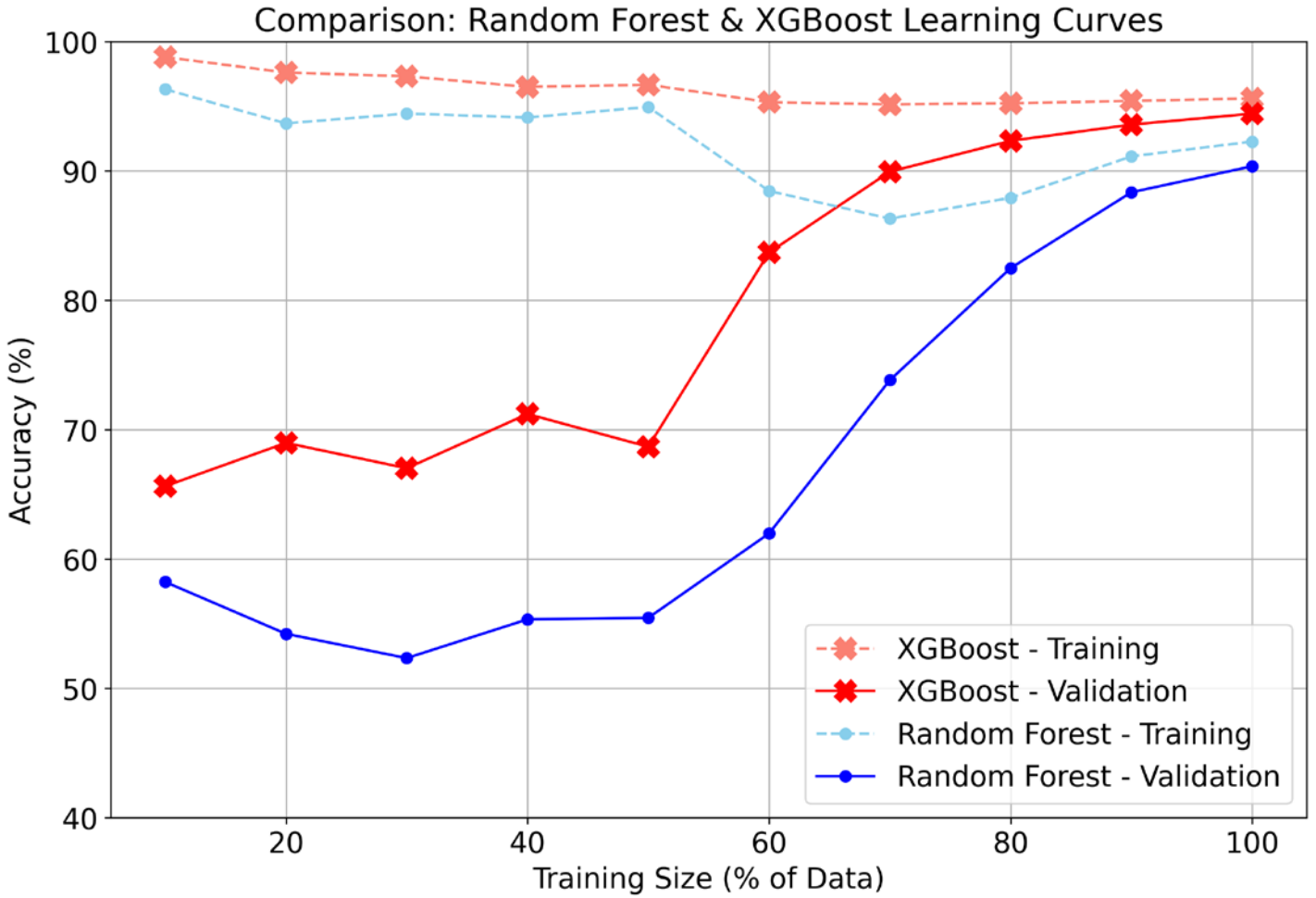
Comparison: Random Forest to eXtreme Gradient Boosting (XGBoost)



Key Takeaways:

- XGBoost demonstrates superior calibration, as reflected in its lower Brier score, meaning it provides more reliable probability estimates.

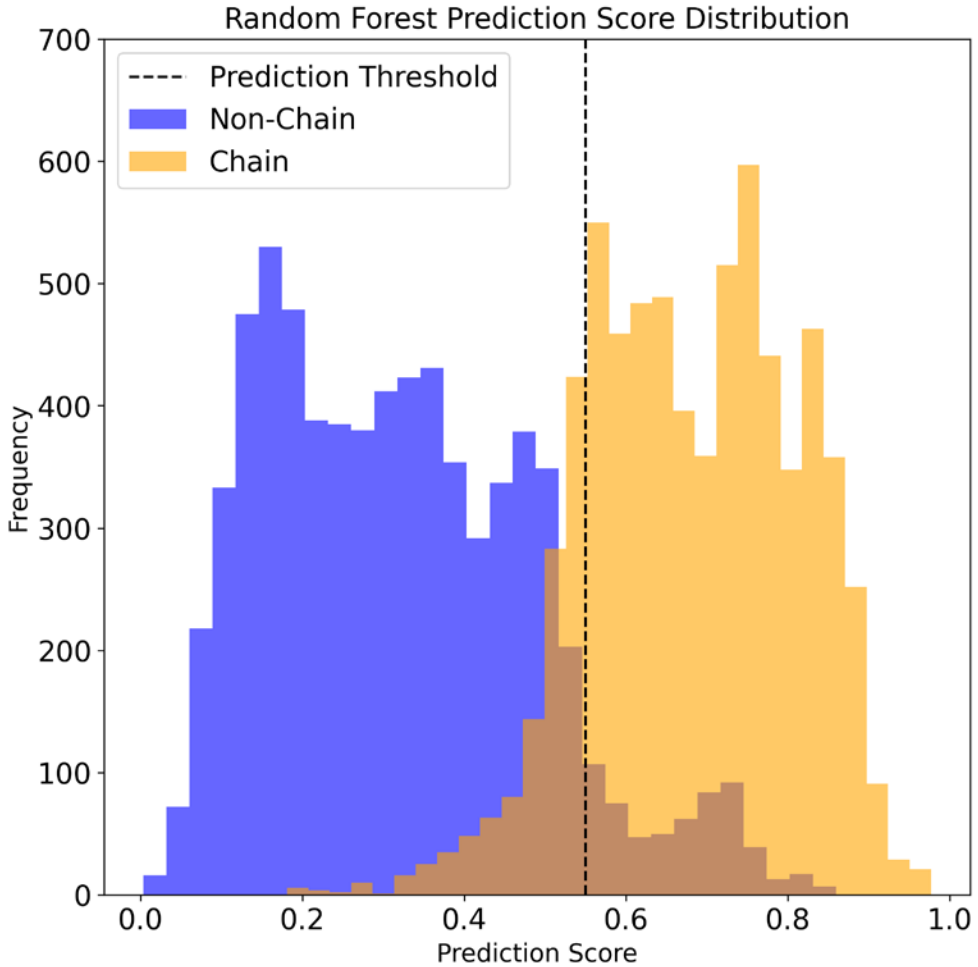
Comparison: Random Forest to eXtreme Gradient Boosting (XGBoost)

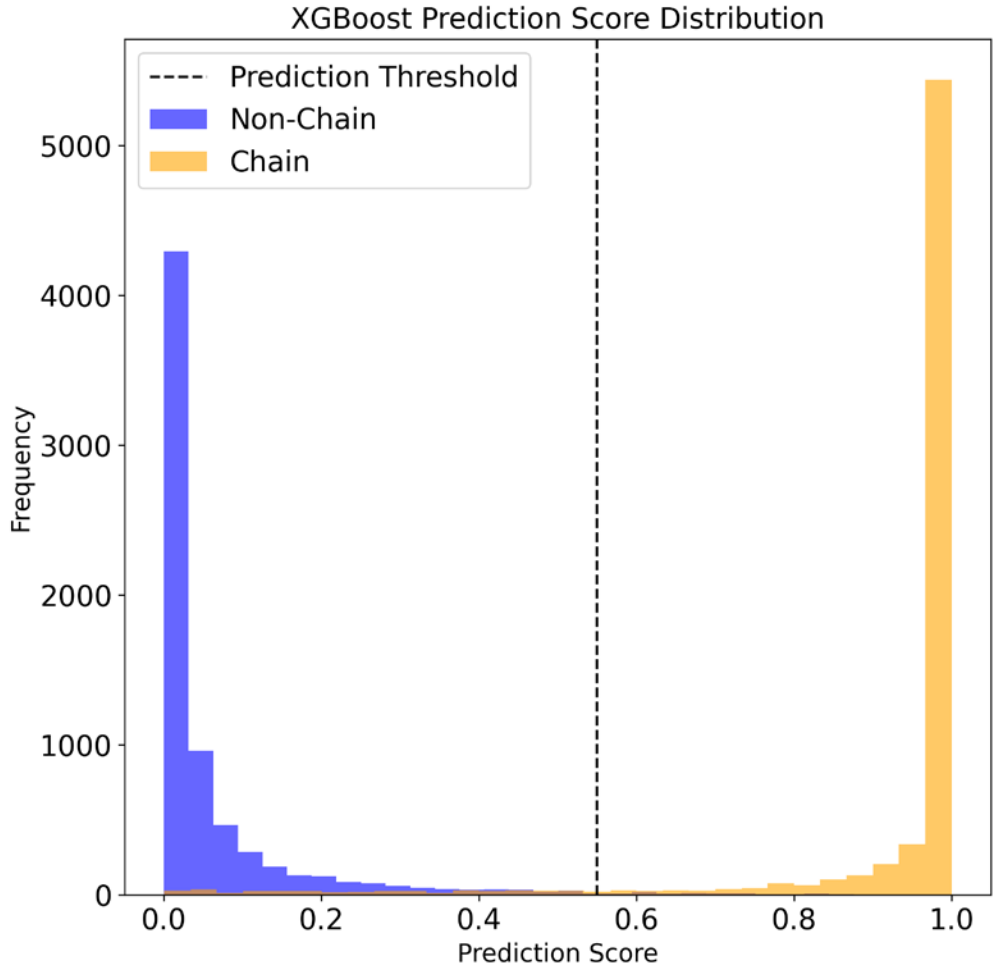


Key Takeaways:

- XGBoost consistently achieves higher accuracy with less data, indicating better learning efficiency.
- XGBoost outperforms Random Forest on validation accuracy.
- Less overfitting (gaps between training and validation curves) with XGBoost.

Comparison: Random Forest to eXtreme Gradient Boosting (XGBoost)





Key Takeaways:

- Polarized (with minimum overlap) in XGBoost predictions (more confidence).
- Random Forest is less confident likely due to errors arising from the highly weighted temperature attribute.

Comparison: Random Forest to eXtreme Gradient Boosting (XGBoost)

Class	Random Forest Classification Report			XGBoost Classification Report		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Non-Chain Aggregate	86%	92%	89%	94%	98%	96%
Chain Aggregate	**91%**	84%	88%	**98%**	94%	96%
Overall Accuracy	**88%**			**96%**		
RMSE	**0.34**			**0.17**		
CV Accuracy	**96%**			**96%**		

	Random Forest Classifier	XGBoost Classifier
Tree Building	**Parallel (bagging)**	**Sequential (boosting)**
Error Handling	**Averages results across many trees**	**Corrects previous tree errors iteratively**
Focus	Reducing variance (stabilizing predictions)	Reducing bias (minimizing error)
Speed	Slower, especially with large datasets	Optimized for faster performance
Overfitting Control	Uses tree pruning and randomness	Built-in regularization

Summary

- Both the Random Forest and XGBoost classifications can accurately discriminate between chain aggregates and non-chain aggregates.

- XGBoost excels in both predictive accuracy and calibration, making it a better tool for classifying chain aggregates from other particle habits imaged by the CPI.

- Integration of advanced sampling and preprocessing strategies further boosts model generalizability in both the Random Forest and XGBoost classifiers.

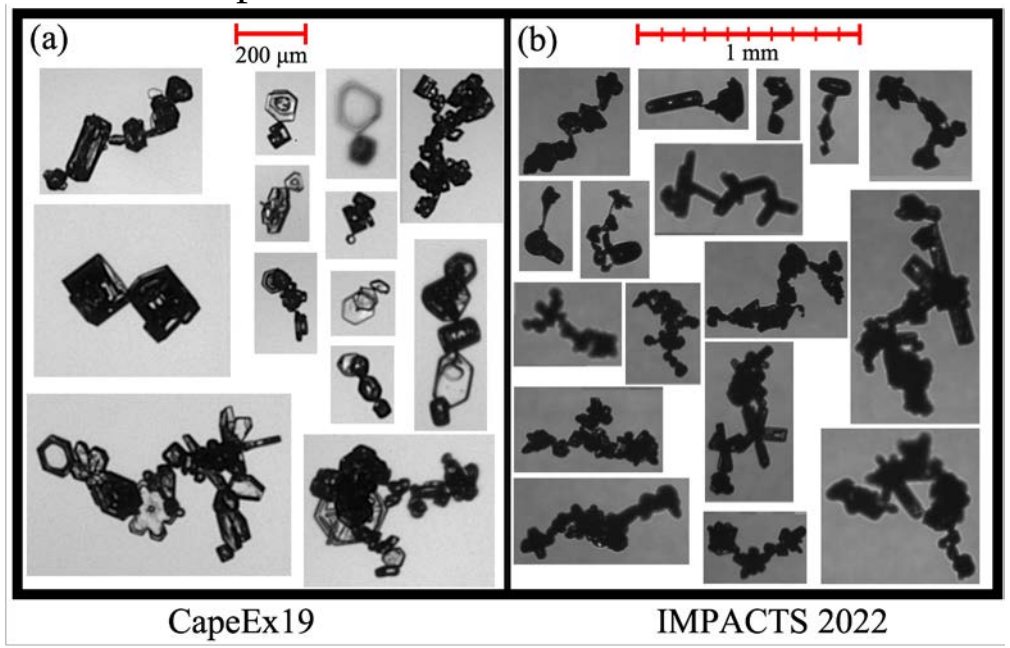
Future Work:

- Apply XGBoost (and possibly the Random Forest concurrently) across all research flights conducted during the IMPACTS field campaign.
- Contextualize chain aggregates within storm life-cycles for improved understanding in the chain aggregation process.

References & Acknowledgements

- Baran, A. J. (2009). A review of the light scattering properties of cirrus. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *110*(14), 1239–1260. https://doi.org/10.1016/j.jqsrt.2009.02.026
- Liou, K. N. (1973). Transfer of Solar Irradiance through Cirrus Cloud Layers., *J. Geophys. Res.*, *78*, 1409–1418.
- Nairy, C. M. (2022). Observations of Chain Aggregates in Florida Cirrus Cloud Anvils on 3 August 2019 during CAPEEX19 (Master's thesis), Dept. of Atmospheric Sciences, University of North Dakota, Grand Forks, North Dakota. Retrieved from https://commons.und.edu/theses/4363/
- Saunders, C. P. R., & Wahab, N. M. A. (1975). The Influence of Electric Fields on the Aggregation of Ice Crystals. *Journal of the Meteorological Society of Japan. Ser. II*, *53*(2), 121–126. https://doi.org/10.2151/jmsj1965.53.2_121
- Schmitt, C. G., and A. J. Heymsfield (2014), Observational quantification of the separation of simple and complex atmospheric ice particles, Geophys. Res. Lett., 41, 1301–1307, doi:10.1002/2013GL058781.
- Stephens, G. L., Tsay, S.-C., Stackhouse, P. W., & Flatau, P. J. (1990). The Relevance of the Microphysical and Radiative Properties of Cirrus Clouds to Climate and Climatic Feedback. *Journal of the Atmospheric Sciences*, *47*(14), 1742–1754. https://doi.org/10.1175/1520-0469(1990)047<1742:TROTMA>2.0.CO;2
- Stith, J. L., Avallone, L. M., Bansemer, A., Basarab, B., Dorsi, S. W., Fuchs, B., et al. (2014). Ice particles in the upper anvil regions of midlatitude continental thunderstorms: the case for frozen-drop aggregates. *Atmospheric Chemistry and Physics*, *14*(4), 1973–1985. https://doi.org/10.5194/acp-14-1973-2014

This research was supported by a NASA research grant to the University of North Dakota. Grant #: 80NSSC19K0328. The IMPACTS dataset is publicly available at the NASA GHRC http://dx.doi.org/10.5067/IMPACTS/DATA101. We would also like to thank Andrew Heymsfield, Stephen Nicholls, Mircea Grecu, Andrew Detwiler, & Patrick Britt for their added expertise in this work.



Thank you.

UNIVERSITY OF NORTH DAKOTA

Extra Slides – Equations and Descriptions of Particle Attributes

- Dmax = Maximum Dimension
- Area Ratio = area ratio from ellipse fit
- Solidity = $\frac{A}{Hull * (2.3 * 1.0e^{-3})^2}$
- Circularity = $\frac{4\pi * A}{P^2}$; (1 → perfect circle)
- Fine Detail Ratio = $\frac{P * (2r * 2.3 * 1.0e^{-3})}{A}$
- Complexity = $\frac{\frac{P * (1 + \sigma)}{\pi}}{\left(2 * \sqrt{A/\pi}\right)}$; (Closer to 1 -> circular, rimed; > 1 -> aggregate, less rimed)
- Compactness = $\frac{P^2}{4\pi * A}$; 1 -> circle (droplet); > 1 -> irregular
- Curl = $\frac{D_{max}}{P - \sqrt{(P)^2 - 16 * A}}$; Measures the degree to which an object is 'curled' up (0 -> circle; > 0 elongated and curly)

Extra Slides – Random Forest

- Param Grid: {n_estimators: 250, max_depth: 12, min_samples_split: 10, min_samples_leaf: 4, max_features: ‘sqrt’, criterion: ‘gini’}

- cv = StratifiedKFold(n_splits=10)

Extra Slides – XGBoost

- num_boost_round = 500
- params = {'objective': 'binary:logistic’, 'max_depth': 8, 'learning_rate': 0.1, 'colsample_bytree': 0.8, 'subsample': 0.8, 'gamma': 0.1, 'reg_alpha': 1, 'reg_lambda': 1, 'eval_metric': 'logloss', # Evaluation metric}

- # Train the model with early stopping
- evals = [(dtest, 'eval'), (dtrain, 'train')]
- bst = xgb_train(params, dtrain, num_boost_round=num_boost_round, evals=evals, early_stopping_rounds=10, verbose_eval=False)