SPATIALLY DISTRIBUTED DATASETS

GLY606 Water Data Analysis & Modeling

Oct 7th 2024

University at Buffalo The State University of New York

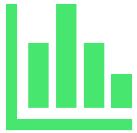
Announcement

- Midterm
  - Time: In class Oct 18th, 2024 (Friday)
  - Length: 50 minutes
  - Format: Open book coding (No communication is allowed)
  - Covered range: Everything until Wednesday's (Oct 9th) lecture
- Reminder
  - Submit the two exercises in HTML format
  - Time: 1 pm Oct 11th, 2024 (Friday)

Recap

Python 101

Numpy, Matplotlib,
Pandas, Urllib



Basic Statistics

Hypothesis testing

Trend analysis

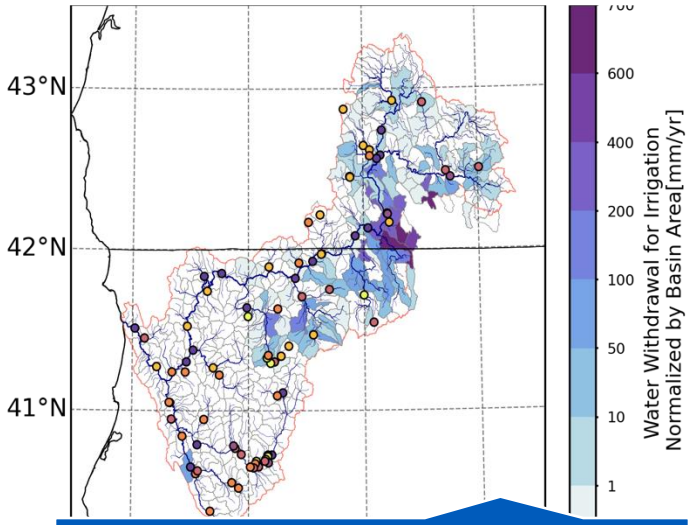
Before today, we studied point-scale data, e.g., time series data for one USGS site…

The streamflow data at one USGS site reflects a lumped information of hydrologic processes upstream of this point.

- Precipitation
- Snowmelt
- Evapotranspiration from vegetations
- …

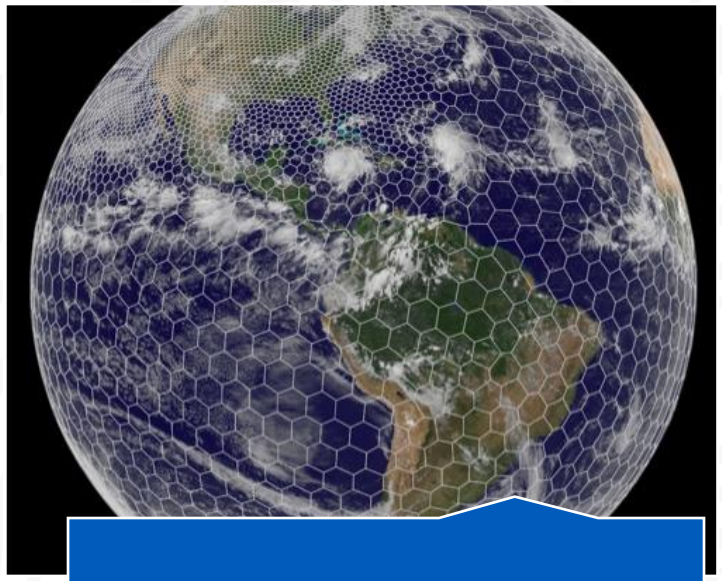
All these relevant hydrologic processes are spatially distributed!!

Spatially distributed datasets



Vector dataset
(Shapefiles)

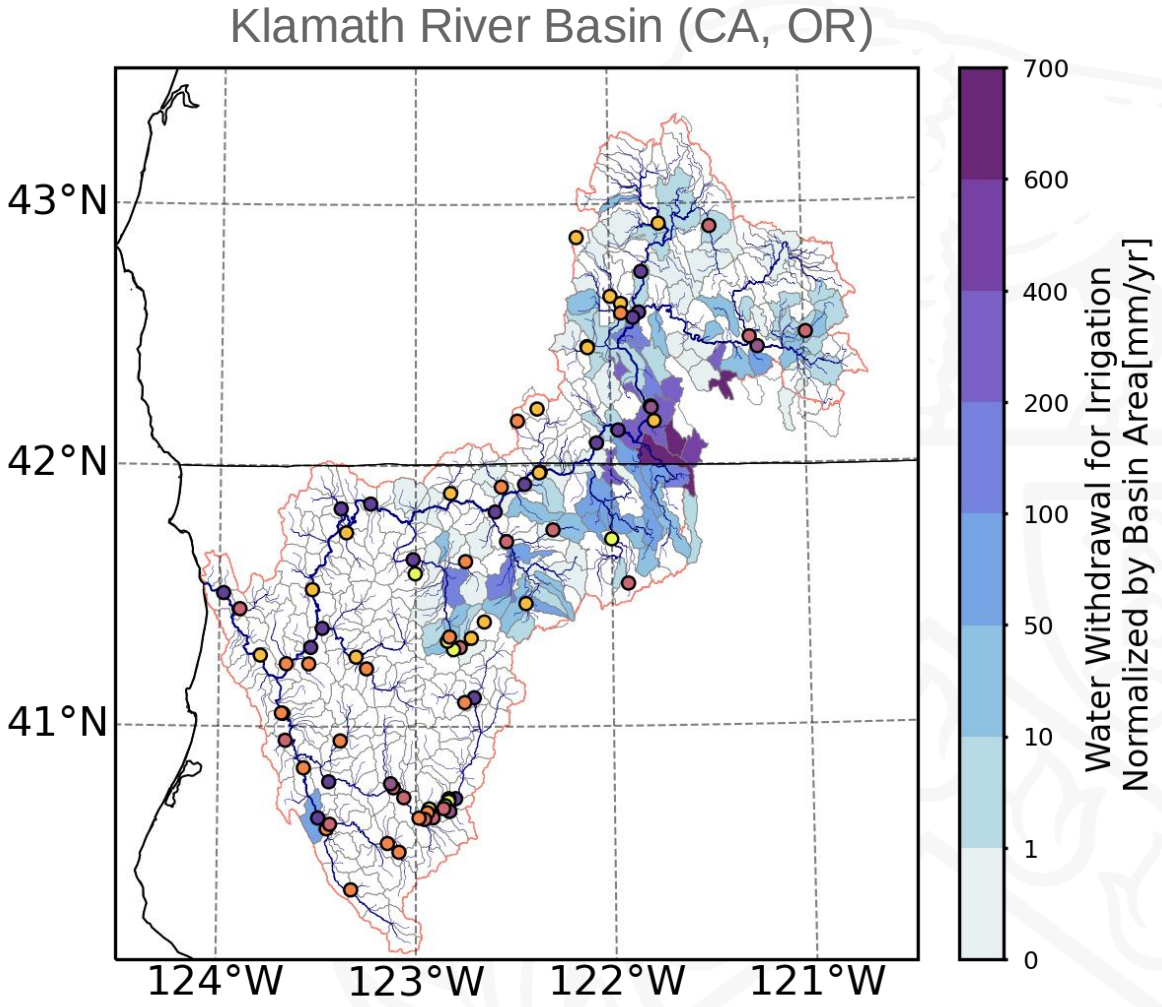
Gridded datasets
(Rectangular/curvilinear)



Unstructured grids

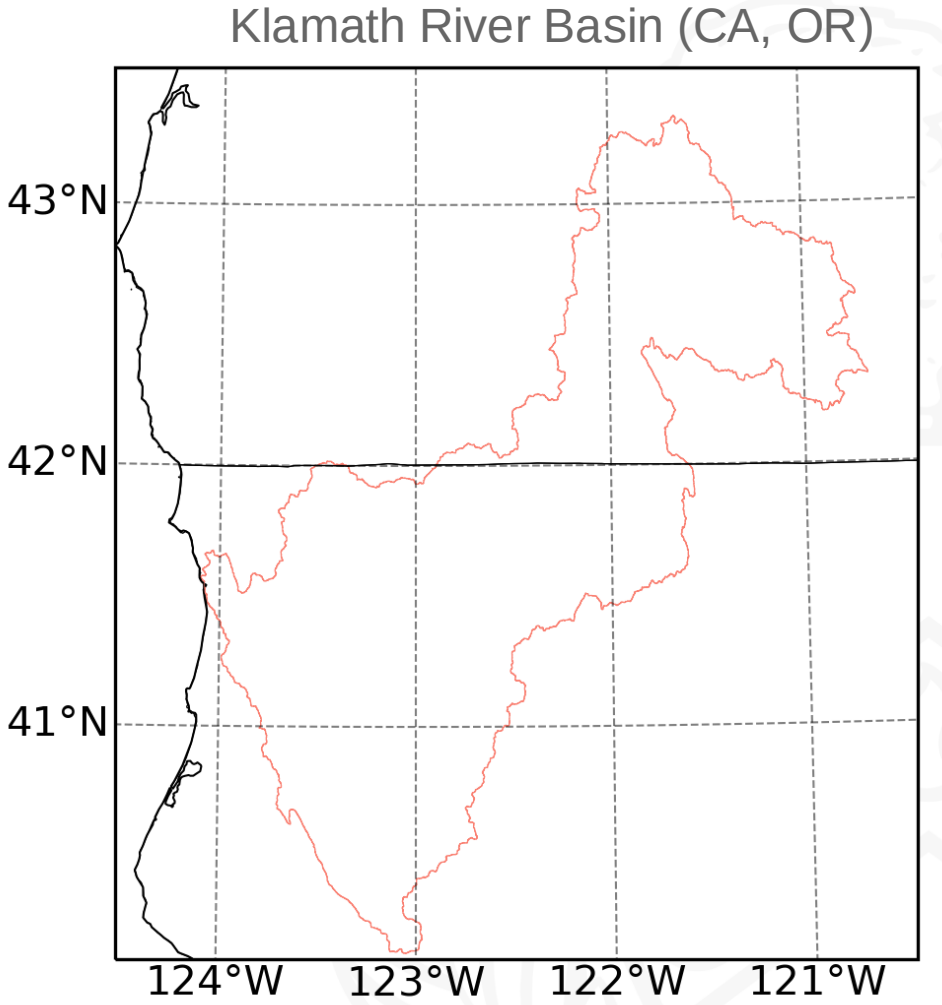
Vector-based data types in hydrology

- Location of site observations (Point)
- River networks (Lines)
- River basin delineation (Polygons)



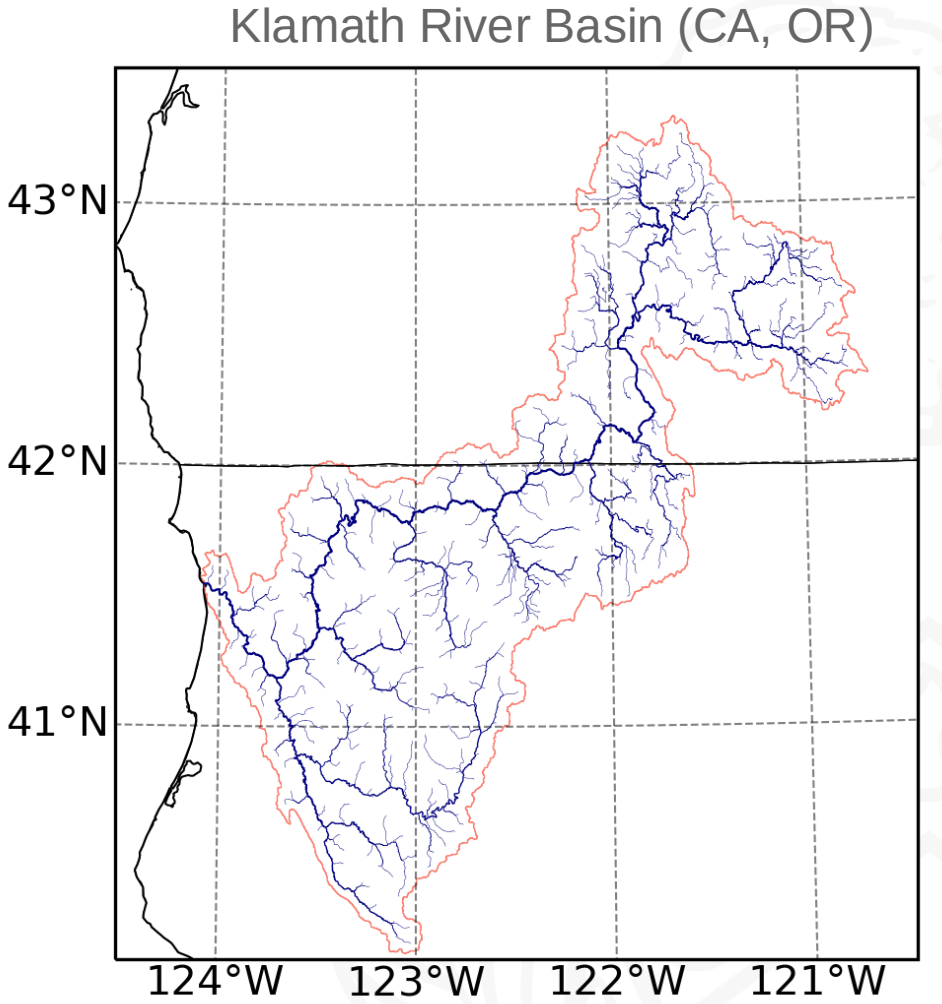
Vector-based data types in hydrology

- Location of site observations (Point)
- River networks (Lines)
- **River basin delineation (Polygons)**



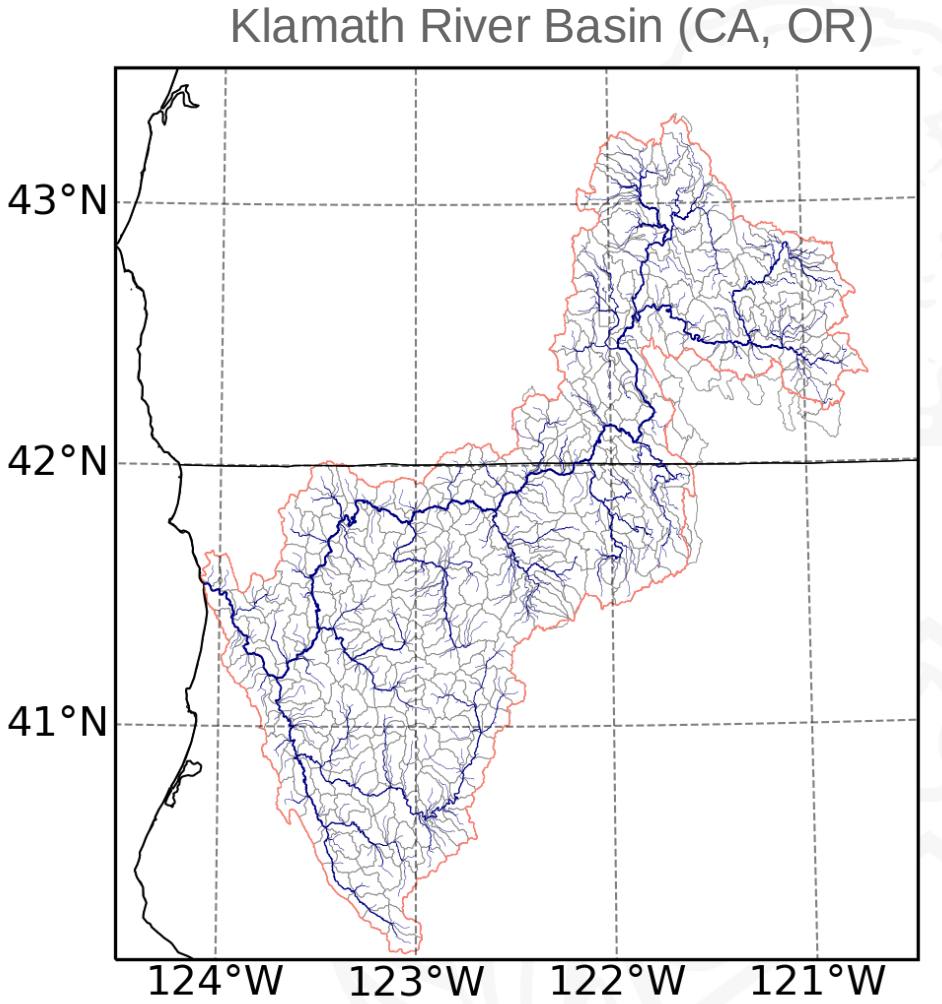
Vector-based data types in hydrology

- Location of site observations (Point)
- **River networks (Lines)**
- River basin delineation (Polygons)



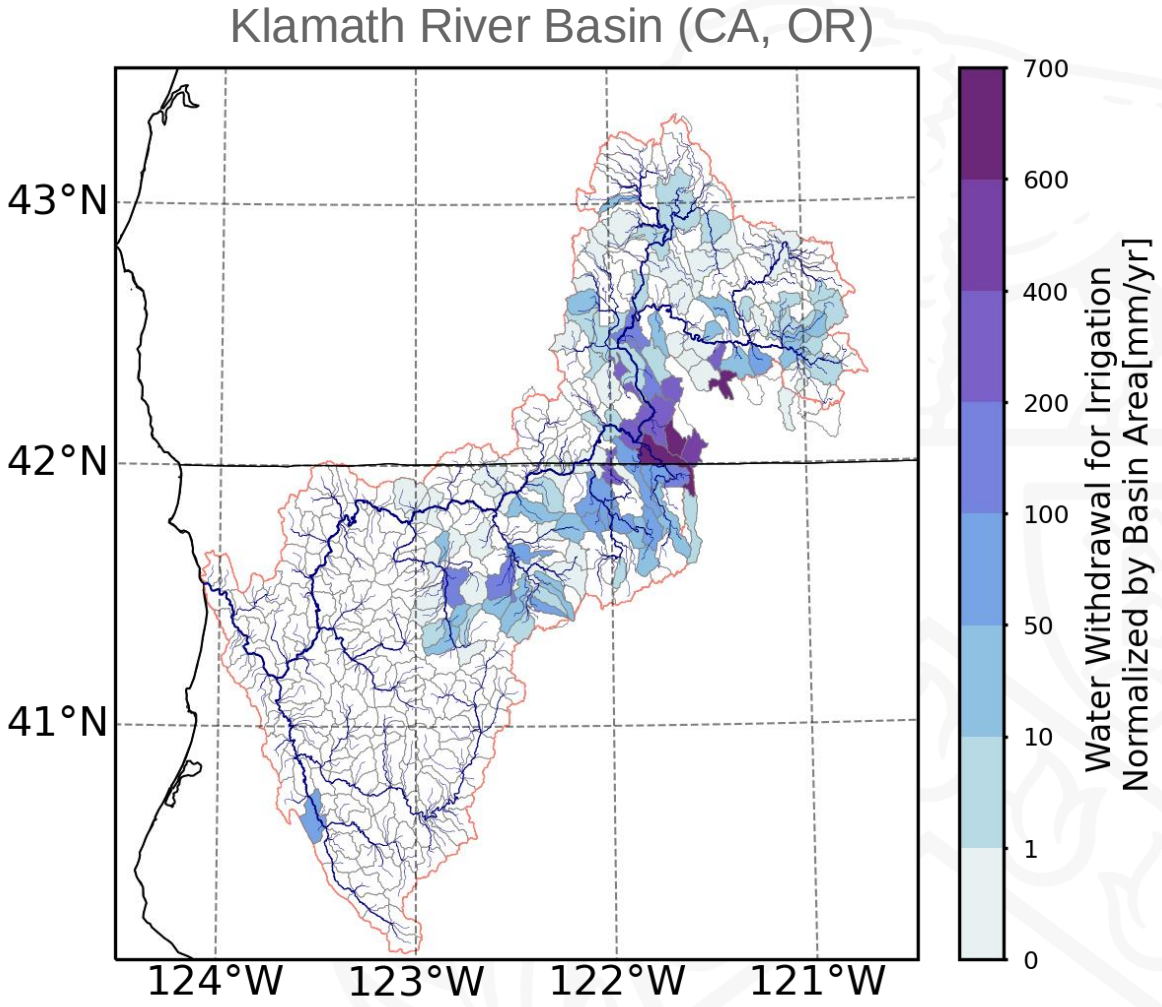
Vector-based data types in hydrology

- Location of site observations (Point)
- River networks (Lines)
- **River basin delineation (Polygons)**



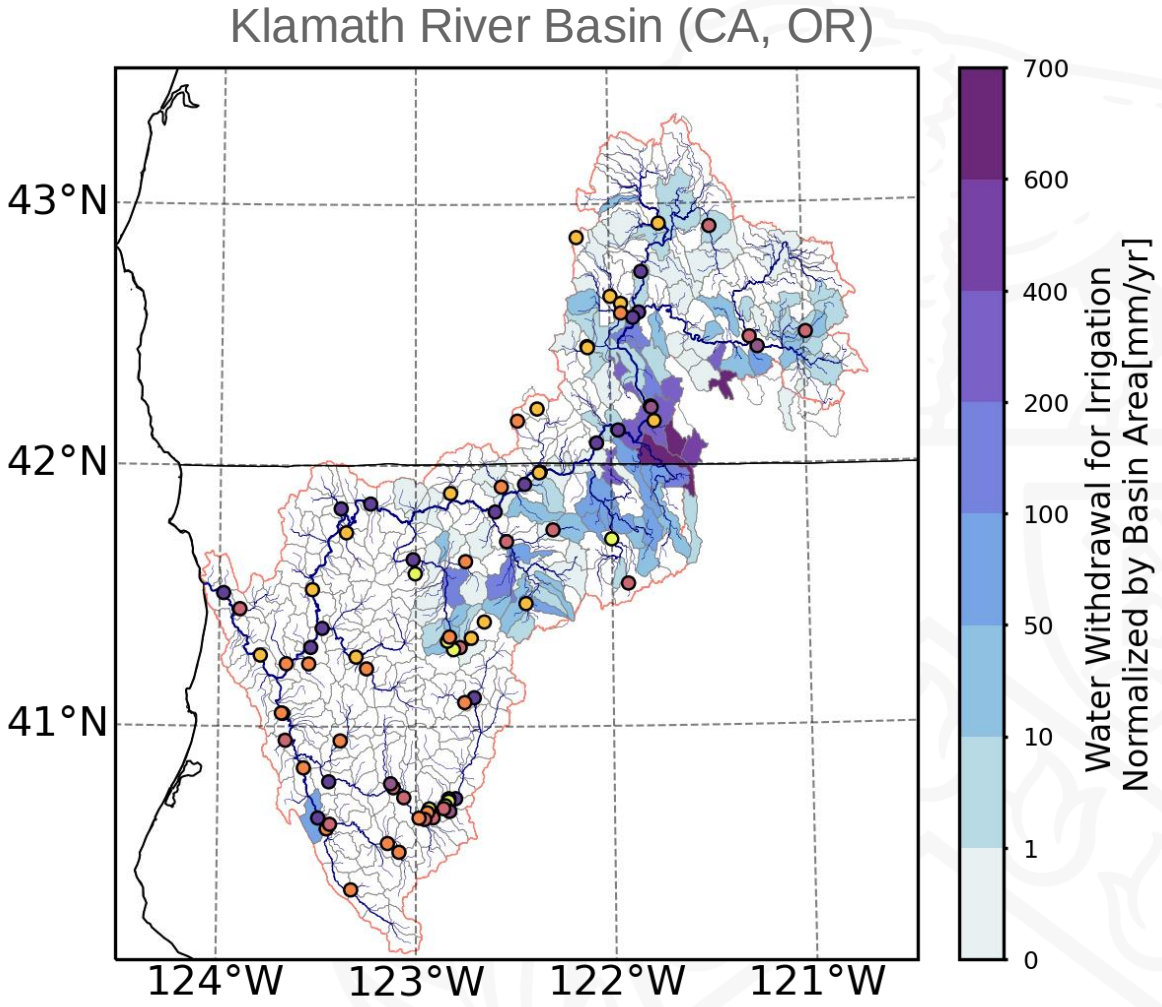
Vector-based data types in hydrology

- Location of site observations (Point)
- River networks (Lines)
- **River basin delineation (Polygons)**



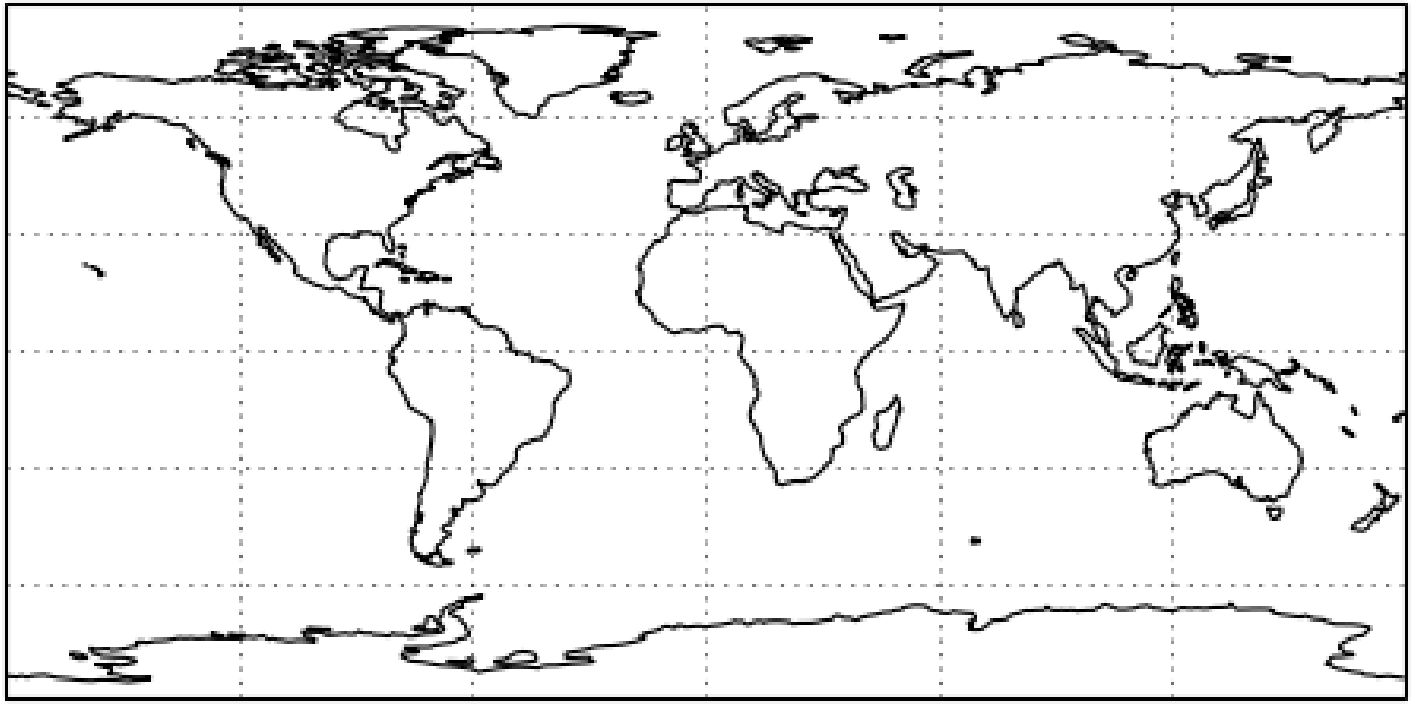
Vector-based data types in hydrology

- **Location of site observations (Point)**
- River networks (Lines)
- River basin delineation (Polygons)

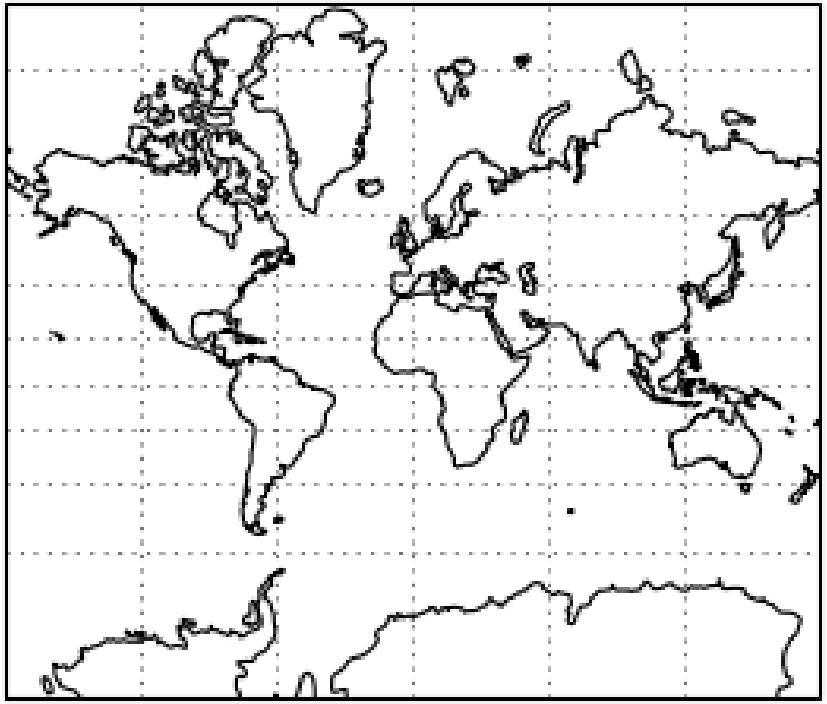


Projection – What does the world look like using different projections?

Plate Carrée



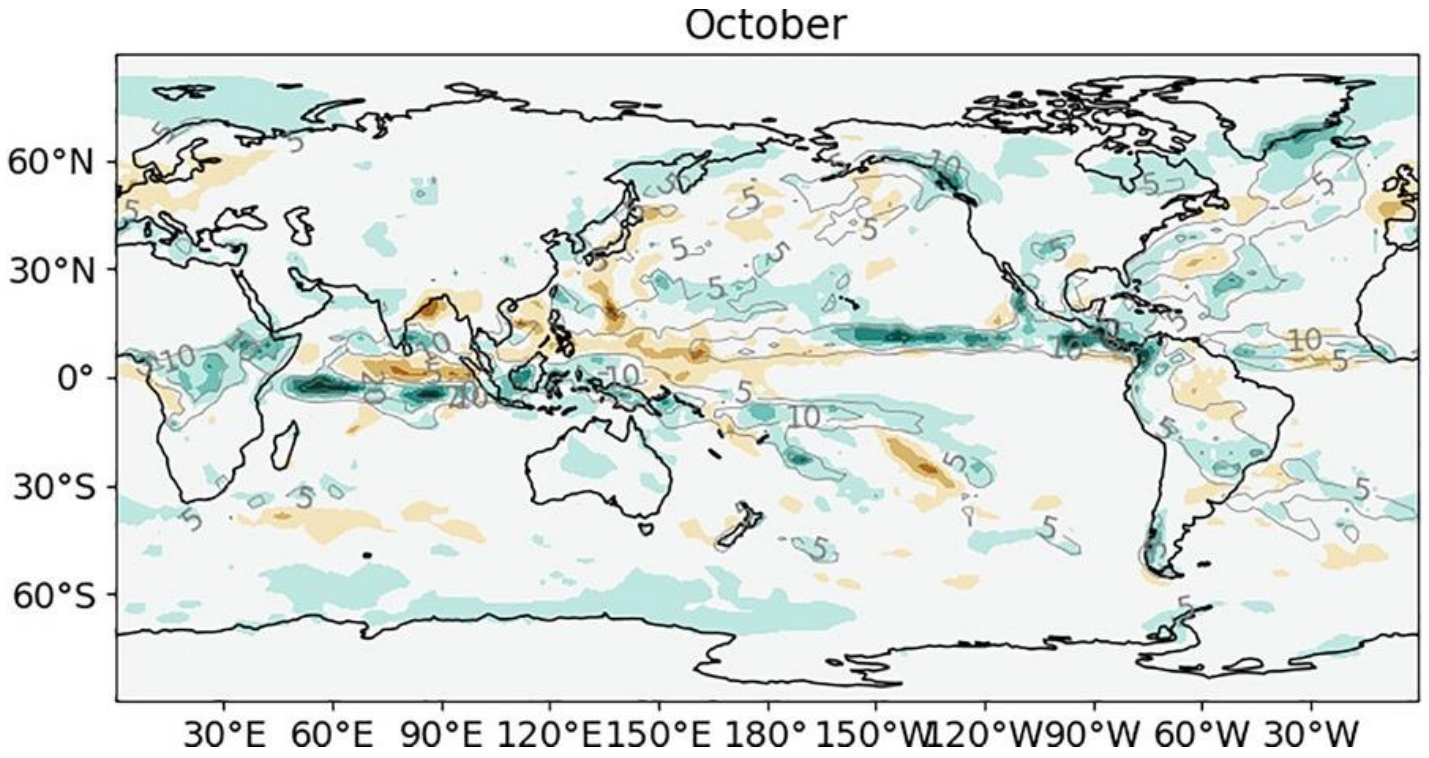
Mercator



https://scitools.org.uk/cartopy/docs/v0.15/crs/projections.html

Projection – What does the world look like using different projections?

Plate Carrée



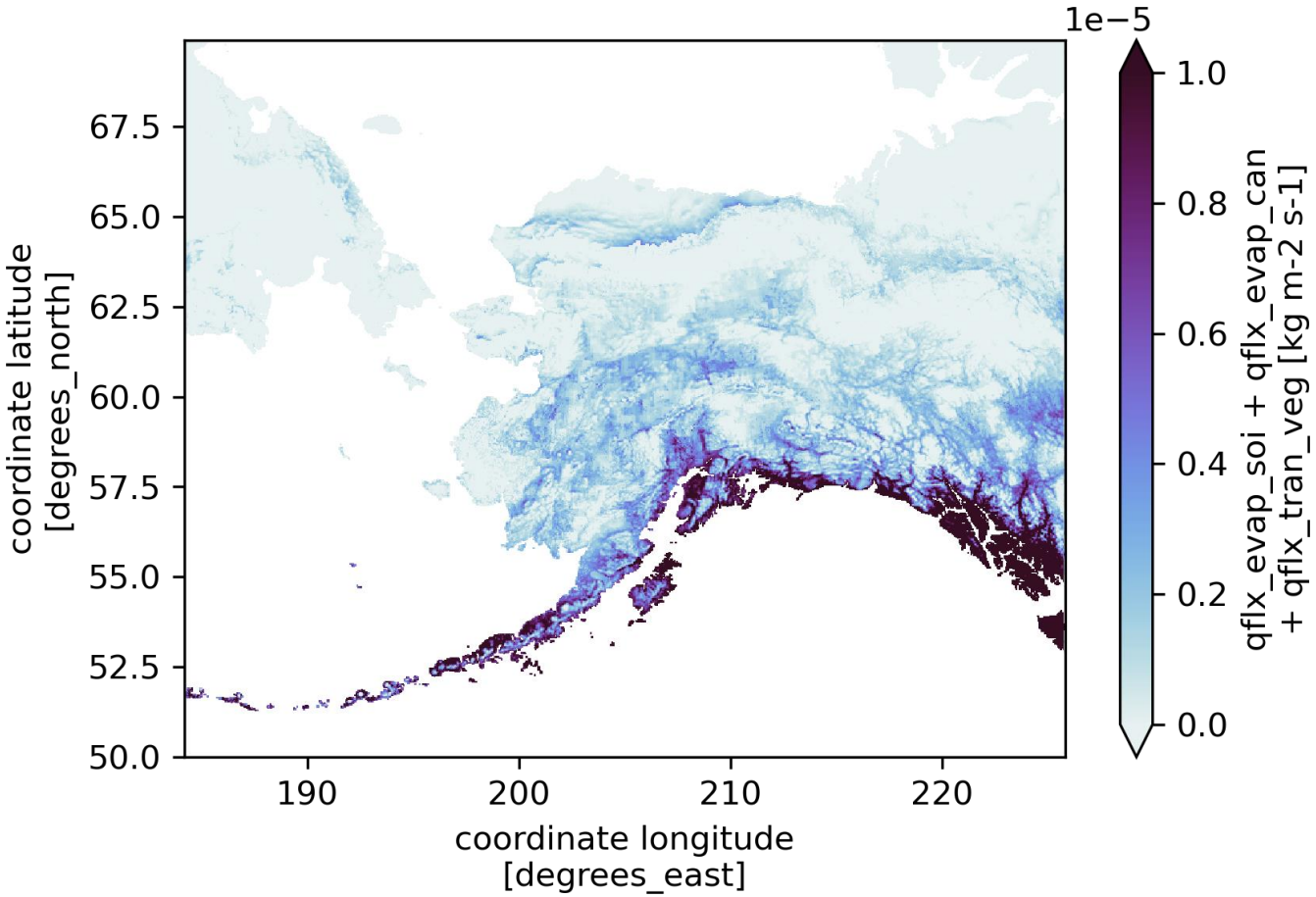
Mercator

https://scitools.org.uk/cartopy/docs/v0.15/crs/projections.html

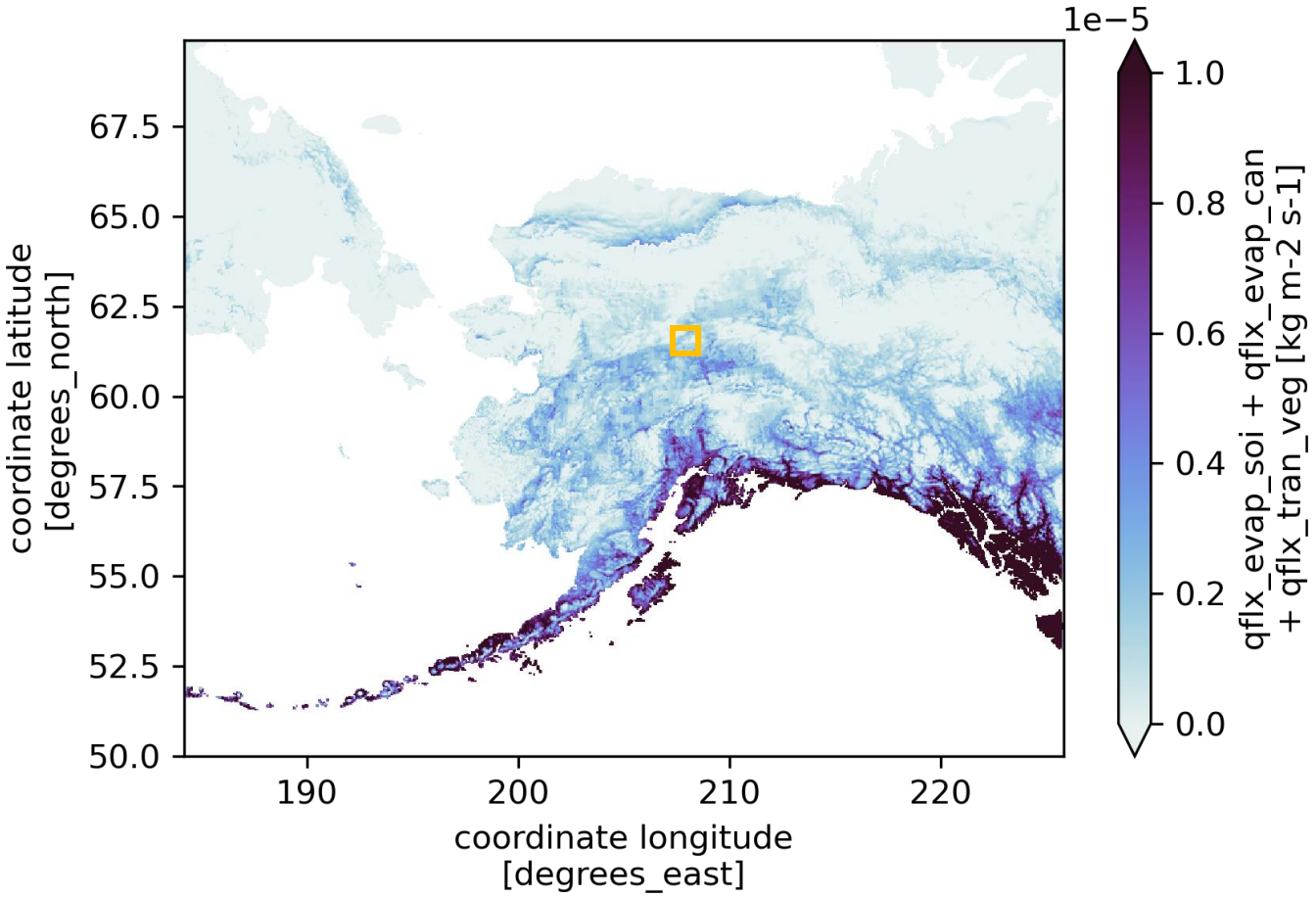
Projection – Why do we need different projections?

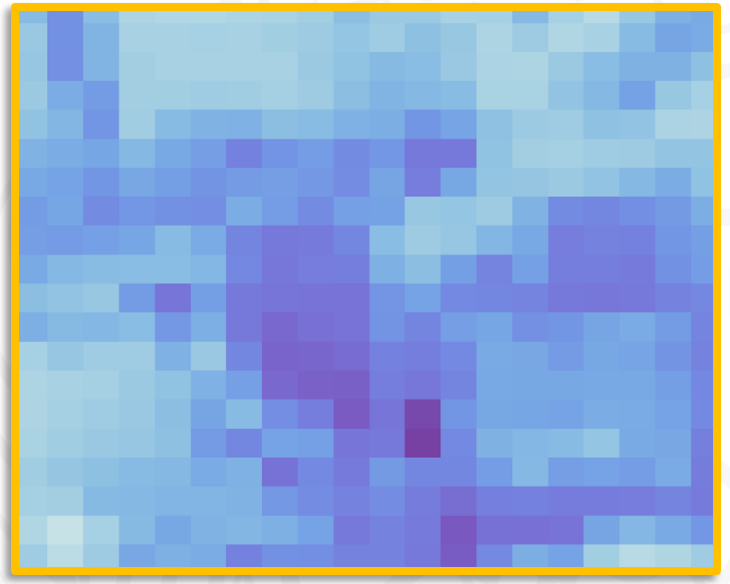
- When it comes to regional maps, we usually choose a projection that is visually pleasing and not distorted too much.

Gridded dataset – What does it look like?



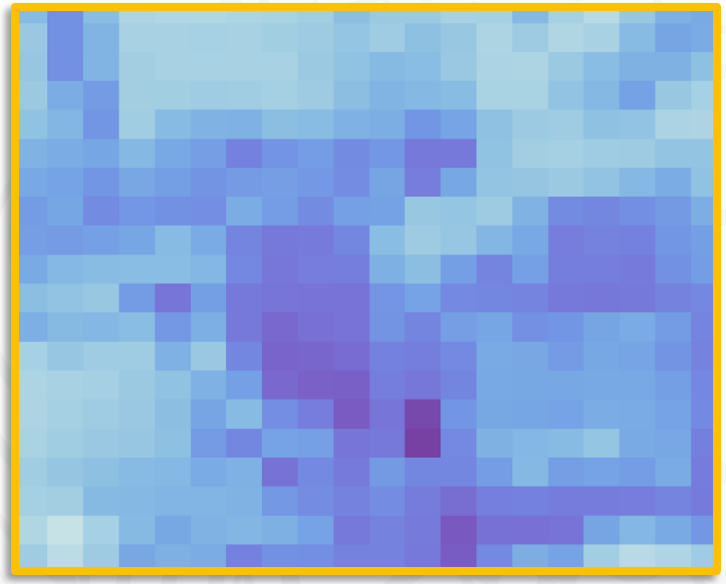
Gridded dataset – What does it look like?



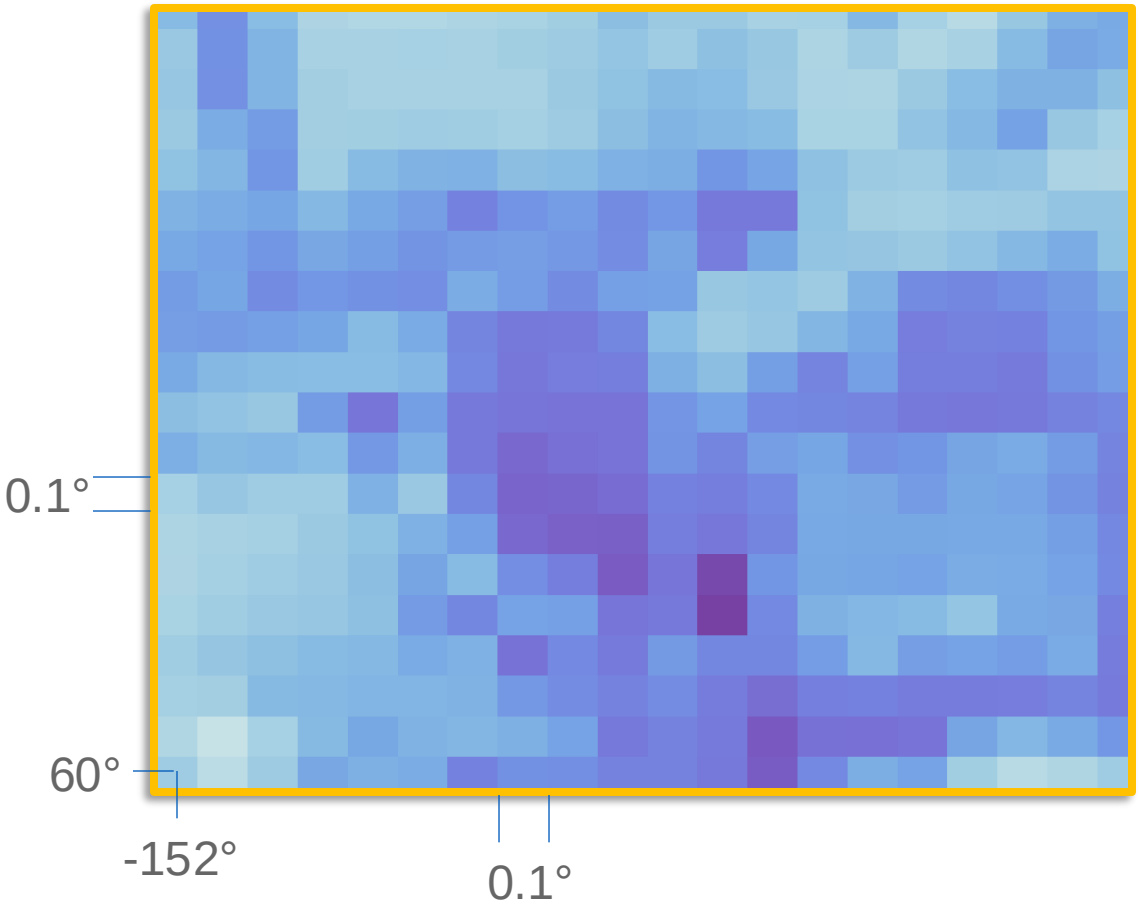


Gridded dataset – What does it look like?

- **Dimensions**
  - Latitudes and Longitudes (Spatially)
  - Time dimension
  - Vertical dimension
    - Above ground: Atmospheric vertical layers
    - Below ground: Soil layers
- **Coordinates**
  - Corresponding values for each dimension that can be used to locate a value
  - For latitude/longitude, its coordinates usually refer to the center of the grid.



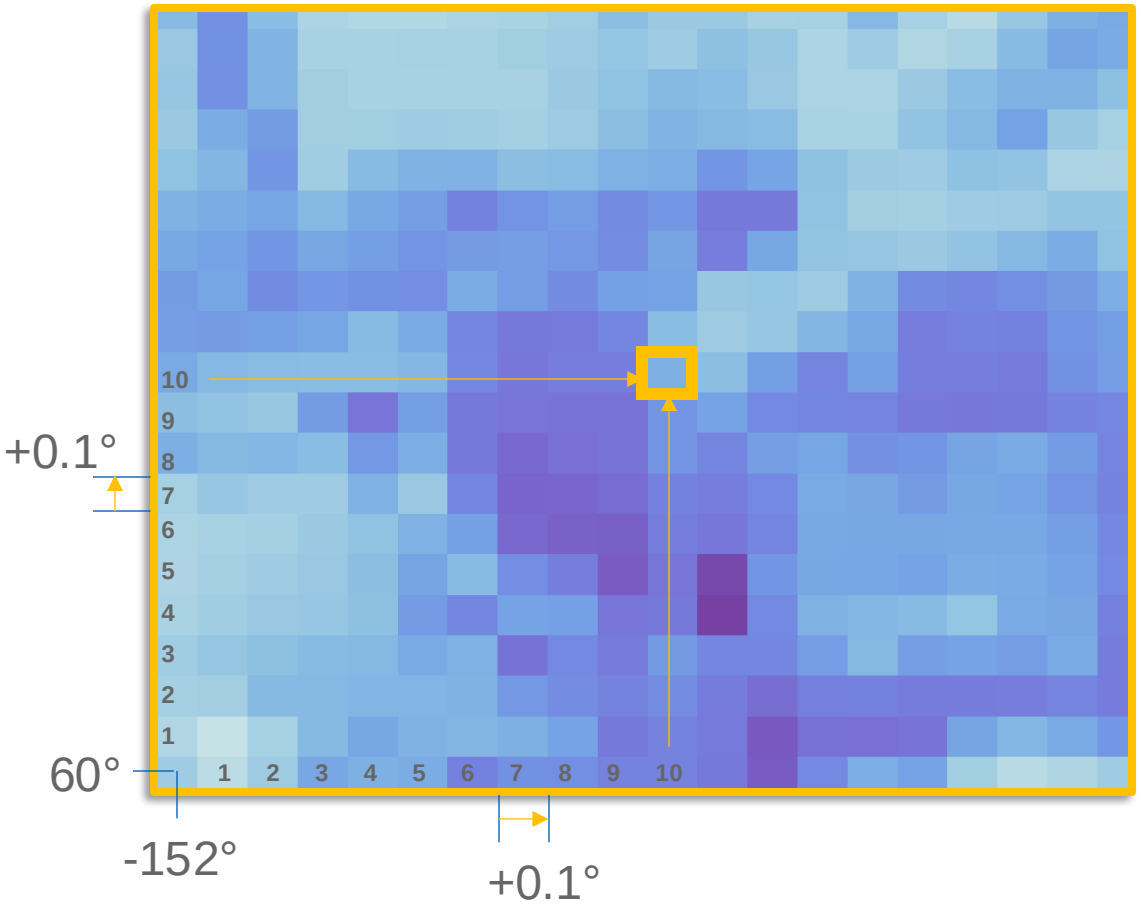
Gridded dataset



The left map shows the total evaporation over a specific region. What are the dimensions for this dataset?

Given the coordinates for the most southeastern corner of this map is 60°N and 152°W, the grid size is 0.1° by 0.1°. How can we get the value of 61°N and 151°W

Gridded dataset



The left map shows the total evaporation over a specific region. What are the dimensions for this dataset?

Given the coordinates for the most southeastern corner of this map is 60°N and 152°W, the grid size is 0.1° by 0.1°. How can we get the value of 61°N and 151°W

Gridded dataset – what type of information is usually stored in gridded datasets?

- **Land surface characteristics**
  - Usually viewed as static, i.e., time-invariant in hydrologic models
  - Examples: land cover information (percent of vegetated area, bare soil, etc.), soil properties…
- **Climate datasets**
  - Usually serves as meteorological forcings, i.e., time-variant variables)
  - Examples: precipitation, air temperature…
- **Output from hydrologic models**
  - Examples: runoff, surface radiation…

Gridded dataset – what are common ways to "generate" gridded datasets?

- Interpolations from observations
- Numerical models
- Satellite Observations
- Reanalysis datasets (Integration of methods mentioned above)

Gridded dataset – ASCII format

- Use a model input as an example
  - In this example, numbers are denoting the river flowing directions

- ncols is the number of columns in the file,
- nrows is the number of rows in the file,
- xllcorner is the longitude of the lower left corner of the grid,
- yllcorner is the latitude of the lower left corner,
- cellsize is the resolution of the grid, and
- NODATA_value is the value that represents missing or unused grid cells.

Flow from each grid cell is given by a by a number:

1.= north
2.= northeast
3.= east
4.= southeast
5.= south
6.= southwest
7.= west
8.= northwest

```
ncols            22
nrows            20
xllcorner        -97.000
yllcorner        38.000
cellsize         0.50
NODATA_value     0
0 0 0 5 5 5 6 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 4.0 4.0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 5 3 5 5 6 7 5 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 5 4.0 5 5 5 5 6 7 5 5 5 7 5 0 0 0 0 0 0
0 0 5 4.0 5 4 5 5 7 7 5 6 7 5 6 0 0 0 0 0 0
4.0 5 7 4.0 3 5 5 7 5 5 7 4 5 6 0 0 0 0 0 0
4.0 3 4.0.4.0.4 3 5 6 7 7 4 5 6 0 0 0 0 0 0
0 4.0 4.0 3 1 2 4.0 4 5 7 5 5 0 0 0 0 0 0 0
0 0 3 4 3 1 8 4 5 4.0 5 5 3 5 6 5 0 0 0 0 0
0 0 0 3 4 5 5 4 5 4 3 5 6 7 7 5 5 7 0 0 0 0
0 0 0 4 3 5 5 3 4.0.4 3 4 3 5 5 7 7 0 0 0 0
0 0 0 4.0 5 4.0.4 3 4.0.4 5 3 5 7 7 0 0 0 0
0 0 0 0 4.0 4.0 4.0 4 5 3 6 7 7 5 5 6 0 5 7
0 0 0 0 0 4.0 3 4 3 5 3 6 7 6 5 7 7 7 7 7 7
0 0 0 0 0 0 0 0 4.0 4 5 7 5 6 7 1 7 1 7 0 0
0 0 0 0 0 0 0 0 3 4.0 5 5 6 7 7 7 6 0 0 0 0
0 0 0 0 0 0 0 0 0 4.0 4 5 7 1 7 7 0 0 0 0 0
0 0 0 0 0 0 0 0 0 3 2 4 5 7 7 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 4.0 5 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Gridded dataset – NetCDF format

Self-Describing. A netCDF file includes information about the data it contains.

Portable. A netCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.

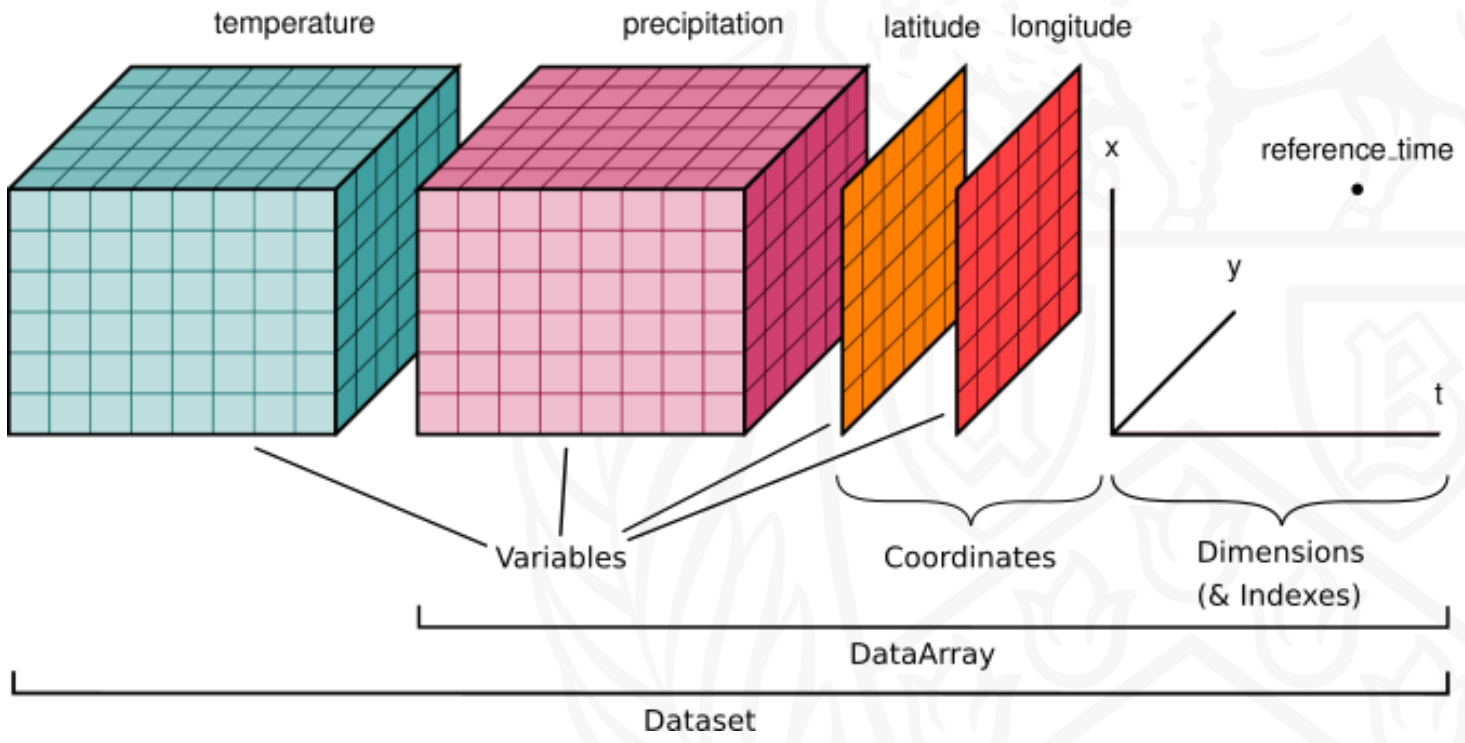
Scalable. Small subsets of large datasets in various formats may be accessed efficiently through netCDF interfaces, even from remote servers.

Appendable. Data may be appended to a properly structured netCDF file without copying the dataset or redefining its structure.

Sharable. One writer and multiple readers may simultaneously access the same netCDF file.

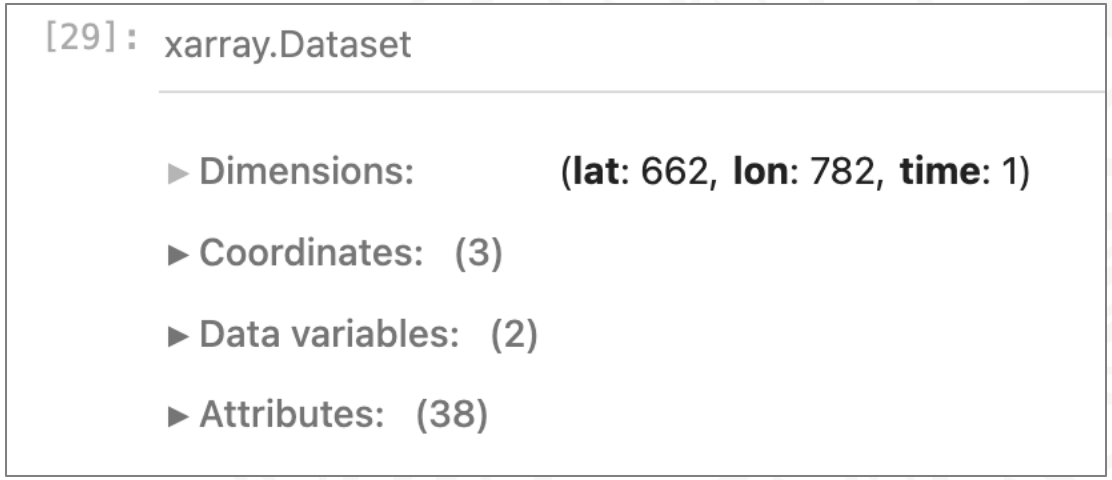
Archivable. Access to all earlier forms of netCDF data will be supported by current and future versions of the software.

Xarray - A python package dealing with N-dimention NetCDF files



Xarray - A python package dealing with N-dimention NetCDF files

```
#Import the packages

import xarray as xr


#read the dataset

ds = xr.open_dataset("path/filename")


#take a brief look of the dataset

ds
```



Xarray - A python package dealing with N-dimention NetCDF files

`Coordinates`

▼ Coordinates:

time	(time)	datetime64[ns]	2064-04-01
lon	(lon)	float32	184.2 184.3 184.3 ... 225.7 225.8
lat	(lat)	float32	50.01 50.04 50.07 ... 69.86 69.89

```
#We will be able to select the data directly
#using its corresponding coordinates
ds.sel(time="2064-04-01",lon=184.2,lat=69.86)
```

Xarray - A python package dealing with N-dimention NetCDF files

```
Data Variables
```

▼ Data variables:

FSNO	(time, lat, lon)	float32	...
long_name :	fraction of ground covered by snow		
units :	unitless		
cell_methods :	time: mean		
QFLX_EVAP_TOT	(time, lat, lon)	float32	...
long_name :	qflx_evap_soi + qflx_evap_can + qflx_tran_veg		
units :	kg m-2 s-1		
cell_methods :	time: mean		

```
#We will be able to select targeted variables
ds["FSNO"].sel(time="2064-04-01",lon=184.2,lat=69.86)
```

Xarray - A python package dealing with N-dimention NetCDF files

`Attributes`

Self-Describing. A netCDF file includes information about the data it contains.

▾ Attributes:

title :	CLM History file information
comment :	NOTE: None of the variables are weighted by land fraction!
Conventions :	CF-1.0
history :	created on 07/29/23 03:27:24
source :	Community Terrestrial Systems Model
hostname :	cheyenne
username :	tcraig
version :	rasm2_2_01_plus
revision_id :	$Id: histFileMod.F90 429032012 - 12 - 2115 : 32 : 10Zmuszala$
case_title :	UNSET
case_id :	NNA.4km.fPGWh.2033.004
Surface_dataset :	surfdata_nna4a.spatial_distrib.all.fillland.220411.T14.cdf5.nc
Initial_condition...	PGW_high_init_hh_4km.clm2.r.2033-06-01-00000.nc
PFT_physiologi...	clm50_params.c210507.nc
ltype_vegetate...	1
ltype_crop :	2
ltype_UNUSED :	3
ltype_landice_...	4
ltype_deep_lake :	5
ltype_wetland :	6
ltype_urban_tbd :	7
ltype_urban_hd :	8

Unstructured grids

Seamless zooming

- MPAS's hexagonal grid system allows the model to zoom in or out across different parts of the globe.

Smooth transitions

- MPAS's unstructured variable resolution meshes can be generated with smoothly varying mesh transitions.

Scalability

- MPAS's numerics scale efficiently across shared and distributed memory on massive parallel supercomputers.