What can climate science bring to eScience?

Wilco Hazeleger
The setting
Tools: climate data sources

- Observed proxy records
- Instrumental in situ data
- Remotely sensed data
- Reanalyzed data from numerical weather models
- Climate model data

Easily more than 100 Pbyte/yr
Climate change past 1000 years

(c) Overlap of reconstructed temperatures
Other relevant data for climate research

• Socio-economic data
  – Population density
  – Emissions of greenhouse gases, aerosols
  – Land use
  – Sectoral data: health, energy, agriculture, mobility,…

• Biogeochemical data
  – Carbon, Nitrogen, Phosphorous
  – Vegetation types
  – Marine species distribution

Typical volume much smaller than climate system data, but very diverse
Tools: numerical climate models

Current resolutions: 100 km and 60 layers in the vertical

~1-10 million lines of code
Petascale hardware

Momentum equations

\[
\frac{\partial U}{\partial t} + \frac{1}{a \cos^2 \theta} \left\{ U \frac{\partial U}{\partial \lambda} + V \cos \theta \frac{\partial U}{\partial \theta} \right\} + \eta \frac{\partial U}{\partial \eta} = +fV - \frac{1}{a} \left\{ \frac{\partial \phi}{\partial \lambda} + RT_v \frac{\partial \ln p}{\partial \lambda} \right\} + P_U + K_U
\]

\[
\frac{\partial V}{\partial t} + \frac{1}{a \cos^2 \theta} \left\{ U \frac{\partial V}{\partial \lambda} + V \cos \theta \frac{\partial V}{\partial \theta} + \sin \theta (U^2 + V^2) \right\} + \eta \frac{\partial V}{\partial \eta} = -fU - \frac{\cos \theta}{a} \left\{ \frac{\partial \phi}{\partial \theta} + RT_v \frac{\partial \ln p}{\partial \theta} \right\} + P_V + K_V
\]

Sub-grid model: "physics"
Numerical diffusion
Principle of a climate simulation

- Present state
- Specify external forcings according to the date
- Model calculates rate of change (tendency)
- State one timestep (30 minutes) later

Sectoral and local climate impacts

Solar radiation
Greenhouse gases
Land use
Aerosols
Data assimilation (fusion)
Scenarios of future CO$_2$ emissions

Moss et al, Nature, 2010
Effect of air pollution policies (2050 vs 2000)
Numerical models: regional impacts

forecast.ewatercycle.org

vd Giessen, Drost et al
Numerical models: High resolution

Nastrom & Gage 1985; Hazeleger et al 2012
Numerical models: ensembles

Selten et al, pers comm.
Numerical model: ensembles
eScience and climate research

Data models
Integrate datasets, analytics

Data-intensive processing
Data from sensors and numerical models (high throughput, large volume)

Numerical models
High resolution, large ensembles, coupled models

Analytics platforms
DBMS network
HPC
Climate model data

Each center multiple Petabytes of data
How to deal with large amounts of data?
(www.climate4impact.eu)

_Taking the processing to the portal – KNMI_

A federated system optimises the use of the limiting resource: _people_.
_No institution can go it alone: data at scale is a global activity based around large national facilities._...
Increasing resolution

IFS Evolution

Resolution halves about every eight years

Courtesy George Mozdzynski, ECMWF
# Increasing resolution

<table>
<thead>
<tr>
<th>IFS model resolution</th>
<th>Envisaged Operational Implementation</th>
<th>Grid point spacing (km)</th>
<th>Time-step (seconds)</th>
<th>Estimated number of cores¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1279 H²</td>
<td>2013 (L137)</td>
<td>16</td>
<td>600</td>
<td>2K</td>
</tr>
<tr>
<td>T2047 H</td>
<td>2014-2015</td>
<td>10</td>
<td>450</td>
<td>6K</td>
</tr>
<tr>
<td>T3999 NH³</td>
<td>2023-2024</td>
<td>5</td>
<td>240</td>
<td>80K</td>
</tr>
<tr>
<td>T7999 NH</td>
<td>2031-2032</td>
<td>2.5</td>
<td>30-120</td>
<td>1-4M</td>
</tr>
</tbody>
</table>

¹ – a gross estimate for the number of ‘IBM Power7’ equivalent cores needed to achieve a 10 day model forecast in under 1 hour (~240 FD/D), system size would normally be ~10 times this number.

2 – Hydrostatic Dynamics

3 – Non-Hydrostatic Dynamics

George Mozdzynski
Climate models on new architectures

Van Werkhoven, Dijkstra et al
Data and climate research

- 100s of Pbytes, distributed, is the norm
- Few producers & many users with different needs & data used for many years
- Accessing, distributing and analyzing the data is the problem (not the generation of it)

- bring compute to the data (limited success so far: scalability, data security, multi-site storage at ESG)

- institutional arrangements needed
HPC and climate

- Higher spatial resolution to resolve relevant scales
- Ensembles to cover parameter space
- Increasing complexity and fusion of data

$\Rightarrow$ Toward exascale: $10^6$ cores in near future easily reached (now $O(10^5)$)

But:
- Load imbalances
- Memory use and bandwidth (internal and external)
- Fault tolerance
- Energy efficiency (Flops/Watt) and wall clock time
- Scalable I/O
eScience and Climate Research

• Tackle the end to end problem!
  – Deal with HPC & Data!
  – The workflow is dominated by data distribution and analysis; less the numerical model
  – Few institutions do the computational work, most climate scientists are data analysts, but they rely on the few data producers

• Manpower lacks to re-design for exascale era and the word is out how to do it

• Let’s learn from each other and look for overarching technologies and solutions
Ample opportunities for eScience in climate research!*

Thank You

* Keep it simple