National Aeronautics and Space Administration

Optimal Strategies for Storing Earth Science Datasets in the Commercial Cloud

Goddard
SPACE FLIGHT CENTER

Dieu My Nguyen, University of Colorado Boulder
Johana Chazaro Cortes, California Baptist University
Marina Dunn, University of California Riverside

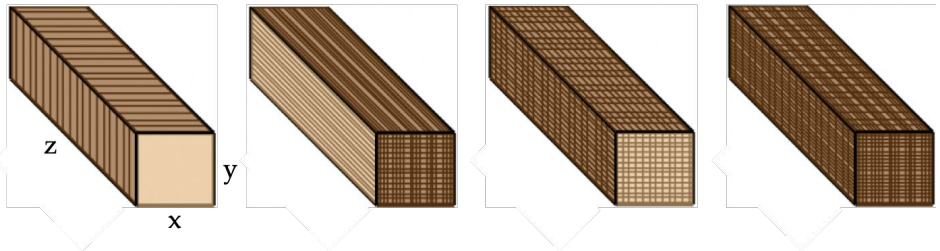
Mentor: Alexey N. Shiklomanov, PhD, GSFC-618

Problem & goal

- Part of EIS Fire Portal project
- How does the chunking scheme affect usage and analysis of multi-dimensional datasets in Zarr format?
- What are the optimal chunking strategies for storing datasets on the cloud?

--- Test performance of different chunking strategies ---

Data cube



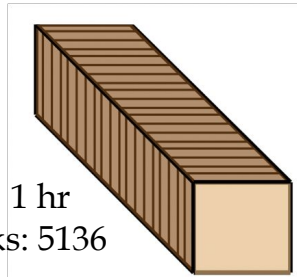
Dataset & chunking strategies

GEOS-FP
dataset in Zarr format

From GEOS - high resolution global atmospheric model

Analyses and forecasts produced in real time

Default chunking scheme

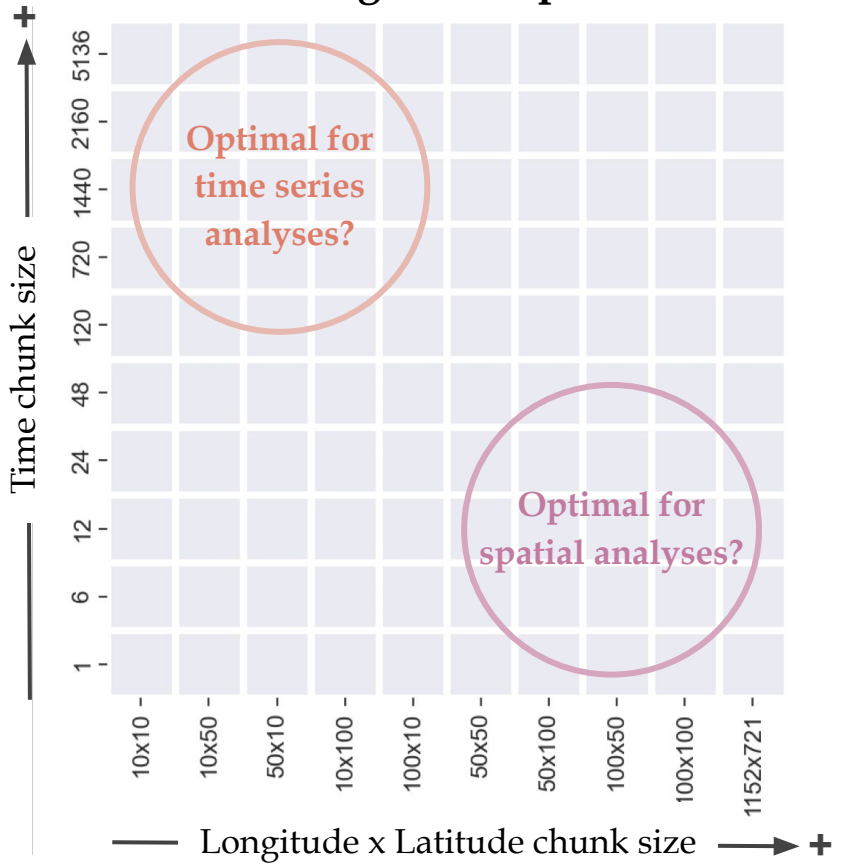


Time
Chunk size: 1 hr
Num chunks: 5136

Latitude
Chunk size: 721 deg
Num chunks: 1

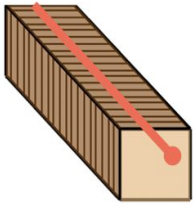
Longitude
Chunk size: 1152 deg
Num chunks: 1

Strategies to explore



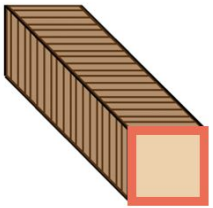
Trade-off between time series and map in CPU time

Task 1. Drawing time series at single location

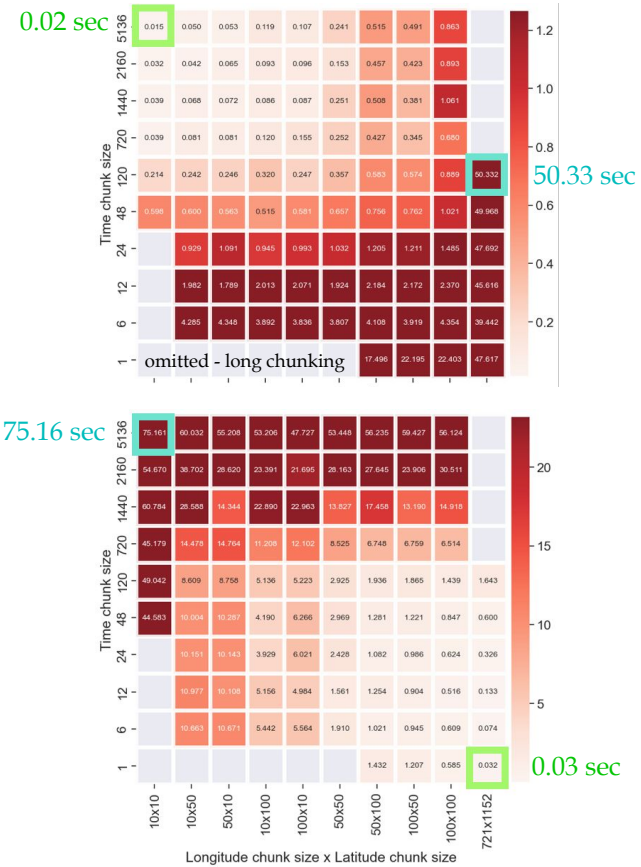


longitude: -122.19
latitude: 47.61

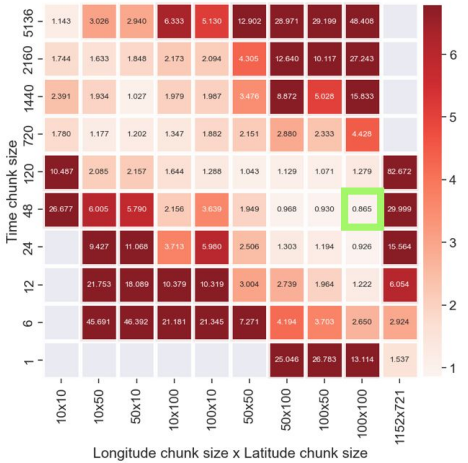
Task 2. Drawing map at 1 time step



time: 2020-06-01
T00:00:00

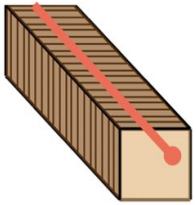


Product of task 1 and task 2



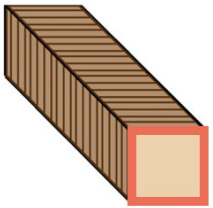
Trade-off between time series and map in peak memory

Task 1. Drawing time series at single location

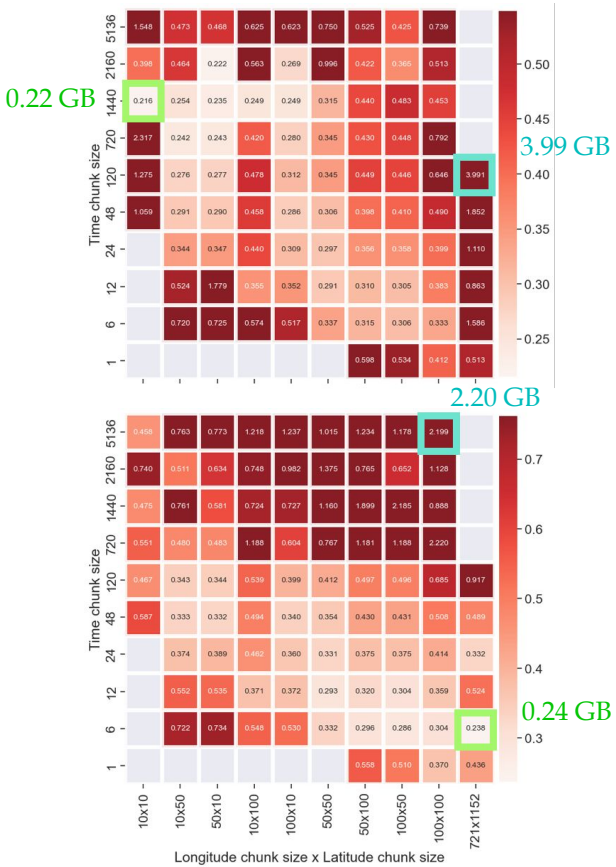


longitude: -122.19
latitude: 47.61

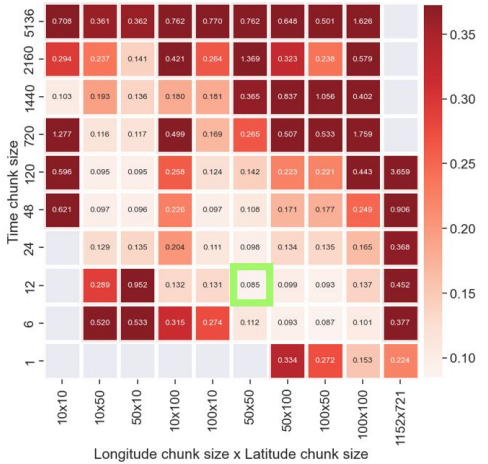
Task 2. Drawing map at 1 time step



time: 2020-06-01
T00:00:00

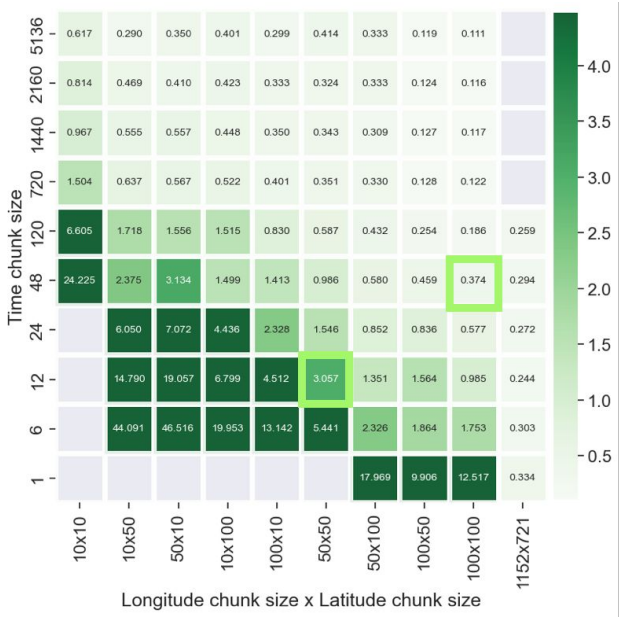


Product of task 1 and task 2

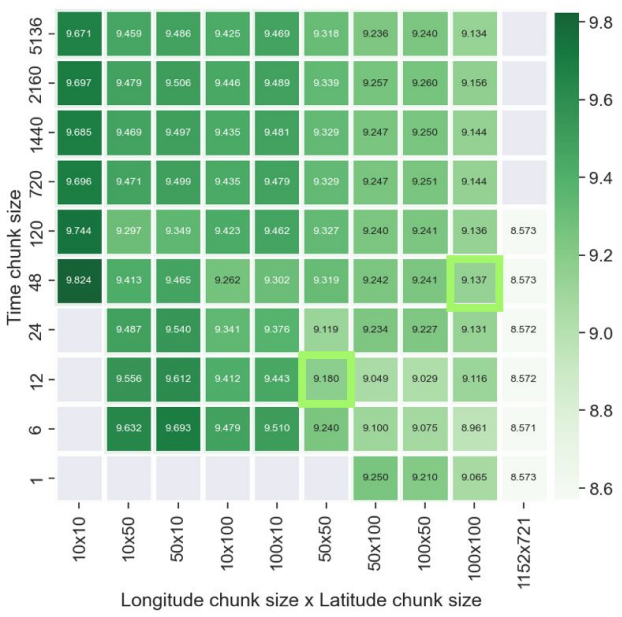


Strategy also affects rechunking time & storage size

Rechunking time (hr)



Storage size (GB)



Some compression effect

Conclusion & Future Directions

- Performance trade-off between optimizing for spatial vs. time-series analyses
- Identified a range of strategies that perform well (and poorly) for both analyses
- Higher-dimensional datasets - more significant trade-offs?
- Consider spatio-temporal analyses that span multiple dimensions simultaneously?
- Submitting manuscript soon!


dieumy.t.nguyen@nasa.gov

dieumynguyen.github.io

Conclusion & Future Directions

- Performance trade-off between optimizing for spatial vs. time-series analyses
- Identified a range of strategies that perform well (and poorly) for both analyses
- Higher-dimensional datasets - more significant trade-offs?
- Consider spatio-temporal analyses that span multiple dimensions simultaneously?
- Submitting manuscript soon!

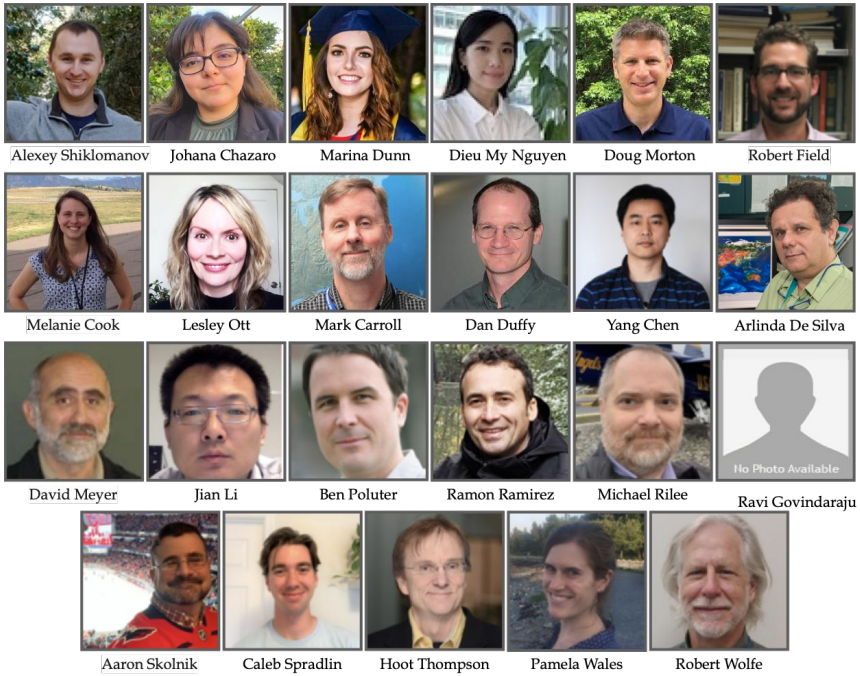
dieumy.t.nguyen@nasa.gov

dieumynguyen.github.io

Thank you

Acknowledgements

EIS-Fire Team



Jeremy Raupp (GSFC/NAVTECA), Sean Harkins, Aimee Barciauskas (DevelopmentSeed), Kata Martin, Joe Hamman, Jeremy Freeman (CarbonPlan), GES DISC