Analysis Ready Data Cubes: Perspectives for Earth System Research

Miguel D. Mahecha$^{1,2}$ & Fabian Gans$^3$

$^{1,2}$rsc4earth.de | twitter: @Rsc4Earth | @miguelmahechag

$^3$bgc-jena.mpg.de | twitter: @MPI_BGC | @meggarto
“Data Cubes” - a cluster of Pilots
Data cubes in the NFDI4Earth pilots

Statistical Learning on data cubes. Winkler et al.

Model evaluation in data cubes. Eyring et al.

EO4Glaciers data cube. Braun et al.

Socioeconomic data cubes. Kraemer et al.

Processing and Visualization. Unnithan et al.
Data cubes emerge everywhere...

https://r-spatial.github.io/stars/
https://eurodatacube.com/
http://www.rasdaman.org/

Earth in a Box
The Earth System Data Lab (ESDL) is a multivariate data set of essential Earth System cubes on a common grid and sharing a common data model.

https://www.earthsystemdatalab.net/
https://www.opendatacube.org/
Overarching aim: Empower big gridded data

Getting data cubes “Analysis Ready”:

- Avoiding complex data splits
- No further preprocessing
- Minimizing access barriers
- Enabling complex exploration
- Visualization

Fig top: Hannes Feilhauer

https://www.youtube.com/ →“ESDL Datacube”

https://www.ecmwf.int/en/about/media-centre/focus/2017/fact-sheet-ensemble-weather-forecasting
One example of “Analysis Ready Data Cubes”

The Earth System is multivariate, and coupled across sub-domains!

- Towards multivariate exploitations
- Dimension-agnostic implementation
- Cube with interactive computing environment
- Mapping arbitrary user defined functions
- Cloud readiness

https://www.earthsystemdatalab.net/

Arbitrarily complex workflows can operate on the cube

Intrinsic dimensionality of land-surface dynamics

- What is the redundancy among all the land-surface variables?
- What are the minimum number of orthogonal dimensions needed?

We need to prepare

Challenges

- Very high-resolution data sets (observations and models!)
- Heterogeneous sources
- Work across repositories
- Multiple data cube solutions
But we are not alone →

Interactive computing infrastructure for your community.

2i2c is a mission-driven non-profit that develops, deploys, customizes, and manages open source tools for interactive computing in research and education.

What’s a 2i2c Hub?

“2i2c Hub is a collection of open source tools that provide interactive computing environments in the cloud.”
Analysis Ready
Cloud Optimized
Data Cubes
Storing spatiotemporal datasets

- Typically in NetCDF or HDF5
- Metadata + data in a single file
- File can be arbitrarily large
- Simple subsetting
- Made for filesystems, random access through seek operations

Figure: Fabian Gans - unpublished
Storing spatiotemporal datasets

- Typically in NetCDF or HDF5
- Metadata + data in a single file
- File can be arbitrarily large
- Simple subsetting
- Made for filesystems, random access through seek operations

Figure: Fabian Gans - unpublished
Storing spatiotemporal datasets

- Typically in NetCDF or HDF5
- Metadata + data in a single file
- File can be arbitrarily large
- Simple subsetting
- Made for filesystems, random access through seek operations

Figure: Fabian Gans - unpublished
Storing spatiotemporal datasets

- Typically in NetCDF or HDF5
- Metadata + data in a single file
- File can be arbitrarily large
- Simple subsetting
- Made for filesystems, random access through seek operations

Figure: Fabian Gans - unpublished
Storing data in the cloud

- Different characteristics than filesystem-based
- Objects in a bucket instead of file hierarchy
- Large latencies
- High data throughput (limited by network bandwidth)
- Access to objects, no seek operations possible
- Highly scalable
Spatiotemporal datasets in the cloud

Figure: Fabian Gans - unpublished
Spatiotemporal datasets in the cloud

**Split by years?**

Bad because of:
- split metadata
- no way to quickly access metadata
- slow time series access

*Figure: Fabian Gans - unpublished*
Spatiotemporal datasets in the cloud

Use a cloud-optimized data format

Zarr - e.g. used in PANGEO
TileDB
Cloud-optimized GeoTiff
HDF5 Cloud

Figure: Fabian Gans - unpublished
Example “zarr” format - truly open

Array Metadata:
- data type
- chunk size
- endianness
- compressor
- filters
- fill value

User Metadata:
- like netcdf attributes
- units
- creator
- long name etc...

```
fgans@atama:/sgross_primary_productivity$ ls -a
.
  1.5.10 3.2.2 4.7.8 6.5.12 8.2.4
..       1.5.11 3.2.3 4.7.9 6.5.13 8.2.5
.zarray  1.5.12 3.2.4 5.0.0 6.5.14 8.2.6
.zattrs  1.5.13 3.2.5 5.0.1 6.5.15 8.2.7
0.0.0    1.5.14 3.2.6 5.0.10 6.5.2 8.2.8
0.0.1    1.5.15 3.2.7 5.0.11 6.5.3 8.2.9
0.0.10   1.5.2  3.2.8 5.0.12 6.5.4 8.3.0
0.0.11   1.5.3  3.2.9 5.0.13 6.5.5 8.3.1
0.0.12   1.5.4  3.3.0 5.0.14 6.5.6 8.3.10
0.0.13   1.5.5  3.3.1 5.0.15 6.5.7 8.3.11
0.0.14   1.5.6  3.3.10 5.0.2  6.5.8 8.3.12
0.0.15   1.5.7  3.3.11 5.0.3  6.5.9 8.3.13
0.0.2    1.5.8  3.3.12 5.0.4  6.6.0 8.3.14
0.0.3    1.5.9  3.3.13 5.0.5  6.6.1 8.3.15
```
User API - simple but powerful

```python
using ESDL, AWSCore, Zarr, Statistics, MultivariateStats, ESDLPlots

[ Info: Precompiling ESDLPlots [d555b242-3f29-57aa-84ea-3df92a135dfd]
 @ Base loading.jl:1278

aws = aws config(creds=nothing, region="eu-de", service name="obs", service host="otc.t-systems.com")
store = S3Store("obs-esdc-v2.0.0", "esdc-8d-0.25deg-184x90x90-2.0.0.zarr", 2, aws)
zarr_group = zopen(store, consolidated = true)
ds = open_dataset(zarr_group)

YAXArray Dataset
Dimensions:
  lat          Axis with 720 Elements from 89.875 to -89.875
  lon          Axis with 1440 Elements from -179.875 to 179.875
  time         Axis with 1782 Elements from 1980-01-05T00:00:00 to 2016-12-30T00:00:00
Variables: soil moisture xco2 leaf area index sensible heat flt c totcol msr stemp free lrt c lrt c potential evaporation evaporation root moiture land surface temperature black sky albedo avhrr precipitation free flt c open water evaporation lrt p srex mask latent energy max air temperture 2m xch4 cth psurf aerosol optical thickness 550 aerosol optical thickness 870 ctt air temperature 2m msr flt free msr lrt evaporativ e stress precipitation era5 aerosol optical thickness 670 snow water equivalent terrestrial ecosystem respiration black sky albedo analysed ss t mask white sky albedo aerosol optical thickness 1600 totcol assim fractional snow cover chlor a gross primary productivity country mask cer free fat c bare soil evaporation flt p par net radiation cot ozone par diff transpiration white sky albedo avhrr totcol free cee surface moisture fat p msr lrt sea ice fraction water vapour interception loss free msr flt c emissions cph ctp min air temperature 2m cfc water mask lwp burt area fat c fapar tip net ecosystem exchange iwp snow sublimation Rg

vars = ["evaporative stress",
          "latent energy",
          "black sky albedo avhrr",
          "fapar tip",
          "root moisture",
          ""]
```
User API - simple but powerful

```plaintext
function sufficient_dimensions(xin::AbstractArray, expl_var::Float64 = 0.95)
    any(ismissing, xin) && return NaN
    npoint, nvar = size(xin)
    means = mean(xin, dims = 1)
    stds = std(xin, dims = 1)
    xin = broadcast((y, m, s) -> s > 0.0 ? (y - m) / s : one(y), xin, means, stds)
    pca = fit(PCA, xin, pratio = 0.999, method = :svd)
    return findfirst(cumsum(principalvars(pca)) / tprincipalvar(pca) .> expl_var)
end
```

```plaintext
cube_int_dim = mapslices(sufficient_dimensions, cube_fill, dims = ("Time", "Variable"))
```

```plaintext
plotMAP(cube_int_dim)
```

[36]:
QUESTIONS?
Spatiotemporal datasets in the cloud

Optimize chunks according to access pattern

Figure: Fabian Gans - unpublished
Spatiotemporal datasets in the cloud

Optimize chunks according to access pattern

Time series

Figure: Fabian Gans - unpublished
Spatiotemporal datasets in the cloud

Optimize chunks according to access pattern

Maps

Figure: Fabian Gans - unpublished
Efficient storage ....

Fabian Gans (in prep) Efficient data cube storage of unlimited size filed

Figure: Fabian Gans - unpublished
Regional cubes, specific cubes, all data in one concept

Figure: Miguel Mahecha - unpublished

Analysis: Felix Behrend - unpublished
Very simple “two-line” operations

\[ f_{\{lat, lon, time\}} : \mathcal{C}(\{lat, lon, time\}) \to \mathcal{C}(\{\}) \].

Mahecha, Gans et al. (2020)
Earth System Dynamics, 11, 201-234.
Very complicated workflows

https://www.earthsystemdatalab.net/

Mahecha, Gans et al. (2020)
Earth System Dynamics, 11, 201-234.
Potential for more complicated parameter estimation

Potential for more complicated parameter estimation


Conclusions

● New in-situ and satellite remote sensing products refine our understanding of Earth system processes

● Flood of downstream data processes require new data analytic approaches

● We are at the edge to do research in digital-twin Earths with unprecedented opportunities - but without solving fundamental issues (physical data consistency, resolutions operationally at the level of true processes understanding etc....) → New ideas wanted