

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

# The Emerging Earth System Data Cube

## Idea and first applications

Guido Kraemer<sup>1</sup>   Miguel D. Mahecha<sup>1</sup>   Fabian Gans<sup>1</sup>   Markus Reichstein<sup>1</sup>  
et al.

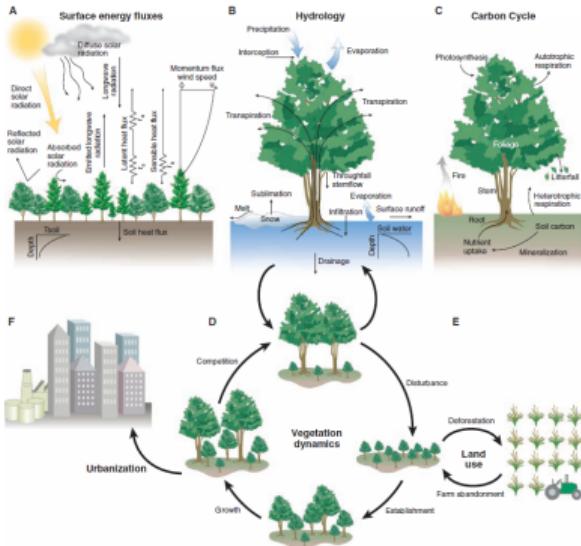
<sup>1</sup>Max Planck Institute for Biogeochemistry, Germany

November 22, 2016

# Our scientific context

## Understanding Atmosphere-Biosphere interactions considering

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



Bonan, G. (2008)

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

# Implications of a “data rich world” for ecologists?

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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?

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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?

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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?

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

$o_1$  : 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”

⋮

$o_N$  : ...

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

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”  
⋮  
 $o_N$  : ...

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

We aim for:

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

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

# Obstacles to explore the data potential

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

$o_1$  : 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”

⋮

$o_N$  : ...

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

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

Introduction  
Opportunities  
Challenges  
**Earth System Data Cube**  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Our work

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

• lat

• lon

• time

• m

# Towards a Earth system data cube

Introduction  
Opportunities  
Challenges  
**Earth System Data Cube**  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

Introduction  
Opportunities  
Challenges  
**Earth System Data Cube**  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

Introduction  
Opportunities  
Challenges  
**Earth System Data Cube**  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

Introduction  
Opportunities  
Challenges  
**Earth System Data Cube**  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

Introduction  
Opportunities  
Challenges  
**Earth System Data Cube**  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

Introduction  
Opportunities  
Challenges  
**Earth System Data Cube**  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Our work

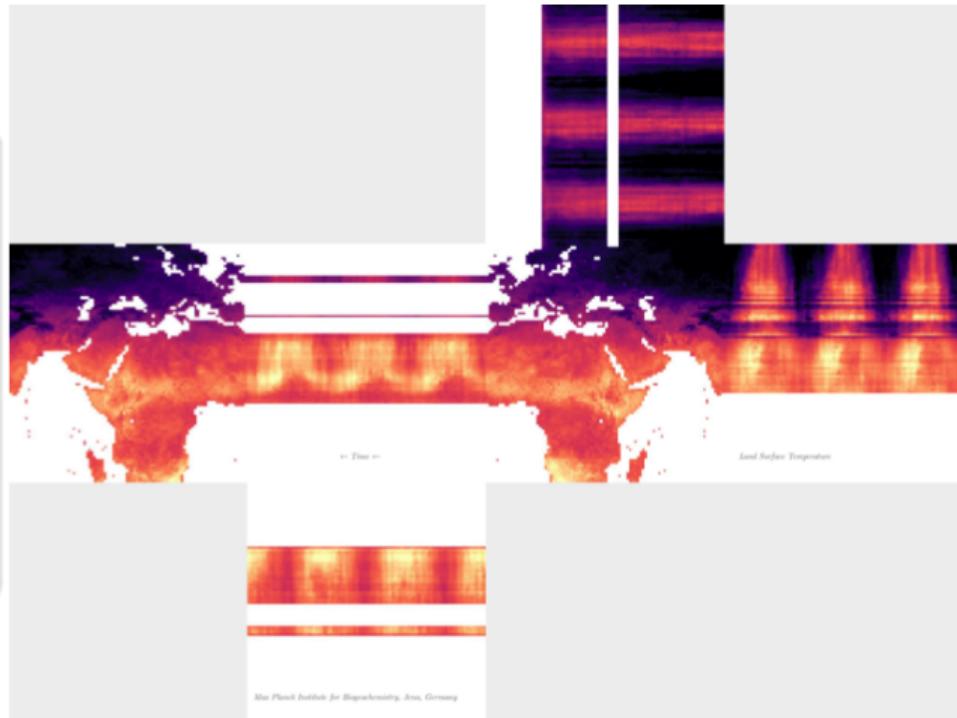
- ▶ Gather all kind of Earth System relevant EO's
- ▶ One cube  $\mathbf{X} = \{x_{u,v,t,m}\}$  with:

*u:* lat

*v:* lon

*t:* time

*m:* variables



# Towards a Earth system data cube

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Our work

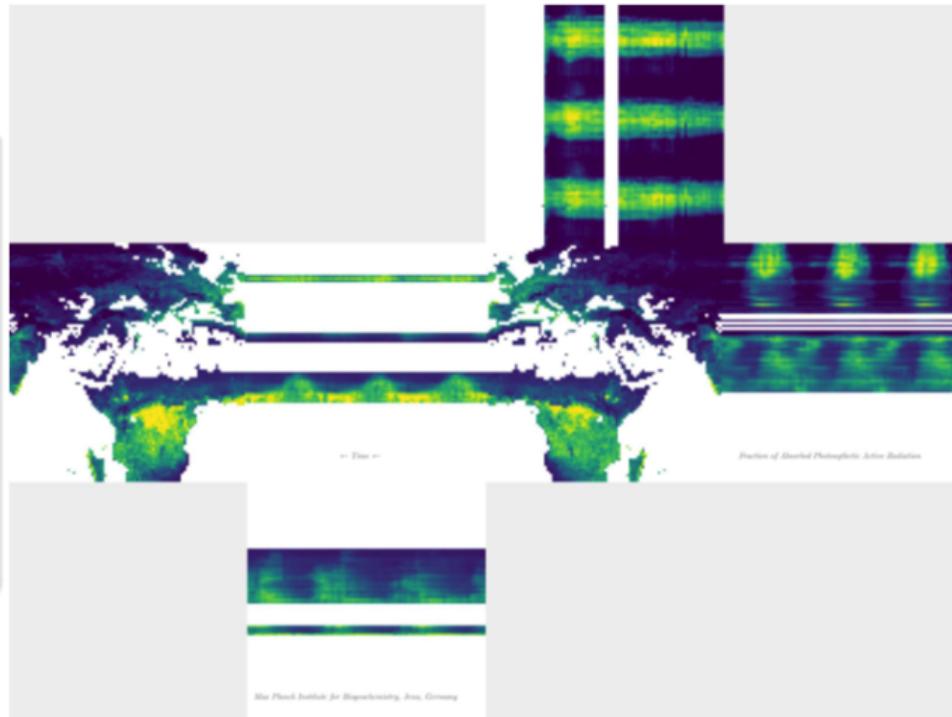
- ▶ Gather all kind of Earth System relevant EO
- ▶ One cube  $\mathbf{X} = \{x_{u,v,t,m}\}$  with:

$u$ : lat

$v$ : lon

$t$ : time

$m$ : variables



# Towards a Earth system data cube

Introduction  
Opportunities  
Challenges  
**Earth System Data Cube**  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Our work

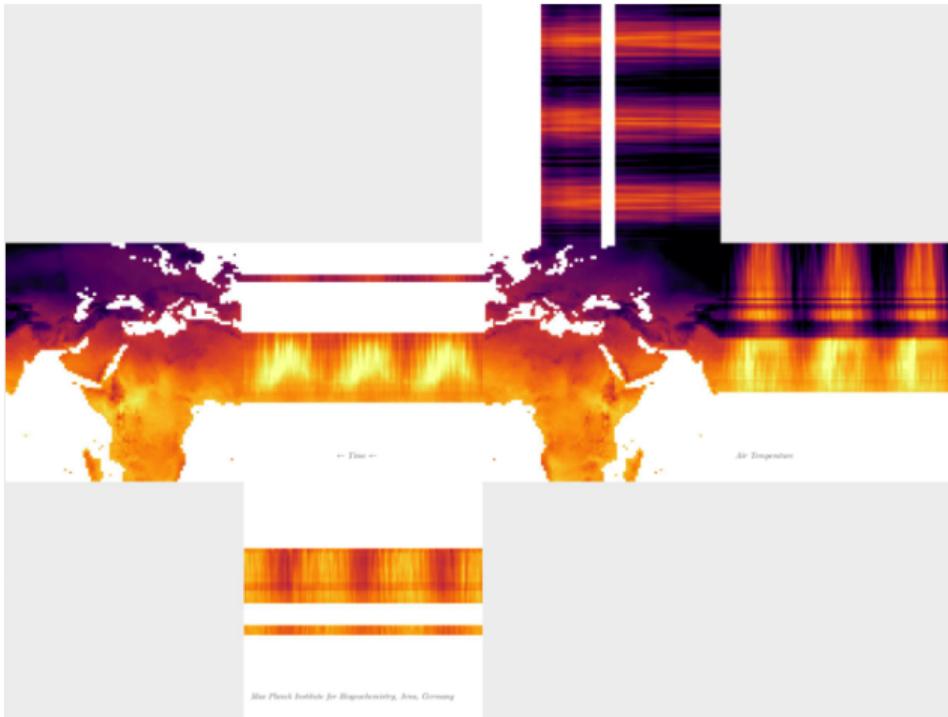
- ▶ Gather all kind of Earth System relevant EO's
- ▶ One cube  $\mathbf{X} = \{x_{u,v,t,m}\}$  with:

$u$ : lat

$v$ : lon

$t$ : time

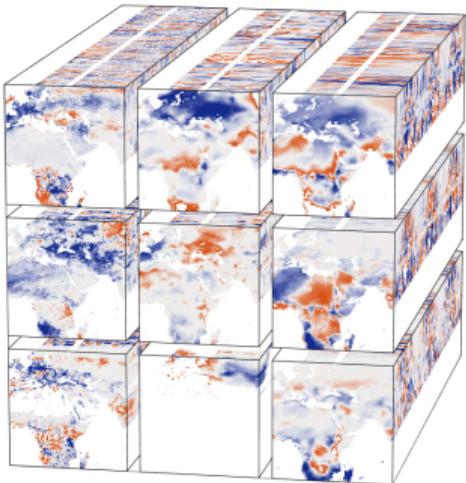
$m$ : variables



# The 3 columns:

---

1. Earth Sys. Data Cube



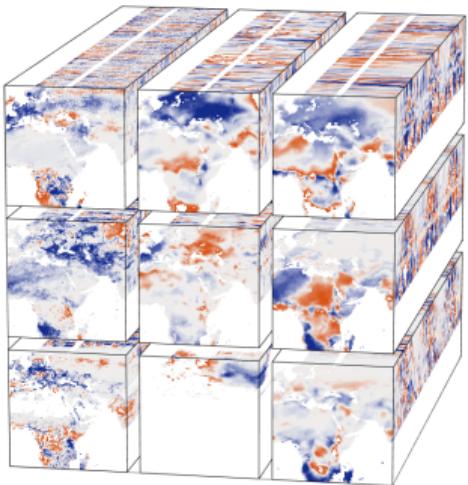
2. Data Analytic Toolkit

3. Scientific Analyses

- [Introduction](#)
- [Opportunities](#)
- [Challenges](#)
- [Earth System Data Cube](#)
- [Data Analytic Toolkit](#)
- [Scientific Perspectives](#)
- [Outlook](#)

# The 3 columns:

## 1. Earth Sys. Data Cube



## 2. Data Analytic Toolkit

GitHub [This repository](#) Search Explore Features Enterprise Pricing Sign up Sign in

CAB-LAB / DataCubeReader.jl | Notebooks | Demo Notebook Jupyter.ipynb

Branch: master | [DataCubeReader.jl](#) | [Notebooks](#) | [Demo Notebook Jupyter.ipynb](#)

Find file Copy path

2 commits

Diff times (1977 files) 58M 00

Row Item History

In [1]:

```
using DataCubeReader
using Images
using DataFrames
using Geoviews
```

Open a datacube and obtain a handle to its data

In [3]:

```
infilepath = "/Users/Epica/verarbeitung/cab-lab-moco/"
datacube = DataCubeReader.read(infilepath)
datacube.datafiles
```

Out[3]:

```
3-element Array{String,1}:
 "precip"
 "deauwaa"
 "precip"
```

Read a full map at a single time step

In [4]:

```
precip = DataCubeReader.getcube(datacube, "precip", (DateTime(2003,1,1), DateTime(2003,1,1)), (-90,90), (-180,180))
```

In [5]:

```
x,y = precip
```

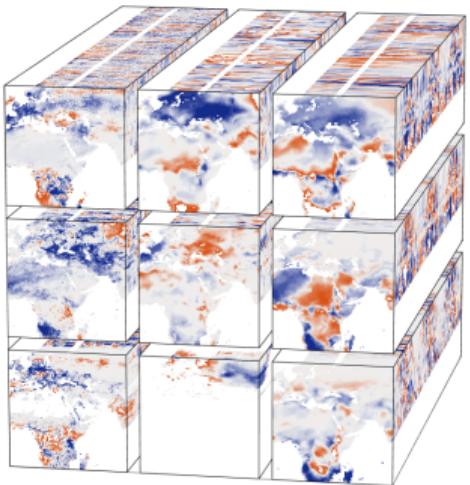
Out[5]:

## 3. Scientific Analyses

- [Introduction](#)
- [Opportunities](#)
- [Challenges](#)
- [Earth System Data Cube](#)
- [Data Analytic Toolkit](#)
- [Scientific Perspectives](#)
- [Outlook](#)

# The 3 columns:

## 1. Earth Sys. Data Cube



## 2. Data Analytic Toolkit

GitHub [This repository](#) Search Explore Features Enterprise Pricing Sign up Sign in

CAB-LAB / DataCubeReader.jl

Branch: master · DataCubeReader.jl / Notebooks / Demo Notebook Jupyter.ipynb

Find file Copy path

1 contributor

Diff Times (1077 files) 58K 00

In [1]:

```
using DataCubeReader
using Images
using DataFrames
using Geckobird
```

Open a datacube and obtain a handle to its data

In [2]:

```
datacube = "https://openstorage.firebaseio/app/cab-lab-ecmwf"
data = DataCube(datacube)
data.datacube_files
```

Out[2]:

```
3-element Array{HTTPResponse,1}:
 [1] "https://openstorage.firebaseio/app/cab-lab-ecmwf"
 [2] "https://openstorage.firebaseio/app/cab-lab-ecmwf"
 [3] "https://openstorage.firebaseio/app/cab-lab-ecmwf"
```

Read a full map in a single time step

In [3]:

```
precip = getcube(data, "Precip", (DateTime(2003,1,1), DateTime(2003,1,1)), (-10, 90), (-180, 180))
```

In [4]:

```
x,y = 1:10
y,x = 1:10
```

Out[4]:

Read the full time series in a specific area only

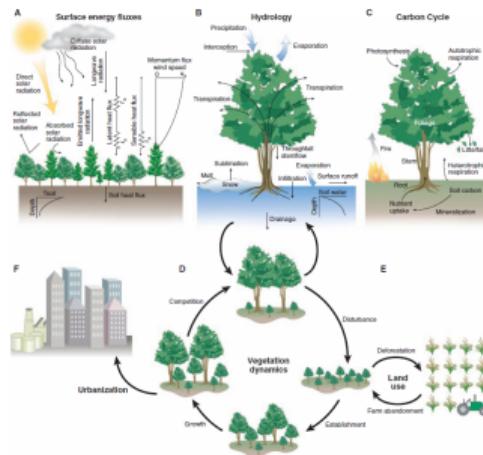
In [5]:

```
precip_specific_area = getcube(data, "Precip", latlon=(0,10), longitude=(-10,20))
precip_specific_area.precip_specific_area.via = "geckobird"
plot(precip_specific_area), y=precip_specific_area.1, c=:blue, line
```

Out[5]:

```
-
```

## 3. Scientific Analyses



Bonan, 2008

- Introduction
- Opportunities
- Challenges
- Earth System Data Cube
- Data Analytic Toolkit
- Scientific Perspectives
- Outlook

# The Earth System Data Cube

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

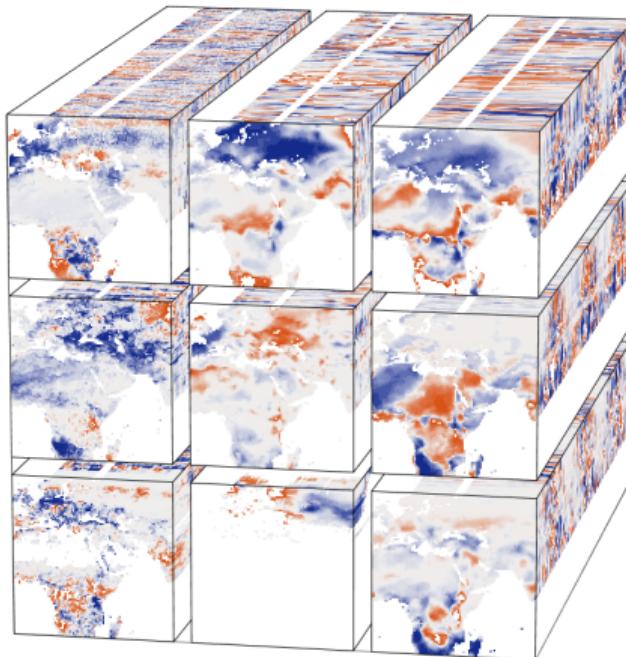
## Anthroposphere

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

*Suggestions most welcome!*

# The Earth System Data Cube

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



[Introduction](#)  
[Opportunities](#)  
[Challenges](#)  
[Earth System Data Cube](#)  
[Data Analytic Toolkit](#)  
[Scientific Perspectives](#)  
[Outlook](#)

# Data Access & Analytic Toolkit

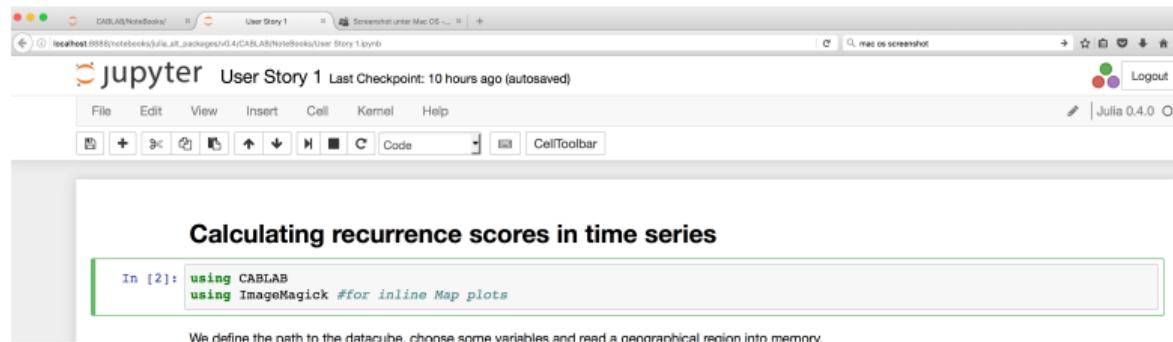
Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Principles:

- ▶ Access
- ▶ Visualization
- ▶ Processing
- ▶ Documentation

## Implementation:

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



The screenshot shows a Jupyter notebook interface running on a Mac OS X system. The title bar indicates the browser is showing a local host notebook at port 8888. The main window displays a single code cell with the following content:

```
In [2]: using CABLELAB
using ImageMagick #for inline Map plots
```

Below the code cell, a note states: "We define the path to the datacube, choose some variables and read a nonorthogonal region into memory."

# Data Access & Analytic Toolkit

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

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

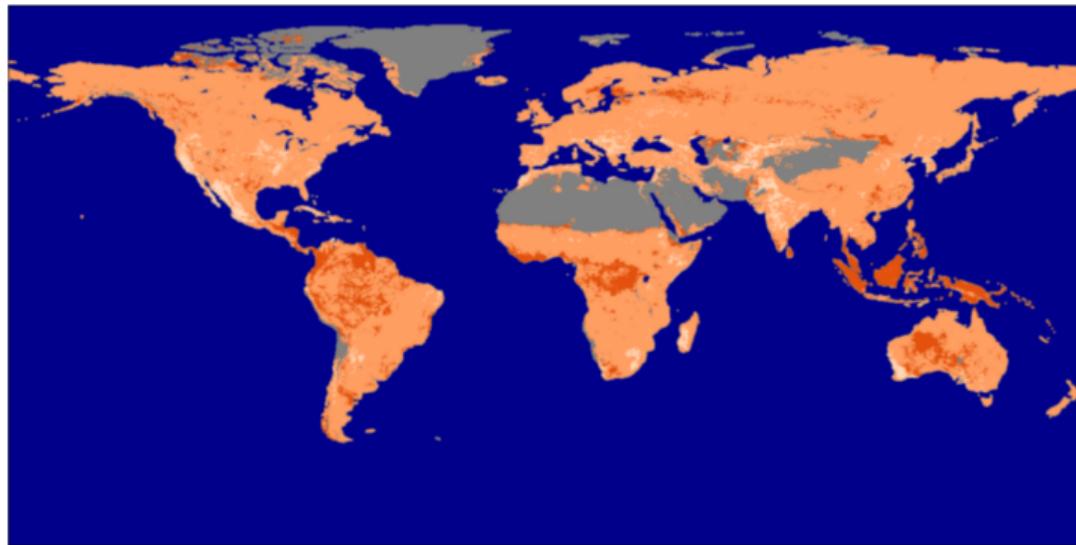
# Data Access & Analytic Toolkit

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

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

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

We can scrutinize the intrinsic dimensionality of the time series

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

Out[17]:

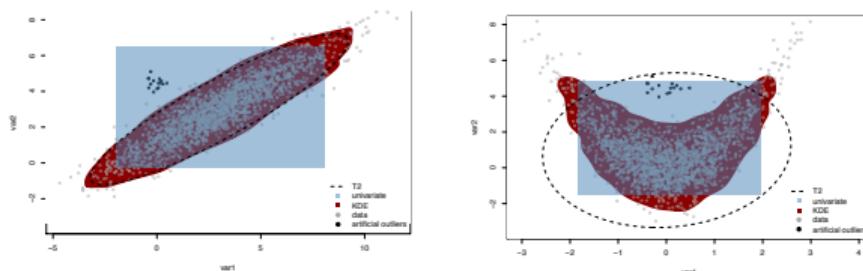


# Avenues of scientific exploitation

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Detecting multidimensional anomalies

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



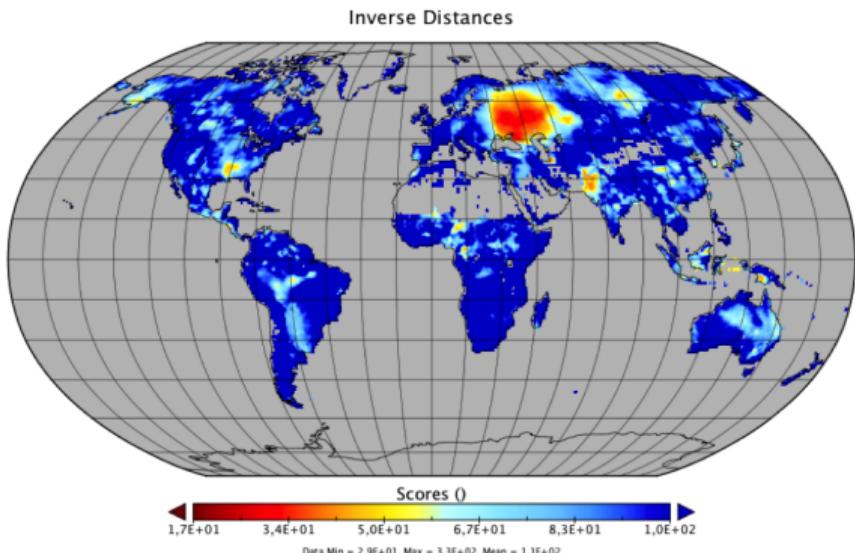
PhD thesis by Milan Flach

# Avenues of scientific exploitation

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Detecting multidimensional anomalies

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



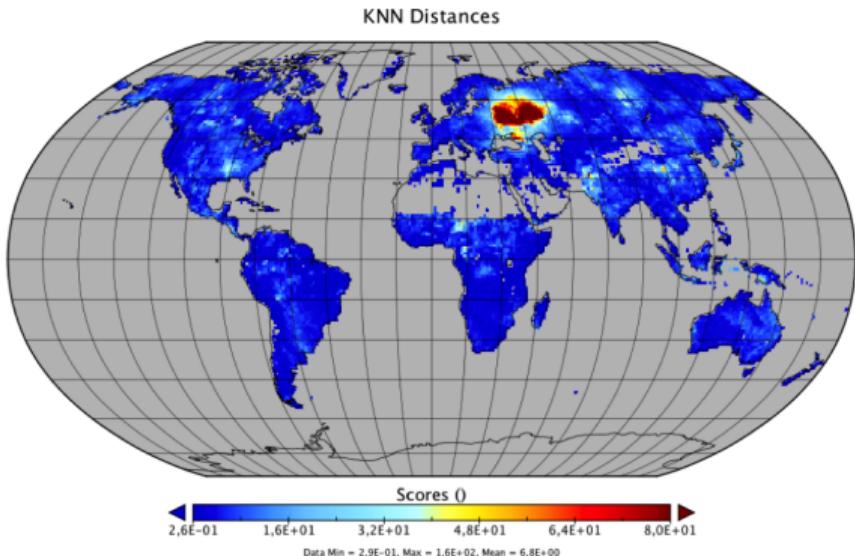
PhD thesis by Milan Flach

# Avenues of scientific exploitation

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Detecting multidimensional anomalies

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

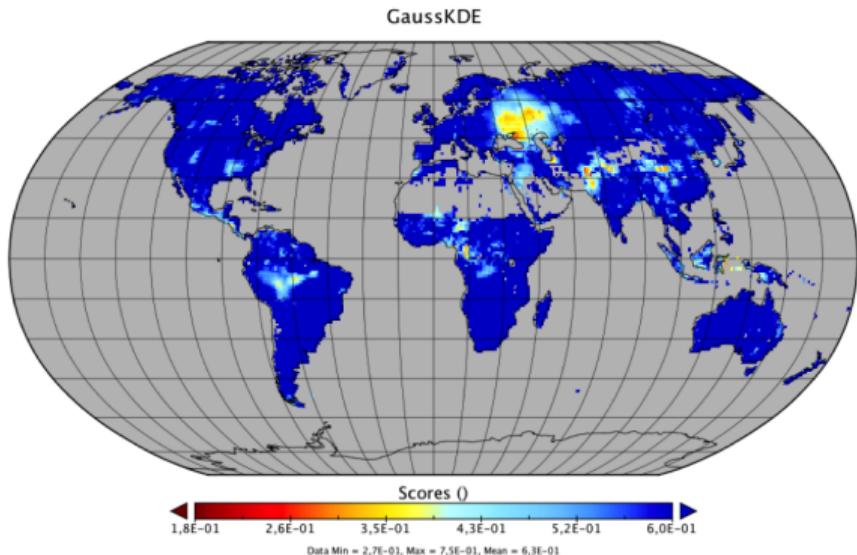


# Avenues of scientific exploitation

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Detecting multidimensional anomalies

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



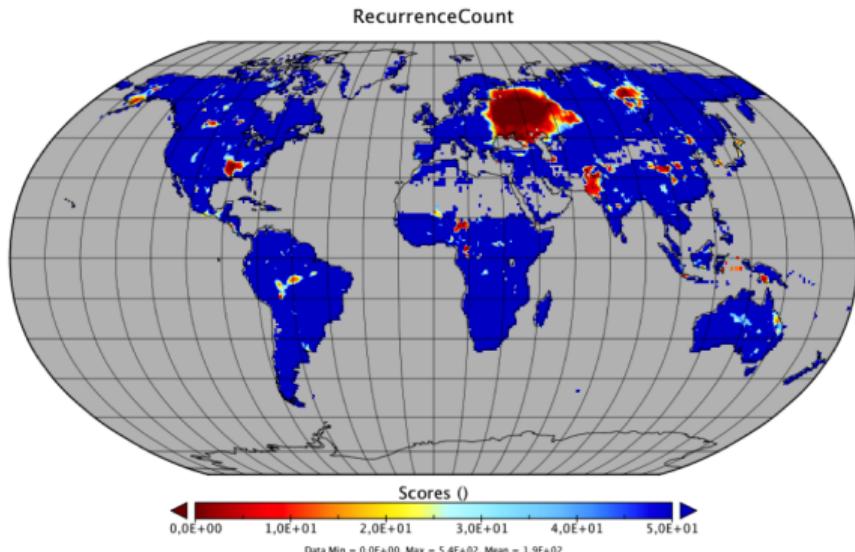
PhD thesis by Milan Flach

# Avenues of scientific exploitation

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Detecting multidimensional anomalies

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



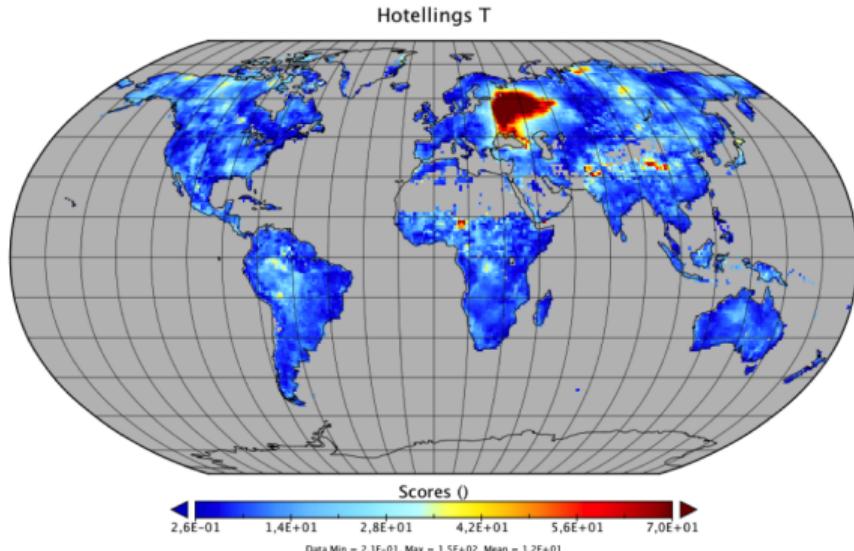
PhD thesis by Milan Flach

# Avenues of scientific exploitation

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## Detecting multidimensional anomalies

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



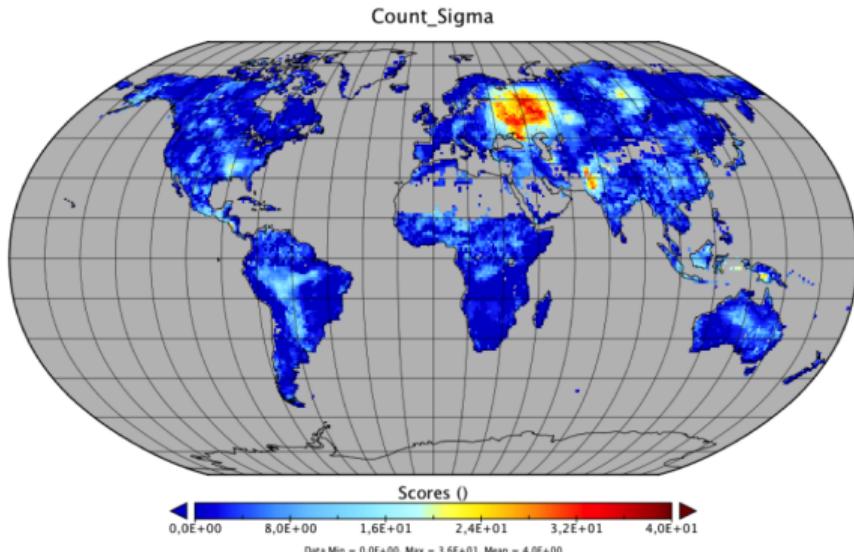
PhD thesis by Milan Flach

# Avenues of scientific exploitation

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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

Introduction  
Opportunities  
Challenges  
Earth System Data Cube  
Data Analytic Toolkit  
Scientific Perspectives  
Outlook

## 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.

## 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.

[Introduction](#)  
[Opportunities](#)  
[Challenges](#)  
[Earth System Data Cube](#)  
[Data Analytic Toolkit](#)  
[Scientific Perspectives](#)  
[Outlook](#)

<http://earthsystemdatacube.net>



Max Planck Institute  
for Biogeochemistry



Stockholm Resilience Centre  
Sustainability Science for Biosphere Stewardship

