HAS Tools:

Overview of Earth Data Repositories

November 15, 2024

The era of big data geoscience is here

(though, of course, not evenly distributed)

THE WALL STREET JOURNAL.

THE FUTURE OF EVERYTHING | DATA

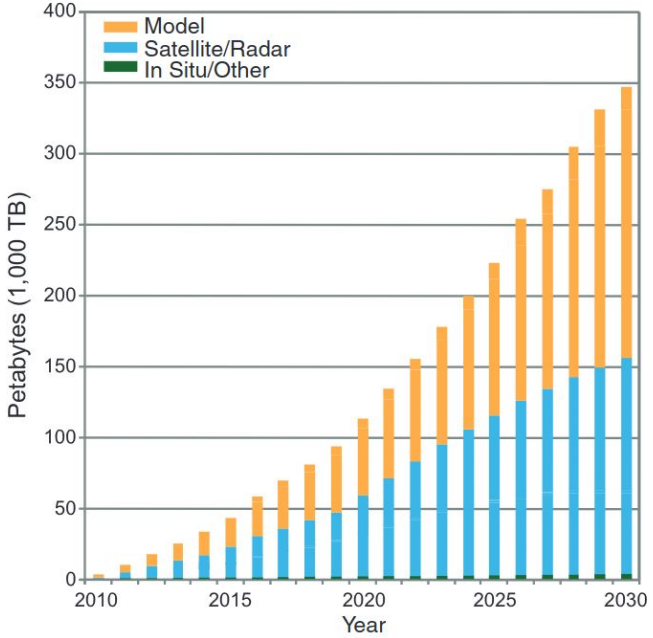
Climate Change Data Deluge Has Scientists Scrambling for Solutions

As earth-observing satellites, aircraft and ocean buoys churn out ever-rising amounts of information about our planet, data managers turn to cloud computing and artificial intelligence

AUTHOR
ROBERT LEE HOTZ

PUBLISHED
DEC. 5, 2021 11:00 AM ET

https://www.wsj.com/articles/climate-change-data-deluge-has-scientists-scrambling-for-solutions-11638720016

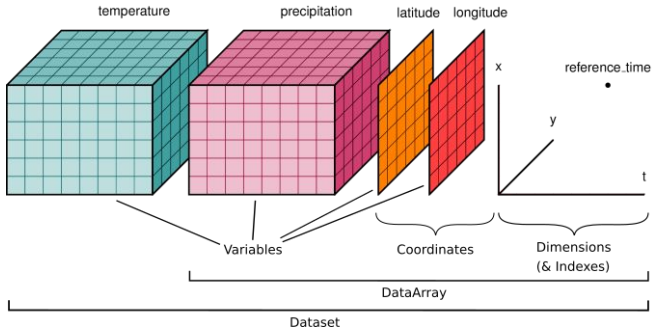


https://doi-org.10.1126/science.1197869

Use of data science & machine learning is already revolutionizing Earth and environment sciences

The Pangeo python data stack of xarray, dask, and zarr enables working with huge, cloud-based datasets

Enables you to scale beyond your laptop, even on your laptop (and beyond)



Zarr

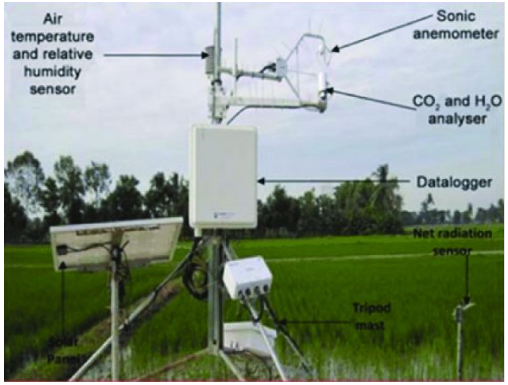
dask

Datasets come in many flavors

Datasets may be:

- Based on in-situ data
- Remotely sensed data
- Historic records/proxies
- Modeled results
- Any combination thereof (reanalysis)

Further, not all datasets are created equal!

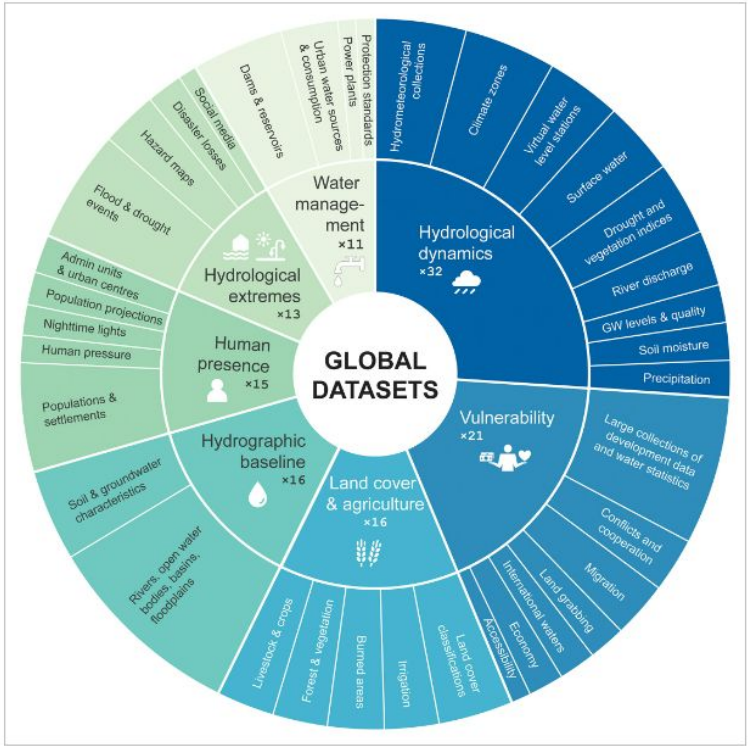


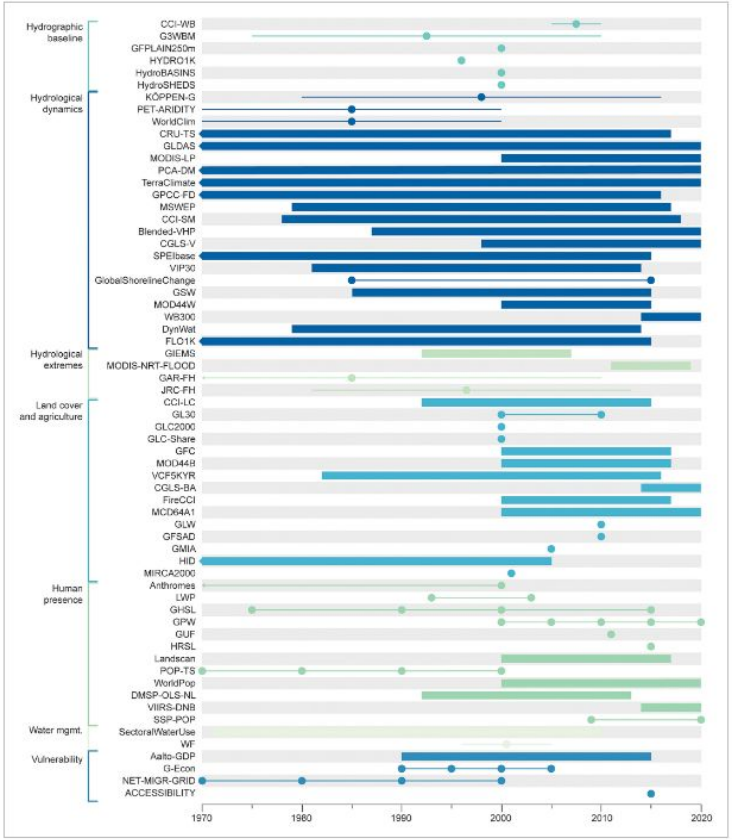
A quick scattershot of datasets I won't be covering, but may be of interest

- FluxNet (Ecosystems & Hydromet)
- Caravan/CAMELS (Catchment Hydrology)
- Airborne Snow Observatory (Snow hydrology)
- MultiRadar/MultiSensor (MRMS, radar precip)
- Integrated Global Radiosonde Archive (Atmospheric soundings)
- SAGE (Seismology)
- PhenoCam (Vegetation phenology)
- CoCoRaHS (Rain, snow, hail)
- National Buoy Data Center (ocean currents, meteorology)
- …the list goes on…

Slide courtesy of Alex Saunders

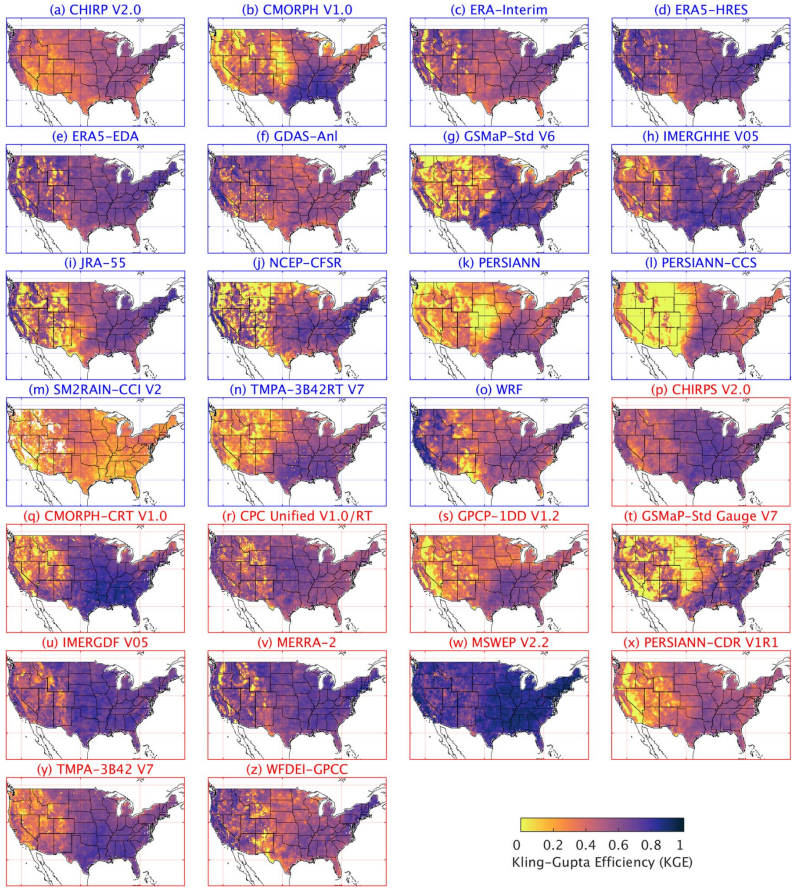
Lindersson et al., 2020 - A review of freely accessible global datasets for the study of floods, droughts and their interactions with human societies





Beck et al., 2019 - Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS

Slide courtesy of Alex Saunders



Name	Details	Data source(s)	Spatial resolution	Spatial coverage	Temporal resolution	Temporal coverage	Reference or website
CHIRPS V2.0	Climate Hazards group InfraRed Precipitation with Stations (CHIRPS) V2.0	G, S, R, A	0.05°	Land, 50° N/S	Daily	1981–NRT[2]	Funk et al. (2015a)
CMORPH-CRT V1.0	CPC MORPHing technique (CMORPH) bias corrected (CRT) V1.0	G, S	0.07°	60° N/S	30 min	1998–2015	Joyce et al. (2004), Xie et al. (2017)
CPC Unified V1.0/RT	Climate Prediction Center (CPC) Unified V1.0 and RT	G	0.5°	Land	Daily	1979–NRT[2]	Xie et al. (2007), Chen et al. (2008)
GPCP-1DD V1.2	Global Precipitation Climatology Project (GPCP) 1-Degree Daily (1DD) Combination V1.2	G, S	1°	Global	Daily	1996–2015	Huffman et al. (2001)
GSMaP-Std Gauge V7	Global Satellite Mapping of Precipitation (GSMaP) Moving Vector with Kalman (MVK) Standard gauge-corrected V7	G, S	0.1°	60° N/S	Hourly	2000–NRT[1]	Ushio et al. (2009)
IMERGDF V05	Integrated Multi-satellitE Retrievals for GPM (IMERG) final run V05	G, S	0.1°	60° N/S	30 min	2014–NRT[3, 4]	Huffman et al. (2014, 2018)
MERRA-2	Modern-Era Retrospective Analysis for Research and Applications 2	G, S, R	~ 0.5°	Global	Hourly	1980–NRT[3]	Gelaro et al. (2017), Reichle et al. (2017)
MSWEP V2.2	Multi-Source Weighted-Ensemble Precipitation (MSWEP) V2.2	G, S, R, A	0.1°	Global	3-hourly	1979–NRT[1]	Beck et al. (2017b, 2019)
PERSIANN-CDR V1R1	Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) Climate Data Record (CDR) V1R1	G, S	0.25°	60° N/S	Daily	1983–2016	Ashouri et al. (2015)
TMPA-3B42 V7	TRMM Multi-satellite Precipitation Analysis (TMPA) 3B42 V7	G, S	0.25°	50° N/S	3-hourly	2000–2017	Huffman et al. (2007)
WFDEI-GPCC	WATCH Forcing Data ERA-Interim (WFDEI) corrected using Global Precipitation Climatology Centre (GPCC)	G, R	0.5°	Land	3-hourly	1979–2016	Weedon et al. (2014)

Sun et al., 2018 - A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons

Slide courtesy of Alex Saunders

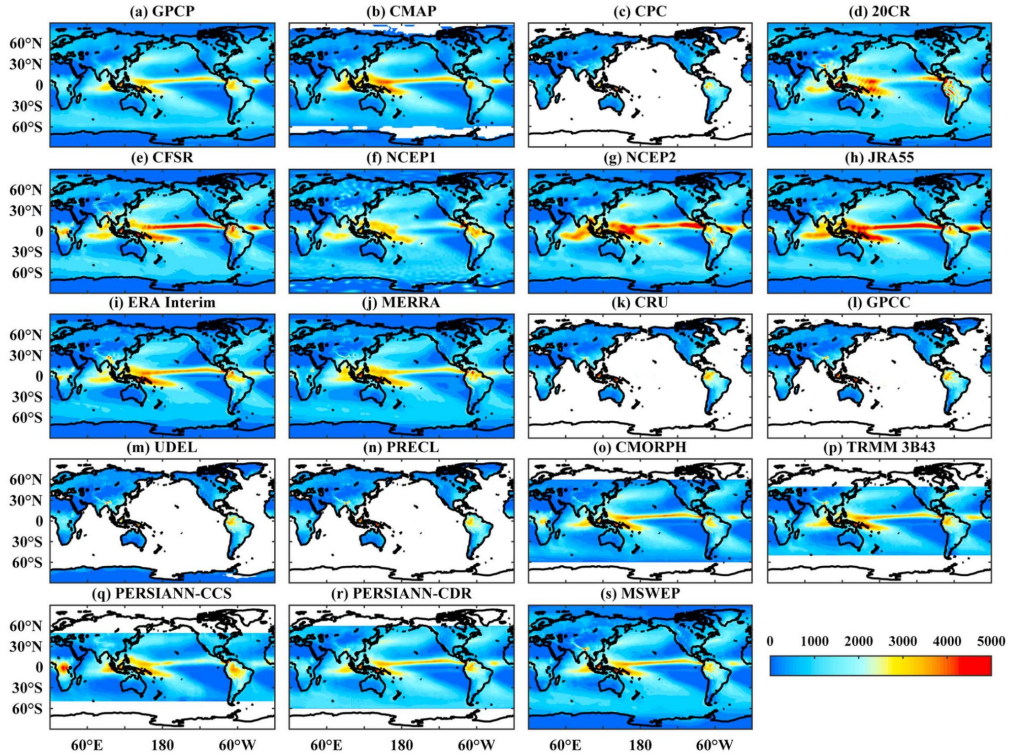
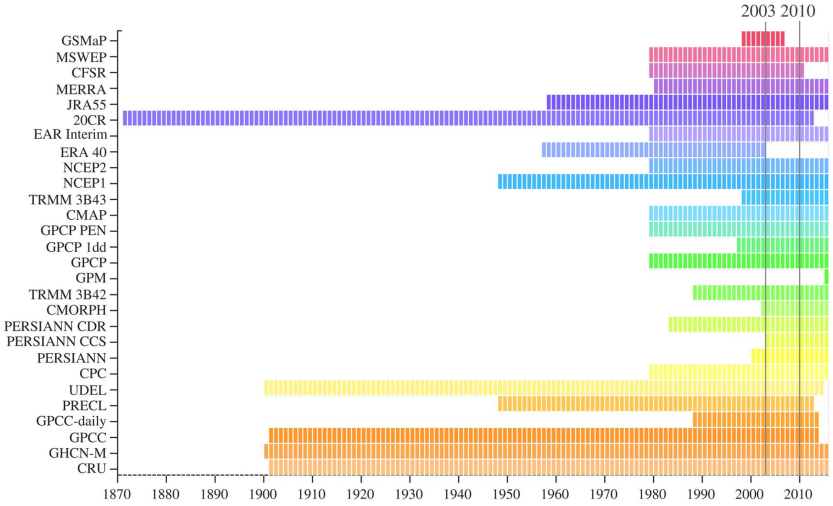


Figure 8. Spatial distribution of the 8 year (2003–2010) average precipitation estimates from different products. Precipitation estimates were based on the original spatial resolution of each data set, without reinterpolation to a unified resolution.



Rasmussen et al., 2023 - CONUS404: The NCAR–USGS 4-km Long-Term Regional Hydroclimate Reanalysis over the CONUS

Slide courtesy of Alex Saunders

Table 2. Qualitative comparison of hydrological model forcing datasets. Good performance means that datasets have higher skill compared to others while suboptimal performance is interpreted relatively to other datasets.

		Gridded station observation based	Observation/ Model fusions	Reanalyses	Convection-Permitting Downscaled Reanalyses	Climate Models	Convection-Permitting Downscaled Climate Models
		PRISM, Livneh, Daymet	AORC, NLDAS, gridMET	NARR, ERA5, MERRA2	CONUS404	CMIP6, CORDEX	CONUS-scale future scenarios
Biases	Systematic differences to in-situ observations						
Realism climate variability	Representation of interannual and decadal variability						
Realism seasonal variability	Representation of seasonal variability						
Realism diurnal variability	Representation of diurnal variabilities	typically daily					
Large-scale extremes	E.g., droughts, heatwaves, pluvial conditions						
Small-scale extremes	E.g., downpours, convective extremes						
Homogeneity	Occurrence of artificial signals in the climate record						
Spatial coverage	Data availability for all CONUS watersheds including over water						
Intervariable consistence	Physical consistency between variables						
Record length	Length of the data record						
Future projections	Availability of future climate projections						

Performance

good | medium | sub-optimal | not available

Slide courtesy of Alex Saunders

References for data intercomparisons

AghaKouchak, A., Behrangi, A., Sorooshian, S., Hsu, K., Amitai, E., 2011. Evaluation of satellite-retrieved extreme precipitation rates across the central United States. Journal of Geophysical Research: Atmospheres 116. https://doi.org/10.1029/2010JD014741

Ali, M.H., Popescu, I., Jonoski, A., Solomatine, D.P., 2023. Remote Sensed and/or Global Datasets for Distributed Hydrological Modelling: A Review. Remote Sensing 15, 1642. https://doi.org/10.3390/rs15061642

Beck, H.E., Pan, M., Roy, T., Weedon, G.P., Pappenberger, F., van Dijk, A.I.J.M., Huffman, G.J., Adler, R.F., Wood, E.F., 2019. Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. Hydrology and Earth System Sciences 23, 207–224. https://doi.org/10.5194/hess-23-207-2019

Beck, H.E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A.I.J.M., Weedon, G.P., Brocca, L., Pappenberger, F., Huffman, G.J., Wood, E.F., 2017. Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. Hydrology and Earth System Sciences 21, 6201–6217. https://doi.org/10.5194/hess-21-6201-2017

Essou, G.R.C., Arsenault, R., Brissette, F.P., 2016. Comparison of climate datasets for lumped hydrological modeling over the continental United States. Journal of Hydrology 537, 334–345. https://doi.org/10.1016/j.jhydrol.2016.03.063

Lindersson, S., Brandimarte, L., Mård, J., Di Baldassarre, G., 2020. A review of freely accessible global datasets for the study of floods, droughts and their interactions with human societies. WIREs Water 7, e1424. https://doi.org/10.1002/wat2.1424

Maggioni, V., Meyers, P.C., Robinson, M.D., 2016. A Review of Merged High-Resolution Satellite Precipitation Product Accuracy during the Tropical Rainfall Measuring Mission (TRMM) Era. https://doi.org/10.1175/JHM-D-15-0190.1

Rasmussen, R.M., Chen, F., Liu, C.H., Ikeda, K., Prein, A., Kim, J., Schneider, T., Dai, A., Gochis, D., Dugger, A., Zhang, Y., Jaye, A., Dudhia, J., He, C., Harrold, M., Xue, L., Chen, S., Newman, A., Dougherty, E., Abolafia-Rosenzweig, R., Lybarger, N.D., Viger, R., Lesmes, D., Skalak, K., Brakebill, J., Cline, D., Dunne, K., Rasmussen, K., Miguez-Macho, G., 2023. CONUS404: The NCAR–USGS 4-km Long-Term Regional Hydroclimate Reanalysis over the CONUS. https://doi.org/10.1175/BAMS-D-21-0326.1

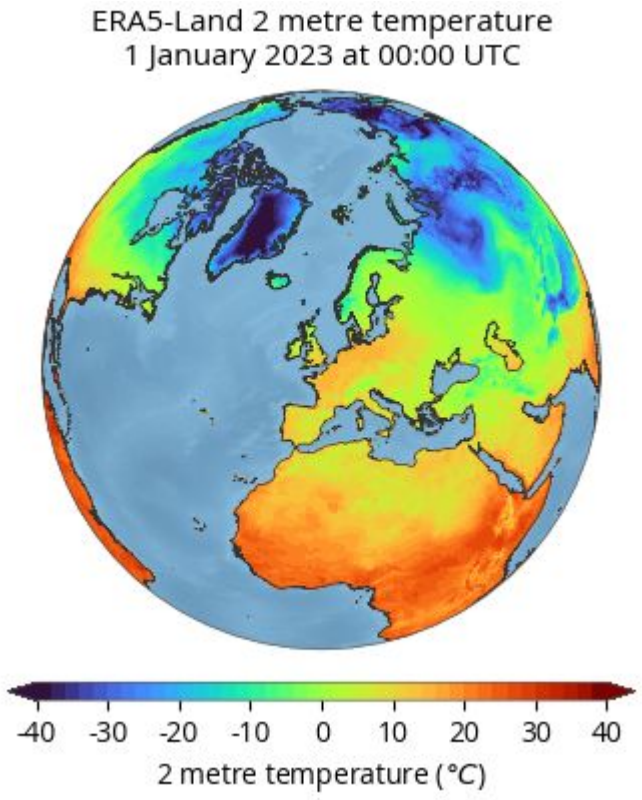
Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., Hsu, K.-L., 2018. A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. Reviews of Geophysics 56, 79–107. https://doi.org/10.1002/2017RG000574

Tarek, M., Brissette, F.P., Arsenault, R., 2020. Large-Scale Analysis of Global Gridded Precipitation and Temperature Datasets for Climate Change Impact Studies. https://doi.org/10.1175/JHM-D-20-0100.1

On reanalysis

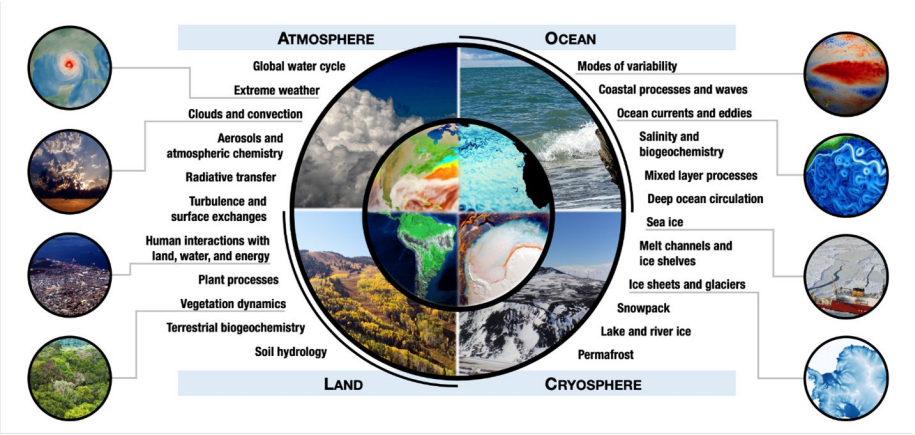
The role of reanalysis:

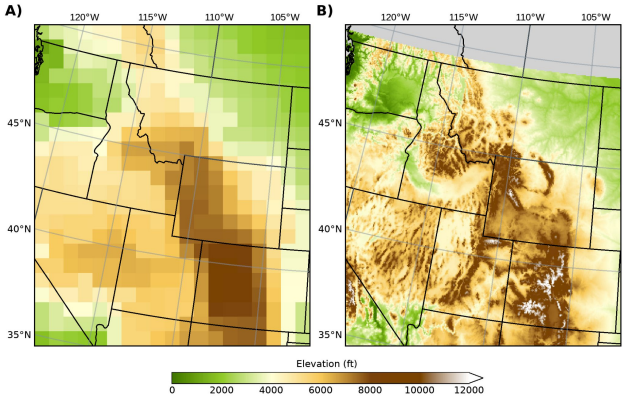
- Learn as much as we possibly can by synthesizing as much data as possible
- Uses model results, historic data, and modern observations
- Produce a spatiotemporally complete view of the Earth System



Climate projections & downscaling

- You have certainly heard of climate projections
- Earth systems models represent the best state of our knowledge
- High process complexity tends to mean low spatial resolution
- To understand regional/local impacts downscaling is done
  - Dynamical vs statistical





Clearly there is a huge amount of data to be analyzed, but how can this be possible?

Increasingly these datasets are available from major cloud providers directly

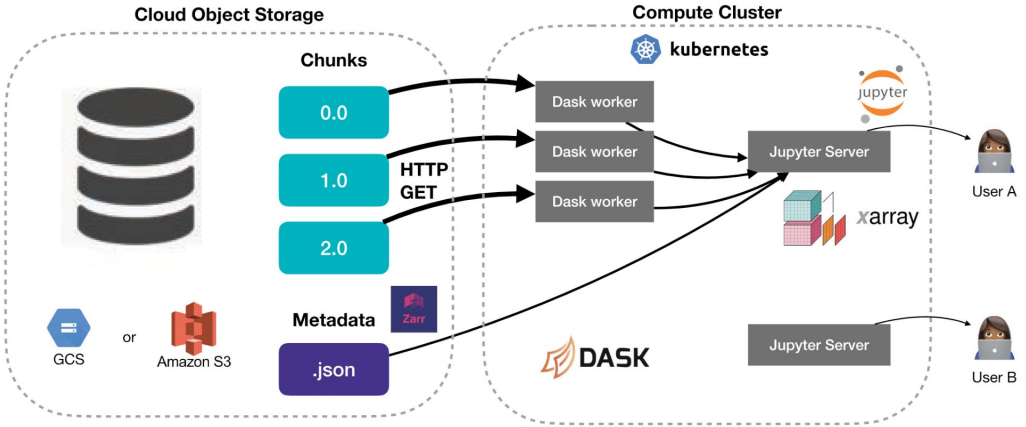
- Historically, data was gate-kept behind either private servers or walled gardens of government agencies
- Data access possible, but required download before any analysis could be done
- Cloud storage means you can access data on the fly (though you may have to pay for access)

Azure Blob Storage

Google Cloud Storage

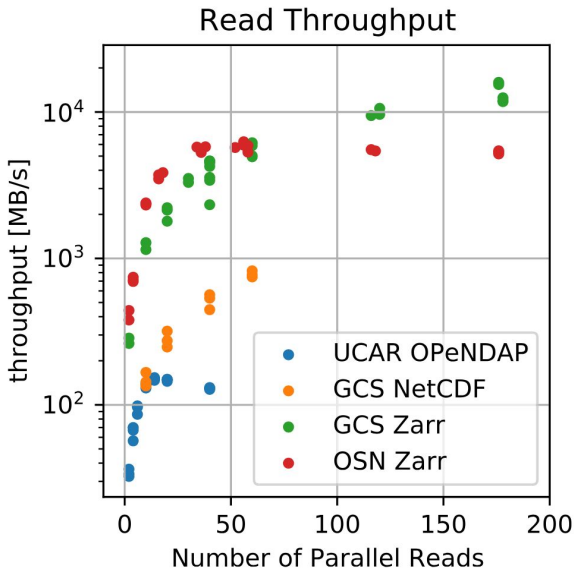
This also means that the field needs to think more about how data is organized

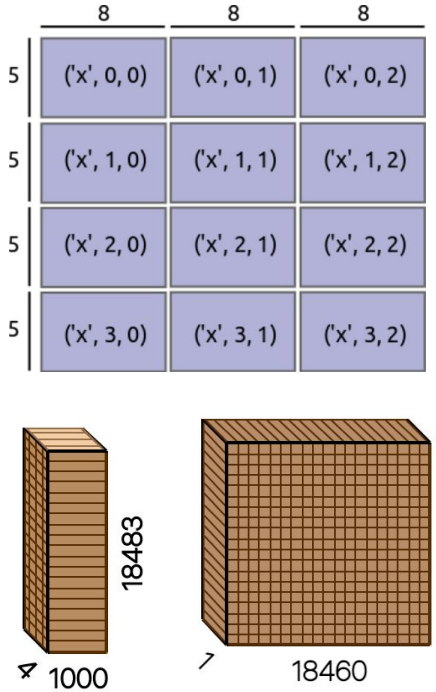
- Accessing data on the fly means it is transferred over the network (aka internet)
- Pro: you don't need to download everything
- Con: this can be slow if data is not laid out appropriately
- Solution:
  Analysis
  Ready
  Cloud
  Optimized Datasets (ARCO)



This also means that the field needs to think more about how data is organized

- What does it mean to make the “data layout” appropriate?
- Data can be “chunked” into regions where the data is easy to pull from storage.
- Chunks split the data up along each of the dimensions into rectangular regions
- The layout of these chunks has a huge impact on analysis efficiency (trading time for space)





What I think you should remember

- Data sparsity is true - especially with respect to any particular hypothesis
- However, data is abundantly being generated, synthesized, and shared
- Make yourself aware of multiple data sources, and be able to work in the big data regime
- Think about how to share your own data and code back into the scientific community