IN33C-0141

AGU 2017 Fall Meeting, New Orleans, LA

Limiting Data Friction by Reducing Data Download Using Spatiotemporally Aligned Data Organization Through STARE

Kwo-Sen Kuo[1,3]
Michael Lee Rilee[2,3]

[1]Bayesics LLC
[2]Rilee Systems Technologies LLC
[3]NASA Goddard Space Flight Center

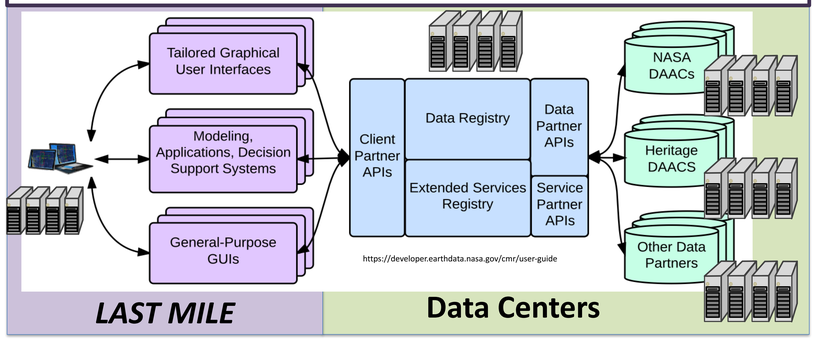
Abstract

- Existing data processing practice limits the volume and variety of relevant geoscience data that can practically be applied to important problems, reducing productivity and limiting quality.
- The SpatioTemporal Adaptive Resolution Encoding (STARE) implements an innovative encoding of geo-spatiotemporal information, originally developed for aligning datasets with diverse spatiotemporal characteristics in an array database.
- The additional encoding of spatial and temporal resolution information in STARE enables comparisons and conditional selections across diverse datasets.
- Spatiotemporal set-operations, e.g. union and intersection, are mapped to efficient integer operations with STARE, and spatial operations are sped by the use of geodesic edges.
- Applied to existing data models (point, grid, spacecraft swath) and corresponding granules, STARE indexes provide a streamlined description usable as geo-spatiotemporal metadata.
- Coupled with large scale, distributed hardware and software, STARE-based data access reduces pre-analysis data preparation costs by offering a convenient means to align different datasets spatiotemporally without specialized effort in parallel computing or distributed data management.

Existing Practice

- File-centric data processing

- Earth Science often requires analyzing multiple, diverse datasets together.
- Existing systems are only scalable for storing, searching, and distributing vast volumes and varieties of data.
  - Researchers as end-users search, order, and download data to their local systems.
- Effort and time spent marshaling, downloading, managing, and locally combining data means **increased hardware/software costs and less time for research**.
  - Researchers must pay for the expensive "*last mile*" to make the data useful and obtain scientific results.
  - Automating such data management and eliminating the need for such end-user data preparation increases resources available for research.



Adapting STARE to Existing Practice

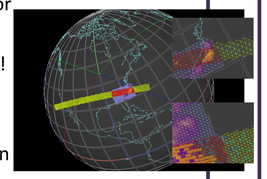
- File-centric practice is not ideal, but STARE still provides value

- **STARE is best leveraged in array database (see summary for other presentations)**
  - A *join* operation will "magically" line up different dataset arrays spatiotemporally for integrative analyses, drastically reducing data preparation effort, boosting analysis productivity
- **STARE can be adapted to existing practice. Two possibilities:**
  - Lookup tables (LUTs) in a STARE companion file for each granule.
    - Forward LUT-granule data array indices (with time) to STARE indices and
    - Reverse LUT-STARE indices to granule data array indices.
  - Use multilevel STARE indices, and hence reduced data volume, as spatiotemporal metadata for each granule.
    - Allows more **flexible and accurate** filtering/subsetting at data discovery phase, e.g. spatiotemporal intersections of dataset(s) with complex geographical regions.
- **A STARE API for calculations and enhancing their usability is in development.**

Use of "STARE Companions"

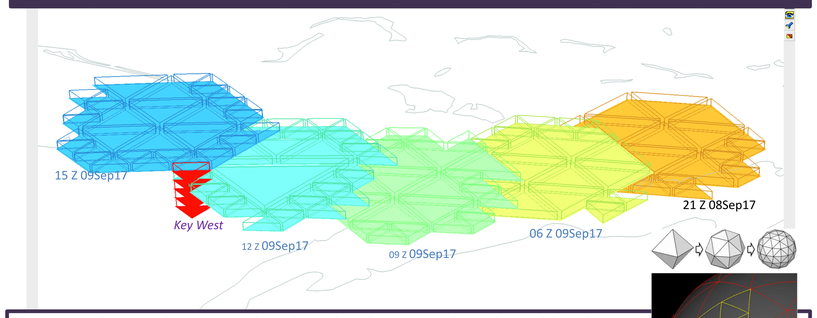
- Environmental conditions of precipitating areas

- **Objective**
  - Where and when the 1st dataset (e.g. a TRMM or GPM dataset) indicates precipitation, find environment conditions, e.g. pressure, temperature, humidity, etc., from a 2nd dataset (e.g. MERRA/-2).
- **Procedure**
  - Find in array(s) of the 1st dataset where/when there is precipitation.
  - Use the Forward LUT in the STARE companions of the 1st dataset to convert its array indices (plus time) to STARE indices.
  - Search in the STARE companions of the 2nd dataset granules for intersections using STARE indices.
  - A much more efficient operation then using lat-lon (and time)!
  - Use the Reverse LUT of the 2nd dataset to convert its STARE indices to its array indices.
  - Voilà! You get the environment conditions for where and when there is precipitation!



Using Multi-Level Spherical Triangles

- All edges shown are segments of great circles, aiding efficiency



- **STARE for spatiotemporal regions for data search and combination**
  - Example above: Hurricane Irma forecast (vertical = time)
  - Multi-resolution shows flexibility of the scheme
  - Geodesic edges speed geometric calculations
  - Very general regions and temporal structure can be supported
  - Useful for metadata and for general spatiotemporal specification
  - Memory and compute efficient
  - Naturally supports efficient data placement and parallel computing
  - Supports processing closer to where data are stored

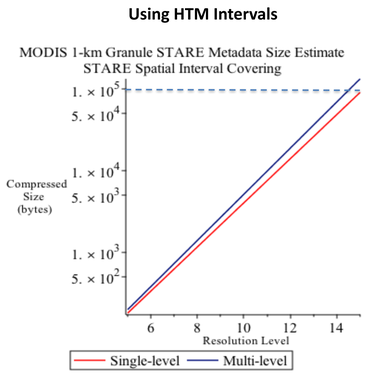
Related Development

- Regridding Infrastructure for combining and comparing diverse datasets

- STARE is merely a step towards realizing convenient and efficient *integrative analysis*.
  - It is a universal indexing scheme for all geophysical data that allows advanced and efficient colocation and subsetting.
- STARE alone is **not** sufficient to accomplish integrative analysis.
  - Need to address the resolution differences between diverse datasets.
    - A cell/pixel/voxel of a coarser resolution dataset contains multiple cells/pixels/voxels of a finer resolution dataset.
    - One-to-many comparison is difficult to interpret!
  - Regridding datasets to a common "grid" is required.
- *NASA Open Geo-Gridding Infrastructure,* **NOGGIn**, is under 2nd year of development in association with NASA LAADS+MODAPS.
  - Similar capabilities have been implemented in an array database.

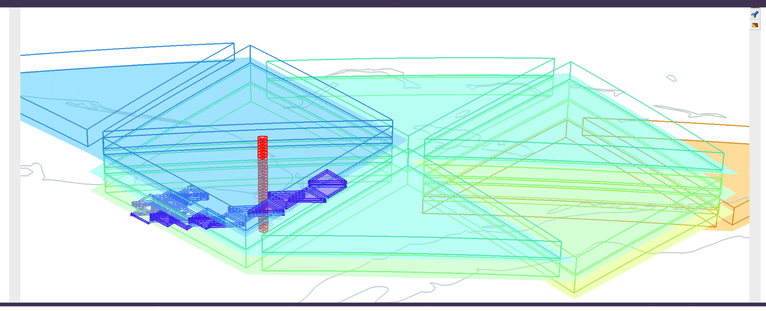
Preliminary Meta Data Volume Estimates

- e.g. MODIS Levels 1&2 Spatial Component



Moving Object DataBase

- MODB Queries – an advanced use of the STARE API



- An **event**, e.g. a hurricane episode, is a moving object, i.e. a spatiotemporal evolution.
- Combined with **event** identification and tracking capabilities, STARE supports more sophisticated MODB type queries in an array database, such as when and where Hurricane Sandy collided with an extratropical cyclone or how many blizzards crossed NY state boundary between midnight and 6 o'clock local time in the last decade.
  - Hurricane Irma forecast in level 5 triangles, Key West in red

Summary

- STARE adapted to Existing Practice

- STARE provides
  - Platform for organizing & combining diverse data
  - Automated mapping to parallel/distributed resources
  - Flexible spatiotemporal representation for data regions and moving objects
  - Efficient geometric and set operations
  - Modest metadata size
  - Adds value to existing practice where time and space require representation
  - Eases building of custom regions and datasets, streamlining file-based processes
  - Path beyond file-based processing to higher-level science query-based processing

Related Presentations:

STARE in Visualization: IN23F-07, IN33C-0141 (this work), in the SciDB array database: IN41B-0035, IN33E-04, and the path to enabling machine learning: IN11E-07.

Acknowledgement
Funding for this research is provided by NASA Earth Science Technology Office (ESTO) through the Advanced Information Systems Technology (AIST) program and by NASA Advancing Collaborative Connections for Earth System Science (ACCESS) program, for which we are very grateful.