AI4ESS Hackathon Introduction

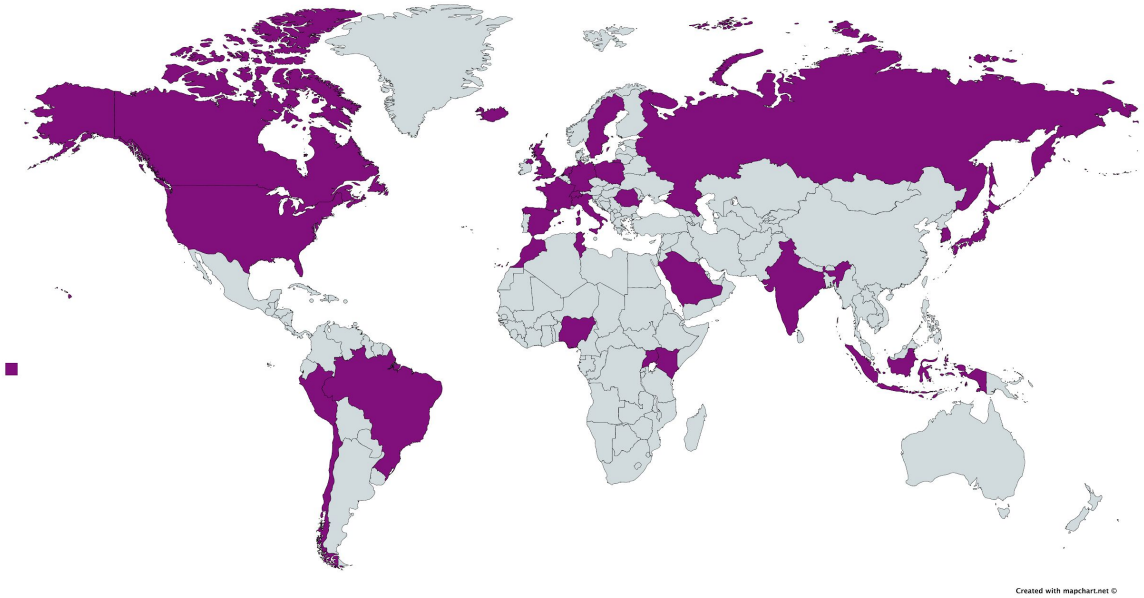
David John Gagne
NCAR

June 22, 2020

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

AI4ESS Hackathon Overview

- The goal of the AI4ESS hackathon is to allow participants to practice their machine learning skills on Earth System Science challenge problems
- Format is useful for both AI and ESS practitioners
  - ESS practitioners get to utilize AI in a domain they are more familiar with
  - AI practitioners get to learn about the special qualities of different ESS datasets

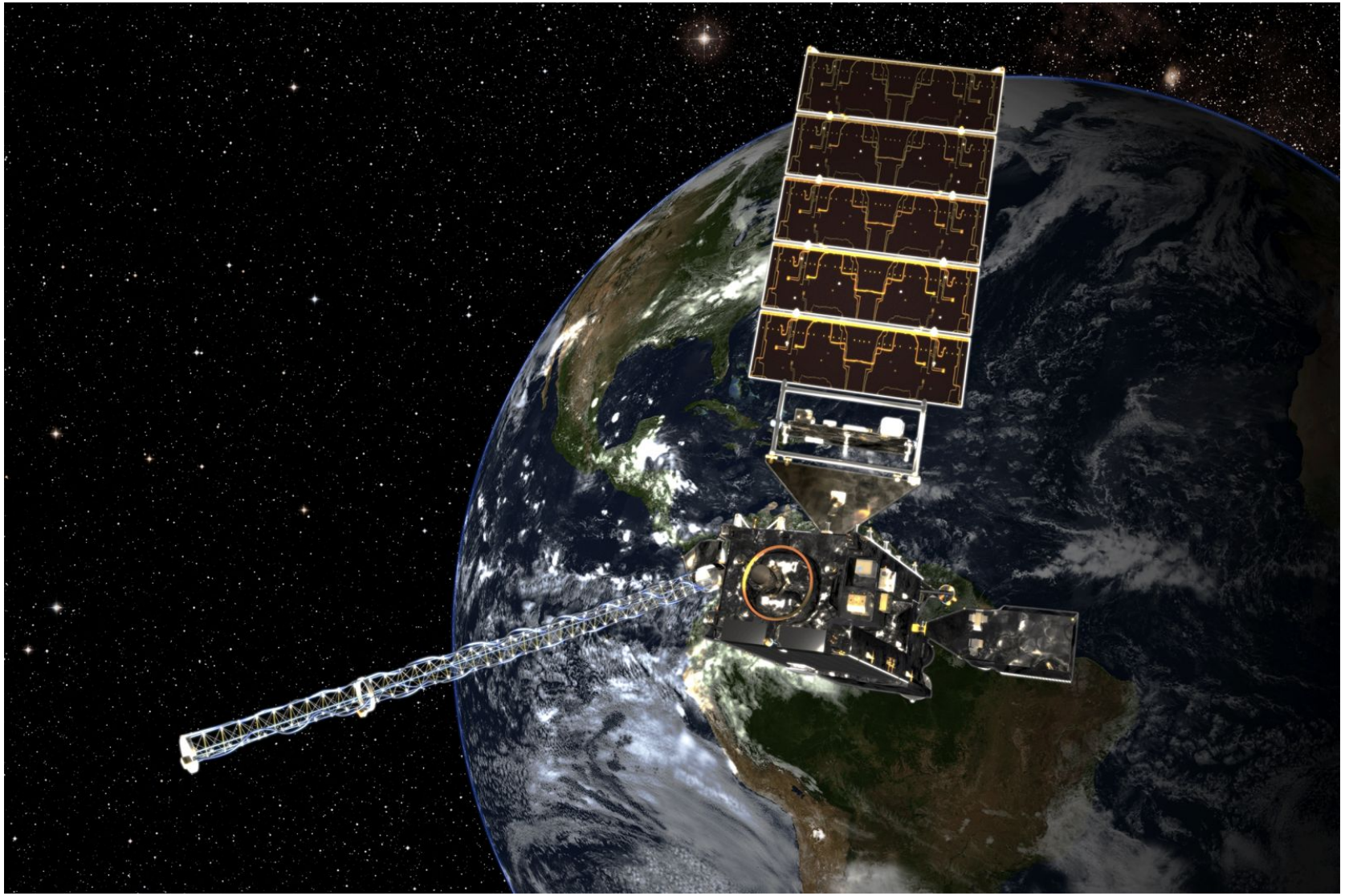


Schedule

Daily Schedule	Description
2 PM to 6 PM Mountain Time	Hackathon Shared Work Period
5 to 6 PM	Slack Open Q & A
Monday-Wednesday 6 PM Mountain Time	Submit Team Notebook
Thursday 6 PM Mountain Time	Submit 2 Google Slides for Friday Presentation

GOES Machine Learning Challenge Problem

David John Gagne, Gunther Wallach, Charlie Becker, Bill Petzke



- The Geostationary Operational Environmental Satellite 16 (GOES-16) is a weather satellite that orbits the Earth
- It can provide a hemispheric, multispectral view of cloud patterns at high space and time resolution through its Advanced Baseline Imager (ABI) camera.
- The satellite holds the Geostationary Lightning Mapper (GLM) instrument that records lightning flashes across the hemispheric view of the satellite

The Challenge

- Lightning kills roughly 30 people per year in the United States (https://www.weather.gov/safety/lightning-victims) and can have large economic impacts by disrupting outdoor work and events and by sparking fires.

- Improved short term prediction of lightning onset can help protect life and property by ensuring that people can get to safety with sufficient lead time.

- The economic impacts of lightning protection practices can also be reduced by improving the prediction of when lightning is expected to end.

Data

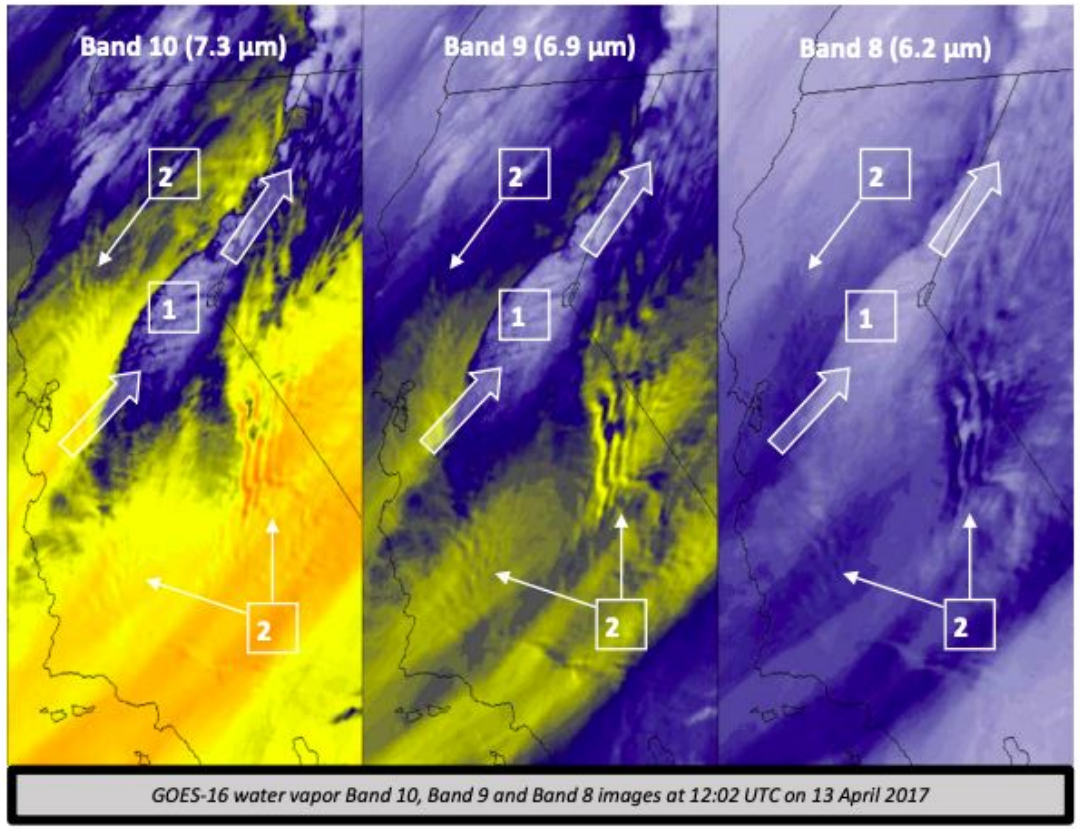
- The Advanced Baseline Imager (ABI), and the lightning counts from the Geostationary Lightning Mapper (GLM) are the two data types we care about from the GOES-16 satellite.
- The ABI camera functions with a spatial resolution of 2km, with a temporal resolution of 5 minutes.
- We select for 32 x 32 sized image patches across our domain (CONUS Midwest), at an unsampled rate of every 20 minutes from 2019-03-02 through 2019-10-01.
- We used these same spatio-temporal patches to aggregate all lightning flash counts within that patch but lagged by one hour. Total aggregated data was output at a daily interval.
- The aggregated data has about 3600 spatio-temporal patches per day, with an X and Y dimension of 32.

Image Interpretation

1 Axis of strong middle / upper tropospheric jet streak

2 Mountain waves downwind of the Coastal Ranges and the Sierra Nevada.

Depending on the topography as well as the atmospheric temperature and moisture profile, mountain waves might show up better on any of the water vapor bands. You may have to apply different enhancements as well.



GOES-16 water vapor Band 10, Band 9 and Band 8 images at 12:02 UTC on 13 April 2017

Dimensions

Dimension Name	Description	Size
Band	ABI Band Number	4 (Bands 8,9,10,14)
Patch	spatio-temporal patch	~3600 per day
X	X-plane	32
Y	Y-plane	32

Potential Input Variables

Variable Name	Units	Description
abi (Band 08)	K	Upper-level Water Vapor
abi (Band 09)	K	Mid-level Water Vapor
abi (Band 10)	K	Lower-level Water Vapor
abi (Band 14)	K	Longwave Window

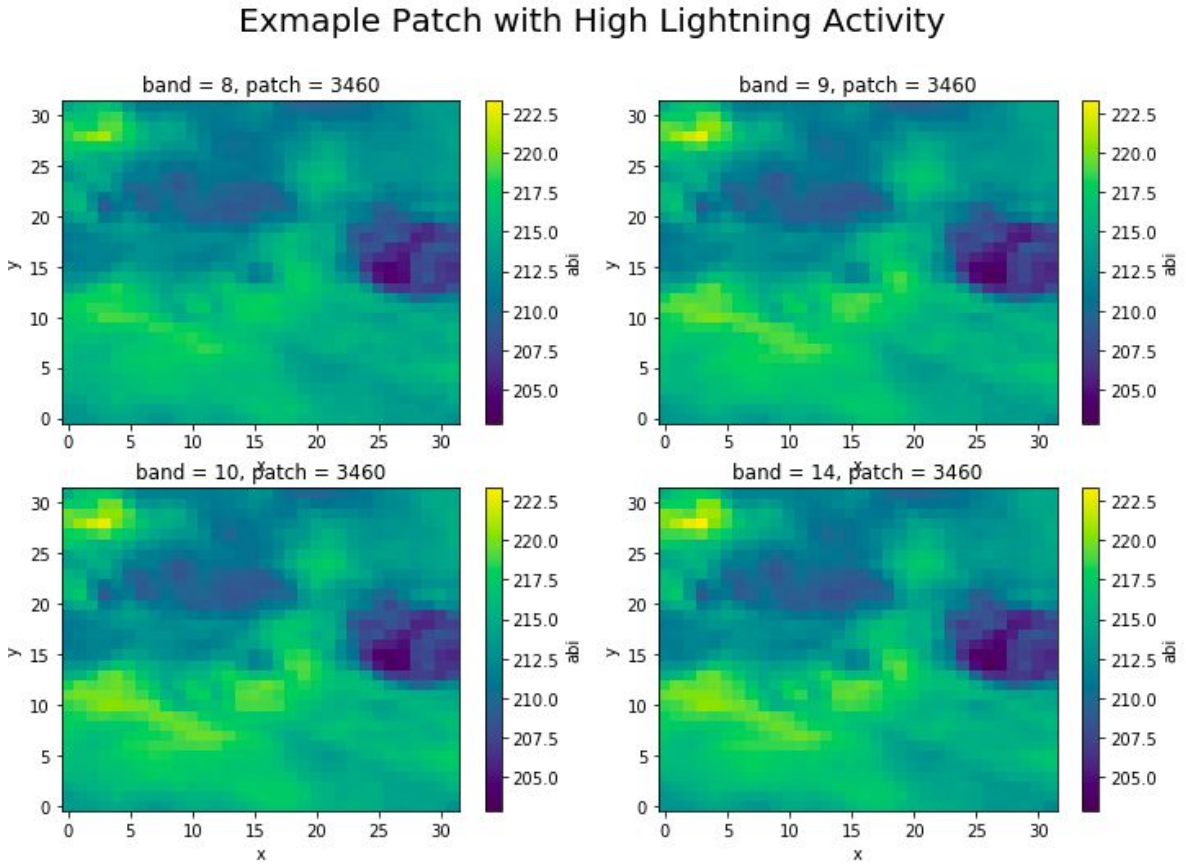
Output Variables

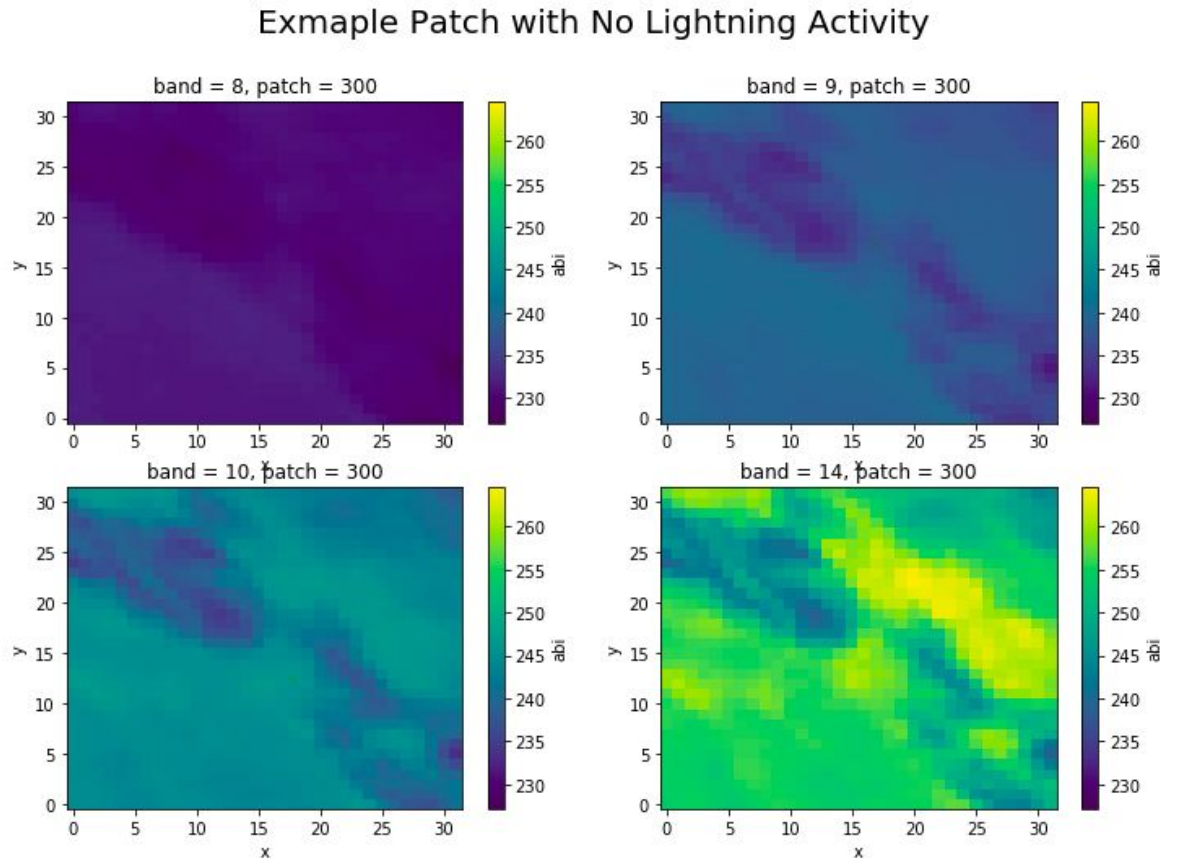
Variable Name	Units	Description
flash_counts	-	Lightning strike count

Example Patch Data

Convolutional Neural Networks (CNNs) work well to capture spatial structural differences.

Residual Networks can be used to increase the effectiveness of the depth of the NN.





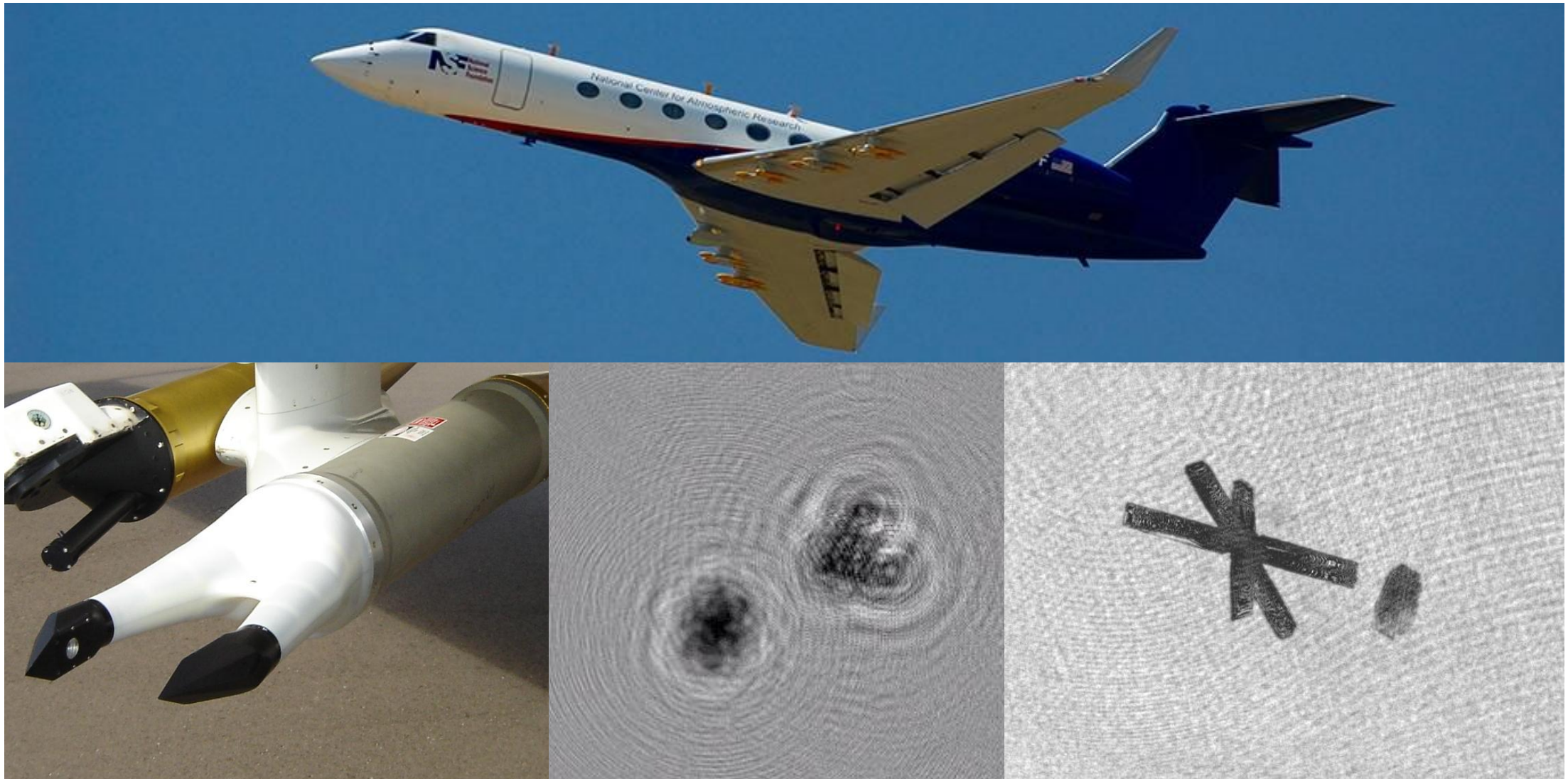
Machine Learning Approaches

- Recently deployed observational systems combined with advances in machine learning have the potential to improve the short-term prediction of lightning by associating broader scale weather patterns with the future occurrence of lightning in that area.

- Combining information from the GOES-16 Satellite in the form of the water vapor bands from the Advanced Baseline Imager (ABI) and lightning counts from the Geostationary Lightning Mapper (GLM), we are able to make machine learning predictions using a Residual Networks (ResNet) architecture.

HOLODEC Machine Learning Challenge Problem

Matt Hayman, Aaron Bansemer, David John Gagne, Gabrielle Gantos, Gunther Wallach, Natasha Flyer

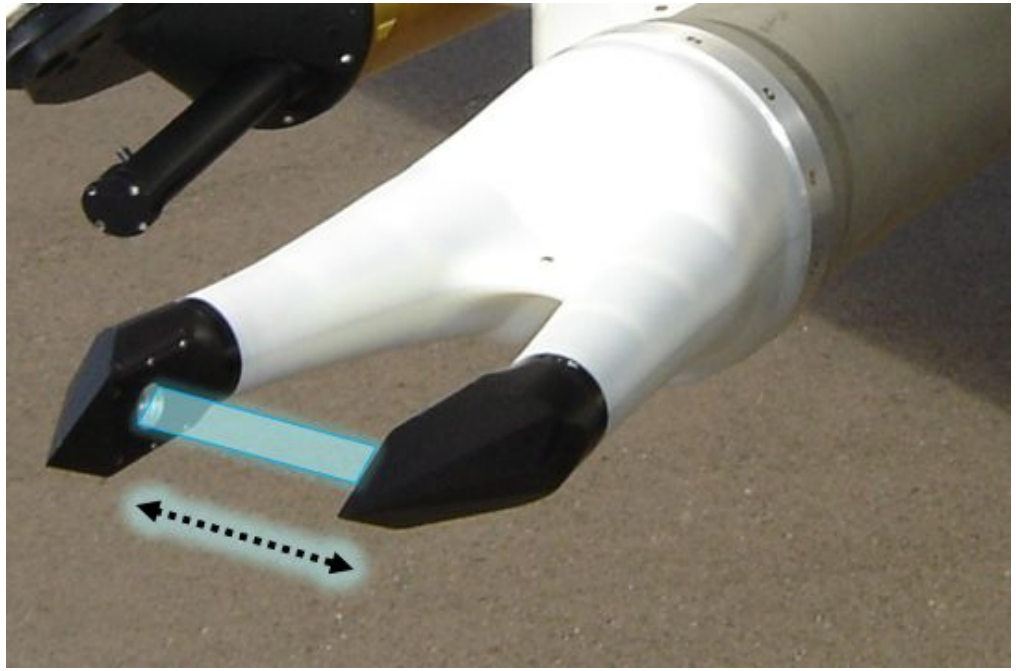


Holographic Detector for Clouds (HOLODEC)

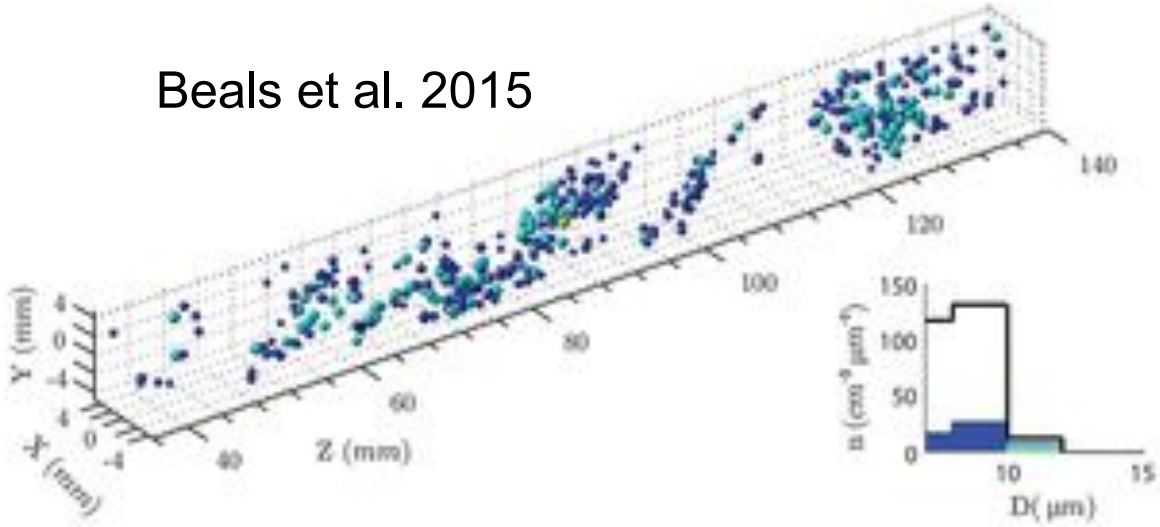
- Airborne instrument that measures liquid droplets and ice crystals in natural clouds
- Droplets sizes, concentrations, and their relative positions influence the formation of drizzle and rain
- Ice crystal size, shape, and concentrations control radiative properties of cirrus
- Used in studies of cloud dynamics, precipitation formation, and radiative transfer

Advantages

- One of the only instruments that can reliably measure mixed-phase (ice crystals and liquid droplet) clouds
- Simultaneously measure all particles in the volume between the arms (13 cm^3) in a single picture
- Allows retrieval of the 3-D position of every cloud particle

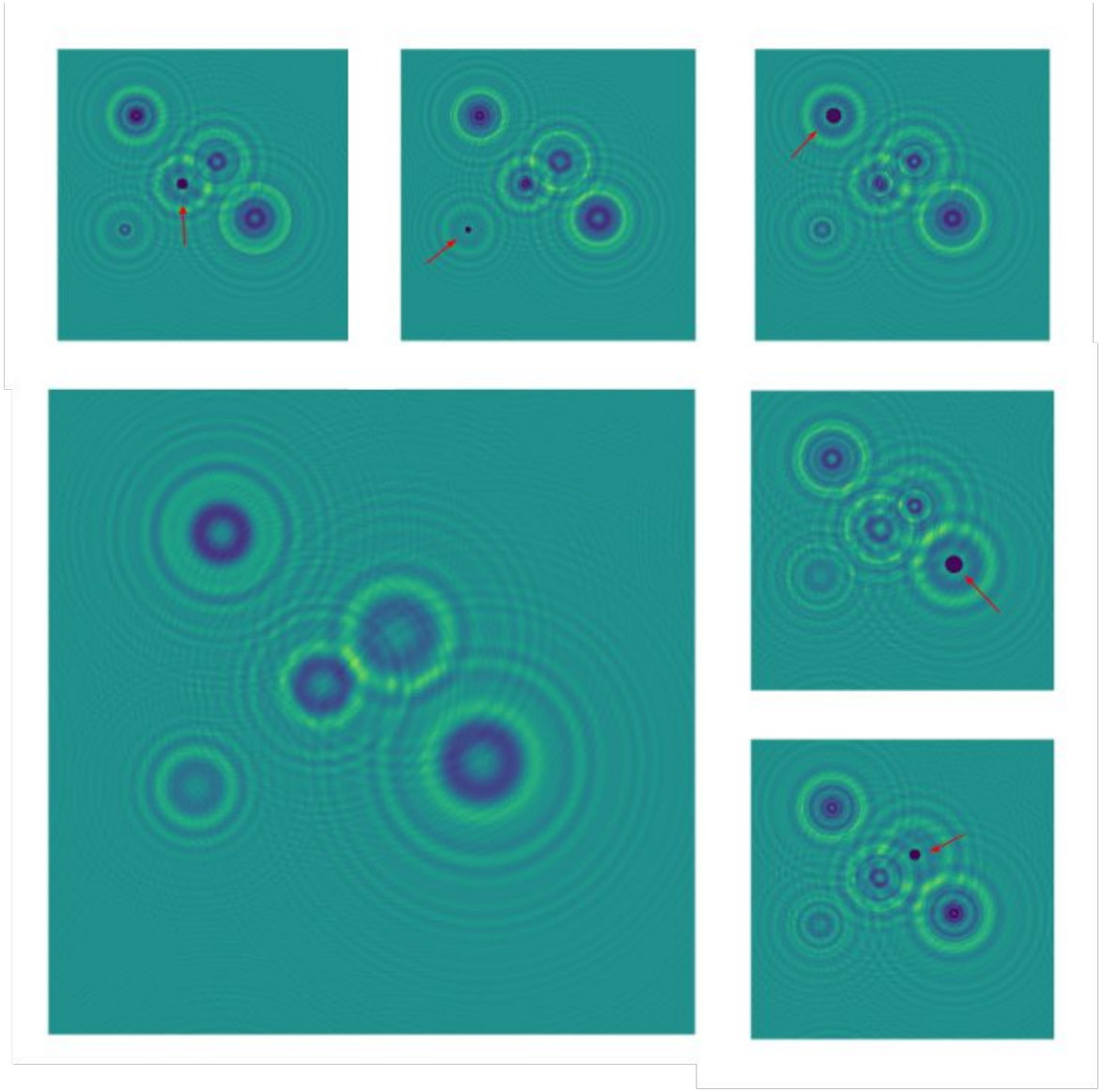


Beals et al. 2015



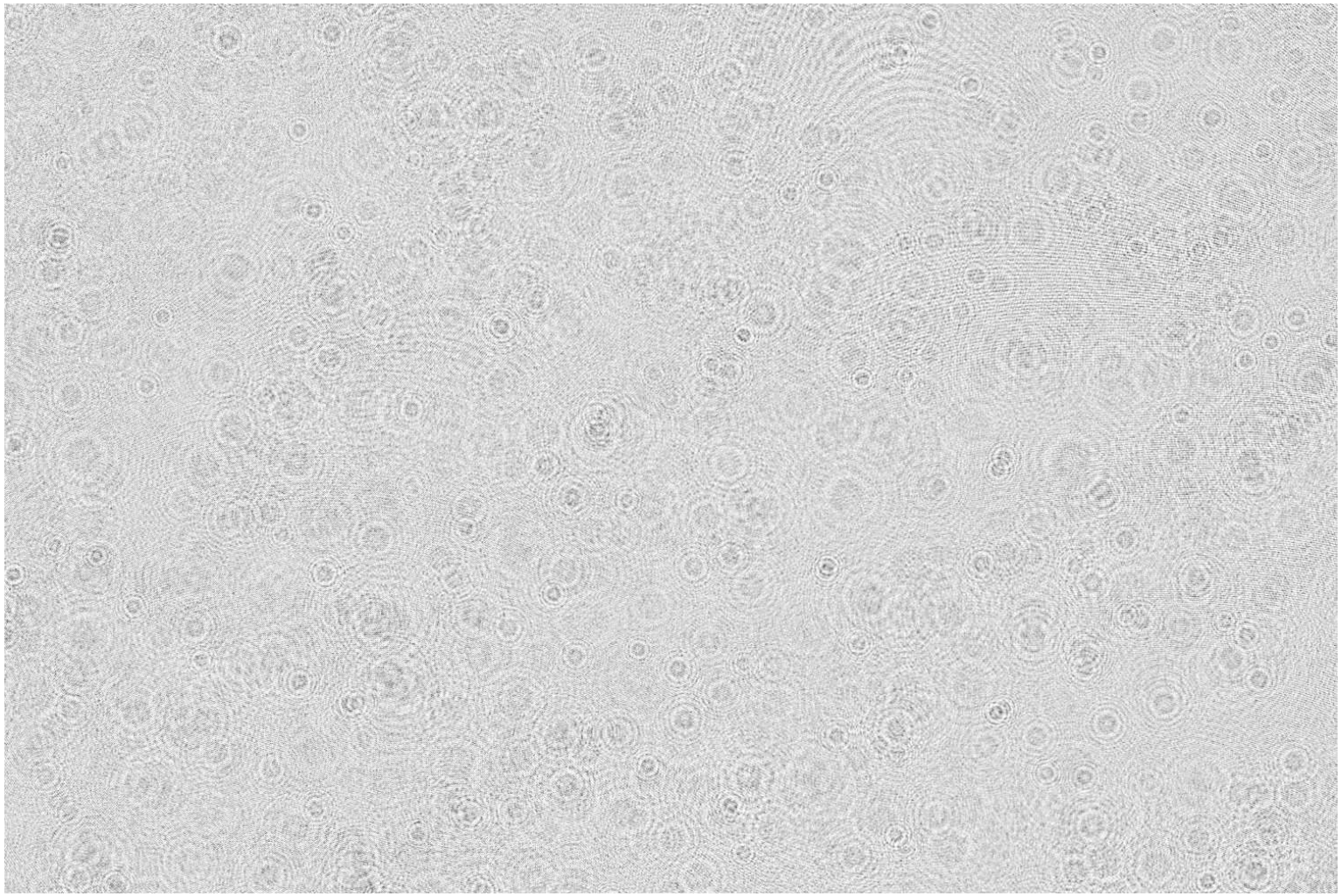
Inline Holographic Imaging

- Microscopic imaging using laser light source, the hologram is simply a 2-D picture
- Particles are intentionally unfocused
- Refocusing performed in software
- Finding best focus returns the 3-D position of the particle (x, y, z) as well as the size (d) and shape
- Example hologram here shows 5 unfocused particle (large picture), and then the refocused image at the z-position of each particle



Instrument Challenges

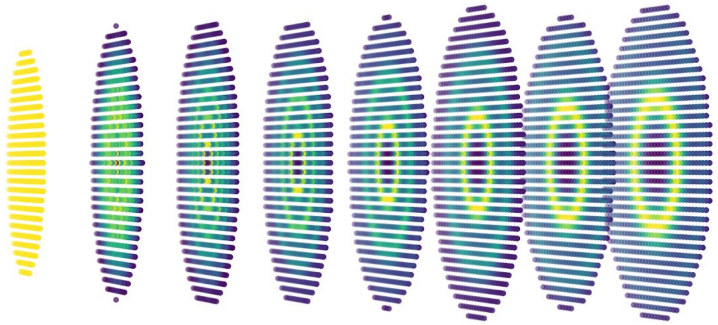
- A single hologram may contain 1000+ particles
- Traditional refocusing is performed 1000 times for each image which searching for particles
- Computationally expensive and labor intensive, up to 2 million core hours per project
- Processing is primary bottleneck in improving probe performance
- Can machine learning perform better?



Hologram in Pacific cumulus, 2015

Datasets

- Synthetic data using simplified holograms
- Circular droplets only
- netCDF format
- Training data files consist of
  - Synthetic grayscale hologram images
  - X-position of each particle
  - Y-position of each particle
  - Z-position (between the arms) of each particle
  - Diameter of each particle
  - Reference variable to link particles to holograms

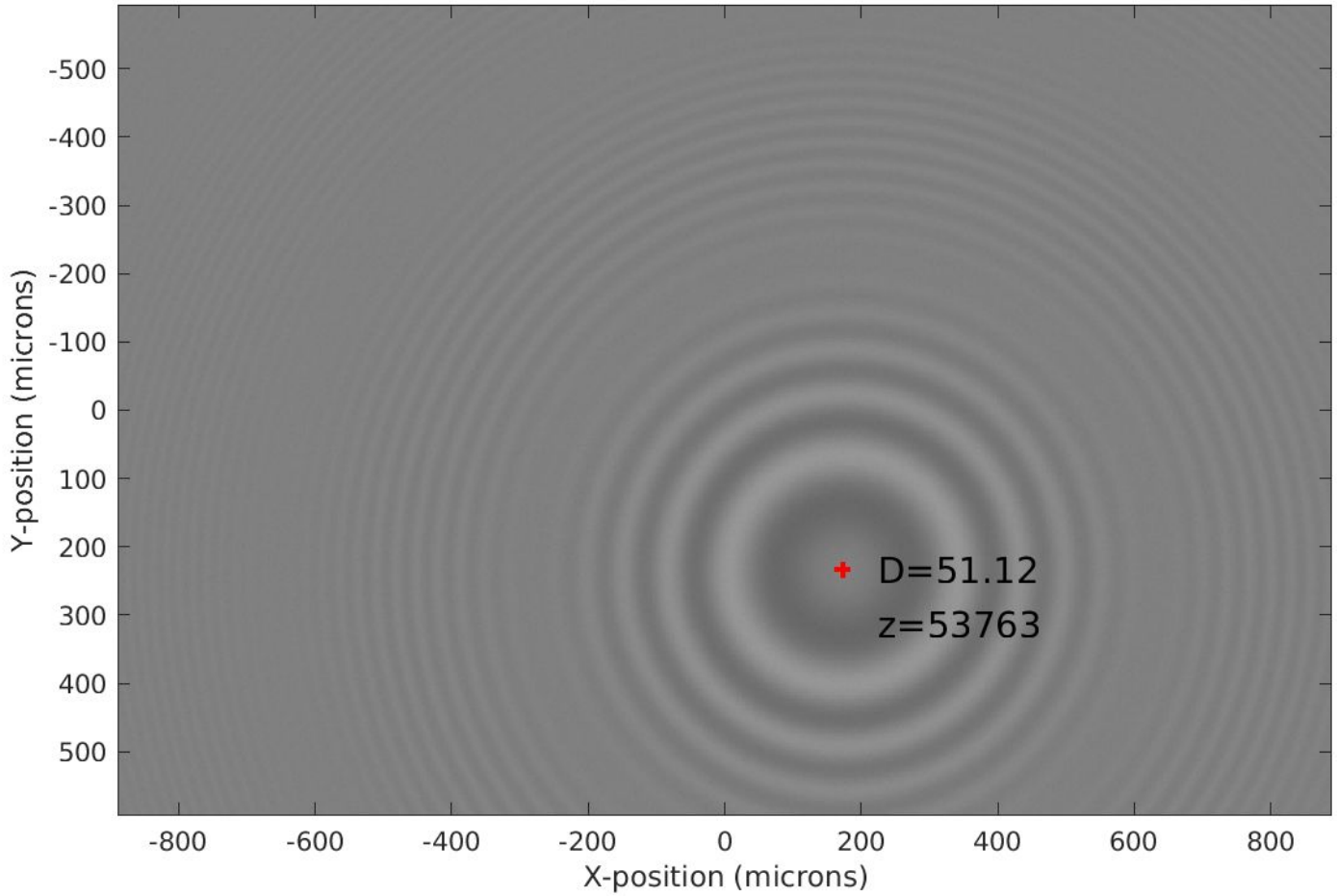


Objective: Can *x*, *y*, *z*, and *D* for each particle be predicted based on the hologram image?

Dataset 1: Single-Particle

Training:
nHolograms = 50,000
nParticles = 50,000

Test and Validation:
nHolograms = 10,000
nParticles = 10,000

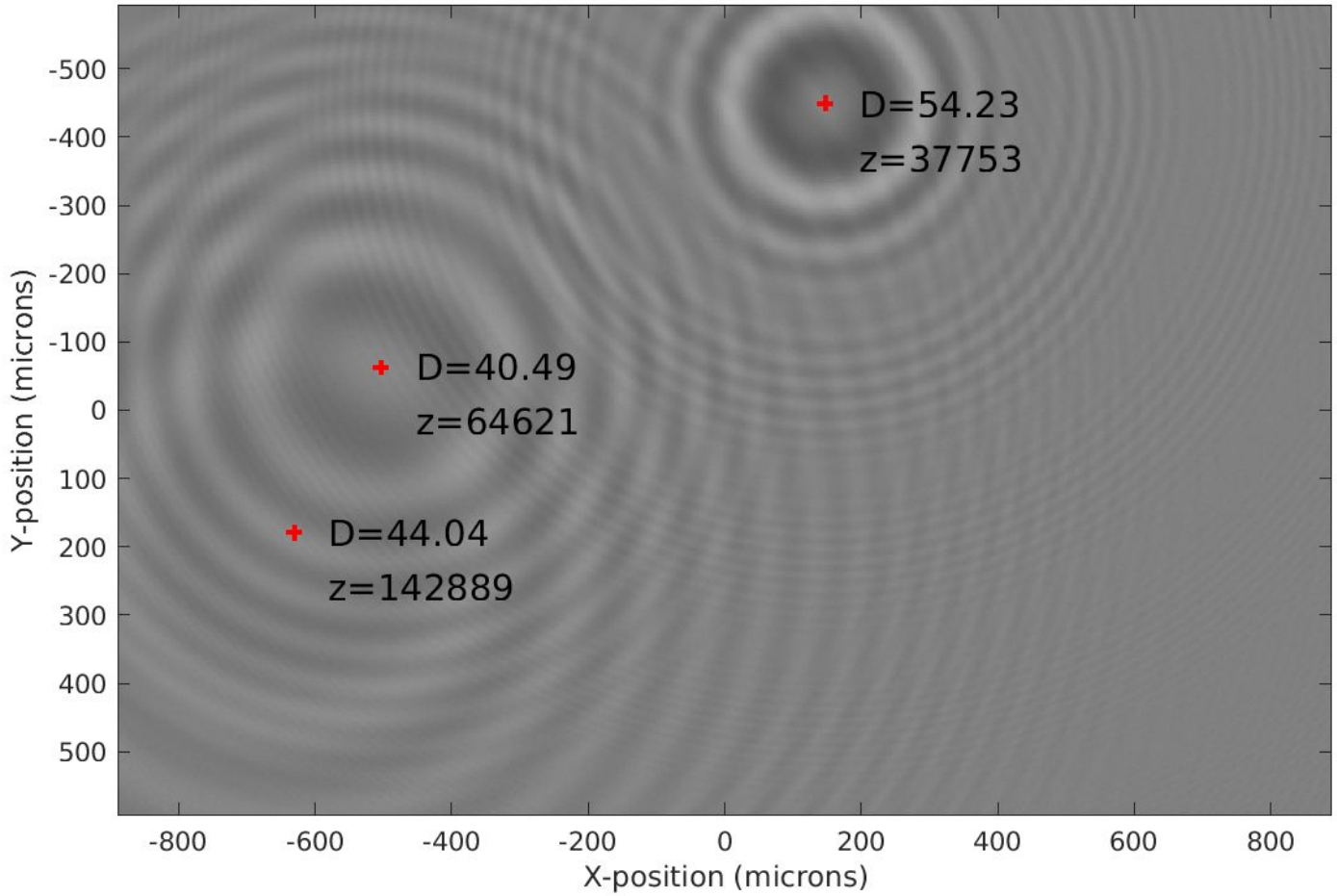


Example single-particle synthetic hologram with x, y, z, and D indicated

Dataset 2: Three-Particle

Training:
nHolograms = 15,000
nParticles = 45,000

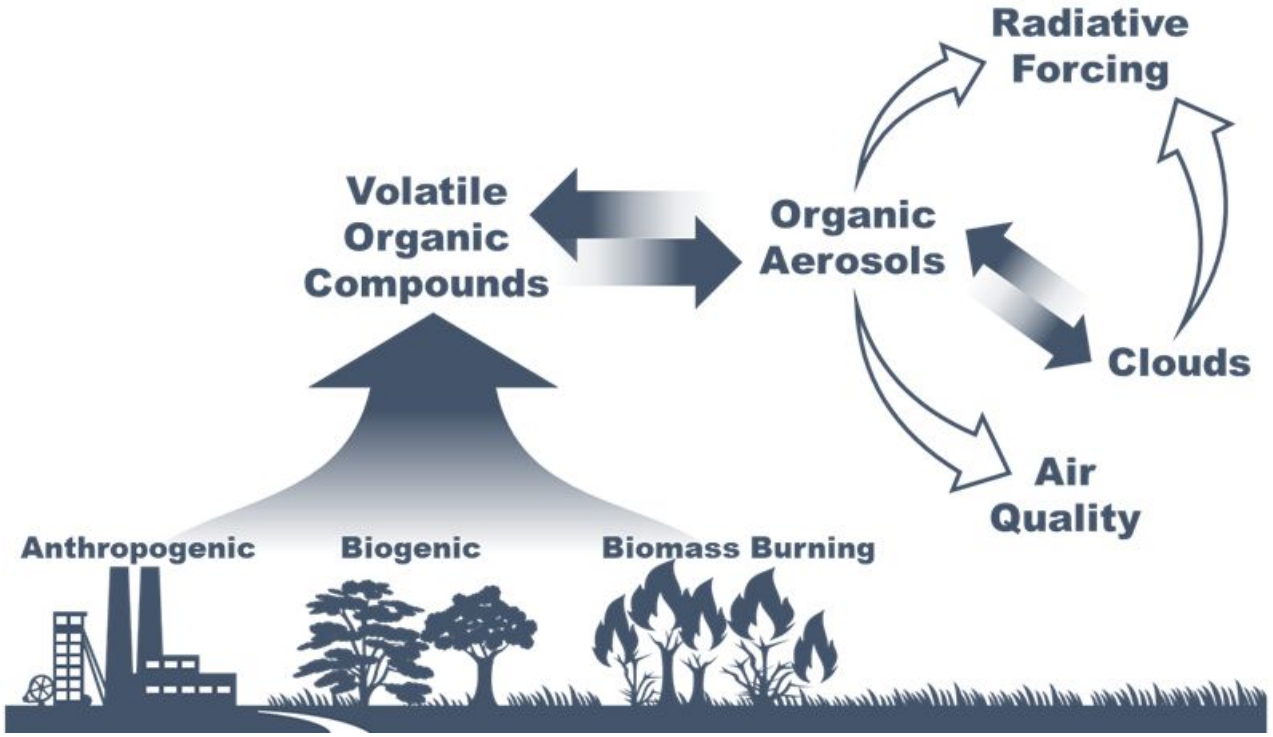
Test and Validation:
nHolograms = 5,000
nParticles = 15,000



Example three-particle synthetic hologram with x, y, z, and D indicated

GECKO-A Machine Learning Challenge Problem

Siyuan Wang, Alma Hodzic, David John Gagne, Keely Lawrence, Charlie Becker, Natasha Flyer



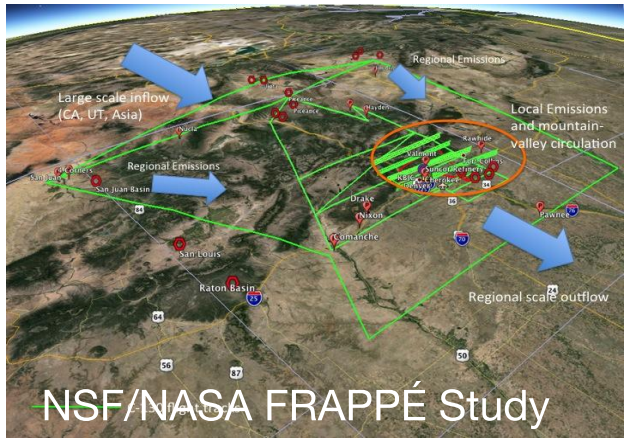
- Natural and anthropogenic sources emit a large number of volatile organic compounds (VOCs).
- VOCs gases undergo complicated chemical reactions and physical processes in the atmosphere, forming organic aerosols.
- ~100 emitted gases but their photochemical oxidation in the atmosphere leads to hundreds of thousands of volatile products that can condense to form organic aerosols

- Organic aerosols have significant direct and indirect radiation effect.
- VOCs and organic aerosols also affect air quality and human health.
- To evaluate the broad impacts of VOCs on air quality, human health, and the climate system, we need to understand the sources and fates of these compounds

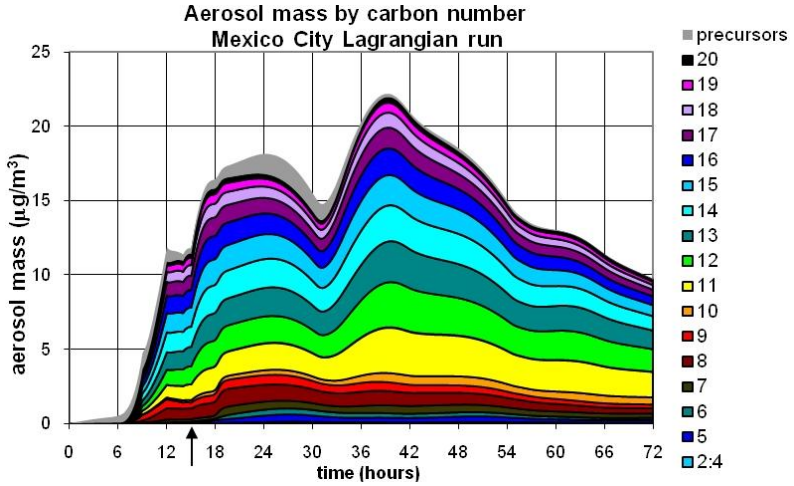
Chemical Mechanisms: Center Role in Chemistry Forecast

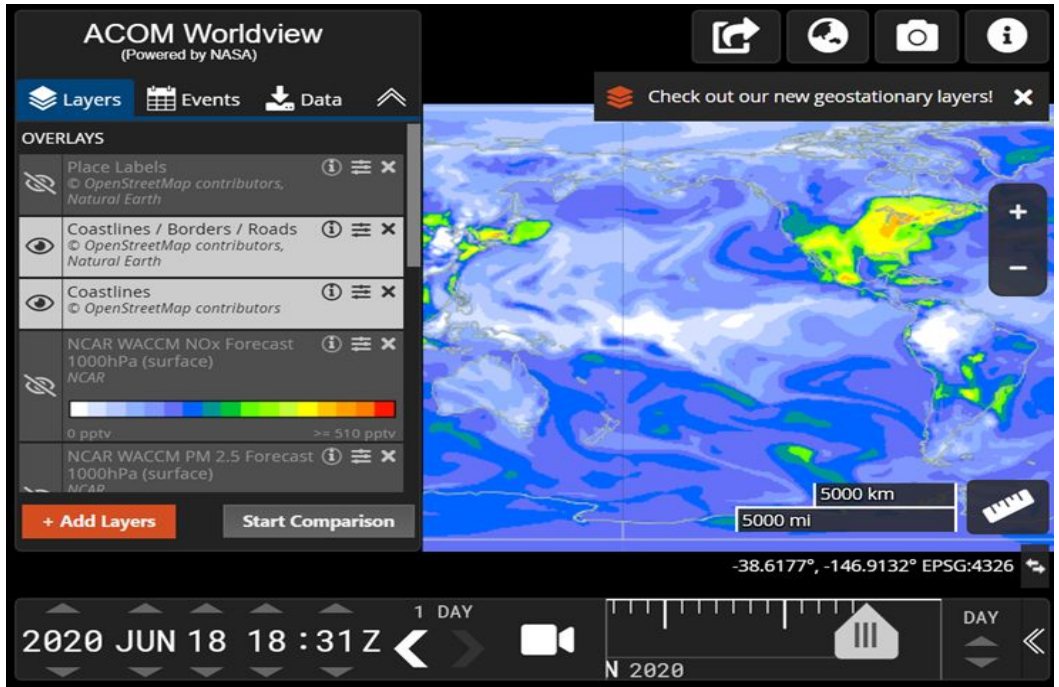
Laboratory experiments

Field Campaigns



Chemical Mechanism

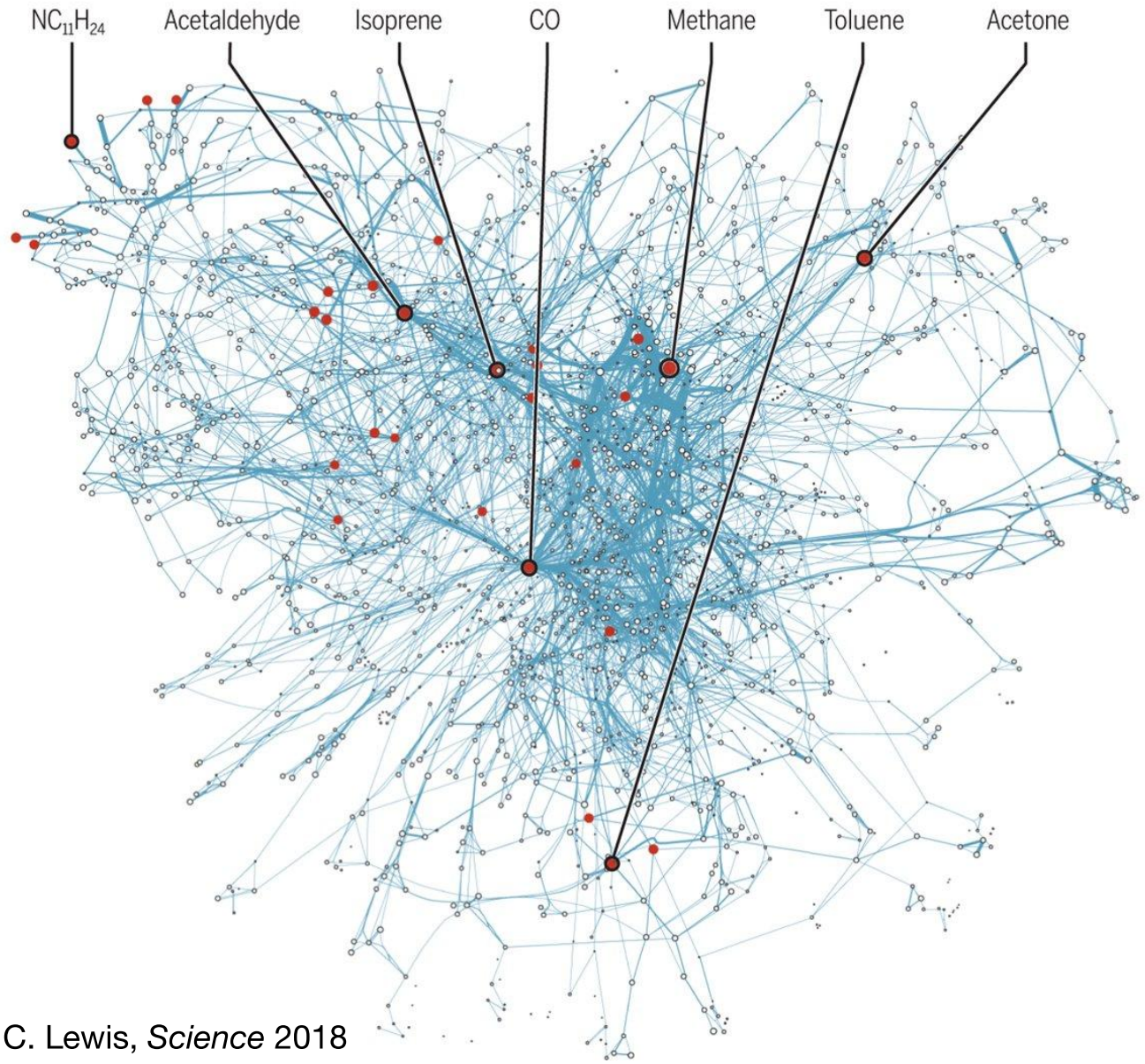




Air Quality & Chemistry-Climate Forecast

dC/dt = Emissions
+ Deposition
+ Transport
+ Chemical Production
+ Chemical Removal

Chemical Mechanisms: Very Complicated!

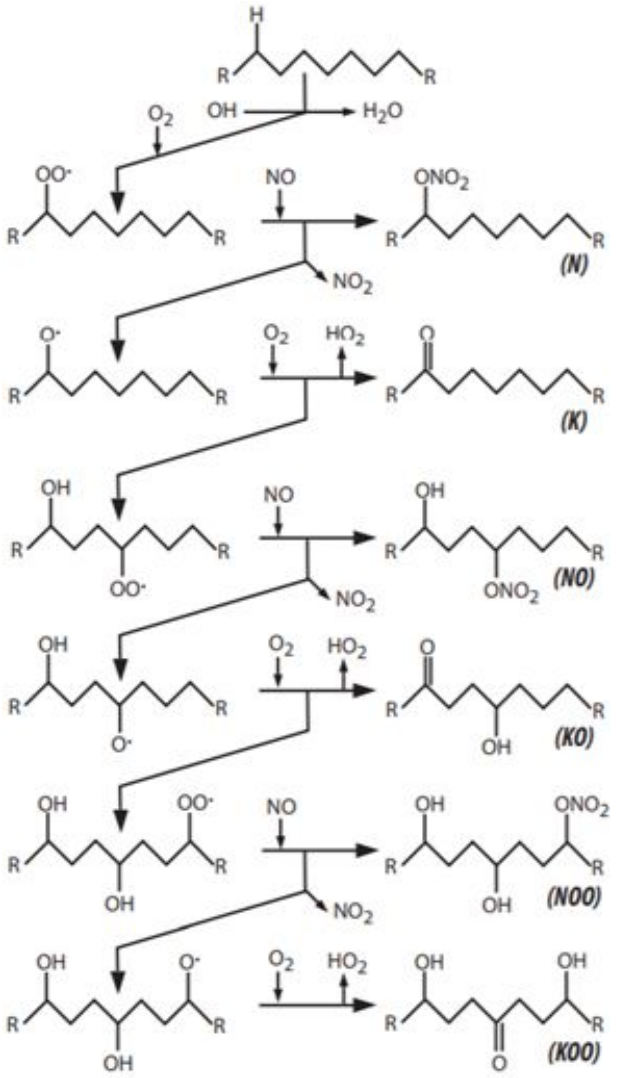


A. C. Lewis, *Science* 2018

- Left figure: cool visual illustration of the Leeds Master Chemical Mechanism (MCM), a near-explicit gas-phase chemical mechanism, with +16,000 chemical reactions
- Although complicated, MCM is constructed based on quite simple framework.
- Yet, most* air quality models and chemistry-climate models cannot afford to run MCM!

* it's not impossible! It's just so expensive that mostly it's not practical and meaningful at all.

GECKO-A: Based on First Principles

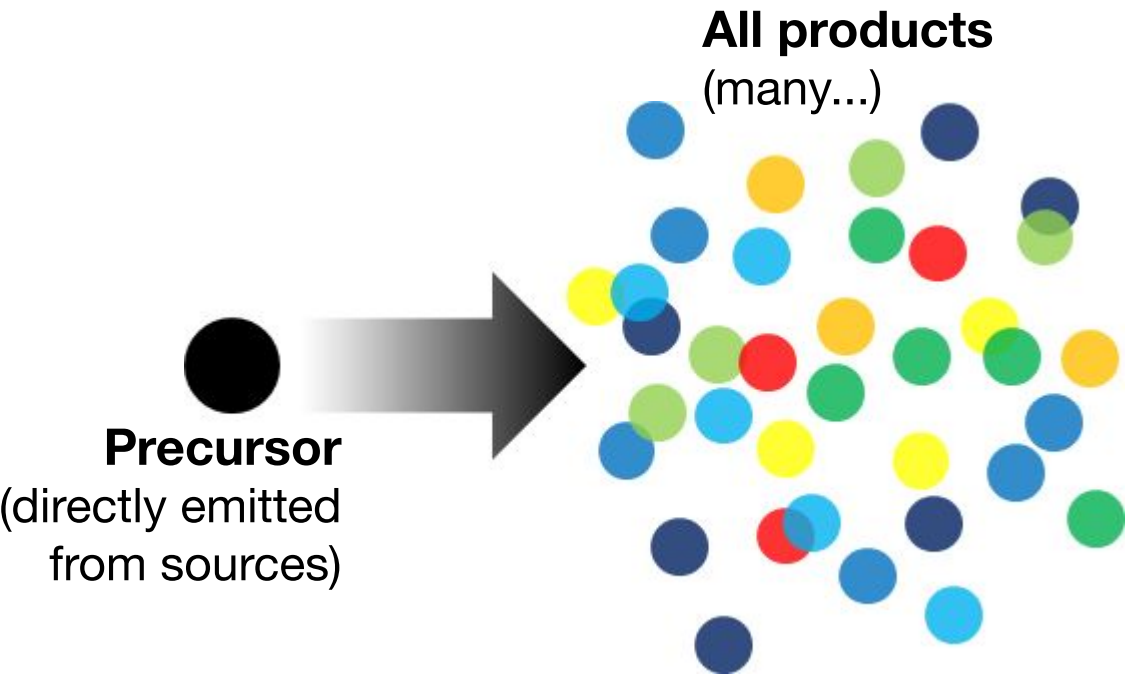


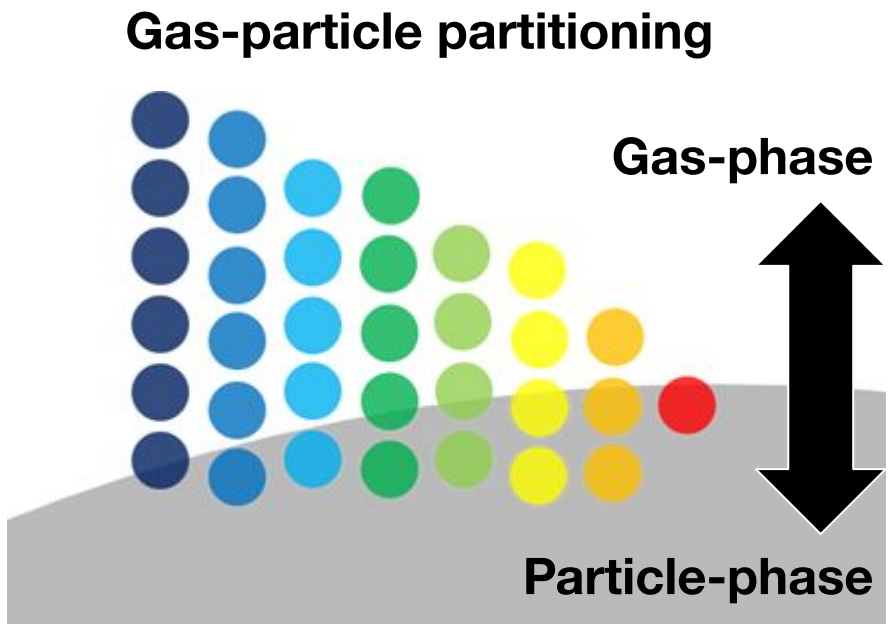
Aumont et al., *Faraday Discuss.* 2013

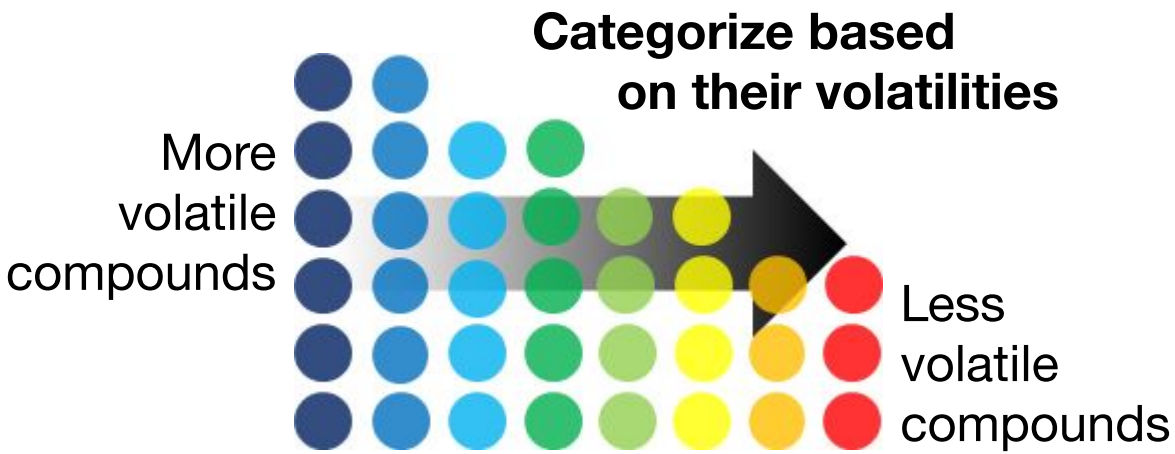
- **GECKO-A: G**enerator of **E**xplicit **C**hemistry and **K**inetics of **O**rganics in the **A**tmosphere
- We tell GECKO-A how atoms, bonds, and functional groups in molecules/radicals behave, then GECKO-A will predict what reactions it may undergo.

- GECKO-A generated chemical mechanisms are extremely large and complicated. For instance -
- a-pinene is a common compound emitted from trees. Its oxidation mechanism involves over 400,000 compounds and over 2,000,000 chemical reactions!
- GECKO-A reveals unique information that none other mechanisms can offer.
- No 3-dimensional models can afford to run GECKO-A in the foreseeable future.

Formation of Organic Aerosols



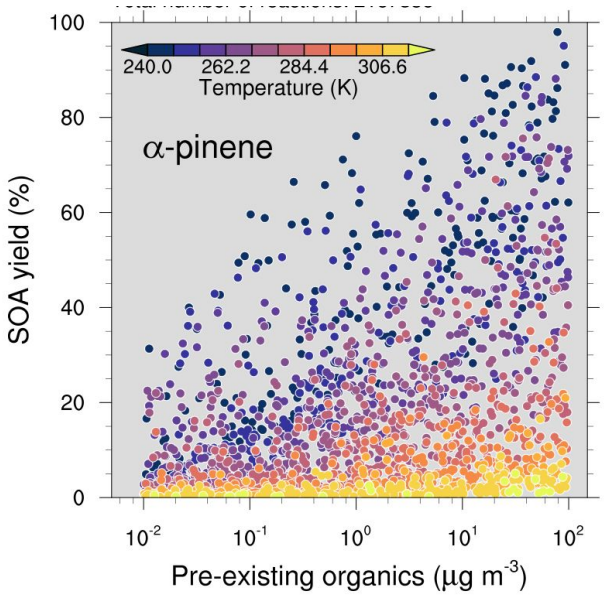




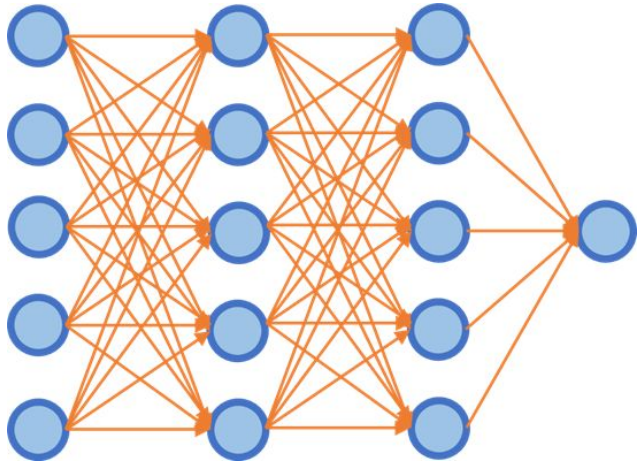
- The products will partition between the gas-phase and particles-phase.
- The partitioning is determined by the property of the molecule (e.g., volatility) and the environmental conditions (e.g., temperature)
- Organic aerosols (OA): very different properties & environmental impacts

GECKO-A Challenge: Build An Emulator For 3-D Models?

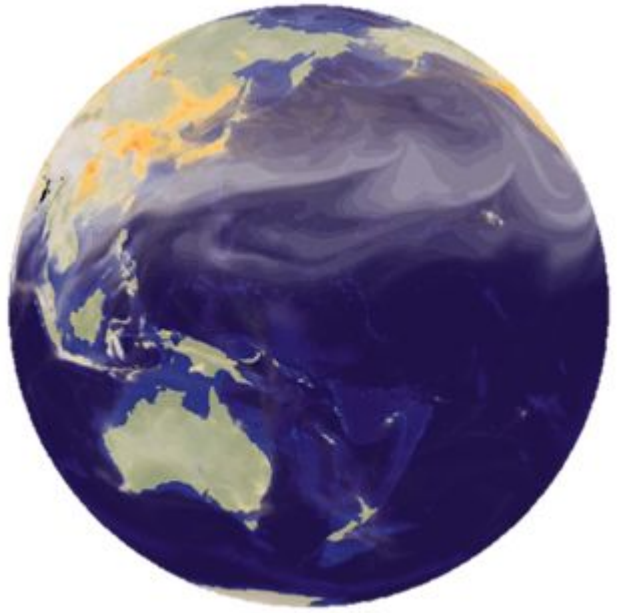
GECKO-A Training Library



Machine-Learning Emulator

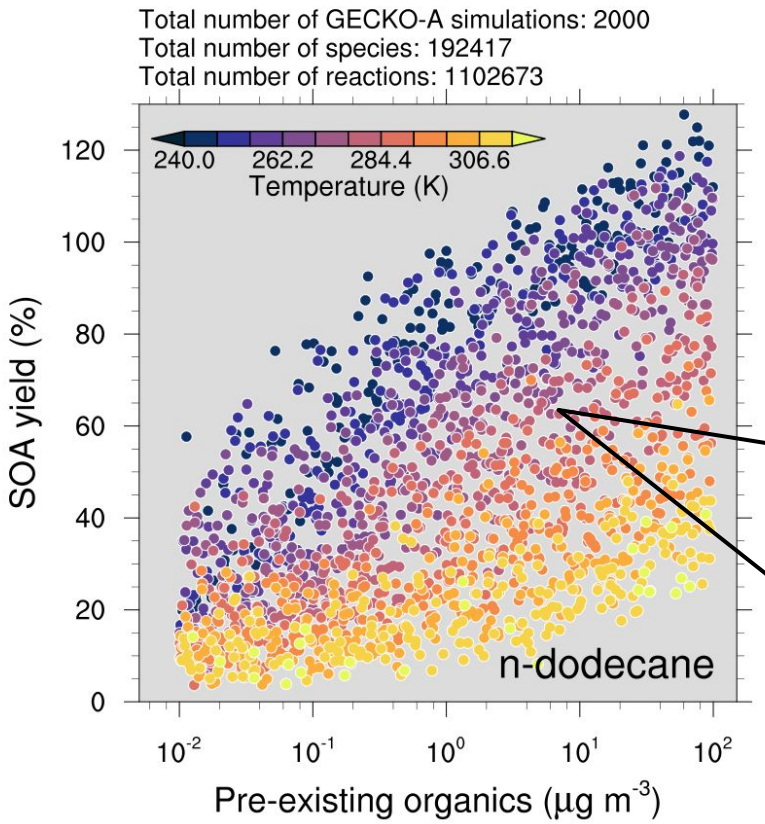


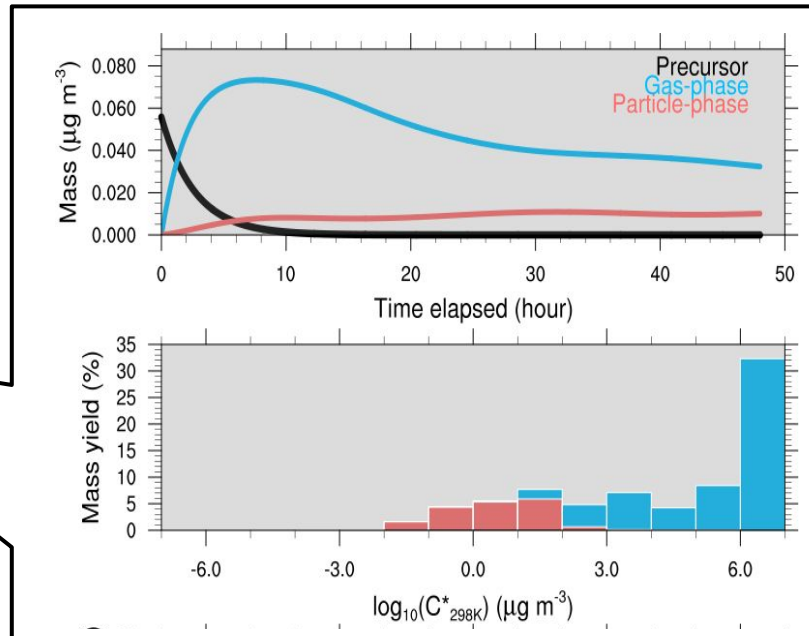
3-D Models



- Many inspiring applications out there: machine-learning emulators using explicit/process-level models, and implementing the trained emulators into large-scale models. Such explicit/process-level models are otherwise too expensive for large-scale models.
- The goal of this project is to train the machine-learning emulator using the “library” generated by the hyper-explicit chemical mechanism, GECKO-A.

Goal: Build Emulator to Predict the Total Organic Aerosol





	Time [s]	Precurso	Gas [ug/	Gas [ug/	Gas [ug/	Gas [ug/	Gas [ug/	Gas [ug/	Gas [ug/	Gas [ug/	Gas [ug/
2	1	3.77E-02	0	0	0	0	0	0	0	0	0
3	301.52	3.76E-02	9.61E-26	1.70E-21	2.46E-22	1.69E-20	2.14E-18	3.40E-15	8.12E-15	1.59E-14	1.01E-12
4	602.04	3.76E-02	4.61E-25	8.13E-21	1.06E-19	2.52E-19	1.14E-14	7.65E-15	1.32E-14	7.54E-14	6.72E-12
5	902.56	3.75E-02	1.14E-24	2.00E-20	1.25E-18	5.27E-19	4.08E-14	1.65E-14	3.10E-14	1.18E-13	2.16E-11
6	1203.08	3.74E-02	2.59E-24	4.54E-20	1.06E-18	6.37E-19	3.92E-14	3.97E-14	9.22E-14	1.51E-13	4.87E-11
7	1503.6	3.74E-02	3.98E-24	6.95E-20	9.74E-19	1.02E-18	4.47E-14	6.41E-14	1.28E-13	1.80E-13	8.95E-11
8	1804.12	3.73E-02	5.72E-24	9.97E-20	8.14E-19	1.16E-18	3.95E-14	9.65E-14	1.69E-13	2.20E-13	1.47E-10
9	2104.64	3.72E-02	8.11E-24	1.41E-19	7.54E-19	1.63E-18	4.27E-14	1.42E-13	1.91E-13	2.64E-13	2.22E-10
10	2405.16	3.72E-02	9.83E-24	1.71E-19	6.16E-19	2.02E-18	3.56E-14	1.91E-13	2.30E-13	2.98E-13	3.17E-10
11	2705.68	3.71E-02	1.10E-23	1.91E-19	5.04E-19	2.57E-18	2.98E-14	2.76E-13	2.54E-13	3.47E-13	4.32E-10
12	3006.2	3.70E-02	1.25E-23	2.16E-19	4.59E-19	3.44E-18	2.97E-14	3.87E-13	2.92E-13	4.23E-13	5.70E-10
13	3306.72	3.70E-02	1.46E-23	2.52E-19	4.38E-19	4.58E-18	3.14E-14	5.36E-13	3.32E-13	4.86E-13	7.30E-10
14	3607.24	3.69E-02	1.69E-23	2.89E-19	4.18E-19	5.98E-18	3.22E-14	7.04E-13	3.68E-13	6.26E-13	9.13E-10
15	3907.76	3.68E-02	1.95E-23	3.34E-19	3.50E-19	7.58E-18	2.74E-14	9.25E-13	4.14E-13	6.92E-13	1.12E-09
16	4208.28	3.68E-02	2.23E-23	3.79E-19	2.69E-18	9.67E-18	6.77E-14	1.16E-12	4.73E-13	1.34E-12	1.35E-09
17	4508.8	3.67E-02	2.42E-23	4.11E-19	5.93E-18	1.22E-17	1.27E-13	1.43E-12	5.26E-13	1.38E-12	1.61E-09
18	4809.32	3.66E-02	2.48E-23	4.21E-19	4.95E-18	1.50E-17	1.15E-13	1.75E-12	5.75E-13	1.49E-12	1.89E-09
19	5109.84	3.66E-02	2.58E-23	4.37E-19	4.22E-18	1.84E-17	1.11E-13	2.14E-12	6.40E-13	1.51E-12	2.19E-09
20	5410.36	3.65E-02	2.77E-23	4.67E-19	3.50E-18	2.18E-17	9.44E-14	2.56E-12	6.96E-13	1.55E-12	2.53E-09

Demo: what the data looks like

GECKO-A Library:

- 2000 GECKO-A simulations: in each run, we run GECKO-A under certain condition for 5 days
- 2000 input files (csv).
- Each file contains: (i) mass of precursors; (ii) mass of products in the gas-phase; and (iii) mass of products in the particle-phase. All (i)-(iii) as a function of time.

Machine Learning the Warm Rain Process

- The warm rain formation process is critical for weather and climate prediction.
- Simply parameterized in large scale models with bulk microphysics
- More detailed treatments computationally expensive
  - Stochastic collection

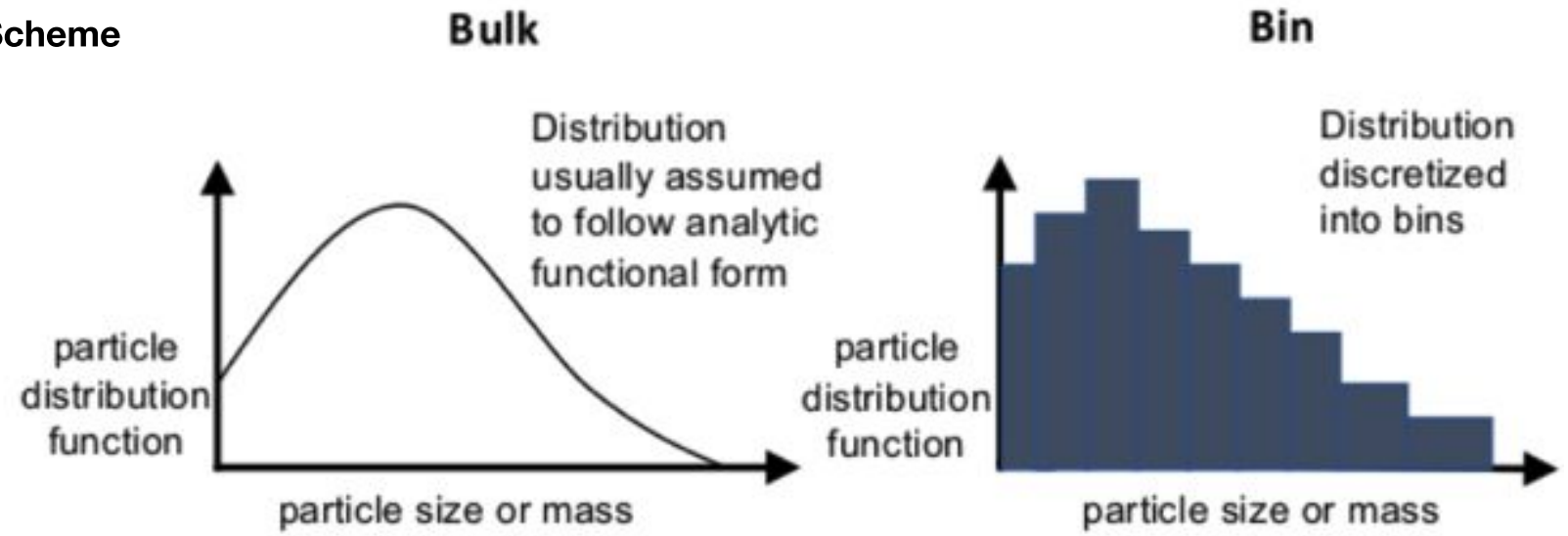
Ask 2 questions:

1. Can we simulate warm rain using alternative methods, what does it do to cloud susceptibility and cloud feedback?
2. Can we use Neural Network emulators (NN) to then speed up this process and reproduce these changes….

Methodology

- Replace existing CAM6-MG2 BULK microphysics warm rain formation process (Khairoutdinov & Kogan 2000: regressions from LES experiments with explicit bin model)
- Instead use the Stochastic Collection Kernel and process from a BIN microphysical model (Tel-Aviv University [TAU] bin model)

Microphysics Scheme Types



Bulk: Efficient, but approximates interactions between drop sizes

Bin: Explicit interactions between sizes closer to physical equations, but too computationally expensive to calculate processes

BULK: Auto-conversion (Ac) & Accretion (Kc)

Khairoutdinov & Kogan 2000: regressions from LES experiments with explicit bin model

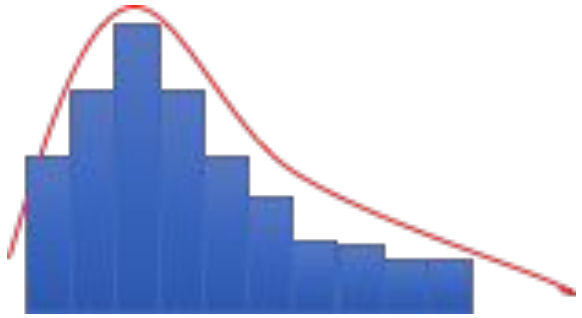
Ac = $$\left(\frac{\partial q_r}{\partial t}\right)_{\text{auto}} = 1350 q_c^{2.47} N_c^{-1.79}, \qquad (29)$$

Kc= $$\left(\frac{\partial q_r}{\partial t}\right)_{\text{accr}} = 67 (q_c q_r)^{1.15}. \qquad (33)$$

- Auto-conversion an inverse function of drop number
- Accretion is a mass only function

Balance of these processes (sinks) controls mass and size of cloud drops

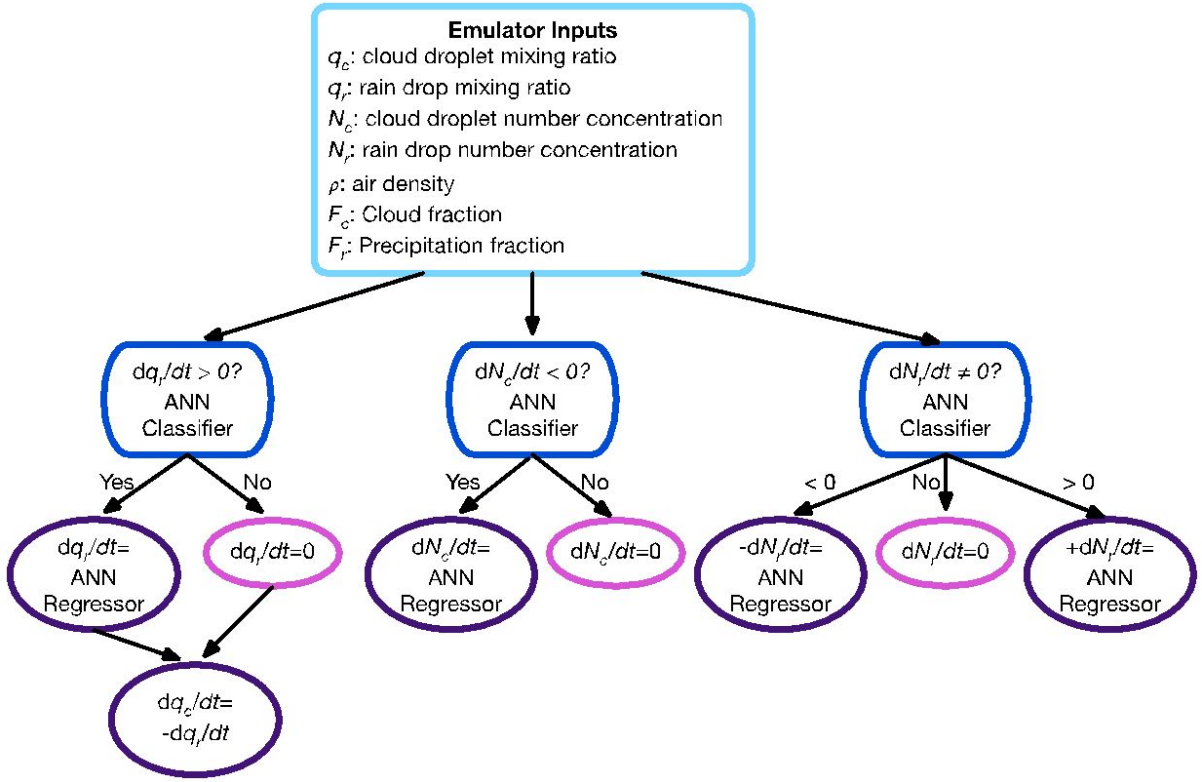
Method



- Break **BULK** size distributions for qc,Nc (liquid) and qr,Nr (rain) into **BINS**
- Run stochastic collection kernel on the bins
- Find minimum between peaks of distributions to separate qc and qr
- Recompose qc,Nc and qr,Nr distributions
- Difference before and after distributions are tendencies for qc,Nc,qr,Nr
- Apply a mass fixer to ensure no loss of mass or negative mass (TAU)
- Then: build a neural network emulator (TAU-ML)
  - What ML methods are being used.

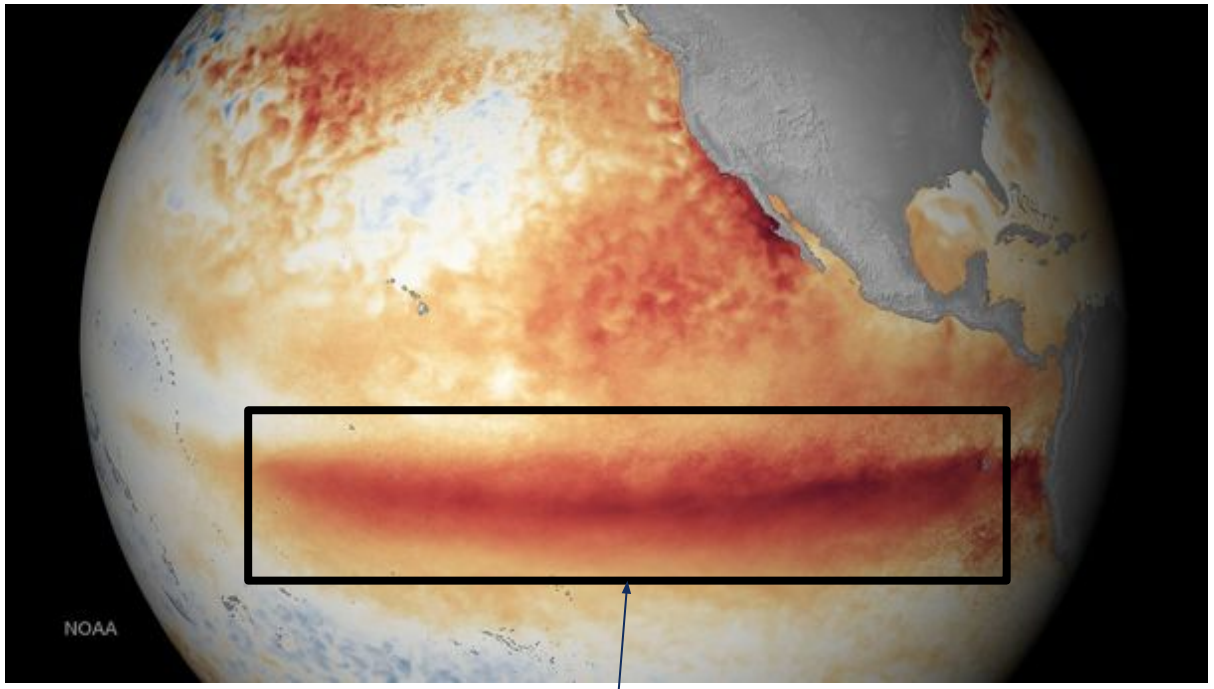
Using machine learning to emulate

1. Run CESM2/CAM6 for two years and obtain instantaneous hourly output
2. Filter and subsample data to find grid points with realistic amount of cloud water
3. Transform and normalize inputs and outputs
4. Train classifier deep neural networks to classify zero and non-zero
5. Train regression deep neural networks to predict non-zero values
6. Evaluate and interpret neural network predictions.



What is El Niño?

- Cycle of warm and cold temperatures in the equatorial Pacific Ocean
- Dominant pattern that influences seasonal temperature
- Broad implications for climate-sensitive sectors, such as energy and agriculture
- How is El Niño measured? *Niño3.4 Index*
  - Rolling 3-month average of sea surface temperatures in the equatorial Pacific



Source: National Oceanic and Atmospheric Administration

Equatorial Pacific Ocean with abnormally warm temperature: El Niño event

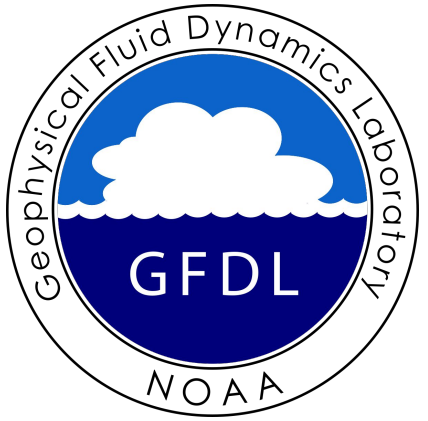
Learning to forecast El Niño

What is the current state of the art?

- Most ENSO forecasts are issued by weather centers, who run physics-based models

Why use neural networks?

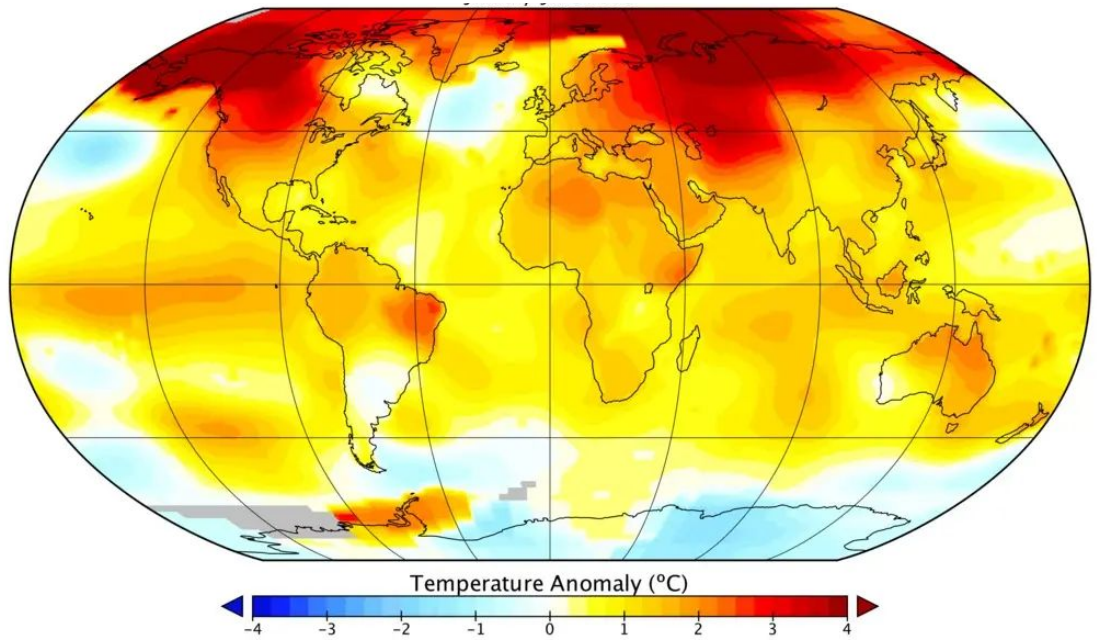
- Potential for more accurate forecasts?
- Lighter computational cost *during inference*
- **Challenge**: limited historical observations to use as training data for a neural network
- **Solution:** train on simulated climate data from Atmosphere-Ocean General Circulation Models (AOGCMs)



What questions will we explore during the hackathon?

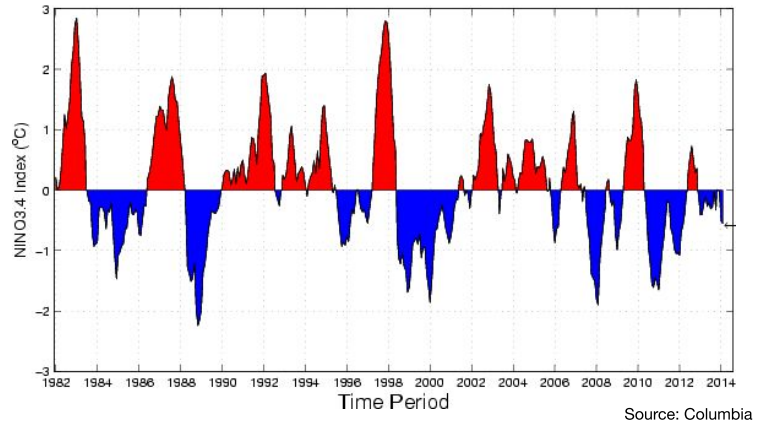
- **Data:** How does an increase in data affect the performance of machine learning?
- **Validation:** how can we ensure that we validate the model rigorously?
- **Ensembling:** What combination of models and training schemes creates the best forecasts?
- **Lead time:** How far ahead can machine learning make skillful predictions?
- **Extendability:** Can we use our neural network architecture to forecast *temperatures on land*?

Predictor Data: surface temperature



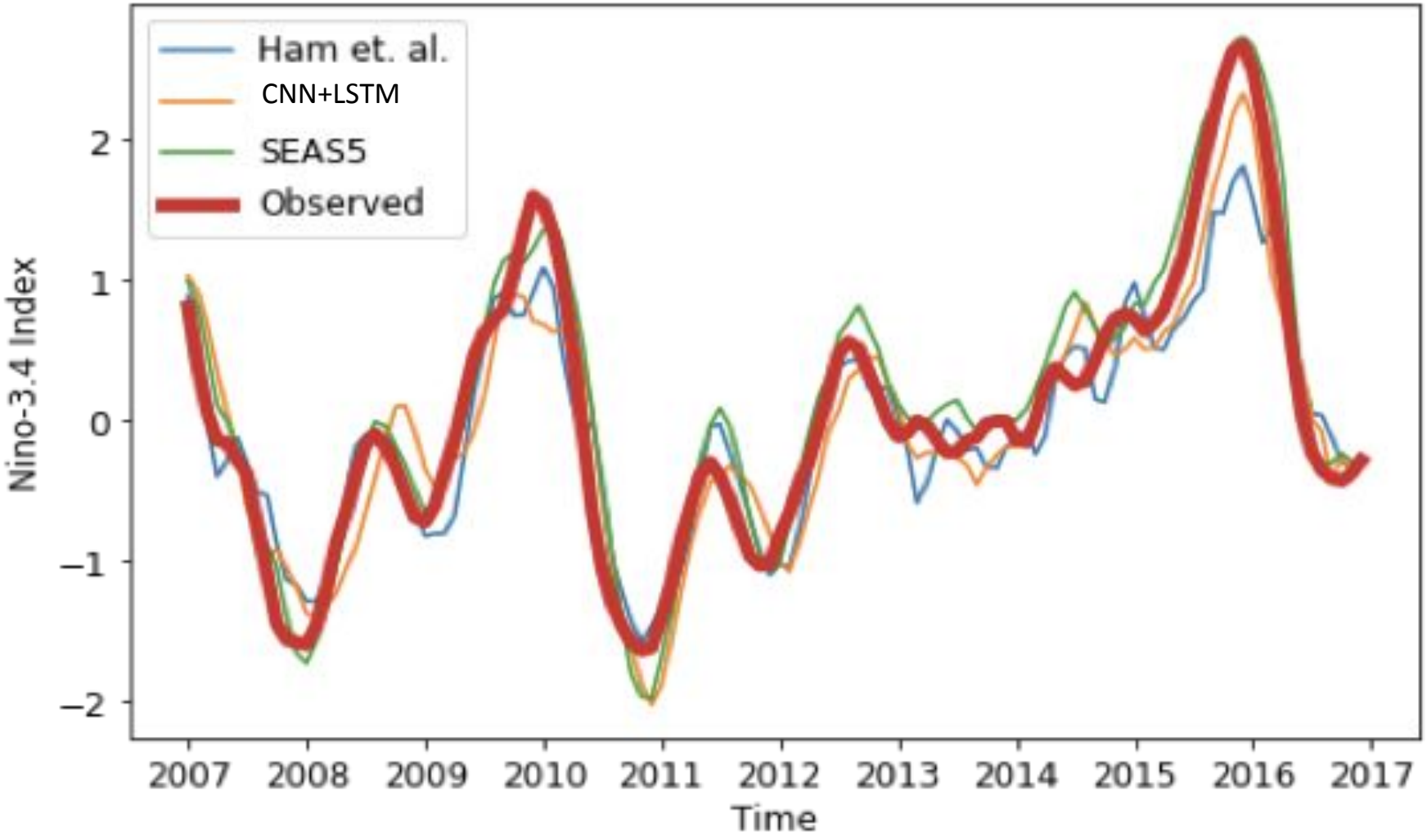
Source: NASA

Target Data:



Source: Columbia University

What might the forecasts look like? (4 month lead time)

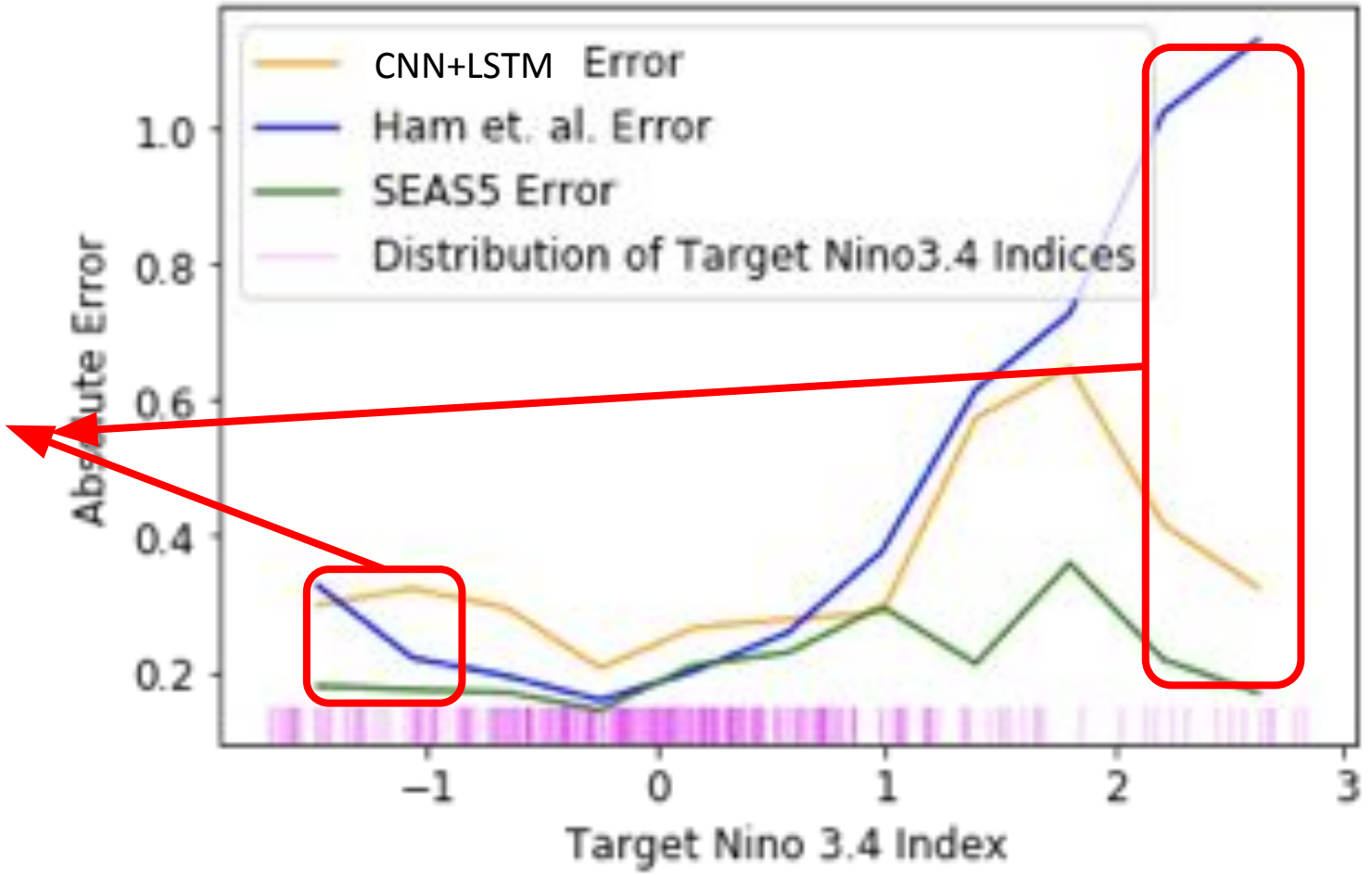


SEAS5: seasonal forecasting model from the European Center for Medium-range Weather Forecasts
CNN+LSTM: a deep learning architecture designed to learn from spatial and time series data

There's still work to do on ENSO forecasting!

- Why work on this problem?

- Deep learning's performance at extreme values of the Niño3.4 index still has room for improvement!



How to Access Jupyterhub

- The hackathon notebooks are on Github at https://github.com/NCAR/ai4ess-hackathon-2020 along with links to Google Colab for each challenge notebook

Hackathon Jupyterhub

- Go to the link emailed to you
- Log in with the Gmail/G-Suite Account you provided at registration
- A loading screen with a progress bar will appear
- Next, a Jupyterlab window will appear
- Enter the ai4ess-hackathon-2020/notebooks directory and open the challenge notebook assigned to your team