Data management in climate modelling

Iraklion
06-Jul-2017

Hannes Thiemann
Deutsches Klimarechenzentrum (DKRZ)
Outline

- Climate modelling: what is it all about?
- Requirements
- Solutions
- ... and the future
Hannes Thiemann

- Geophysicist
- At DKRZ since 1991
- Member of Data Management department
- Long term archiving
- Joined EUDAT in 2011
DKRZ – our mission

**DKRZ** - to provide high performance computing platforms, sophisticated and high capacity data management, and superior service for premium climate science.

**DKRZ** – to provide a unique combination of world-class computer power and expert personnel to enable superior climate modelling.
Climate vs. weather

- Weather is the state of the atmosphere at a given time and place
  - Cold / warm
  - Wet / dry
  - ...

- Climate is defined as the average of weather over several decades or longer
Climate system

- **SPACE**
  - Net solar (short-wave) radiation
  - Net terrestrial (long-wave) radiation

- **ATMOSPHERE**
  - Clouds
  - Precipitation
  - Absorption
  - Reflection
  - Emission
  - Air-ice interactions
  - Air-ocean interactions
  - Wind

- **OCEAN**
  - Sea-ice
  - Currents
  - Ice-ocean interactions

- **OCEAN**
  - Lakes and rivers

- **GROUND**
  - Snow and ice
  - Runoff
  - Volcanic gases and particles
  - Land surface processes
  - Human activities

- **Max-Planck-Institut für Meteorologie**

Hannes Thiemann
06.07.2017
Model of the atmosphere

- Equations describing the behavior of liquids and gases
- Conservation rates (impulse, energy, mass)
- Radiation
- Boundary conditions (earth geometry, path parameters, land-sea distribution, ...)

Hannes Thiemann
Model of the atmosphere

Equations not directly solvable

\[
\begin{align*}
\frac{\partial U}{\partial t} - (f + \xi) \cdot V + \eta \frac{\partial U}{\partial \eta} + \frac{R_d T_v}{a} \frac{\partial \ln P}{\partial \lambda} + \frac{1}{a \partial \lambda} (\phi + E) &= P_U + K_U \\
\frac{\partial V}{\partial t} + (f + \xi) \cdot U + \eta \frac{\partial V}{\partial \eta} + \frac{R_d T_v}{a} (1 - \mu^2) \frac{\partial \ln P}{\partial \mu} + \frac{(1 - \mu^2)}{a} \frac{\partial (\phi + E)}{\partial \mu} &= P_V + K_V \\
\frac{\partial T}{\partial t} + \frac{U}{\alpha (1 - \mu^2)} \frac{\partial T}{\partial \lambda} + \frac{V}{\alpha} \frac{\partial T}{\partial \mu} + \eta \frac{\partial T}{\partial \eta} - \frac{\kappa T_y \omega}{(1 + (\delta - 1) q_v) P} &= P_T + K_T \\
\frac{\partial q_v}{\partial t} + \frac{U}{\alpha (1 - \mu^2)} \frac{\partial q_v}{\partial \lambda} + \frac{V}{\alpha} \frac{\partial q_v}{\partial \mu} + \eta \frac{\partial q_v}{\partial \eta} &= P_{q_v} + K_{q_v} \\
\frac{\partial q_w}{\partial t} + \frac{U}{\alpha (1 - \mu^2)} \frac{\partial q_w}{\partial \lambda} + \frac{V}{\alpha} \frac{\partial q_w}{\partial \mu} + \eta \frac{\partial q_w}{\partial \eta} &= P_{q_w} + K_{q_w}
\end{align*}
\]
- approximate solutions using iterative procedures on 3D grid
Model resolution

Resolution 500km
Model resolution

Resolution 500km

Resolution 250 km

Tenfold computing time
Model resolution

- **Modell T21**
  - Resolution 500km

- **Modell T42**
  - Resolution 250 km

- **Modell T63**
  - Resolution 180km

- **Modell T106**
  - Resolution 110 km
Tenfold increase of computing time

> sixfold increase of data amount
Data amounts

- **Horizontal grid resolution of the model**
  - **T42**: $128 \times 64 = 8192$ points per global field
  - **T106**: $160 \times 320 = 51200$ points per global field

- **Required storage (GRIB Format)**
  - Horizontal field (**access unit**):
    - **17.1 kB** (T42)
    - **100.1 kB** (T106)
  
  - Filesize for monthly accumulated results with 6 hour storage interval and 300 2d variables (**physical unit**):
    - **616 MB** (T42)
    - **3500 MB** (T106)

  - **240 Jahre model integration (**logical unit**):**
    - **1.7 TB** (T42)
    - **10 TB** (T106)
What have we learned so far:

- What is climate
- What are climate models
- Data amounts
Model evaluation

Climate models must be evaluated very critically by the scientists.

Impact on data management
- Comparison with observational data
- Even more data produced
- More data collected at selected sites
CMIP (Coupled Model Intercomparison Project)

- CMIP has led to a better understanding of past, present and future climate change and variability. Since 1995 CMIP has developed in phases.
- Generates input for Intergovernmental Panel on Climate Change (IPCC)
CMIP5

- Generates input for Intergovernmental Panel on Climate Change, Fifth Assessment Report (IPCC-AR5)

- Approx. 20 Experiments
- 20 modelling centres

- Output consists of 4.2 Petabytes organized in a few million files
CMIP6

• Generates input for Intergovernmental Panel on Climate Change, Sixth Assessment Report (IPCC-AR6)
  – Handful of common experiments
  – Ensemble of CMIP-Endorsed Model Intercomparison Projects (MIPs)
CMIP6

Eyring et al.
CMIP6

- Output consists of some tens of petabyte of data.
- Organized in approx. 250 Mio files.
Challenges

- Replication
- Archiving
- Access
- Versioning
- Long-Term
- Disks
- Standardisation
- Citation
- Metadata
- Selection
- Network
- High-Availability
- Backup
- Money
- Rejection
- Governance
Climate research

Typical questions in climate research

- At the end of the century: by how much will global temperature differ from the one at the beginning of the century.
- What is the probability that the winter in Iraklion in 2050 will be drier than the average winter these days.
Interdisciplinary scientific usage

Impacts on data management

- Different presentation of data needed
- Longer data availability
- Support
Society, politics and business

Impact on data management

- Different presentation of data needed
- Support
- Transparency of processes required
Paradigm shift

Climate model output transformed into community resources.

Implications (e.g.)

- Equal and easy access to large disparate datasets.
- Enable scientists to evaluate models through a common interface regardless of the data’s location.
- Research landscape is distributed
- Data volumes too large to be managed at single site
- Data is used interdisciplinary
- Data becomes community resource.
Infrastructures

- Scope
- Sustainability
- Governance
- Technical implementation
- Business model
ESGF Comparison to Other Archives

<table>
<thead>
<tr>
<th>NOMADS</th>
<th>DAACs</th>
<th>Google Earth Engine</th>
<th>ESGF</th>
</tr>
</thead>
</table>

- **Data Management**
- **Distributed Search**
- **Federation**
- **Analysis and Visualization**
- **Provenance Capture**
- **Security**
- **Network**
- **Compute Facilities**
- **Dynamic Resource**
- **Data Transfer**
- **Long-Tail Publication**
- **Data Citation**
- **Machine Learning**

Legend:
- Current Capabilities
- Future Capabilities
ICON

- able to conduct production simulations on target grids as large as $10^{10}$ grid elements
- grid spacing finer than 400 m (100 m being the target)
- capable of efficiently using a diversity of advanced high performance computing resources
- scaling of the model to many tens, possibly hundreds of thousands of cores
Data amounts

GRIB Format, $10^{10}$ grid elements

- Horizontal field (access unit):
  - 17.1 kB (T42)
  - 100.1 kB (T106)
  - 10 GB ($10^{10}$)

- Filesize for monthly accumulated results with 6 hour storage interval and 300 2d variables (physical unit):
  - 616 MB (T42)
  - 3500 MB (T106)
  - 360 TB ($10^{10}$)

- 240 years model integration (logical unit):
  - 1.7 TB (T42)
  - 10 TB (T106)
  - 1 EB ($10^{10}$)
Thank you.