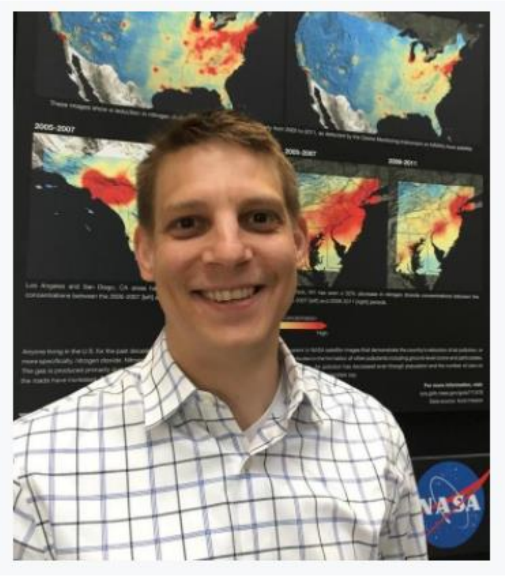
ABoVE Science Cloud Webinar

Presentations by Dr. Christoph Keller, Matthew Stroud, and Dr. Paul Montesano

September 17th, 2020

Machine Learning to Simulate Atmospheric Chemistry

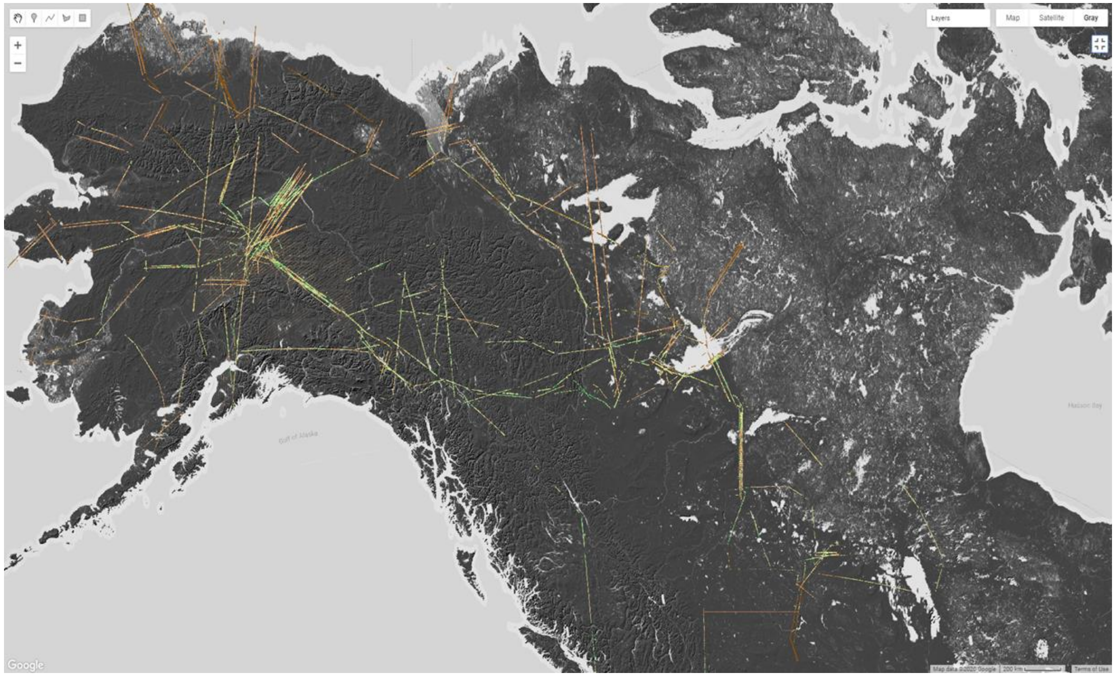
- https://www.nccs.nasa.gov/news-events/nccs-highlights/simulation_using_ML



Demonstration of the GPU Cluster on ADAPT

- https://www.nccs.nasa.gov/nccs-users/instructional/adapt-instructional/adapt-gpu/

Gridded LVIS L2 vegetation structure estimates for ABoVE analyses



ABoVE Science Cloud

ASC Updates

ASC Updates

- ADAPT downtimes – Monthly, the Wednesday after the second Tuesday of the month.
  - October 14th
  - November 1th

- ADAPT Issues? Contact support@nccs.nasa.gov

Announcements

- Success stories you would like to share? Email Liz Hoy elizabeth.hoy@nasa.gov
- How do I cite the ASC in my publications? Use language similar to:

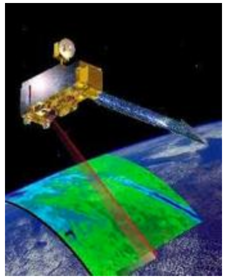
"Resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center."

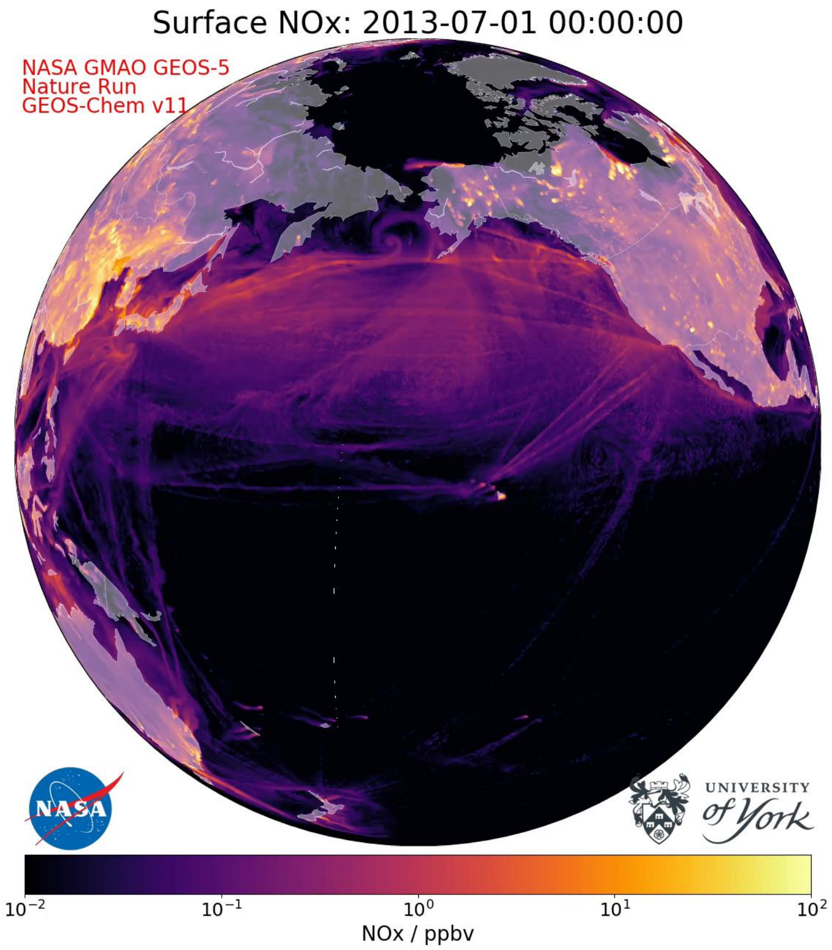
Atmospheric Chemistry Modeling using Machine Learning

Christoph A. Keller
NASA Global Modeling and Assimilation Office (GMAO)
Universities Space Research Association (USRA)

Mat J. Evans
Wolfson Atmospheric Chemistry Laboratories, University of York
National Centre for Atmospheric Sciences, University of York

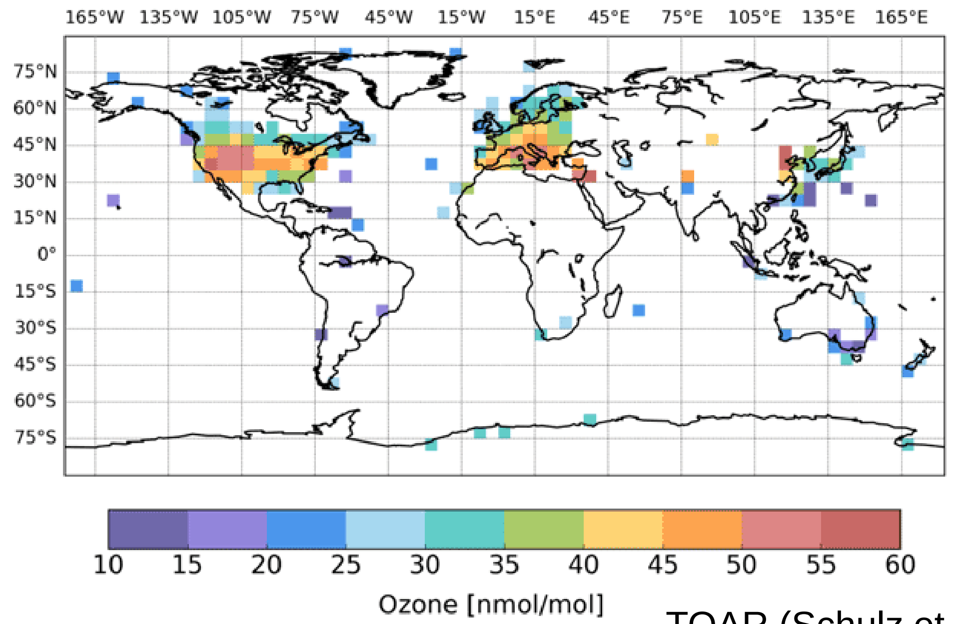
We can now simulate atmospheric composition with unprecedented level of detail



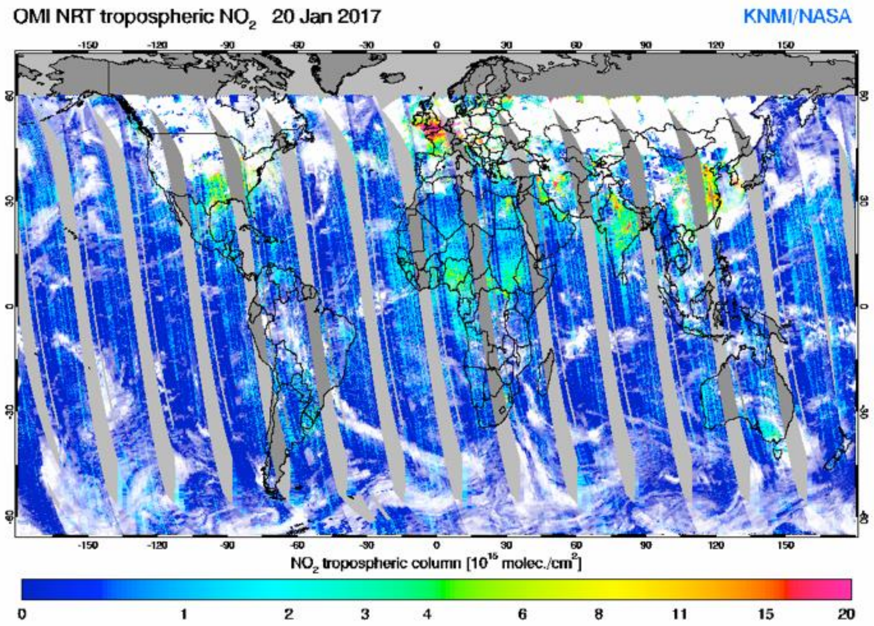


christoph.a.keller@nasa.gov

Need models to fill gaps in observations



TOAR (Schulz et al., 2017)

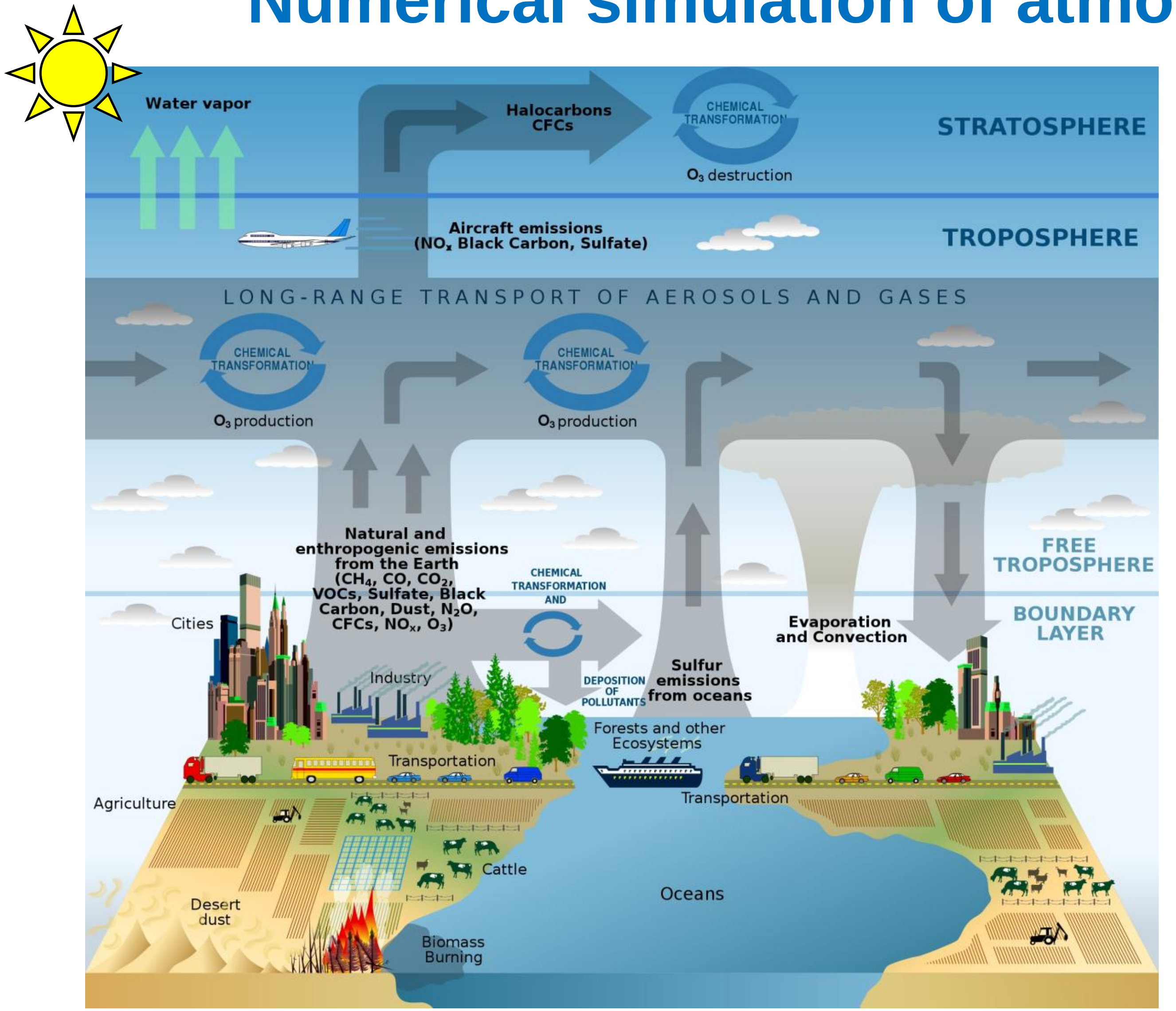


Surface observations are not global

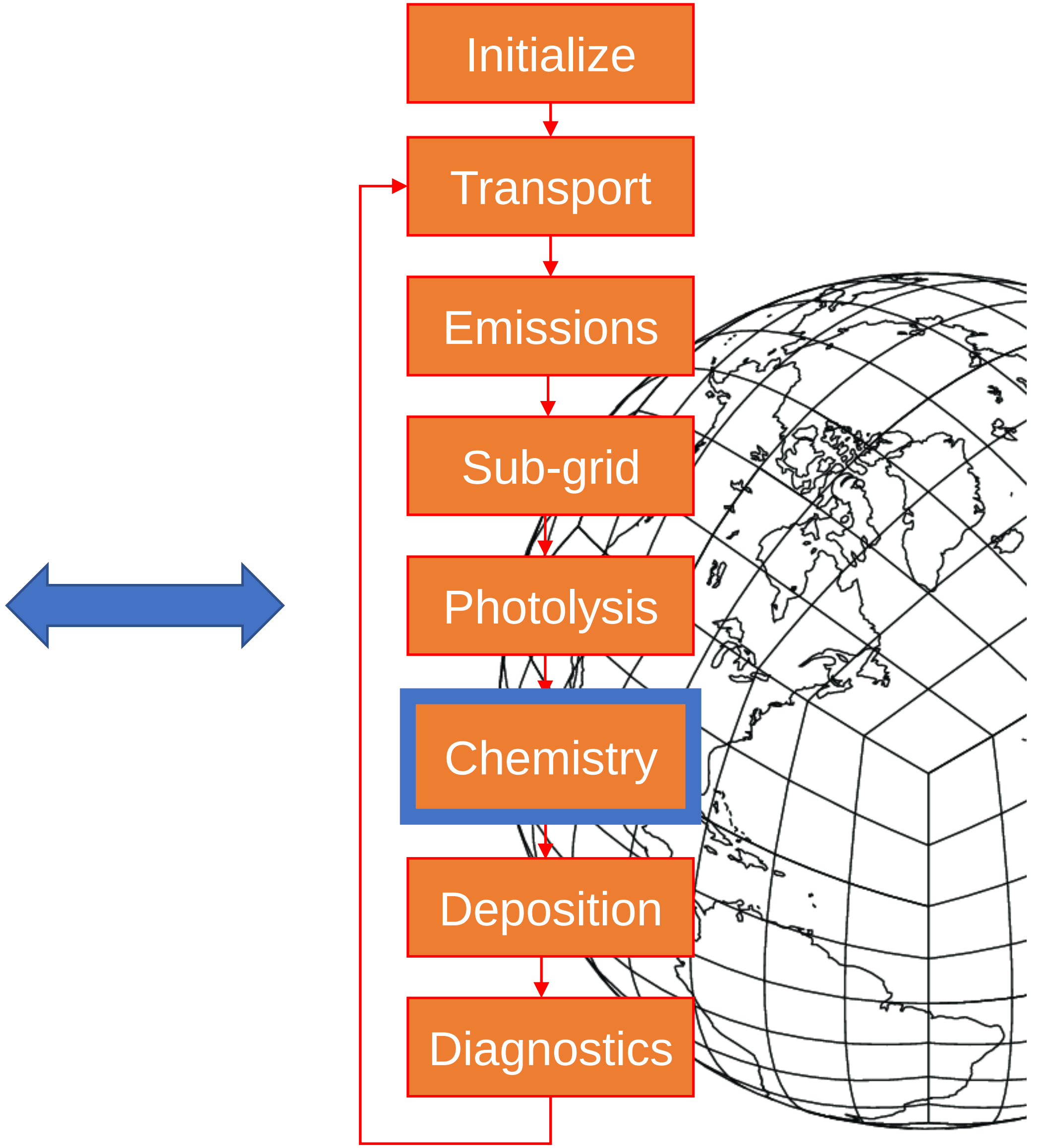
Satellite observations are also discontinuous

christoph.a.keller@nasa.gov

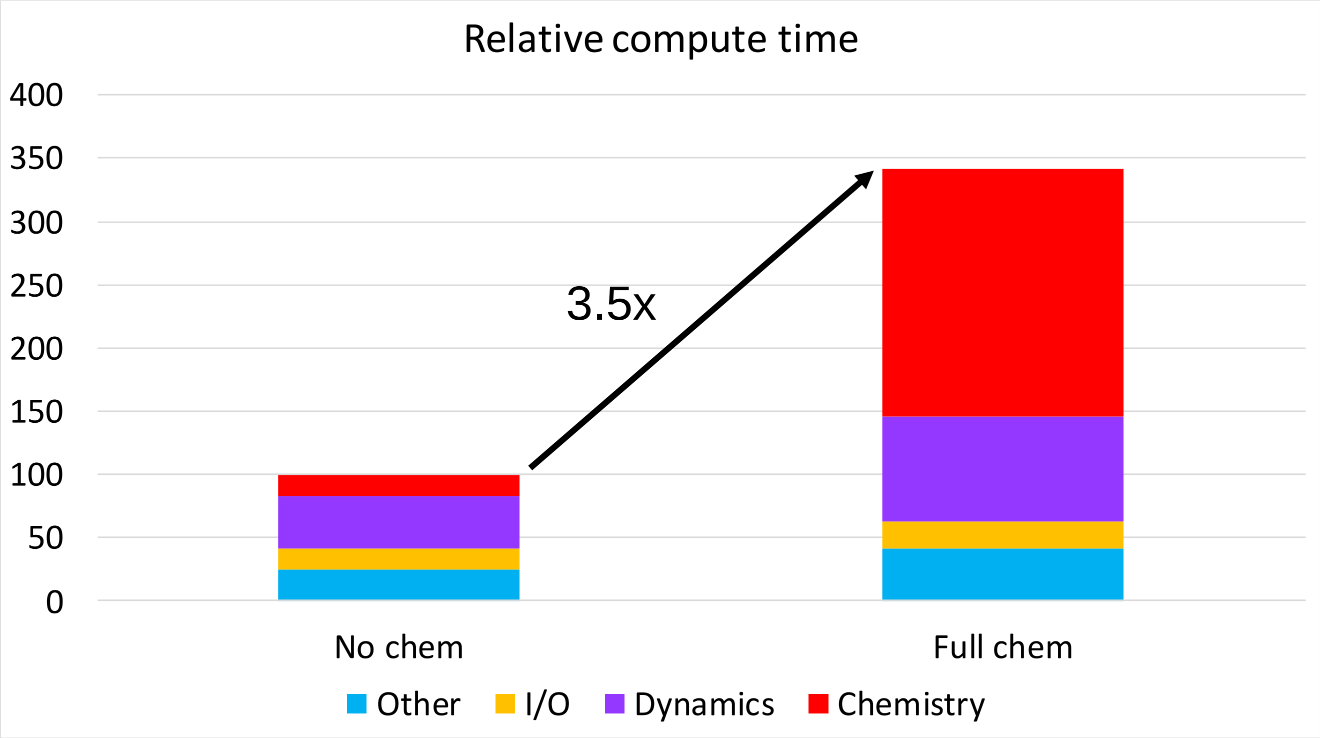
Numerical simulation of atmospheric composition



https://digital.library.unt.edu/ark:/67531/metadc11954/m1/37/



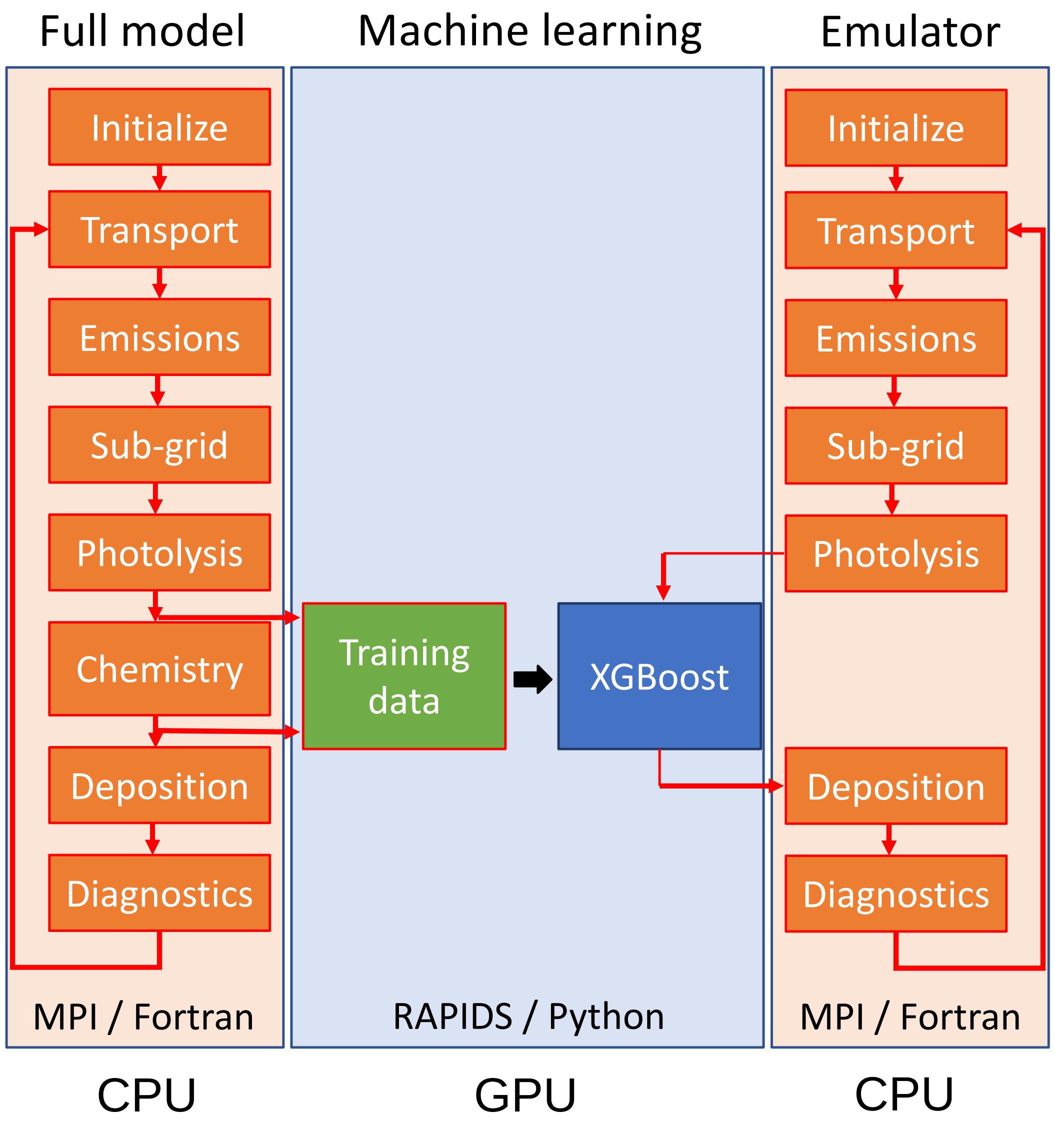
Atmospheric chemistry models are computationally expensive



- High-resolution chemistry simulation requires >1000 CPU's
- Throughput: approx. 20 simulation days in 24 hours
- Outputting the full chemical state: ~1.5 TB / simulation day

christoph.a.keller@nasa.gov

Replace slow chemical integrator with machine learning model



christoph.a.keller@nasa.gov

Use machine learning to emulate chemical transformations in the atmosphere

Inputs

Output

Meteorology:
- 7 variables

Chemistry:
- 143 chemical species
- 91 photolysis rates

ML →

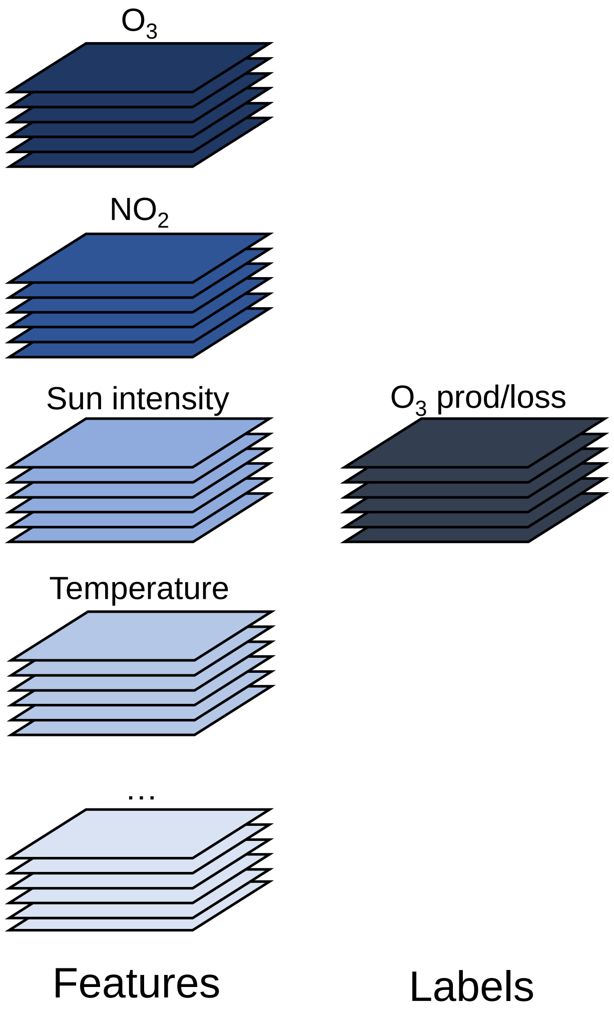
Chemical production / loss

- ➢ Algorithm: extreme gradient boosted decision trees (XGBoost)
- ➢ Train separate algorithm for each species

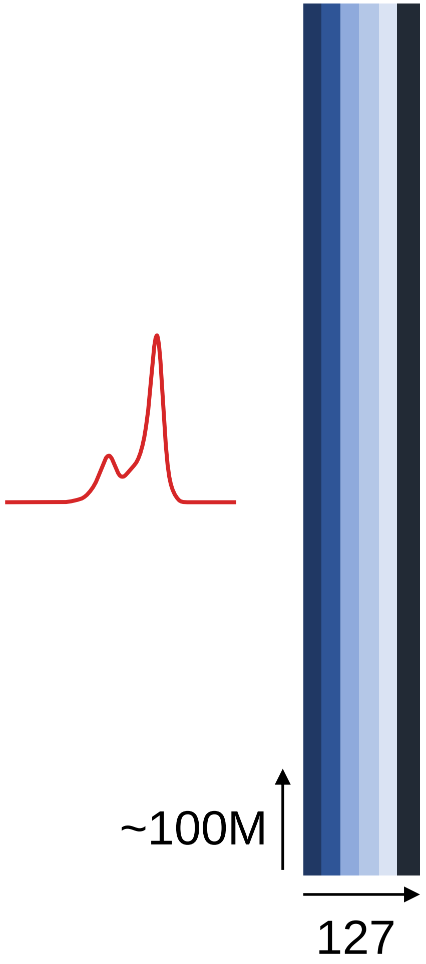
christoph.a.keller@nasa.gov

Machine learning workflow

Training data (netCDF format)



Subsample & flatten, write to csv (xarray)



Train (XGBoost):

- Read csv, convert to DMatrix
- Train

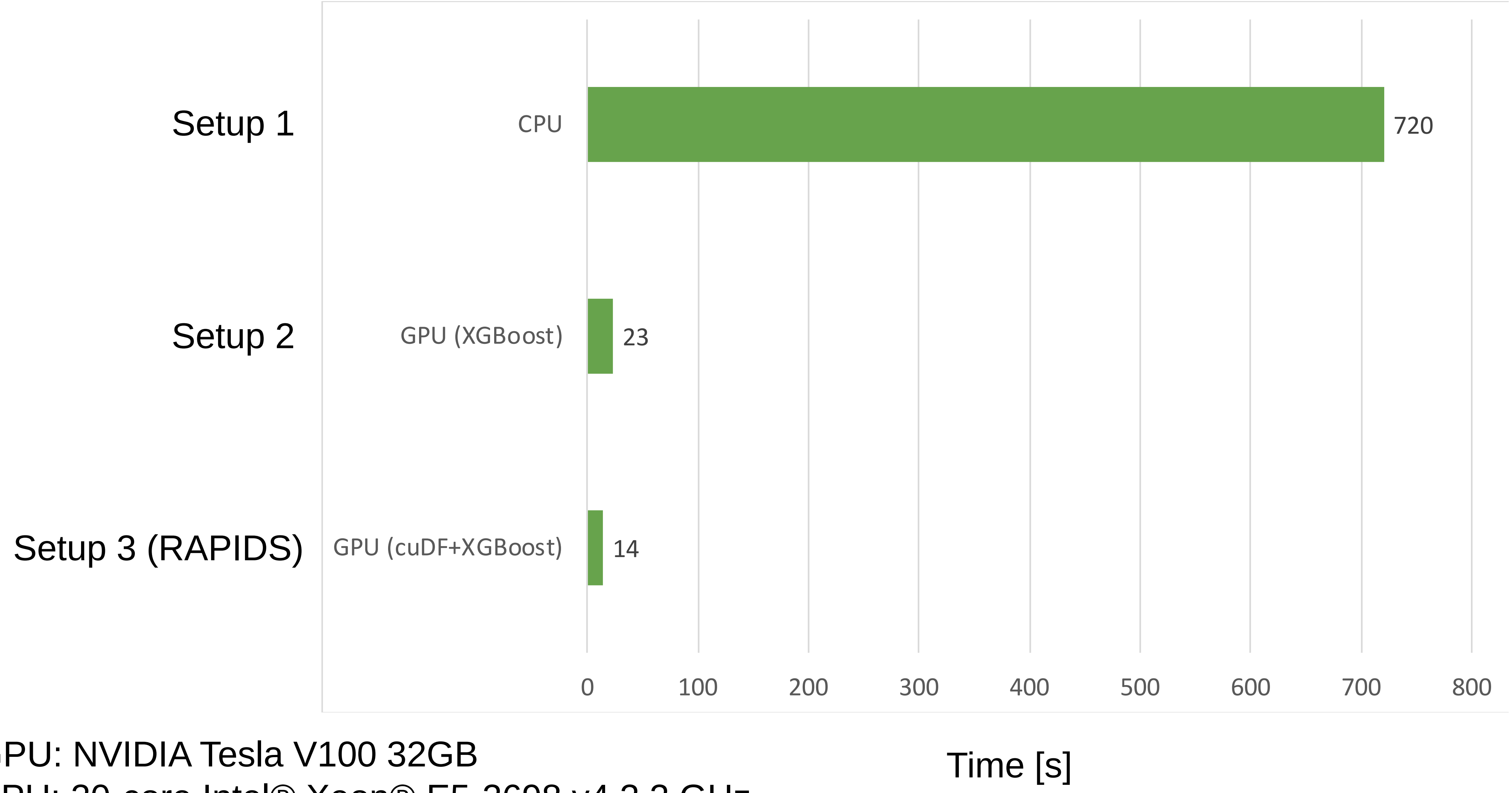
Setup 1
Read on CPU (Intel Haswell)
Train on CPU

Setup 2:
Read on CPU
Train on GPU (V100)

Setup 3 (RAPIDS):
Read on GPU (cuDF/cuIO)
Train on GPU (dask-XGBoost)

christoph.a.keller@nasa.gov

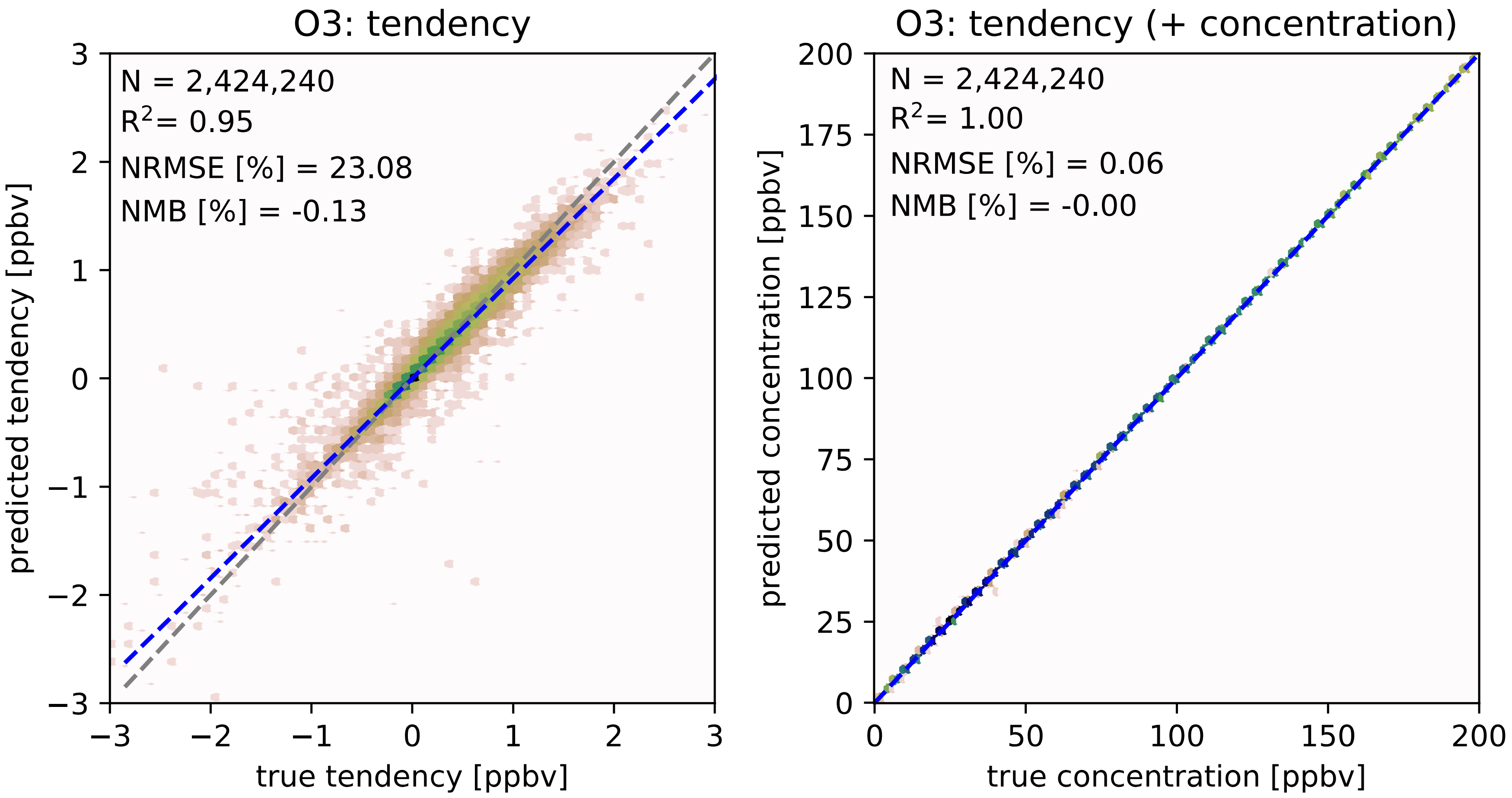
XGBoost training benchmarks



GPU: NVIDIA Tesla V100 32GB
CPU: 20-core Intel® Xeon® E5-2698 v4 2.2 GHz

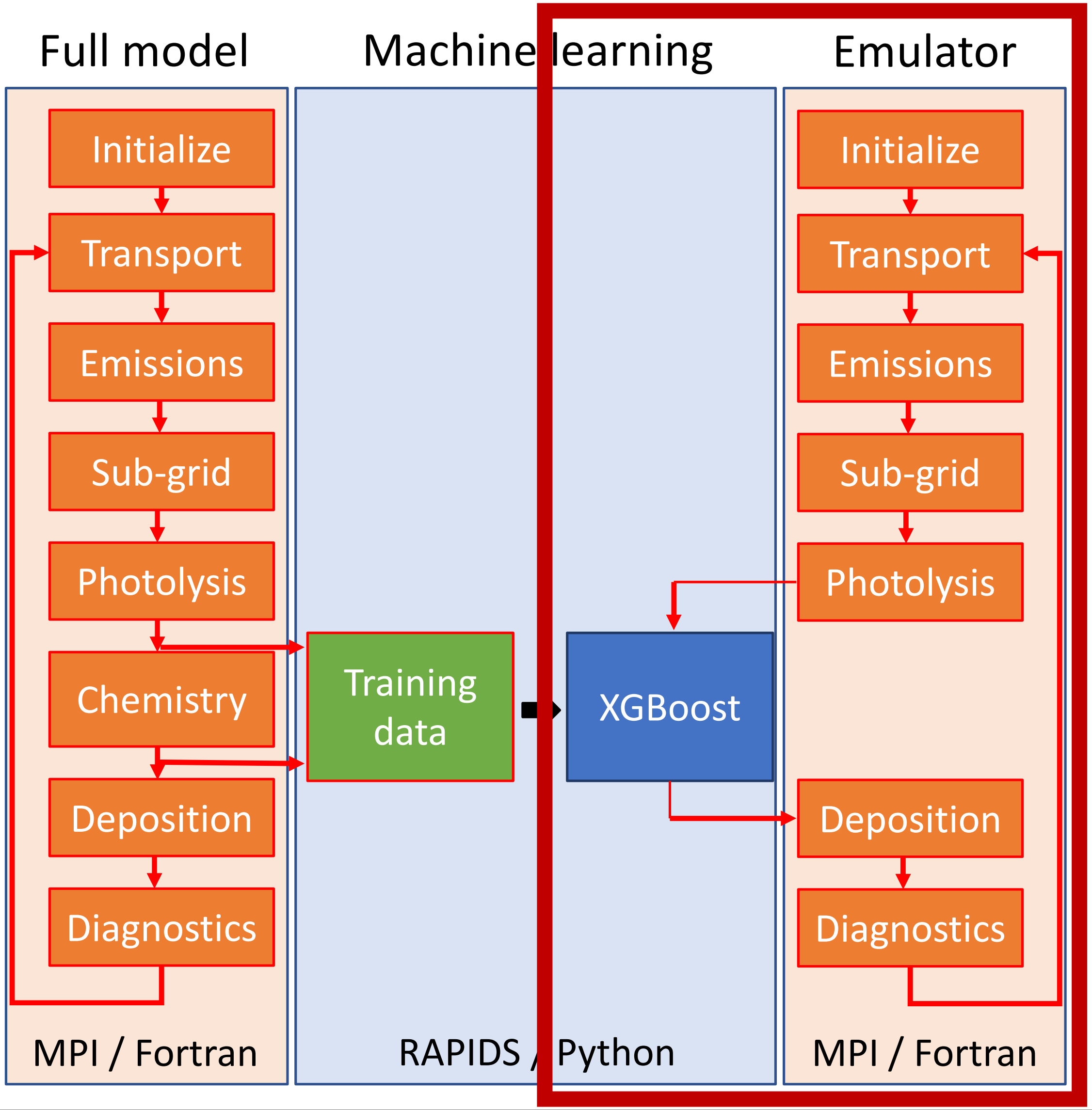
christoph.a.keller@nasa.gov

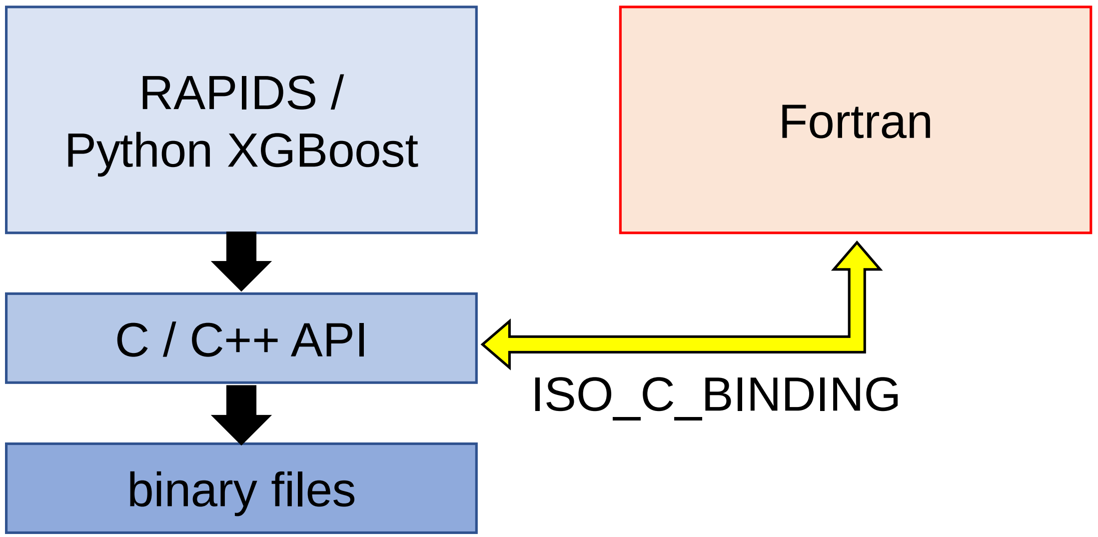
Random forest / XGBoost reproduce target concentrations well (single-step prediction)



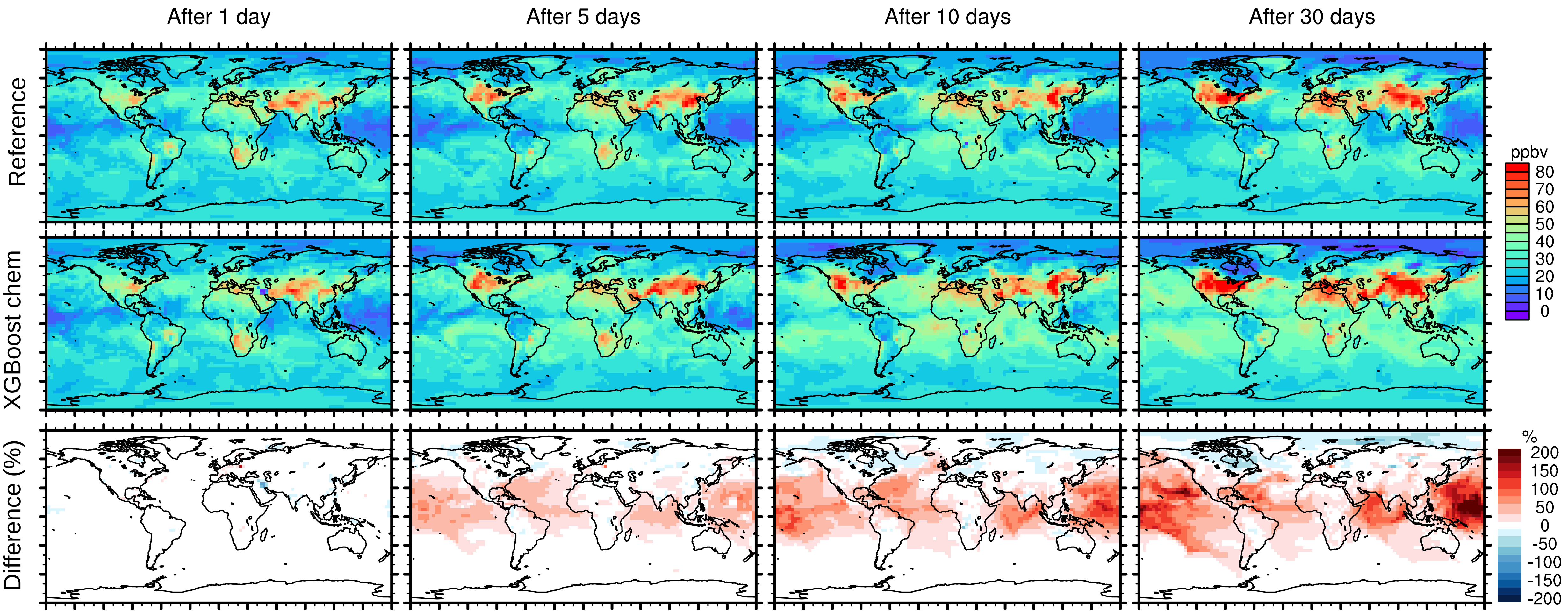
christoph.a.keller@nasa.gov

1-month simulation with XGBoost emulator

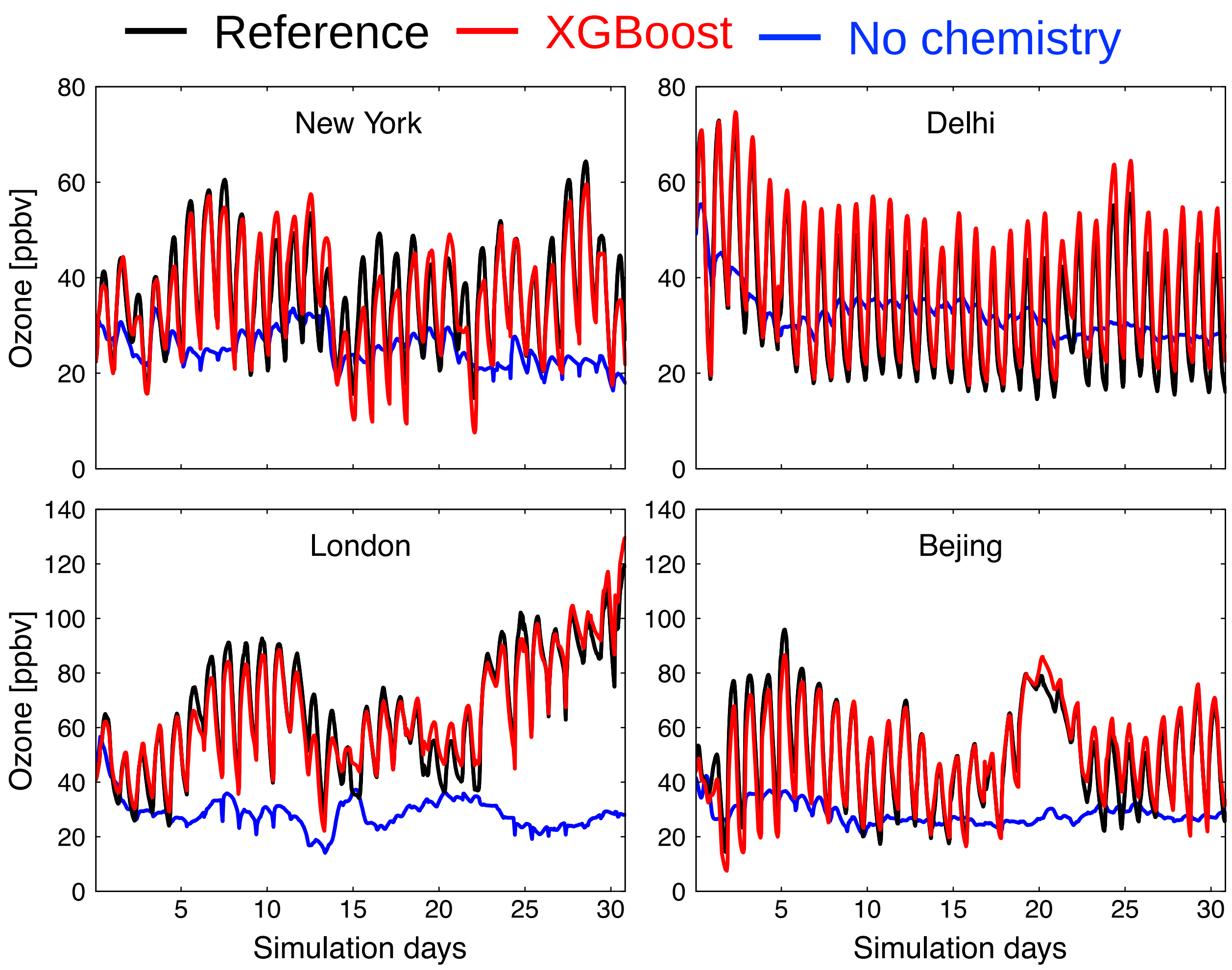




Emulator model is generally accurate, but overestimates ozone concentrations over remote regions

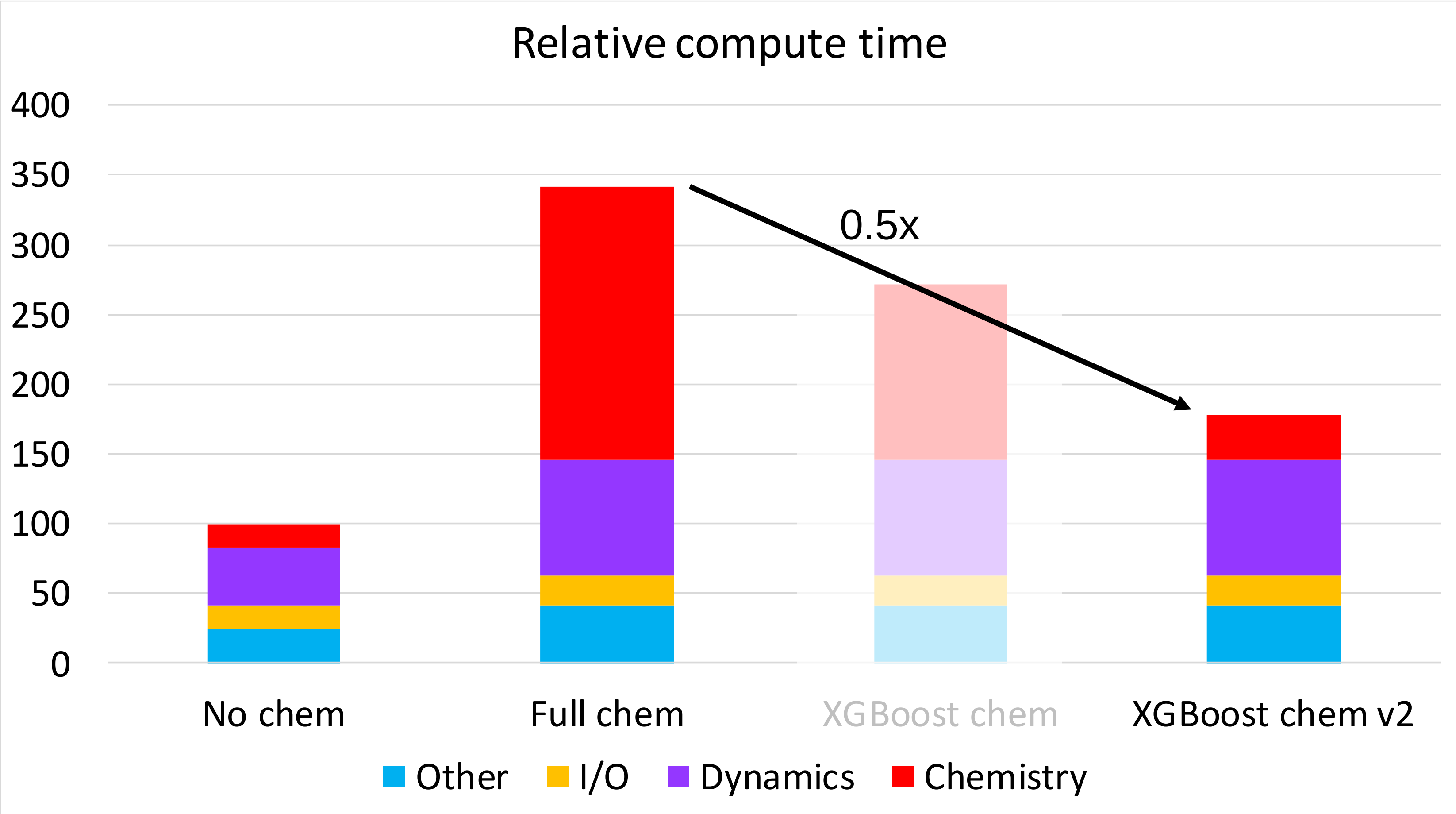


Surface concentrations over polluted regions are well reproduced by ML model



christoph.a.keller@nasa.gov

Model speedup with optimized XGBoost model



- XGBoost chemistry model is now ~2 times faster than reference model
- Chemistry >6x faster than before, dynamics becomes bottleneck

christoph.a.keller@nasa.gov

Summary

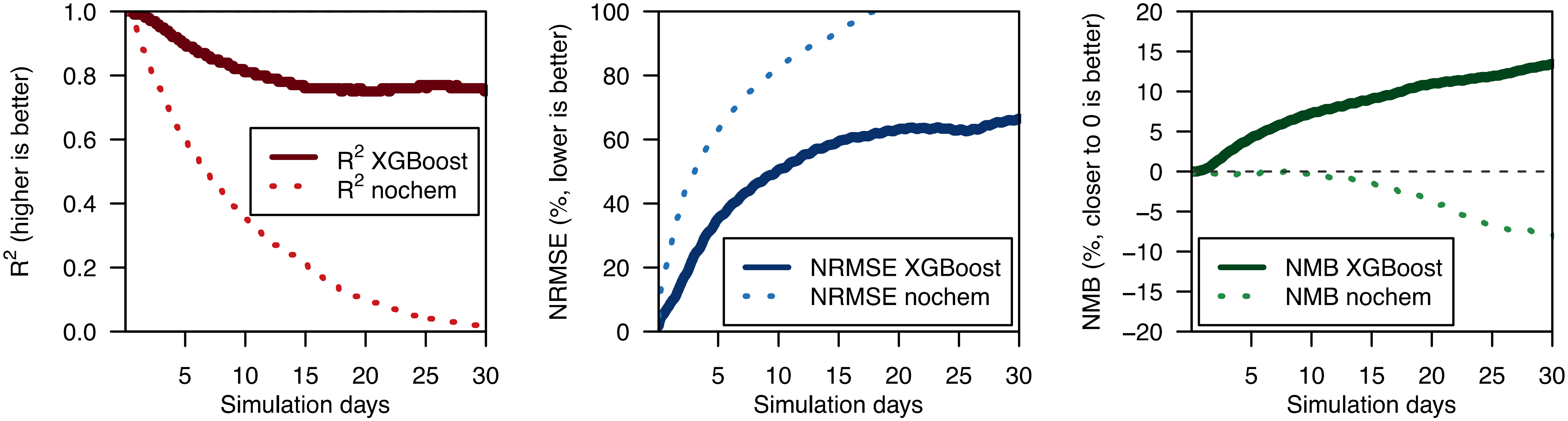
- Machine learning can help speed up air quality models by at least 2-5x
- Benefits:
  - Better use of satellite observations
  - Improve (short to medium-term) air quality forecasts
- Ongoing work:
  - Train on very large data sets (>1 TB)
  - Better coupling between CPU and GPUs (model side)
  - Dynamics for >200 chemical species is still slow

Keller and Evans: Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10, GMD, 2019.

christoph.a.keller@nasa.gov

christoph.a.keller@nasa.gov

Machine learning model remains stable over the long-term



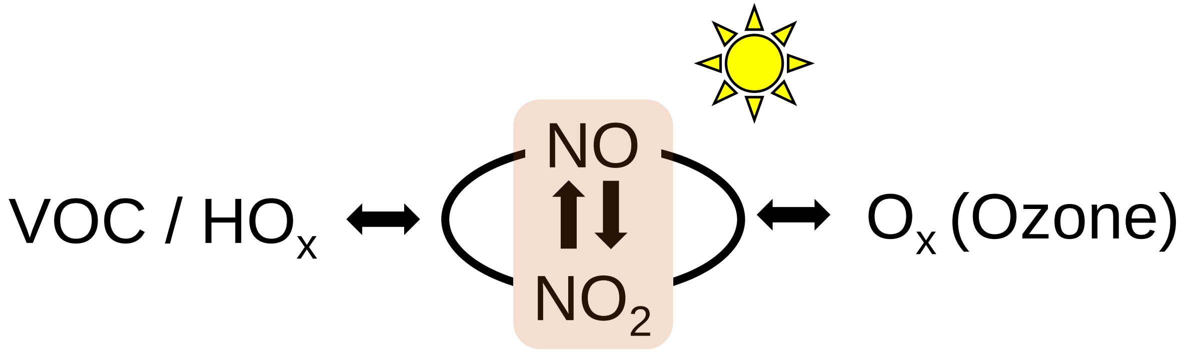
Impose chemical constraints on ML model to improve (long-term) accuracy

1. Distinguish between short-term vs. long-term species

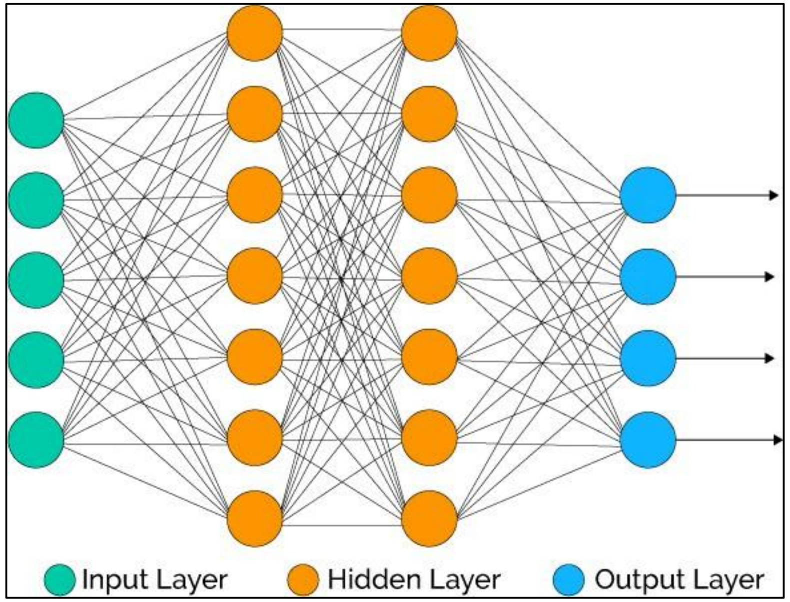
Long-lived (tendencies): $[X_i]_{T+\Delta T} = [X_i]_T + f(\,\mathbf{k, J, [X]}\,)$

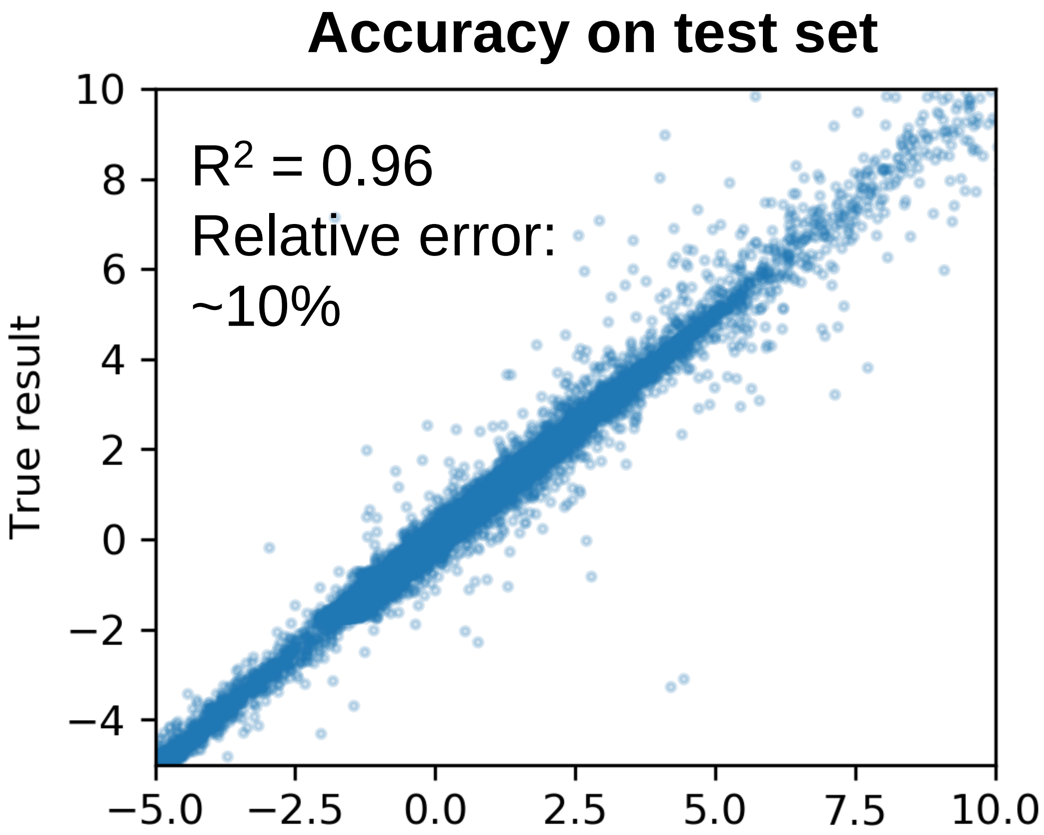
Short-lived (steady state): $[X_i]_{T+\Delta T} = f(\,\mathbf{k, J, [X]}\,)$

2. Predict NO + NO_2 combined (NOx family approach)



Neural network has weaker performance than tree methods

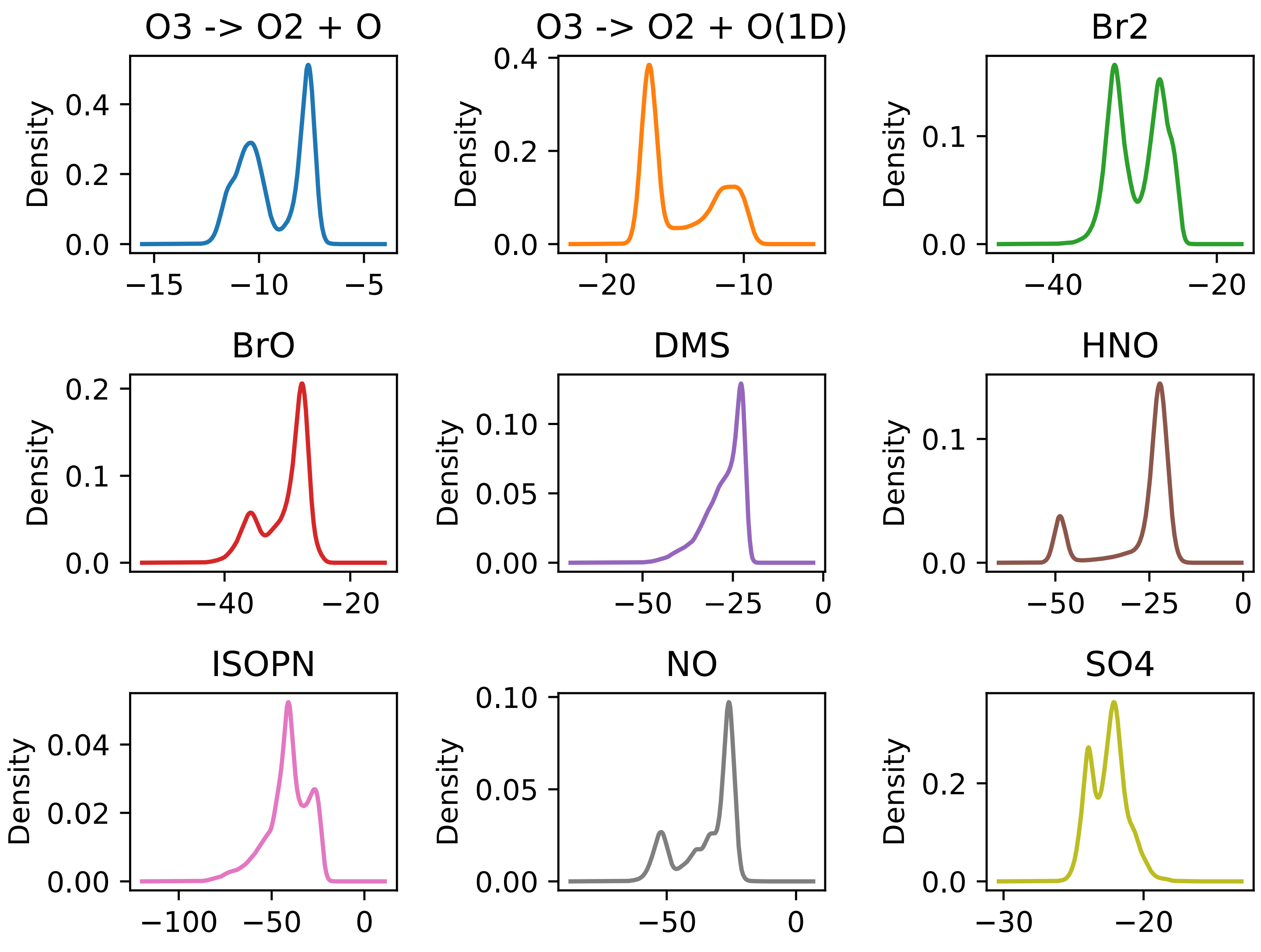




Courtesy of J. Zhuang, M. Kelp (Harvard)

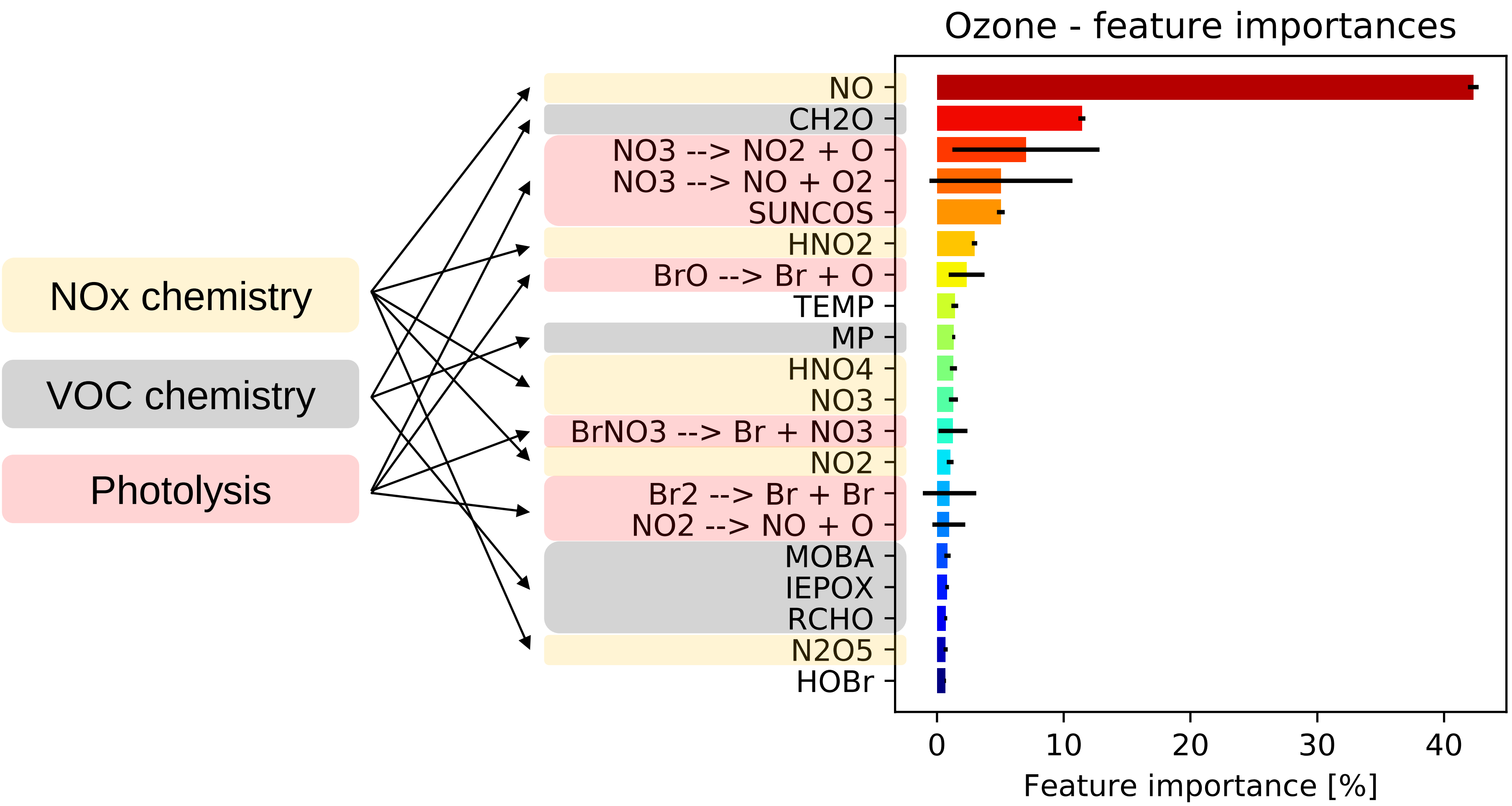
- 2 hidden layers, 300 neurons each
- Normalization? Split mechanism (e.g. day/night)?

Many input features have multiple modes



christoph.a.keller@nasa.gov

Random forest / XGBoost solutions reflect known features of chemical kinetics



christoph.a.keller@nasa.gov

NCCS GPU Clusters

Hardware

- Legacy ML Cluster:
    - 24x2 CPU Cores
    - 512GB Memory
    - 2 Nodes with 4x16GB NVIDIA V100 GPUs each
    - 1 Node with 8x16GB NVIDIA V100 GPUs
- (New) GPU Cluster:
    - 20x2 CPU Cores
    - 768GB Memory
    - 22 Nodes with 4x32GB NVIDIA V100 GPUs

Software

- CentOS 7
- NVIDA Software
  - CUDA
  - cuDNN
  - NCCL
- Anaconda3
  - We have a few curated environments with some of the standard ML frameworks.
  - Users can create and customize personal environments.
  - https://www.nccs.nasa.gov/nccs-users/instructional/Anaconda
  - https://www.nccs.nasa.gov/sites/default/files/Anaconda_on_Adapt_0.pdf
- Modules (Compilers/MPI/Misc)
- JupyterHub
  - jh-ml.nccs.nasa.gov

Gaining Access

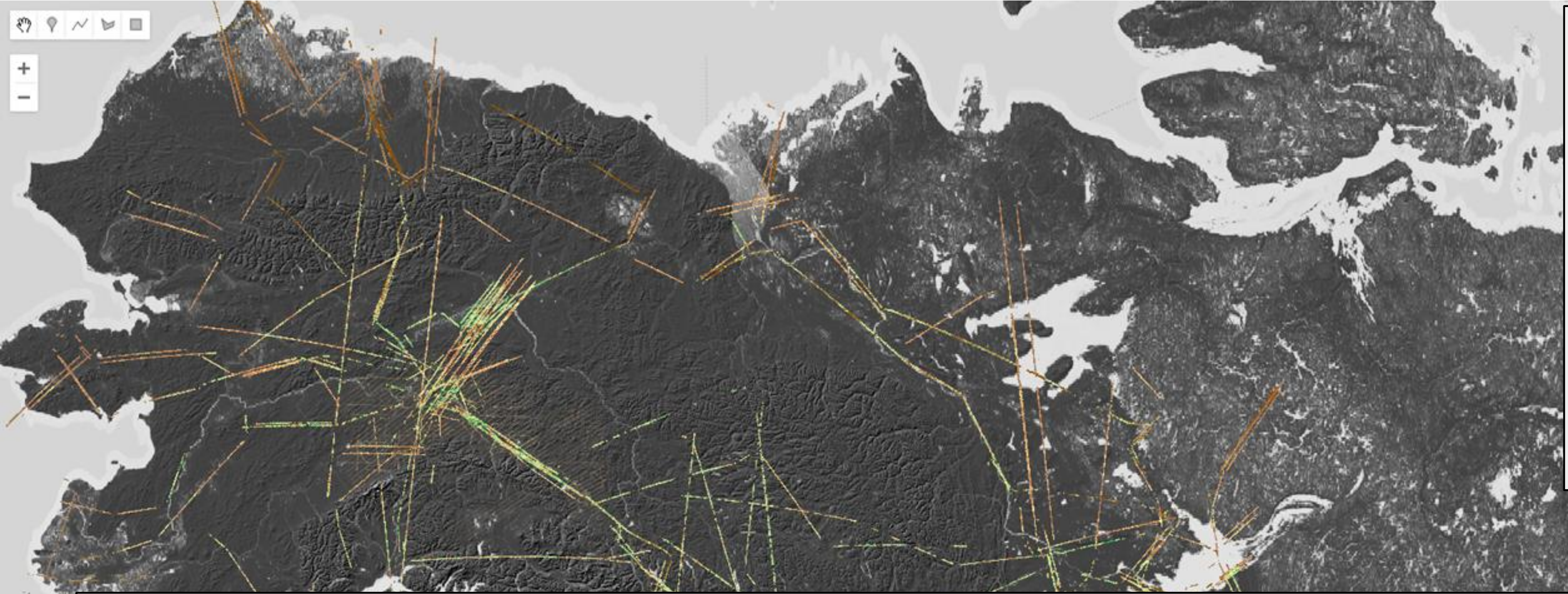
- I believe that you are just a support ticket away from access as existing users. Just ask for access to the ADAPT GPU resources by emailing support@nccs.nasa.gov

Using the GPUs

- GPUs are accessible through Slurm only.
  - sbatch
  - salloc

Questions?

https://www.nccs.nasa.gov/nccs-users/instructional/adapt-instructional/adapt-gpu/



ABoVE Vegetation Structure & Function Working Group

Paul Montesano
Matt Macander
Amanda Armstrong
Laura Duncanson
John Armston
Chris Neigh
Abhishek Chatterjee
Ben Poulter
Jan Eitel
Naiara Pinto
et al.

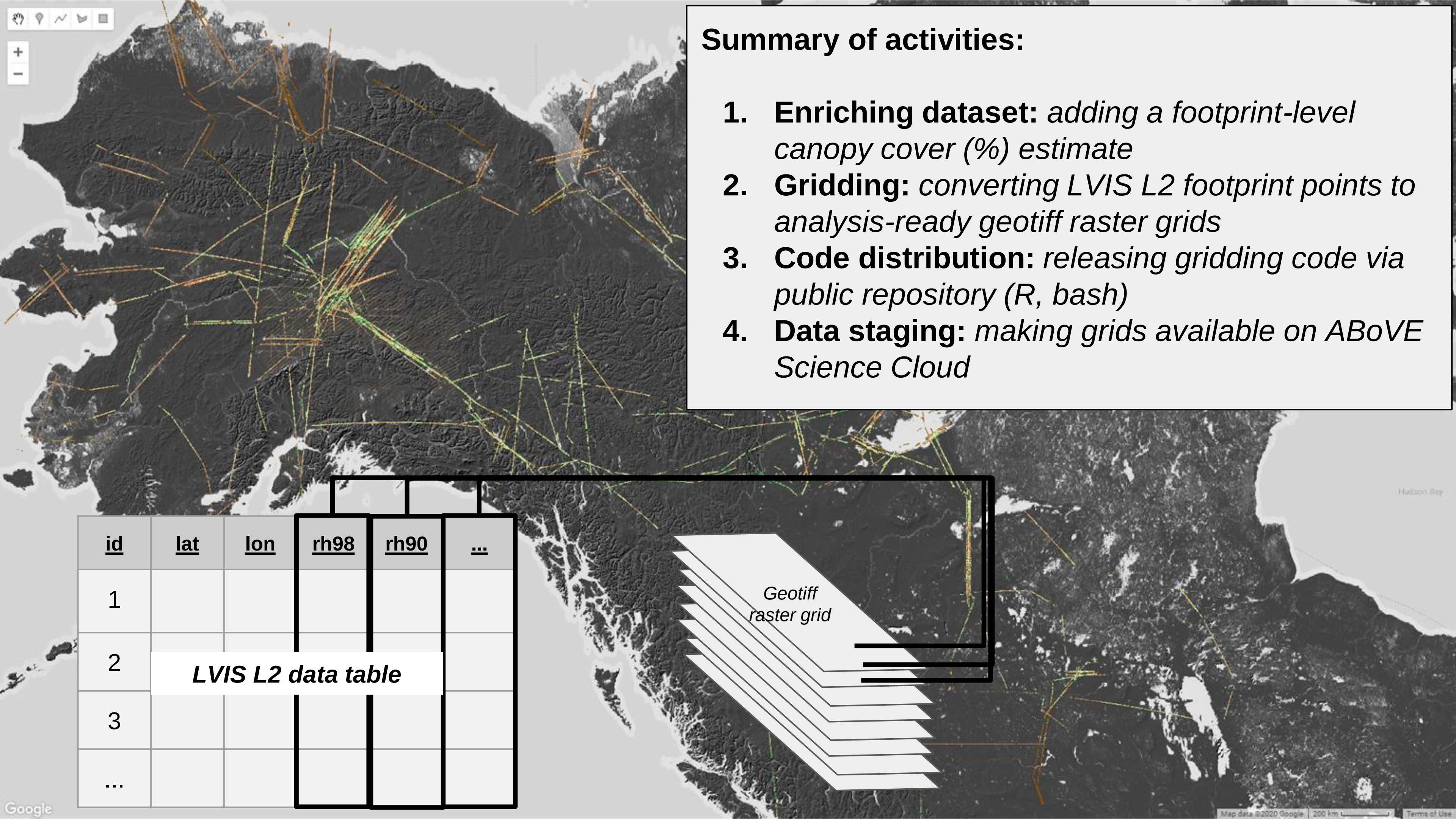
Gridded LVIS L2 vegetation structure estimates for ABoVE analyses

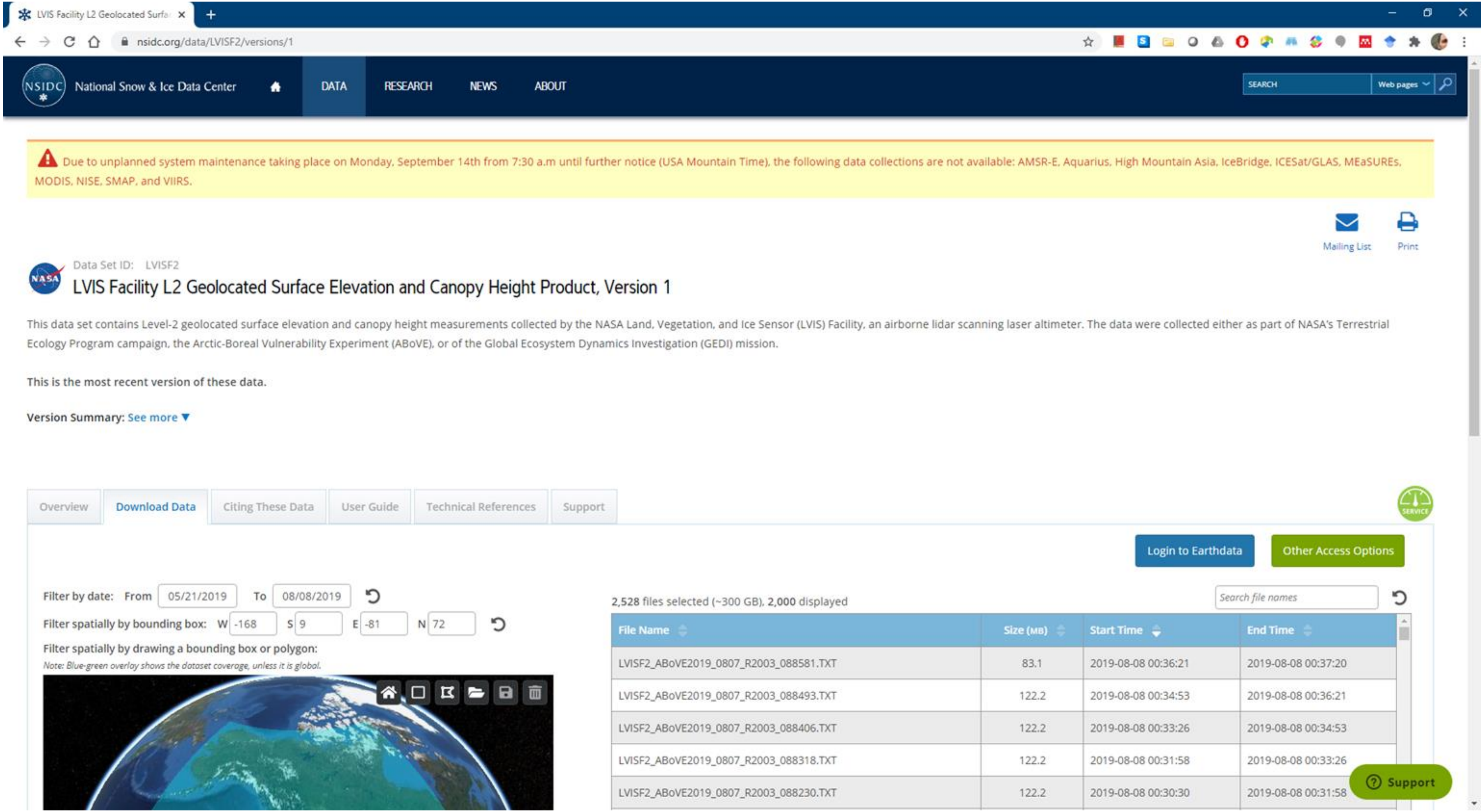
Paul Montesano | paul.m.montesano@nasa.gov
Matt Macander | mmacander@abrinc.com

Enriching LVIS Level 2 data of forest structure with estimates of cover and processing to analysis-ready grids for storage on ABoVE Science Cloud

Summary of activities:

1. **Enriching dataset:** *adding a footprint-level canopy cover (%) estimate*
2. **Gridding:** *converting LVIS L2 footprint points to analysis-ready geotiff raster grids*
3. **Code distribution:** *releasing gridding code via public repository (R, bash)*
4. **Data staging:** *making grids available on ABoVE Science Cloud*



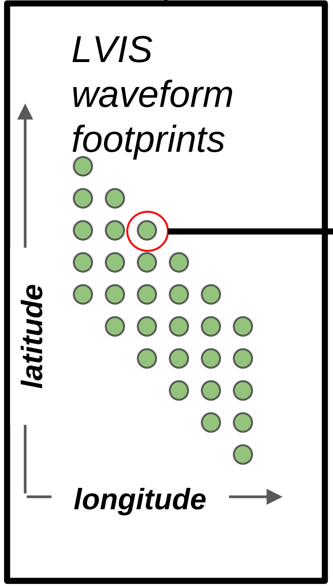

LVIS Facility L2 Geolocated Surfa
nsidc.org/data/LVISF2/versions/1
NSIDC National Snow & Ice Data Center
DATA
RESEARCH
NEWS
ABOUT
SEARCH
Web pages
Due to unplanned system maintenance taking place on Monday, September 14th from 7:30 a.m until further notice (USA Mountain Time), the following data collections are not available: AMSR-E, Aquarius, High Mountain Asia, IceBridge, ICESat/GLAS, MEaSUREs, MODIS, NISE, SMAP, and VIIRS.
Mailing List
Print
Data Set ID: LVISF2
LVIS Facility L2 Geolocated Surface Elevation and Canopy Height Product, Version 1
This data set contains Level-2 geolocated surface elevation and canopy height measurements collected by the NASA Land, Vegetation, and Ice Sensor (LVIS) Facility, an airborne lidar scanning laser altimeter. The data were collected either as part of NASA's Terrestrial Ecology Program campaign, the Arctic-Boreal Vulnerability Experiment (ABoVE), or of the Global Ecosystem Dynamics Investigation (GEDI) mission.
This is the most recent version of these data.
Version Summary: See more
Overview
Download Data
Citing These Data
User Guide
Technical References
Support
Login to Earthdata
Other Access Options
Filter by date: From 05/21/2019 To 08/08/2019
Filter spatially by bounding box: W -168 S 9 E -81 N 72
Filter spatially by drawing a bounding box or polygon:
Note: Blue-green overlay shows the dataset coverage, unless it is global.
2,528 files selected (~300 GB), 2,000 displayed
Search file names
File Name
Size (MB)
Start Time
End Time
LVISF2_ABoVE2019_0807_R2003_088581.TXT 83.1 2019-08-08 00:36:21 2019-08-08 00:37:20
LVISF2_ABoVE2019_0807_R2003_088493.TXT 122.2 2019-08-08 00:34:53 2019-08-08 00:36:21
LVISF2_ABoVE2019_0807_R2003_088406.TXT 122.2 2019-08-08 00:33:26 2019-08-08 00:34:53
LVISF2_ABoVE2019_0807_R2003_088318.TXT 122.2 2019-08-08 00:31:58 2019-08-08 00:33:26
LVISF2_ABoVE2019_0807_R2003_088230.TXT 122.2 2019-08-08 00:30:30 2019-08-08 00:31:58
Support

What is LVIS L2 lidar data?

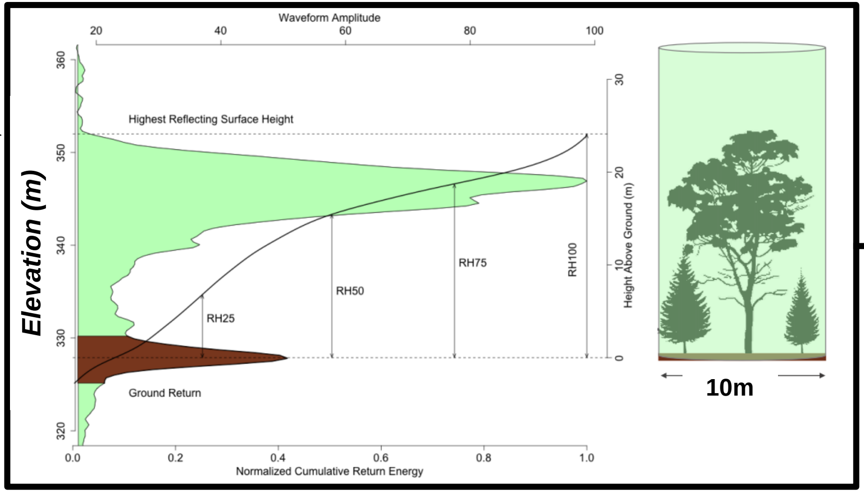
https://nsidc.org/data/LVISF2/versions/1

Data Set ID: LVISF2

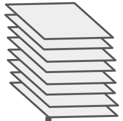
LVIS Facility L2 Geolocated Surface Elevation and Canopy Height Product, Version 1



Waveform example

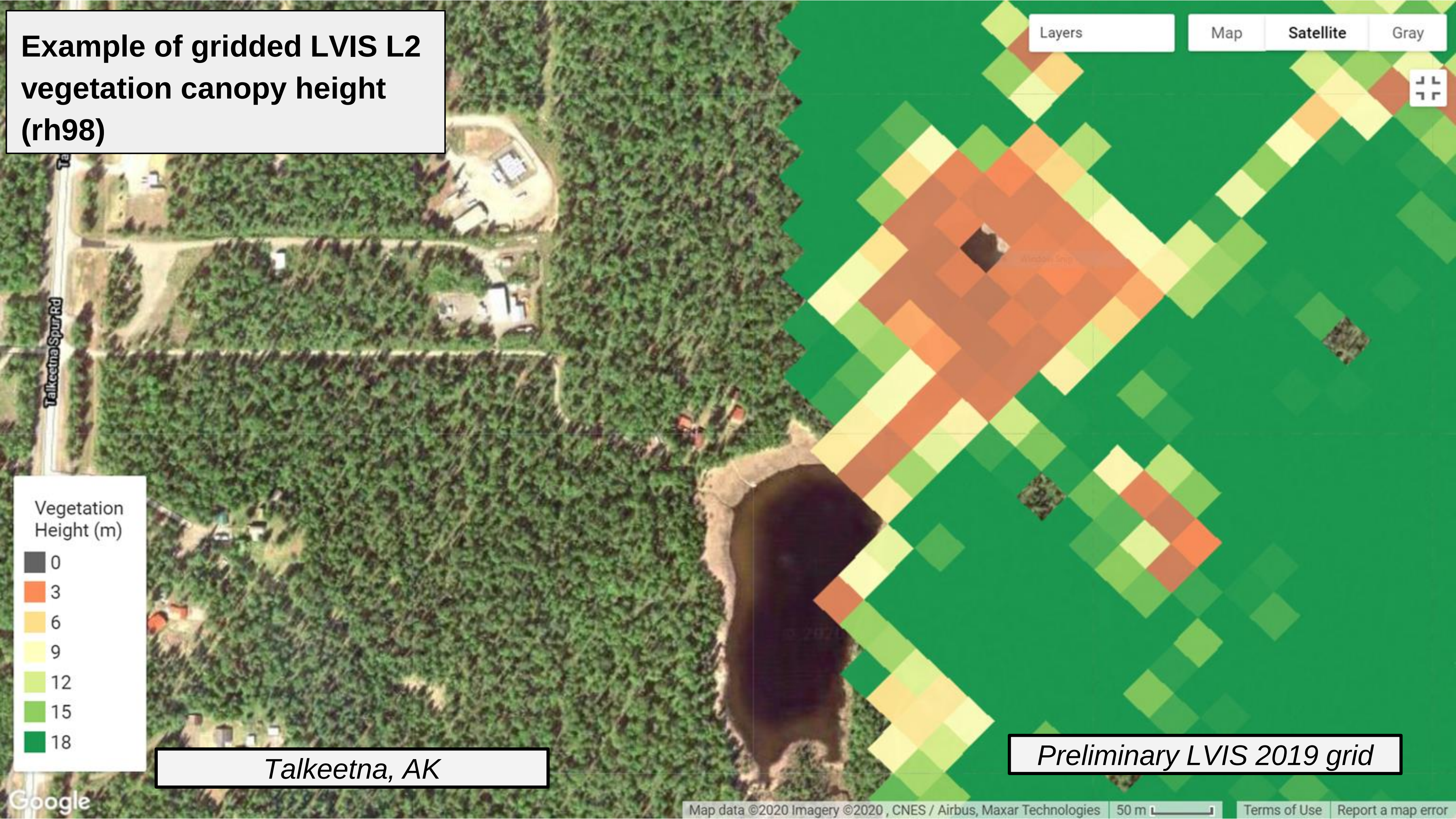


Gridded metrics (geotiff)



id	lat	lon	rh98	rh90	...
1					
2					
3					
...					

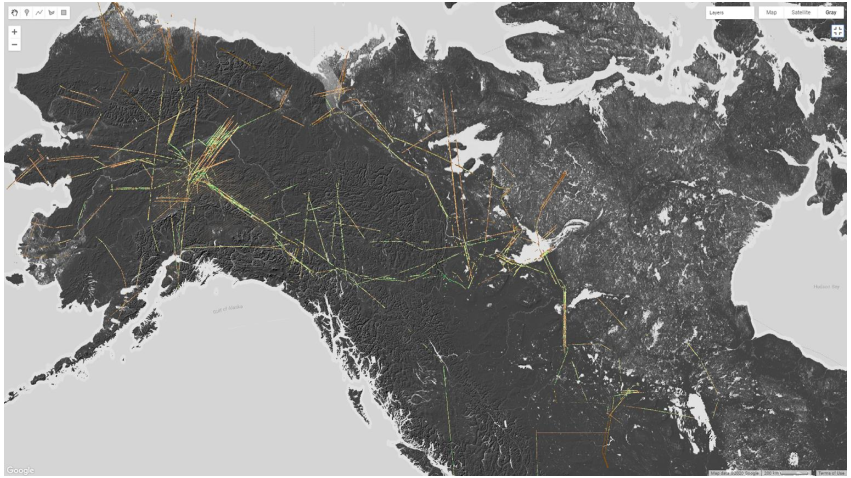
Example of gridded LVIS L2 vegetation canopy height (rh98)



Talkeetna, AK

Preliminary LVIS 2019 grid

ABoVE LVIS grids: accessible on ADAPT (ASC)



Main directory:

/att/pubrepo/ABoVE/intermediate/LVIS/

code/ → *R code will be placed here*

Organized by year of airborne campaign:

2019/

2017/

Each year's 'output' directory holds 3 subdirs:

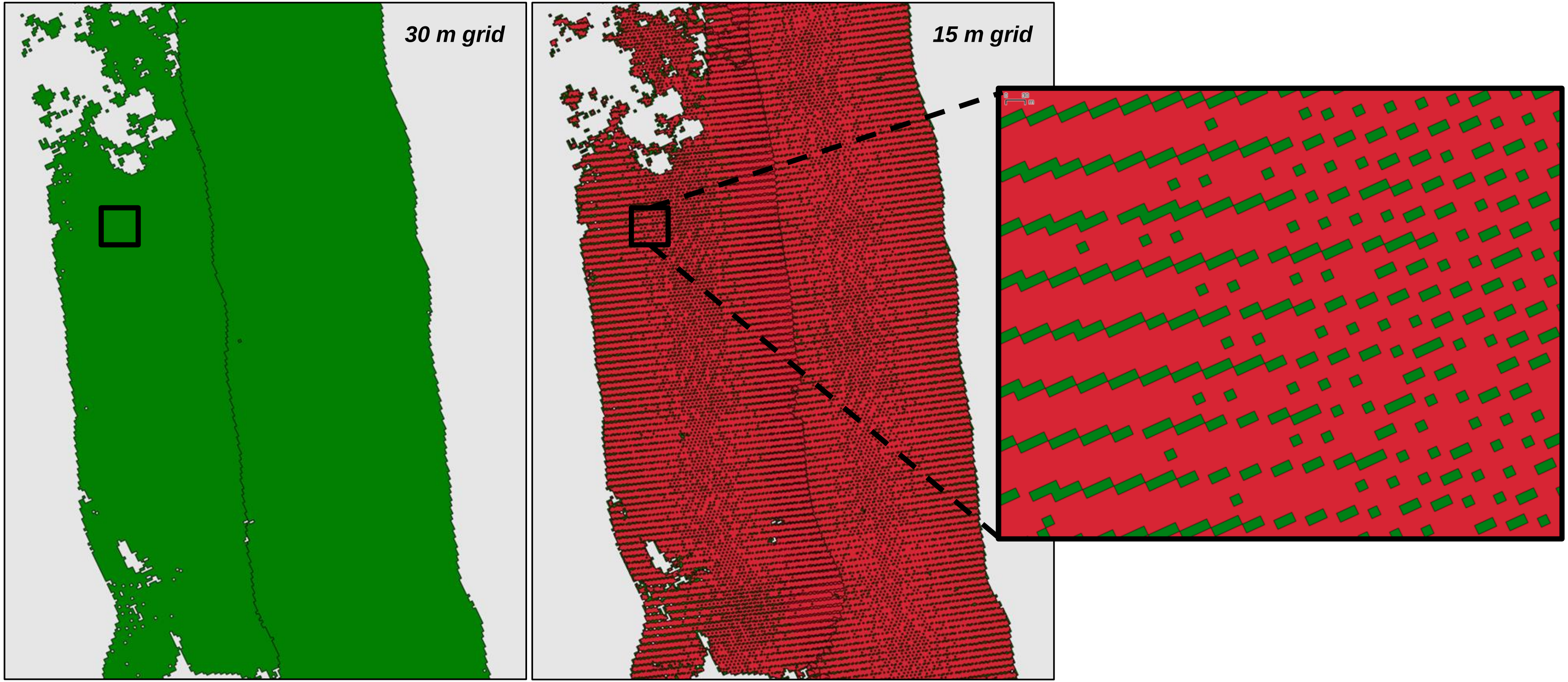
2019/output/

_footprints/ → spatial 'footprints' of each gridded flightline (shp | kml)

metrics/ → gridded LVIS L2 metrics by ***flightline name***

metrics_vrts/ → virtual mosaics of all flightlines by ***metric name***

_**footprints/**: spatial footprints (shp | kml)



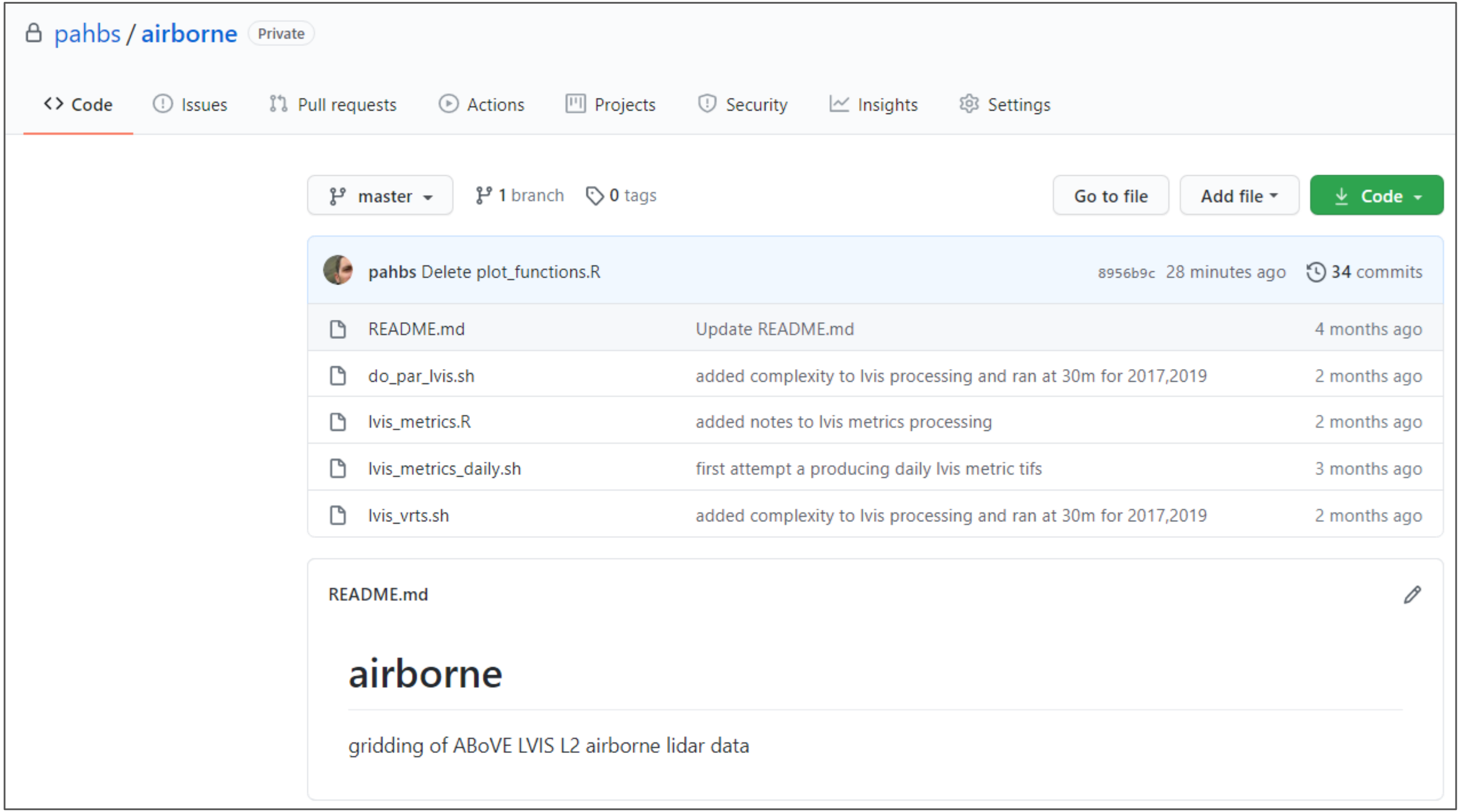
metrics/: grids organized by LVIS flightline name

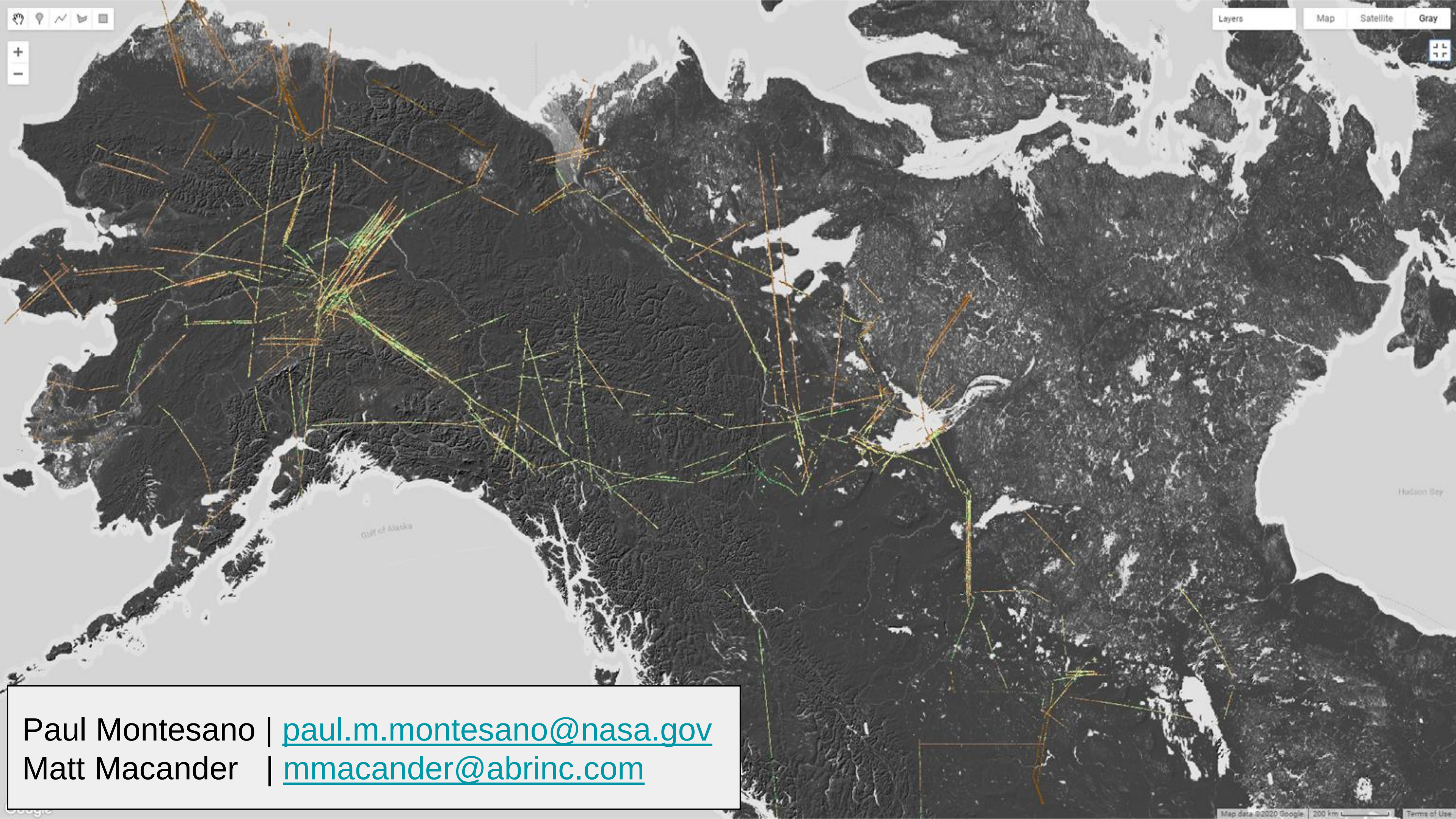
31 grids of metics for each flightline name

- 23 relative height metrics (*RH*.tif)
- 5 canopy cover proxies (*retge*.tif)
- 1 canopy complexity (*COMPLEXITY*.tif)
- 1 ground elevation (*ZG*.tif)
- 1 LVIS footprint count (*lvis_pt_cnt*.tif)

```
[pmontesa@adaptlogin101 15]$ ls -lht LVISF2_ABoVE2019_0723_R2003_061278*
-rw-r--r-- 1 pmontesa s1782 110K Jul 15 12:48 LVISF2_ABoVE2019_0723_R2003_061278_retge_5p00_15m.tif
-rw-r--r-- 1 pmontesa s1782 124K Jul 15 12:47 LVISF2_ABoVE2019_0723_R2003_061278_retge_2p00_15m.tif
-rw-r--r-- 1 pmontesa s1782 127K Jul 15 12:45 LVISF2_ABoVE2019_0723_R2003_061278_retge_1p50_15m.tif
-rw-r--r-- 1 pmontesa s1782 128K Jul 15 12:44 LVISF2_ABoVE2019_0723_R2003_061278_retge_1p37_15m.tif
-rw-r--r-- 1 pmontesa s1782 135K Jul 15 12:42 LVISF2_ABoVE2019_0723_R2003_061278_retge_1p00_15m.tif
-rw-r--r-- 1 pmontesa s1782 617K Jul 15 12:41 LVISF2_ABoVE2019_0723_R2003_061278_RH100_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 607K Jul 15 12:41 LVISF2_ABoVE2019_0723_R2003_061278_RH099_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 598K Jul 15 12:40 LVISF2_ABoVE2019_0723_R2003_061278_RH098_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 597K Jul 15 12:40 LVISF2_ABoVE2019_0723_R2003_061278_RH097_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 595K Jul 15 12:39 LVISF2_ABoVE2019_0723_R2003_061278_RH096_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 591K Jul 15 12:39 LVISF2_ABoVE2019_0723_R2003_061278_RH095_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 576K Jul 15 12:39 LVISF2_ABoVE2019_0723_R2003_061278_RH090_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 566K Jul 15 12:38 LVISF2_ABoVE2019_0723_R2003_061278_RH085_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 556K Jul 15 12:38 LVISF2_ABoVE2019_0723_R2003_061278_RH080_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 552K Jul 15 12:38 LVISF2_ABoVE2019_0723_R2003_061278_RH075_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 539K Jul 15 12:37 LVISF2_ABoVE2019_0723_R2003_061278_RH070_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 535K Jul 15 12:37 LVISF2_ABoVE2019_0723_R2003_061278_RH065_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 534K Jul 15 12:37 LVISF2_ABoVE2019_0723_R2003_061278_RH060_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 534K Jul 15 12:36 LVISF2_ABoVE2019_0723_R2003_061278_RH055_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 534K Jul 15 12:36 LVISF2_ABoVE2019_0723_R2003_061278_RH050_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 521K Jul 15 12:35 LVISF2_ABoVE2019_0723_R2003_061278_RH045_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 532K Jul 15 12:35 LVISF2_ABoVE2019_0723_R2003_061278_RH040_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 530K Jul 15 12:35 LVISF2_ABoVE2019_0723_R2003_061278_RH035_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 541K Jul 15 12:34 LVISF2_ABoVE2019_0723_R2003_061278_RH030_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 536K Jul 15 12:34 LVISF2_ABoVE2019_0723_R2003_061278_RH025_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 538K Jul 15 12:34 LVISF2_ABoVE2019_0723_R2003_061278_RH020_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 540K Jul 15 12:33 LVISF2_ABoVE2019_0723_R2003_061278_RH015_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 542K Jul 15 12:33 LVISF2_ABoVE2019_0723_R2003_061278_RH010_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 742K Jul 15 12:33 LVISF2_ABoVE2019_0723_R2003_061278_ZG_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 675K Jul  8 12:38 LVISF2_ABoVE2019_0723_R2003_061278_COMPLEXITY_mean_15m.tif
-rw-r--r-- 1 pmontesa s1782 135K Jul  8 12:37 LVISF2_ABoVE2019_0723_R2003_061278_lvis_pt_cnt_15m.tif
[pmontesa@adaptlogin101 15]$ ls -lht LVISF2_ABoVE2019_0723_R2003_061278* | wc -l
31
[pmontesa@adaptlogin101 15]$ ls -lht LVISF2_ABoVE2019_0723_R2003_061278*RH* | wc -l
23
[pmontesa@adaptlogin101 15]$ pwd
/att/pubrepo/ABoVE/intermediate/LVIS/2019/output/metrics/15
[pmontesa@adaptlogin101 15]$
```

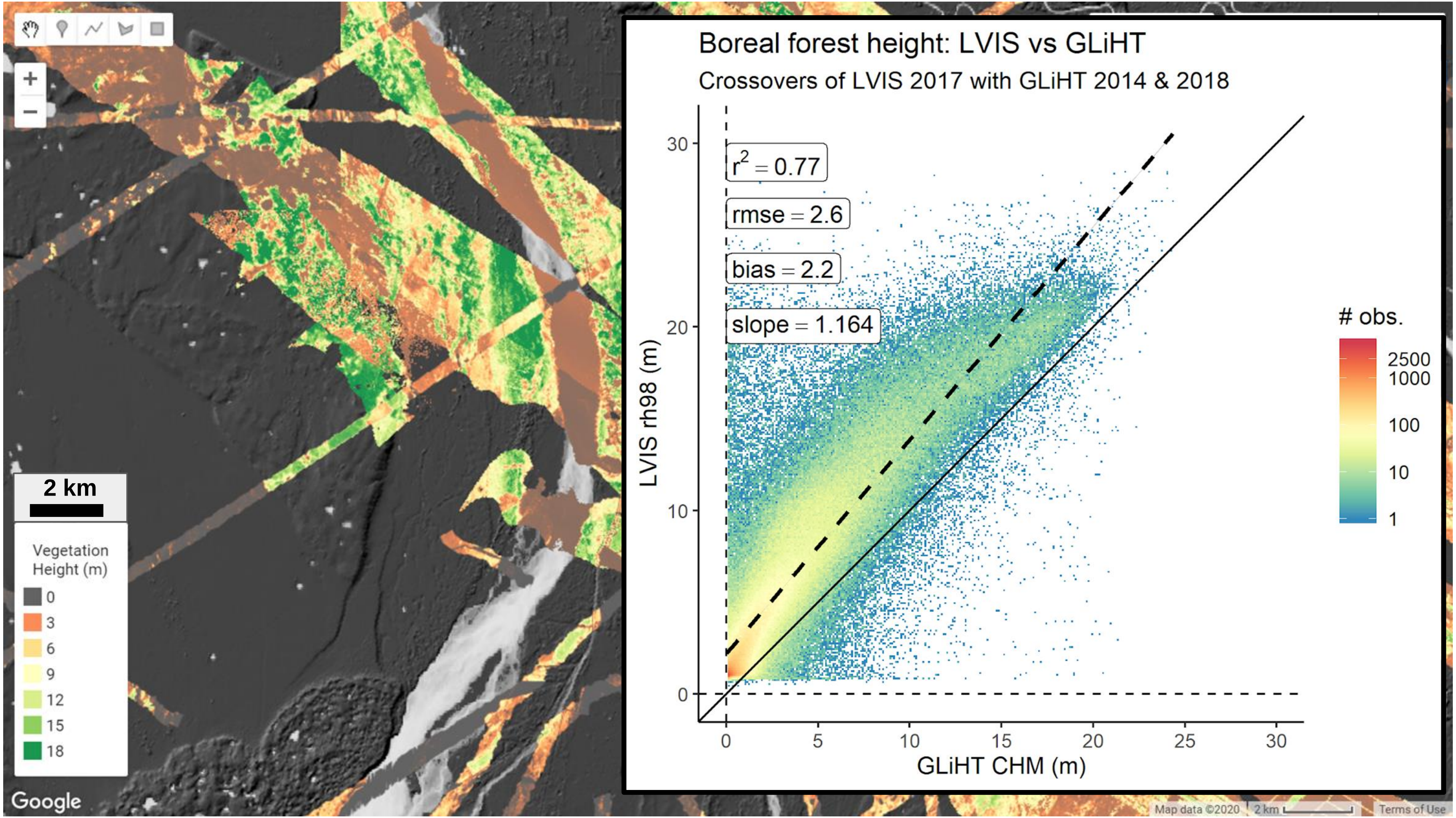
code/: R and shell scripts



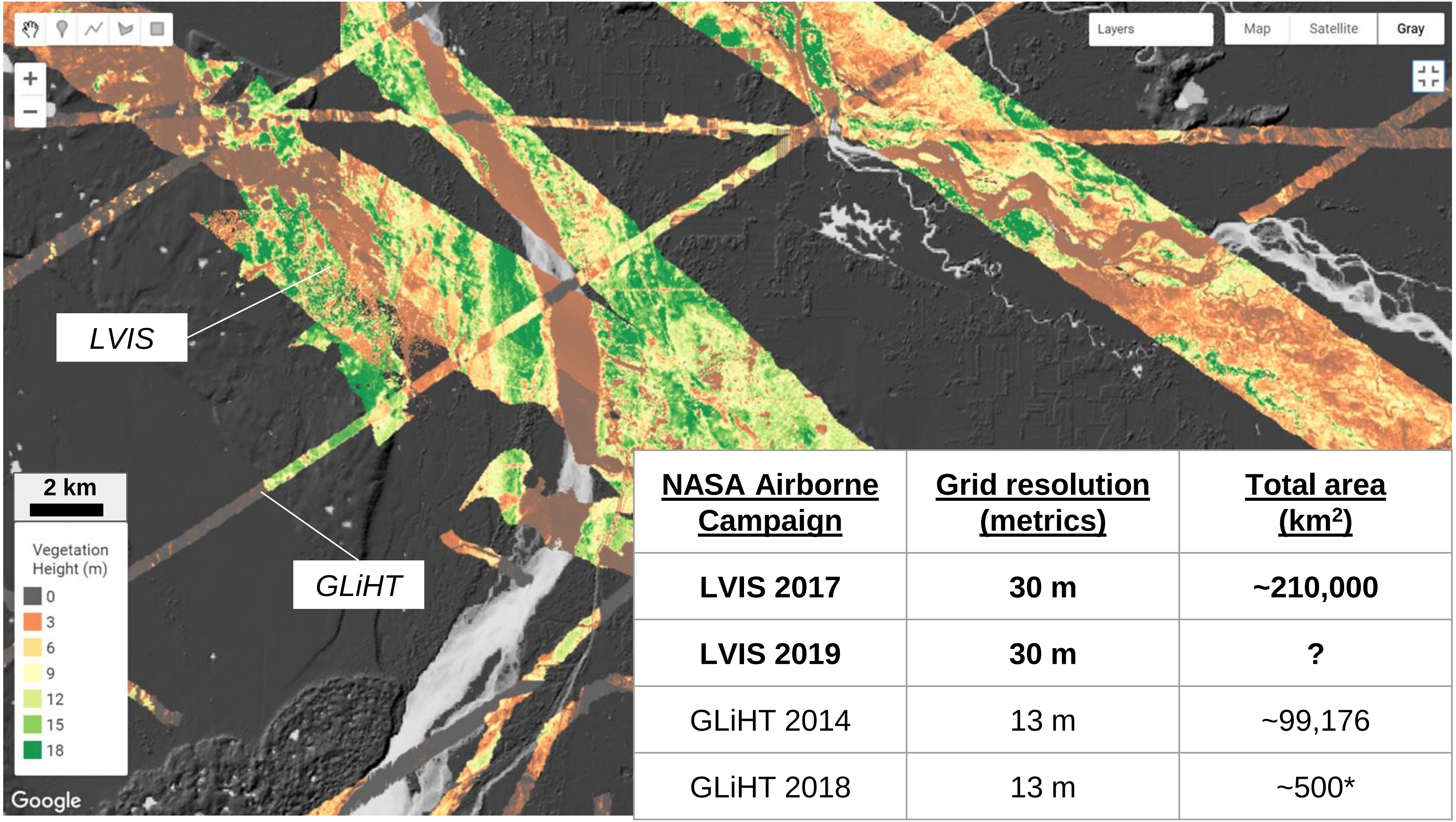


Paul Montesano | paul.m.montesano@nasa.gov
Matt Macander | mmacander@abrinc.com

Extra slides below



2 km
Vegetation Height (m)
0
3
6
9
12
15
18
Google
Boreal forest height: LVIS vs GLiHT
Crossovers of LVIS 2017 with GLiHT 2014 & 2018
r² = 0.77
rmse = 2.6
bias = 2.2
slope = 1.164
LVIS rh98 (m)
GLiHT CHM (m)
0
5
10
15
20
25
30
obs.
2500
1000
100
10
1
Map data ©2020
2 km
Terms of Use



NASA Airborne Campaign	Grid resolution (metrics)	Total area (km^2)
LVIS 2017	**30 m**	**~210,000**
LVIS 2019	**30 m**	**?**
GLiHT 2014	13 m	~99,176
GLiHT 2018	13 m	~500*

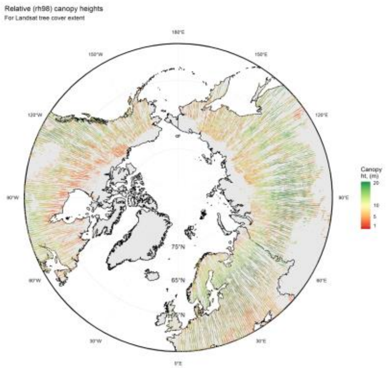
LVIS-Facility and LVIS-Classic on the G-V

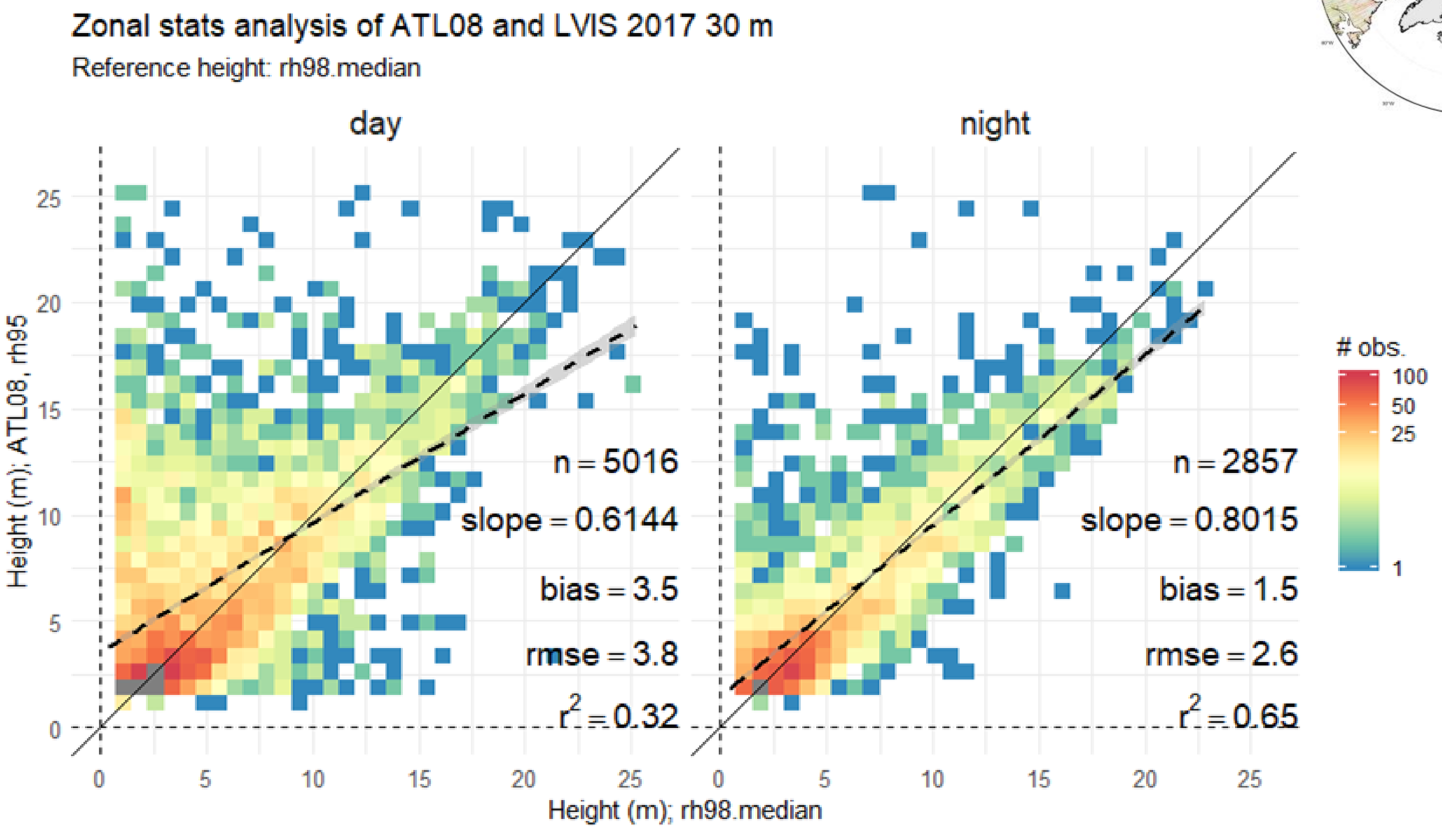
- ABoVE 2019: fly both LVIS-Facility and LVIS-Classic on NASA G-V
- FOV of both instruments will be positioned so they overlap on the ground
- Camera system in LVIS-Classic view port (~3sec interval)
- Coverage per hour in the G-V (once at altitude): 1440 km2 (200m/s, 2km swath). 38,000' flight altitude, 8-10 hour flights.
  - Coverage per hour once at altitude: 1440km^2 (200m/s, 2km swath)
- Note: LVIS Classic and Facility have the same build configuration, but Facility is the updated/modernized version of Classic

- **LVIS-Facility Instrument**
  - Swath width (0' – 3000'): 2 km?
  - Nominal Footprint diameter: 10 m
  - Footprint spacing: 10m along and 10m across track
  - Laser specs: 1064 nm wavelength, 5ns pulsewidth, rate: 4000 Hz.
  - Digitization of outgoing and return waveforms at 1000MHz

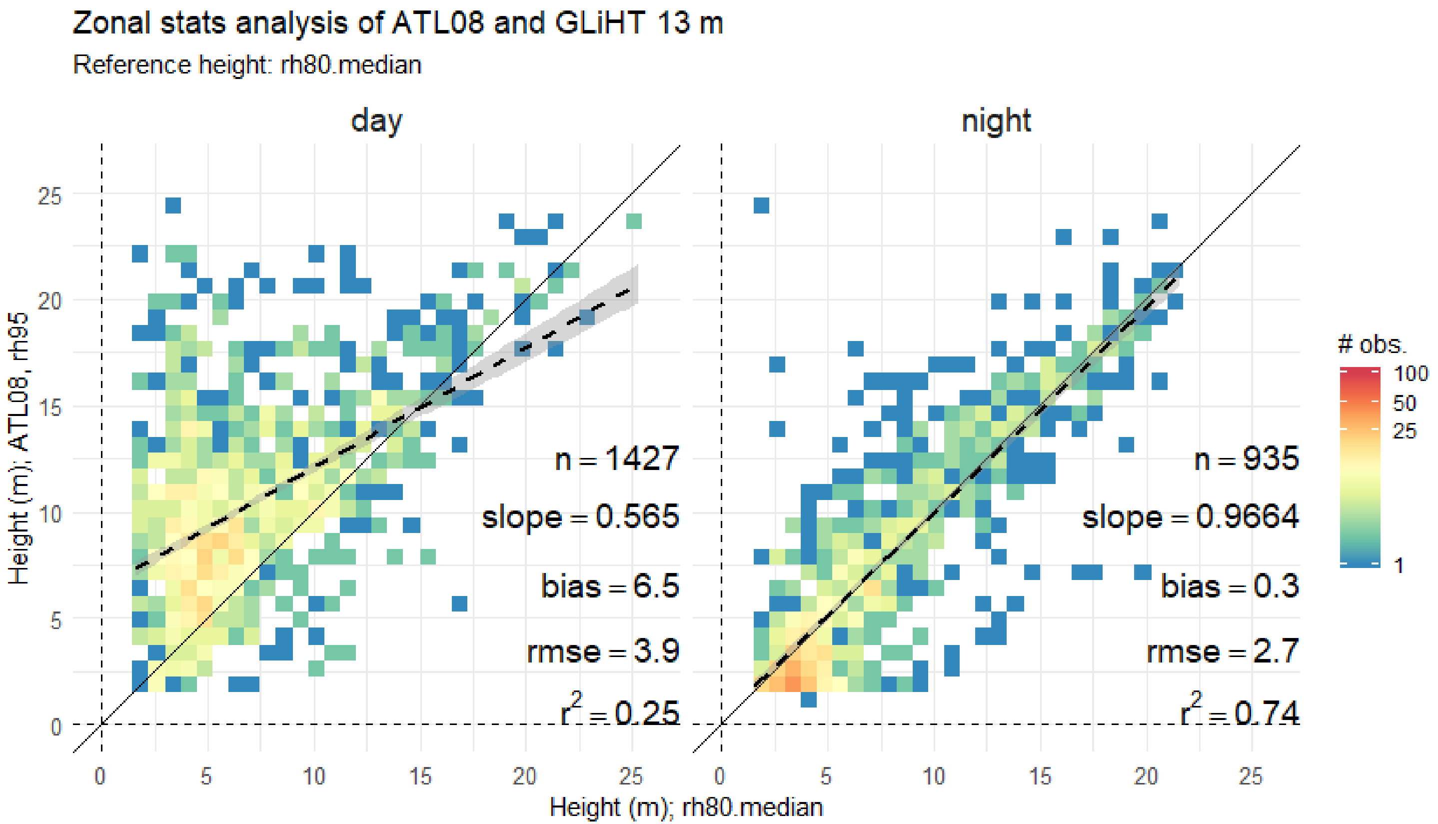
- **LVIS-Classic Instrument**
  - Swath width (0' – 3000'): 2.1 -2.3 km?
  - Nominal Footprint diameter: 20 m
  - Footprint spacing: 17m along and 17m across track
  - Laser specs: 1064 nm wavelength, 9 ns pulsewidth, rate: 1000 Hz.
  - Digitization of outgoing and return waveforms at 1000MHz

LVIS validation of height from ICESat-2 (v2)

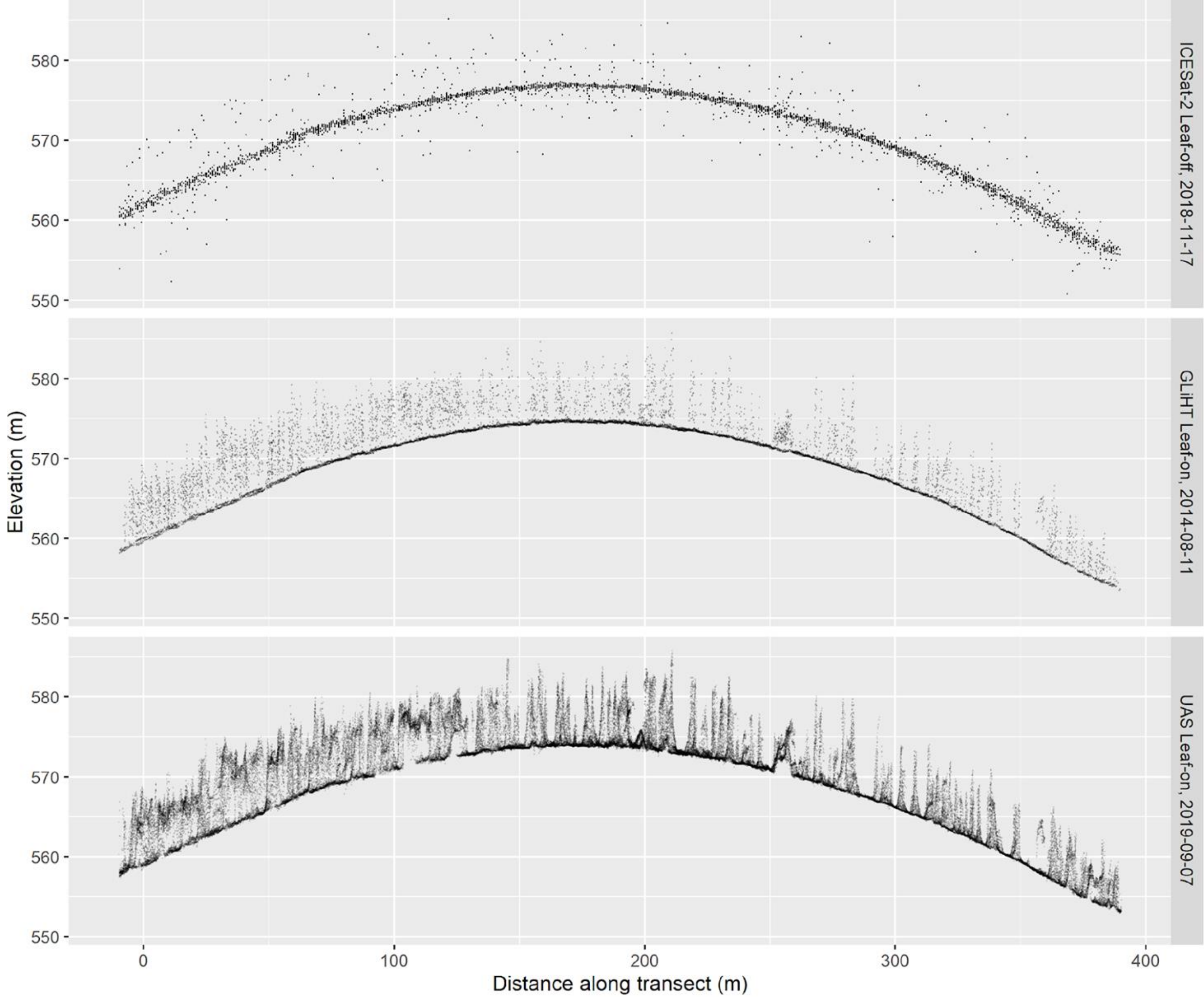




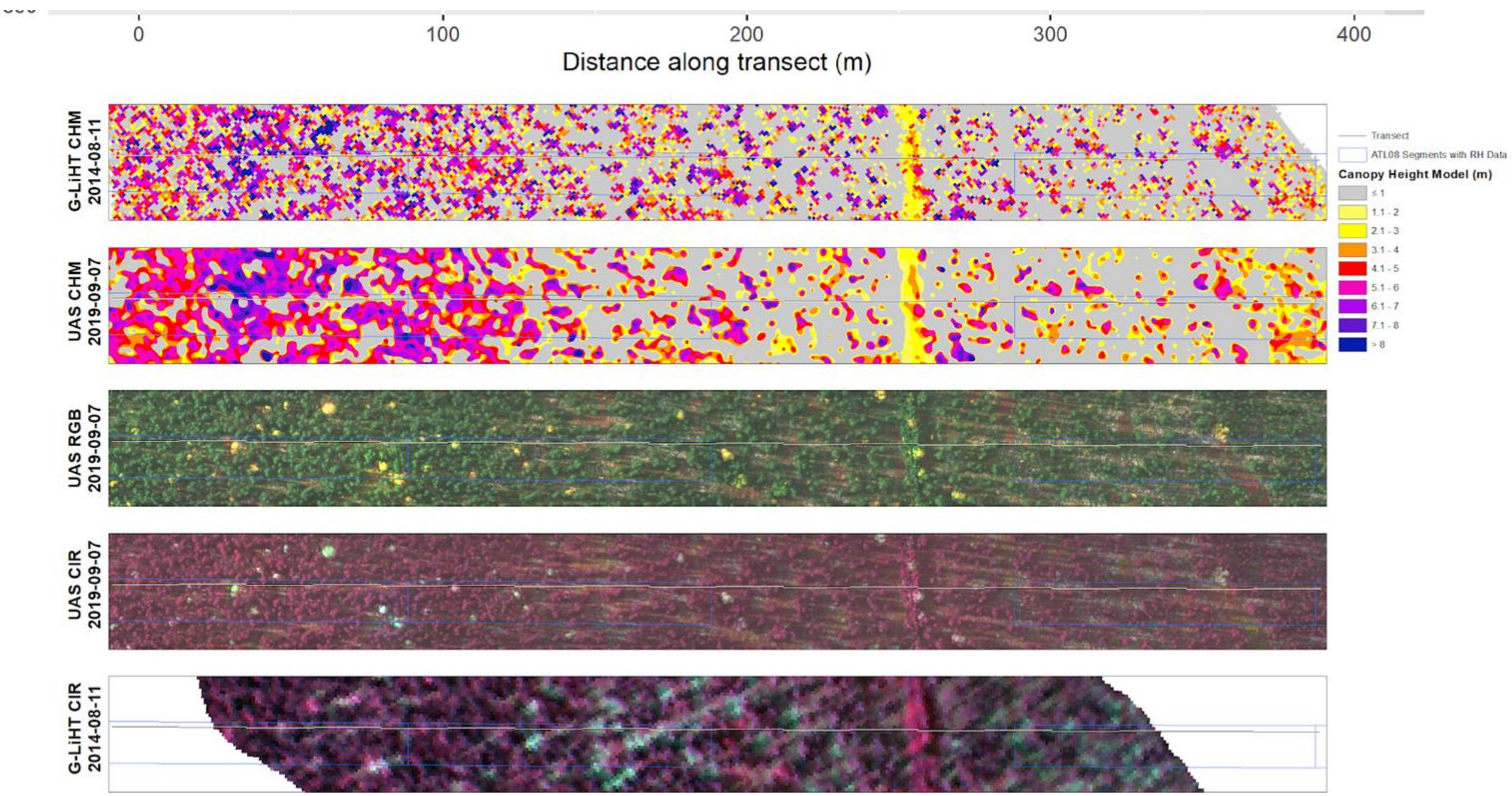
GLiHT validation of height from ICESat-2 (v2)



ICESat-2 (v2), GLiHT and UAS

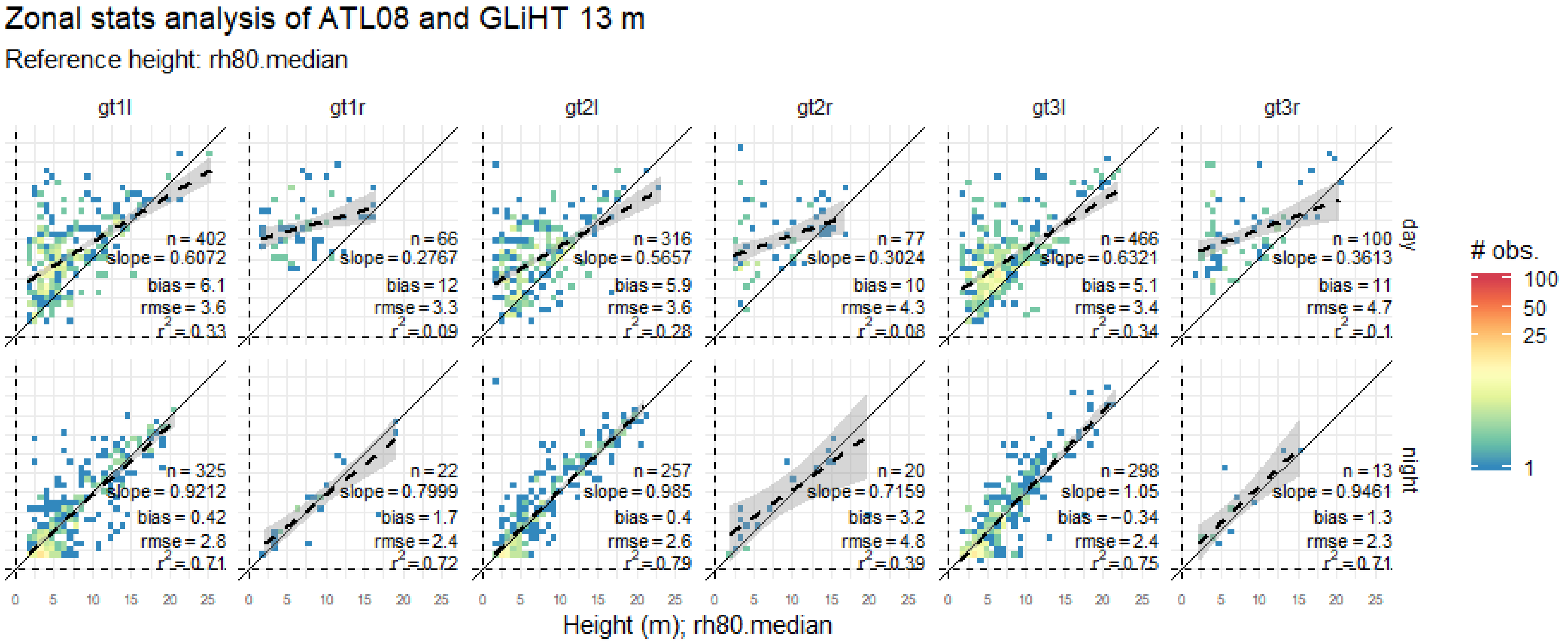


GLiHT and UAS, Canopy Structure and Composition

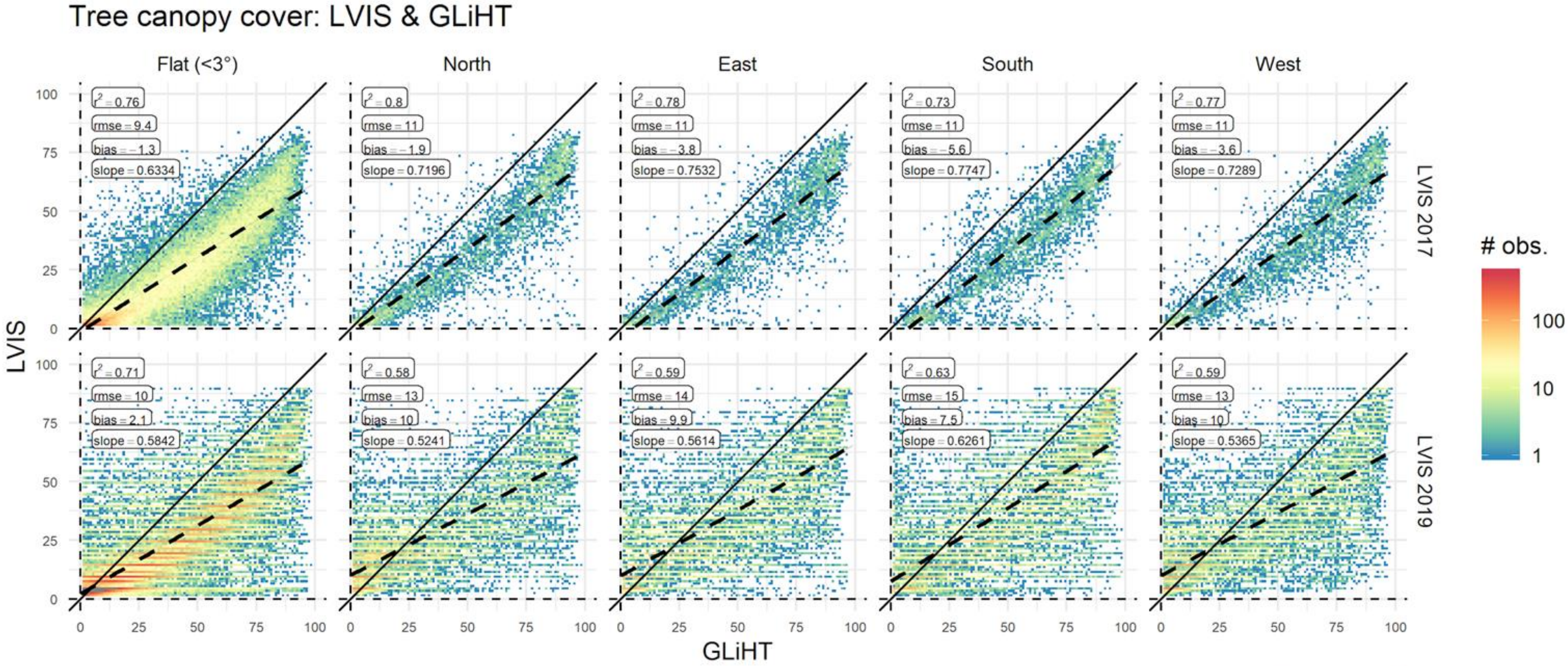


Nighttime low power may be useful for small trees

Note: in June 2020, this will be updated and re-evaluated with v3 ATL08



What is LVIS tree canopy cover?



Source: above_lidar_chk.Rmd

Deriving boreal forest growth rates

LVIS gridded height estimates linked to Landsat forest age are used to evaluate forest growth variations across the boreal.

PI Chris Neigh

Co-Is: Paul Montesano, Ben Poulter, Joe Sexton

Synthesis: Abhishek Chatterjee