Managing planetary scale data on the cloud

How to use AWS and open source software to develop applications to support decision makers

developmentSEED

Introductions

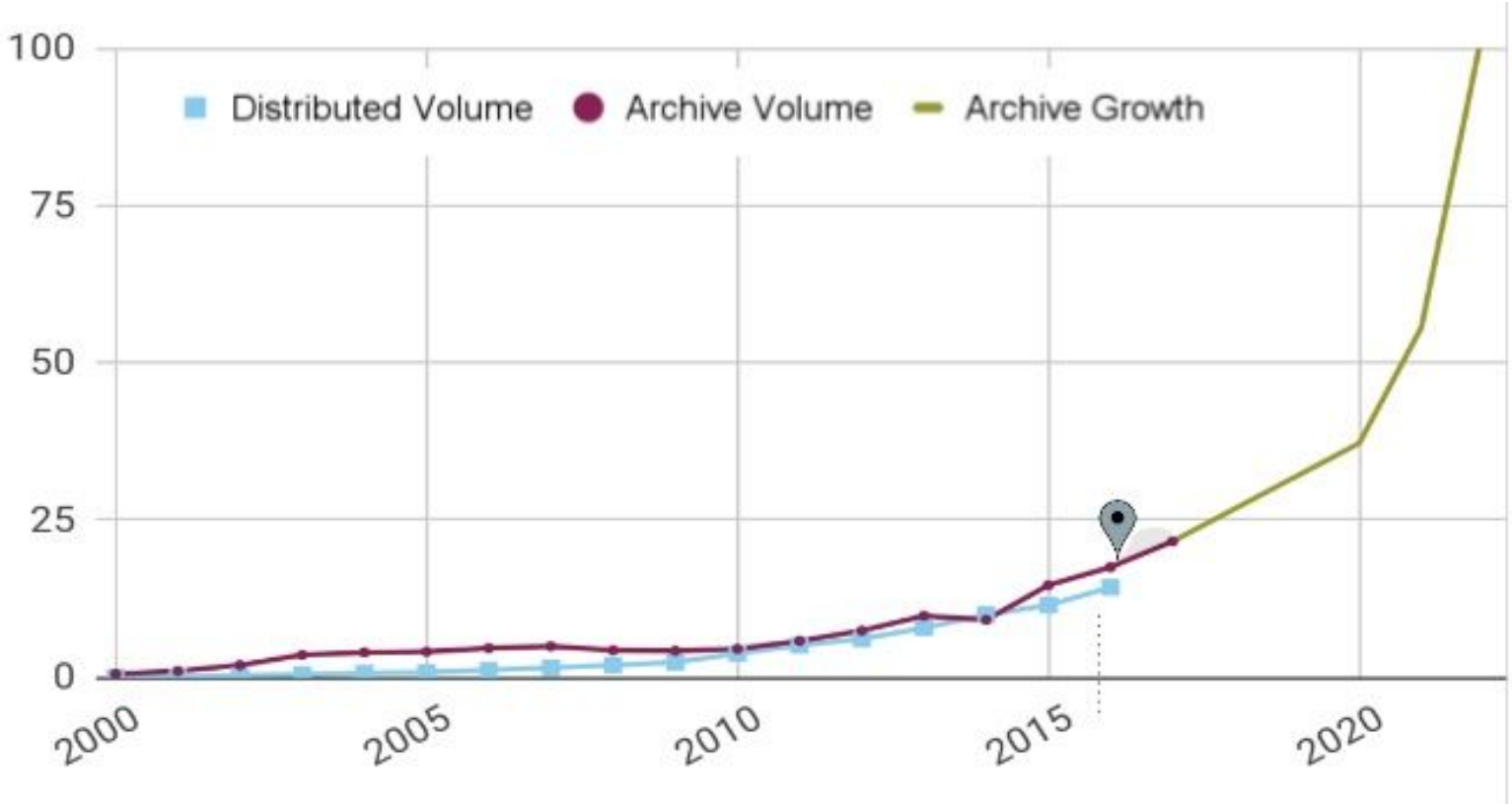
Aimee Barciauskas
Data Engineer
@_aimeeb
aimee@developmentseed.org

Ian Schuler
CEO
@ianschuler
ian@developmentseed.org

NASA’s Commitment to open data

Since 1994, the ESDS Program has committed to the full and open sharing of Earth science data obtained from NASA instruments to all users.

EOSDIS Big Data Evolution



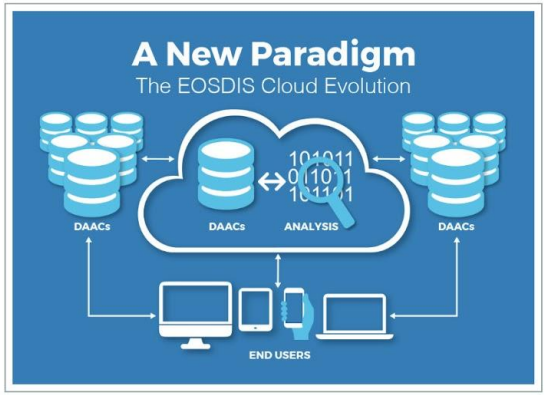
A New Paradigm

The EOSDIS Cloud Evolution

Earth Science Data in the Cloud: The EOSDIS Cumulus Project

As part of the ongoing evolution of EOSDIS data and services, testing and prototyping are underway to see how DAAC data collections can be archived and disseminated using the commercial cloud.

Josh Blumenfeld, EOSDIS Science Writer



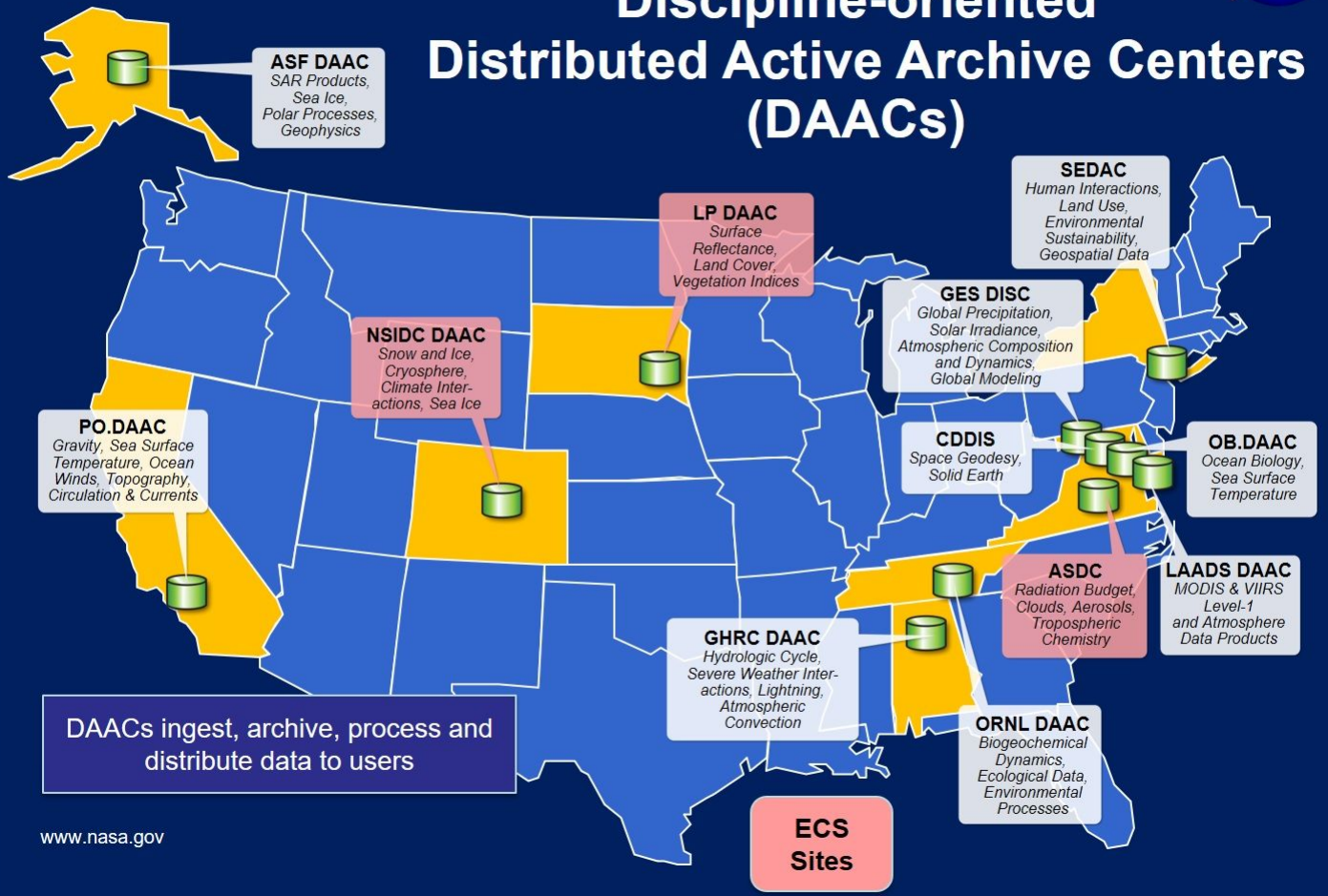
Clouds in the sky constantly grow and shrink as they adjust to evolving atmospheric conditions. A cloud computing environment, like an atmospheric cloud, also easily can adjust to evolving conditions, expanding or contracting as needed based on data storage requirements and the needs of data users. This flexibility helps make the commercial cloud a viable option for archiving and disseminating large volumes of data or for managing data holdings that are expected to change rapidly over a short amount of time.

NASA's Earth Observing System Data and Information System (EOSDIS) is responsible for a data collection that is both large in volume and projected to grow rapidly over the next several years. From its current size of almost 22 petabytes (PB), the volume of data in the EOSDIS archive

What is Cumulus?

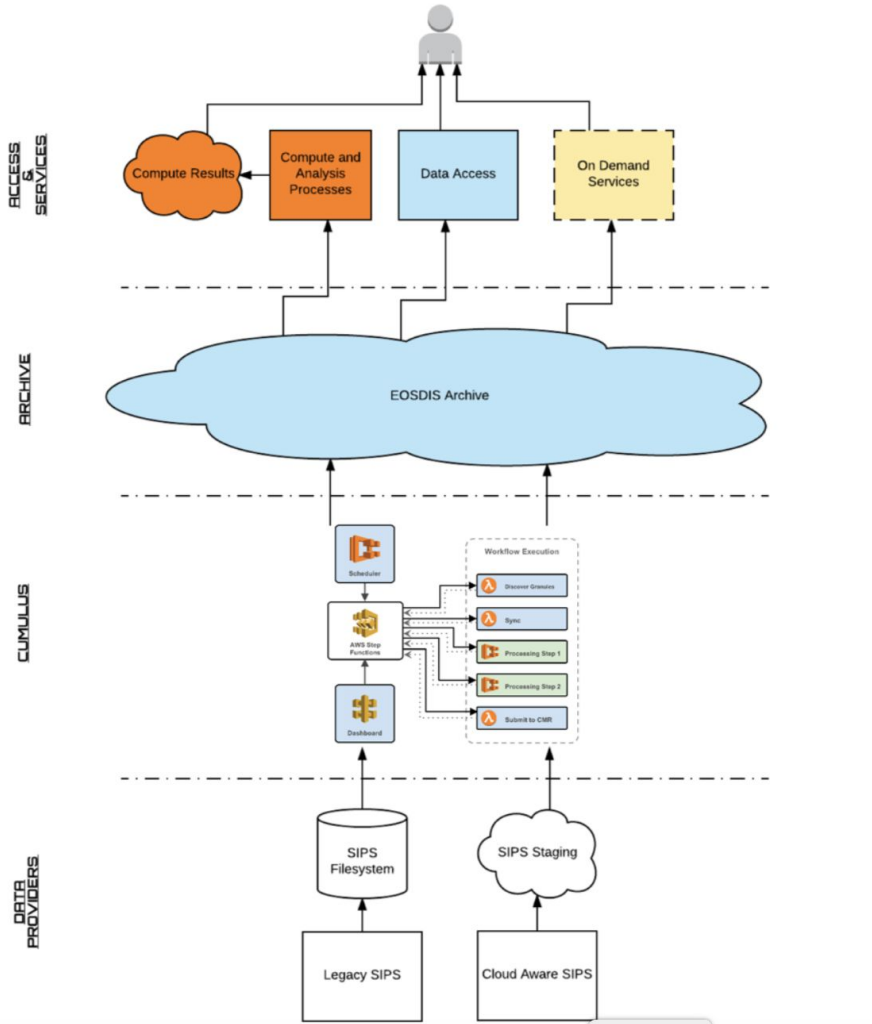
National Aeronautics and Space Administration

Discipline-oriented Distributed Active Archive Centers (DAACs)



DAACs ingest, archive, process and distribute data to users

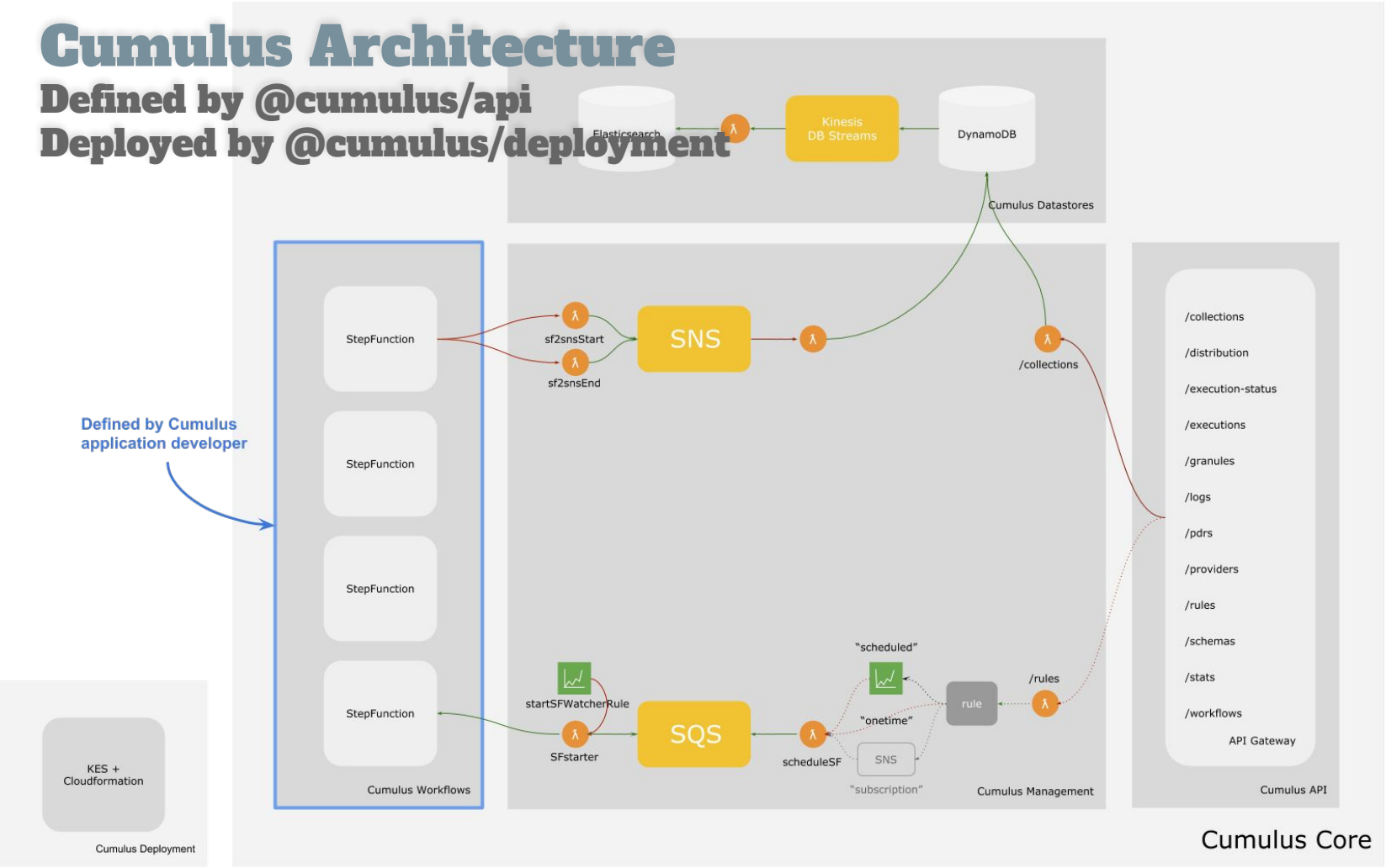
www.nasa.gov


ACCESS & SERVICES
Compute Results
Compute and Analysis Processes
Data Access
On Demand Services
ARCHIVE
EOSDIS Archive
CUMULUS
Scheduler
AWS Step Functions
Dashboard
Workflow Execution
Discover Granules
Sync
Processing Step 1
Processing Step 2
Submit to CMR
DATA PROVIDERS
SIPS Filesystem
SIPS Staging
Legacy SIPS
Cloud Aware SIPS

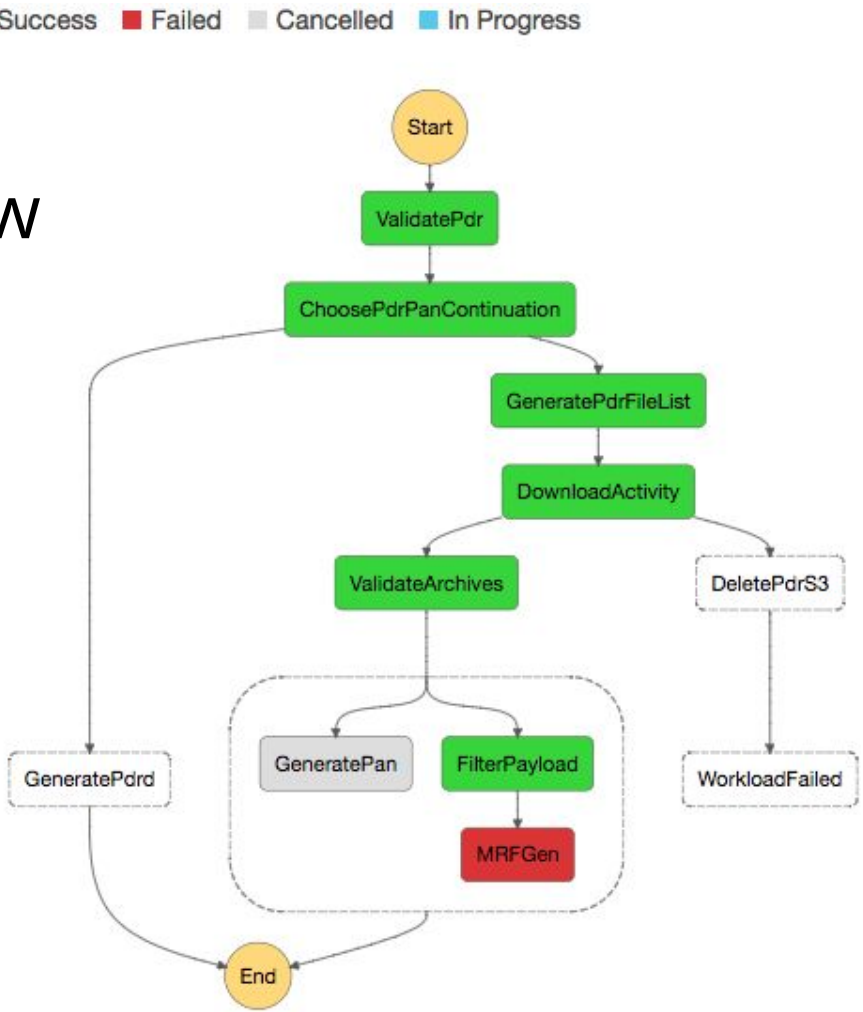
Preview

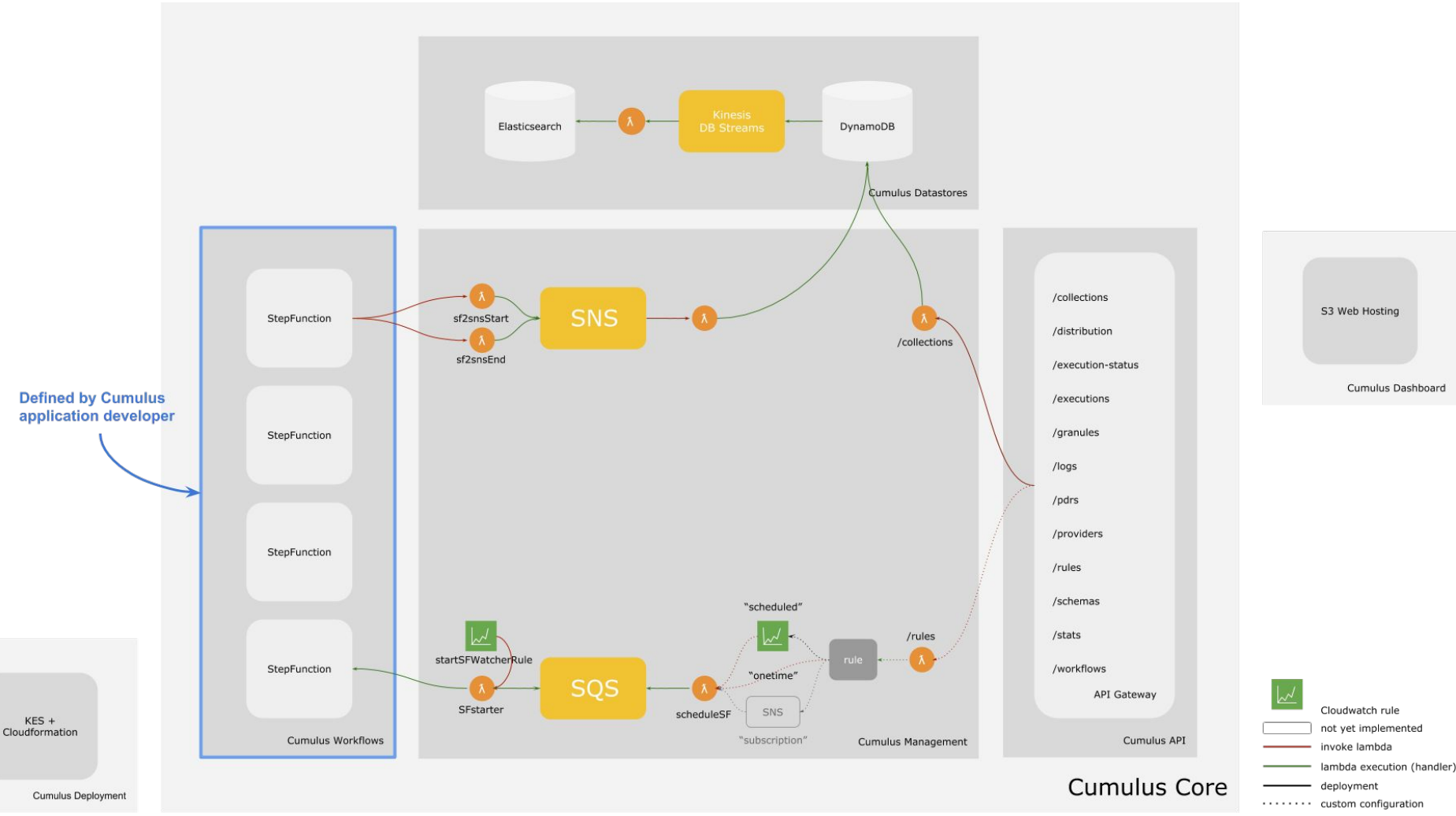
Cumulus Architecture

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

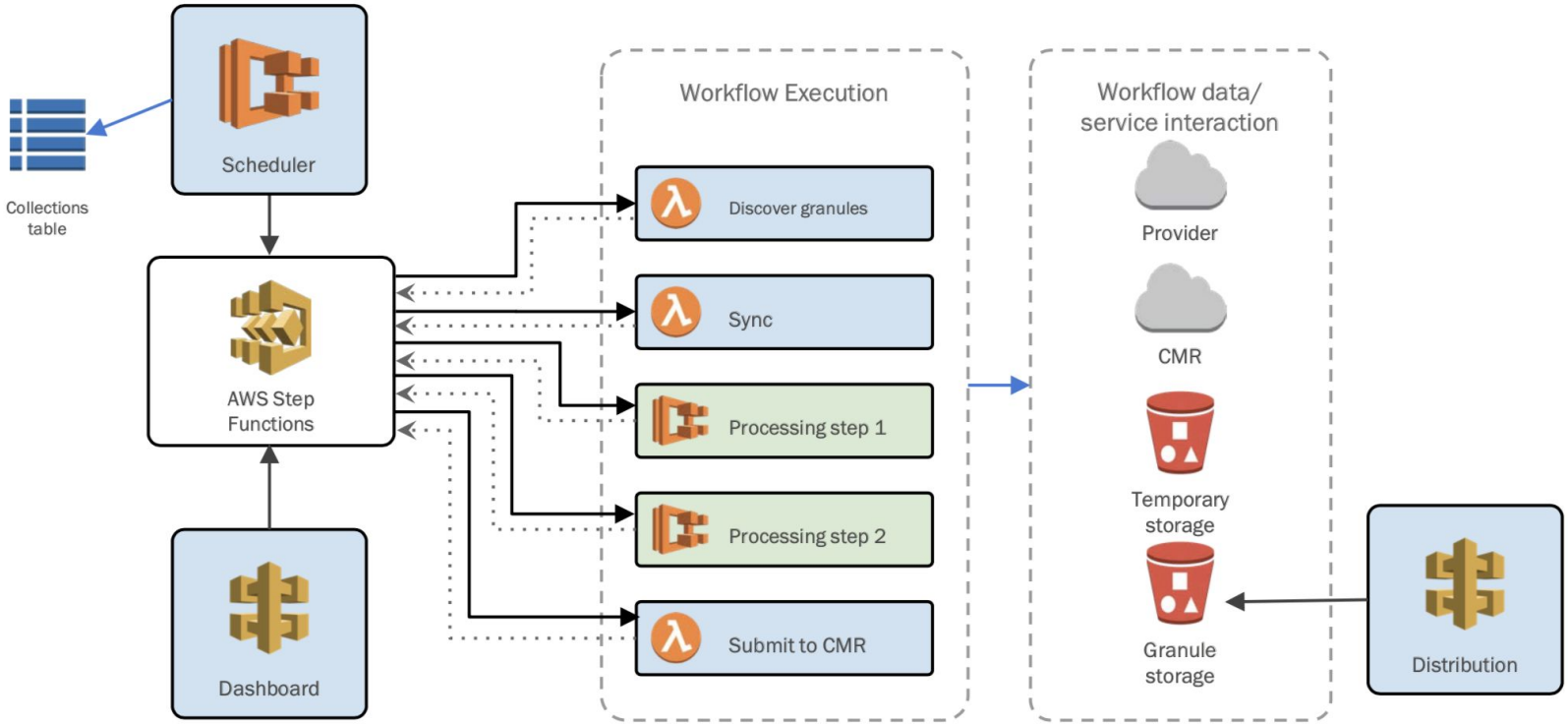


NASA's Global Browse Imagery Ingest Workflow




Elasticsearch
Kinesis
DB Streams
DynamoDB
Cumulus Datastores
Defined by Cumulus
application developer
StepFunction
StepFunction
StepFunction
StepFunction
Cumulus Workflows
sf2snsStart
sf2snsEnd
SNS
/collections
/collections
/distribution
/execution-status
/executions
/granules
/logs
/pdrs
/providers
/rules
/schemas
/stats
/workflows
API Gateway
S3 Web Hosting
Cumulus Dashboard
"scheduled"
startSFWatcherRule
"onetime"
rule
/rules
SFstarter
SQS
scheduleSF
SNS
"subscription"
Cumulus Management
Cumulus API
KES +
Cloudformation
Cumulus Deployment
Cumulus Core
Cloudwatch rule
not yet implemented
invoke lambda
lambda execution (handler)
deployment
custom configuration

Ingest & Archive with AWS Step Functions



EOSDIS

Cumulus is a collection of resources for deploying and configuring a data pipeline in the cloud.

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/discover-granules**
- **cumulus-dashboard:** Code to generate and deploy the dashboard for the Cumulus API.

Why Cumulus?

1. Leverages AWS Serverless, which gets us:
    a. Reduced devops work or limits security risks associated with managing servers
    b. Scales and is fault tolerant out of the box
2. Features a rich API for triggering, scheduling and monitoring workflows
3. Dashboard offers user interface for underlying API
4. API and dashboard come with configurable OAuth integration
5. Supported by NASA Cumulus Core development team
6. **Modular:** Has many components but can be configured for different use cases

Cumulus Applications

Cumulus outside of NASA

Partner	**Project**	**Using Cumulus to**
GEO	GEO GLAM (Global Agricultural Modeling)	Discover and transfer MODIS tiles
WRI	Air Quality Model Live	Produces air quality model results in near real-time
NHC	Hurricane Intensity Estimation using Machine Learning on GOES imagery	Generate hurricane intensity predictions

Want to learn more?

nasa / cumulus

Cumulus Framework + Cumulus API

11,349 commits 53 branches 40 releases 1 environment 22 contributors View license

Branch: master New pull request Create new file Upload files Find File Clone or download

Jkovarik Merge pull request #876 from nasa/CUMULUS-670-docs ... Latest commit 4fd10fe 13 hours ago

.circleci	remove yarn e2e command from circeci command	5 months ago
.github	Update PR template	26 days ago
bin	Fix typo	14 days ago
docs	Update docs/data-cookbooks/setup.md	19 hours ago
example	Set config to shared bucket	4 days ago
packages	Update packages/api/models/schemas.js	19 hours ago
tasks	Merge remote-tracking branch 'origin/master' into CUMULUS-670-feature...	6 days ago
travis-ci	Merge branch 'master' into Cumulus-1163	13 days ago
website	Merge branch 'master' into release-1.11.3-2	8 days ago
.eslint-ratchet-high-water-mark	Ratchet eslint	8 months ago
.eslintignore	Merge branch 'AddEslintPluginNode' into AddAsyncOperationsEndpoint	7 months ago
.eslintrc.json	Disable node/no-missing-require rule, which duplicates the functional...	4 months ago
.gitallowed	Fix formatting of .gitallowed [skip-integration-tests]	5 months ago

Cumulus

Project Description

This Cumulus project seeks to address the existing need for a "native" cloud-based data ingest, archive, distribution, and management system that can be used for all future Earth Observing System Data and Information System (EOSDIS) data streams via the development and implementation of Cumulus. The term "native" implies that the system will leverage all components of a cloud infrastructure provided by the vendor for efficiency (in terms of both processing time and cost). Additionally, Cumulus will operate on future data streams involving satellite missions, aircraft missions, and field campaigns.

This documentation includes both guidelines, examples and source code docs.

The documentation is accessible at https://nasa.github.io/cumulus

Navigating the Cumulus Docs

- Cumulus API Documentation - here
- Cumulus Developer Documentation - here - Readme's throughout the main repository.
- General Cumulus Documentation - here <- you're here
- Data Cookbooks - here
- Operator Docs - here

Contributing

Cumulus API

Show examples in: cURL Python JS

The Cumulus API allows developers to interact with the Cumulus Framework, such as monitoring status or creating, editing, and deleting records. This is the same API that powers the Cumulus dashboard.

By utilizing this API, a developer can integrate with the Cumulus framework in any language or environment; although interacting with Cumulus through the Cumulus dashboard may be appropriate for many end users, for some use cases it's best to have the flexibility of a web-accessible API.

The API accepts and responds with JSON payloads at various HTTPS endpoints.

In order to use these endpoints, you must include authentication information in your HTTPS request; authentication is explained in the following section.

The following table lists the query string parameters that can be used with most of the Cumulus API endpoints. `{fieldName}` is a stand-in for any of the fields in the record, and for nested objects dot notation can be used; for example, valid `fieldName`s include: `pdrName`, `status`, and `recipe.processStep.description`.

query string parameter	description
`limit={number}`	number of records to be returned by the API call; default is `1`, maximum is `100`
`page={number}`	page number, 1-indexed; default is `1`
`sort_by={fieldName}`	which field to sort by; default is `timestamp`
`order={asc|desc}`	whether to sort in `asc` or `desc` order
`prefix={value}`	`startsWith` search of the `granuleId`, `status`, `pdrName`, `collectionName`, and `userName` fields
`fields={fieldName1, fieldName2}`	which fields to return, separated by a comma
`{fieldName}={value}`	exact value match for the given field

Cumulus Code + Documentation

- Cumulus Core Repository → https://github.com/nasa/cumulus
- Cumulus documentation → https://nasa.github.io/cumulus
- Cumulus Confluence Space → wiki.earthdata.nasa.gov/display/CUMULUS/
- DAACs' Cumulus Deployments → git.earthdata.nasa.gov/projects/CUMULUS
- Integration tests (good example stack) → github.com/nasa/cumulus/.../example

Earth AI

Search projects

Help Donate

label-maker 0.3.2

```
pip install label-maker
```

Data preparation for satellite machine learning

Navigation

- Project description
- Release history

Project Description

Label Maker

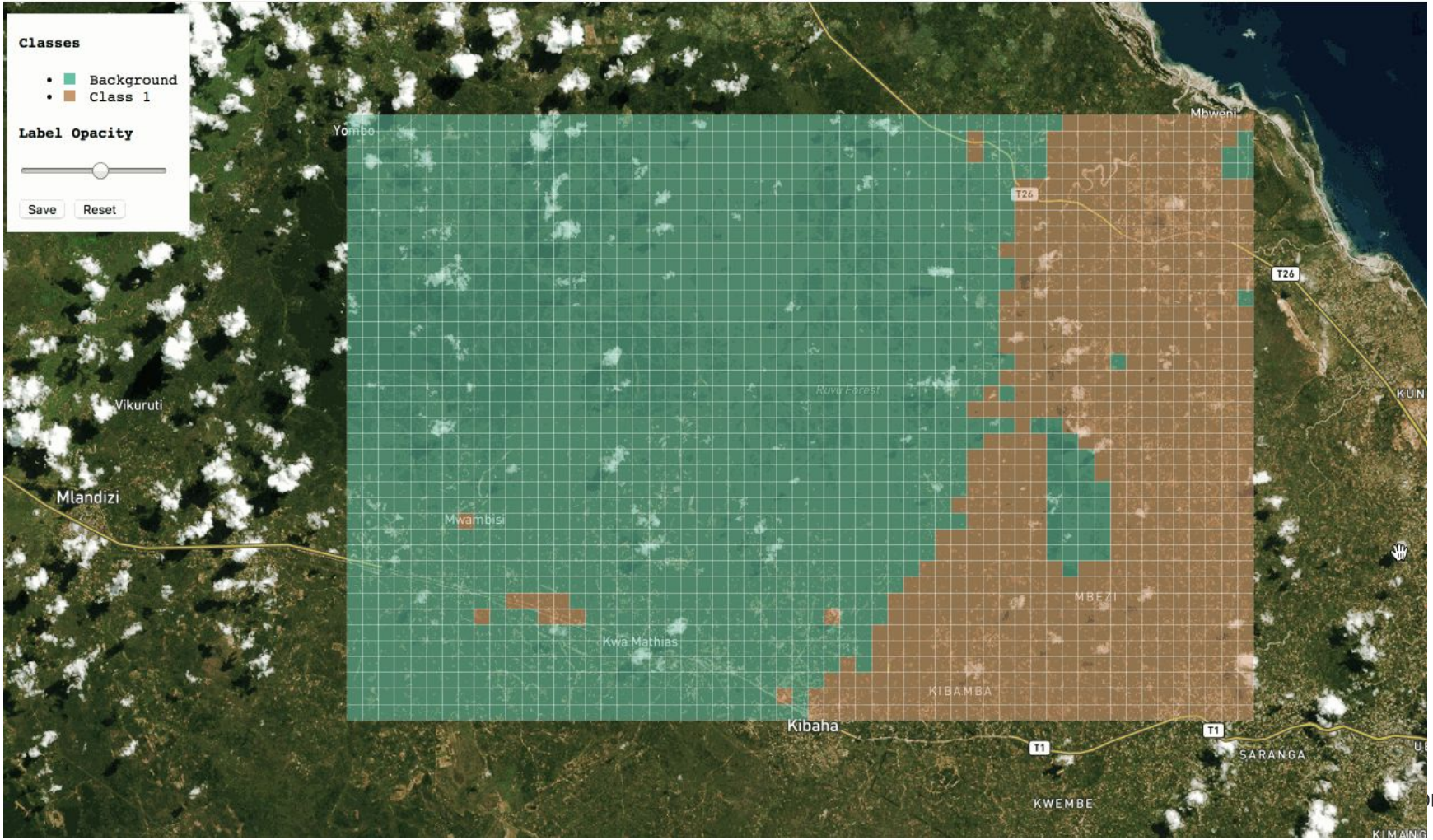
Data Preparation for Satellite Machine Learning

```
{
  "country": "united_republic_of_tanzania",
  "bounding_box": [38.83563,-6.78309,39.142055,-6.57952],
  "zoom": 16,
  "classes": [
    { "name": "Populated Area", "filter": ["has", "building"] }
  ],
  "imagery": "http://a.tiles.mapbox.com/v4/mapbox.satellite/{z}/{x}/{y}.jpg?access_token
  "background_ratio": 1,
  "ml_type": "classification"
}
```

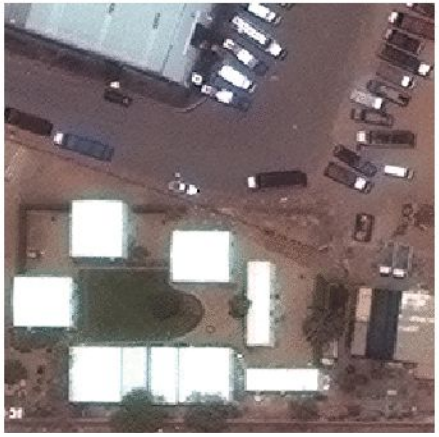
Use Label Maker and Amazon SageMaker to automatically map buildings in Vietnam

Development Seed
Jan 22, 2018 · 7 min read

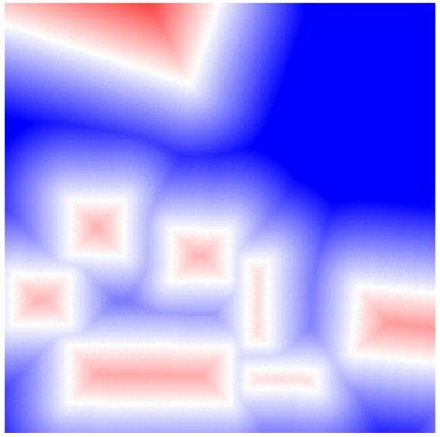
Zhuangfang Yi, PhD Explains how to quickly train and deploy an MXNet on Amazon Web Services


Classes
Background
Class 1
Label Opacity
Save
Reset
Yombo
Mbweni
T26
Vikuruti
Kuvu Forest
Mlandizi
Mwambisi
MBEZI
Kwa Mathias
KIBAMBA
Kibaha
T1
SARANGA
KWEMBE

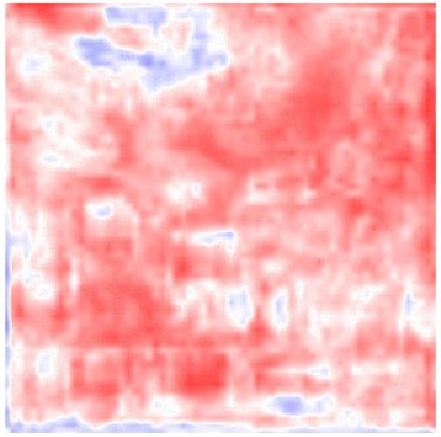
Raster image



Ground truth

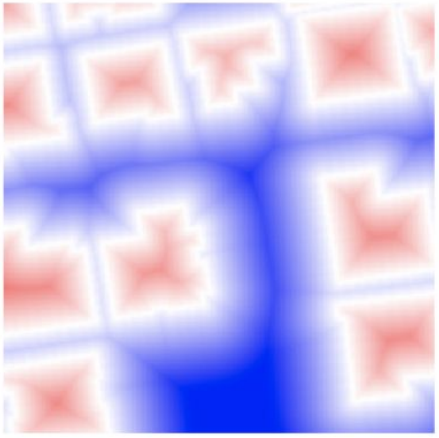


Predicted Mask

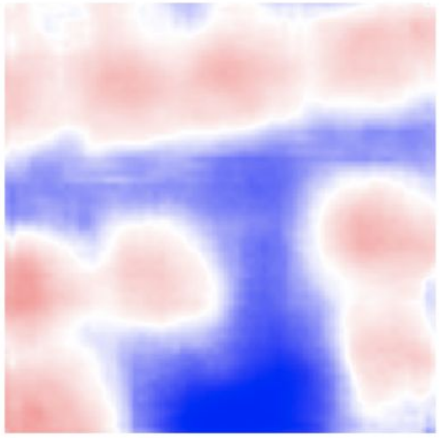


Raster image

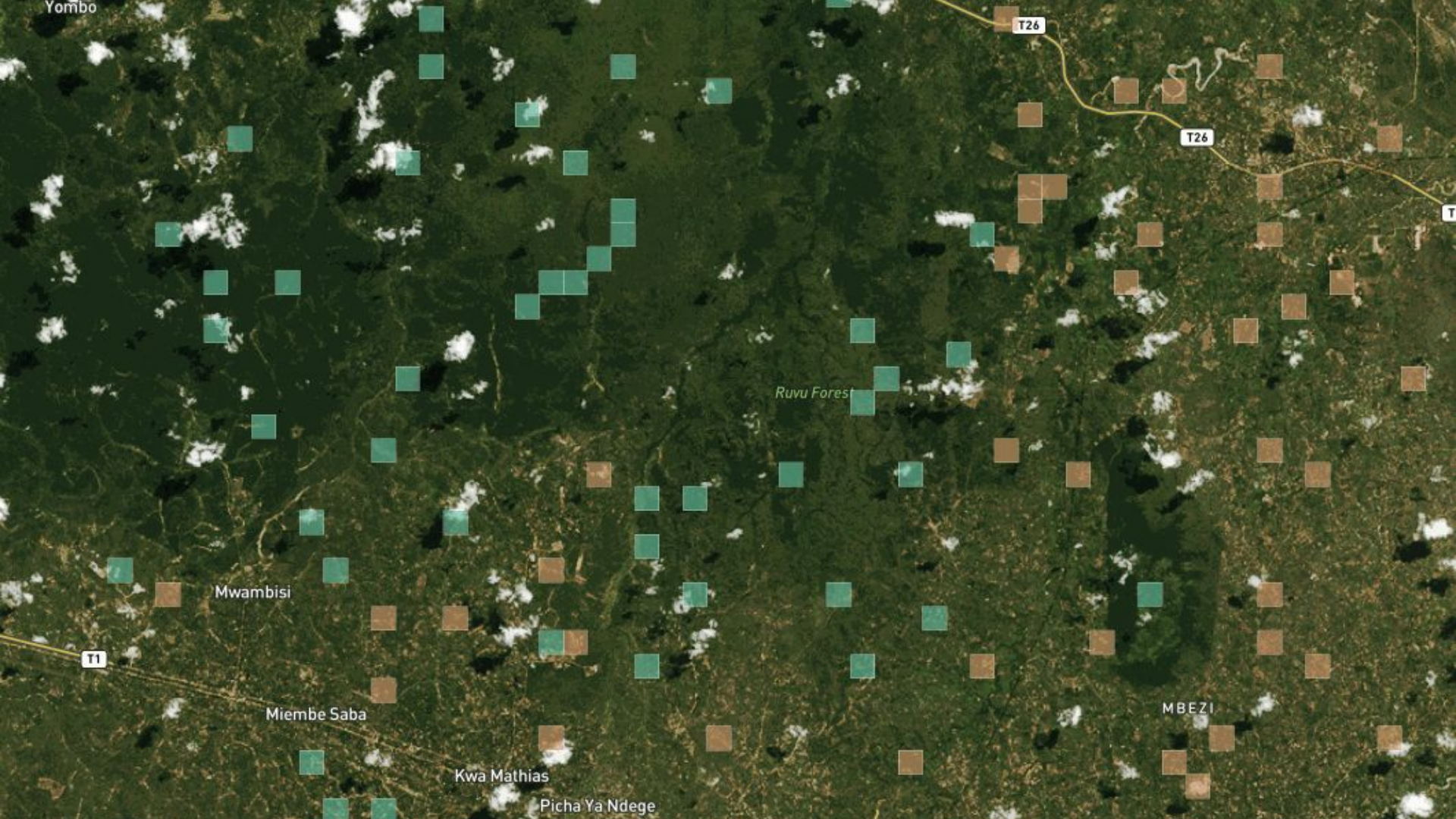
Ground truth

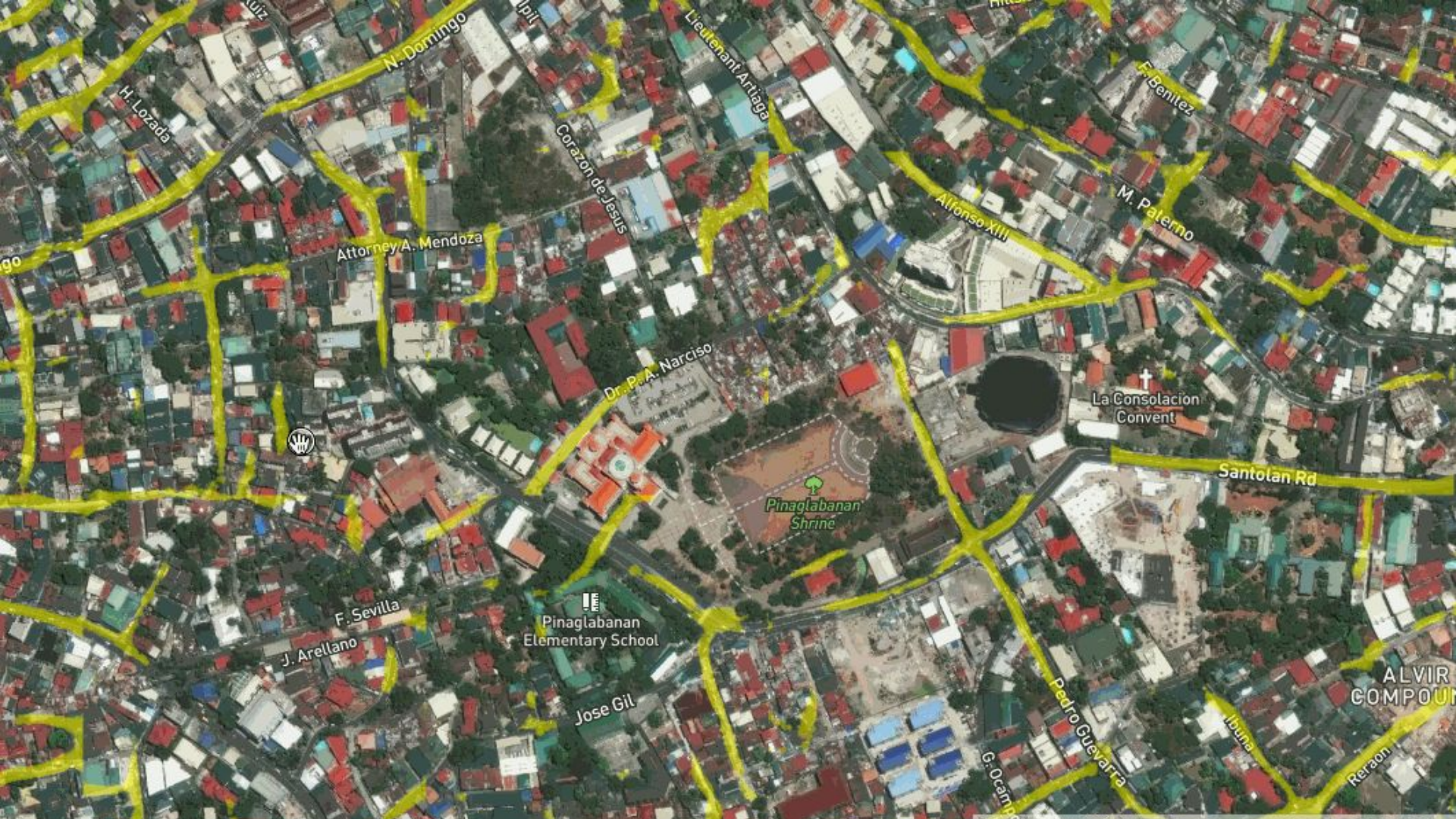


Predicted Mask

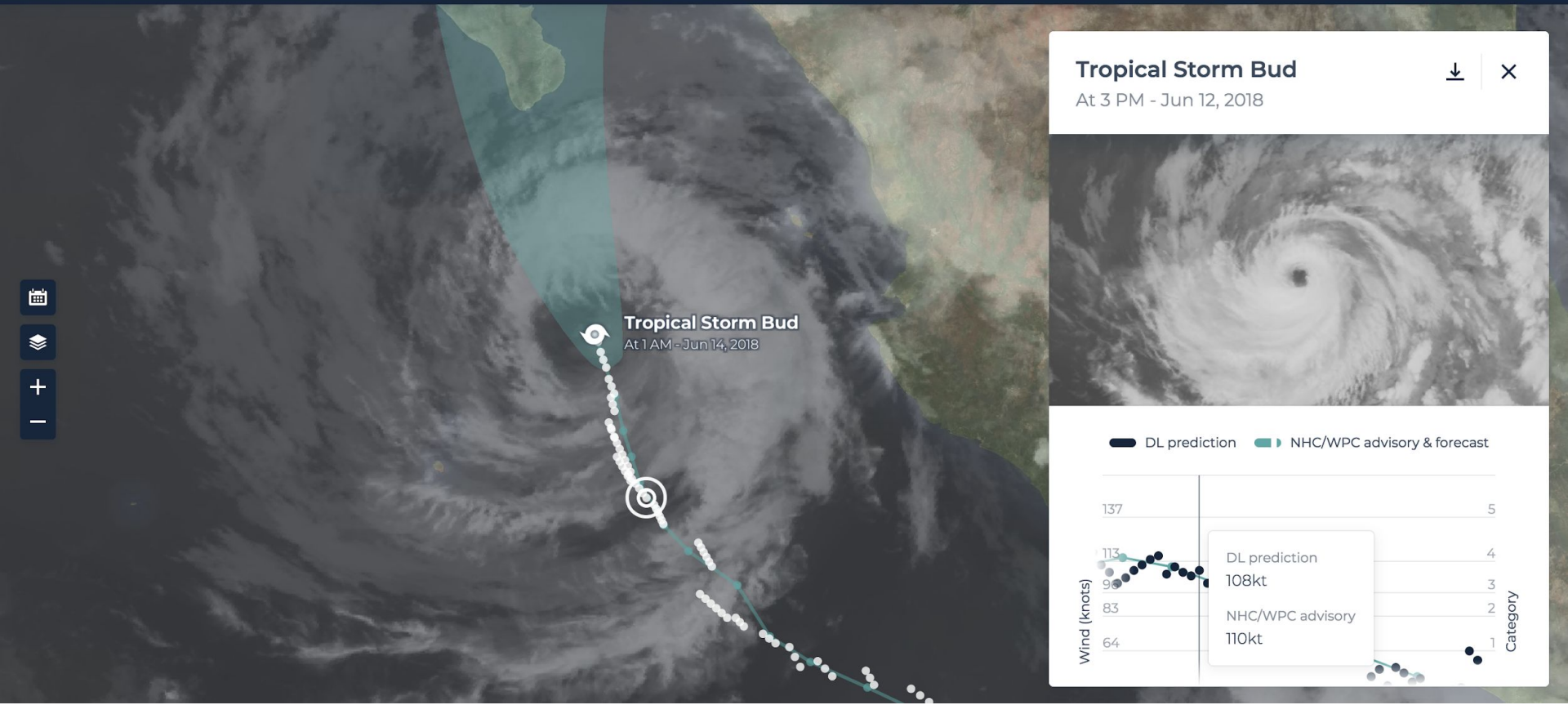


developmentSEED


Yombo
T26
T26
Ruvu Forest
Mwambisi
T1
Miembe Saba
Kwa Mathias
Picha Ya Ndege
MBEZI

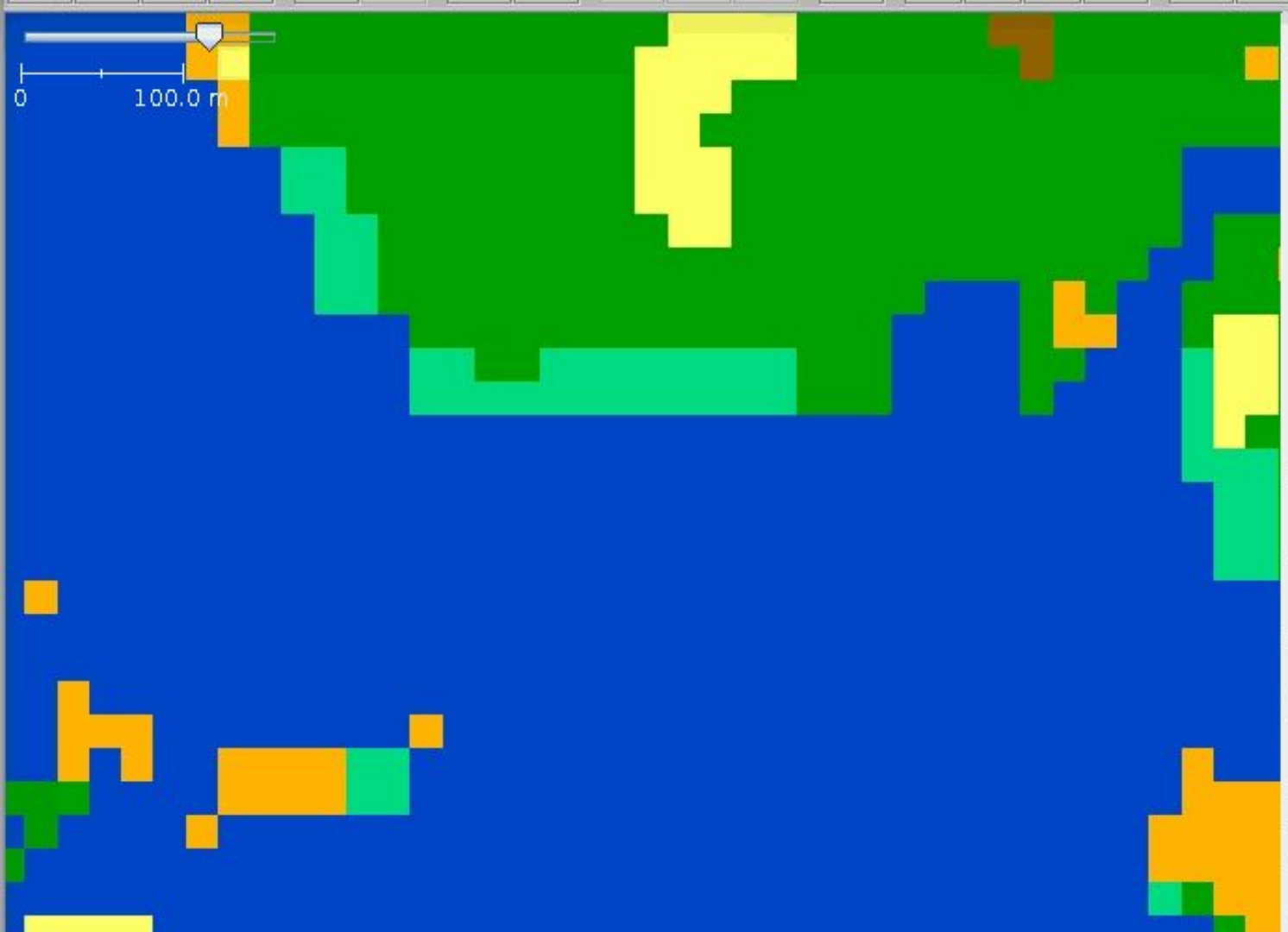

N. Domingo
H. Lozada
Corazon de Jesus
L'eutenant Artiaga
Attorney A. Mendoza
Dr. P.A. Narciso
Alfonso XIII
M. Paterno
F. Benitez
La Consolacion Convent
Santolan Rd
Pinaglabanan Shrine
Pinaglabanan Elementary School
F. Sevilla
J. Arellano
Jose Gil
Pedro Guevarra
Ibuna
Rerason

Products for Decisionmakers



Tropical Storm Bud

At 3 PM - Jun 12, 2018



Layers
Measurements
wetland.geojson
LAND COVER
Bing aerial imagery
malawi-2017-05-11
malawi-2017-10-25
malawi-2017-10-18
malawi-2017-09-23
malawi-2017-09-16
malawi-2017-08-31
malawi-2017-08-22
malawi-2017-08-06

Tags / Memberships
Add
Edit

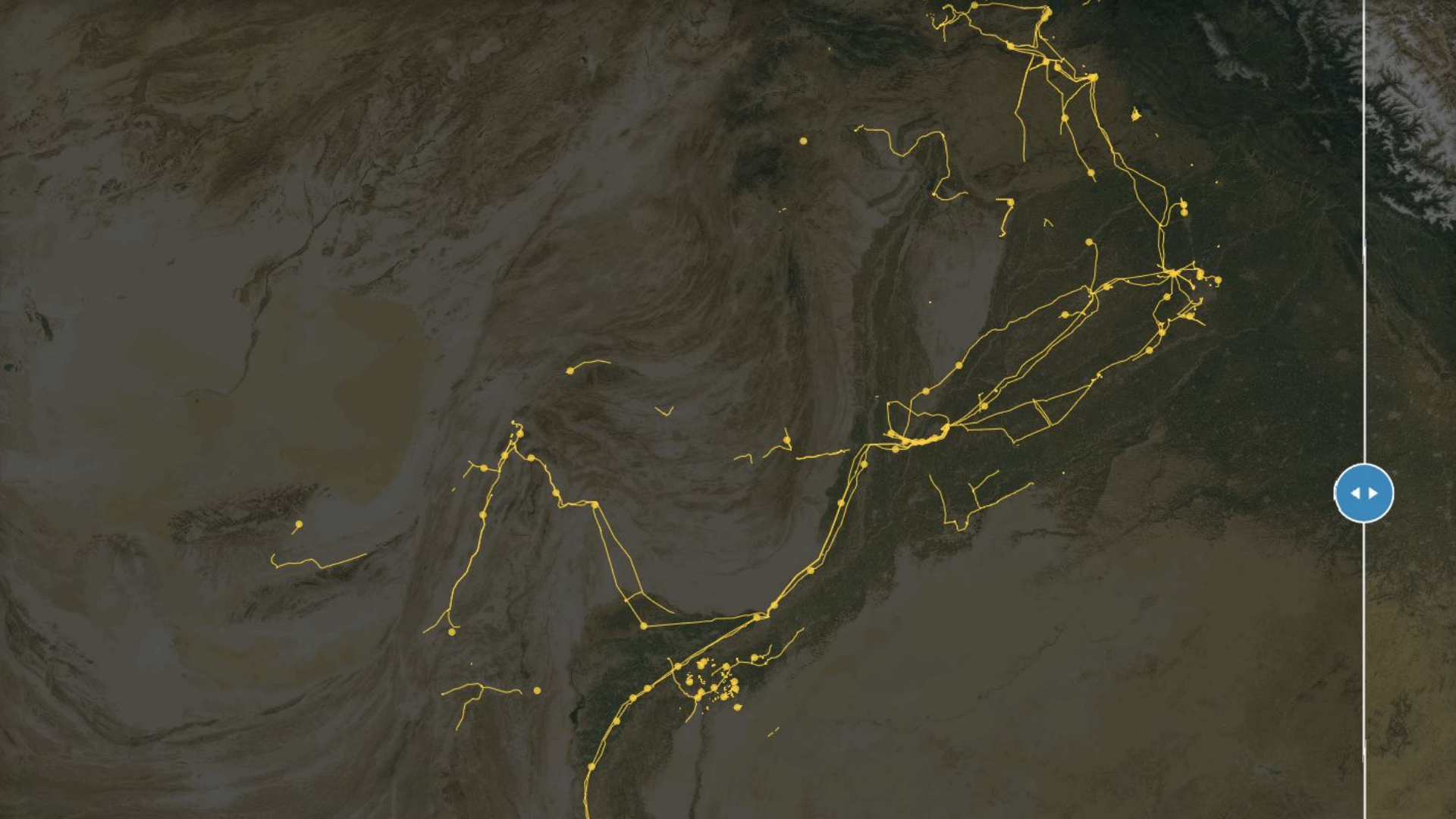
Validation Results
S...
L...
V...

Filter Hidden:0 Disable
E
H
Text
A...
E...
D...

Map Paint Styles

OpenStreetMap obj in

Measured values



UrChn


Latest Change Detection and OSM Edits
Madrid
Alcobendas
Torrejón de Ar
Madrid
Alcorcón
Móstoles
Leganés
Getafe
labrada
Select a Grid Data Source:
OSM Density
Sq. Km of OSM Buildings
less
more