Weather and Climate Computing Futures in the context of European Competitiveness

Bryan Lawrence
NCAS

... ten minutes, not enough time to cover ICT issues specific to predictability of weather, but they’re essentially the same as the climate problem with the added requirement of “timeliness” putting pressure on networks, data acquisition, computing and analysis ...
Outline

Ten minutes covering:

(1) Context, why we care (competitiveness), and an intro to models.
(2) Objectives, numbers, strategic objectives.
(3) One real case study.
(4) Conclusions: competitiveness requires European scale infrastructure, that is, computing AND networks targeted at data analysis as well as data production!

Warning: some subliminal slides, but I'll slow down for the slides of greatest import for this workshop!
Disease vectors and Water Security?

How will climate change affect the global distribution of malaria?

How will climate change affect the frequency of drought conditions and hence security of water supply and biological diversity?

Drought, Floods, or both?
Landslides and transport systems

How will climate change affect the incidence of road and rail closures due to landslides?

How can network and transport design be improved to adapt to environmental change?
Extreme winds and rainfall: Extratropical Cyclones

July 2007 Tewkesbury flood: 3B€ loss

Jan 2007 Windstorm Kyrill: 6B€ loss

Expensive occurrences!

Some local, some global events but European exposure more or less wherever these events occur (e.g. the insurance industry)!

How will the frequency, intensity, and location of these events change in the future?
Types of models: “Global Climate Model” (GCM)

Fully Coupled.

All components interact via two-way fluxes of relevant quantities.

Image: from J. Lafeuille, 2006
The world in global climate models
What can we afford (technically)?

(Many versions of this slide exist, this one from J. Kinter's presentation to the world modelling summit 2008)
And the science chooses?

<table>
<thead>
<tr>
<th>WEATHER AND CLIMATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPLICATION</td>
</tr>
<tr>
<td>CLIMATE CHANGE</td>
</tr>
<tr>
<td>OCEANOGRAPHY and MARINE FORECASTING</td>
</tr>
<tr>
<td>METEOROLOGY, HYDROLOGY and AIR QUALITY</td>
</tr>
</tbody>
</table>

… progress down all axes: complexity, resolution, bigger ensembles for longer, more and better evaluation (greater role for EO and data assimilation), all of which are DATA INTENSIVE!
# All that computation, All that Data!

<table>
<thead>
<tr>
<th>Key numbers for Climate Earth System Modelling</th>
<th>2012</th>
<th>2016</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horizontal resolution of each coupled model component (km)</td>
<td>125</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>Increase in horizontal parallelisation wrt 2012 (hyp: weak scaling in 2 directions)</td>
<td>1</td>
<td>6.25</td>
<td>156.25</td>
</tr>
<tr>
<td>Horizontal parallelization of each coupled model component (number of cores)</td>
<td>1,00E+03</td>
<td>6,25E+03</td>
<td>1,56E+05</td>
</tr>
<tr>
<td>Vertical resolution of each coupled model component (number of levels)</td>
<td>30</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>Vertical parallelization of each coupled model component</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Number of components in the coupled model</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Number of members in the ensemble simulation</td>
<td>10</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Number of models/groups in the ensemble experiments</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Total number of cores (4x6x7x8x9) (Increase:)</td>
<td>8,00E+04</td>
<td>1,00E+06</td>
<td>1,56E+09</td>
</tr>
<tr>
<td>Data produced (for one component in Gbytes/month-of-simulation)</td>
<td>2,5</td>
<td>26</td>
<td>1302</td>
</tr>
<tr>
<td>Data produced in total (in Gbytes/month-of-simulation)</td>
<td>200</td>
<td>4,167</td>
<td>1,302,083</td>
</tr>
</tbody>
</table>

(Adapted from the PRACE scientific update case, courtesy of Sophie Valcke)
A European Infrastructure

Recommendations:

1. Provide a blend of HPC facilities ranging from national machines to a world class computing facility suitable for climate applications, which, given the workload anticipated, may well have to be dedicated to climate simulations.

2. Accelerate the preparation for exascale computing, e.g. by establishing closer links to PRACE and by developing new algorithms for massively parallel many-core computing.

3. Ensure data from climate simulations are easily available and well documented, especially for the climate impacts community.

4. Build a physical network connecting national archives with transfer capacities exceeding Tbits/sec.

5. Strengthen the European expertise in climate science and computing to enable the long term vision to be realized.

The future requires cooperation at the European level!!
The nation should (9 bullet points, precise for this meeting):
1. Evolve to a common national software infrastructure that supports a diverse hierarchy of different models for different purposes …
2. Convene … forum … promotes tighter coordination and more consistent evaluation …
3. Nurture a unified weather-climate modeling effort …
5. Sustain the availability of state-of-the-art computing systems for climate modeling
8. Enhance the national and international IT infrastructure that supports climate modeling data sharing and distribution
Recommendation 8:

Growth rate of climate model data archives is exponential, and maintaining access to this data is a growing challenge!

... the climate research community and decision makers and other user communities desire to analyse and use (simulation and observational) data in increasingly sophisticated ways.

These two trends imply growth in resource demands that cannot be managed in ad-hoc way. Instead

Data-sharing infrastructure ... should be systematically supported as an operational backbone for climate research and serving the use community.

Without substantial research effort into new methods of storage, data dissemination, data semantics and visualization, all aimed at bringing analysis and computation to the data, rather than trying to download the data and perform analysis locally, it is likely that data might become frustratingly inaccessible to users!

J T Overpeck et al. Science 2011;331:700-702
Politics (this slide verbatim from breakout at Exascale meeting 2011)

The exascale data handling problem is not just about the lack of s/w, it's also about sustained s/w investment. European initiatives are often not sustained long enough to be competitive with other (American, and probably soon, Chinese) offerings.

Data life cycle is expected to be long. Analysis s/w will need to have longevity. 3-5 year funding life cycle is not representative of the reality of big data handling.

Need to ask the question as to whether European funding can be better spent enhancing existing (foreign) s/w (which has sustained investment) rather than building products aimed at being competitive (but without sustained investment).

− Invest in collaboration, rather than competition may yield better results?
UPSCALE

Pl: Vidale (NCAS)

Project: 150 million core hours at HERMIT! (~20% of the machine for a year!)

Some have advertised it as the largest single compute project ever!

But it's not just the largest single “Compute” project ever, it's one of the largest ever distributed data analysis problems ever!

As of last week, UPSCALE had moved roughly 250 TB of data via GEANT to the JASMIN super data cluster in the UK, half of which had been moved on to the UK Met Office! Running at an average of 1 TB/day, peaking at 6 TB/day! (I.e sustaining over ~ Gbit/s for many hours at a time)
Tropical Cyclone Tracks: Transits per month

Roberts & Vidale et al
Tropical Cyclone Tracks: Transits per month

HodGEM3–A GA3.0
1985–1995

HodGEM3–A GA3.0
1979–2009

Track Density from obs and reanalyses (transits/month)

Roberts & Vidale et al
Handling UPScale: JASMIN

JASMIN*: Joint Analysis System (NERC+Met Office)

4.6 PB (usable, 6.6 raw) fast disk (Panasas)

Connected to petascale tape store

350+ cores compute, configured so that I/O is very very far from saturation!

Support for parallel data analysis via both batch and interactive jobs.

Support for remote and local virtualisation via VMware.

Connected to SuperJanet at 10 Gbit/s for general traffic, plus additional light paths to UK HPC and Dutch clients.

Much of the JASMIN data store is already committed (e.g. CMIP5 archive, 1-3 PB!), but UPScale data will be archived and analysed on JASMIN.

First example of JASMIN data analysis advantage:

- seasonal cyclone tracking analysis job for 25km run (running on 7 months of data) used to take 56 hours wall clock (longer than the model took) … now takes 22 hours … and we still haven’t looked at parallelisation.

* JASMIN is deployed alongside CEMS, and these are the specs of the complete system. See http://arxiv.org/abs/1204.3553
Dedicated Light Paths to HPC and users!

Moving away from download everywhere! Service provision includes (or will include): JASMIN analysis systems, and KNMI portals!

1 and 2 Gbit/s now, 10 Gbit/s Planned!

... but even JASMIN is a tiny step towards what we will need to cope with exascale climate science!
Summary

Weather and Climate computing continue to need tier-0 computing resources (as well as tier-1, 2, 3 et al).
- it's possible such tier-0 computing will need to be configured specially for this application domain!

Data archive and analysis is becoming more and more of a bottleneck, hindering both science, and application in commerce and policy.
- the UK example suggests topic specific archive and analysis systems as well!

We need to go beyond shared networks like Geant and support topic specific dedicated networks (a la CERN).
- On demand lightpaths?

We need to continue to invest in shared software development but consider hard issues of
- sustainability and
- how to ensure delivery of topic specific archival services which are pan European.

We need to implement the ENES foresight strategy!