Accessibility Considerations for Very Large Datasets

Puneet Kishor
University of Wisconsin-Madison and Creative Commons

Acknowledgments to

CODATA for inviting me, Creative Commons for funding my trip, University of Wisconsin-Madison for paying my salary, and most importantly, the US Federal Goverment for making all the data available to anyone, anywhere without any pre-conditions

Research context: ecosystem process modeling of very large terrestrial ecosystems

Information by numbers

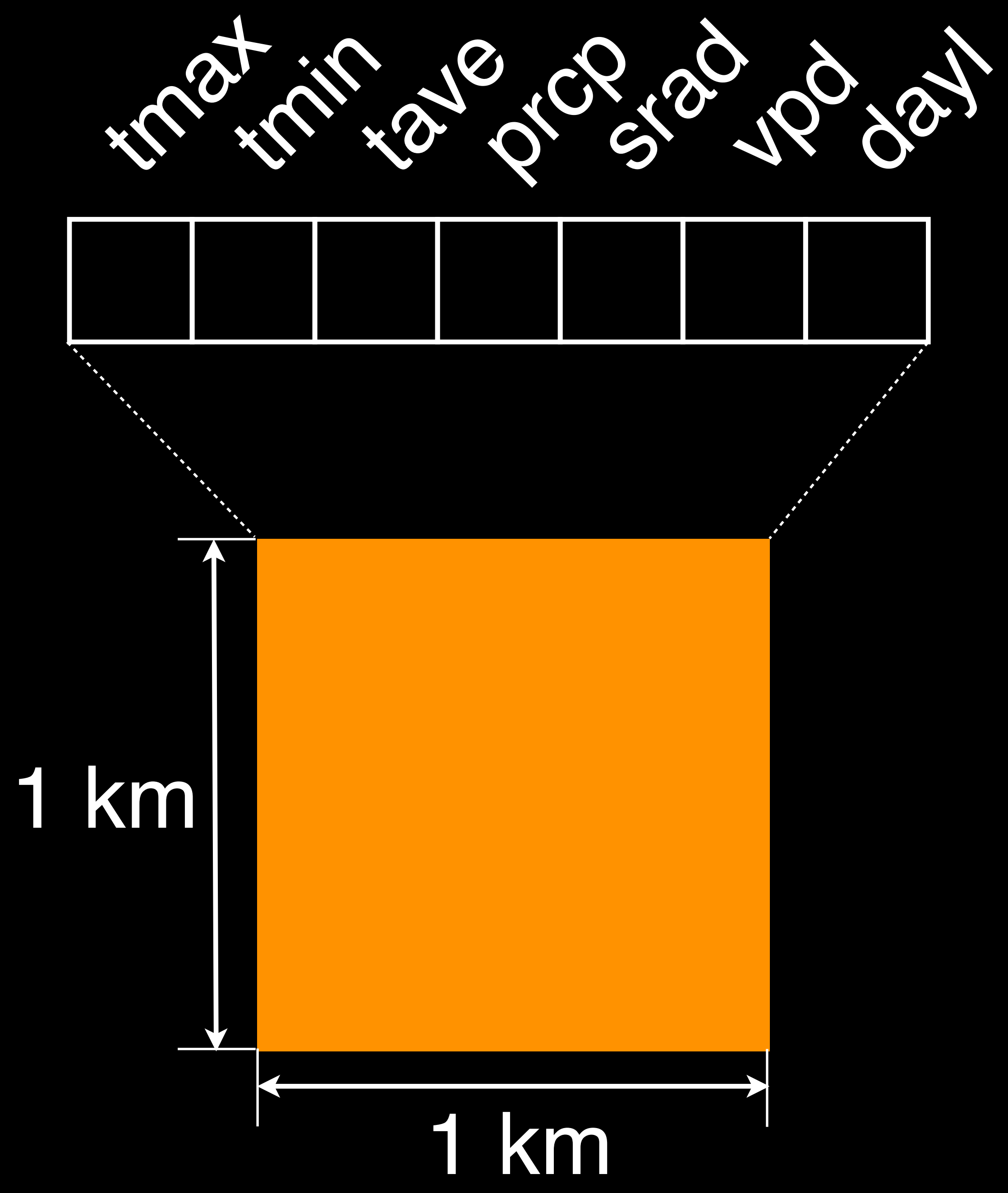
7

tmax tmin tave prcp srad vpd dayl

daily variables

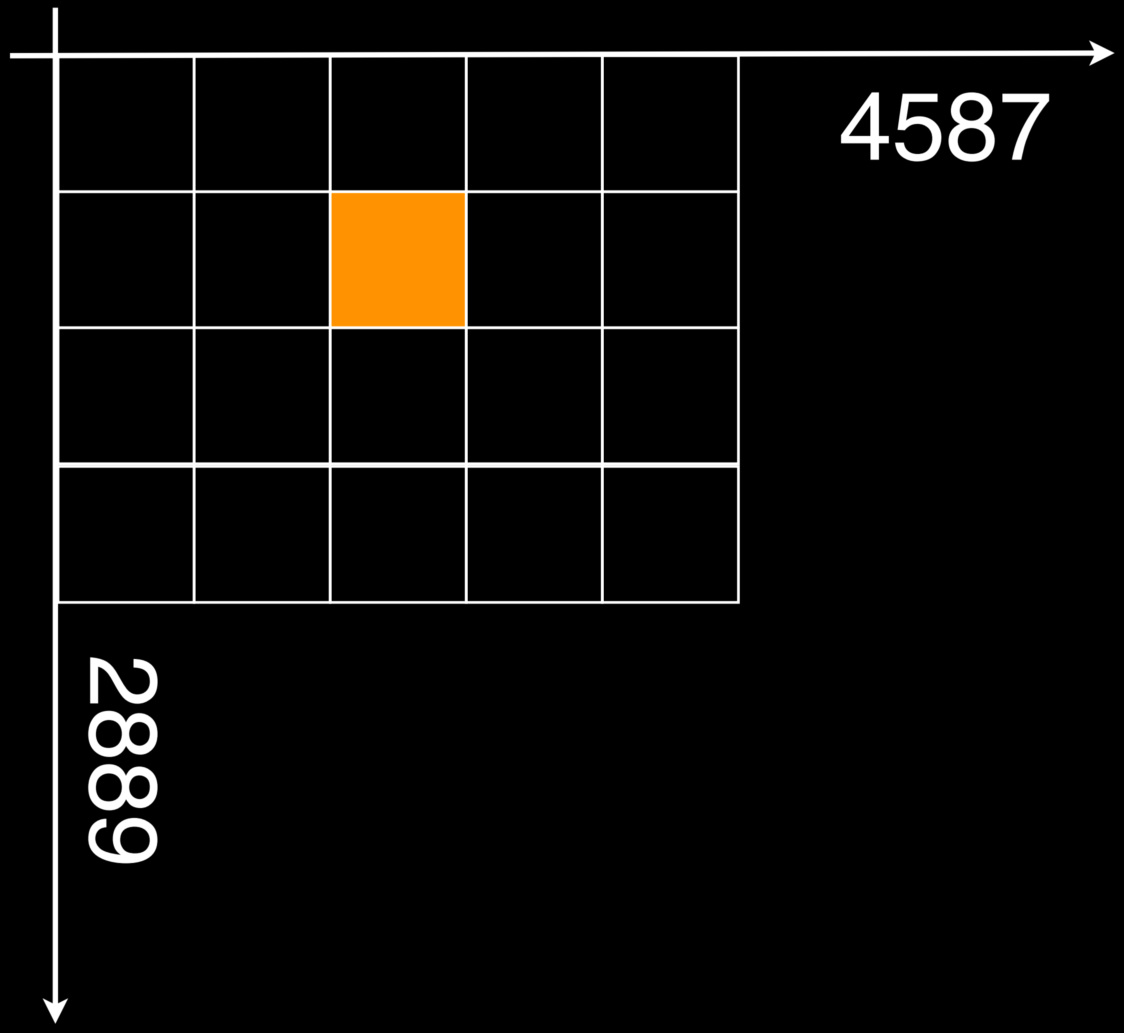
1

km^2 cell

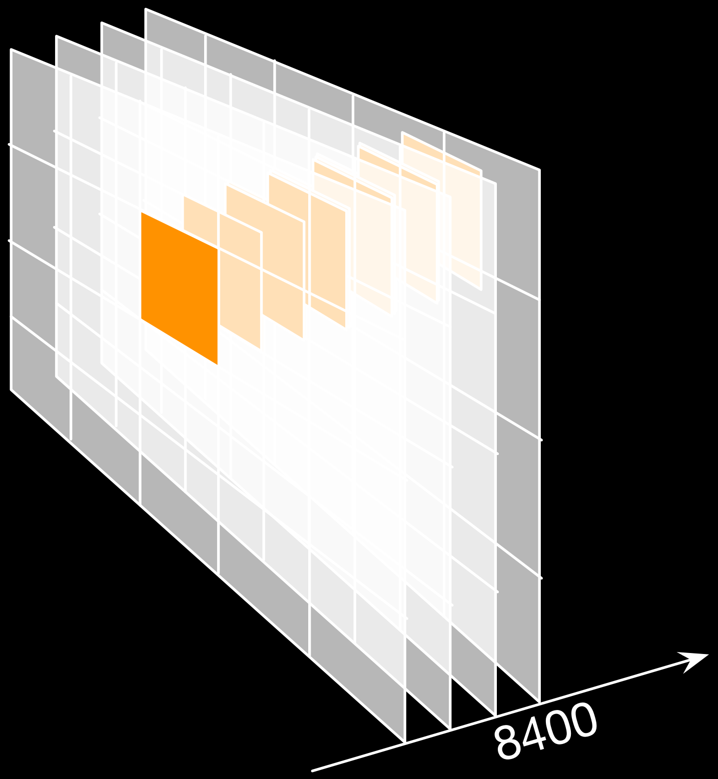


13.25

million cells



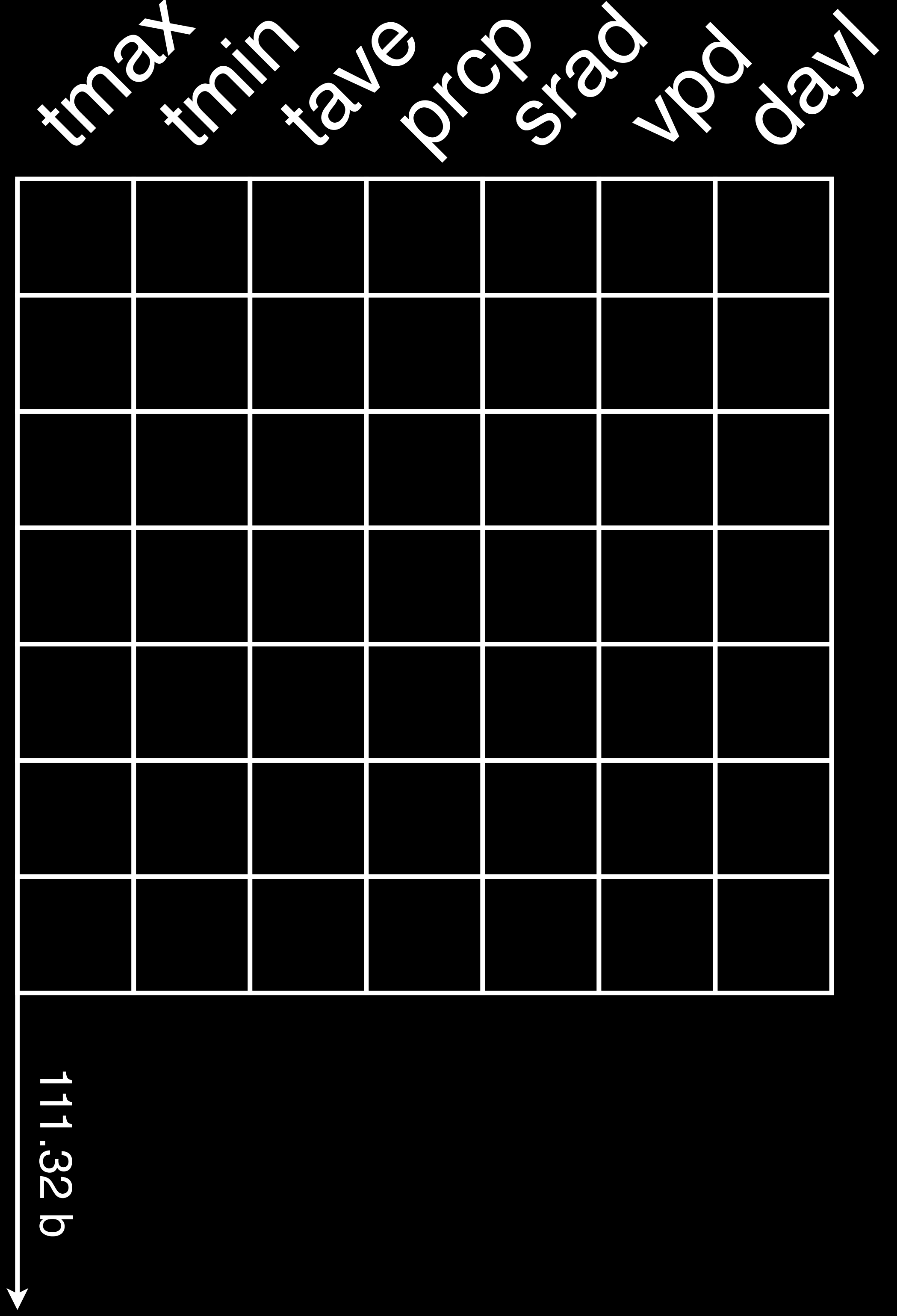
8401



days

$111^{.32}$

billion septets



725$^{.78}$

raw gigabytes

10

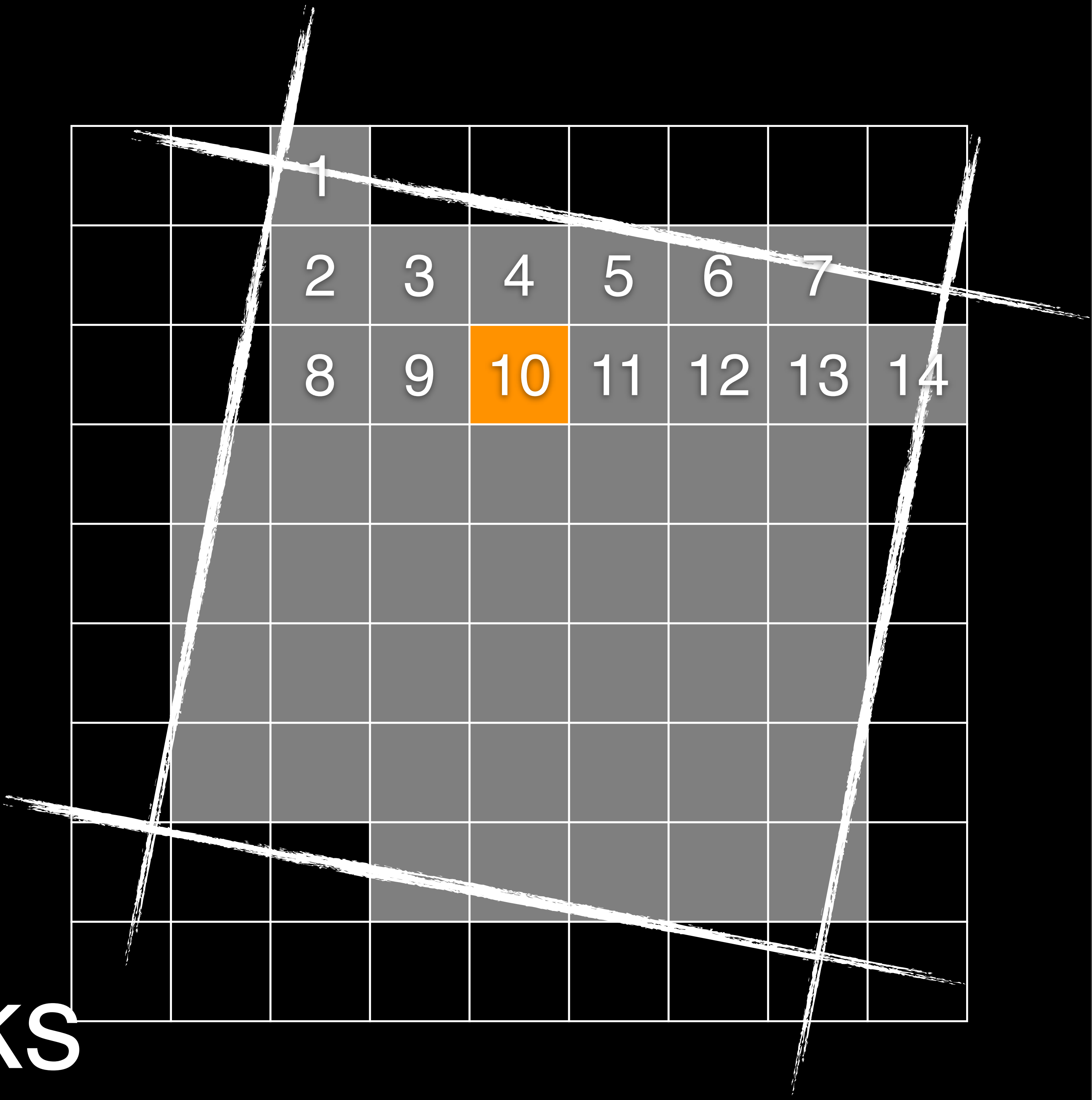
times as much in a database

84

GB of NetCDF format
in tar gzipped archives

4°

square chunks

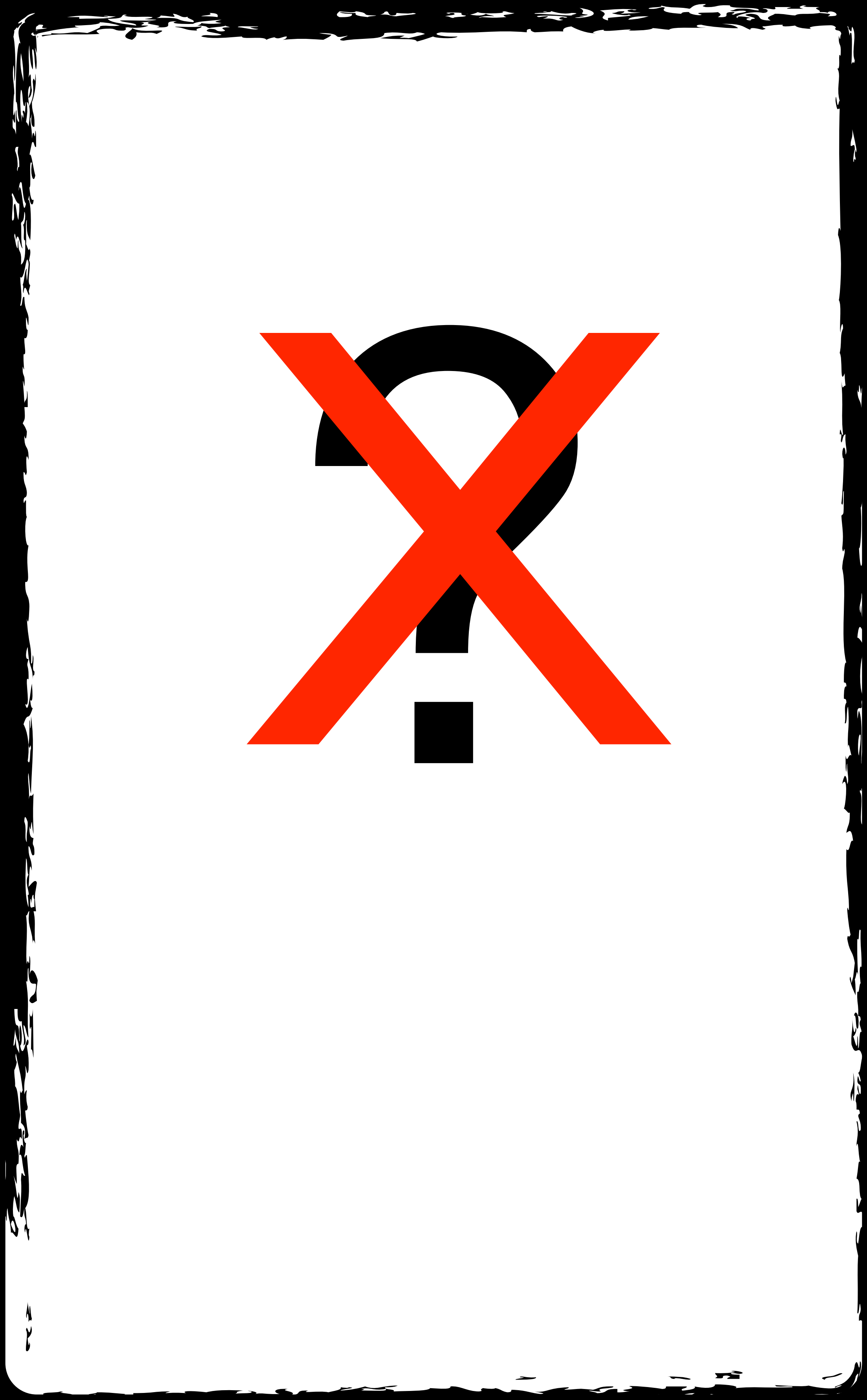


“½”

incomplete
documentation

0

ways to query the data



“10”

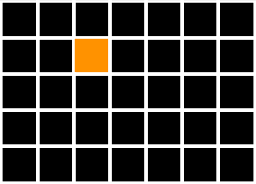
times the work to unpack the data

1. Acquire NetCDF file of lat/lon values for each cell from the weather data 1 km^2 estimates
2. Dump lat/lon values to CSV with Panoply
3. Import into ArcMap as XY data
4. Export as shapefile
5. Assign WGS84 datum to shapefile in ArcCatalog
6. Reproject to Lambert Spherical (“US National Albers Equal Area”)
7. Separate by 2x2 degree tile using "tile_num" attribute (so grid will match the netCDF met files) using defination query in ArcMap and exporting to individual shapefiles (256 tiles) as "mask".
8. Open lambert points in qGIS and make 1km grid (shapefile) for each 2x2 tile
9. Assign projection to output (EPSG:2163)
10. Add each new grid shapefile (one at a time) to ArcMap with 2x2 Grid as separate layer
11. Select by location (select from grid x that intersect mask x)
12. Export selected features of grid x (now will be numbered sequentially by record in a way that matches the met NetCDF “ncells”)
13. Clean up: delete extra fields from qGIS (ID,MAXX,MINX,MAXY,MINY) add ncell_id (FID +1) block_id, block_name

Many kinds of queries

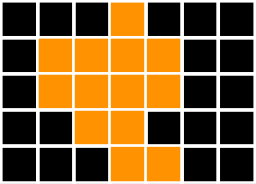
f<variable> <location> <point in time>

avg(srad) at x,y on Dec 2, 2001



tmin for area on May 19, 1992

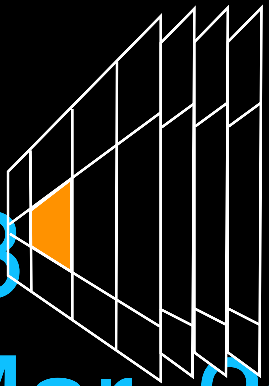
tmax at x,y on May 19, 1992



f<variable> <point location> <duration of time>

tave at x,y during the first quarter of 1983

sum(vpd) at x,y during the last week of Mar, 2003



accessible |akˈsesəbəl|

adjective

1 (of a place) able to be reached or entered : *the town is **accessible by** bus | the building has been made **accessible to** disabled people.*

- (of an object, service, or facility) able to be easily obtained or used : *making learning opportunities more **accessible to** adults.*
- easily understood : *his Latin grammar is lucid and accessible.*
- able to be reached or entered by people in wheelchairs : *it provides specialized features such as nonslip floors and accessible entrances.*

2 (of a person, typically one in a position of authority or importance) friendly and easy to talk to; approachable : *he is more accessible than most tycoons.*

Accessible information is easy to: find, determine what one can do with it, acquire, and use

Factors that affect accessibility: law; technology; culture; semantics; and economics

Law makes sharing permissible; technology makes it possible; culture makes it acceptable; semantics make it understandable; and economics affordable

It is permissible, acceptable, and affordable to access public sector information, but not necessarily possible or understandable

Goals of the new storage: make the information technologically and semantically accessible

Allow access by providing user-interface, application programming interface and documentation