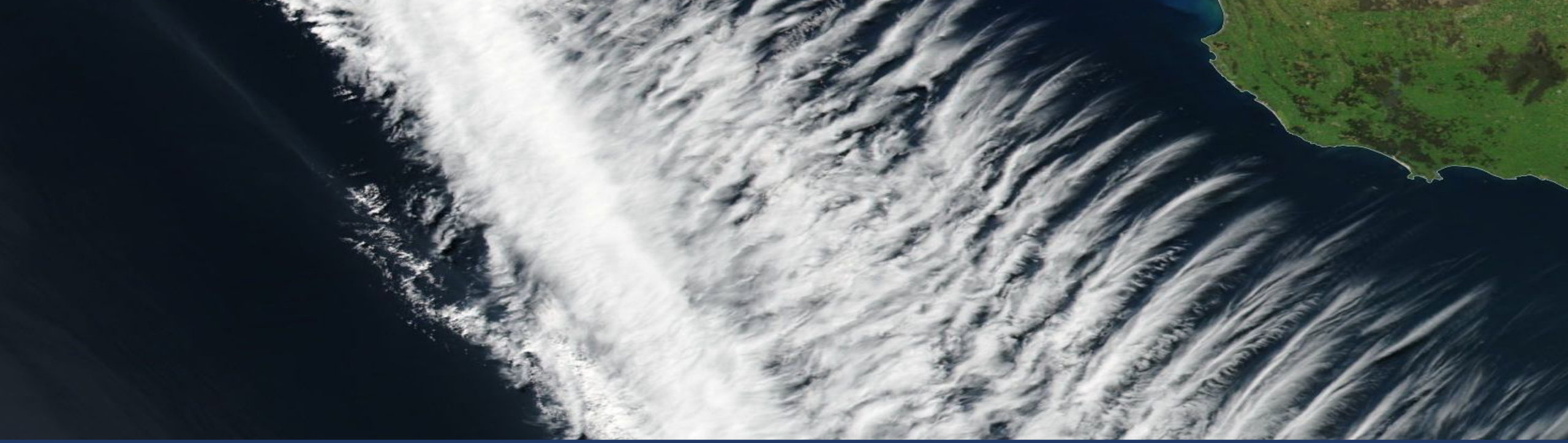


Birds Eye View on Analytics from Operational NASA Data Centers Perspective

Dr. Rahul Ramachandran
NASA/MSFC

Biographical information

Experience

- IMPACT Project Office, Project Manager, NASA/MSFC (July 2018)
- DAAC Manager, NASA/MSFC, 2013 - 2018
- Principal Research Scientist III, University of Alabama in Huntsville, 2011- 2013
- Team Lead, Earth Science Informatics, Sr. Spatial Systems Architect, Oak Ridge National Laboratory, 2010-2011

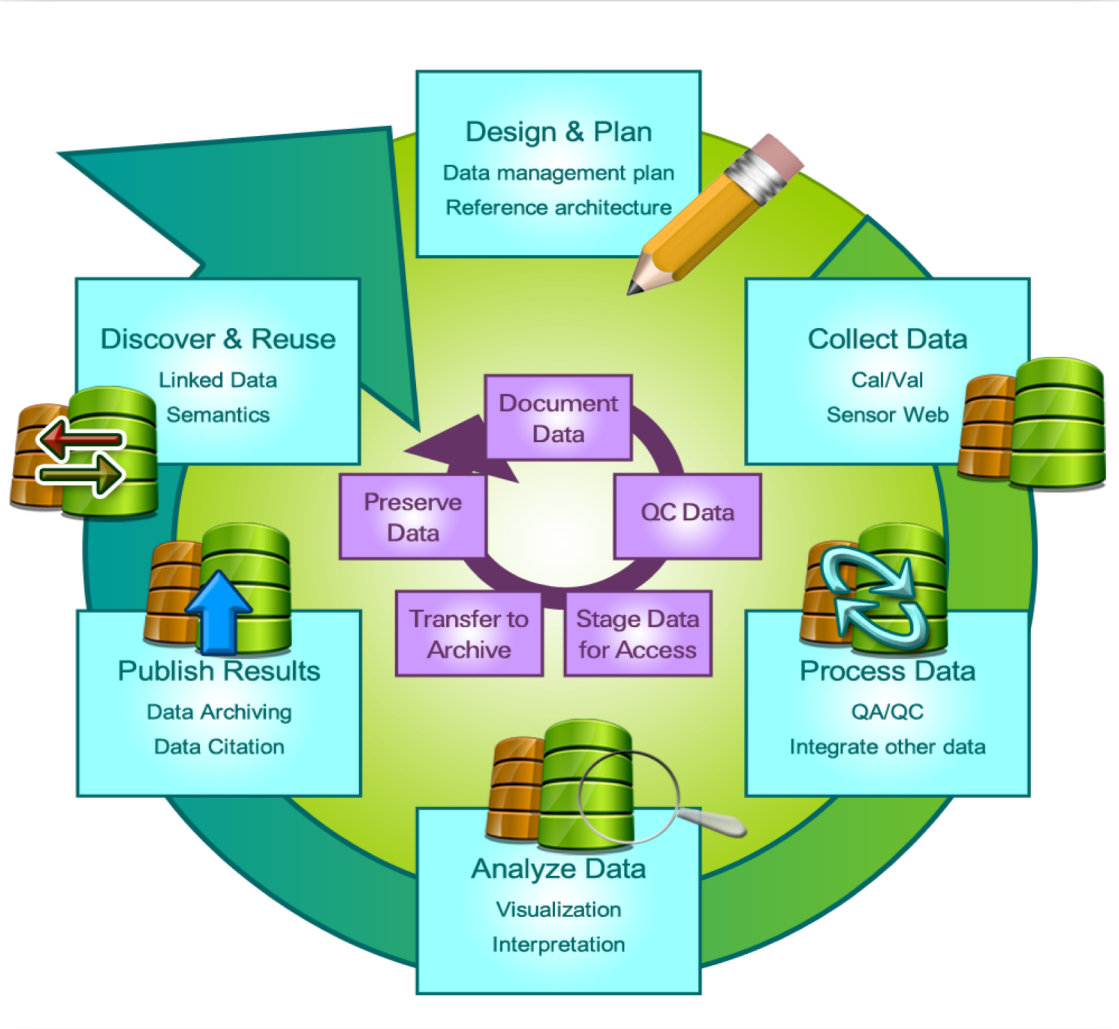
Awards

- Presidential Early Career Award for Scientists and Engineers (PECASE), 2009

Appointments

- Member Editorial Board, CODATA Data Science Journal 2015- present
- Chair, Earth Science Informatics Technical Committee, IEEE Geoscience & Remote Sensing Society, 2013-2015
- Adjunct Professor, Department of Atmospheric Science, University of Alabama in Huntsville, 2012-2013
- Member, Artificial Intelligence Applications to Environmental Science, American Meteorological Society Committee (2010 – 2013)

Technology as a Research Enabler



Informatics and Data Science components in both the "data" (inner) and the "knowledge generation" (outer) lifecycles.

Foundational data services:

- Design and develop new data and information architecture, components, and systems
- Improve information quality, standards, data/information related processes, and develop best practices

Knowledge generation services:

- New algorithms and tools to support information/knowledge extraction

Khalsa, Siri Jodha S, and Rahul Ramachandran. 2013. "Earth Science Informatics Comes of Age." *IEEE Geoscience and Remote Sensing Magazine*, December.

Ramachandran, Rahul, and Siri Jodha S. Khalsa. 2015. "Moving from Data to Knowledge: Challenges and Opportunities." *IEEE Geoscience and Remote Sensing Magazine*, July.

Outline

1. Overview of NASA Earth Science Data System Program
2. Moving towards the Cloud
    1. Cumulus – Ingest/Archive System
3. Cloud Analytics Workshop Results
4. Deep Learning Applications
    1. Hurricane Intensity Estimation
    2. Transverse Cirrus Band Detection
    3. Use our NASA data for your DL experiment!

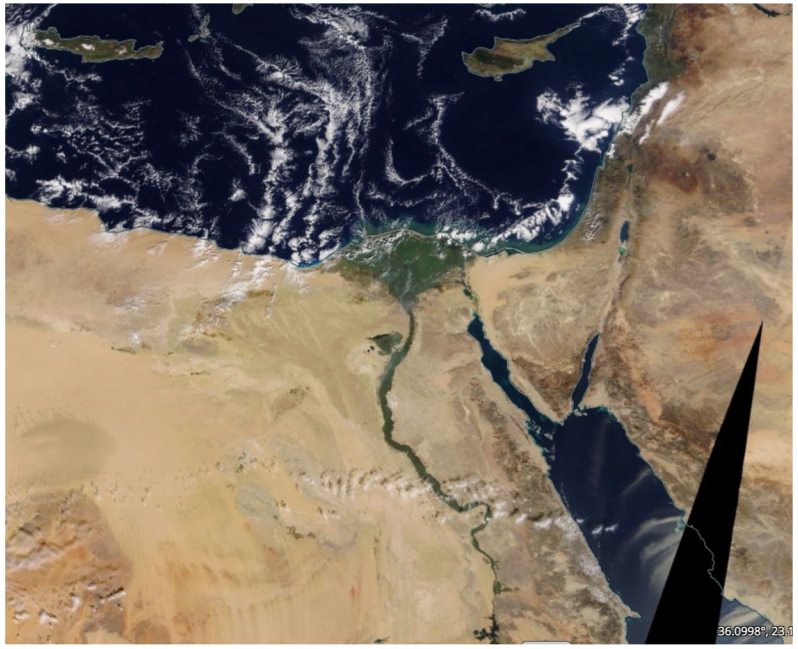
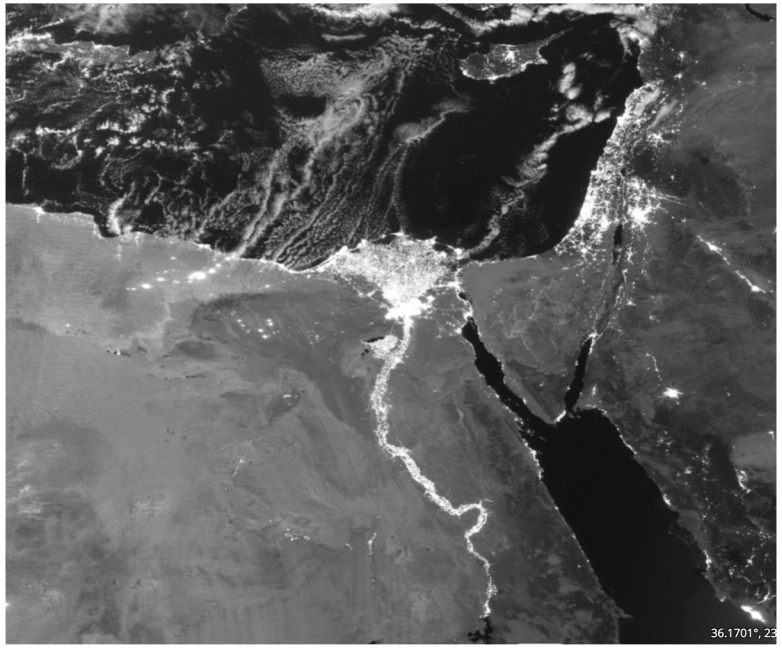
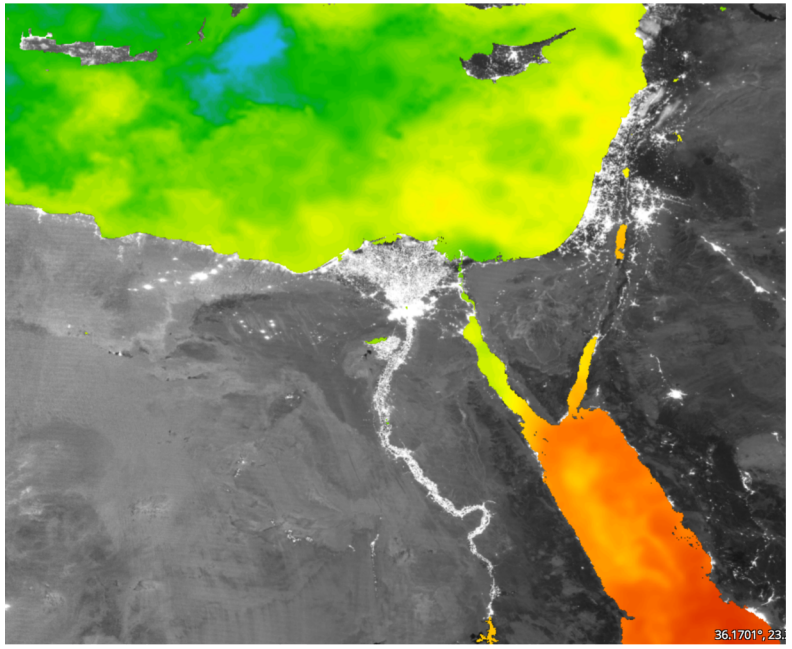


"Advance knowledge of Earth as a system to meet the challenges of environmental change, and to improve life on our planet"

- NASA Earth Science Objective

1. NASA's Earth Science Data Systems

Kevin Murphy
NASA HQ/ESDS PE

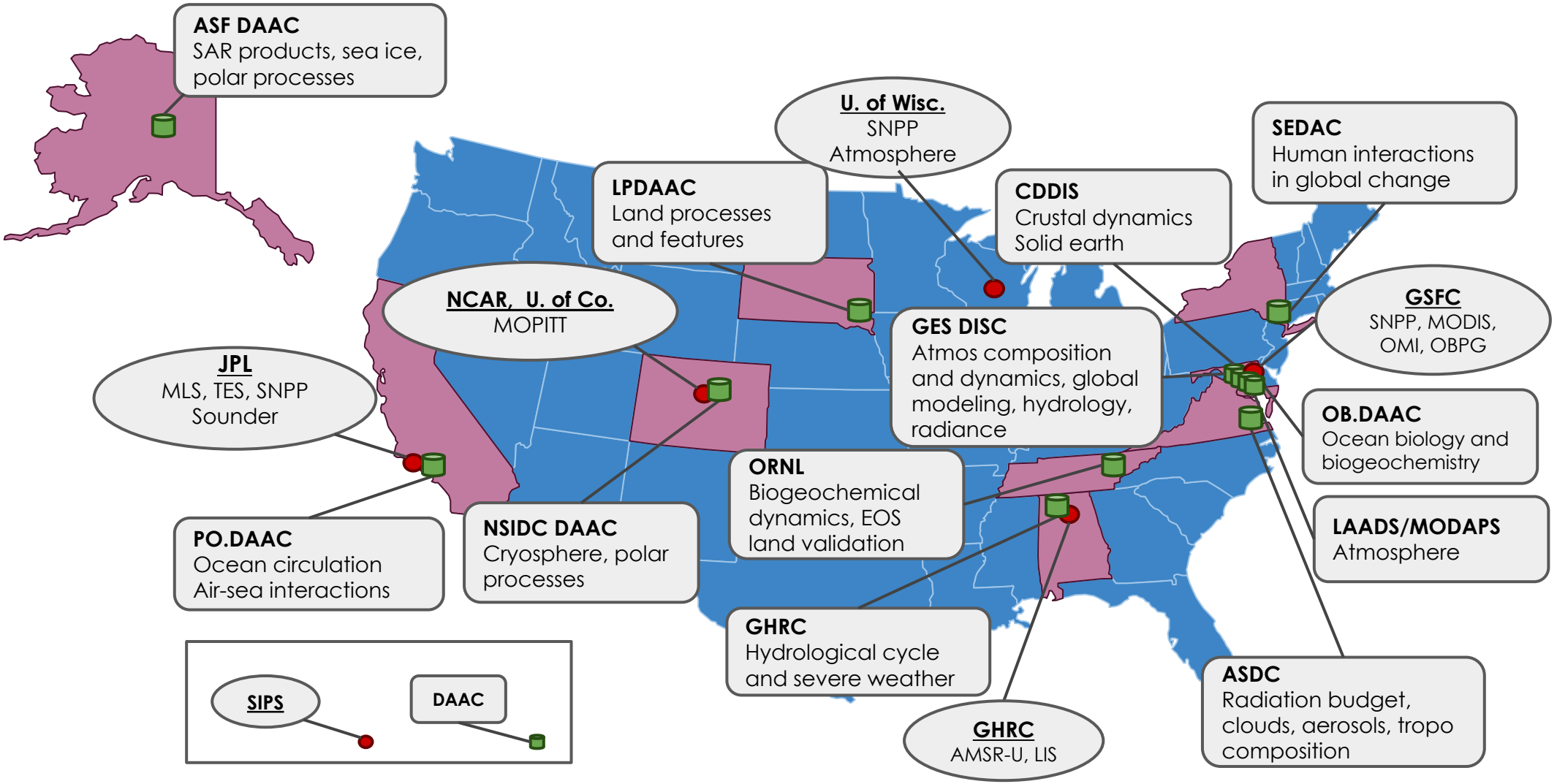
Single largest repository of Earth Science Data, integrating **multivariate/heterogeneous** data from diverse observational platforms.

Earth Science Open Data Policy

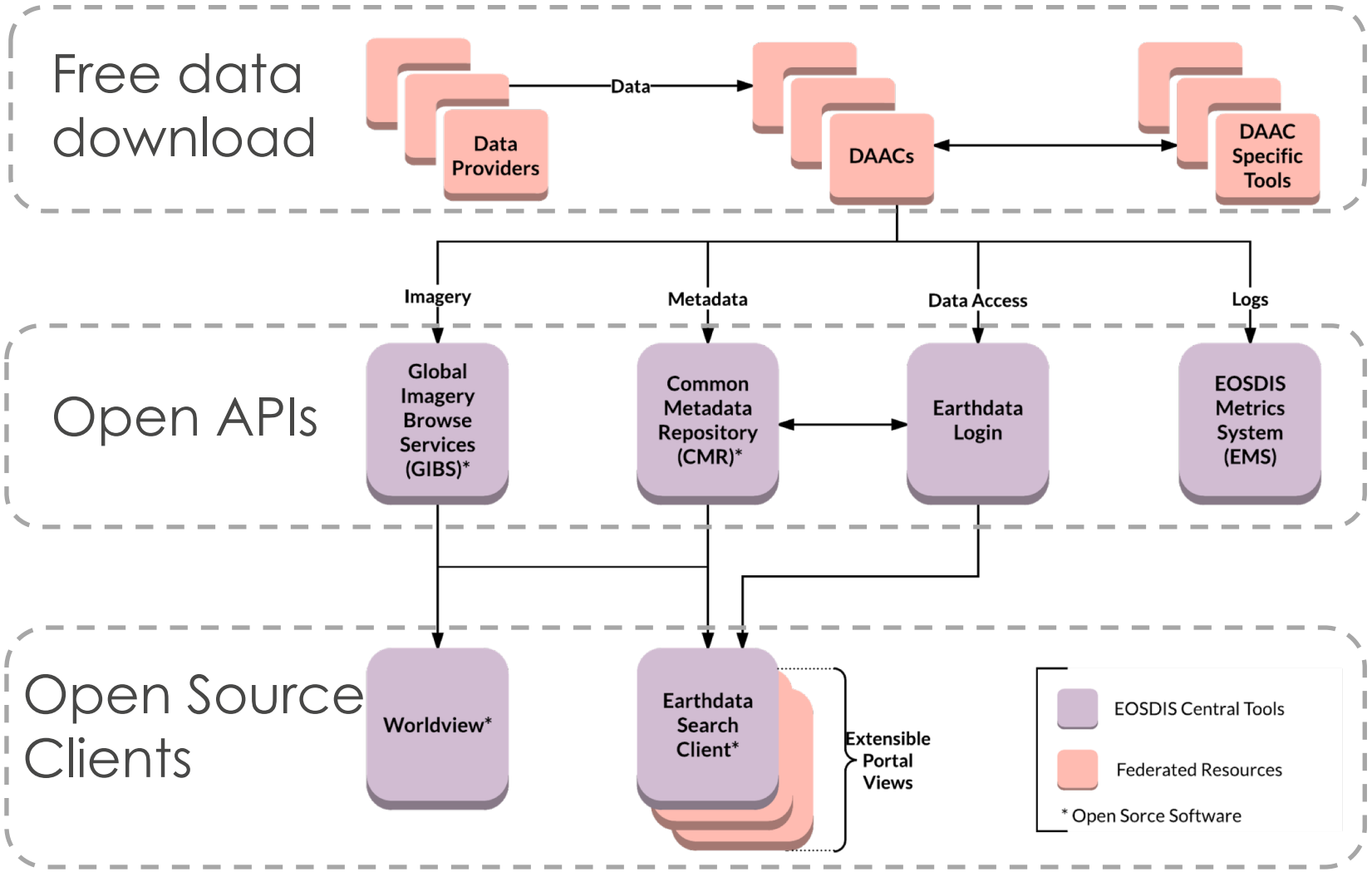
NASA's earth observation data is **collected continuously.** For over half a century these invaluable records of earth processes have provided a critical resource for scientists and researchers.

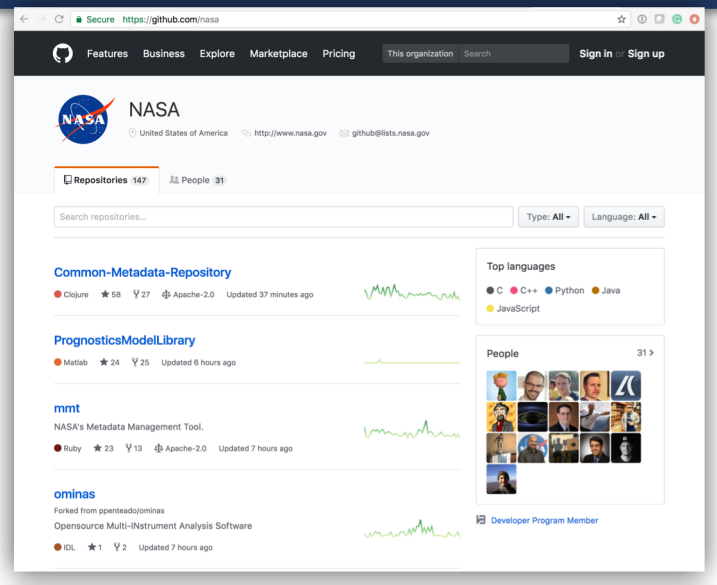
Since 1994 NASA earth science data have been free and open to all users for any purpose as quickly and as practical after instrument checkout and calibration.

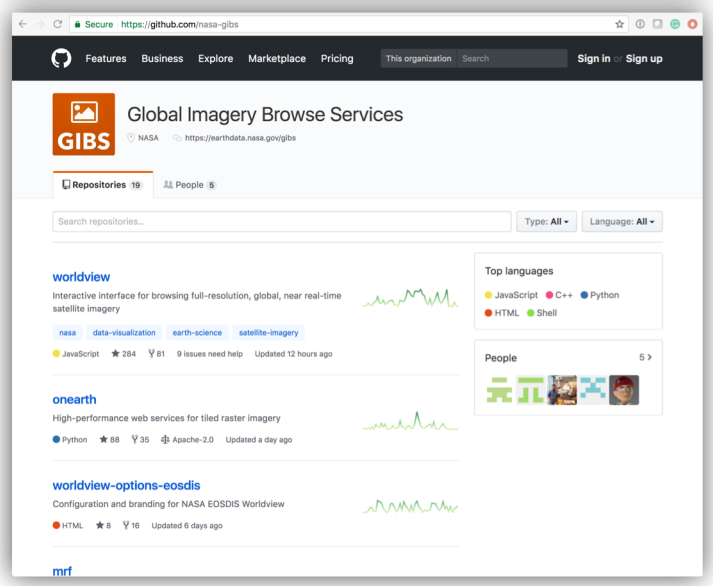
Data is Produced and Managed by Science Discipline Experts



Open Data, Open Source, Open Services





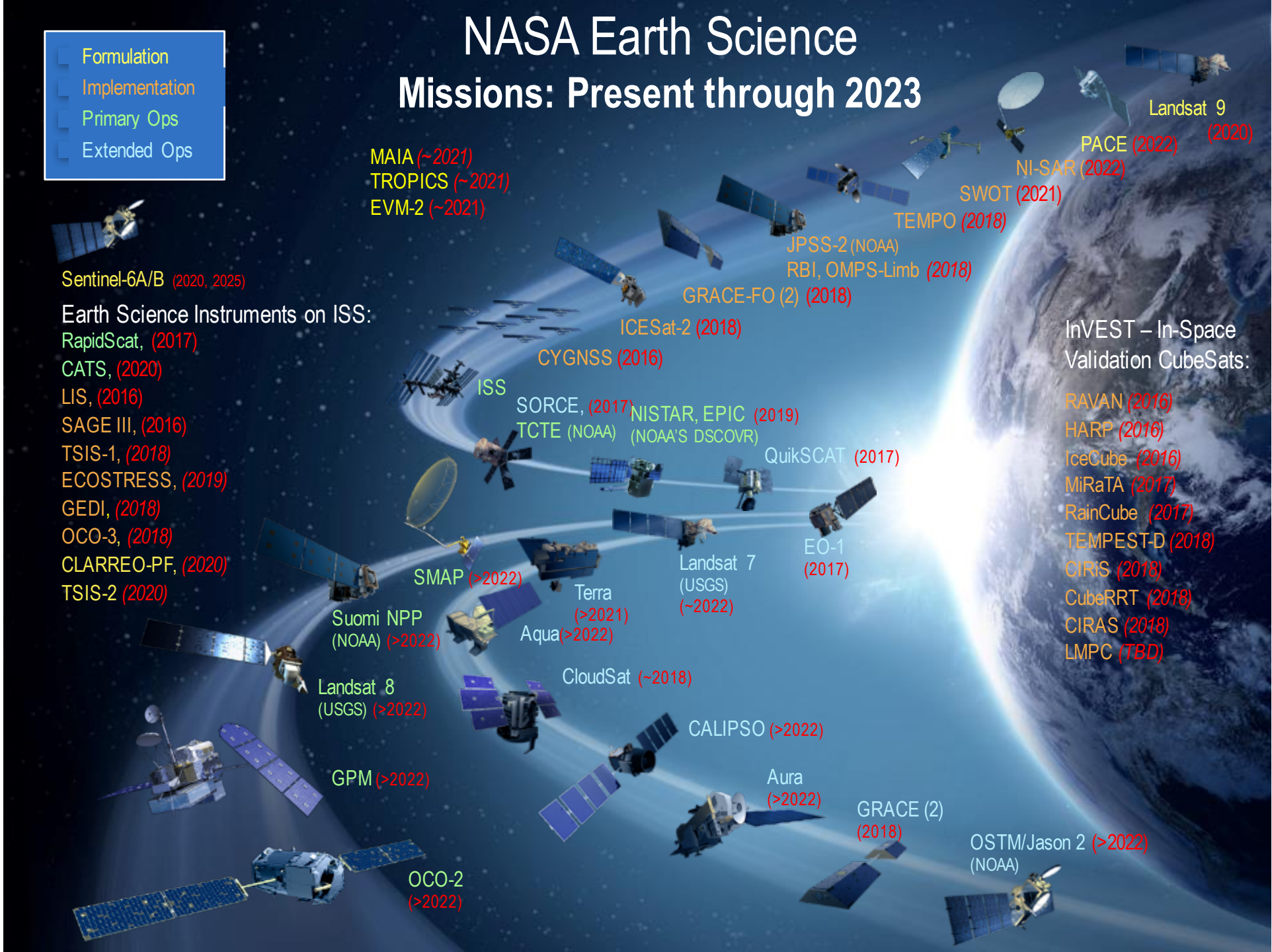


5 Years from Today

New instruments and missions.

2017 NRC Earth Science Decadal Survey - National Imperatives for the Next Decade and Beyond

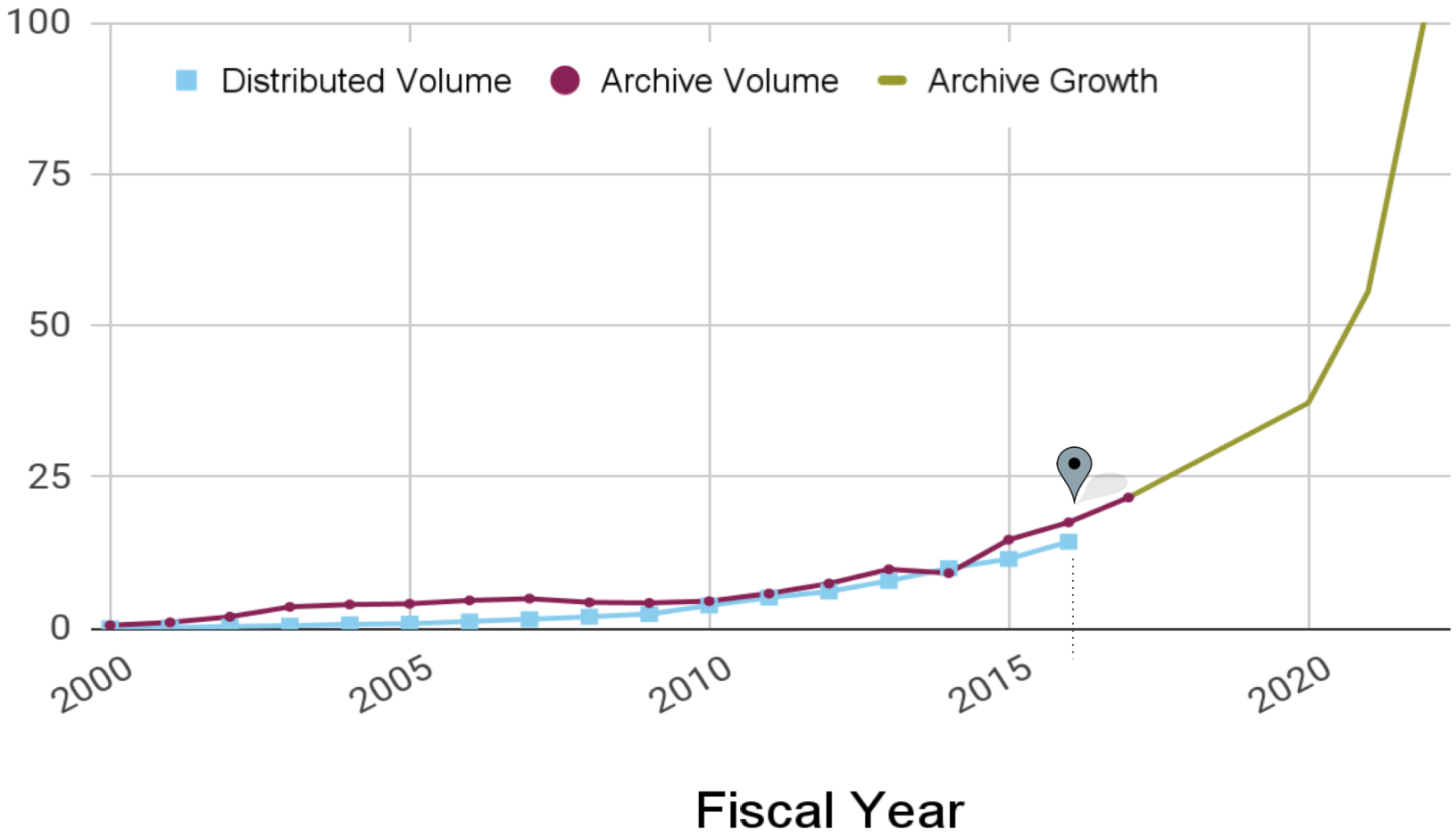
User expectations continue to evolve.



Data System Evolution

The current architecture will not be cost effective as the annual ingest rate increases from **4** to **50PB/year**

EOSDIS is developing **open source cloud native software** for reuse across the agency, throughout the government and for any other user.



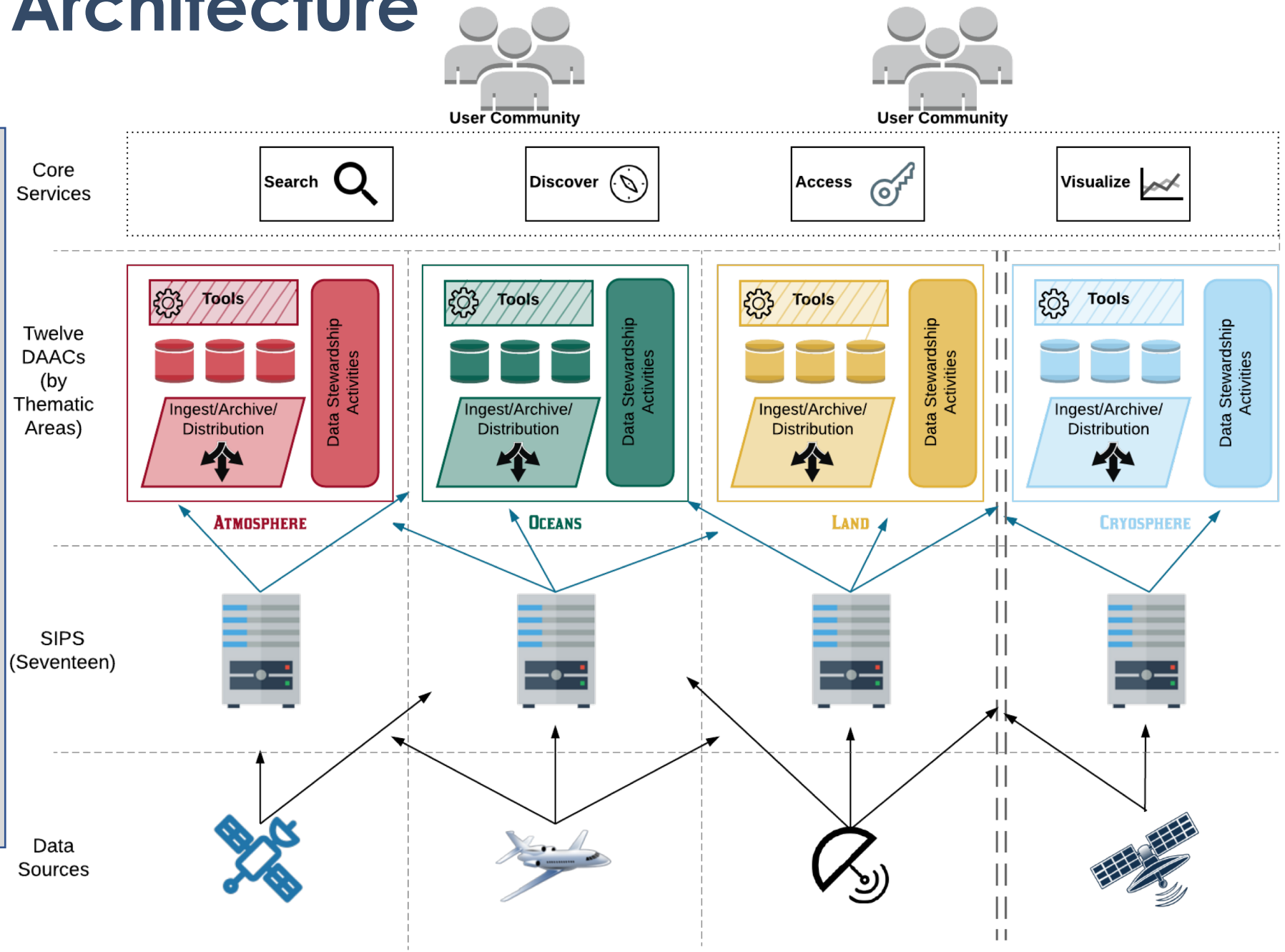
Cloud offers benefits like the ability to analyze data at scale, analyze multiple data sets together easily and avoid lengthy expensive moves of large data sets allowing scientists to work on data "in place"

Current Architecture

Optimized for archival, search and distribution

Expert user support

Easily add new data producers



Uneven levels of service and performance

Significant time for interface coordination

No processing in place

Balanced Path Forward

- Maintain what Earth Science Data Systems do well
  - Data stewardship expertise to support science community needs.
  - Flexibility to incorporate new data sources and data producers with little disruption.
  - Provide free and open access to data products and services all users with open APIs and software.
  - High level of customer satisfaction
- Improve user functionality and data stewardship efficiency
  - Ability to conduct regional and global processing without need for data management
  - Consistent level of service and performance across the system

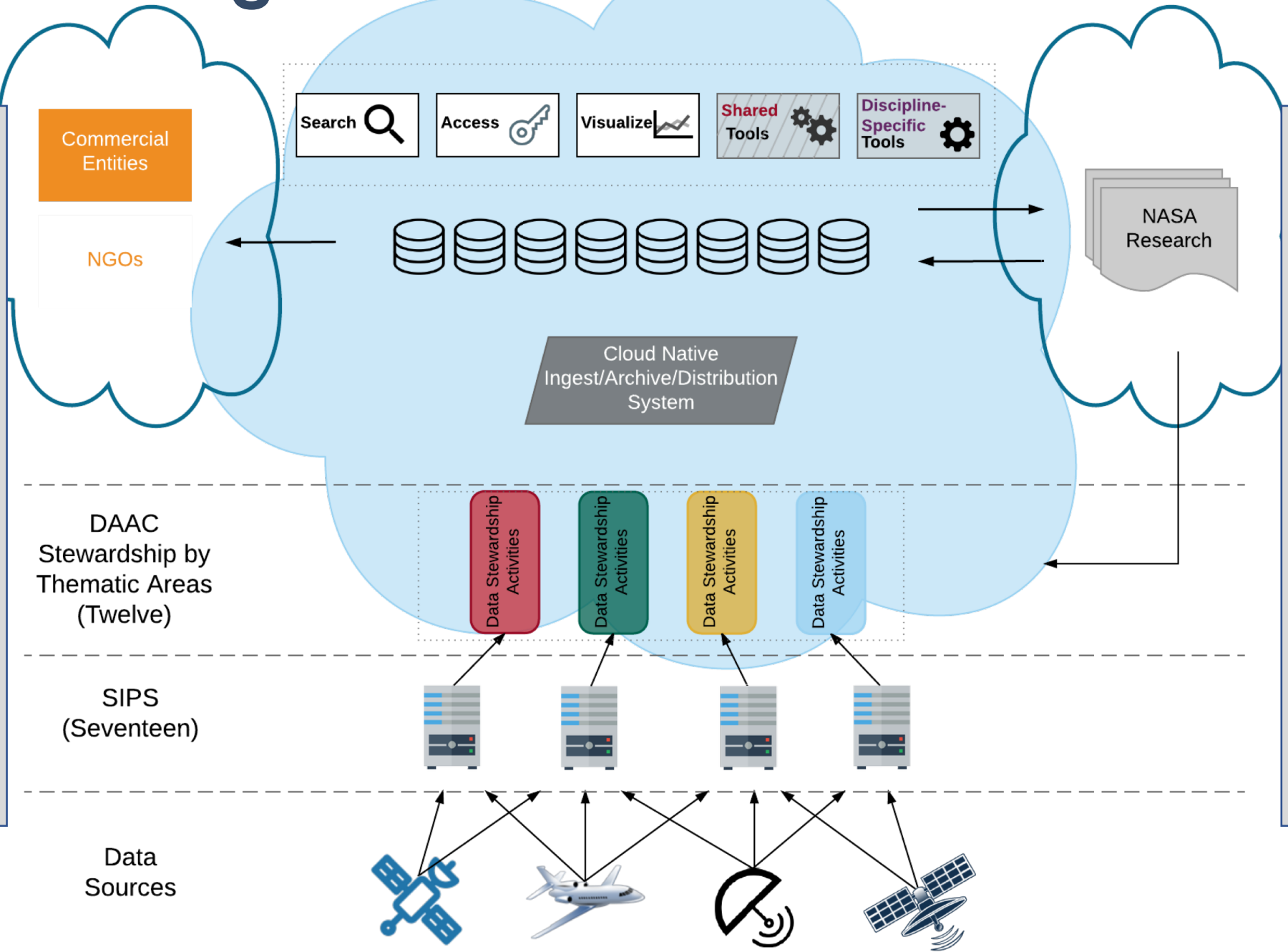
Evaluating Commercial Cloud Architecture

Processing next to data – buy by the yard for **anyone**

Data available to all DAACs and users

Expert user support

Easily add new data producers



Development coordination

Cost management

Security, ITAR, SBU, EAR99

Business processes and skillsets

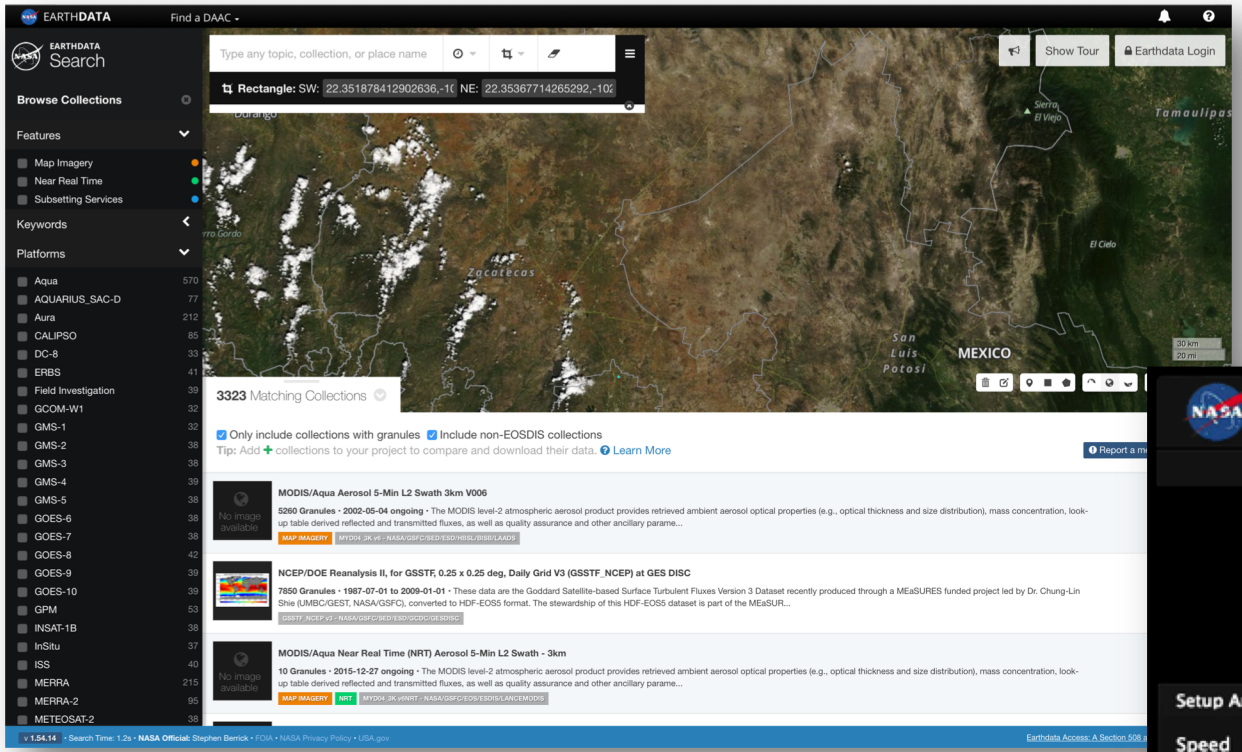
Lock-in



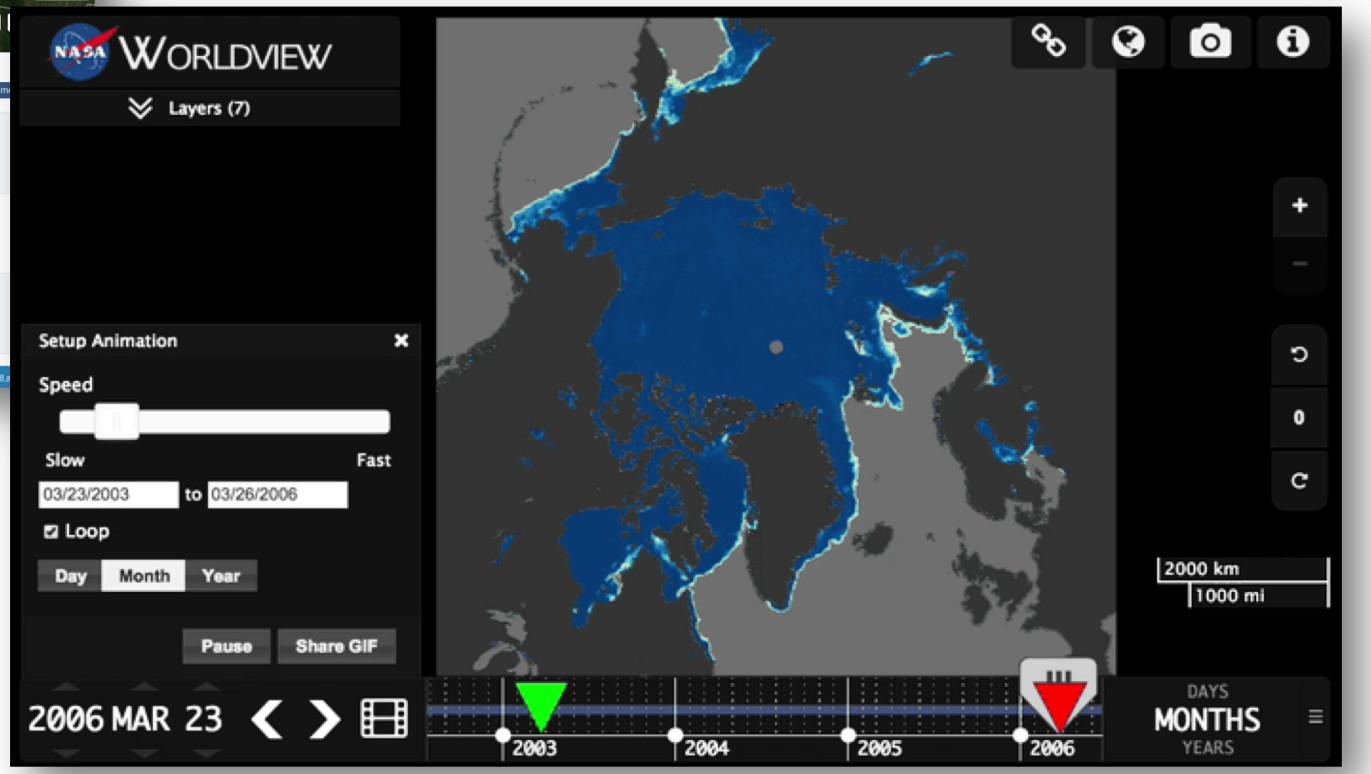
Earthdata Cloud 2021

- Improve the efficiency of NASA's data systems operations
- Prepare for planned high-data-rate missions
- Increase opportunity for researchers and commercial users to process PBs of data quickly without the need for data management/movement

Cross Discipline Data Access



Data-centric users
https://search.earthdata.nasa.gov



Imagery-centric users
https://worldview.earthdata.nasa.gov

2. Moving Towards the Cloud

2.1 Cumulus: NASA's cloud based distributed active archive center system

Rahul Ramachandran, Katie Baynes, Kevin Murphy, Alireza Jazayeri, Ian Schuler, Dan Pilone

NASA MSFC, NASA GSFC, NASA HQ, Development Seed, Element 84

Ramachandran, R., Baynes, K., Murphy, K., Jazayeri, A., Shuler, I., & Pilone, D. (2017). Cumulus: NASA'S Cloud Based Distributed Active Archive Center Prototype. In *IEEE Geoscience and Remote Sensing Symposium*. Fort Worth, Texas.

Cumulus Project

Goal: Design and develop a functional "light weight" data ingest, archive, and distribution "cloud native" framework

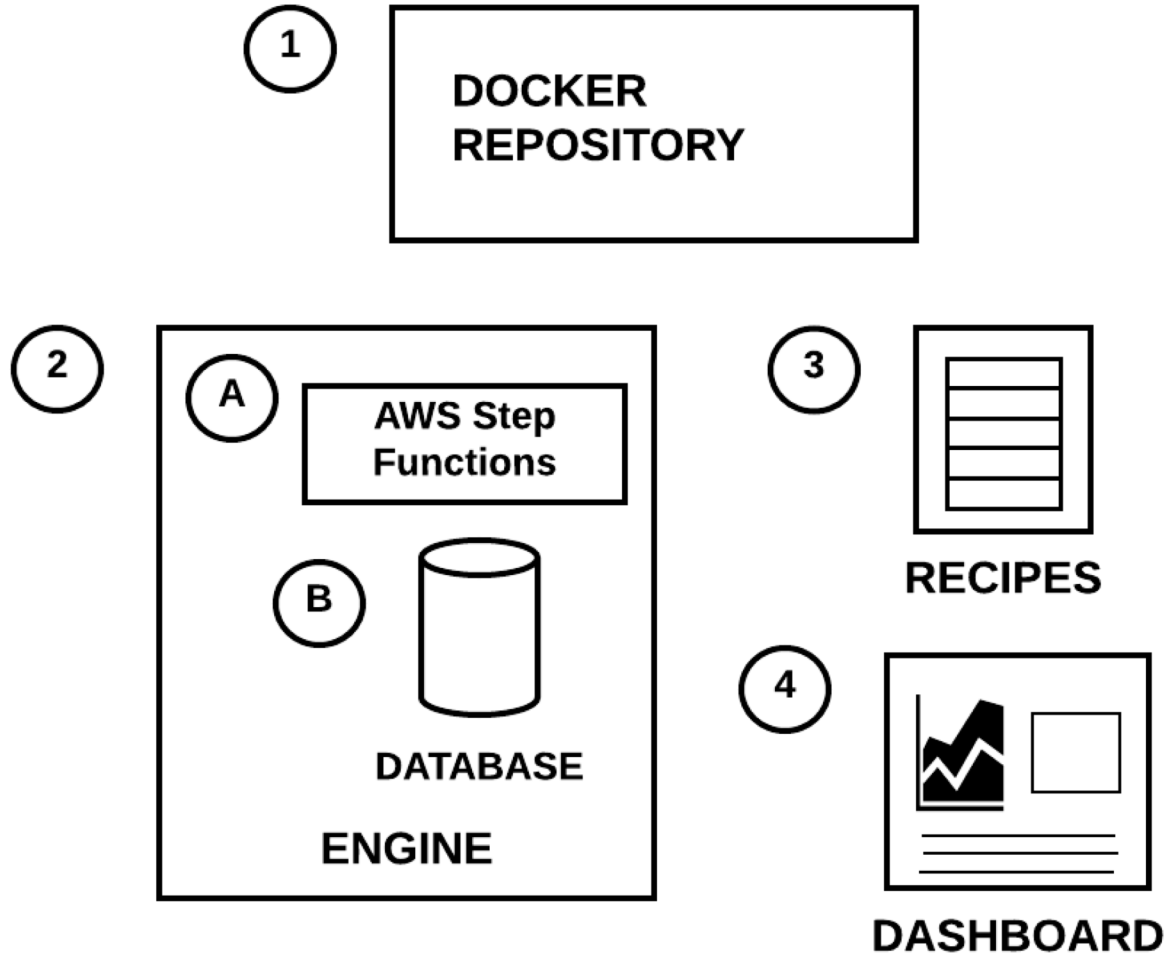
Specific Objectives:

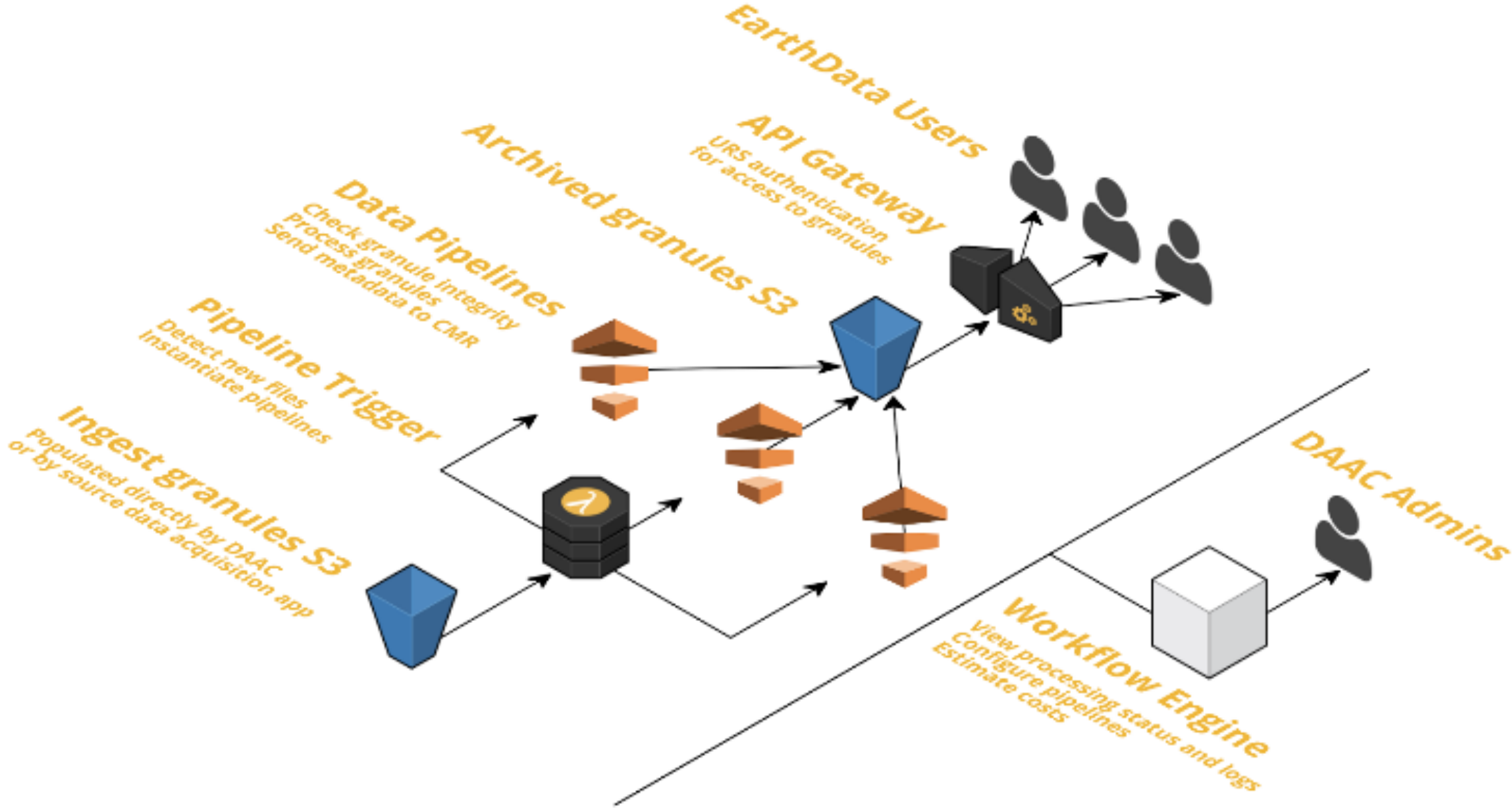
- Provide core DAAC functions (ingest/processing/archive) so they can be performed on a commercial cloud
- Operate within NASA's security compliance policies
- Provide cost estimates from running a few data streams
- Provide options for future operational strategy
- Provide a planning tool for DAACs to enable their transition to Cumulus
- Provide guidance for future Cumulus operations by DAACs

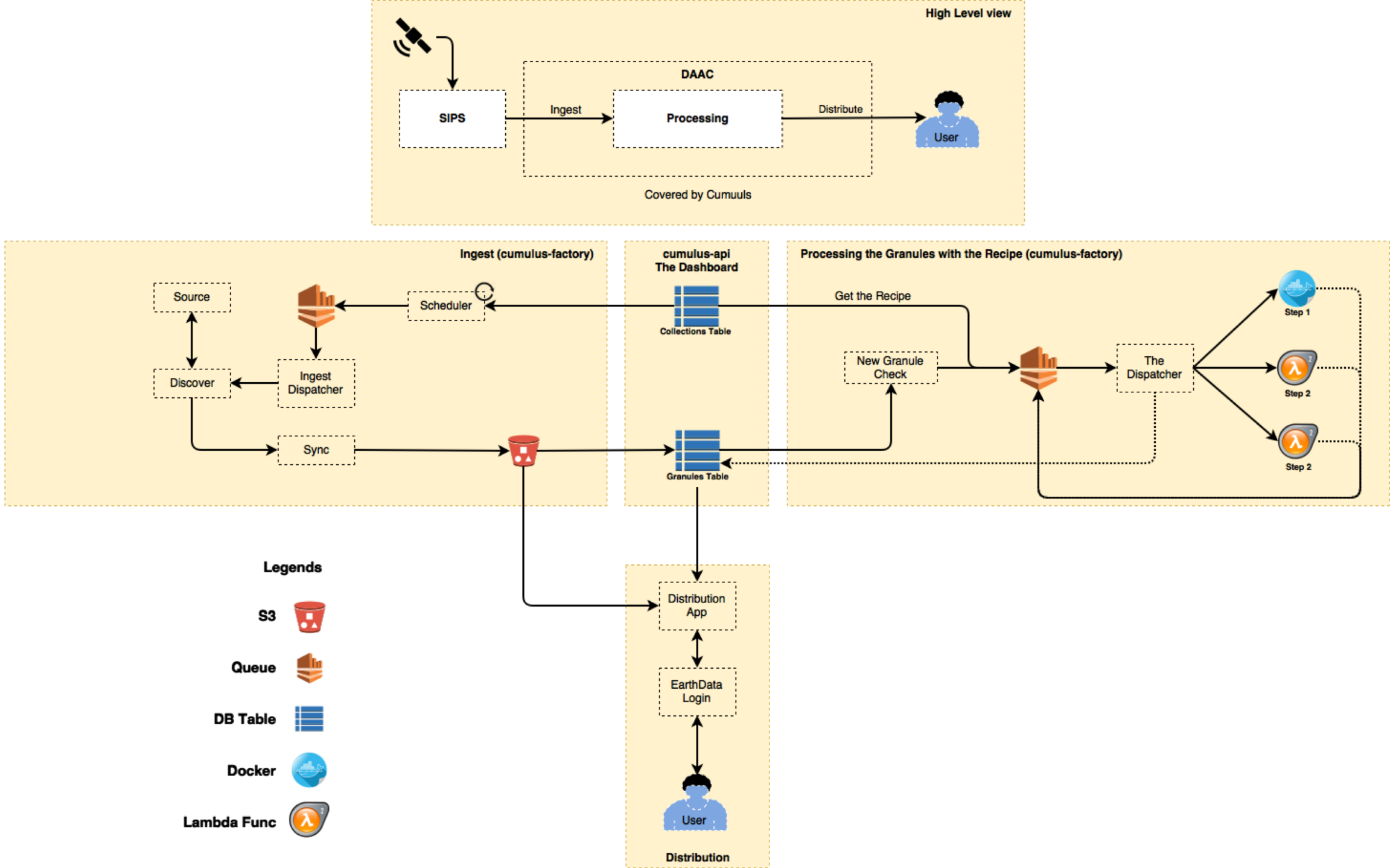
What is Cumulus?

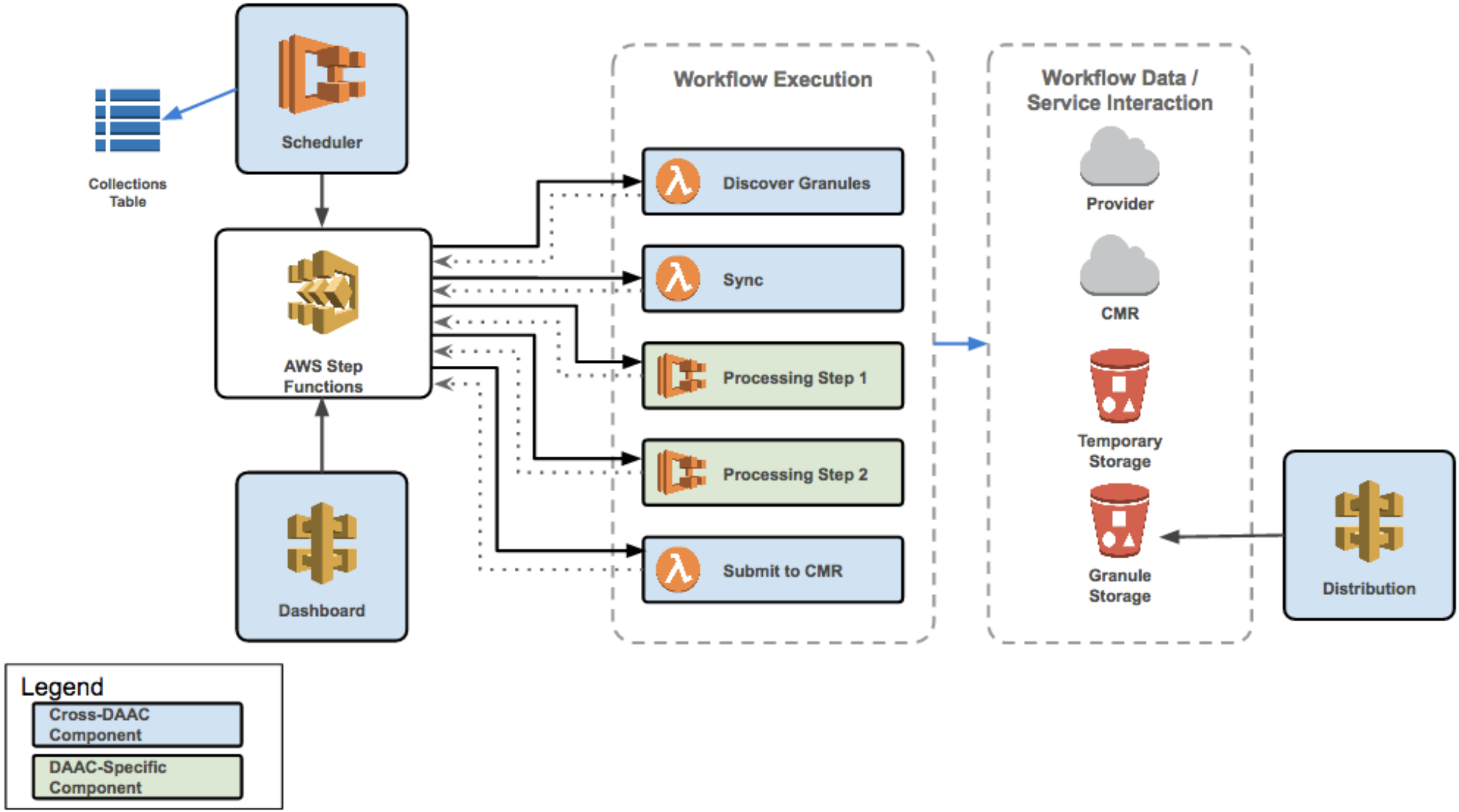
A lightweight framework consisting of:

- **Tasks Repository** (stand-alone functions) that can constitute a discrete action. Tasks are invoked as a Lambda function or Service. They have a common format for payload and output to support chaining.
- **Orchestration engine** (AWS Step Functions) that controls invocation of Tasks via the recipe. Engine handles fault tolerance and scaling.
  - **Database** to store task status, logs, and other system state information
- **Recipe(s)** or Configuration file(s) that define the ingest, processing, publication, and archive operations
- **Dashboard** to allow operators to create and execute recipes. The dashboard also allows operators to check the status of executions, track errors, etc.

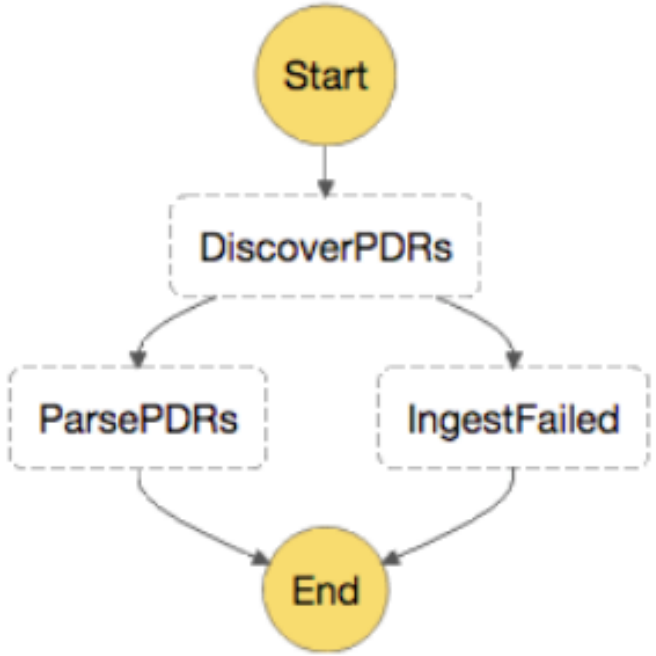


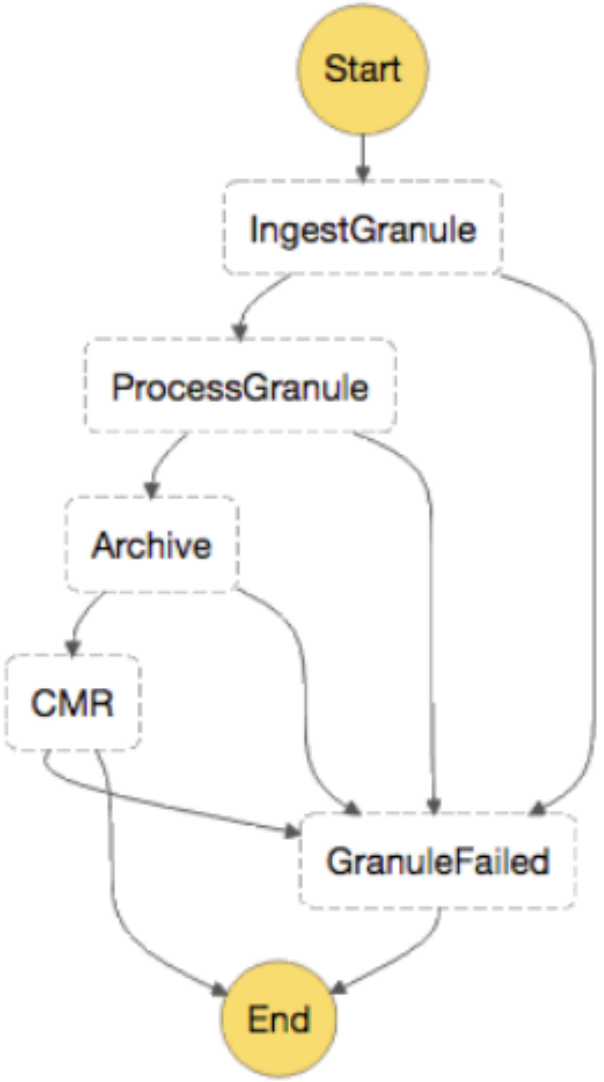

EarthData Users
API Gateway
URS authentication
for access to granules
Archived granules S3
Data Pipelines
Check granule integrity
Process granules
Send metadata to CMR
Pipeline Trigger
Detect new files
Instantiate pipelines
Ingest granules S3
Populated directly by DAAC
or by source data acquisition app
DAAC Admins
Workflow Engine
View processing status and logs
Configure pipelines
Estimate costs


High Level view
SIPS
Ingest
DAAC
Processing
Distribute
User
Covered by Cumuuls
Ingest (cumulus-factory)
Source
Scheduler
Discover
Ingest Dispatcher
Sync
cumulus-api
The Dashboard
Collections Table
Granules Table
Processing the Granules with the Recipe (cumulus-factory)
Get the Recipe
New Granule Check
The Dispatcher
Step 1
Step 2
Step 2
Legends
S3
Queue
DB Table
Docker
Lambda Func
Distribution App
EarthData Login
User
Distribution


Scheduler
Collections Table
AWS Step Functions
Dashboard
Workflow Execution
Discover Granules
Sync
Processing Step 1
Processing Step 2
Submit to CMR
Workflow Data / Service Interaction
Provider
CMR
Temporary Storage
Granule Storage
Distribution
Legend
Cross-DAAC Component
DAAC-Specific Component

Step Functions





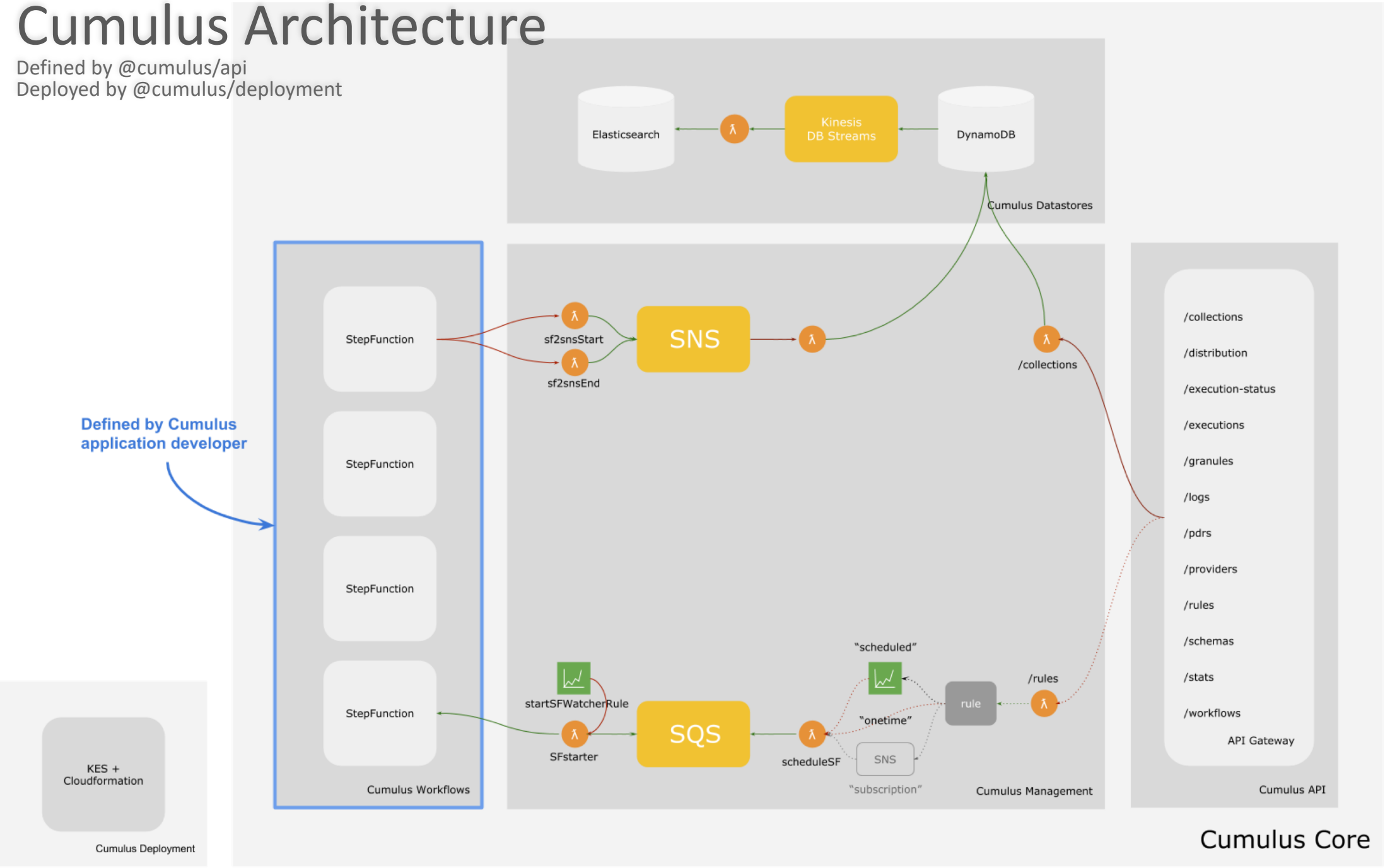
Cumulus is a collection of resources

These resources are:

- **@cumulus/deployment:** A node module for creating a Cumulus deployment. A Cumulus deployment is comprised of 4 AWS Cloudformation stacks. Each Cumulus application will have it's own cloudformation stacks.
- **@cumulus/api:** A node module for deploying the Cumulus API and other AWS resources required to run **Cumulus workflows**.
- **Node modules for common tasks** to be run as part of Cumulus Workflows, for example @cumulus/parse-pdr
- **cumulus-dashboard:** Code to generate and deploy the dashboard for the Cumulus API.

Cumulus Architecture

Defined by @cumulus/api
Deployed by @cumulus/deployment



Dashboard

0	7	0	51.79s	0	14	4
Errors	Collections	Granules Processed in the Past Hour	Average Processing Time	Pending Tasks	Running Tasks	Queued Messages

Granules Updated Today (121K)

0	45	0	0
Granules Ingesting	Granules Processing	Granules Updating CMR	Granules Archiving
121K	93		
Granules Completed	Granules Failed		

3. Workshop on "Enabling Analytics in the Cloud for Earth Science Data"

Rahul Ramachandran, Chris Lynnes, Andy Bingham, Brandi Quam
NASA MSFC, NASA GSFC, NASA LARC, NASA JPL

Ramachandran, R., Lynnes, C., Bingham, A., & Quam, B. (2018). Enabling Analytics in the Cloud for Earth Science Data. In *NASA Workshop*. Annapolis, MD: NASA.

About the Workshop

- The NASA Earth Science Data Systems Program sponsored a workshop on 21-23 February 2018 where the participants discussed the convergence of Big Data, Cloud Computing, and Analytics.
- These discussions clustered around six themes.

Strategic Alignment of Cloud Analytic Efforts

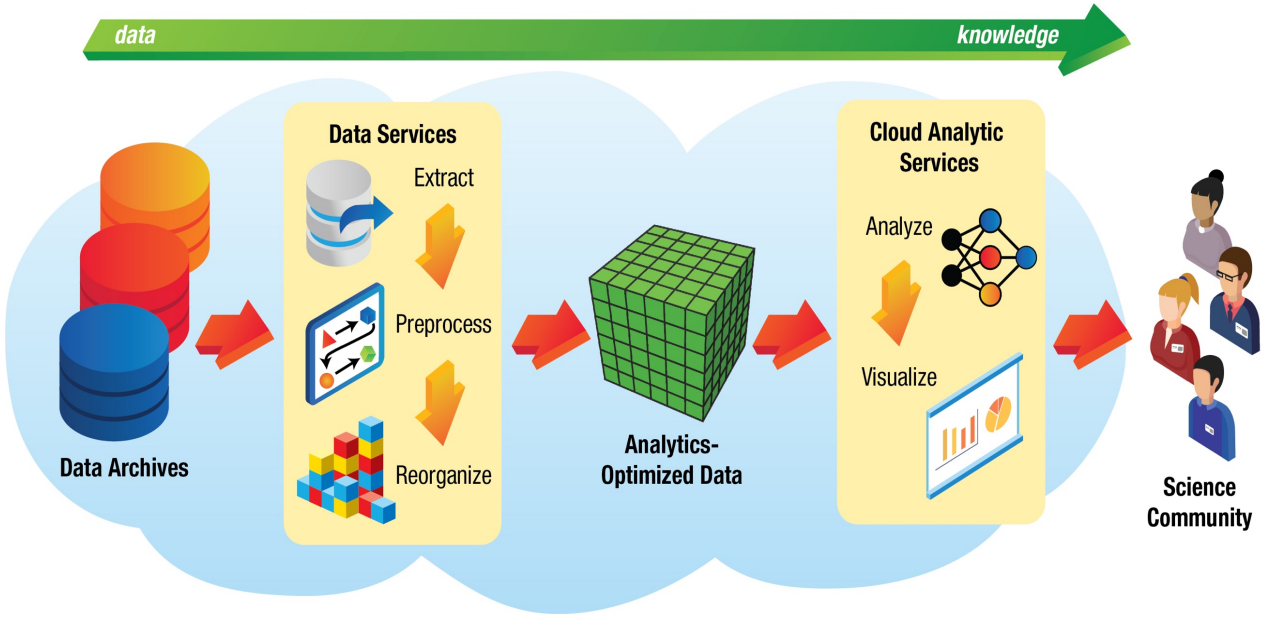
- Closer coordination among NASA programs and awareness of others' initiatives, priorities, and investments are necessary to prevent duplication of effort

- Partnerships with the private sector to leverage existing and future commercial cloud analytic platforms are recommended

Reference Architecture for Cloud Analytics

- Community has not yet developed a consensus around a single architecture or even common terms and definitions.
- The adoption of a service-based architecture will alleviate the need for a common architecture
- One area of agreement:
  - ***Need to build intermediate cloud data stores*** to create views into the data archive that are optimized for fast and scalable data analytics.

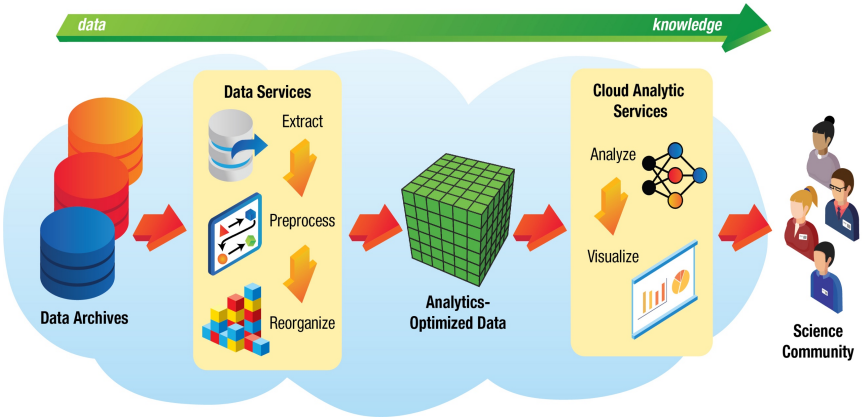
Analytics-Optimized Data



- Identified "Analytics Optimized Data Stores" (AODS) as a solution to Big Data Volume and Variety challenges
- AODS are data stored to minimize the need for data-wrangling for a large user community which enable fast access
- AODS are optimized, cost-effective storage structures enabling iterative queries relevant to users
- Building AODS to enable new analytic tools and services is viewed as a best path forward.

Analytics-Optimized Data

- **Specific Goals:**
  - Is a subset of a DAAC holding
  - Designed to serve a need for a specific community



- **Data Preprocessing Criteria:**
  - Data is preprocessed using consistent and well known standard algorithms and can cover a spectrum of steps:
    - Subsetting, coregistration, calibration etc.
    - Coregistration is required
  - Stores a usable lineage capturing any transformations

- **Data Engineering Criteria**
  - Restructured and/or distributed to enable data parallel computation analysis needs of the users (query optimized)
  - Has an index to enable fast access
  - Storage solution carefully considers cost/performance trade-off

Reuse of Cloud Analytics-Related Service

- Cloud computing model of reuse via services should be extended to encompass future cloud analytics-related services
- Significant capabilities should be constructed as services to enable reuse.
- Establishment of Community Best Practices in architecting new systems/tools based on services is recommended.

Wider Deep Learning (DL) Adoption

- Challenge for wider adoption of DL is no longer the lack of usable algorithms, tools, or compute resources but ***rather the dearth of sufficient training (labeled) data in ES***.
- Access to training data would entice DL practitioners to tackle ES problems.
- Research and Development efforts should be developed to create new strategies for generating and distributing labeled training data in ES

Cloud Adoption

- Cloud adoption by the user community is not a foregone conclusion
- Efforts should be directed toward fostering that adoption.
- Training programs for end users to learn and understand the realm of possibility in cloud analytics should be developed.

4. Deep Learning Applications

4.1 Tropical Cyclone Intensity Estimation Using Deep Convolutional Neural Networks

Manil Maskey, Dan Cecil, Rahul Ramachandran, and JJ Miller
NASA Marshall Space Flight Center

Pradhan, Ritesh, Ramazan Aygun, Manil Maskey, Rahul Ramachandran (2017). Tropical Cyclone Intensity Estimation Using Deep Convolutional Neural Network. *IEEE Transactions on Computational Imaging, Special is*.

Overview

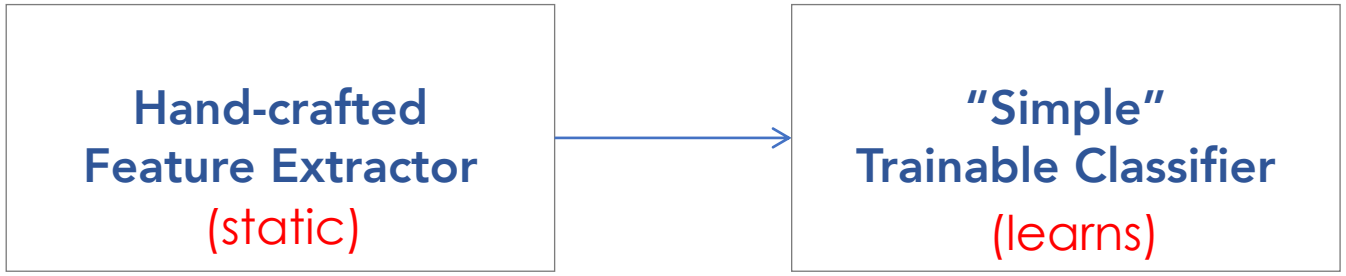
- Deep Learning and Convolutional Neural Network
- CNN for Tropical Cyclone Intensity Estimation
- Preliminary results
- Work in progress

Deep Learning

- A subfield of machine learning
- Algorithms inspired by function of the brain
- Scales with amount of training data
- Powerful tool without the need for feature engineering
- Suitable for many Earth Science applications

Traditional Image Classification Approach

- Image Features: Color, Texture, Edge histogram,…
- "Shallow" architecture
- Experts define features

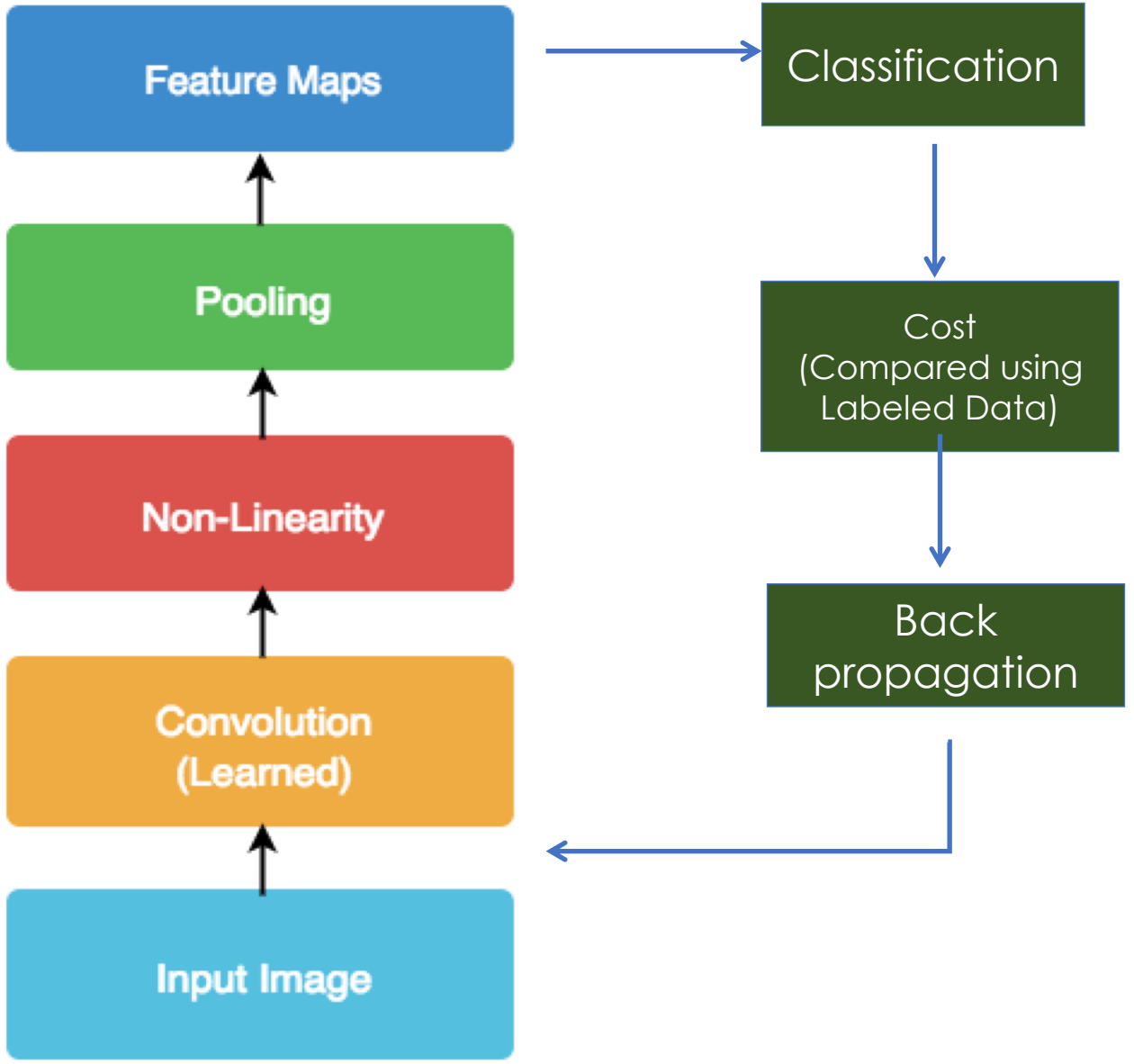


“DEEP” Architecture

- Features are key to recognition
- Deep Learning
  - Hierarchical Learning
  - Modeled after human brain
  - Process information through multiple stages of transformation and representation

Convolutional Neural Network

- Input image – labeled training data
- Convolution Layers – filters are applied across input images (start with random filters)
- Non-linearity – a bias function so that the network is not remembering but rather generalizing
- Pooling – subsampling of the output so that the images do not grow exponentially
- Final output images are passed through a traditional neural network for classification
- Classification results are compared using a loss function to determine error
- Based on error the weights and filters are adjusted using gradient descent
- Iterate the process until the error is below some threshold



Convolutional Layer

Input (7x7), pad of 1

0	0	0	0	0	0	0
0	1	2	1	0	1	0
0	2	1	1	2	1	0
0	1	2	1	2	2	0
0	2	2	2	1	0	0
0	0	1	1	1	2	0
0	0	0	0	0	0	0

Kernel (3x3), stride of 2

1	1	-1
0	-1	1
-1	0	1

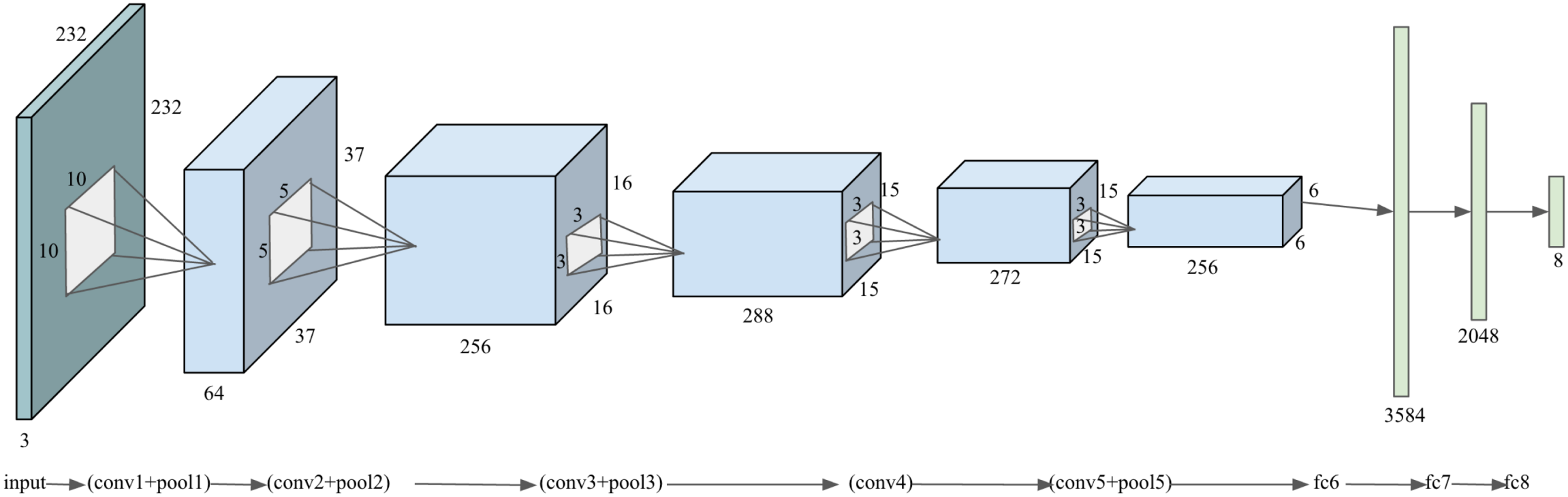
Output (3x3)

2	0	-3
4	0	0
1	3	-1

- Stride (s)
  - Jump/step with which filters move across width/height of input volume
- Padding (p)
  - Amount of wrapping used in input
- Output size (W_o) = **$(W_i - k + 2*p)/s + 1$**

2D Convolution, Single Slice

Network Architecture



Tropical Cyclone Intensity Estimation

- **The Dvorak Technique**
  - Vernon Dvorak (1970s)
  - Satellite-based method
  - Cloud system measurement
  - Development patterns corresponds to T-number
- **Deviation-Angle Variation Technique (DAVT)**
  - Piñeros et al. (2008)
  - Variance for quantification of cyclones
  - Calculates using center (eye) pixel
  - Directional gradient statistical analysis of the brightness of images

Issues

- Subjective/Uncertainty
- Lack of generalizability
- Inconsistency
- Complexity

15 UTC 10 Oct 17 NHC advisory on Tropical Storm Ophelia

*"Dvorak intensity estimates range from T2.3/**33 kt** from **UW-CIMSS** to T3.0/**45 kt** from **TAFB** toT4.0/**65 kt** from **SAB**. For now, the initial intensity will remain at 45 kt, which is an average of the scatterometer winds and all of the other available intensity estimates."*

Observation:
Two human experts at TAFB and SAB differed by 20 knots in their Dvorak analyses, and the automated version at the University of Wisconsin was 12 kt lower than either of them!

Can we objectively predict wind speed from satellite images?

Data

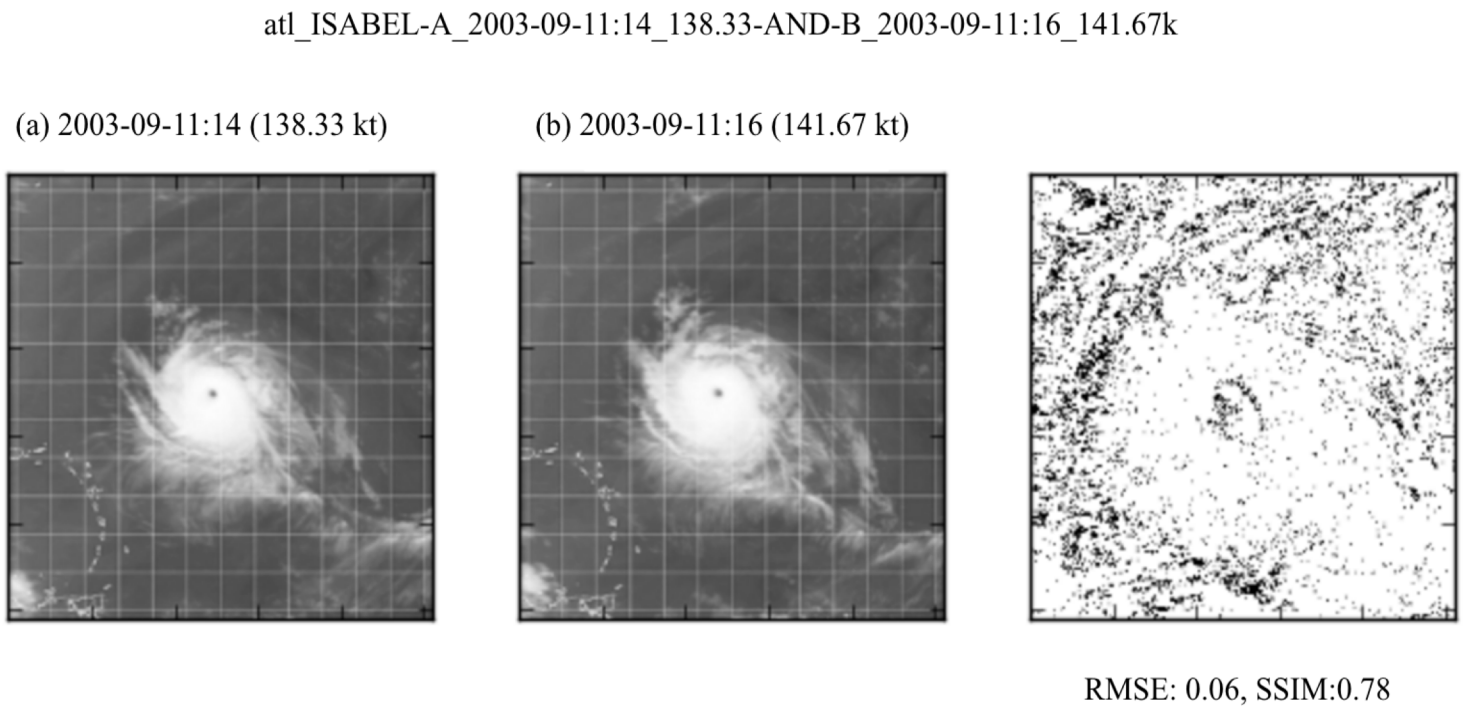
- **Images**
  - US Naval Research Laboratory (http://www.nrlmry.navy.mil/tcdat)
  - From 1998 to 2014
  - Images at 15 minutes interval
- **Cyclone data**
  - National Hurricane Center (http://www.nhc.noaa.gov) (HURDAT and HURDAT2)
  - Hurricane Research Division (http://www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html)
  - Every 6 hours
- **98 cyclones collected over Pacific and Atlantic regions**
  - 68 from Atlantic
  - 30 from Pacific

Storms

Region/Basin	Year	Cyclones
Atlantic	1998	Mitch
	2003	Isabel
	2004	Ivan
	2005	Emily, Katrina, Rita, Wilma
	2007	Dean, Felix
	2010	Alex, Bonnie, Colin, Danielle, Earl, Fiona, Five, Gaston, Igor, Julia, Karl, Lisa, Matthew, Nilcole, Otto, Paula, Richard, Shary, Tomas, Two
	2011	Arlene, Bret, Cindy, Don, Emily, Franklin, Gert, Harvey, Irene, Jose, Katia, Lee, Maria, Nate, Ophelia, Philippe, Rina, Sean, Ten
	2012	Alberto, Beryl, Chris, Debby, Ernesto, Florence, Gordon, Helene, Isaac, Joyce, Kirk, Leslie, Michael, Nadine, Oscar, Patty, Rafael, Sandy, Tony
	2014	Edouard
Pacific	2002	Elida, Fausto, Hernan, Kenna
	2005	Jova, Kenneth
	2006	Bud, Daniel, Ioke, John, Lane
	2007	Flossie
	2008	Hernan, Norbert
	2009	Felicia, Guillermo, Jimena, Rick
	2010	Celia, Darby
	2011	Adrian, Dora, Eugene, Hilary, Jova, Kenneth
	2012	Bud, Emilia, Miriam, Paul

Data Augmentation

- Interpolate to increase even more
- 2 hours interpolated image differences

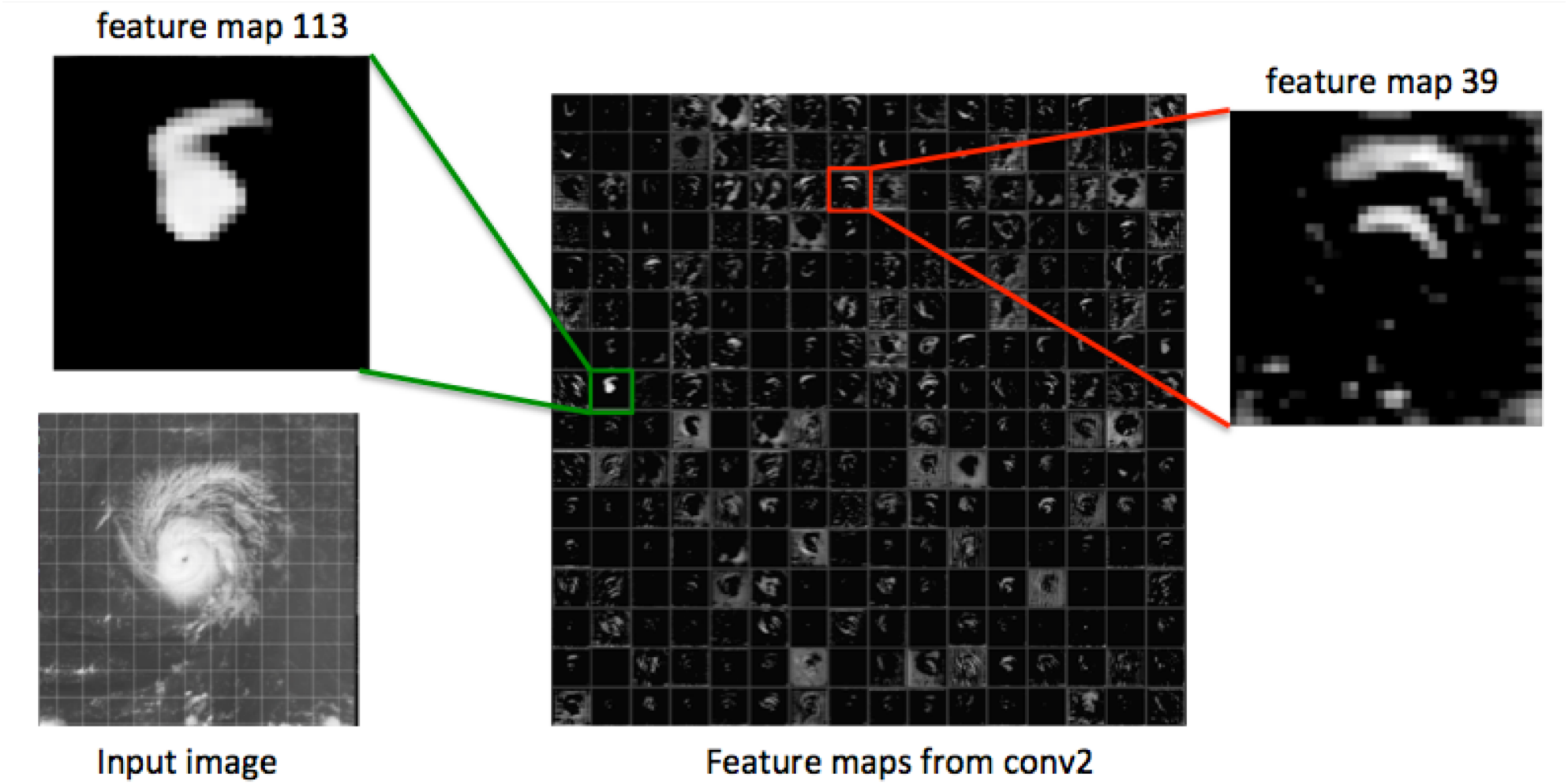


2 hour interpolated image differences

Training, Test and Validation

Hurricane Category	Train	Validation	Test	Total
H1	3314	1104	1816	6234
H2	1860	620	994	3474
H3	1848	616	992	3456
H4	1886	628	1032	3546
H5	603	201	306	1110
NC	126	42	54	222
TD	6363	2121	3576	12060
TS	9863	3288	5575	18726
Total	**25863**	**8620**	**14345**	**48828**

Visualization



Feature maps from second convolution

Initial Performance

- Model with around 90% of **validation** accuracy
- Tested against 14,345 test images (Atlantic + Pacific)
  - Confusion Matrix
  - Classification Report
  - Accuracy
  - RMS Intensity Error

Accuracy

- Top-1: exact-hits
- Top-2: exact-hits + 2nd-hits

	Total Counts	**Accuracy**
Top-1	11571	80.66%
Top-2	13695	95.47%

Category	**Total**	**Top-1**	**2^{nd} hit**	**Top-2**
NC	54	32	15	47
TD	3576	3174	364	3538
TS	5575	4838	665	5503
H1	1816	1235	432	1667
H2	994	614	215	829
H3	992	657	212	869
H4	1032	816	148	964
H5	306	205	73	278
Total	**14345**	**11571**	**2124**	**13695**

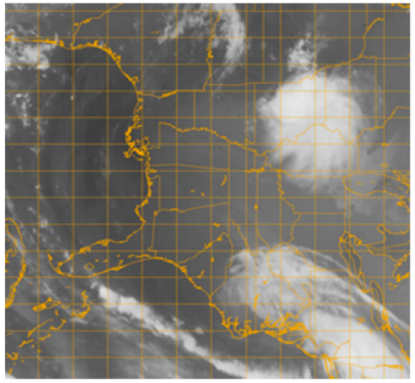
Error Metrics

- Our model
  - Across Atlantic and Pacific
  - Achieved RMSE of 9.19*kt*
- North Atlantic
  - Piñeros et al. (2011): 14.7kt
  - Ritchie et al. (2012): 12.9kt
- North Pacific
  - Ritchie et al. (2014): 14.3kt

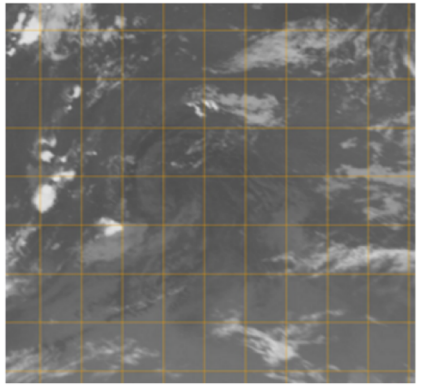
Category	RMSE	MAE
NC	10.14	6.19
TD	6.59	2.18
TS	7.68	2.71
H1	12.17	6.59
H2	12.43	6.82
H3	12.44	6.31
H4	10.50	4.09
H5	10.08	5.32
Total Average	**9.19**	**3.77**

Sample Correct Classifications

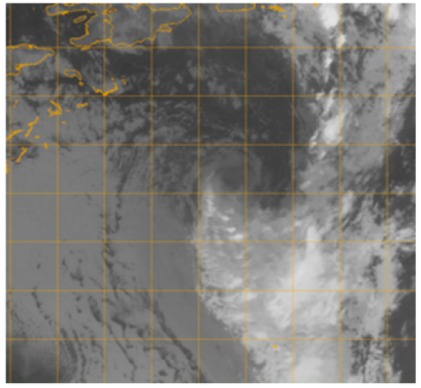
True Positives



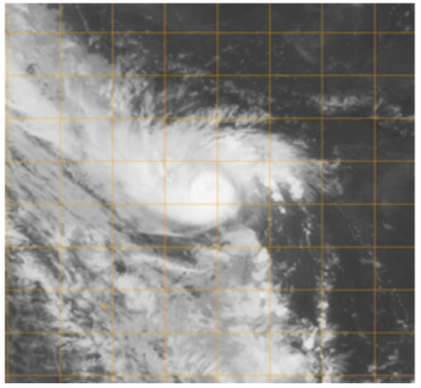
(a) NC: ['NC': 99.4]



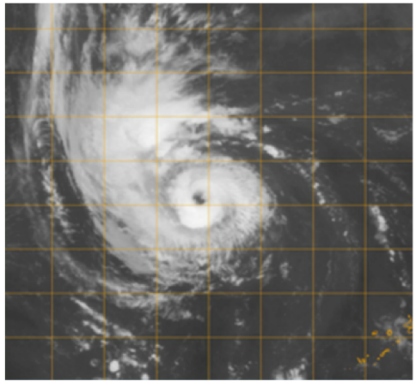
(b) TD: ['TD': 87.46]



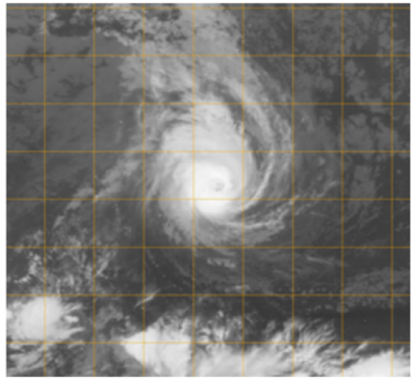
(c) TS: [TS: 100]



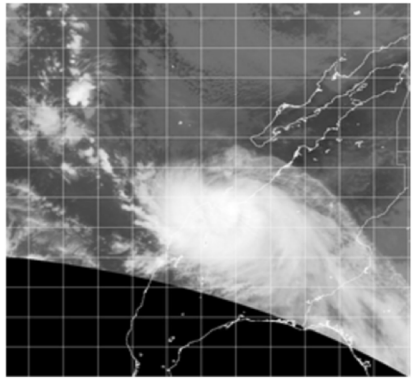
(d) H1: [H1: 56.8]



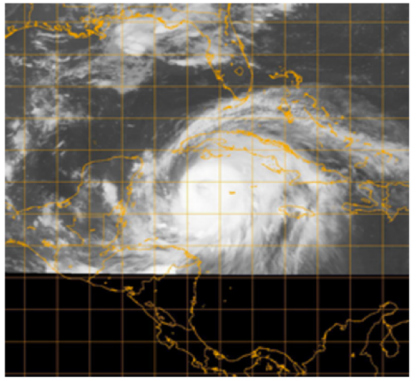
(e) H2: [H2: 78.54]



(f) H3: [H3: 95.73]



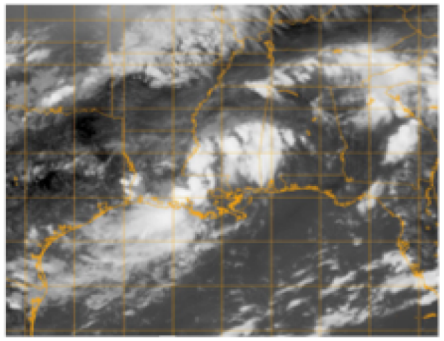
(g) H4: [H4: 86.04]



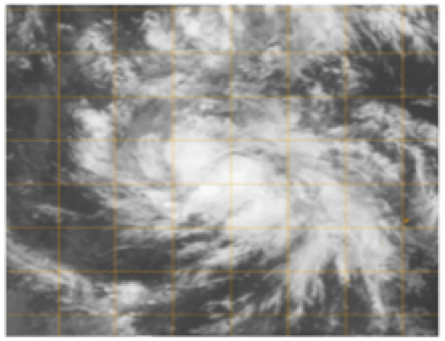
(h) H5: [H5: 58.26]

Sample Incorrect Classifications

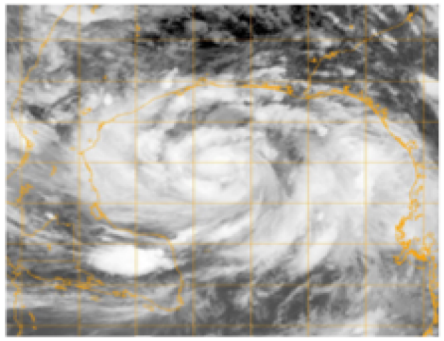
False Negatives



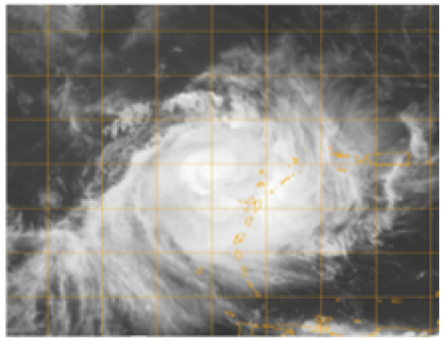
(a) NC: [TD --> 99.98] [TS --> 0.01]



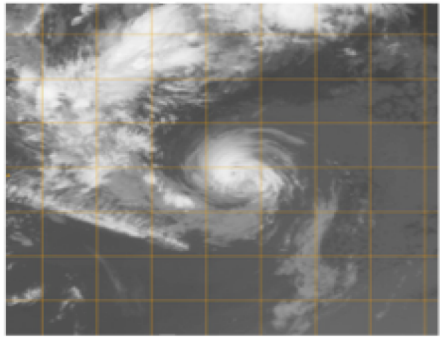
(b) TD: [TS --> 96.7] [H1 --> 3.03]



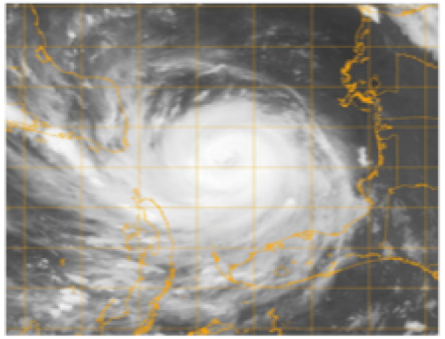
(c) TS: [H1 --> 97.93] [H2 --> 1.33]



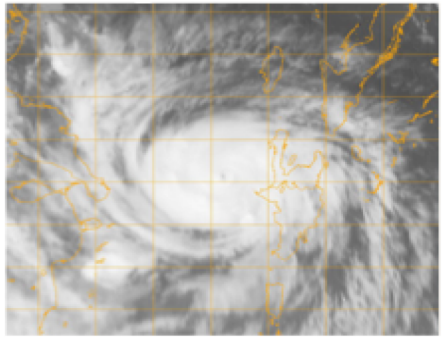
(d) H1: [H3 --> 61.31] [H2 --> 23.06]



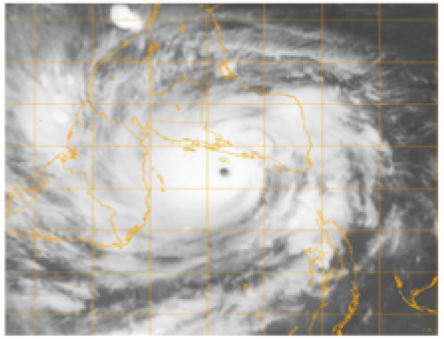
(e) H2: [TS --> 100.0] [H1 --> 0.0]



(f) H3: [H4 --> 97.32] [H5 --> 2.22]



(g) H4: [H2 --> 54.0] [H3 --> 36.79]



(h) H5: [H4 --> 99.71] [H3 --> 0.13]

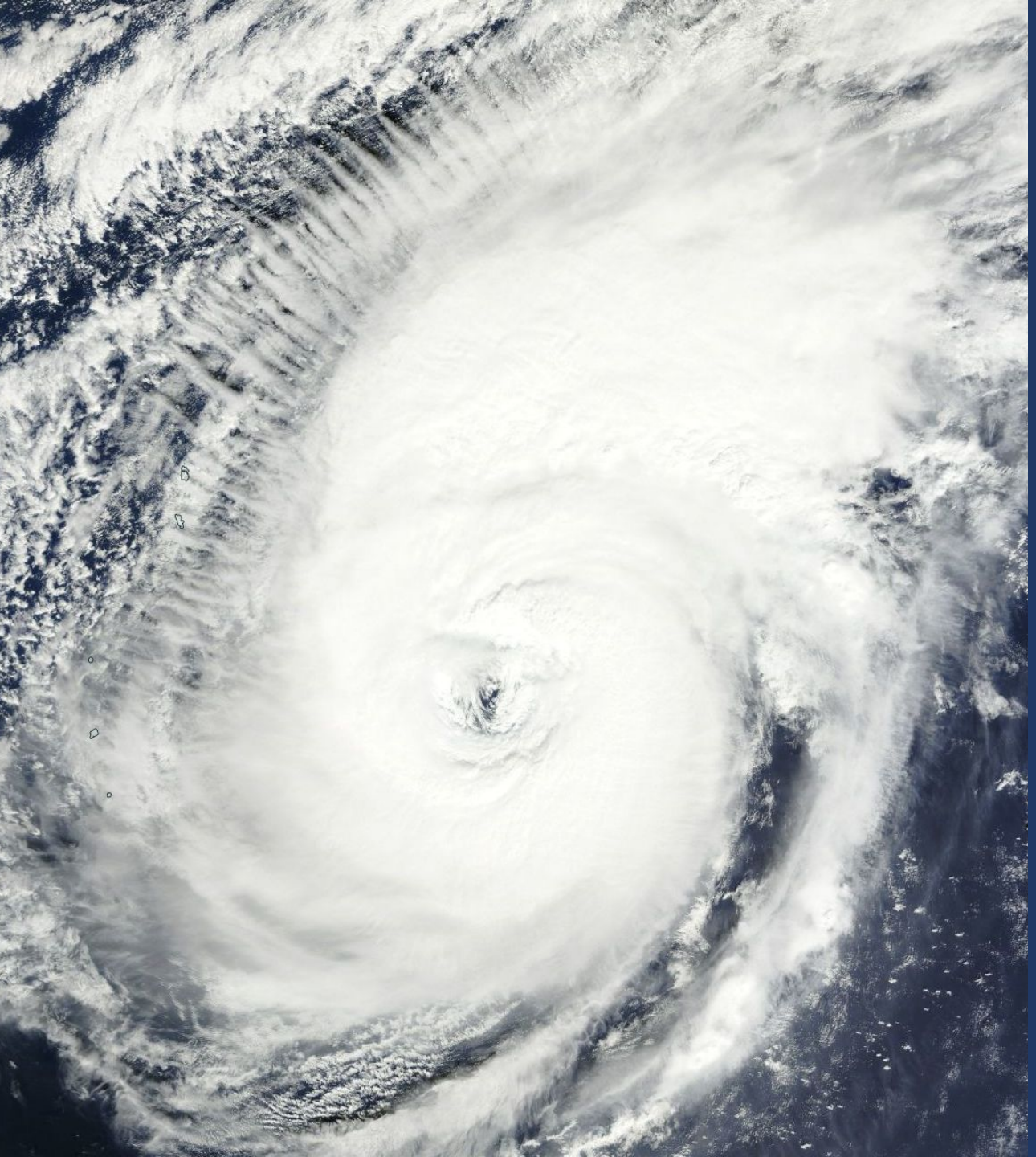
Work In Progress

- Hurricane intensity estimation portal
- Use of passive microwave dataset
- Use of atmospheric conditions

4.2 Detection of Transverse Cirrus Bands in Satellite Imagery Using Deep Learning

JJ Miller
NASA Marshall Space Flight Center

Miller, J.J., Udaysankar Nair, Rahul Ramachandran, Manil Maskey (2018). Detection of Transverse Cirrus Bands in Satellite Imagery Using Deep Learning. *Computers & Geosciences*.



Motivation

- Transverse cirrus bands (TCB) associated with aviation turbulence
- No global climatology studies on TCB occurrence
- No consensus on formation mechanisms
- Detailed information on TCB occurrence could be very useful information for the aviation industry
- Automated method to detect TCBs in satellite imagery would be very useful to operational forecasting
- Few applications of deep learning on identifying atmospheric phenomena
    - Only large scale phenomena

Research Questions

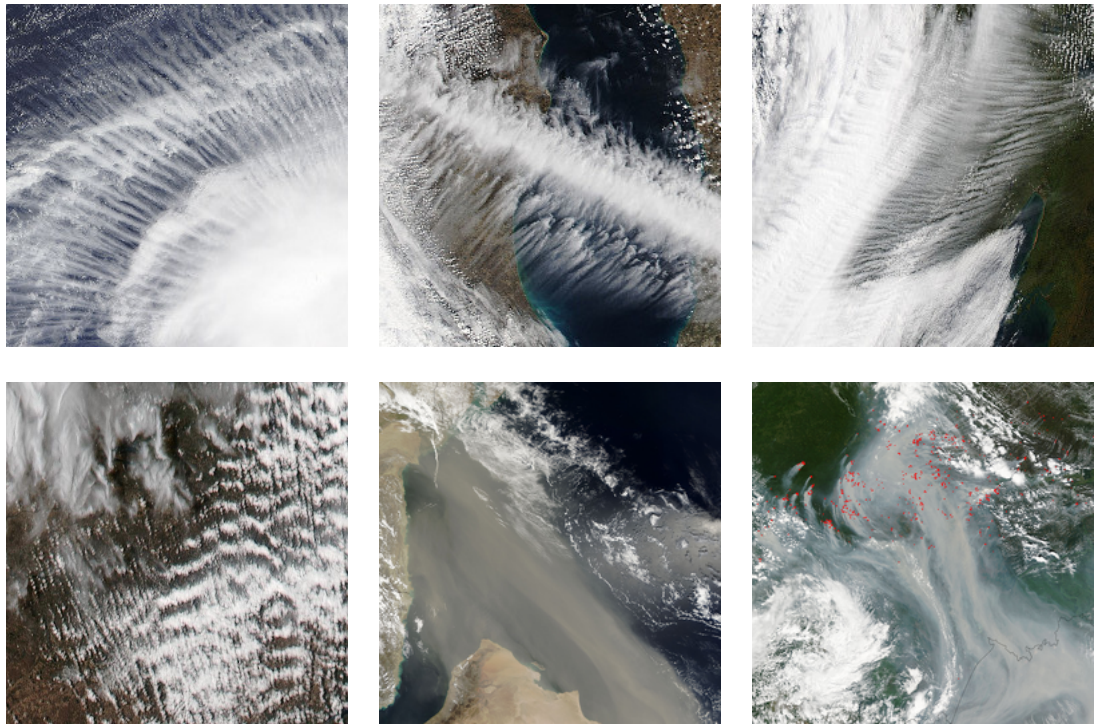
1. Can a Convolutional Neural Network (CNN) that was trained for ordinary image classification be retrained to detect transverse cirrus bands in true-color MODIS satellite imagery?
2. In which regions do TCBs most often form and do these regions change seasonally?
3. From what parent phenomena do TCBs most often form?

Transverse Cirrus Bands

- TCBs are irregularly spaced cirrus clouds that form perpendicular to a jet stream axis
- Observed in multiple different weather systems
  - Extratropical cyclones (ETC)
  - Mesoscale convective system (MCS)
  - Tropical Cyclones
  - Jet streaks
- Sometimes confused with gravity waves
- Pilots use them as a general "rule of thumb" for areas of CAT (Ellrod 1989)
- Majority of the studies focus on TCB association with MCSs and ETCs
- Sometimes referred to as "radial bands"

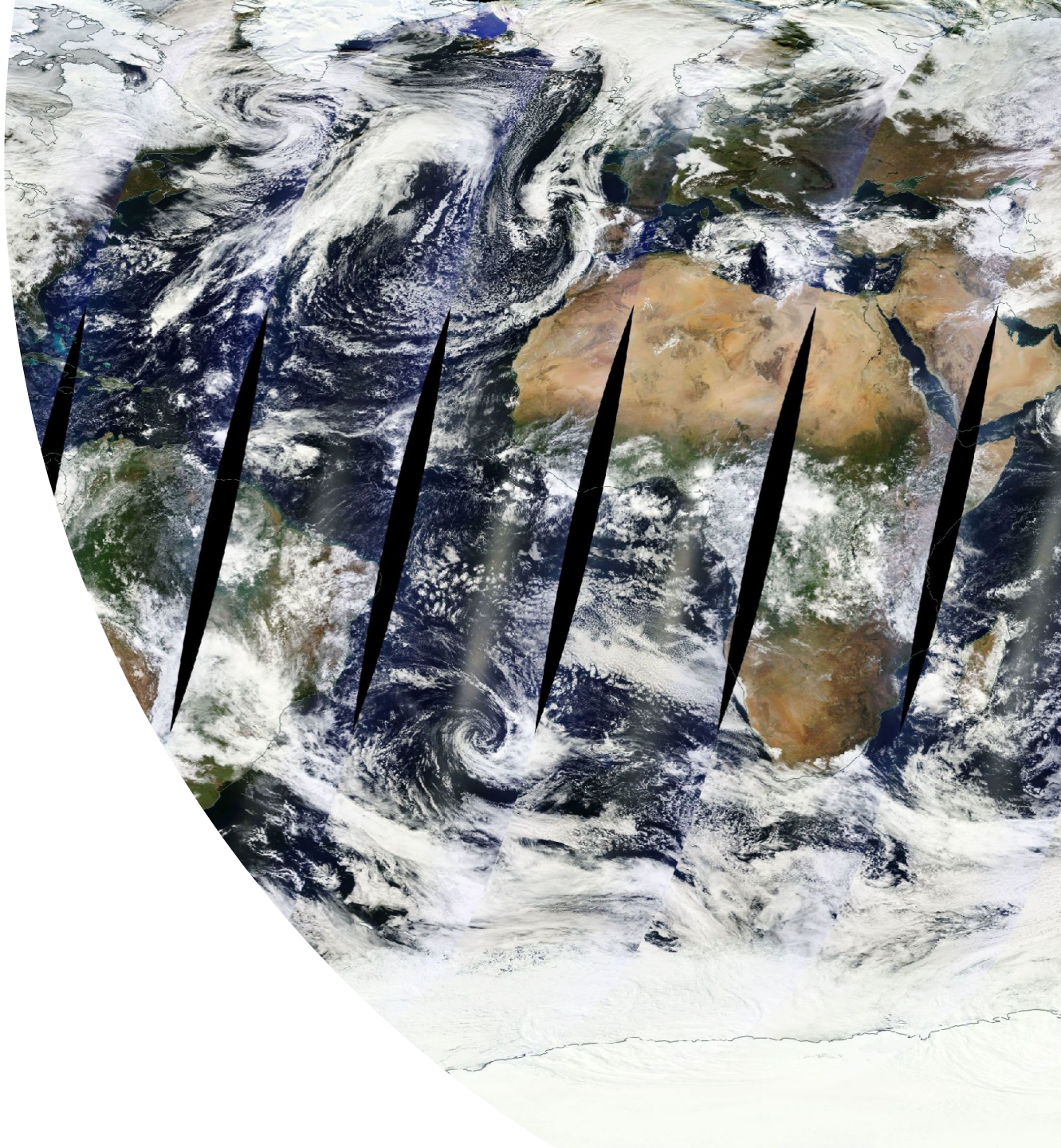
Training Data

- MODIS and VIIRS RGB true color composite
- Manually obtained from NASA Worldview and a pre-existing dataset (Maskey et al. 2017)
- Images with TCBs, images without TCBs, and images with features that look similar to TCBs
- Total of 5579 images
- Split into Training (60%), validation (20%), and test (20%) sets
- Images resized to 256x256



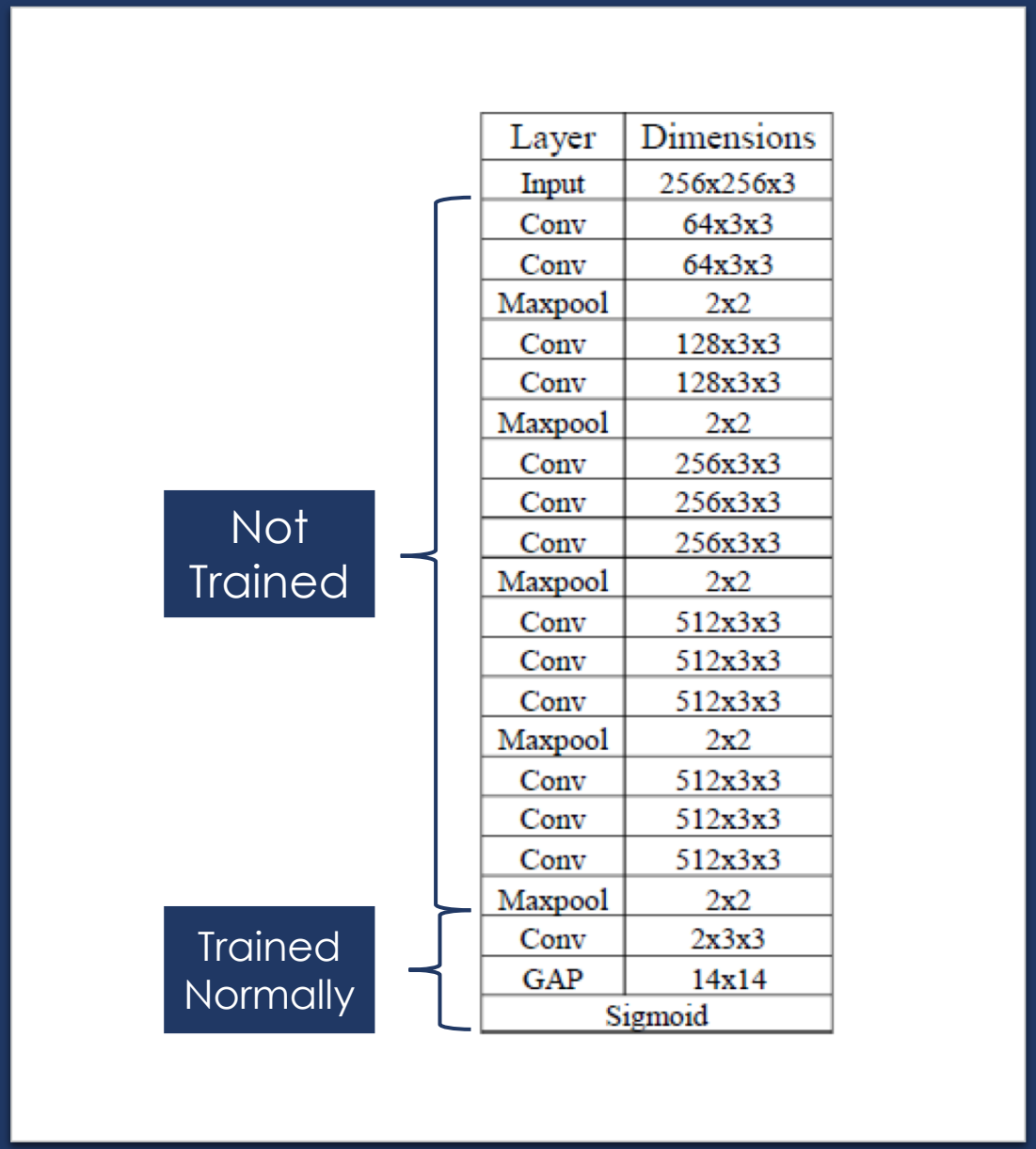
Climatology Data

- Three years (2013, 2014, 2015) of MODIS Terra RGB true color composite imagery tiles
- Obtained from Global Imagery Browse Service (GIBS)
- 1 km resolution
- Roughly 250,000 tiles per year
- Tiles with missing data were removed



Training Methodology

- Visual Geometry Group (VGG16)
  - Simonyan and Zisserman (2015)
- Modified for CAM application
  - Replaced fully-connected layers with 1 conv layer and 1 GAP layer
- Used transfer learning
- First 12 layers were fixed
- Built using Keras library
- Learning rate: $1x10^{-4}$
- Batch size: 46



Layer	Dimensions
Input	256x256x3
Conv	64x3x3
Conv	64x3x3
Maxpool	2x2
Conv	128x3x3
Conv	128x3x3
Maxpool	2x2
Conv	256x3x3
Conv	256x3x3
Conv	256x3x3
Maxpool	2x2
Conv	512x3x3
Conv	512x3x3
Conv	512x3x3
Maxpool	2x2
Conv	512x3x3
Conv	512x3x3
Conv	512x3x3
Maxpool	2x2
Conv	2x3x3
GAP	14x14
Sigmoid	

Performance Metrics

- **Accuracy**
  - The standard
- **F1-Score**
  - Often used in machine learning
- **Matthews Correlation Coefficient (MCC)**
  - Calculated from confusion matrix
  - Not influenced by unbalanced data
  - Ranges from -1 to 1

TP=True Positive; TN=True Negative
FP=False Positive; FN=False Negative

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{P(Precision)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{R(Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} = 2\frac{\text{P}\times\text{R}}{\text{P} + \text{R}}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

Network Evaluation on Test Set

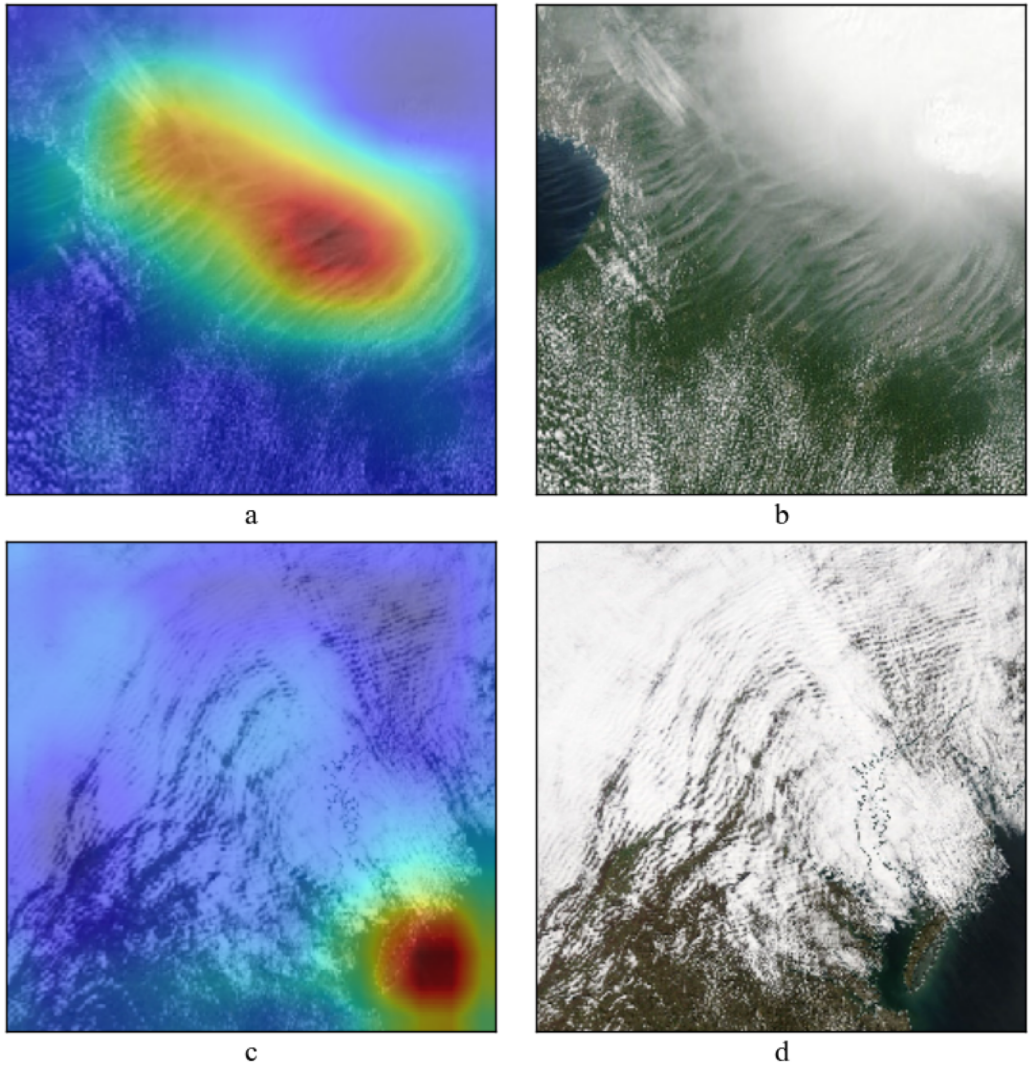
- Trained for 122 epochs
  - Validation loss: 0.11
  - Validation accuracy: 97.4%
- Accuracy on the test set: 94.4%
- CNN performance compared with RF classifier
  - Implemented using Scikit-Learn
  - 500 estimators
  - Use maximum of 500 features
  - Maximum node size of 20
- MCC shows the CNN performs much better than RF classifier

		Prediction		
		Bands	Not Bands	
Truth	Bands	296	33	329
	Not Bands	28	730	767
		324	772	

	MCC	F1	Accuracy
RF	0.54	0.79	80.7%
CNN	0.87	0.94	94.4%

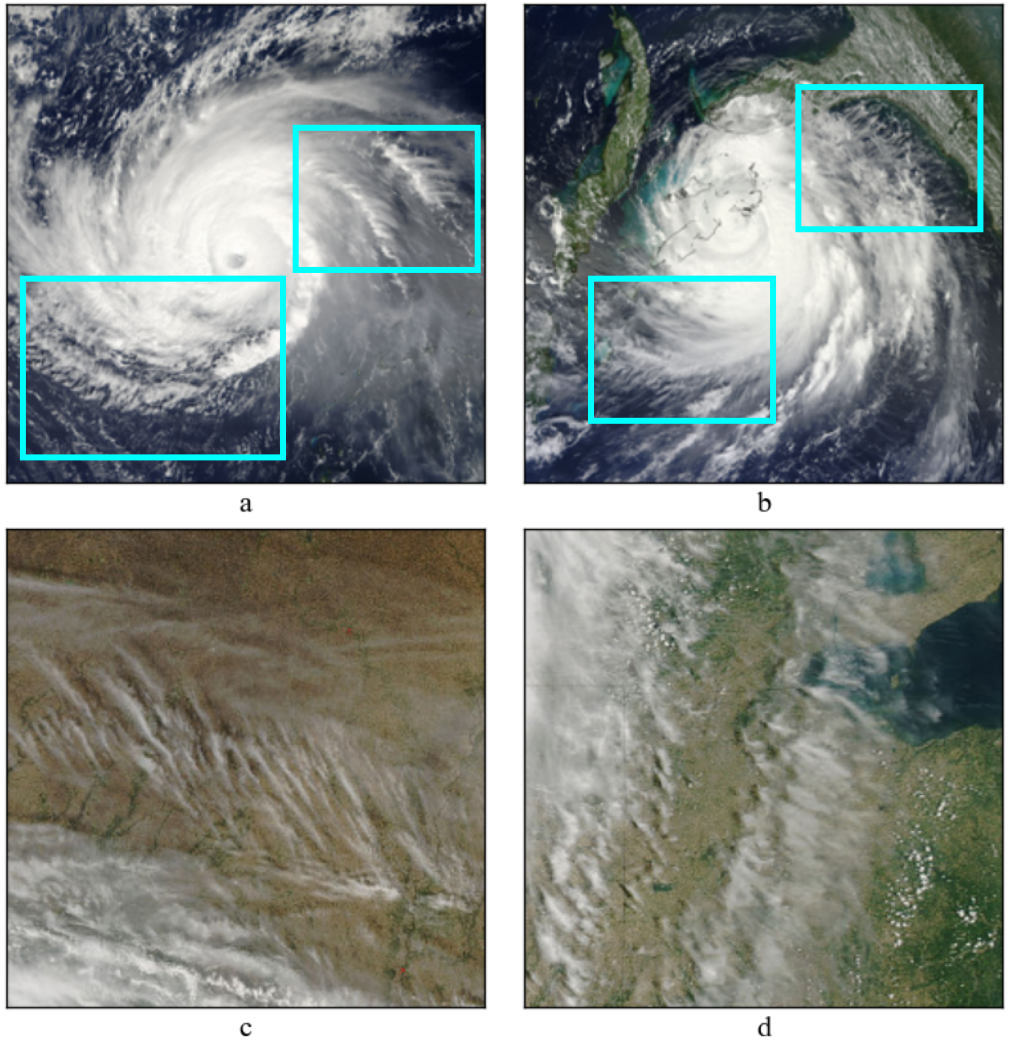
Class Activation Map Analysis

- CAMs are used to highlight the features of an image that are most discriminative for the classification (Zhou et al. 2015)
- Helps to better understand what the CNN is "looking at"
- CAMs suggest that the network appears to be "looking" for co-occurrence of TCBs with parent phenomena
- CNN does not appear to be fooled by the presence of gravity waves



Incorrectly Classified Images

- Network performed poorly with images where TCBs were small relative to entire image
- Network performed poorly when the bands appeared to be dissipating or poorly formed
- Likely caused by a lack of training data representing these types of TCBs

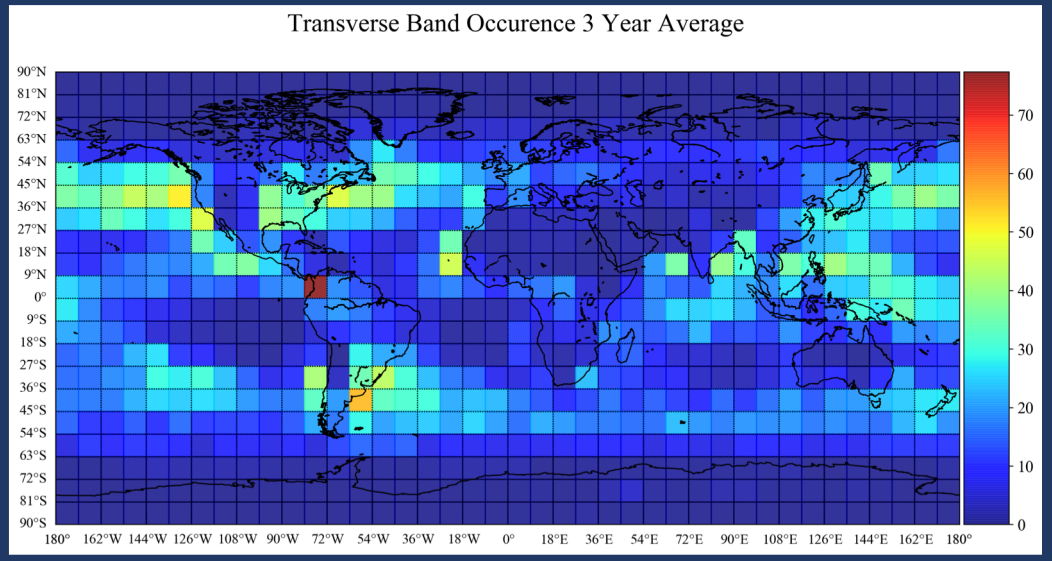


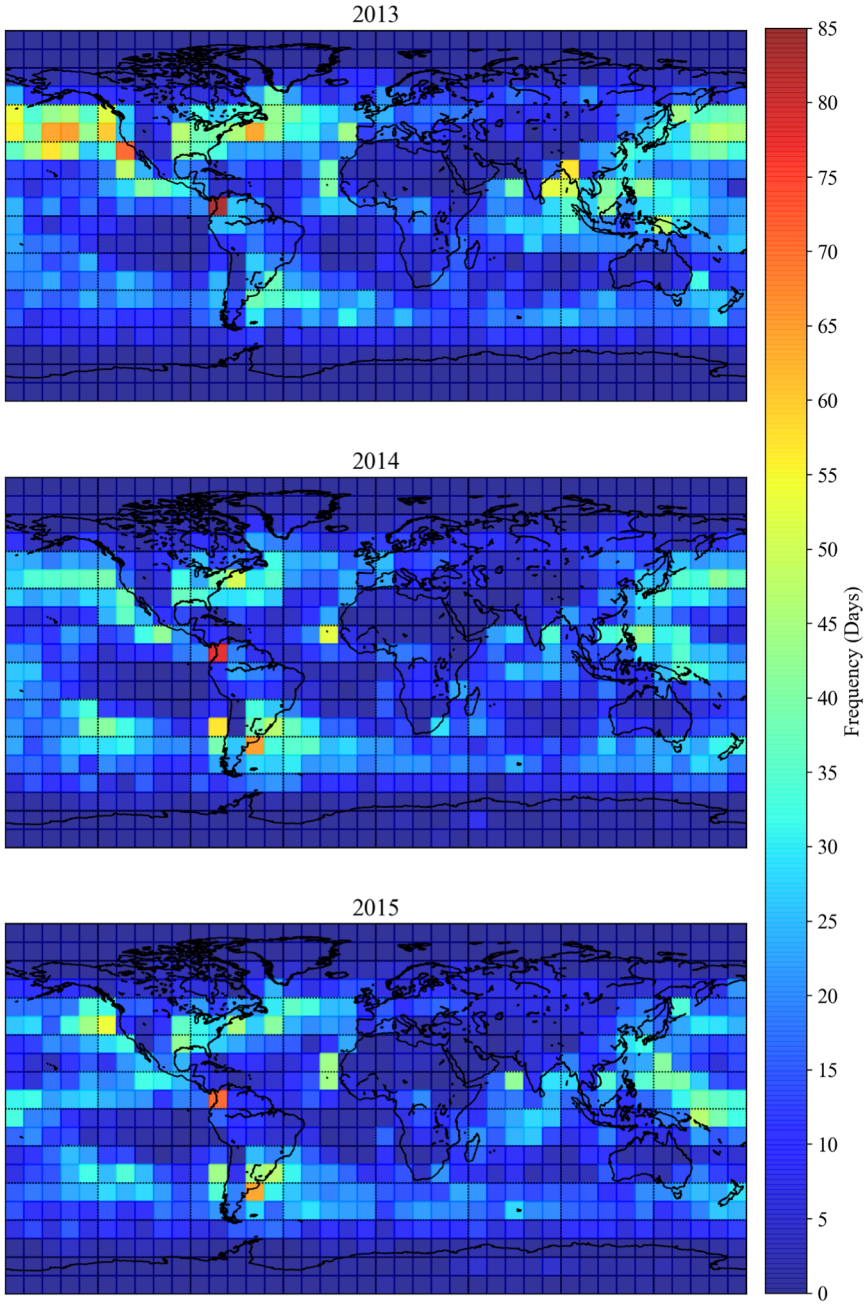


TCB Climatology

3 Year Climatology

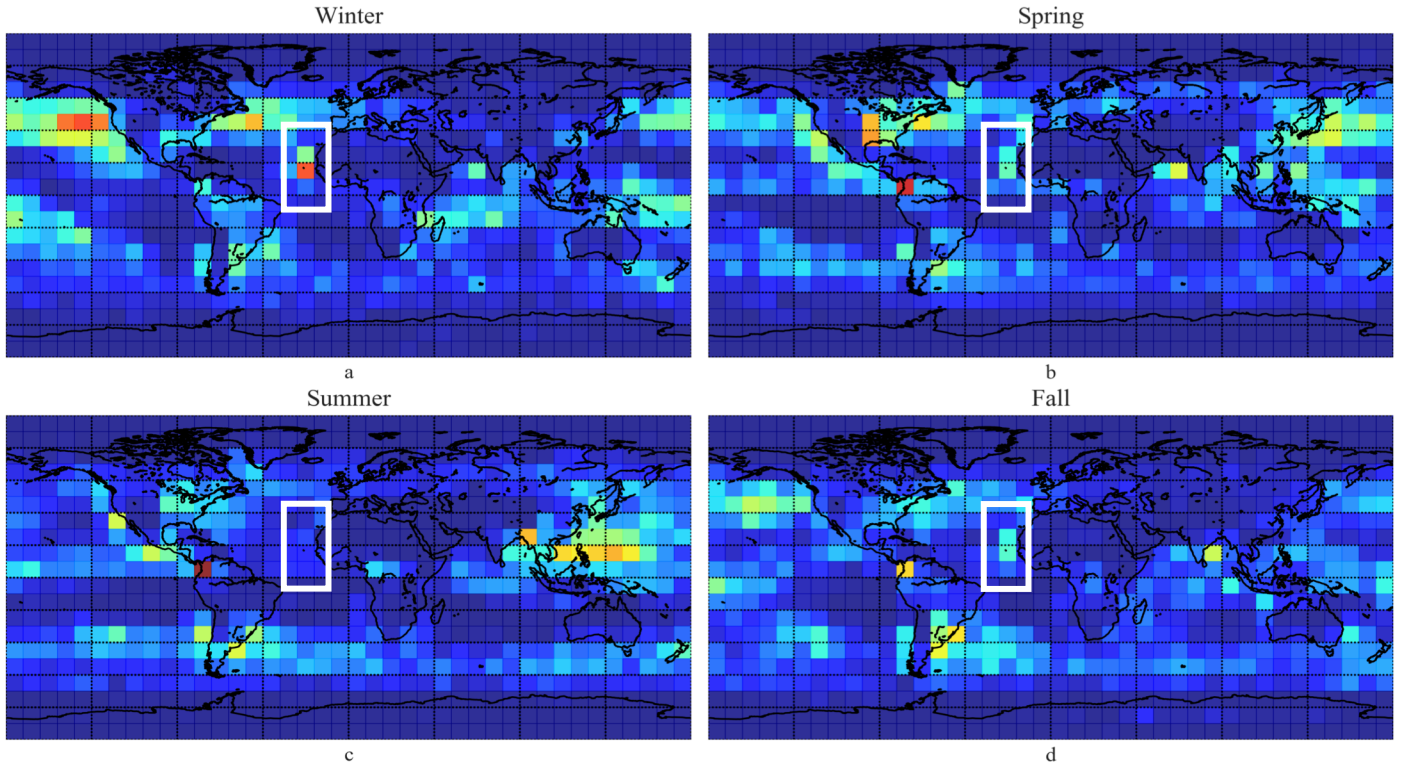
- Patterns are similar all three years
- Regions with the most TCB occurrence match up with northern/southern hemisphere jet streams
- Regions high occurrence:
  - Eastern North America
  - North Pacific
  - North Atlantic
  - Eastern South America
  - South Atlantic/Southern Ocean
- North American region supports findings from Lenz et al. (2009)





Seasonal Variation

- Winter (JFM), Spring (AMJ), Summer (JAS), Fall (OND)
- Patterns match up well with climatology study on jet-core occurrence (Pena-Ortiz et al. 2013)
  - North western coast of Africa in winter/spring



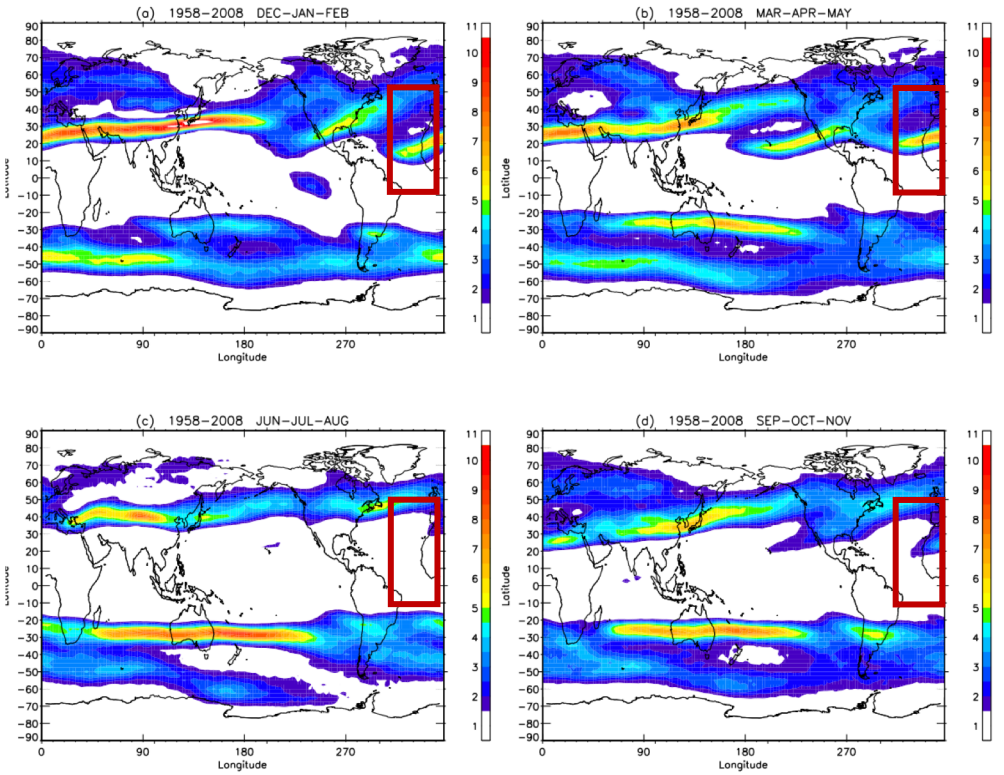
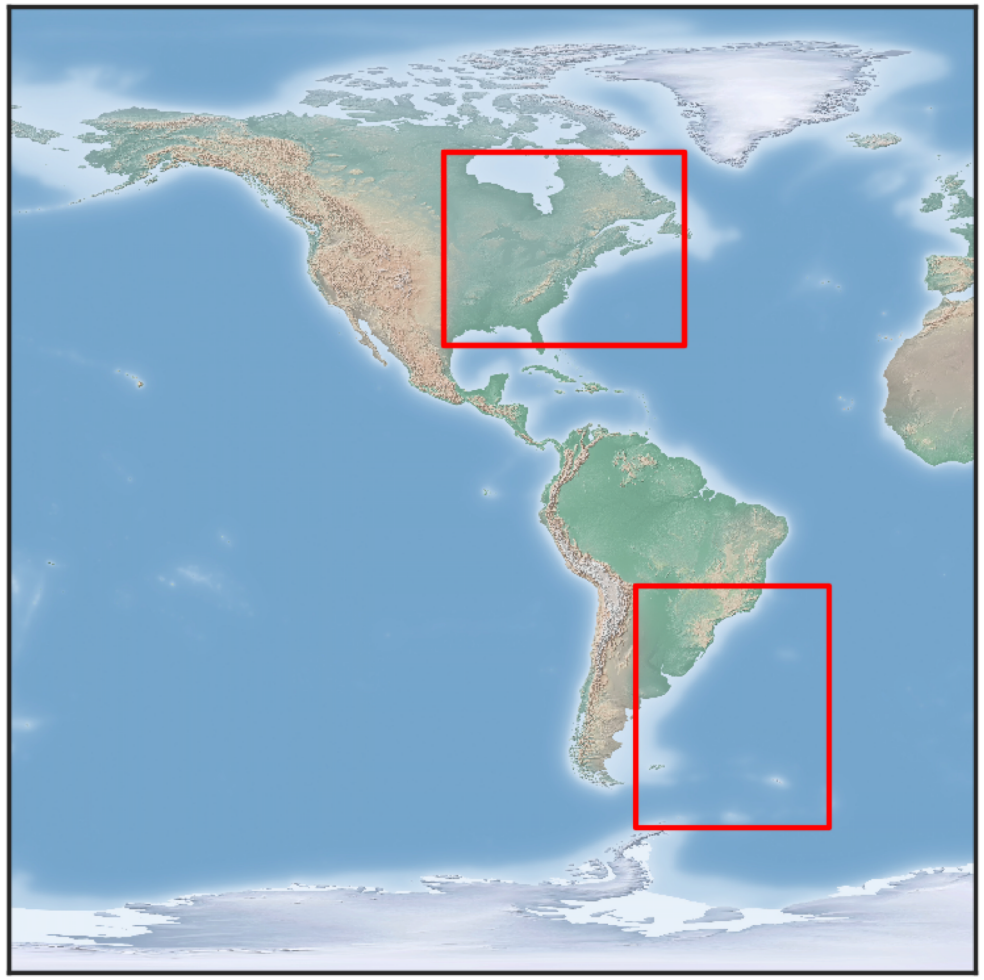
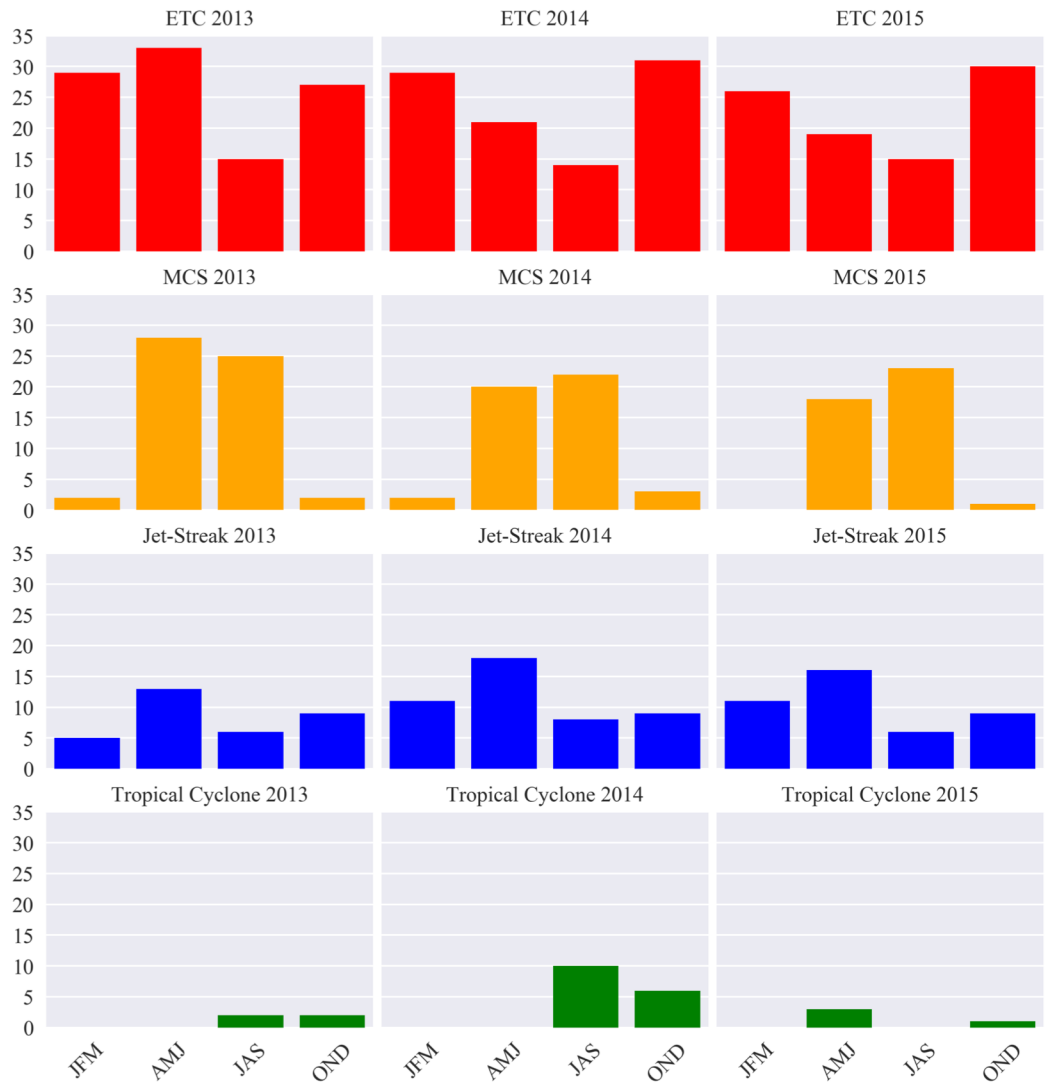


Figure 2 in Pena-Ortiz et al. 2013

Parent Phenomena Analysis

- North America, east of the Rocky Mountains
  - Access to historical surface and upper level maps for verification
- South America, east of the Andes Mountains
  - Similar phenomena to North America
  - Access to historical upper level maps for verification



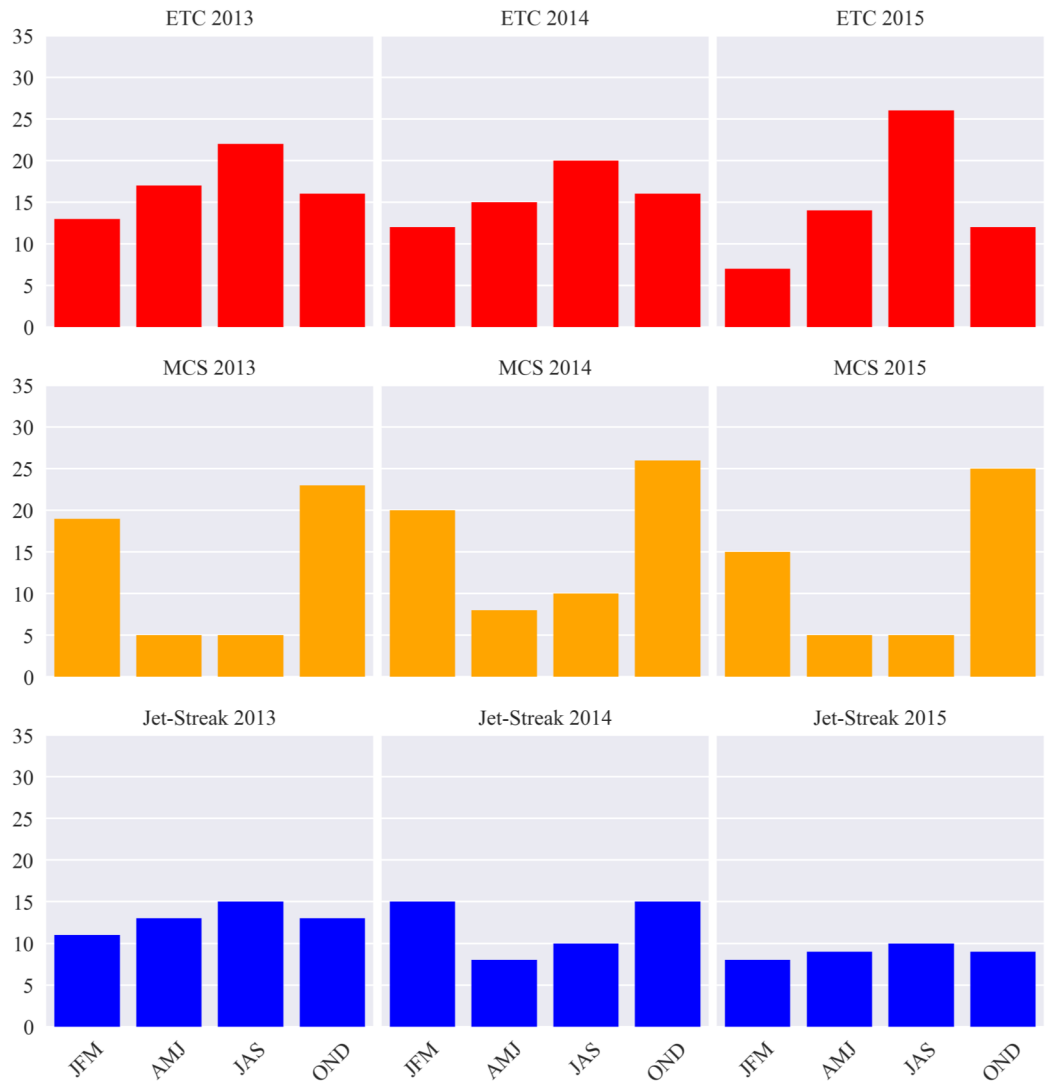


Eastern North America

- All 4 parent phenomena were identified
- Most frequent parent phenomena:
  1. ETCs: 96 tiles per year
  2. MCSs: 49 tiles per year
  3. Jet-streaks: 40 tiles per year
  4. Tropical cyclones: 8 tiles per year
- Season where TCBs are most common:
  1. Spring (AMJ) 63 tiles
  2. Summer (JAS) 49 tiles
  3. Fall (OND) 43 tiles
  4. Winter (JFM) 38 tiles
- ETCs account for 49.83%
  - Maximum in OND
- MCSs account for 25.17%
  - Maximum in JAS
- Jet streaks account for 20.86%
  - Maximum in AMJ
- Tropical cyclones account for 4.14%
  - Maximum in JAS

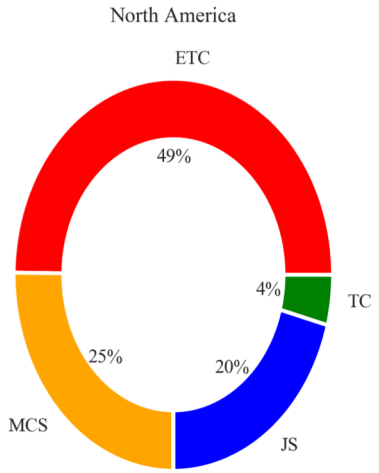
Eastern South America

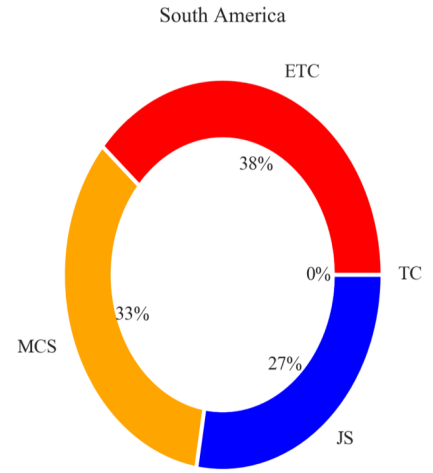
- 3 out of 4 parent phenomena identified
- Most frequent parent phenomena:
  1. ETCs: 63 tiles per year
  2. MCSs: 55 tiles per year
  3. Jet-streaks: 45 tiles per year
- Season where TCBs are most common:
  1. Spring (OND) 52 tiles
  2. Winter (JAS) 41 tiles
  3. Summer (JFM) 40 tiles
  4. Fall (AMJ) 31 tiles
- ETCs account for 38.62% of all tiles identified
  - Maximum in JAS
- MCSs account for 33.74% of all tiles identified
  - Maximum in OND
- Jet streaks account for 27.64% of all tiles identified
  - Maximum in OND

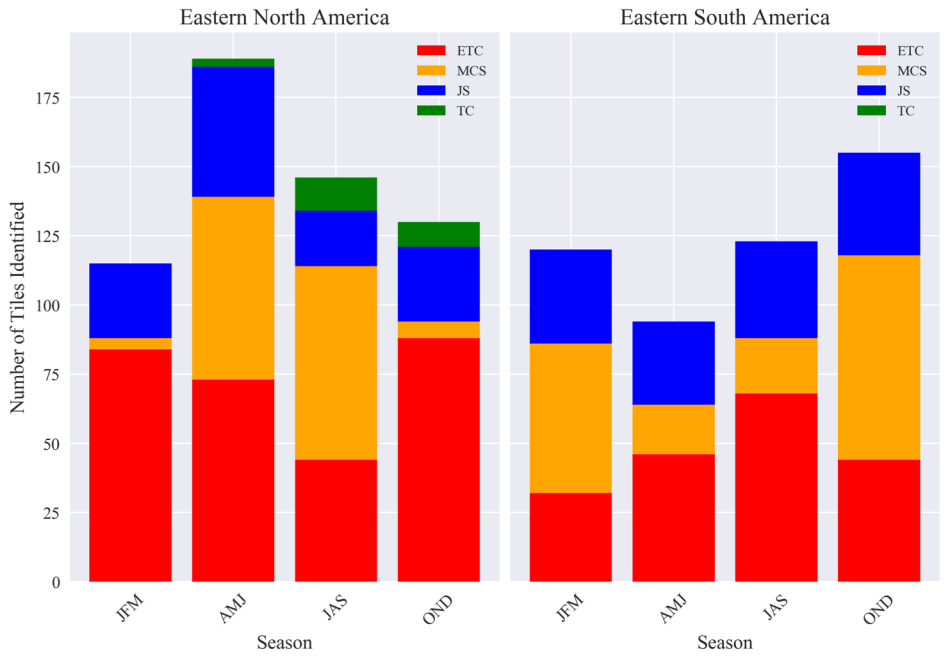


Region Comparison

- Both regions see a maximum in TCB occurrence in spring
- North America sees a minimum in the winter (JFM)
- South Americas minimum is in the fall (AMJ)
- Both regions experience more TCBs in the warm months
- Nearly half (49%) of all tiles in North America were associated with ETCs (38% in South America)
- South America experienced more tiles associated with MCSs
- North America had a higher total of tiles containing TCBs (580 vs 492)







4.3 Use our labeled data for your DL experiments

- Problem: Dearth of sufficient training (labeled) data in Earth science. Access to training data is required to entice DL practitioners to tackle Earth science problems.

- Solution: Provide labeled training data in Earth science focused on phenomena

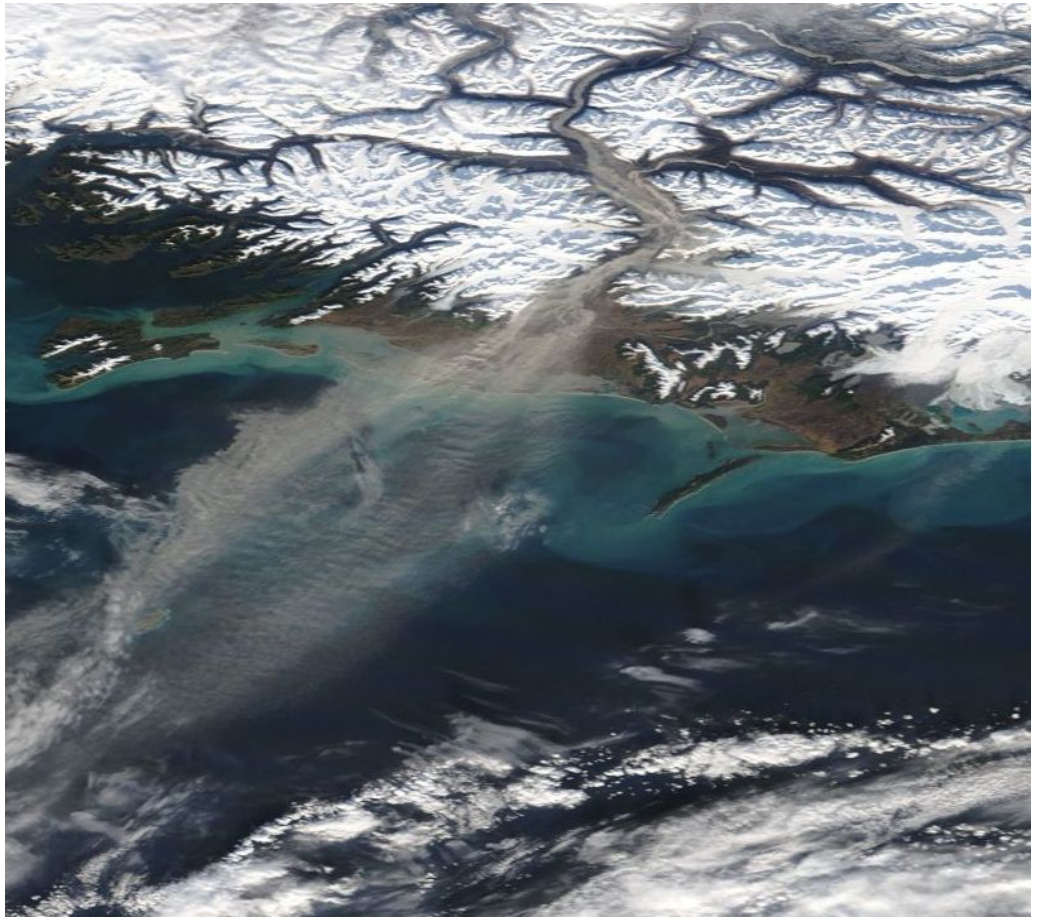
Maskey, M., Ramachandran, R., & Miller, J. J. (2017). Deep Learning for Phenomena-based Classification of Earth Science Images. *Journal of Applied Remote Sensing*.

Volcanic Ash clouds



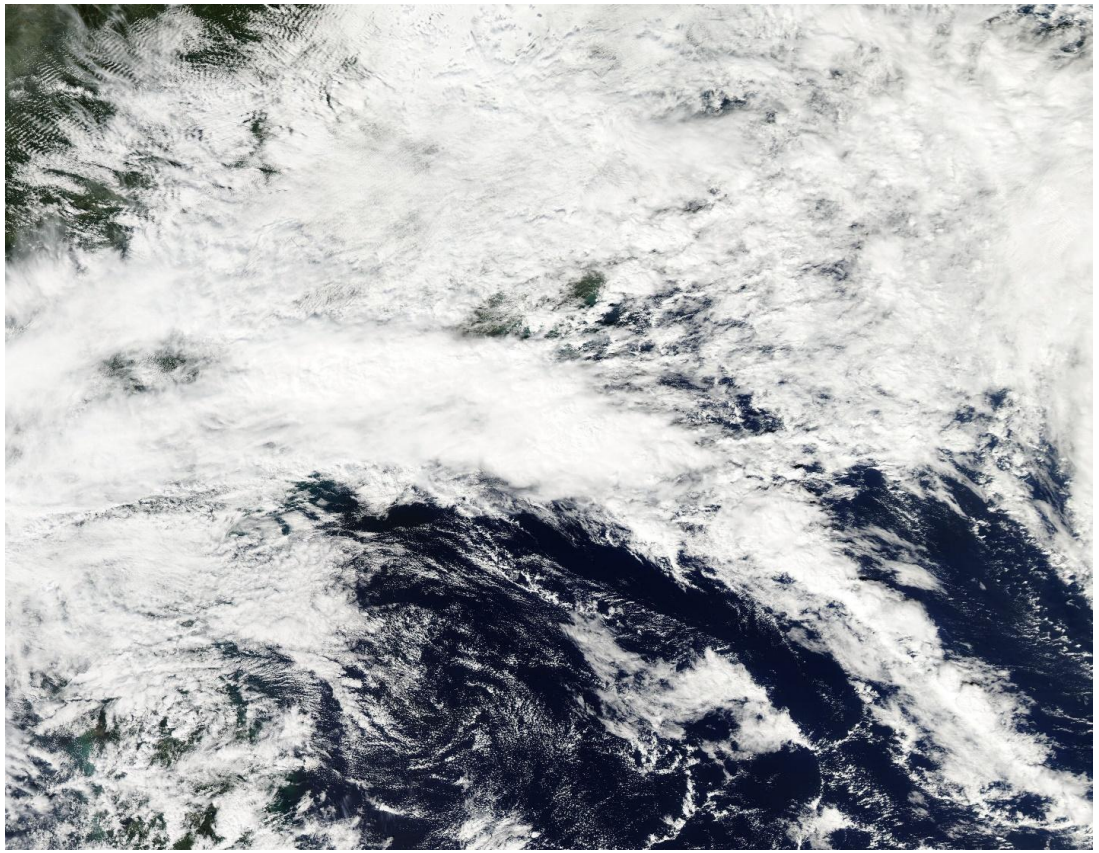
Volcanic eruptions, particularly ones with an expansive ash plume, can lead to significant issues in air travel and can even have impacts on global climate/temperatures in extreme cases. The image below is from the Eyjafjallajökull volcanic eruption on April 17, 2010 (MODIS-Aqua).

Glacial Silt



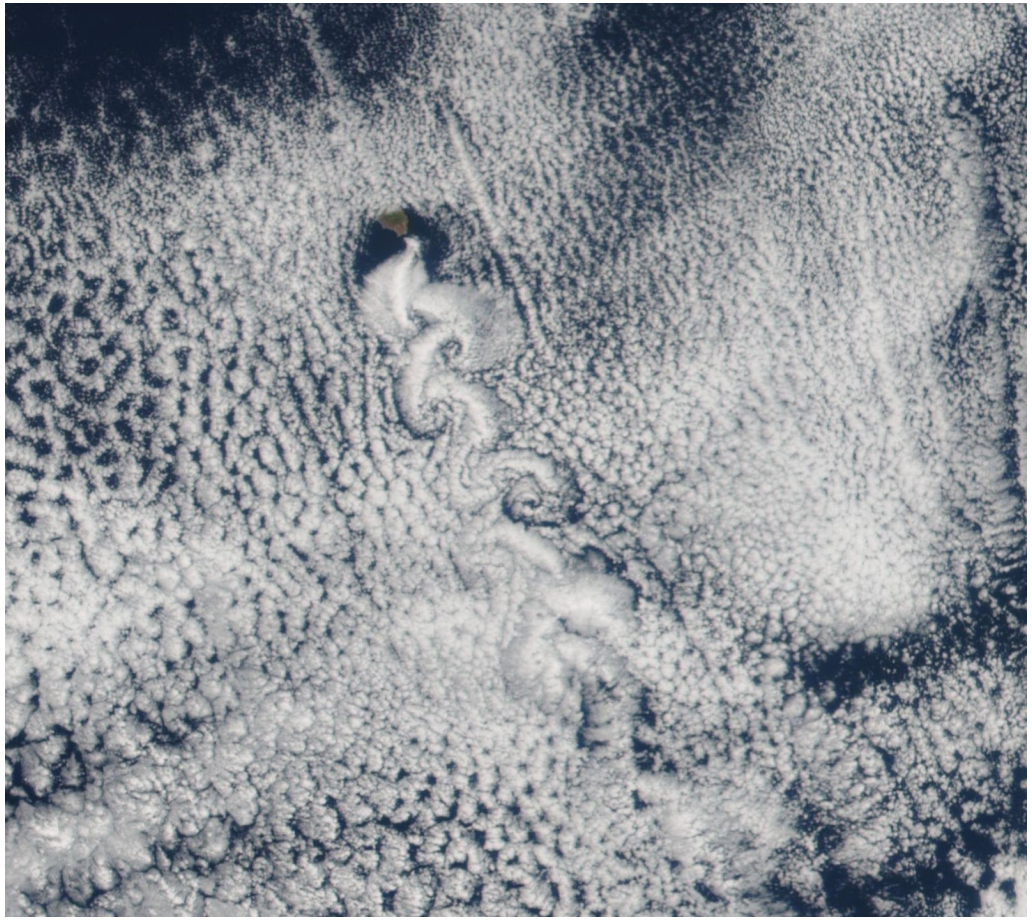
Sediment rich glacial meltwater intrusions can be identified in satellite imagery by turquoise colored water where glacial rivers deposit into oceans or lakes. The image is from Cordova, Alaska on September 20, 2017 (MODIS-TERRA).

Atmospheric Rivers

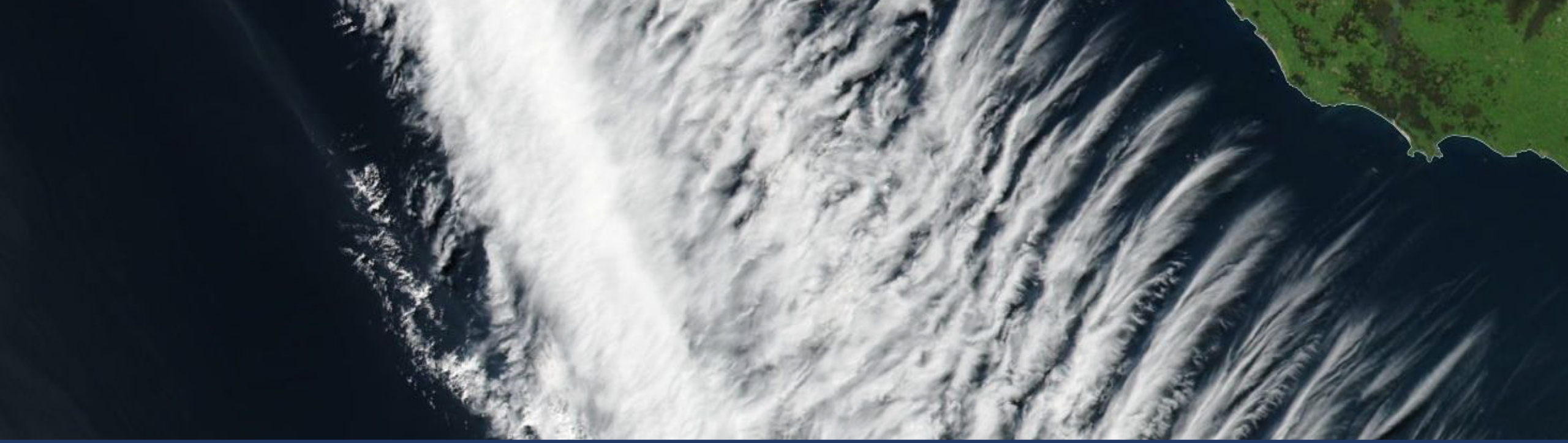


Prolonged heavy rainfall events associated with anomalously high precipitable water and efficient microphysical properties that can cause significant flooding. The below is from the South Carolina coast on October 04, 2015 (MODIS)

Von Karman Vortices

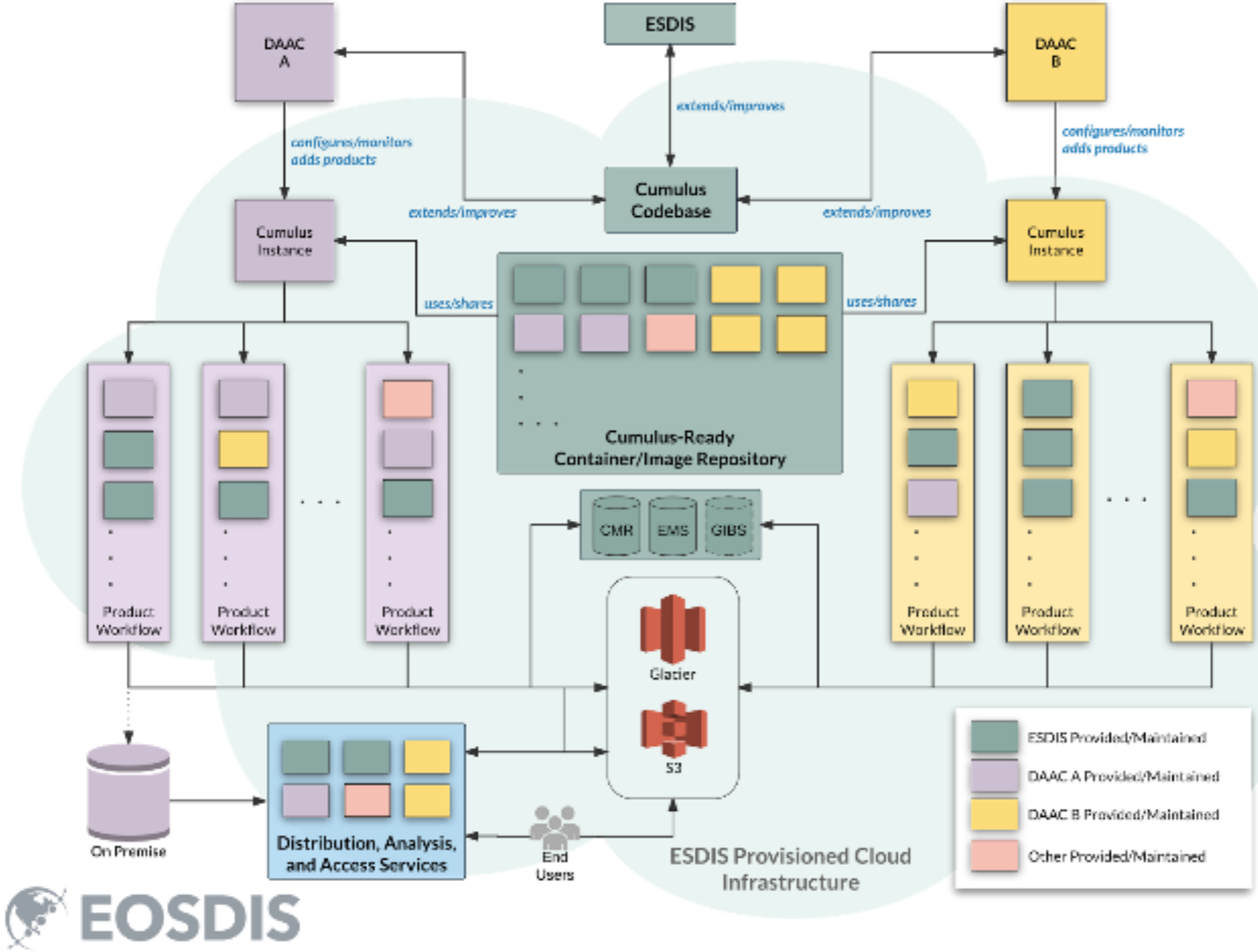


Indication of atmospheric instability. Commonly found over isolated islands. An example image SW of the Baja Peninsula over Isla Socorro on May 30, 2017 (VIIRS).



Use NASA data!

Questions?
rahul.ramachandran@nasa.gov


DAAC A
ESDIS
DAAC B
extends/improves
configures/monitors
adds products
Cumulus Codebase
extends/improves
extends/improves
configures/monitors
adds products
Cumulus Instance
Cumulus Instance
uses/shares
uses/shares
Cumulus-Ready Container/Image Repository
CMR
EMS
GIBS
Product Workflow
Product Workflow
Product Workflow
Product Workflow
Product Workflow
Product Workflow
Glacier
S3
On Premise
Distribution, Analysis, and Access Services
End Users
ESDIS Provisioned Cloud Infrastructure
ESDIS Provided/Maintained
DAAC A Provided/Maintained
DAAC B Provided/Maintained
Other Provided/Maintained
EOSDIS

Cumulus Trade-Offs

Advantages

- Maintains thematic data stewardship expertise to support specific community needs
- Collocates computations and data to enable new science and application of large scale analytics
  - Enables opportunities for other communities to innovate (Commercial, NGO, etc.)
  - Collocation of data supports cross-thematic interdisciplinary science
- Common framework reduces redundant tools/services; enables sharing; enforces use of standards, uniform policies and processes
- Brings to bear economies of scale
- Enables consistent performance for services

Disadvantages

- Perceived lack of control over data (Golden copy issues)
- Will require a systematic transition
- Requires modifications in existing operations
  - New governance policies and procedures are needed
- Retraining existing staff with a different skill set
- Cost of Dockerizing existing code.
- Possible vendor lock-in

Valley Fog



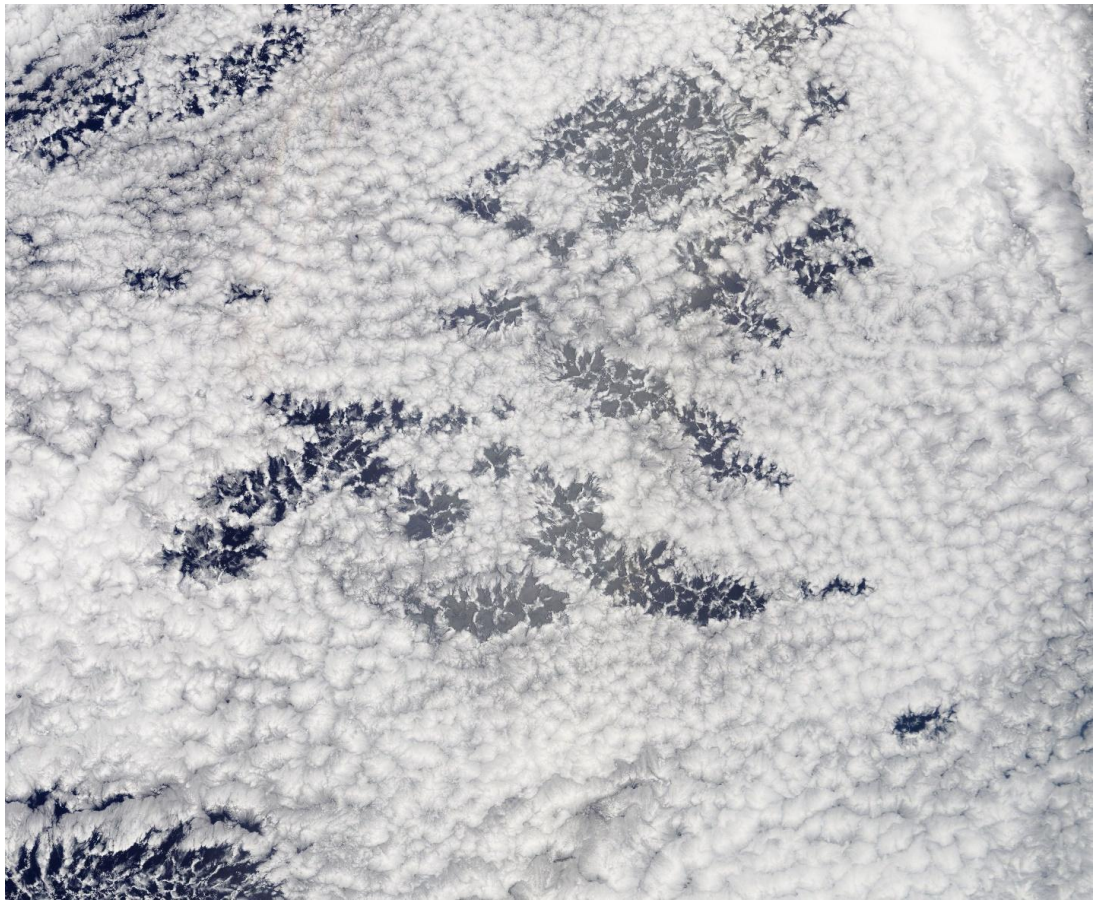
- Can be an indication of soil moisture and surface winds within mountainous terrain. In clear sky conditions soil moisture differences between terrain and valleys can lead to ground fog formation. The image is from Central California on January 25, 2011 (MODIS).

Sea/Lake Breeze



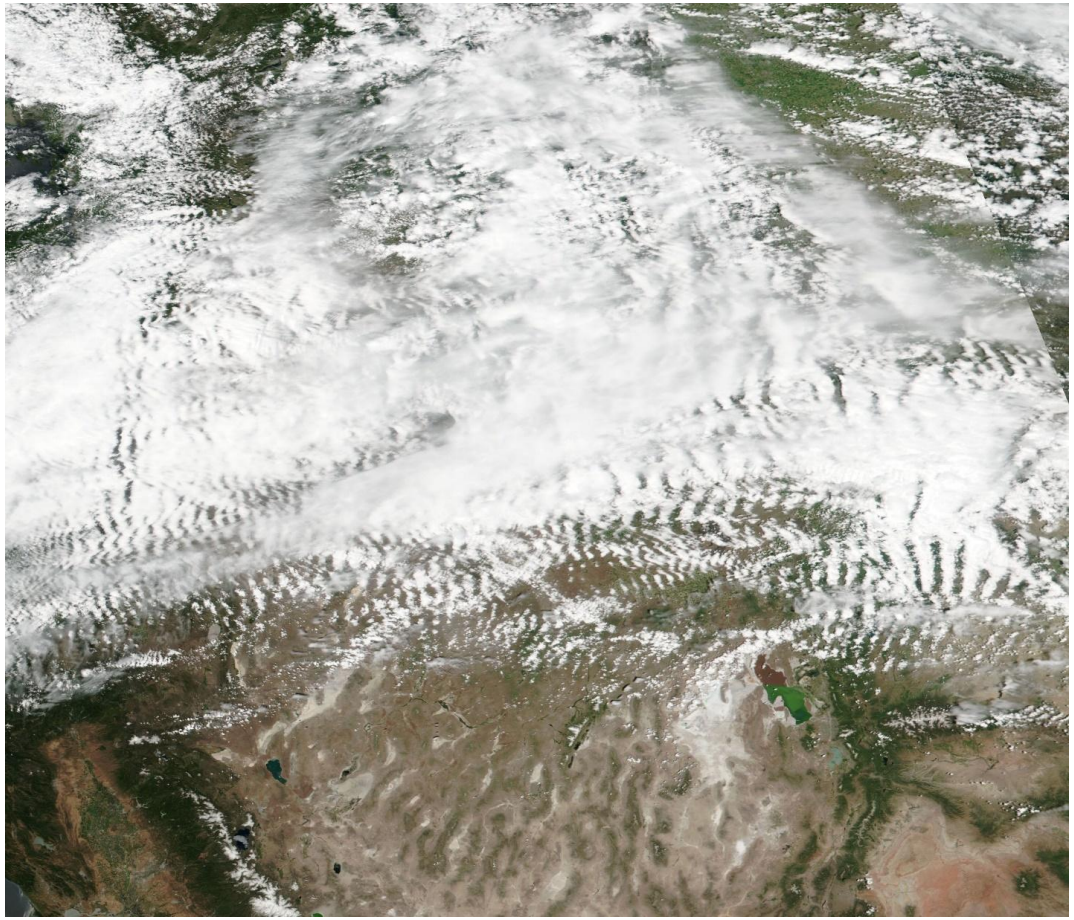
- Commonly found near large bodies of water caused by different heat capacities between water and land. Interaction between sea breeze front and horizontal convective rolls (cloud streets) is important mechanism for convective initiation in coastal settings. The image is from the Florida Peninsula on July 30, 2016 (MODIS).

Open/Closed Convection



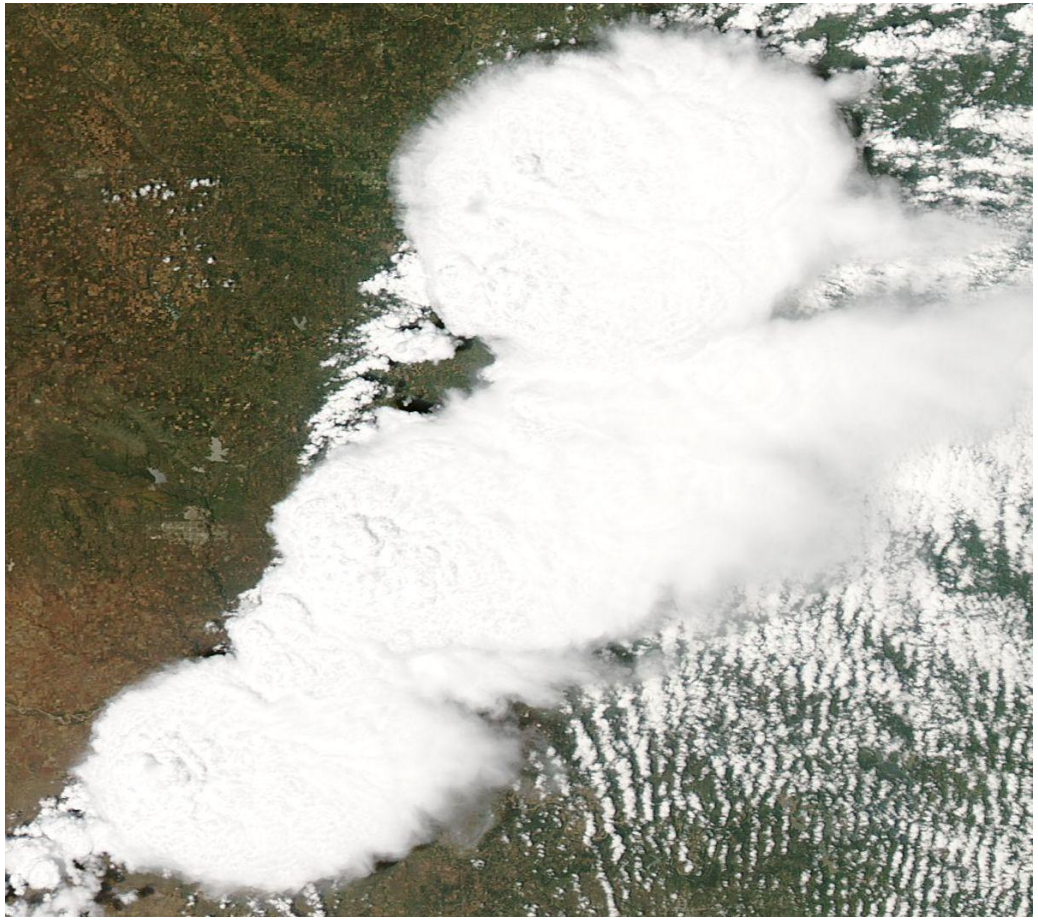
- Typically occurs over oceans and can be indicative of spatial distribution of vertical motion and surface characteristics. Example image west of the Baja Peninsula on May 30, 2017 (MODIS).

Lee Waves



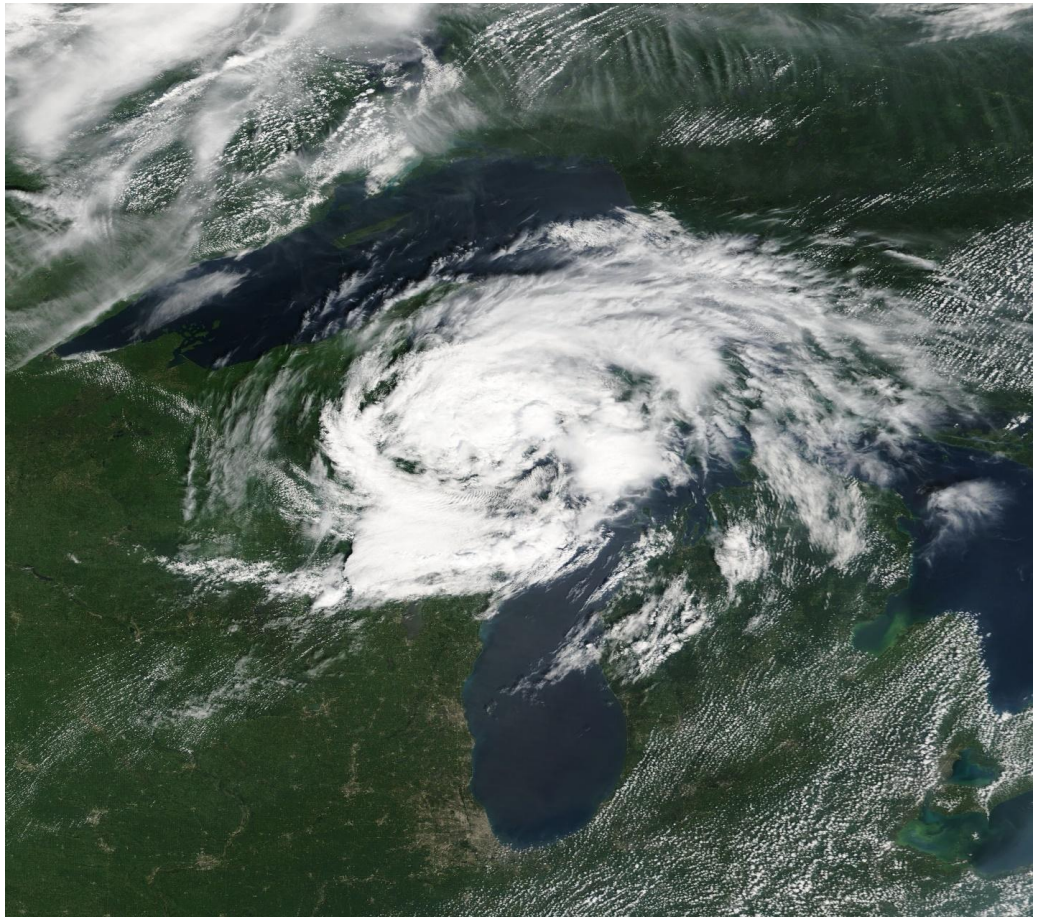
- Occur in statically stable environments east of barrier topography when air is forced up over the barrier. The image below is from the western United States from June 16, 2017 (VIIRS)

Supercells



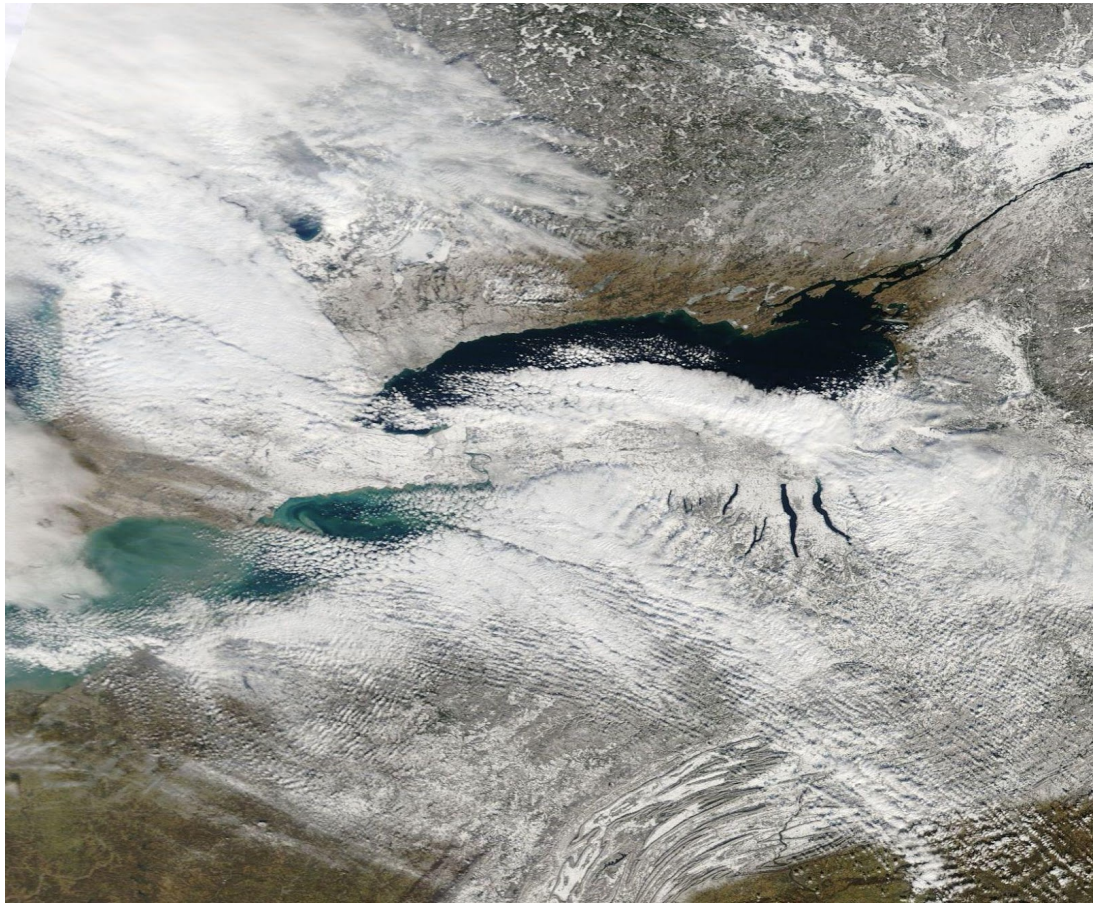
- Thunderstorms with a rotating updraft that can produce tornadoes and significant damage at the surface. Aloft overshooting tops and spreading anvils make them easy to detect from satellite imagery. Typically occur in the late spring in the Midwestern and Southeastern United States. The image below is from May 20, 2013 over Central Oklahoma (MODIS)

Mesoscale Convective Vortex/Clusters



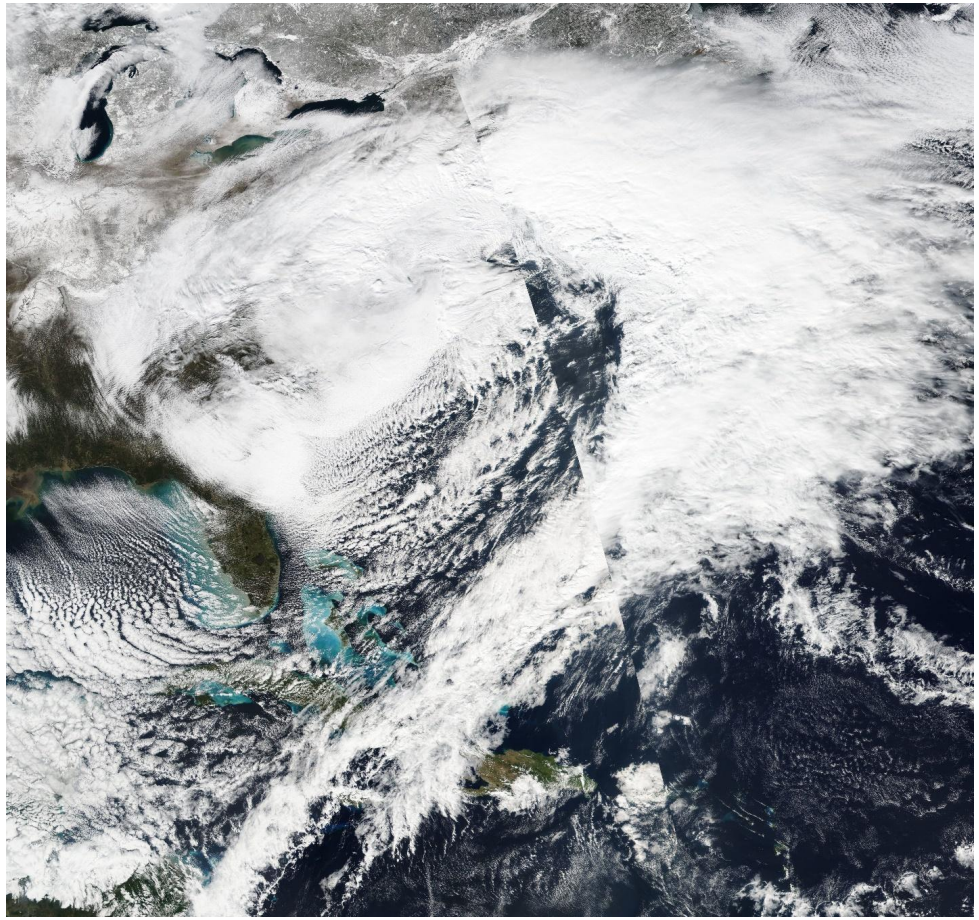
- Mesoscale convective vortices produce unique cloud features and sometimes heavy rainfall with an appearance of a smaller scale tropical cyclone. The image is from the upper Midwest and the UP of Michigan on August 14, 2009 (MODIS)

Lake Effect Snow Bands



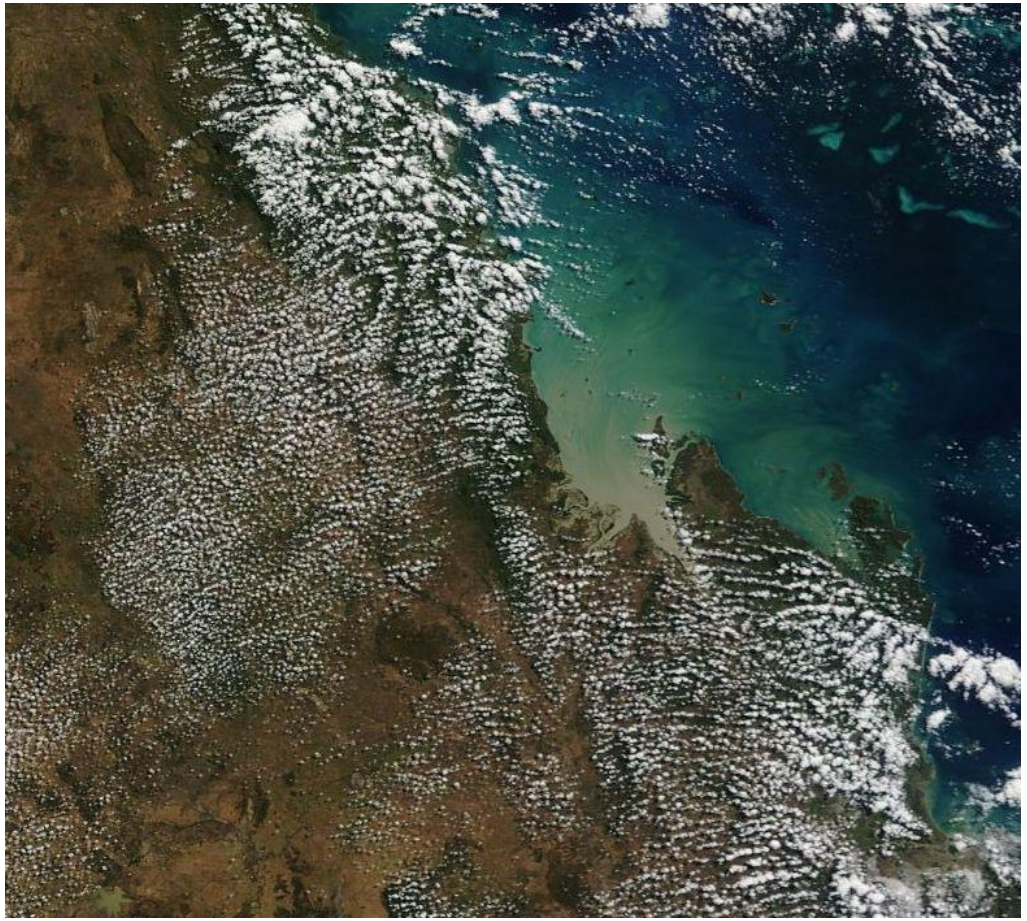
- Commonly found over large lake bodies in the mid latitudes in the fall to early winter caused by the different heat capacities between land and water. The image is from Lake Ontario and Lake Erie on January 18, 2013 (MODIS).

Extratropical Cyclones



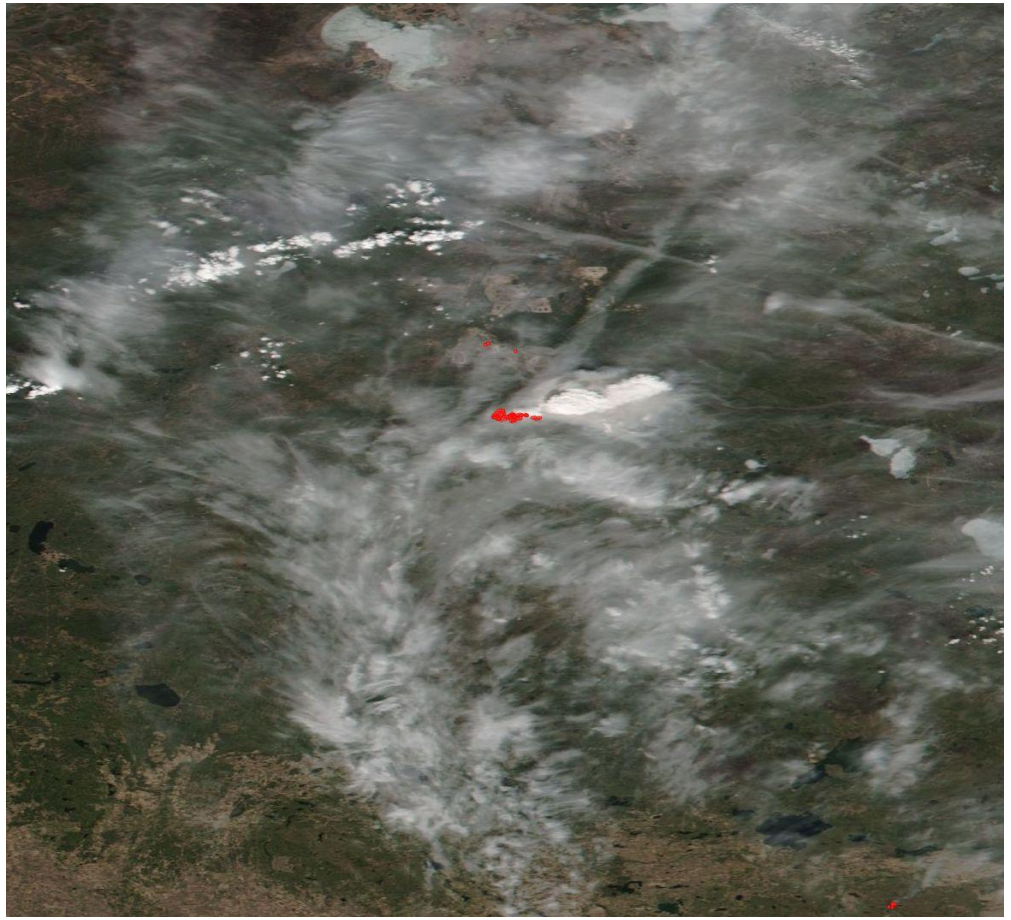
- Baroclinically forced cyclones responsible for most mid-latitude weather systems. Can be found globally, typically poleward of ~30°N/S. The image is from an extratropical cyclone from January 23, 2016 of the east coast of the United States (VIIRS)

Cloud Streets



- Long roll-like clouds typically oriented parallel to the mean wind and found in certain convective environments. The orientation of the horizontal convective rolls can have significant implications in coastal convective initiation. The image is from Queensland, Australia on May 08, 2016 (MODIS).

Pyrocumulus Clouds



- Thermal instability associated with fires can generate strong updrafts supporting cumulus cloud development near the surface hotspot. The image below is from the Ft. McMurray wildfire in Alberta, Canada on May 03, 2016 (VIIRS).