The Emerging Earth System Data Cube

Idea and first applications

Guido Kraemer[1] Miguel D. Mahecha[1] Fabian Gans[1] Markus Reichstein[1]
et al.

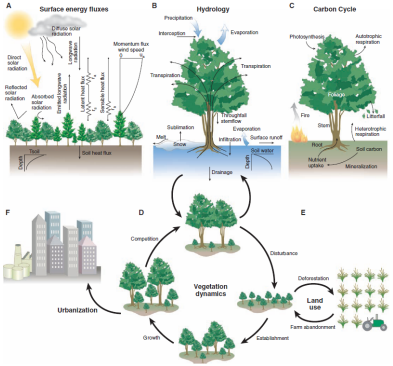
[1]Max Planck Institute for Biogeochemistry, Germany

November 22, 2016

Our scientific context

Understanding Atmosphere-Biosphere interactions considering

- *Climate change*
- *Extreme anomalies*
- *Land use change*
- . . .



Bonan, G. (2008)

Implications of a “data rich world” for ecologists?

Increasing

- Data Availability
  - in-situ
  - spatiotemporal
- Resolution
  - spatial
  - temporal
  - spectral

New Chances

- describing global environmental features
- addressing classical questions more systematically across ecosystems

This may require
new analytic techniques
and
methodological tools.

Implications of a “data rich world” for ecologists?

Increasing

- Data Availability
  - in-situ
  - spatiotemporal
- Resolution
  - spatial
  - temporal
  - spectral

New Chances

- describing global environmental features
- addressing classical questions more systematically across ecosystems

This may require
new analytic techniques
and
methodological tools.

Implications of a “data rich world” for ecologists?

Increasing

- Data Availability
  - in-situ
  - spatiotemporal
- Resolution
  - spatial
  - temporal
  - spectral

New Chances

- describing global environmental features
- addressing classical questions more systematically across ecosystems

This may require
new analytic techniques
and
methodological tools.

Implications of a “data rich world” for ecologists?

Increasing

- Data Availability
  - in-situ
  - spatiotemporal
- Resolution
  - spatial
  - temporal
  - spectral

New Chances

- describing global environmental features
- addressing classical questions more systematically across ecosystems

This may require
new analytic techniques
and
methodological tools.

Obstacles to explore the data potential

o_1 : Archive dispersal
o_2 : Access difficulties
o_3 : Inconsistent formatting
o_4 : Inconsistent naming
o_5 : Inconsistent resolution
o_6 : Inconsistent flags
o_7 : Data use policies

o_8 : Workflow documentation
o_9 : Workflow sharing
o_{10} : Workflows reproduction
o_{11} : Versioning
o_{12} : “living data”

$\vdots$

o_N : …

$$\text{Scientific Discovery} \propto \frac{1}{\sum_i^N o_i}$$

We aim for:

$$\arg\max_{o_i} (\text{Scientific Discovery}) = \arg\min_{o_i} \sum_{i=1}^{N} o_i$$

Obstacles to explore the data potential

o_1 : Archive dispersal
o_2 : Access difficulties
o_3 : Inconsistent formatting
o_4 : Inconsistent naming
o_5 : Inconsistent resolution
o_6 : Inconsistent flags
o_7 : Data use policies

o_8 : Workflow documentation
o_9 : Workflow sharing
o_{10} : Workflows reproduction
o_{11} : Versioning
o_{12} : “living data”

$\vdots$

o_N : …

$$\text{Scientific Discovery} \propto \frac{1}{\sum_i^N o_i}$$

We aim for:

$$\arg\max_{o_i} (\text{Scientific Discovery}) = \arg\min_{o_i} \sum_{i=1}^{N} o_i$$

Obstacles to explore the data potential

- o_1 : Archive dispersal
- o_2 : Access difficulties
- o_3 : Inconsistent formatting
- o_4 : Inconsistent naming
- o_5 : Inconsistent resolution
- o_6 : Inconsistent flags
- o_7 : Data use policies
- o_8 : Workflow documentation
- o_9 : Workflow sharing
- o_{10} : Workflows reproduction
- o_{11} : Versioning
- o_{12} : “living data”
- $\vdots$
- o_N : …

$$\text{Scientific Discovery} \propto \frac{1}{\sum_i^N o_i}$$

We aim for:

$$\arg\max_{o_i} (\text{Scientific Discovery}) = \arg\min_{o_i} \sum_{i=1}^{N} o_i$$

Towards a Earth system data cube

Our work

- Gather all kind of Earth System relevant EOs
- One cube $\mathbf{X} = \{x_{u,v,t,m}\}$ with:
  - u: lat
  - v: lon
  - t: time
  - m: variables

Towards a Earth system data cube

Our work

- Gather all kind of Earth System relevant EOs
- One cube $\mathbf{X} = \{x_{u,v,t,m}\}$ with:

u: lat
v: lon
t: time
m: variables

Towards a Earth system data cube

Our work

- Gather all kind of Earth System relevant EOs
- One cube $\mathbf{X} = \{x_{u,v,t,m}\}$ with:

u: lat
v: lon
t: time
m: variables

Towards a Earth system data cube

Our work

- Gather all kind of Earth System relevant EOs
- One cube $\mathbf{X} = \{x_{u,v,t,m}\}$ with:

u: lat
v: lon
t: time
m: variables

Towards a Earth system data cube

Our work

- Gather all kind of Earth System relevant EOs
- One cube $\mathbf{X} = \{x_{u,v,t,m}\}$ with:

u: lat
v: lon
t: time
m: variables

Towards a Earth system data cube

Our work

- Gather all kind of Earth System relevant EOs
- One cube $\mathbf{X} = \{x_{u,v,t,m}\}$ with:

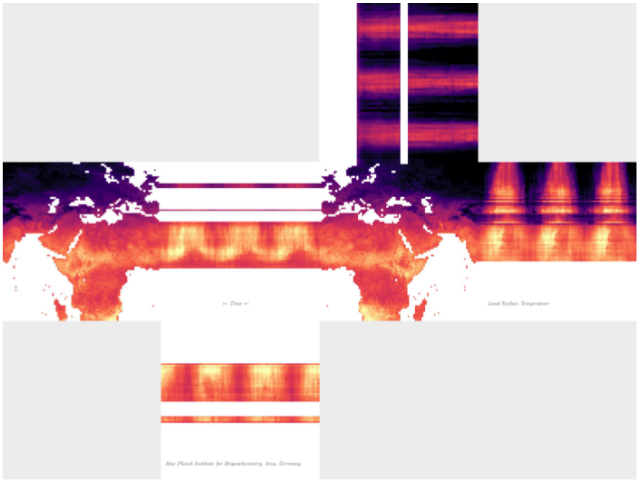
u: lat
v: lon
t: time
m: variables

Towards a Earth system data cube

Our work

- Gather all kind of Earth System relevant EOs
- One cube $\mathbf{X} = \{x_{u,v,t,m}\}$ with:

u: lat
v: lon
t: time
m: variables

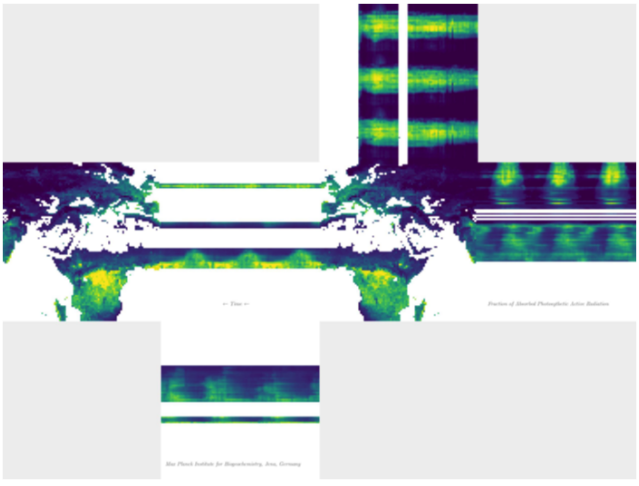


Towards a Earth system data cube

Our work

- Gather all kind of Earth System relevant EOs
- One cube $\mathbf{X} = \{x_{u,v,t,m}\}$ with:

u: lat
v: lon
t: time
m: variables

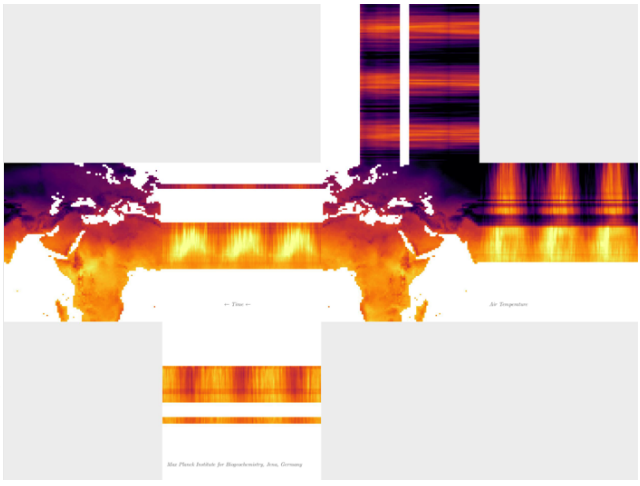


Towards a Earth system data cube

Our work

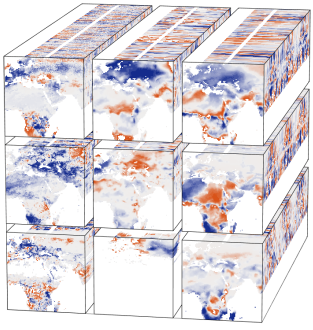
- Gather all kind of Earth System relevant EOs
- One cube $\mathbf{X} = \{x_{u,v,t,m}\}$ with:

u: lat
v: lon
t: time
m: variables



The 3 columns:

1. Earth Sys. Data Cube

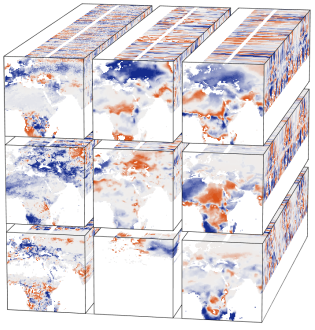


2. Data Analytic Toolkit

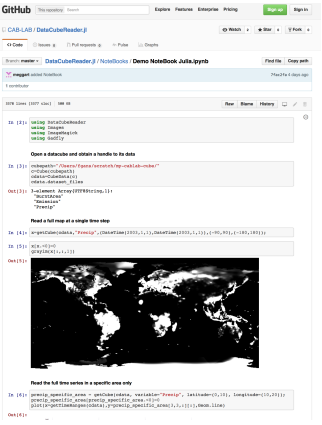
3. Scientific Analyses

The 3 columns:

1. Earth Sys. Data Cube



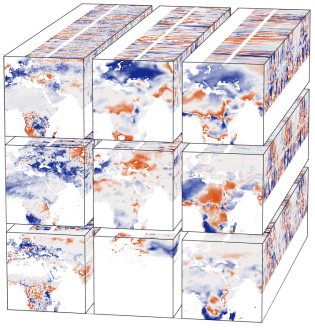
2. Data Analytic Toolkit



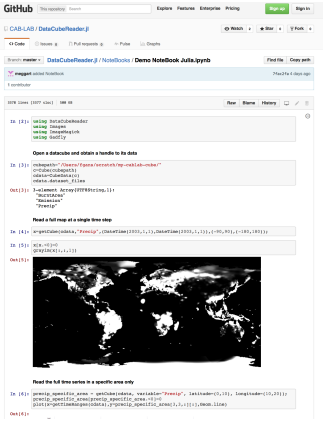
3. Scientific Analyses

The 3 columns:

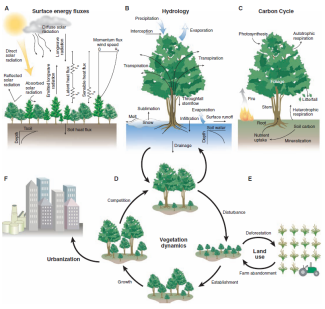
1. Earth Sys. Data Cube



2. Data Analytic Toolkit



3. Scientific Analyses



Bonan, 2008

The Earth System Data Cube

Atmosphere

- Radiation
- Aerosols
- Temperature
- Precipitation
- Vapor
- Ozone
- Evaporation
- Albedo
- Snow

Biosphere

- Burned Area
- Evapotranspiration
- Gross Primary Production
- Ecosystem Respiration
- Land Surface Temperature/Moisture
- Latent Energy
- Soil Moisture
- Snow Water Equivalents

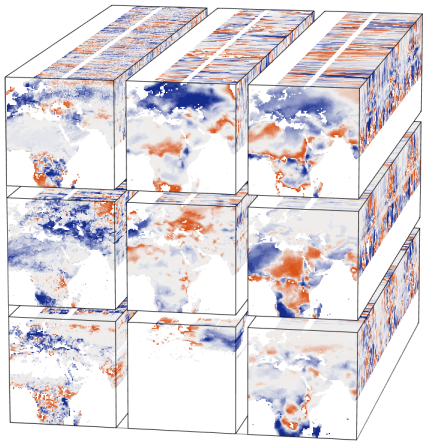
Anthroposphere

- World Development Indicators
- Other national statistics
- Night time light emission

Suggestions most welcome!

The Earth System Data Cube

- Global extent
- Different Resolutions
  - 0.083°, 533 GB
  - 0.25°, 60 GB
  - …
- Convenience aggregations e.g. to national levels
- Consistent temporal sampling (8-daily for 2001-2011)



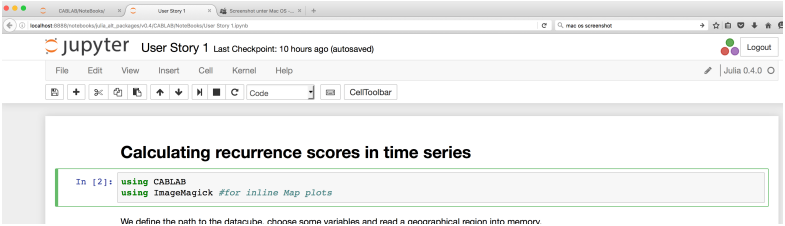
Data Access & Analytic Toolkit

Principles:

- Access
- Visualization
- Processing
- Documentation

Implementation:

- Jupyter multi-language notebooks (Julia, R, Python)
- Publishing/sharing workflows
- Local or remote access



Data Access & Analytic Toolkit

Convenience functions condense processing steps:

```
In [4]: @time cube_filled     = map(gapFillMSC,cdata,46,max_cache=1e7);
        18.017533 seconds (14.40 M allocations: 877.965 MB, 1.70% gc time)

In [5]: @time cube_anomalies  = map(removeMSC,cube_filled,46,max_cache=1e7);
        3.923924 seconds (5.05 M allocations: 198.965 MB, 3.20% gc time)

In [6]: @time cube_normalized = map(normalize,cube_anomalies,max_cache=1e7);
        3.708907 seconds (2.90 M allocations: 154.698 MB, 1.22% gc time)

In [7]: @time scores          = map(recurrences,cube_normalized,5.0,5,zeros(Float32,506,506),max_cache=1e7);
        61.147698 seconds (47.40 M allocations: 1.997 GB, 0.92% gc time)
```

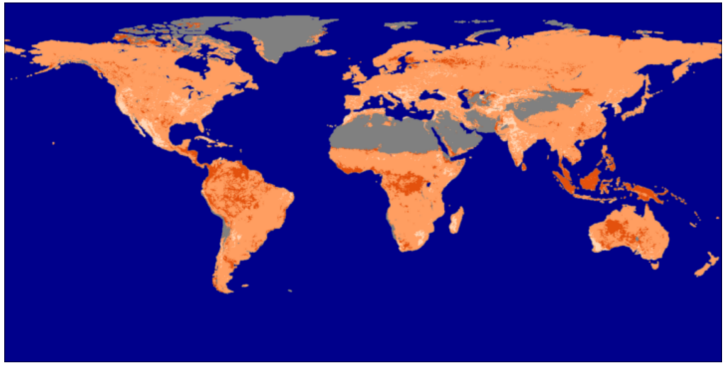
- Result of each processing step stored on disk as temporary cubes
- Workflows can be shared, guarantee transparency and reproducibility
- Parallel processing
- Out-of-memory data

Data Access & Analytic Toolkit

We can scrutinize the intrinsic dimensionality of the time series

```
In [16]: plotMAP(qualitypca,dmin=1.f0,dmax=5.f0)
```

Out[16]:

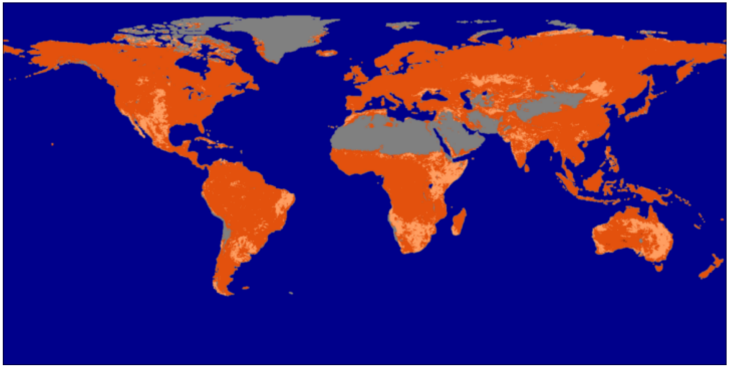


Data Access & Analytic Toolkit

We can scrutinize the intrinsic dimensionality of the time series

```
In [17]: plotMAP(qualitypcaanom,dmin=1.f0,dmax=5.f0)
```

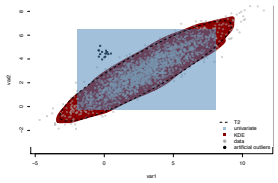
Out[17]:

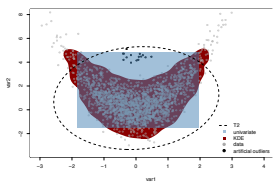


Avenues of scientific exploitation

Detecting multidimensional anomalies

1. Generically
2. Invariant to periodic patterns
3. In space and time
4. . . .



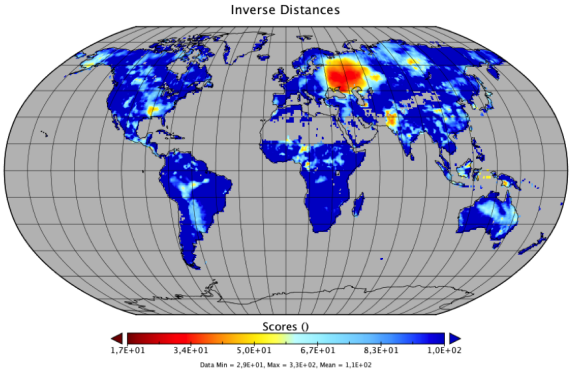


PhD thesis by Milan Flach

Avenues of scientific exploitation

Detecting multidimensional anomalies

1. Generically
2. Invariant to periodic patterns
3. In space and time
4. . . .

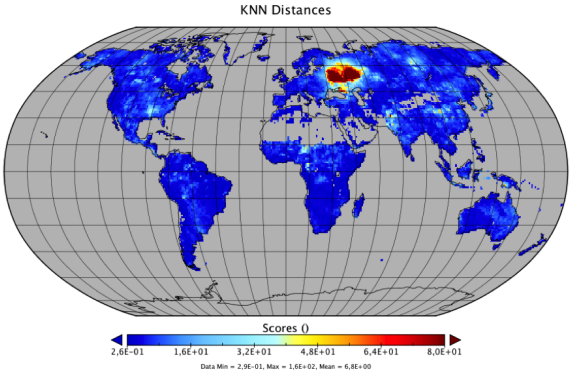


PhD thesis by Milan Flach

Avenues of scientific exploitation

Detecting multidimensional anomalies

1. Generically
2. Invariant to periodic patterns
3. In space and time
4. . . .

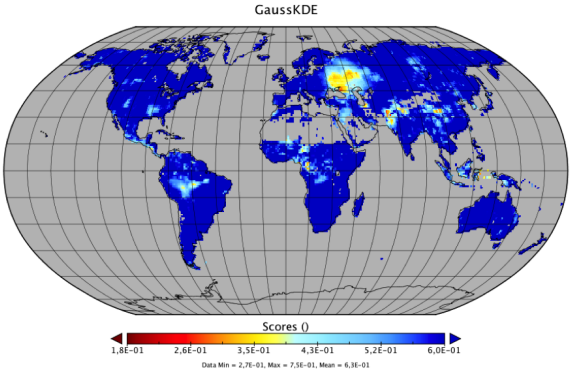


PhD thesis by Milan Flach

Avenues of scientific exploitation

Detecting multidimensional anomalies

1. Generically
2. Invariant to periodic patterns
3. In space and time
4. . . .

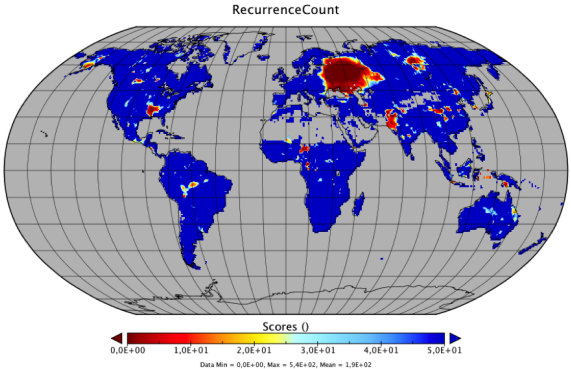


PhD thesis by Milan Flach

Avenues of scientific exploitation

Detecting multidimensional anomalies

1. Generically
2. Invariant to periodic patterns
3. In space and time
4. . . .

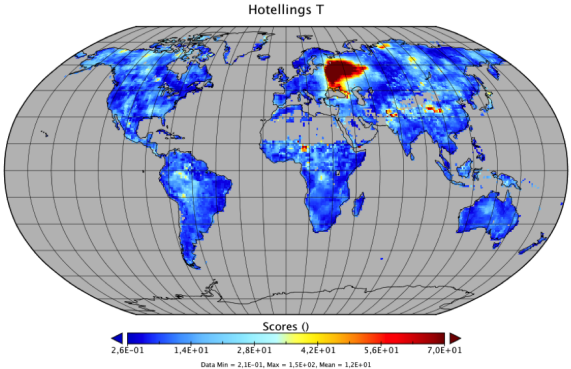


PhD thesis by Milan Flach

Avenues of scientific exploitation

Detecting multidimensional anomalies

1. Generically
2. Invariant to periodic patterns
3. In space and time
4. . . .

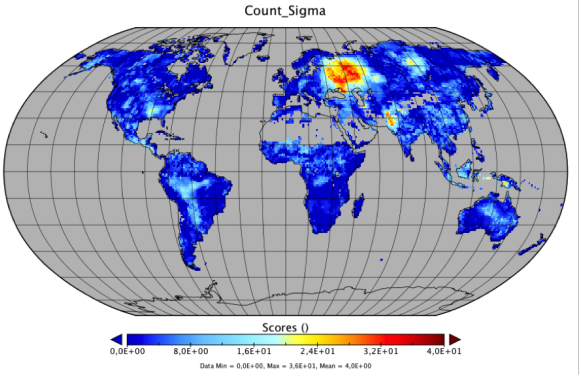


PhD thesis by Milan Flach

Avenues of scientific exploitation

Detecting multidimensional anomalies

1. Generically
2. Invariant to periodic patterns
3. In space and time
4. . . .



PhD thesis by Milan Flach

The emerging Earth system data cube

Development

- Open Source: `https://github.com/CAB-LAB`
- DataCube: current version is 0.2.0
- CABLAB.jl: current version is 0.2

Science

- Minimizing obstacles to scientists.
- Systematic descriptions of global patterns.
- Spatiotemporal/high-dimensional scientific discovery in Earth system data.
- We invite you to discuss with us the implementation of stories - simple ones, complex ones, and we would be of course happy to find new collaborations here.

The emerging Earth system data cube

Development

- Open Source: `https://github.com/CAB-LAB`
- DataCube: current version is 0.2.0
- CABLAB.jl: current version is 0.2

Science

- Minimizing obstacles to scientists.
- Systematic descriptions of global patterns.
- Spatiotemporal/high-dimensional scientific discovery in Earth system data.
- We invite you to discuss with us the implementation of stories - simple ones, complex ones, and we would be of course happy to find new collaborations here.

The emerging Earth system data cube

Development

- Open Source: `https://github.com/CAB-LAB`
- DataCube: current version is 0.2.0
- CABLAB.jl: current version is 0.2

Science

- Minimizing obstacles to scientists.
- Systematic descriptions of global patterns.
- Spatiotemporal/high-dimensional scientific discovery in Earth system data.
- We invite you to discuss with us the implementation of stories - simple ones, complex ones, and we would be of course happy to find new collaborations here.

The emerging Earth system data cube

Development

- Open Source: `https://github.com/CAB-LAB`
- DataCube: current version is 0.2.0
- CABLAB.jl: current version is 0.2

Science

- Minimizing obstacles to scientists.
- Systematic descriptions of global patterns.
- Spatiotemporal/high-dimensional scientific discovery in Earth system data.
- We invite you to discuss with us the implementation of stories - simple ones, complex ones, and we would be of course happy to find new collaborations here.

The emerging Earth system data cube

Development

- Open Source: `https://github.com/CAB-LAB`
- DataCube: current version is 0.2.0
- CABLAB.jl: current version is 0.2

Science

- Minimizing obstacles to scientists.
- Systematic descriptions of global patterns.
- Spatiotemporal/high-dimensional scientific discovery in Earth system data.
- We invite you to discuss with us the implementation of stories - simple ones, complex ones, and we would be of course happy to find new collaborations here.

http://earthsystemdatacube.net