

Investigating Access Performance of Long Time Series with Restructured Big Model Data

Suhung Shen^{1,2}, Dana M. Ostrenga^{1,3}, Bruce E. Vollmer¹, Dave Meyer¹

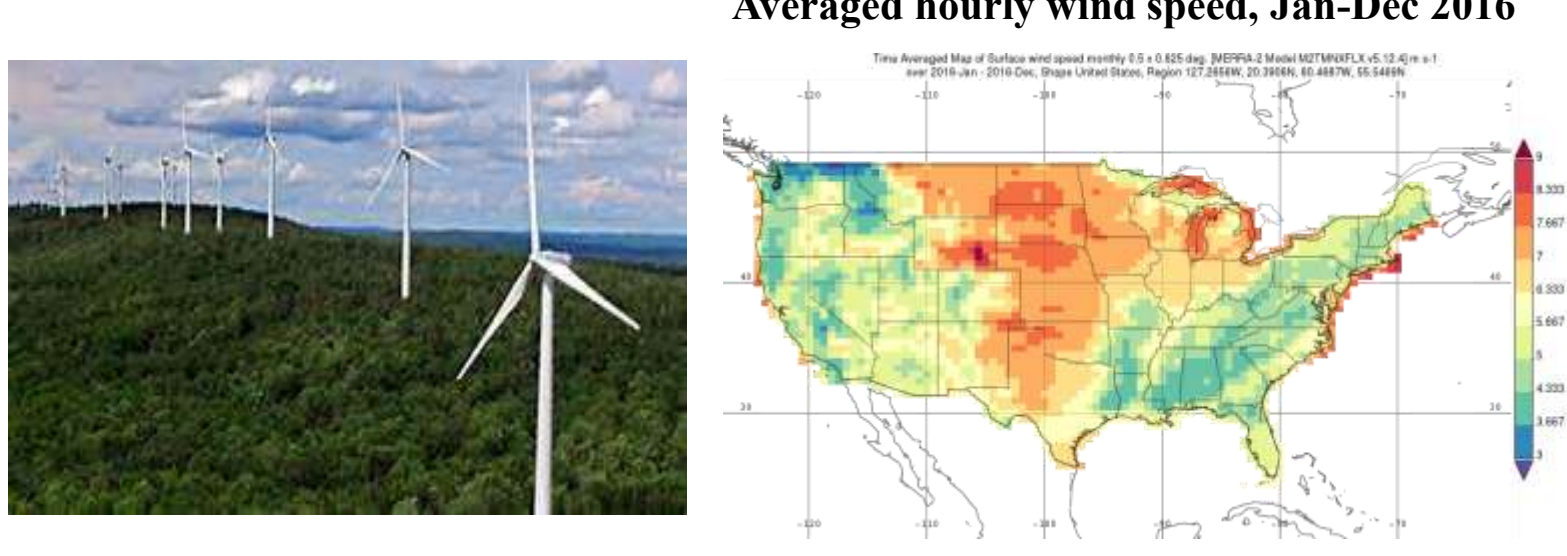
suhung.shen@nasa.gov ¹NASA Goddard Space Flight Center, ²George Mason University, ³ADNET

Abstract

Data sets generated by models are substantially increasing in volume, due to increases in spatial and temporal resolution, and the number of output variables. Many users wish to download subsetted data in preferred data formats and structures, as it is getting increasingly difficult to handle the original full-size data files. For example, application research users – such as those involved with wind or solar energy, or extreme weather events – are likely only interested in daily or hourly model data at a single point (or for a small area) for a long time period, and prefer to have the data downloaded in a single file. With native model file structures, such as hourly data from NASA Modern-Era Retrospective analysis for Research and Applications Version-2 (MERRA-2), it may take over 10 hours for the extraction of parameters-of-interest at a single point for 30 years. The NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) is exploring methods to address this particular user need. One approach is to create value-added data by reconstructing the data files. Taking MERRA-2 data as an example, we have tested converting hourly data from one-day-per-file into different data cubes, such as one-month, or one-year. Performance is compared for reading local data files and accessing data through interoperable services, such as OPeNDAP. Results show that, compared to the original file structure, the new data cubes offer much better performance for accessing long time series. We have noticed that performance is associated with the cube size and structure, the compression method, and how the data are accessed. An optimized data cube structure will not only improve data access, but also may enable better online analysis services.

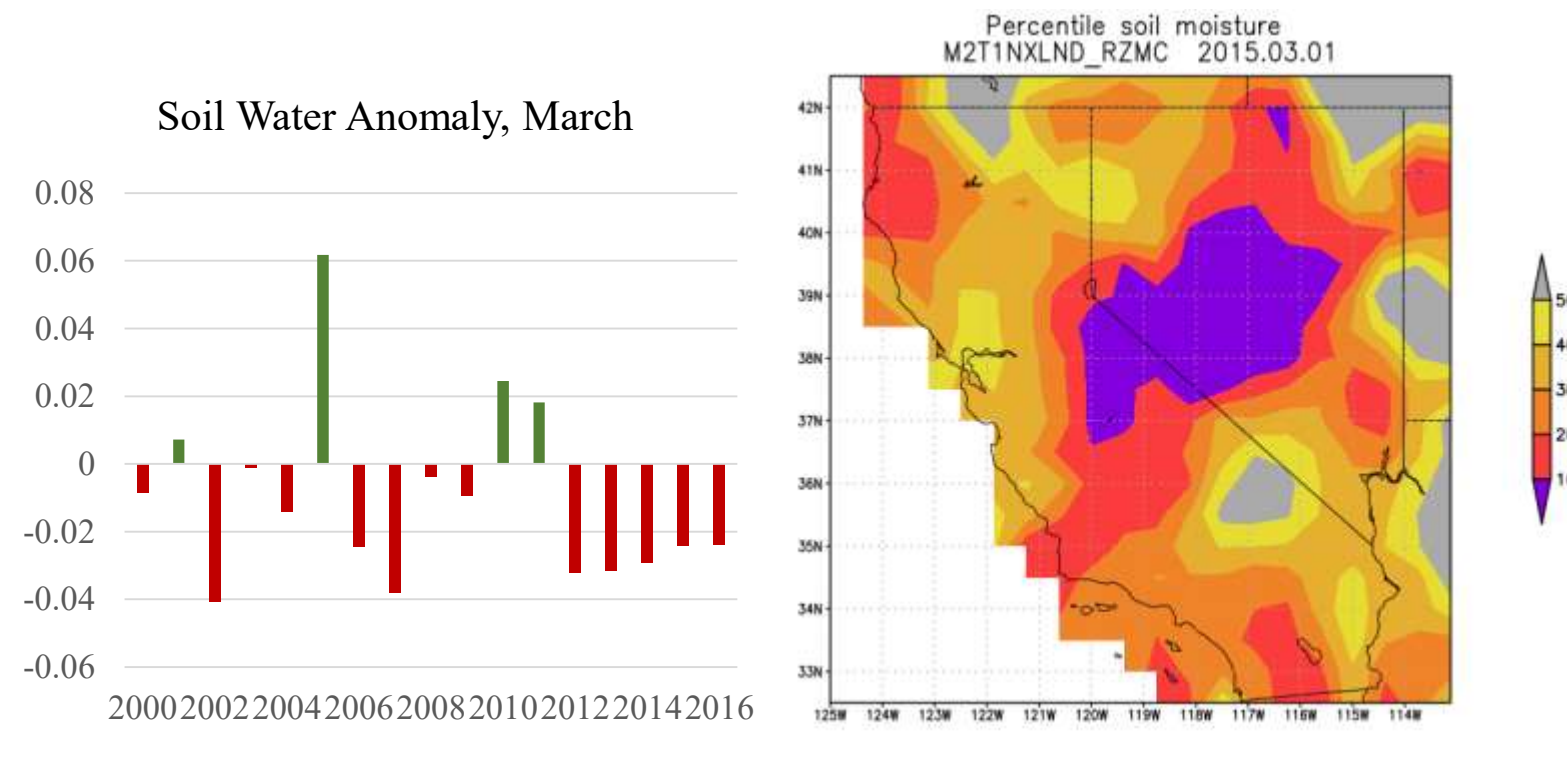
Example of MERRA-2 long time-series used in application research

Application in Wind Energy



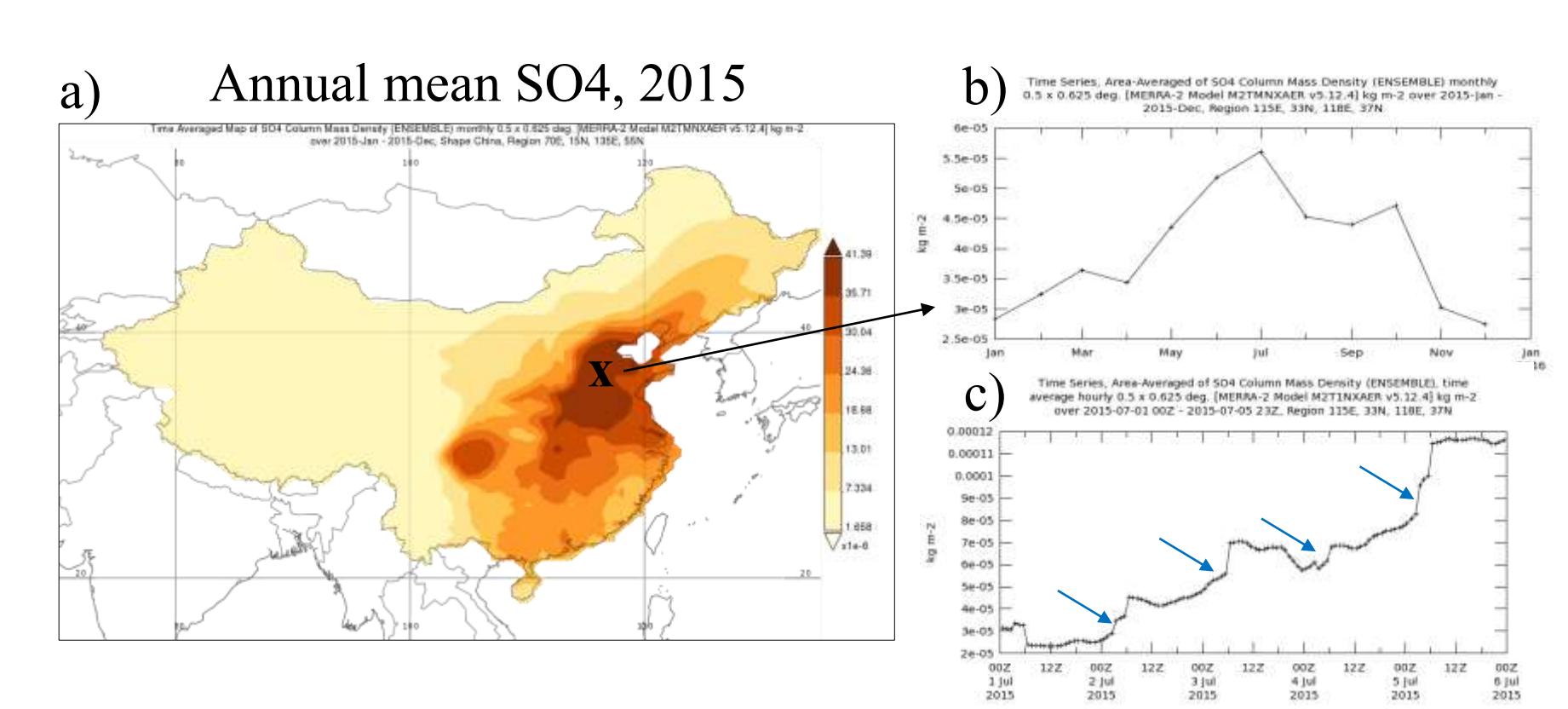
- High temporal resolution wind at 80m (the normal turbine height) for validating data and computing statistics, such as extreme, daily mean, min, max wind speed and direction information
- Wind vertical profile in the boundary layer for studying boundary layer stability which may affect power generation
- Temperature and moisture near surface, for managing the on/off state of the power-grid system

Application in Drought Events



Monthly root zone water anomaly of March from 2000 to 2016 (left), and example of percentile map of root zone water on 2015.03.01 (right), calculated by using hourly data. The climatology base period used is 1980.01.01 to 2014.12.31.

Application in Air Quality



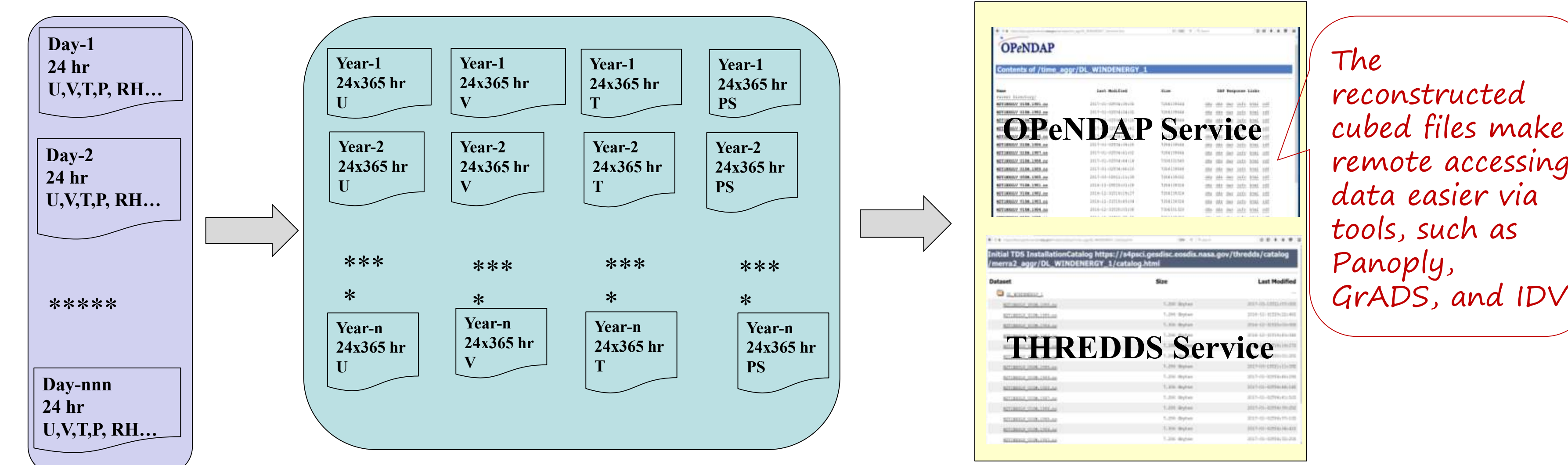
a) Year 2015 annual mean of SO4 column mass density over China; b) Area mean of 2015 monthly SO4 over East central China (33°-37°N, 115°-118°E), showing seasonal variations; c) hourly time series of SO4 for July 1-5, 2015 over the same area as in b), indicating diurnal variations.

Examples of User's Needs:

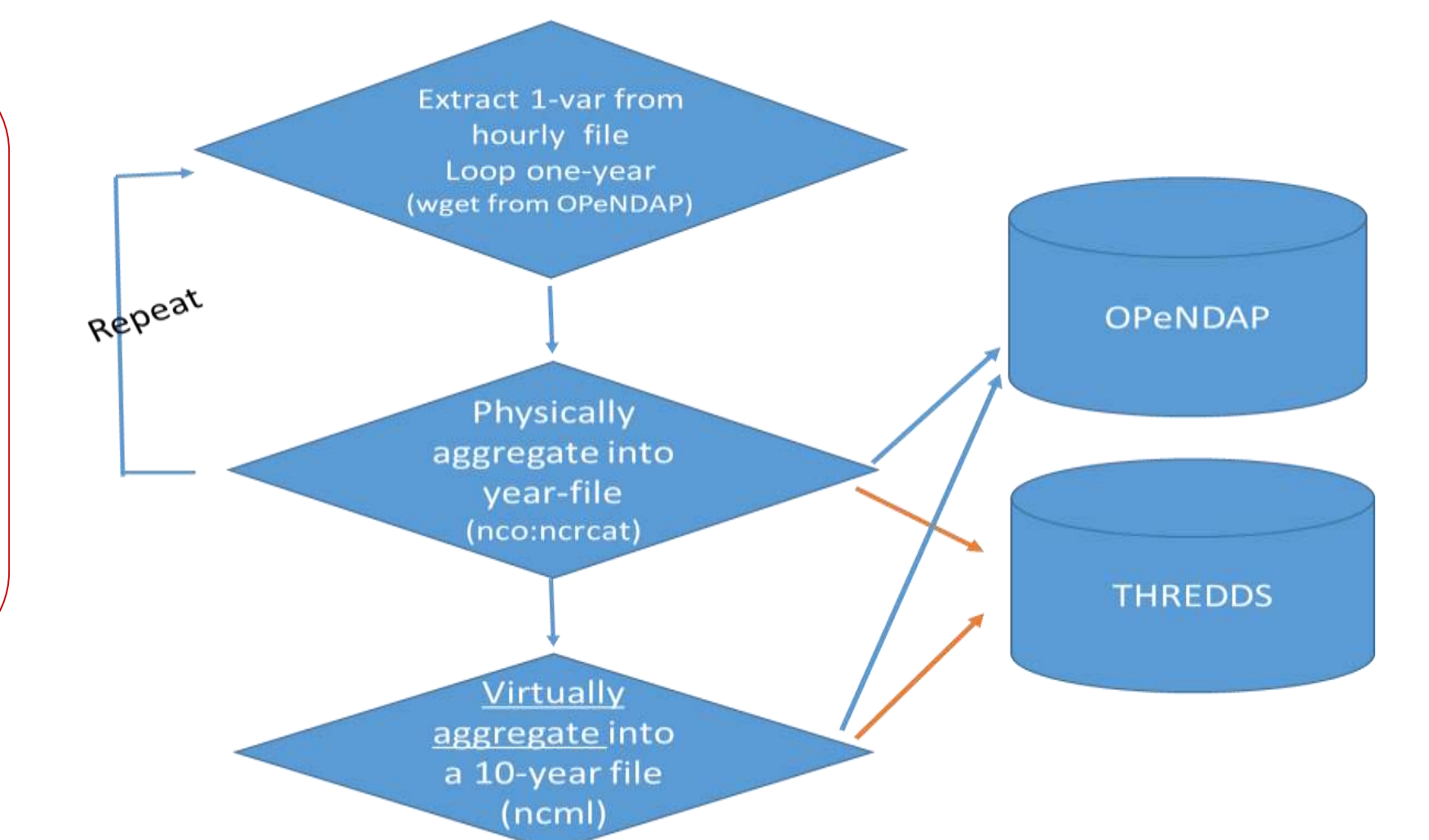
- To download a single file containing long-term **hourly time series** at one point, or a small area, in ASCII format easy and fast.
- To download time series of daily statistics, such as mean, min, max, and variation over an interested region

Performance Comparison of Downloading Time Series from Reconstructed and Current Archived Files

Day-multiple variables → Year (or month) single variable (selected)



Example of data reconstruction workflow



About MERRA-2 Reanalysis Data


<https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/docs/>

MERRA-2 is a NASA reanalysis data set for the satellite era focused on historical analyses of meteorology, atmospheric chemistry, land, ocean, and aerosols data. The data are available for a broad range of weather and climate time scales and places. They include the NASA Earth Observation System (EOS) suite of observations along with GPS-Radio Occultation datasets in a climate context.

- **Temporal Coverage:** 1980-present
- **Temporal Resolution:** Hourly, 3-Hourly, Monthly, Monthly diurnal
- **Spatial Coverage:** Global
- **Spatial Resolution:** 0.5° x 0.625° (361x576, L1, L42, L72, L73)
- **Number of Product Groups:** 95
- **Data Format:** NetCDF-4

Finding Data from NASA GES DISC

[http://disc.gsfc.nasa.gov/datasets?keywords="MERRA-2"](http://disc.gsfc.nasa.gov/datasets?keywords=)



The product landing page contains:

- **Product Summary**
- **Documentation**
 - ✓ User's guide
 - ✓ File specific
 - ✓ Key references
 - ✓ Tools
- **Data Citation:**

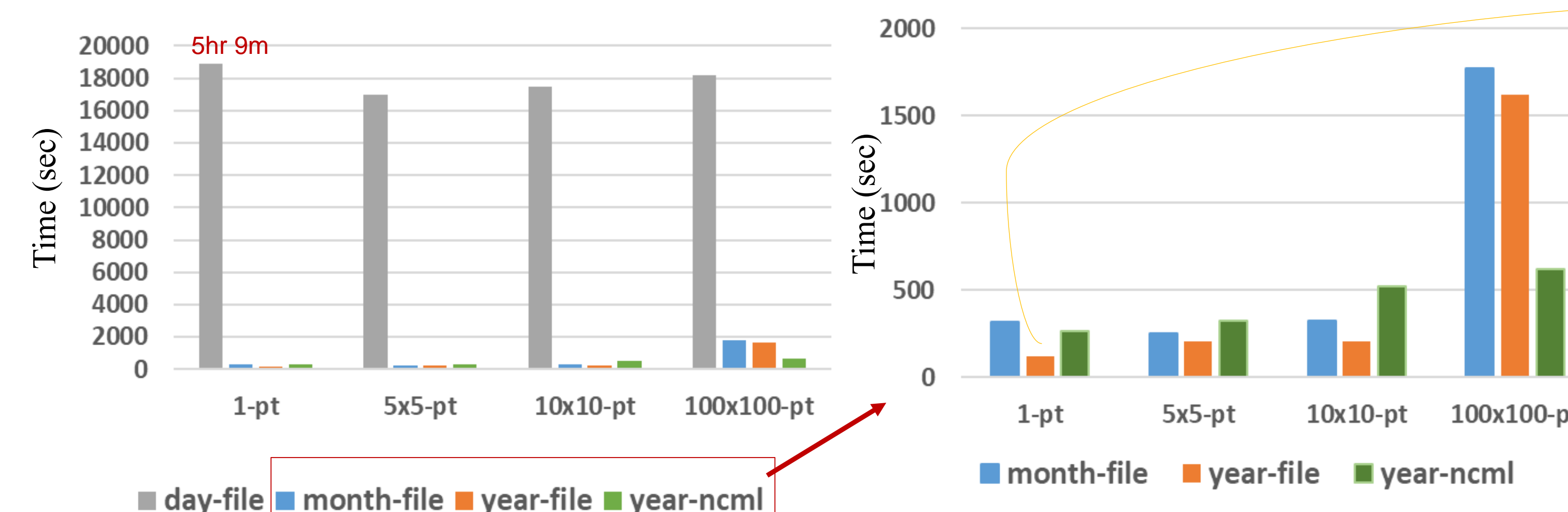
Cite the dataset in publications e.g. Global Modeling and Assimilation Office (GMAO)(2015), MERRA-2 tavgM_2d_aer_Nx: 2d, Monthly mean, Time-averaged, Single-Level, Assimilation, Aerosol Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed [Data Access Date] 10.5067/FH9A0MLJPC7N

Data Access Methods:

- ✓ MERRA-2 subsetter
- ✓ Direct download (HTTPS)
- ✓ OPeNDAP
- ✓ GDS
- ✓ Giovanni: visualization
- ✓ Data Recipes (step-by-step instructions on accessing, reading, & viewing data with various data tools)

Table: Comparison of Accessing 1-pt data

Time span	Current Archived		Reconstructed	
	OPeNDAP (original day-cube)	GDS (virtual time aggregating day-cube)	OPeNDAP (physical year-cube)	OPeNDAP (ncml aggregating year-cube)
1 month (720 time steps)	real 2m35.53s user 0m0.69s sys 0m0.415s	real 0m17.953s user 0m0.019s sys 0m0.019s	real 0m0.314s user 0m0.032s sys 0m0.013s	real 0m34.712s user 0m0.023s sys 0m0.019s
1 year (8784 time steps)	real 31m27.32s user 0m8.282s sys 0m4.822s	real 1m59.725s user 0m0.025s sys 0m0.017s	real 0m33.876s user 0m0.034s sys 0m0.022s	real 0m34.528s user 0m0.023s sys 0m0.018s
10 years (87672 time steps)	Estimated: 5hr 9m	Failed (if single access) real 25m44.115s user 0m0.244s sys 0m0.133s	real 5m12.437s user 0m0.296s sys 0m0.178s	Failed (if single access) real 4m25.336s user 0m0.123s sys 0m0.094s



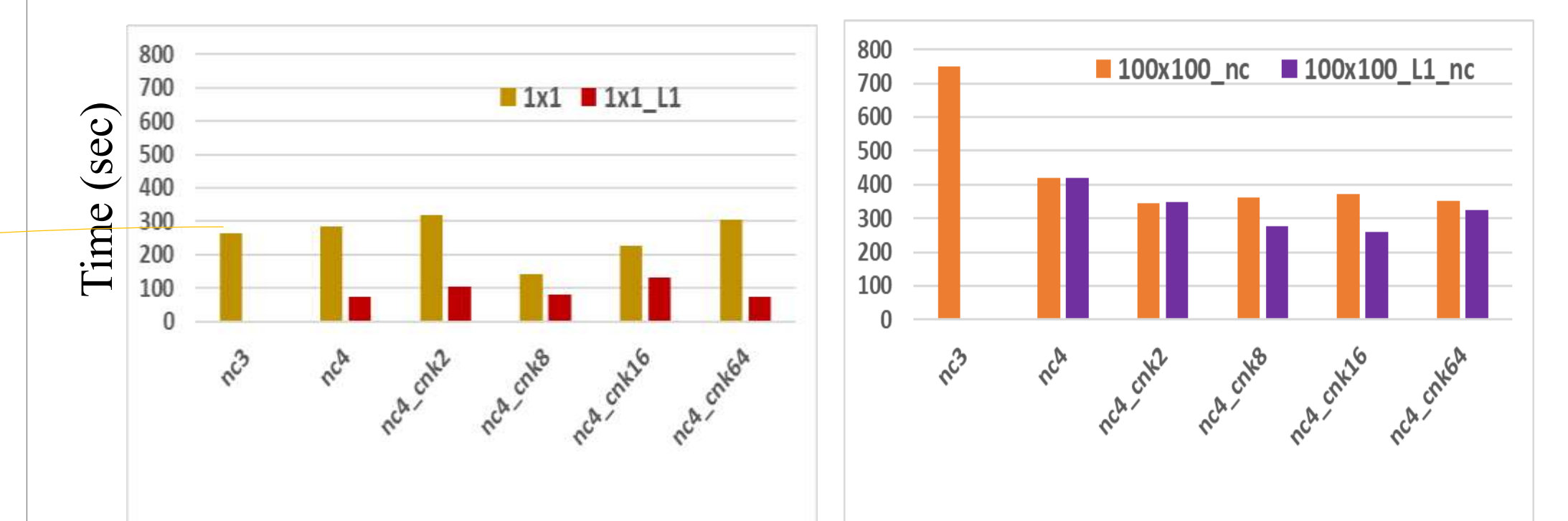
Figures below are the time (unit=sec) used to download 10-years hourly time series of wind in ASCII format of a 1x1 (single point), 5x5, 10x10, and 100x100 points from daily (day-file), monthly (month-file), yearly (year-file) cubed files, and virtual time aggregated with ncml tool (year-ncml).

Impact of Internal Chunking and Compression

Test data: hourly wind data (M2T1NXSLV_U10M)
yearly cube array size: 8760x361x576 [time x lat x lon]

L1: Compression → -L 1

Cnk2: --cnk_dmn XDim,2 --cnk_dmn YDim,2 → tile_size 8760x180x288
Cnk8: --cnk_dmn XDim,8 --cnk_dmn YDim,8 → tile_size 8760x45x72
Cnk16: --cnk_dmn XDim,16 --cnk_dmn YDim,16 → tile_size 8760x22x36
Cnk64: --cnk_dmn XDim,64 --cnk_dmn YDim,64 → tile_size 8760x6x9



Performance tests by varying chunking size for uncompressed and compressed data in cases downloading 10-years single point data in ASCII (left), and 10-years 100x100 sized region in NetCDF format (right). Results suggest that it is possible to download 30 years of time series in ~7 min for a single point and ~20 min for 100x100 array from the global yearly cubed data served in OPeNDAP.