



Lukas Ciepielowski

Wissenschaftliche Datenformate

08.01.2019

Gliederung

1. Wissenschaftliche Datenformate
2. NetCDF
3. Bearbeiten der Daten mit CDOs
4. Visualisierung der Daten
5. Zusammenfassung
6. Literatur

1.1 Wissenschaftliche Datenformate

Allgemeine Informationen

- Entstehen aus Anforderungen
- Standards (in der Wissenschaft) sehr wichtig
- Meta-Daten
- Daten werden in N-Dimensionalen Arrays gespeichert

1.2 Wissenschaftliche Datenformate

Voraussetzungen

- Kompakte Speicherung (Komprimierung)
- Portabilität
- Schneller Zugriff
- Software zur Bearbeitung
- Programmierschnittstelle (API) für verschiedene Sprachen



Abb. : [1]

Anforderungen an die Hardware

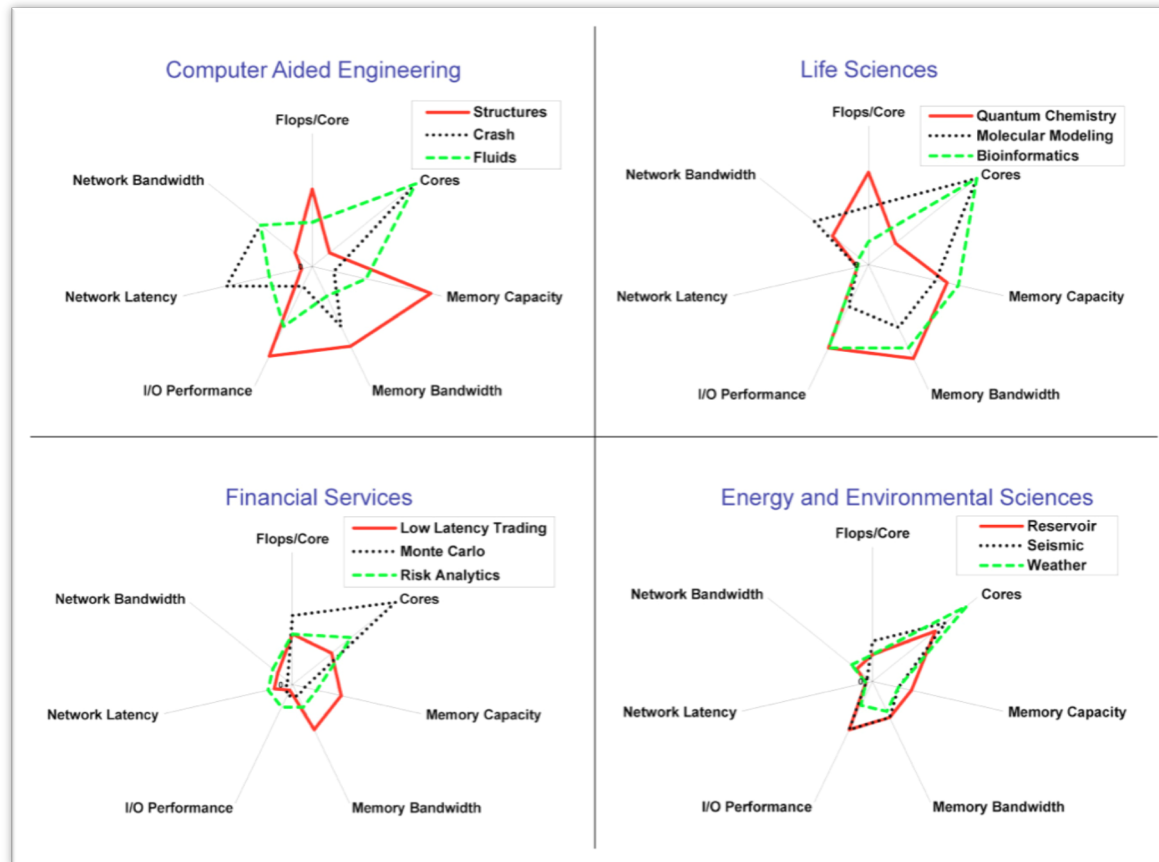


Abb. : [2] https://www.ibm.com/it-infrastructure/images/resources/power/IT-infrastructure-HPC_HPDA-figure4.jpg

1.3 Wissenschaftliche Datenformate

Beispiele für Datenformate

- Klimatologie → NetCDF (**Net**work **C**ommon **D**ata **F**ormat)
- Biologie → PDB (**P**rotein **D**ata **B**ank) file format
 - 3D Beschreibung von Molekülen
 - Daten werden in Tabellenform gespeichert
- Medizin → DICOM (**D**igital **I**maging and **C**ommunications in Medicine)
 - Standard in der Medizin
 - DICOM Datensatz dient als Container, auch Metadaten werden gespeichert
 - **Informationen** : Bilddaten, mathematische und behandlungsspezifische Informationen

Zuwachs der Datenmenge in der Klimatologie

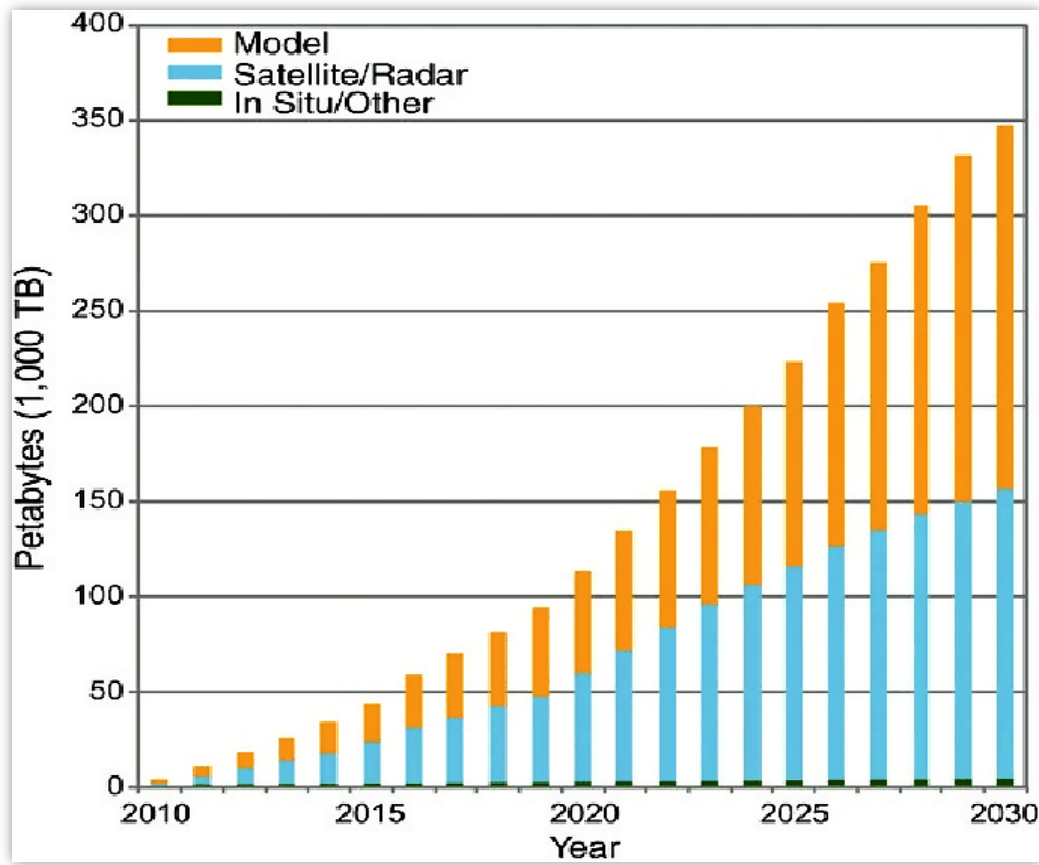


Abb. : [3]

https://www.researchgate.net/profile/Sebastian_Bathiany/publication/312475518/figure/fig4/AS:451512070610944@1484660274143/Projected-increase-in-the-amount-of-data-from-climate-models-remote-sensing-and-in-situ.png

2. NetCDF

1. Struktur
2. NetCDF File Format
3. NetCDF-4/HDF5

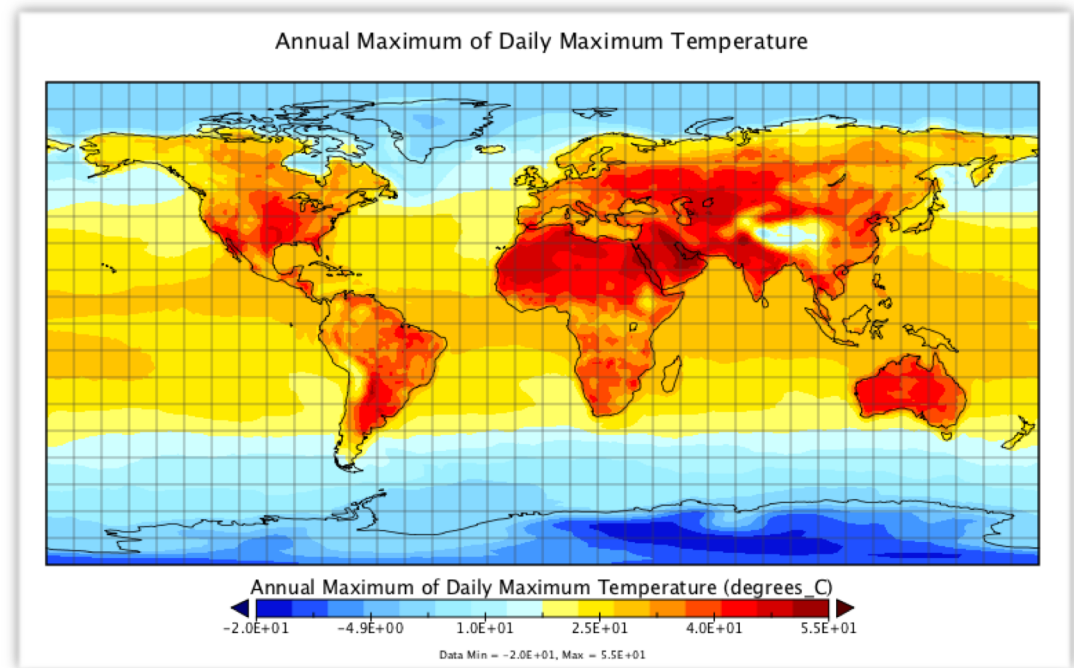


Abb. : [4]

NetCDF Daten sind...

- Selbstbeschreibend
- Übertragbar
- Skalierbar



Abb. : [5]

2.1 NetCDF

Dimensionen

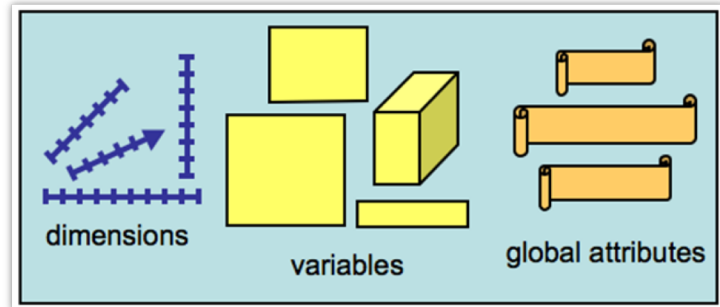
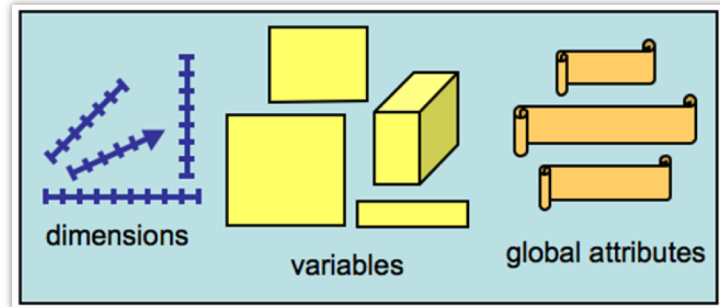


Abb. : [6]

- Enthält Name und Größe
- Durch Dimensionen werden physikalische Größen dargestellt
 - Zeit, Breite, Länge oder Höhe
- Ganzzahliger positiver Wert
- In einer netCDF Datei kann es nur eine Dimension der Größe UNLIMITED geben
 - Ausnahme bei NetCDF-4

2.1 NetCDF

Variablen



- Array mit Werten desselben Datentyps Abb. : [6]
- Besteht aus Namen, Typ und Form, welche durch die angegebene Dimension festgelegt wird
- „Record variable“ = Variablen mit min. einer unlimitierten Dimension
- Anzahl der Dimensionen wird als Rang bezeichnet
 - Rang 0 = Skalar
 - Rang 1 = Vektor
 - Rang 2 = Matrix

2.1 NetCDF

Koordinatenvariablen

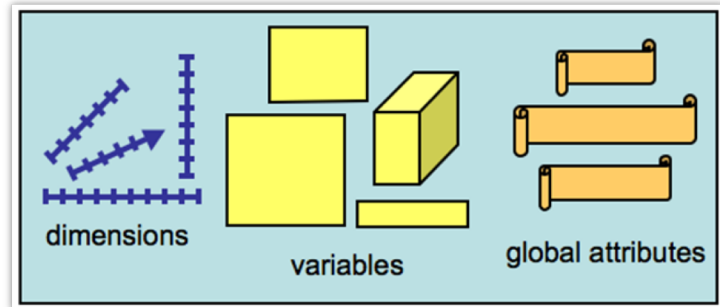


Abb. : [6]

- Gleicher Name wie eine Dimension
- Vektor Variable mit der identischen Form zu ihrer Dimension
- Keine besondere technische Bedeutung
- **ABER** : werden von NetCDF-Tools besonders behandelt
- Werden genutzt, um physikalische Koordinaten in einem Koordinatensystem abzubilden

2.1 NetCDF *Attribute*

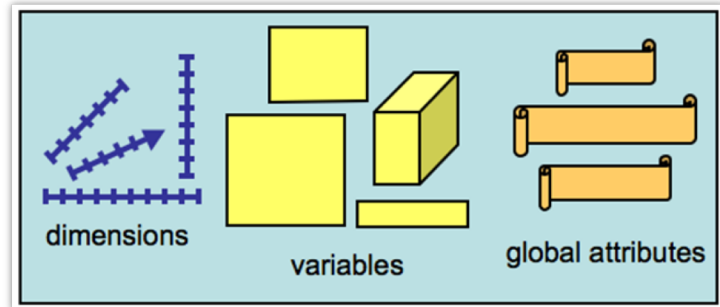


Abb. : [6]

- Werden genutzt um Daten über Daten zu speichern (Meta-Daten)
- Meistens Informationen über eine Variable
- Attribute werden mit dem Namen der Variable und einem eigenen Namen versehen
- **Globale Attribute** : liefern Informationen über die gesamte NetCDF-Datei

NetCDF Data Model

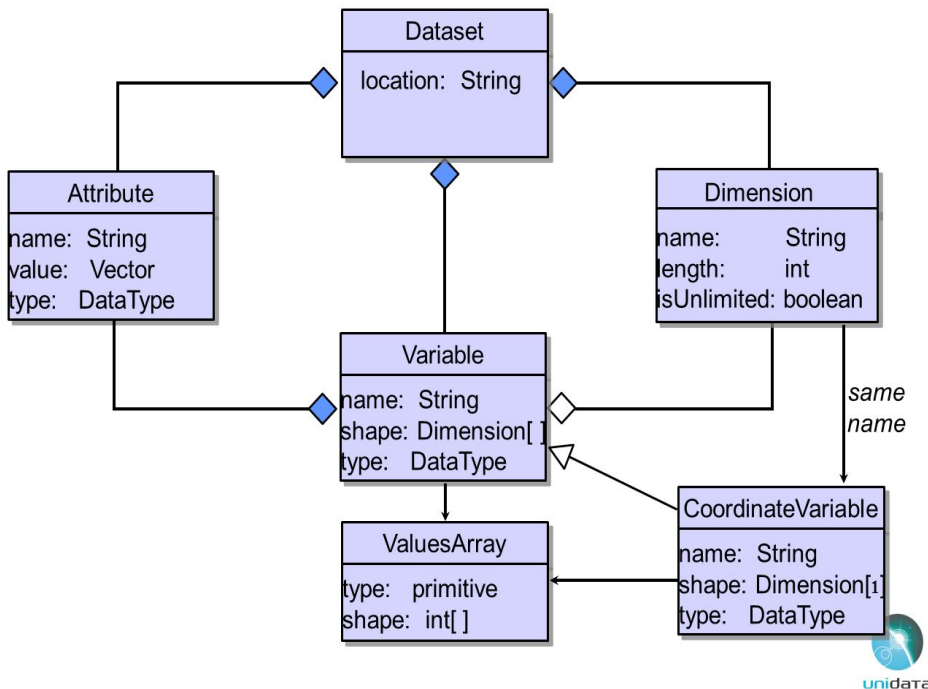


Abb. : [7]

Beispiel einer NetCDF Datei in der CDL-Notation

```

netcdf filename {
dimensions:
    lat = 3 ;
    lon = 4 ;
    time = UNLIMITED ; // (2 currently)

variables:
    float lat(lat) ;
        lat:long_name = "Latitude" ;
        lat:units = "degrees_north" ;
    float lon(lon) ;
        lon:long_name = "Longitude" ;
        lon:units = "degrees_east" ;
    int time(time) ;
        time:long_name = "Time" ;
        time:units = "days since 1895-01-01" ;
        time:calendar = "gregorian" ;
    float rainfall(time, lat, lon) ;
        rainfall:long_name = "Precipitation" ;
        rainfall:units = "mm yr-1" ;
        rainfall:missing_value = -9999.f ;

// global attributes:
    :title = "Historical Climate Scenarios" ;
    :Conventions = "CF-1.0" ;

data:
    lat = 48.75, 48.25, 47.75;
    lon = -124.25, -123.75, -123.25, -122.75;
    time = 364, 730;
    rainfall =
        761, 1265, 2184, 1812, 1405, 688, 366, 269, 328, 455, 524, 877,
        1019, 714, 865, 697, 927, 926, 1452, 626, 275, 221, 196, 223;
}
    
```

Annotations in the CDL notation example:

- Coordinate variable:** Points to the `lat(lat)` variable definition.
- Variable attribute:** Points to the `time:calendar = "gregorian"` attribute.
- Global attribute:** Points to the `:Conventions = "CF-1.0"` global attribute.

Abb. : [8]

2.2 NetCDF File Format

- NetCDF Classic Format (CDF-1)
 - Default Format von NetCDF
 - Beschränkt auf 2 GiBs
 - Ziel ist dabei die Portabilität → wird empfohlen
- NetCDF 64-bit Offset Format (CDF-2)
 - Beschränktheit der Größe wird entlastet
 - Variablen mit einer Größe von 4 GiB möglich
- NetCDF 64-bit Data Format (CDF-5)
 - 64-bit Integers

2.3 NetCDF-4/HDF5

- Setzt auf das HDF5 Format auf
- Größe der Dateien nicht beschränkt
- Mehrere *unlimitierte Dimensionen* möglich
- Gruppen sind möglich



Vorteile von HDF5 werden direkt genutzt

3. Bearbeiten der Daten mit CDOs

- Sammlung von Werkzeugen zur Bearbeitung und zum Auslesen von wissenschaftlichen Daten
 1. Informationen über Inhalte
 2. Modifizieren
 3. Auswerten
- Anwendbar für NetCDF und GRIB Files

3. Bearbeiten der Daten mit CDOs

1. Informationen über Inhalte :

- Anzeigen der Variablen : `cdo showname infile`
- Grid description : `cdo griddes infile`



```
gridtype : gaussian
gridsize : 2048
xname    : lon
xlongname : longitude
xunits   : degrees_east
yname    : lat
ylongname : latitude
yunits   : degrees_north
xsize    : 64
ysize    : 32
```

Abb. : [9]

2. Modifizieren :

- Variablen Name ändern : `cdo chname, temp, t2m infile outfile`
- Definieren einer Zeitachse : `cdo settaxis,1987-01-16,12:00:00,1mon infile outfile`

3. Bearbeiten der Daten mit CDOs

- Interpolation : `cdo remapbil,n32 infile outfile`

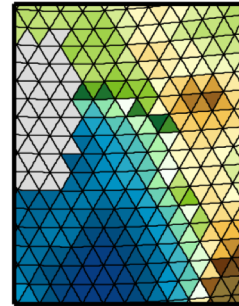
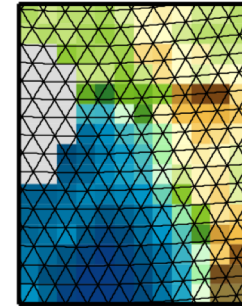
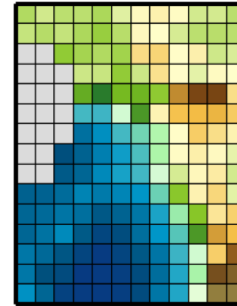
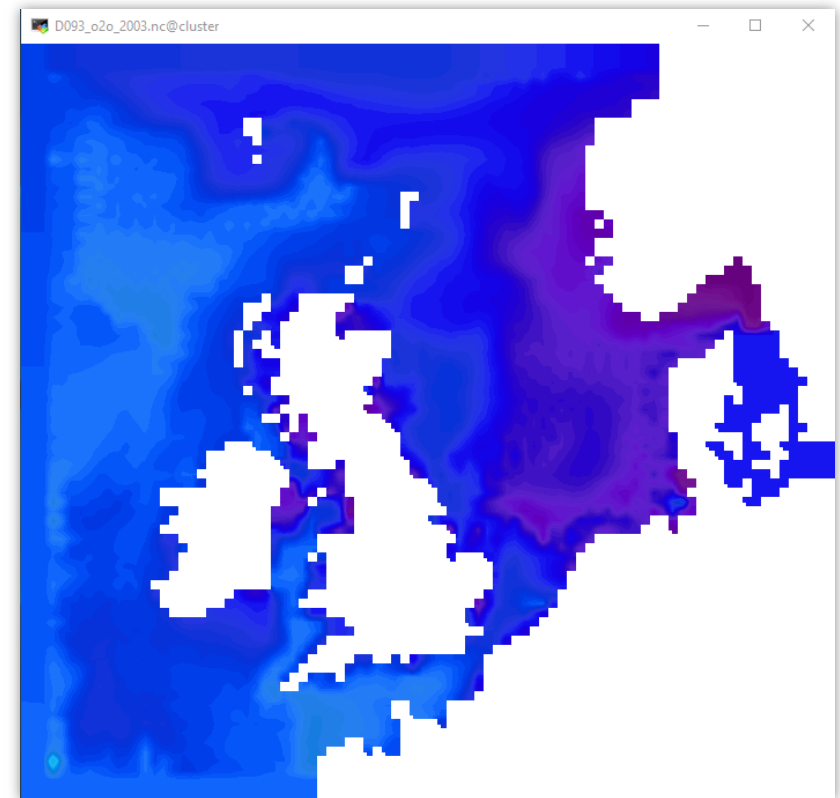


Abb. : [10]

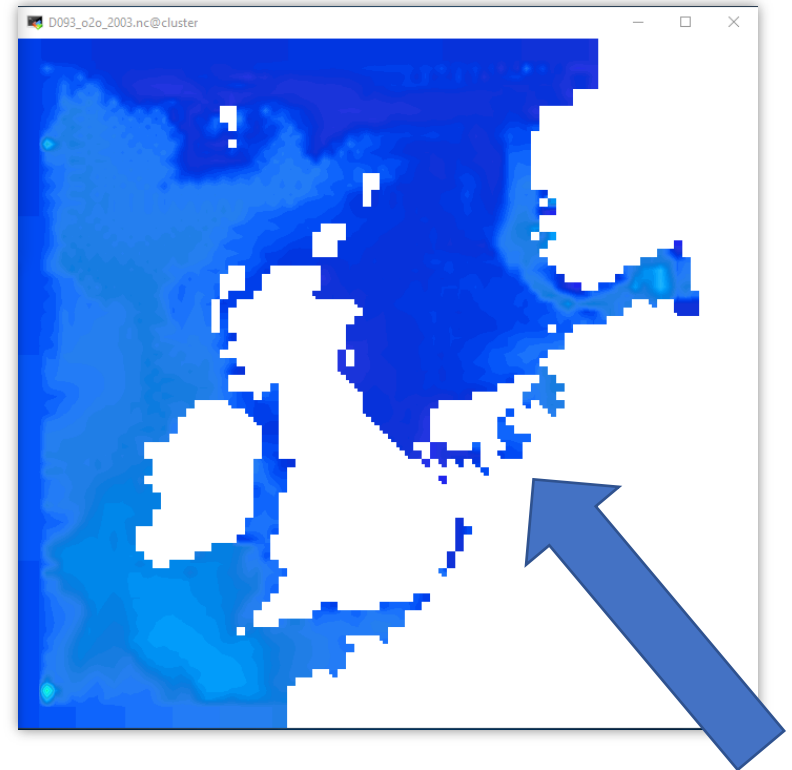
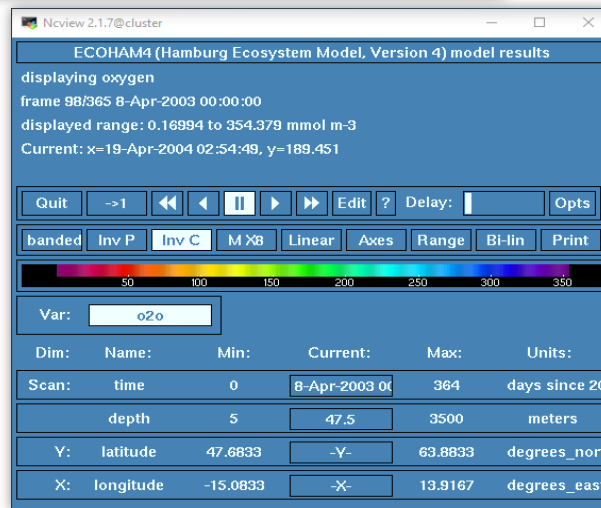
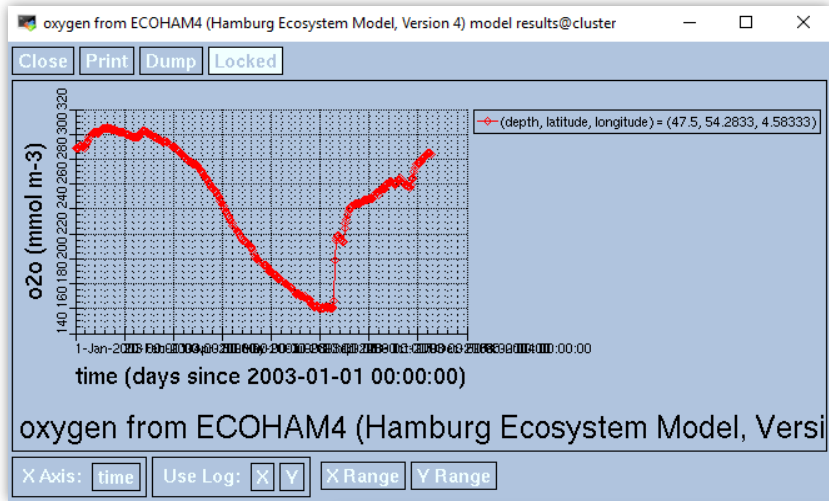
3. Auswerten :

- Durchschnittswert pro Stunde : `cdo hourmean infile outfile`
- Durchschnitt mehrerer Dateien berechnen : `cdo ensmean infile[1-6] outfile`
- Median mehrerer Dateien berechnen : `cdo enspctl, 50 infile[1-6] outfile`

4. Visualisierung der Daten



Quelle: H. Lenhart pers. comm.



Quelle: H. Lenhart pers. comm.

5. Zusammenfassung

- Wissenschaftliche Datenformate entstehen aus den Anforderungen
- Gewisse Voraussetzungen sollten gegeben sein
- NetCDF :
 - Bereich der Klimatologie
 - Dimensionen, Variablen, Attribute
 - NetCDF-4/HDF5
- CDOs bieten eine Sammlung von Werkzeugen an
- Visualisierung der Daten durch Programme möglich

6. Literatur

Informationsquellen

1. https://www.unidata.ucar.edu/software/netcdf/workshops/2012/third_party/CDO.html
2. <http://desktop.arcgis.com/de/arcmap/10.3/manage-data/netcdf/what-is-netcdf-data.htm>
3. https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_introduction.html
4. <https://code.mpimet.mpg.de/projects/cdo/wiki/Tutorial>
5. <https://code.mpimet.mpg.de/projects/cdo/embedded/index.html#x1-5800002.12>
6. Kull, Meelis: „Scientific Data Formats“, <http://kodu.ut.ee/~varmo/tday-kaariku/kull-slides.pdf>
7. [https://en.wikipedia.org/wiki/Protein_Data_Bank_\(file_format\)](https://en.wikipedia.org/wiki/Protein_Data_Bank_(file_format))
8. https://de.wikipedia.org/wiki/Digital_Imaging_and_Communications_in_Medicine
9. [https://en.wikipedia.org/wiki/List_of_file_formats#Scientific_data_\(data_exchange\)](https://en.wikipedia.org/wiki/List_of_file_formats#Scientific_data_(data_exchange))

6. Literatur

Informationsquellen

10. Schulzweida, Uwe: CDO User's Guide, März 2013
11. Kostede, Simon: Performanceanalyse der Ein- /Ausgabe des Ökologiemodells ECOHAM5, Masterarbeit, Universität Hamburg, S.10 & S. 18-20

6. Literatur

Bildquellen

- [1] <https://intland.com/wp-content/uploads/2016/10/requirements-software-obsolete.png>
- [2] <https://www.ibm.com/it-infrastructure/images/resources/power/IT-infrastructure-HPC-HPDA-figure4.jpg>
- [3] <https://www.researchgate.net/profile/Sebastian-Bathiany/publication/312475518/figure/fig4/AS:451512070610944@1484660274143/Projected-increase-in-the-amount-of-data-from-climate-models-remote-sensing-and-in-situ.png>
- [4] <https://live.osgeo.org/images/netcdf-dataset-screenshot.png>
- [5] <https://pages.uncc.edu/techne/wp-content/uploads/sites/93/2013/12/netcdf.png>
- [6] <https://www.unidata.ucar.edu/software/netcdf/workshops/2011/datamodels/images/nccfile.png>

6. Literatur

Bildquellen

- [7] https://images.slideplayer.com/32/10075718/slides/slide_9.jpg
- [8] <http://desktop.arcgis.com/de/arcmap/10.3/manage-data/netcdf/GUID-41416463-CDBA-4D6F-A9C4-57D745DDE958-web.gif>
- [9] Schulzweida, Uwe: CDO User's Guide, März 2013, Seite 23
- [10] <https://code.mpimet.mpg.de/projects/cdo/embedded/figures/remapbil-.png>