Max-Planck-Institut
für Biogeochemie

Wir und die Daten: Ein Beitrag aus der Perspektive der Biogeochemie & Erdystemanalyse

Martin Jung

Miguel Mahecha

Markus Reichstein

Aspekte

- Womit befassen wir uns?
- Wie sehen unsere Daten und Datenmanagmentansätze aus?
- Womit haben wir zu kämpfen?

NOAH Visualization

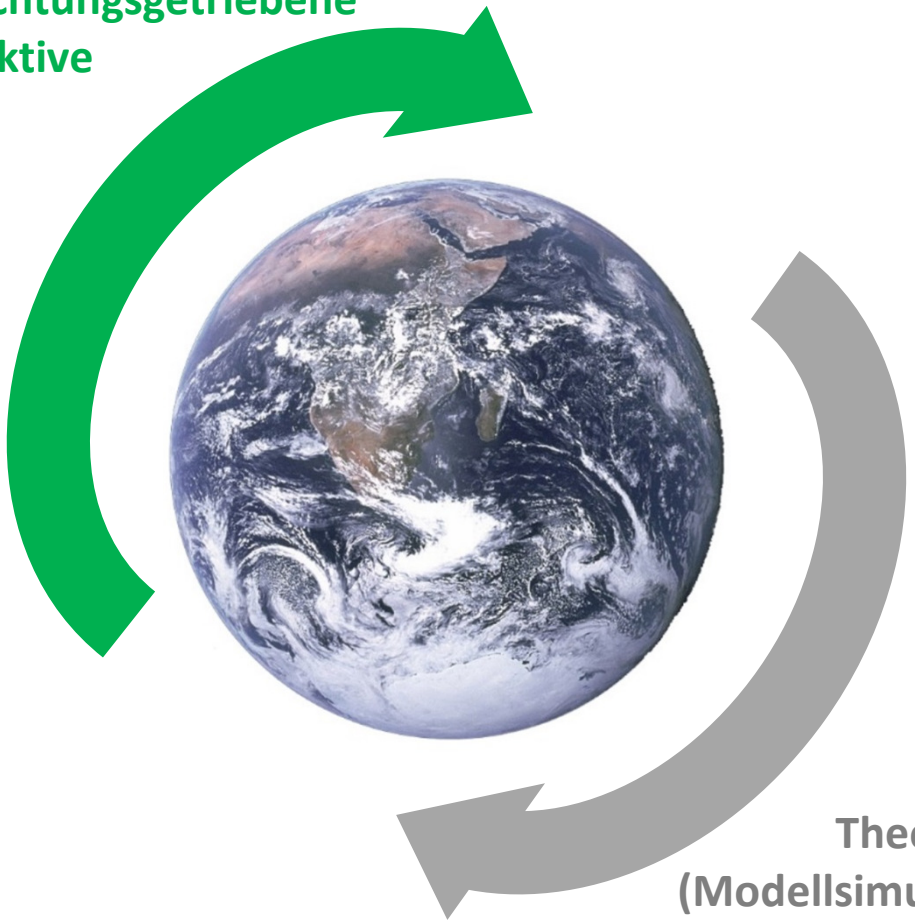
System Erde

https://www.mpg.de/mpforschung

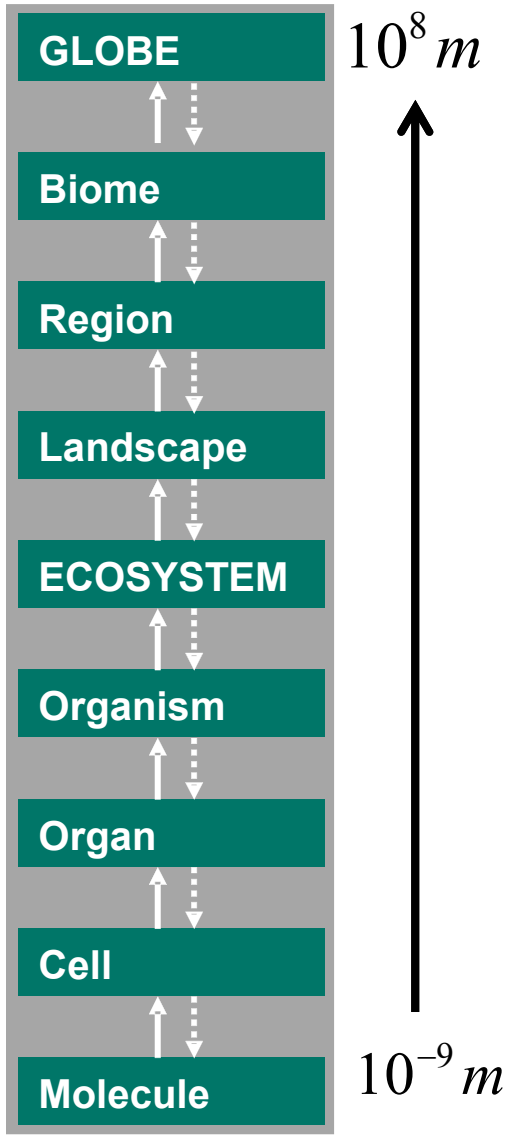
Wie funktioniert das Erdsystem?

Focus: Biosphäre

Beobachtungsgetriebene Perspektive



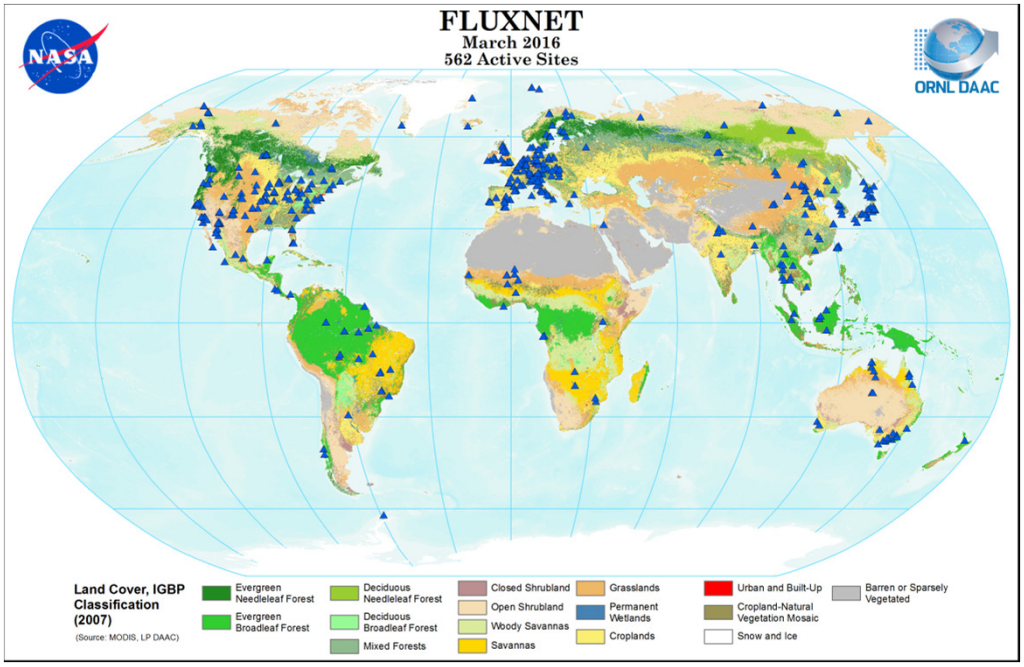
Theoretische (Modellsimulations-) Perspektive



Was für Daten?

„**Stationsdaten**“: (meist) Zeitserien verschiedener in-situ gemessener Variablen (z.B. Kohlenstoffaustausch zwischen Land und Atmosphere) + diverse meta-Daten

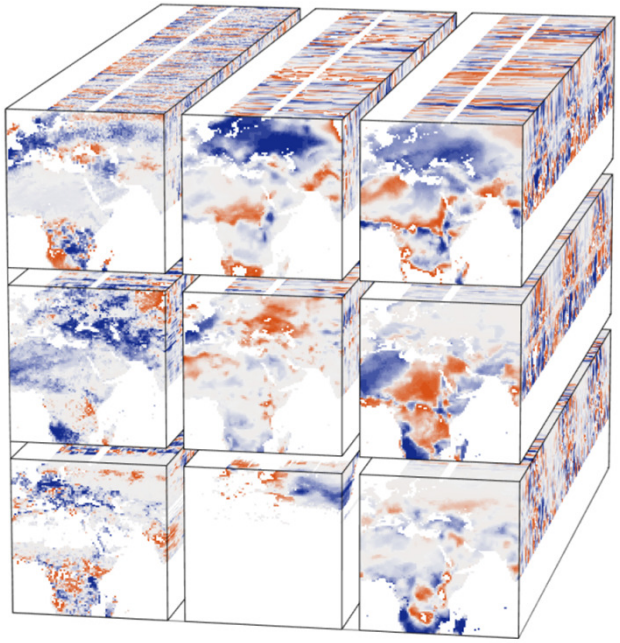
Quellen: „Kollaborative Netzwerke“

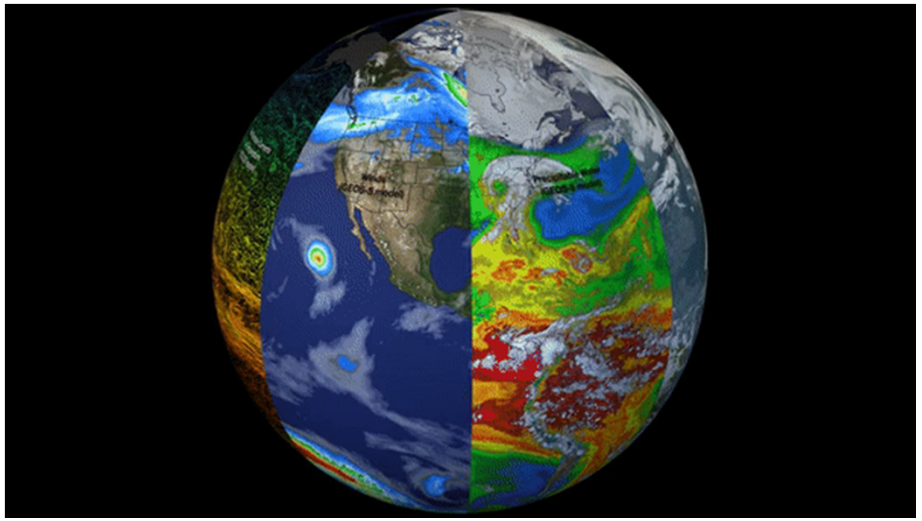


Was für Daten?

„Data Cubes“: Multivariate, räumlich und zeitlich aufgelöste „Gitter“ (grids) → „Videos von Karten der Erde“

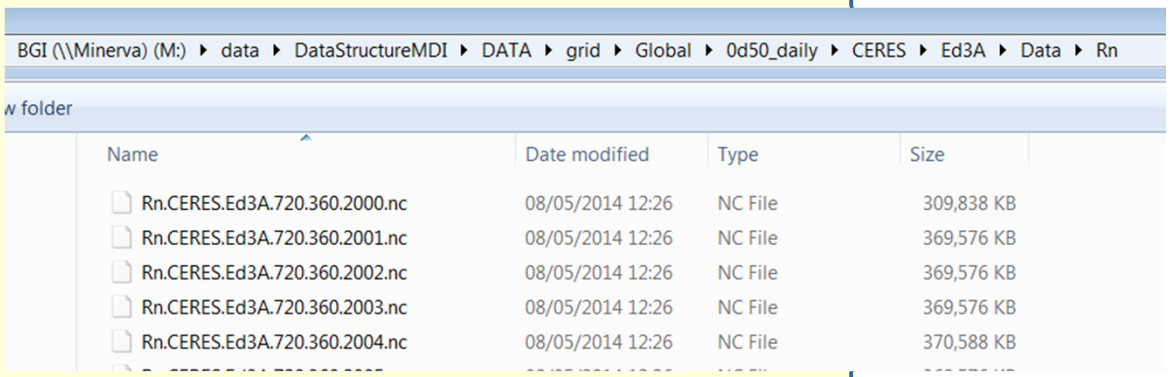
Quellen: Satellitenerdbeobachtung, Erdsystemmodelle





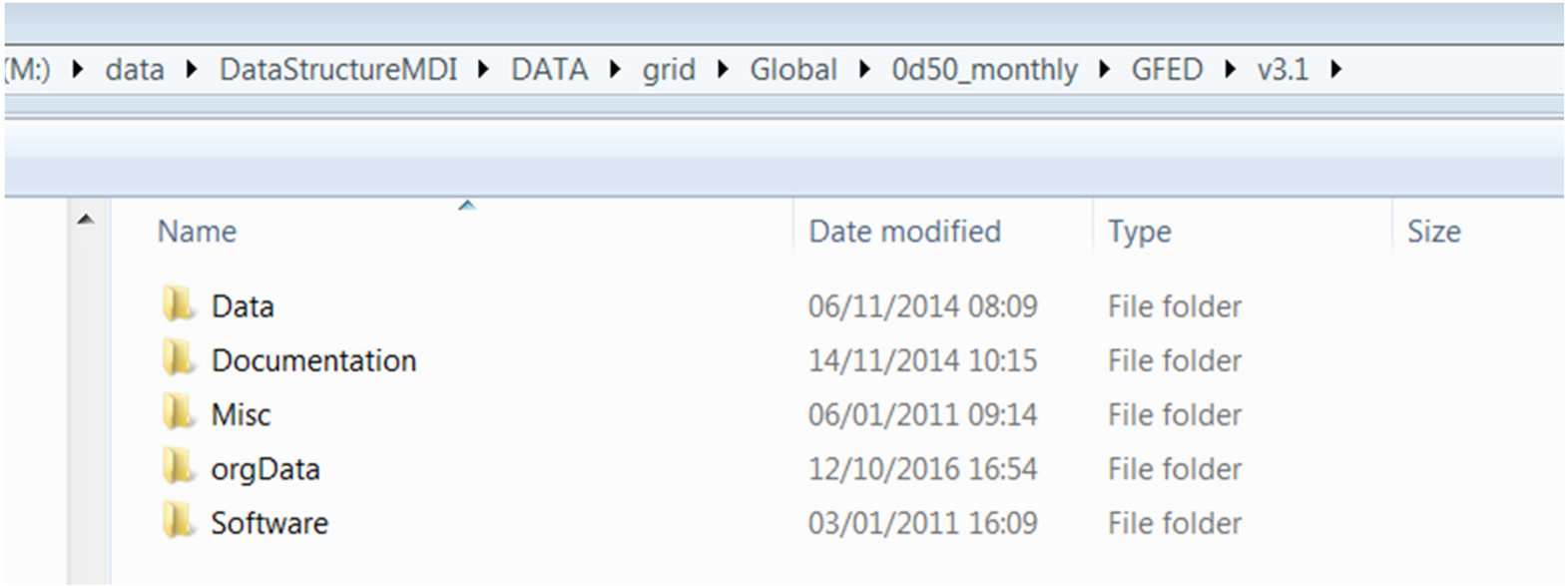
Wie sehen solche Daten aus?

```
uweber@pc027:~/data/grid/Global/0d50_daily/CERES/Ed3A/Data/Rn> ncdump -h Rn.CERES.Ed3A.720.360.2006.nc
netcdf Rn.CERES.Ed3A.720.360.2006 {
dimensions:
        lon = 720 ;
        lat = 360 ;
        time = 365 ;
variables:
        float Rn(time, lat, lon) ;
                Rn:long_name = "Net Radiation" ;
                Rn:units = "W m-2" ;
                Rn:scale_factor = 1.f ;
                Rn:add_offset = 0.f ;
                Rn:missing_value = -9999.f ;
                Rn:_FillValue = -9999.f ;
        double lon(lon) ;
                lon:long_name = "longitude" ;
                lon:standard_name = "longitude" ;
                lon:units = "degrees_east" ;
                lon:valid_range = -180.f, 180.f ;
        double lat(lat) ;
                lat:long_name = "latitude" ;
                lat:standard_name = "latitude" ;
                lat:units = "degrees_north" ;
                lat:valid_range = -90.f, 90.f ;
        double time(time) ;
                time:units = "days since 1582-10-14 00:00" ;
                time:calendar = "gregorian" ;

// global attributes:
                :title = "Net Radiation calculated from CERES input)" ;
                :provided_by = "http://ceres.larc.nasa.gov/index.php" ;
                :created_by = "ulrich weber (uweber@bgc-jena.mpg.de)" ;
                :method = "Rn = (lwdown - lwup) + (swdown - swup)" ;
                :version = "Ed3A" ;
                :history = "Thu May  8 12:26:59 2014" ;
                :download = "http://ceres-tool.larc.nasa.gov/ord-tool/jsp/SYN1degSelection.jsp" ;
                :reference = "http://ceres.larc.nasa.gov/documents/DQ_summaries/CERES_SYN1deg_Ed3A_DQS.pdf" ;
}
```



Wie verwalten wir unsere Daten?

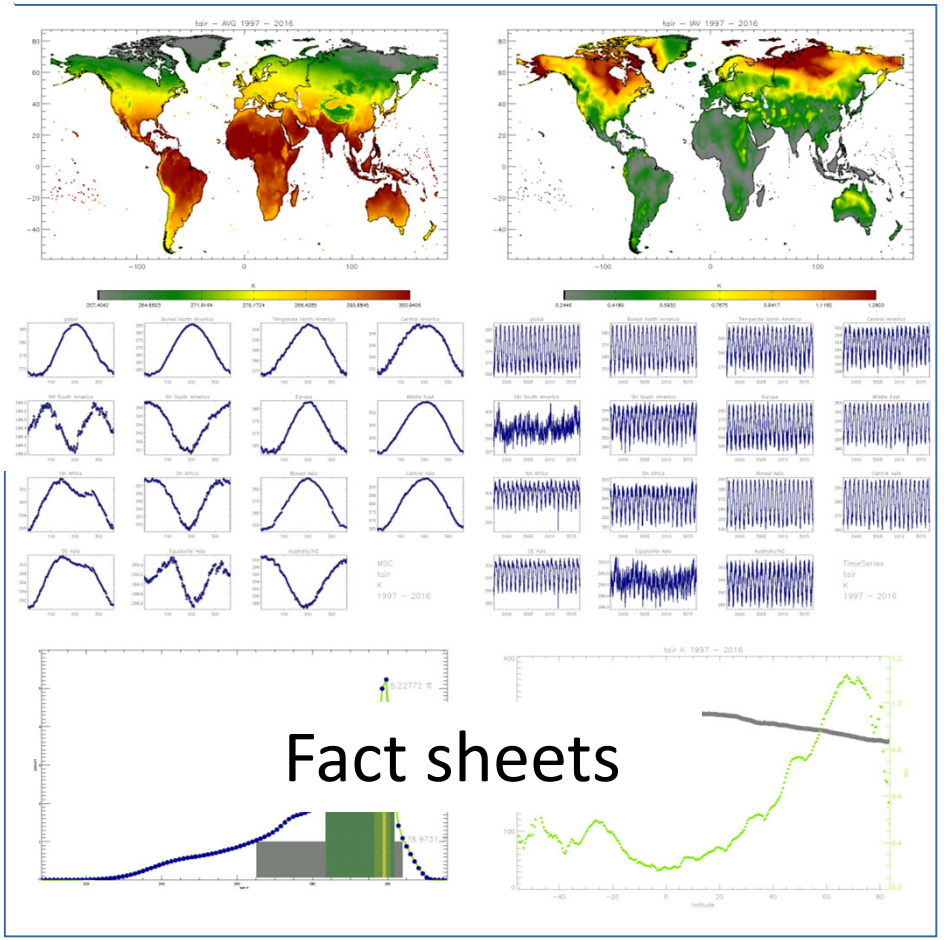
Ordnerstruktur:
[räumlicher Ausschnitt]→[raum-zeitliche Auflösung]→[Daten-Produkt]→[Version]



- Konsistentes Dateiformat (ncdf)
- Standardisierte Attributnamen
- Konventionen wie intern abgelegt

Datenstruktur Werkzeuge

Suchen



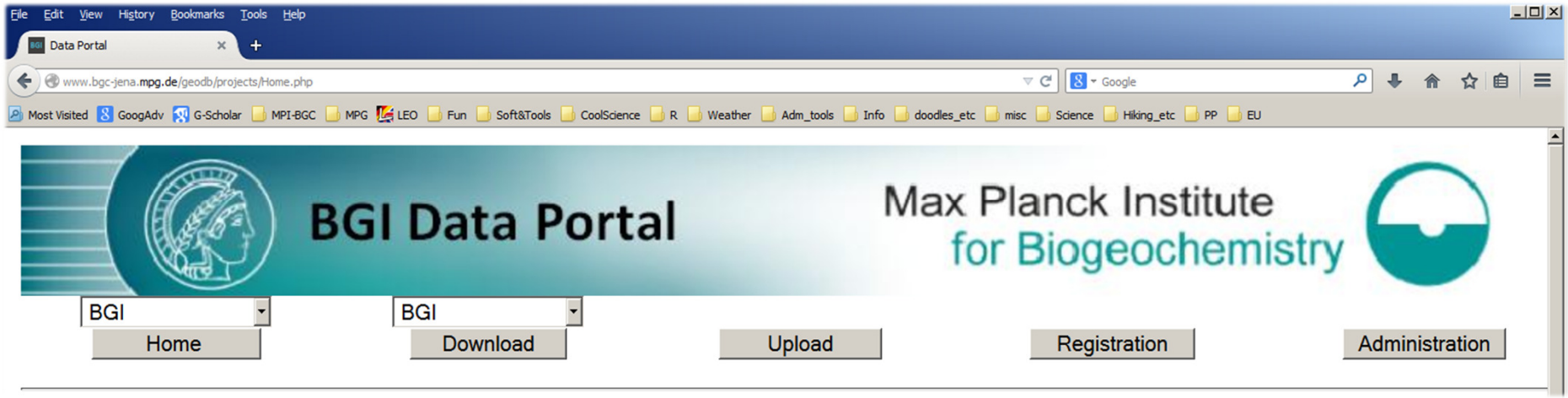
Fact sheets

In Entwicklung: code für robuste und effiziente Verarbeitung

Was machen wir mit den Daten?

→ Finden
→ Akquirieren
→ Vorprozessieren für Datenstruktur
→ Verwenden
→ neue Datenprodukte → publizieren & verteilen (DOIs via Bibliothek und MPDL)

} Data Manager



ML training
Predictor variables
ML prediction

Ensemble estimation

Prediction

Training & cross-validation

Feature selection

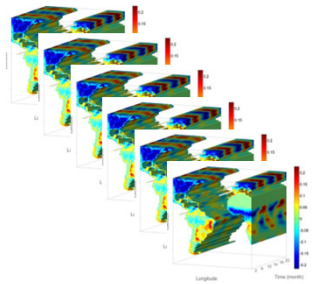
Flux data processing

RS@0.083° & 8-daily

RS+METEO@0.5° & daily

3 Ensembles
LE (9ML × 3EBC),
H (9ML × 3EBC),
Rn (8ML)

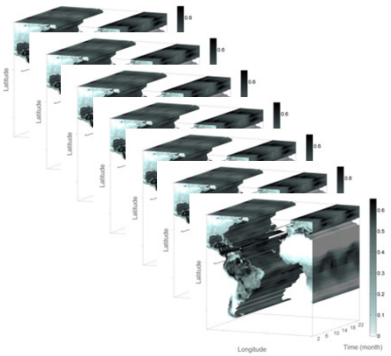
12 Ensembles
4GC × LE (3ML × 3EBC),
4GC × H (3ML × 3EBC),
4GC × Rn (3ML)



84 Global products
(3ML × 7fluxes × 4GC)

Global MODIS d

Global climate data
(GC)



FLUXCOM
4 Jahre Arbeit

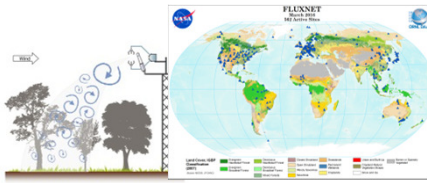
9 Machine learning methods (ML)

Satellite based predictor variables

meteorological and satellite based predictor variables

7 Measured/energy balance corrected (EBC) fluxes:
Latent energy (LE, LE_{RES}, LE_{BWR})
Sensible heat (H, H_{RES}, H_{BWR})
Net radiation (Rn)

Tower meteo



LETTER

doi:10.1038/nature09396

Recent decline in the global land evapotranspiration trend due to limited moisture supply

Martin Jung[1], Markus Reichstein[1], Philippe Ciais[2], Sonia I. Seneviratne[3], Justin Sheffield[4], Michael L. Goulden[5], Gordon Bonan[6], Alessandro Cescatti[7], Jiquan Chen[8], Richard de Jeu[9], A. Johannes Dolman[9], Werner Eugster[10], Dieter Gerten[11], Damiano Gianelle[12], Nadine Gobron[13], Jens Heinke[11], John Kimball[14], Beverly E. Law[15], Leonardo Montagnani[16], Qiaozhen Mu[17], Brigitte Mueller[3], Keith Oleson[6], Dario Papale[18], Andrew D. Richardson[19], Olivier Roupsard[20], Steve Running[17], Enrico Tomelleri[1], Nicolas Viovy[2], Ulrich Weber[1], Christopher Williams[21], Eric Wood[4], Sönke Zaehle[1] & Ke Zhang[14]

LETTER

doi:10.1038/nature20780

Compensatory water effects link yearly global land CO_2 sink changes to temperature

Martin Jung[1], Markus Reichstein[1,2], Christopher R. Schwalm[3], Chris Huntingford[4], Stephen Sitch[5], Anders Ahlström[6,7], Almut Arneth[8], Gustau Camps-Valls[9], Philippe Ciais[10], Pierre Friedlingstein[11], Fabian Gans[1], Kazuhito Ichii[12,13], Atul K. Jain[14], Etsushi Kato[15], Dario Papale[16], Ben Poulter[17], Botond Raduly[16,18], Christian Rödenbeck[19], Gianluca Tramontana[16], Nicolas Viovy[10], Ying-Ping Wang[20], Ulrich Weber[1], Sönke Zaehle[1,2] & Ning Zeng[21,22]

LETTER

doi:10.1038/nature13731

Global covariation of carbon turnover times with climate in terrestrial ecosystems

Nuno Carvalhais[1,2], Matthias Forkel[1], Myroslava Khomik[1,3], Jessica Bellarby[4,5], Martin Jung[1], Mirco Migliavacca[1,6], Mingquan Mu[7], Sassan Saatchi[8], Maurizio Santoro[9], Martin Thurner[1], Ulrich Weber[1], Bernhard Ahrens[1], Christian Beer[1,10], Ales [illegible] James T. Randerson[7] & Markus Reichstein[1]

Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate

Christian Beer,[1*] Markus Reichstein,[1] Enrico Tomelleri,[1] Philippe Ciais,[2] Martin Jung,[1] Nuno Carvalhais,[1,3] Christian Rödenbeck,[4] M. Altaf Arain,[5] Dennis Baldocchi,[6] Gordon B. Bonan,[7] Alberte Bondeau,[8] Alessandro Cescatti,[9] Gitta Lasslop,[1] Anders Lindroth,[10] Mark Lomas,[11] Sebastiaan Luyssaert,[12] Hank Margolis,[13] Keith W. Oleson,[7] Olivier Roupsard,[14,15] Elmar Veenendaal,[16] Nicolas Viovy,[2] Christopher Williams,[17] F. Ian Woodward,[11] Dario Papale[18]

CARBON CYCLE

The dominant role of semi-arid ecosystems in the trend and variability of the land CO_2 sink

Anders Ahlström,[1,2‡] Michael R. Raupach,[3†] Guy Schurgers,[4] Benjamin Smith,[1] Almut Arneth,[5] Martin Jung,[6] Markus Reichstein,[6] Josep G. Canadell,[7] Pierre Friedlingstein,[8] Atul K. Jain,[9] Etsushi Kato,[10] Benjamin Poulter,[11] Stephen Sitch,[12] Benjamin D. Stocker,[13,14] Nicolas Viovy,[15] Ying Ping Wang,[16] Andy Wiltshire,[17] Sönke Zaehle,[6] Ning Zeng[18]

CrossMark click for updates

PNAS PLUS

Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence

Luis Guanter[a,1,2], Yongguang Zhang[a,1], Martin Jung[b], Joanna Joiner[c], Maximilian Voigt[a], Joseph A. Berry[d], Christian Frankenberg[e], Alfredo R. Huete[f], Pablo Zarco-Tejada[g], Jung-Eun Lee[h], M. Susan Moran[i], Guillermo Ponce-Campos[i], Christian Beer[j], Gustavo Camps-Valls[k], Nina Buchmann[l], Damiano Gianelle[m], Katja Klumpp[n], Alessandro Cescatti[o], John M. Baker[p], and Timothy J. Griffis[q]

Probleme und neue Konzepte

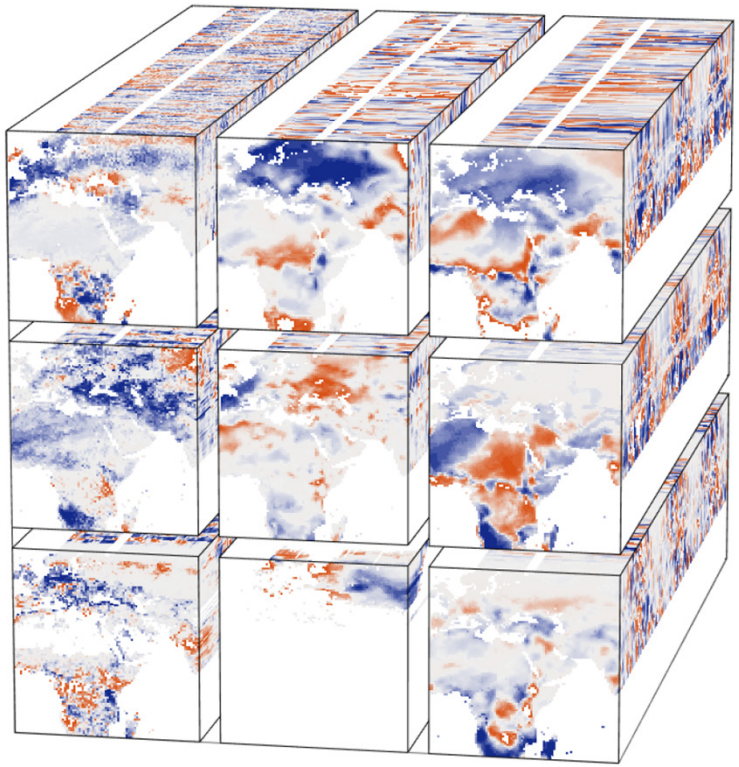
Erstellung, Dokumentation, Publikation und Teilen von Prozessierungs- und Analyseketten (workflows)

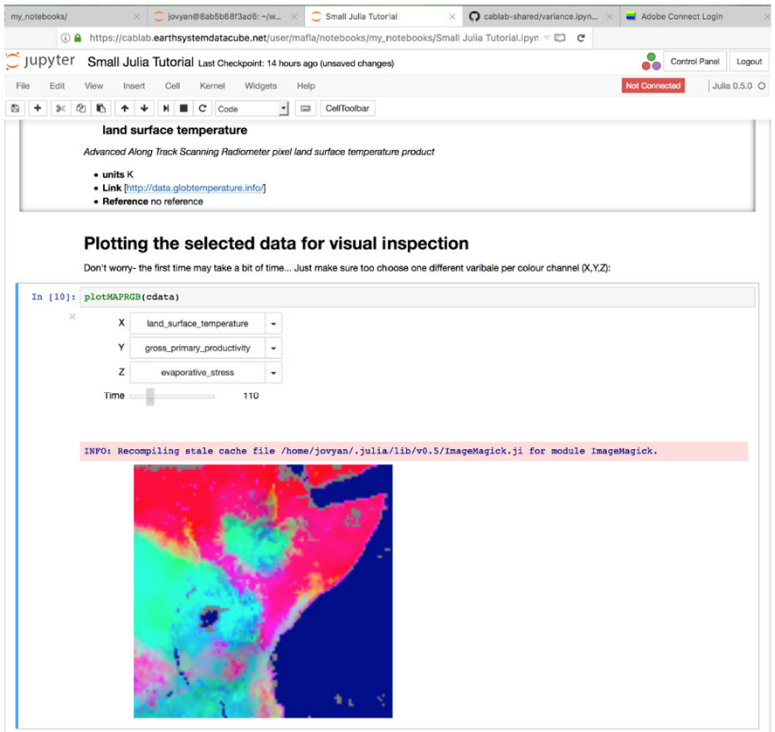
→ Für Transparenz und Reproduzierbarkeit
→ Für Wiederverwendbarkeit
→ Für effektives und effizientes Arbeiten („Wissenschaft“ vs Softwareentwicklung und Datamanagement)

Warum wird oft nicht der code eines Projektes publiziert?
- Oft schwer lesbar und wiederverwendbar – großer Anteil betrifft performantes Verarbeiten und nicht „Wissenschaft“ (z.B. Parallelisierung, Kopplung an Hardware)
- Oft schlecht dokumentiert

Earth System Data Cube

Harmonisierte Daten + Analysewerkzeug





http://earthsystemdatacube.net/

Das data cube Analysewerkzeug

- Viele Standardoperationen sind effizient implementiert (high-performance computing läuft im Hintergrund)
- Neue Funktionen können dazu gefügt werden
- Workflows können einfach dokumentiert und geteilt werden (z.B. via Jupiter notebooks)
- Julia, Python, R interfaces

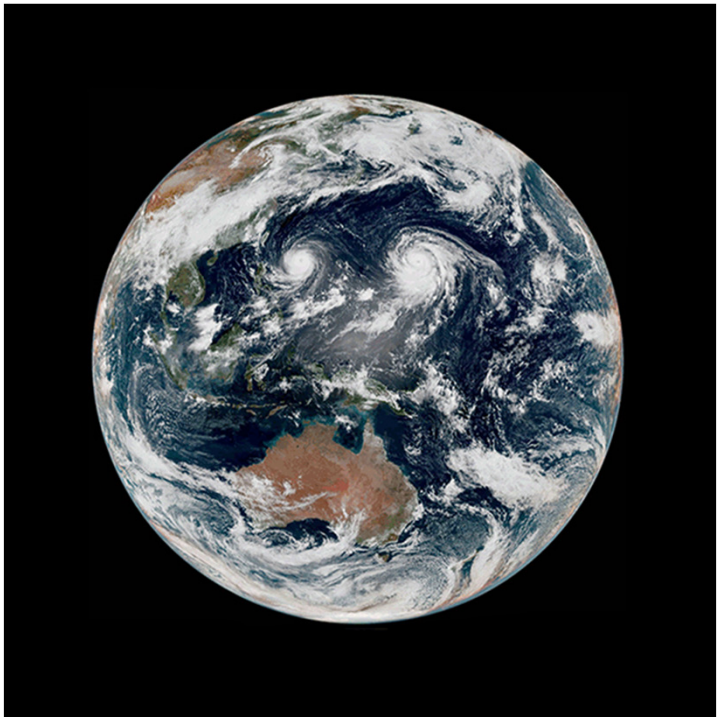
```
In [4]: @time cube_filled     = map(gapFillMSC,cdata,46,max_cache=1e7);
        18.017533 seconds (14.40 M allocations: 877.965 MB, 1.70% gc time)

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)
```

Probleme und neue Konzepte

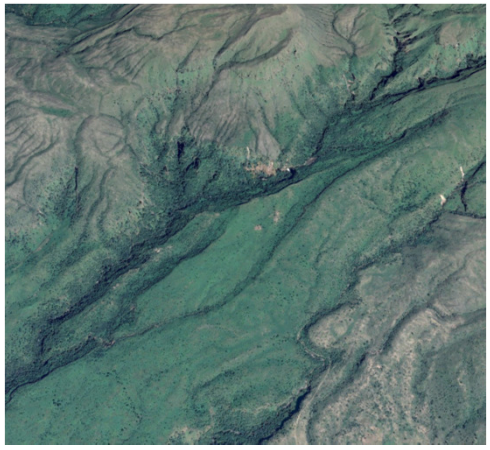
Too Big Data

Mehrere hundert Terrabytes an Erdsystemdaten werden täglich von Satelliten aufgezeichnet

z.B. Himawari-8 produziert ca.
350 GB/Tag



Hohe zeitliche Auflösung (Minuten)



Hohe räumliche Auflösung (Meter)

Cloud computing – aber wie und wo?

Klassische Ansatz: Datenakquise → Vorprozessierung → Analyse … funktioniert nicht mehr (bzw. sehr uneffektiv)

„Neue“ Ansatz: Cloud computing → Implementierung von workflows in big data clouds (z.B. Google Earth Engine, Amazon …)

- enthalten meist nicht alle benötigten (großen) Daten
- keine Garantie für Reproduzierbarkeit (Daten „veralten“ und werden nicht für immer vorgehalten)
- erfordert neues „know-how“
- kostenintensiv (?)
- IT Sicherheit (?)

"Sollen wir alles auf Google-Earth Engine rechnen, oder gibt es eine MPG Lösung..?"

Vielen Dank für Ihre Aufmerksamkeit!

mjung
mmahecha
mreichstein
@bgc-jena.mpg.de