

CMIP6: Data and Computing Issues

GO-ESSP 2015

Cosener's House, Abingdon UK

V. Balaji

NOAA/GFDL and Princeton University

26 February 2015

Outline

- 1 Global data infrastructure and the WIP
- 2 Computational constraints
- 3 Analysis on distributed archives
- 4 GFDL perspective on CMIP6
- 5 Summary

The global data infrastructure underpinning MIPs

- MIPs, and in general any science involving cross-model comparisons, critically depend on the global data **infrastructure** – the “vast machine” (Edwards 2010) – making this sort of data-sharing possible.
- Infrastructure should not be a research project.
- Infrastructure should be treated as such by the national and international research agencies, but it is instead funded piecemeal, as a soft-money afterthought. This places the system at risk (NRC 2012: “A National Strategy for Advancing Climate Modeling”, ISENES-2 Infrastructure Strategy document, 2012.)

Role of WGCM and its infrastructure panel

- Provide scientific guidance and requirements for the GDI; exert greater influence over its design and features.
- Provide standards governance allowing for orderly evolution of standards.
- Provide design templates (e.g CMOR extensions) for groups designing MIPs and work to ensure their conformance to standards.
- Work with academies and publishers to require adequate data citation and recognition for data providers.
- Intercede with national agencies to provision data infrastructure with adequate and stable long-term funding.

We expect this to be a non-trivial commitment of time and effort by Panel members.

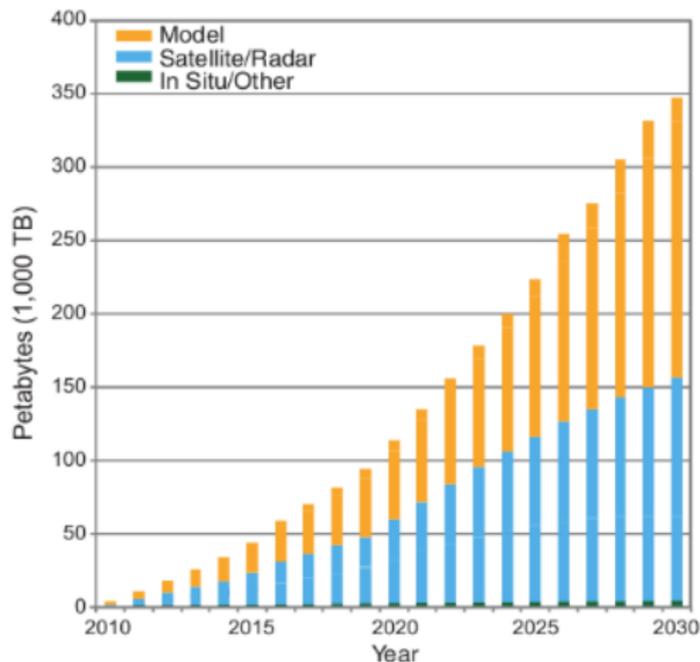
Acknowledgements: Proposal initially prepared by V. Balaji and Karl Taylor, with input and revisions made by co-authors Eric Guilyardi, Michael Lautenschlager, Bryan Lawrence, and Dean Williams.

WIP: The WGCM Infrastructure Panel

- Chaired by V. Balaji (Princeton/GFDL) and K. Taylor (PCMDI).
- Strategy to develop a series of "position papers" on global data infrastructure and its interaction with the scientific design of experiments. These will be presented to WGCM annual meeting.
 - **protocol document** for the "endorsed MIPs" **delivered**. Working with CMIP panel and MIP sponsors on CMIP6 data request.
 - **data access** policies: would open access simplify the technical design of the infrastructure?
 - **data citations**. Developing and promoting a path to data citations using DOIs and the emerging data journals, such as ESSD, Nature Scientific Data.
 - projected **data volumes** for CMIP6, strategies for managing the growth path
- Infrastructure issues that impinge on science design for CMIP6 will be handled through close involvement of the WIP and CMIP panel (e.g. joint papers)
- Interest from other WCRP working groups! (WGSIP, WGNE)

CMIP6: data explosion?

Overpeck et al. (2011), *Science* forecast a 100-fold increase in data volume over 10 years.



Complexity, resolution, ensemble size

Computational increases can be applied along 3 axes: resolution, complexity, ensemble size.

- **resolution**: where an N^3 growth in computing is applied to (x, y, t) leading to only N^2 growth in archive in (x, y) : thus $A \sim C^{\frac{2}{3}}$!
- **complexity**, as new subsystems and feedbacks are added to comprehensive earth system models;
- **UQ**, as we build ensembles of simulations to sample uncertainty, both in our knowledge and representation, and of that inherent in the chaotic system. In particular, we are interested in characterizing the "tail" of the PDF (weather extremes) where a lot of climate risk resides.

Analysis of GFDL models: results

Model	Resolution	Cmplx.	SYPD	CHSY	Coupler	Load Imb.	I/O	MBloat	ASYPD
CM2.6 S	A0.5L32 O0.1L50	18	2.2	212,465	5.71%	20%		12%	1.6
CM2.6 T	A0.5L32 O0.1L50	18	1.1	177,793	1.29%	60%	24%	12%	0.4
CM2.5 T	A0.5L32 O0.25L50	18	10.9	14,327	17%	0%			6.1
FLOR T	A0.5L32 O1L50	18	17.9	5,844	0%	57%	5.1%	31%	12.8
CM3 T	A2L48 O1L50	124	7.7	2,974	0.5%	41%	14.76%	3%	4.9
ESM2G S	A2L24 O1L50	63	36.5	279	8.91%	1%		34%	25.2
ESM2G T	A2L24 O1L50	63	26.4	235	2.63%	22%		34%	11.4

- More details are available (layout on MPI/thread, aggregate I/O per CH or SD, platform, optimization, cost per component...)
- Is this a basis for a cross-model comparison of performance (CPMIP, anyone?) for a common understanding of the roadblocks to performance?

Preliminary cross-model comparisons

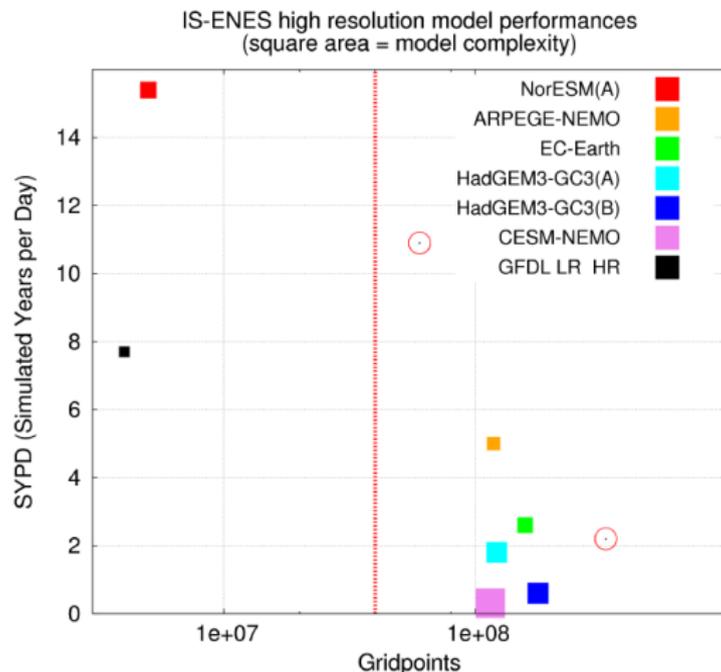
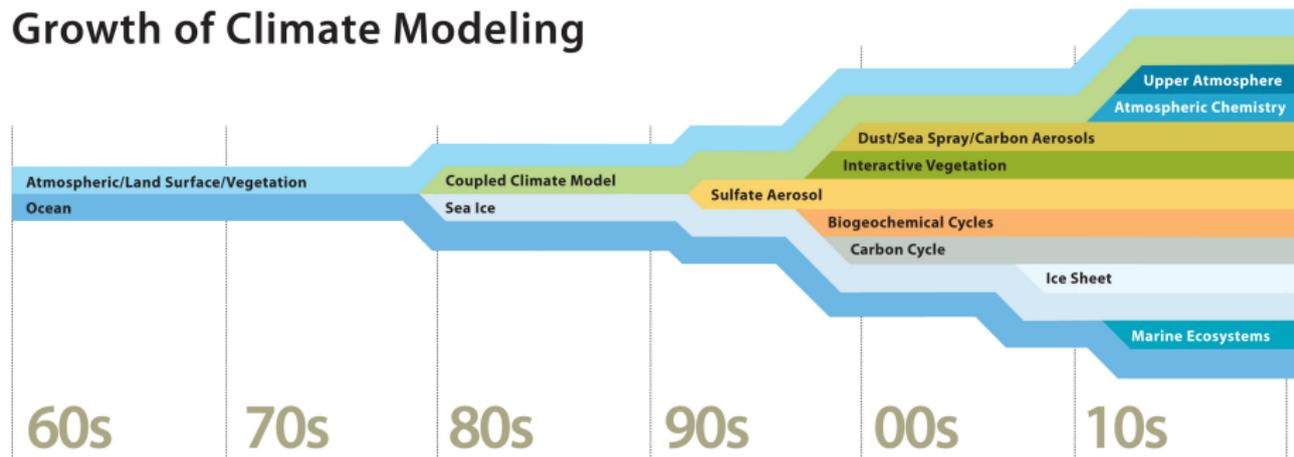


Figure courtesy Eric Maisonnave, Joachim Biercamp, Giovanni Aloisio and others on the ISENES2 team.

Complexity in ESMs

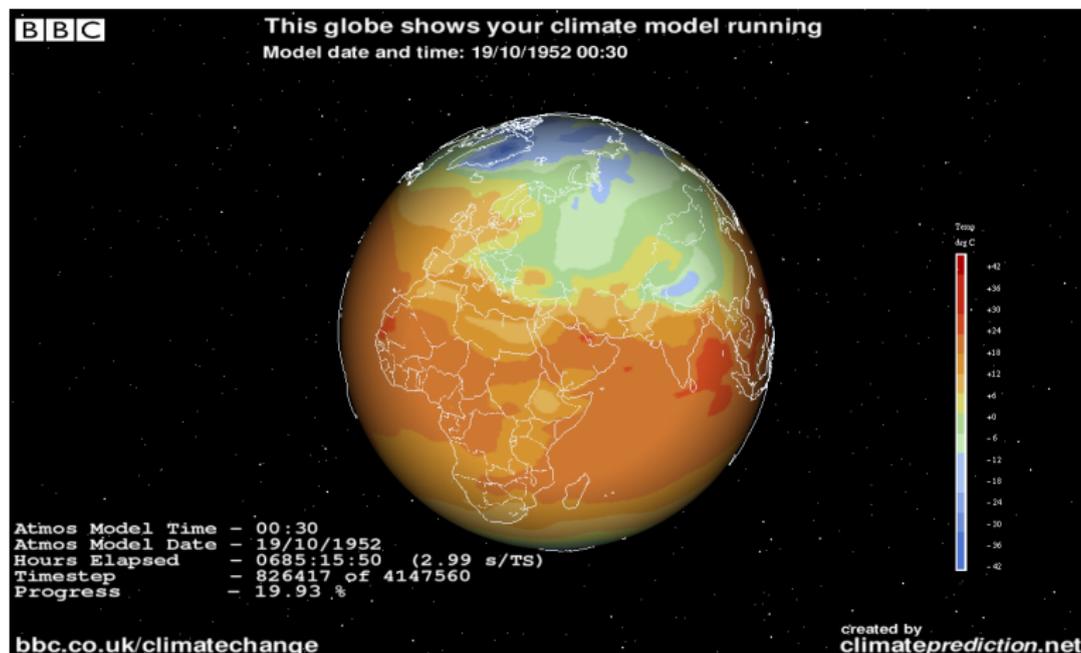
Growth of Climate Modeling



Note slope is “piecewise-constant” over 1-2 CMIP cycles!

Figure courtesy UCAR Climate FAQ.

Uncertainty: large ensembles?



Extreme value analysis can require very large ensembles ($N \sim 1000$) but CMIP6 design uses comprehensive ESMs and $N \sim 3 - 10$.
Figure courtesy climateprediction.net.

ExArch: Climate analytics on distributed exascale data archives

Martin Juckes, V. Balaji, B.N. Lawrence, M. Lautenschlager, S. Denvil, G. Aloisio, P. Kushner, D. Waliser, S. Pascoe, A. Stephens, P. Kershaw, F. Laliberte, J. Kim, S. Fiore

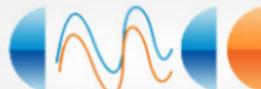
UCLA



UNIVERSITY OF
TORONTO



Princeton
University



Centro Euro-Mediterraneo
per i Cambiamenti Climatici

The G8 Exascale Research Initiative

- A joint initiative by research councils of Canada, France, Germany, Japan, Russia, UK, and USA;
- Research into exploitation of exascale computational resources;
- Focus on 10-year time horizon;
- A unique opportunity for funded international collaboration;
- But with restrictions – only funding to 7 participating countries; restricted eligibility within participating countries;

Bringing analysis to data: what's involved

The projected frequency of intense tropical cyclones in some region of the globe for input into an impacts model?

- Query execution:
- evaluation of provenance and quality control meta-data to determine which datasets to include;
- dispatch of queries to processing nodes, negotiating authentication and access control layers;
- Key issue is the use of **user-developed** analytic scripts;
- collection of results from the processing nodes, evaluation of return codes for fault detection;
- further calculations to combine collected results;
- archive results for re-use; delivery of processed results to the end-user, perhaps in deferred fashion if the associated computation needs to be scheduled on a "cloud".

Waiting to see if Phase II will be funded.

GFDL perspectives on CMIP6

- We will tailor models to the available computing: models are analyzed for both capacity (CHSY) and speed (SYPD) and computing set aside for all CMIP6 runs (DECK and MIPs).
- We expect standards (variable lists, CMOR, DRS) to be frozen 15 July 2015 per current schedule from CMIP panel. Necessary for building our CMIP6 workflow.
- We expect limited direct use of ESGF by scientists: since CMIP5 we've maintained an internal mirror, and built a toolset to populate it (based on **synchro_data** from IPSL). Current size is 330 TB.
- Theoretically could be ready to begin runs in early 2016.

Summary

- CMIP6 probably will see uptick in number of models, resolution: may be flat in complexity and ensemble size.
- We do not expect a huge increase in data volumes, as between CMIP3 and CMIP5, perhaps 2-5X. This estimate is disputed: we will publish the data volume estimate later in 2015.
- Full data request and associated standards and conventions expected by mid-2015. An important issue under consideration is defining the separate needs of “community” users and “downstream” users: e.g an archive of high-value variables on a common grid and calendar.
- User feedback indicates that an analysis framework for distributed archives is still not clearly in view.
- ESGF governance is still not in place.
- Considerations of a “CMIP6 Operations Team” are underway.
- WIP meets at GOESSP-2015.