G8 ExArch: Climate analytics on distributed exascale data archives

Investigators:

M. Juckes, STFC, UK (PI)	G. Aloisio, CMCC
Balaji, Princeton/GFDL	B. Lawrence, STFC, UK
M. Lautenschlager, DKRZ	P. Kushner, Toronto
S. Denvil, IPSL	D. Waliser, UCLA/JPL

Support:

ANR, DFG, NSERC/CRSNG, NSF, RCUK

Presented by Paul Kushner (paul.kushner@utoronto.ca)
Earth, Atmospheric and Planetary Physics
Department of Physics
University of Toronto

Outline

- Introduction

  Exascale data challenge of computational climate science.

  ExArch goals and research team

- Approach and activities

  Discussion of ExArch components (server side processing, etc.).

- Conclusion

  Summary, referring to G8 Agencies points of inquiry.

Introduction

Computational climate science is synchronized with *IPCC Assessment Reports* (**AR**) and *Coupled Model Intercomparison Projects* (**CMIP**) for Earth System Models (**ESM**).

Computational climate science is synchronized with *IPCC Assessment Reports* (**AR**) and *Coupled Model Intercomparison Projects* (**CMIP**) for Earth System Models (**ESM**).

Over the last 30 years, CMIP ESMs have become more realistic, . . .

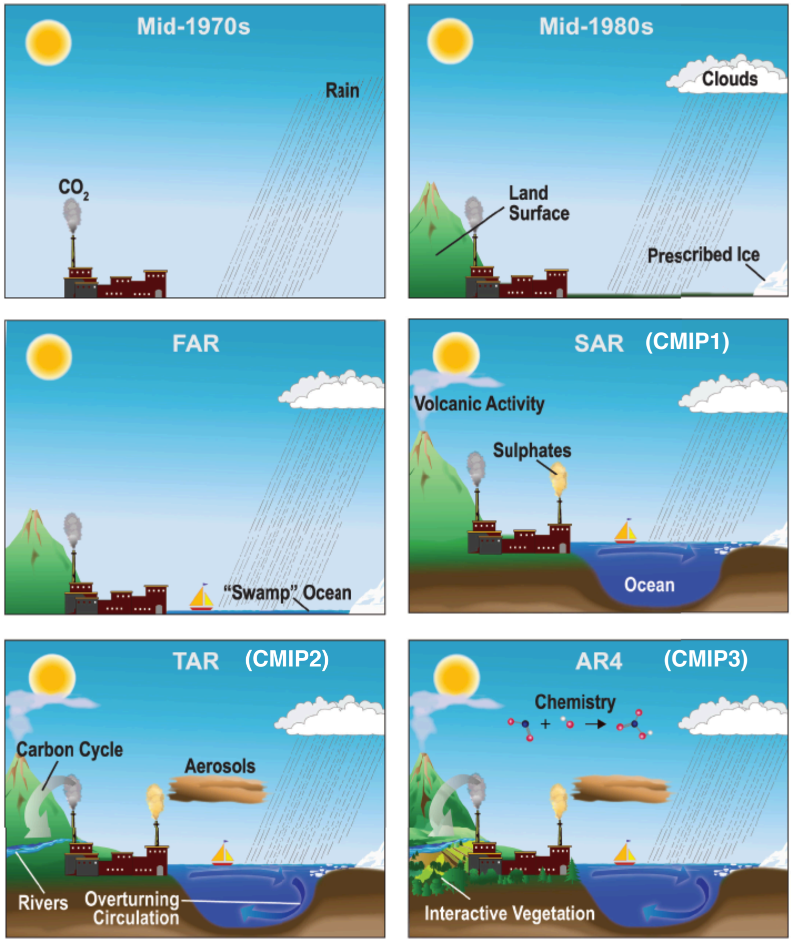


Figure 1.2. *The complexity of climate models has increased over the last few decades. The additional physics incorporated in the models are shown pictorially by the different features of the modelled world.*

Computational climate science is synchronized with *IPCC Assessment Reports* (**AR**) and *Coupled Model Intercomparison Projects* (**CMIP**) for Earth System Models (**ESM**).

Over the last 30 years, CMIP ESMs have become more realistic, . . .

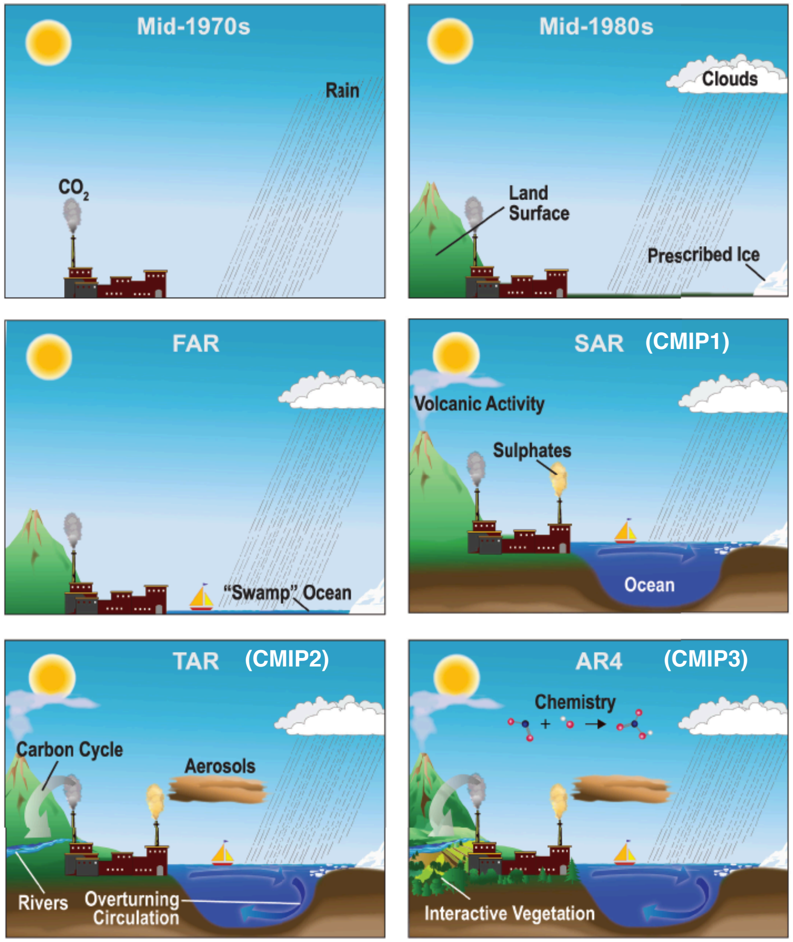
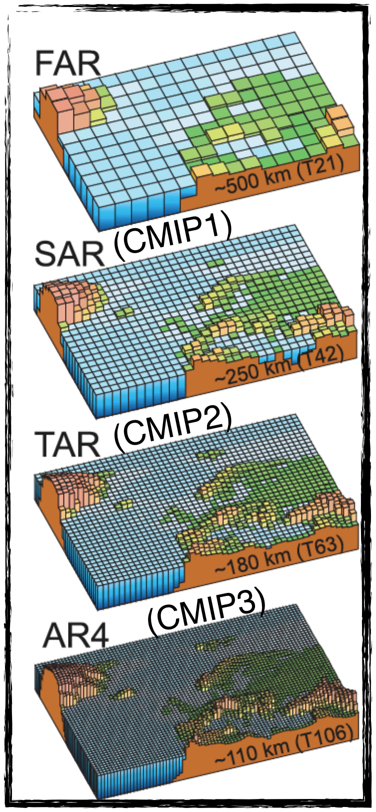


Figure 1.2. *The complexity of climate models has increased over the last few decades. The additional physics incorporated in the models are shown pictorially by the different features of the modelled world.*

. . . better resolved . . .



Computational climate science is synchronized with *IPCC Assessment Reports* (**AR**) and *Coupled Model Intercomparison Projects* (**CMIP**) for Earth System Models (**ESM**).

Over the last 30 years, CMIP ESMs have become more realistic, . . .

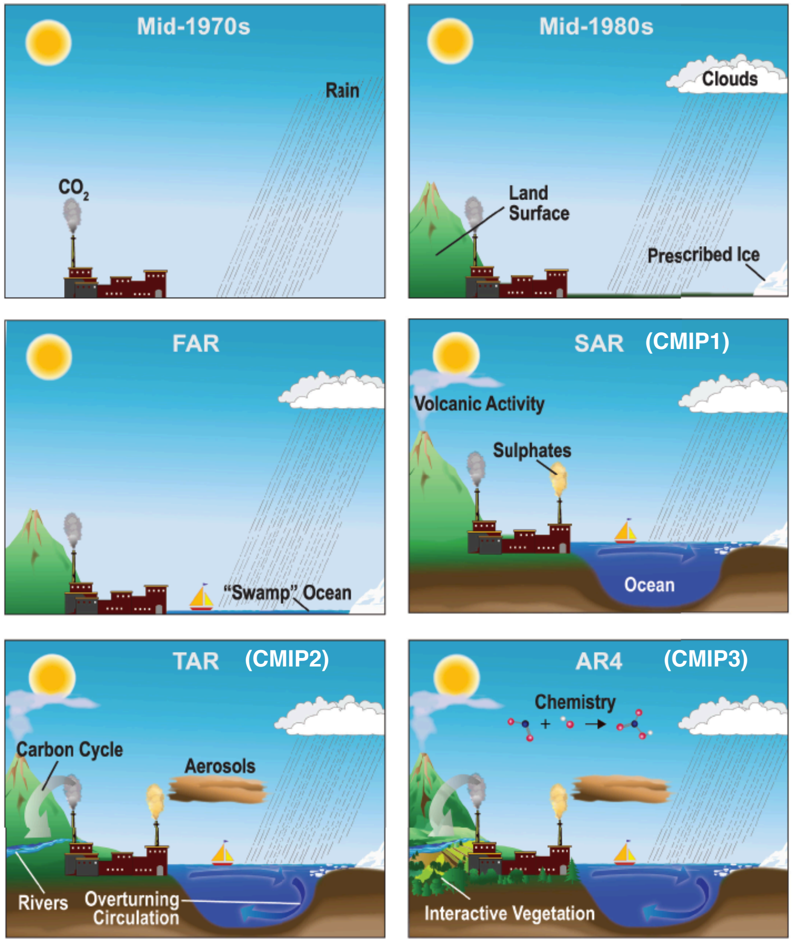
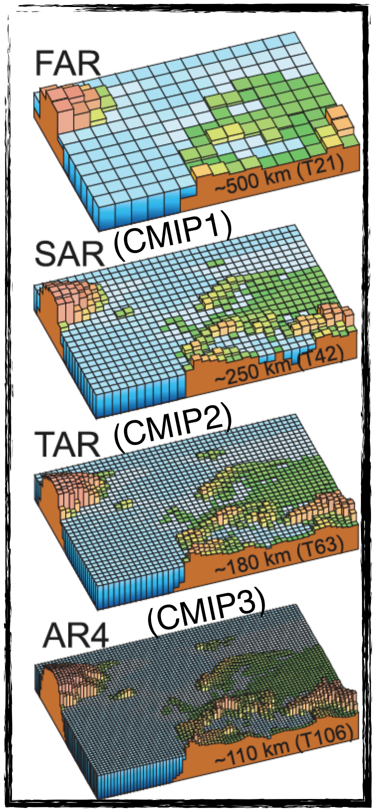
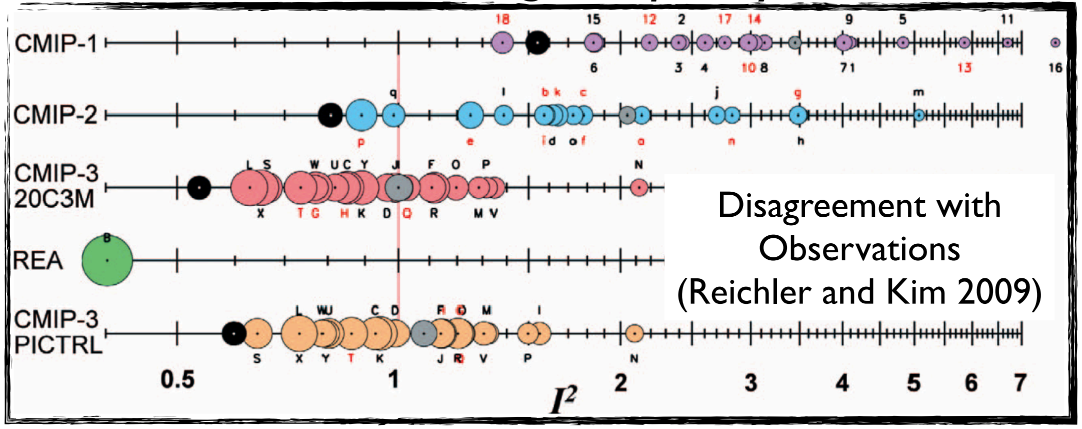


Figure 1.2. *The complexity of climate models has increased over the last few decades. The additional physics incorporated in the models are shown pictorially by the different features of the modelled world.*

. . . better resolved . . .



. . . and higher quality.



Disagreement with Observations (Reichler and Kim 2009)

Computational climate science is synchronized with *IPCC Assessment Reports* (**AR**) and *Coupled Model Intercomparison Projects* (**CMIP**) for Earth System Models (**ESM**).

Over the last 30 years, CMIP ESMs have become more realistic, . . .

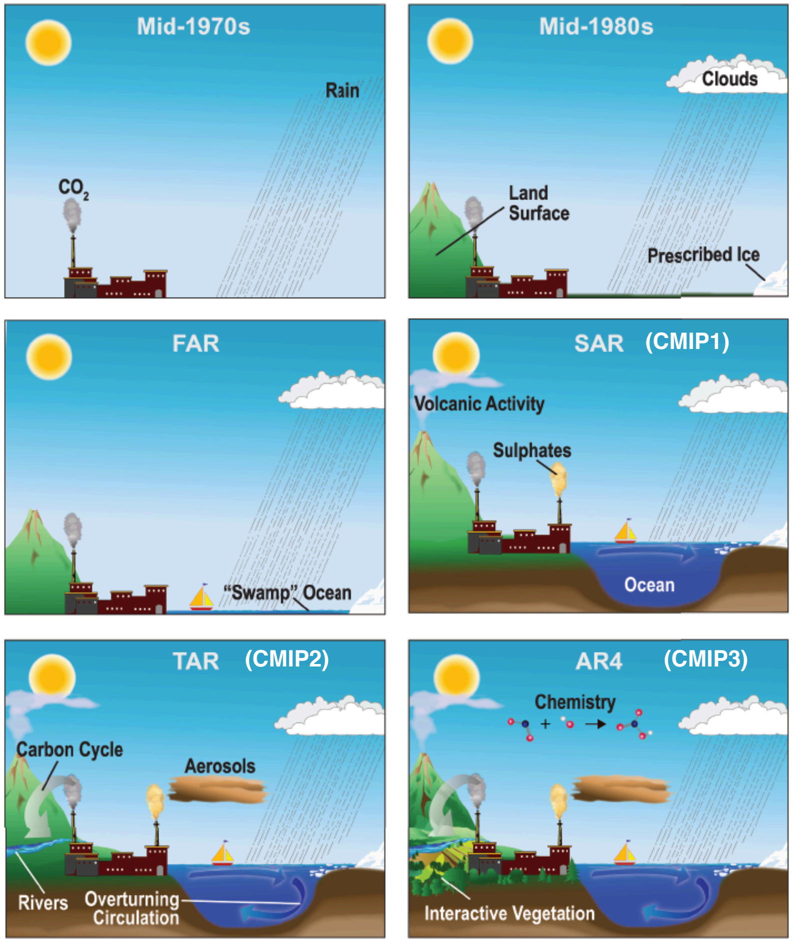
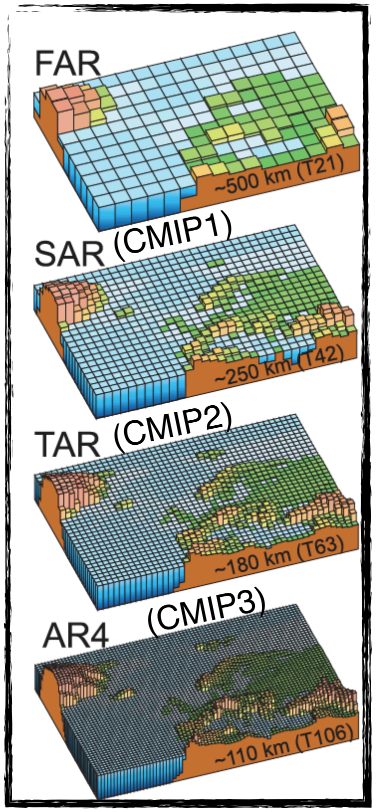


Figure 1.2. *The complexity of climate models has increased over the last few decades. The additional physics incorporated in the models are shown pictorially by the different features of the modelled world.*

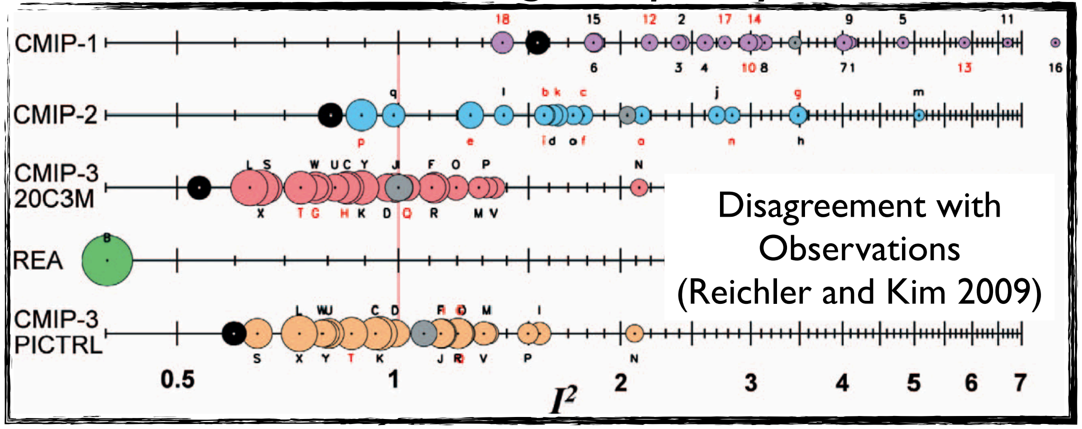
. . . better resolved . . .



Accelerating model development and applications pose an immediate data challenge.

This challenge provides a preview of the future as computing moves to exascale.

. . . and higher quality.



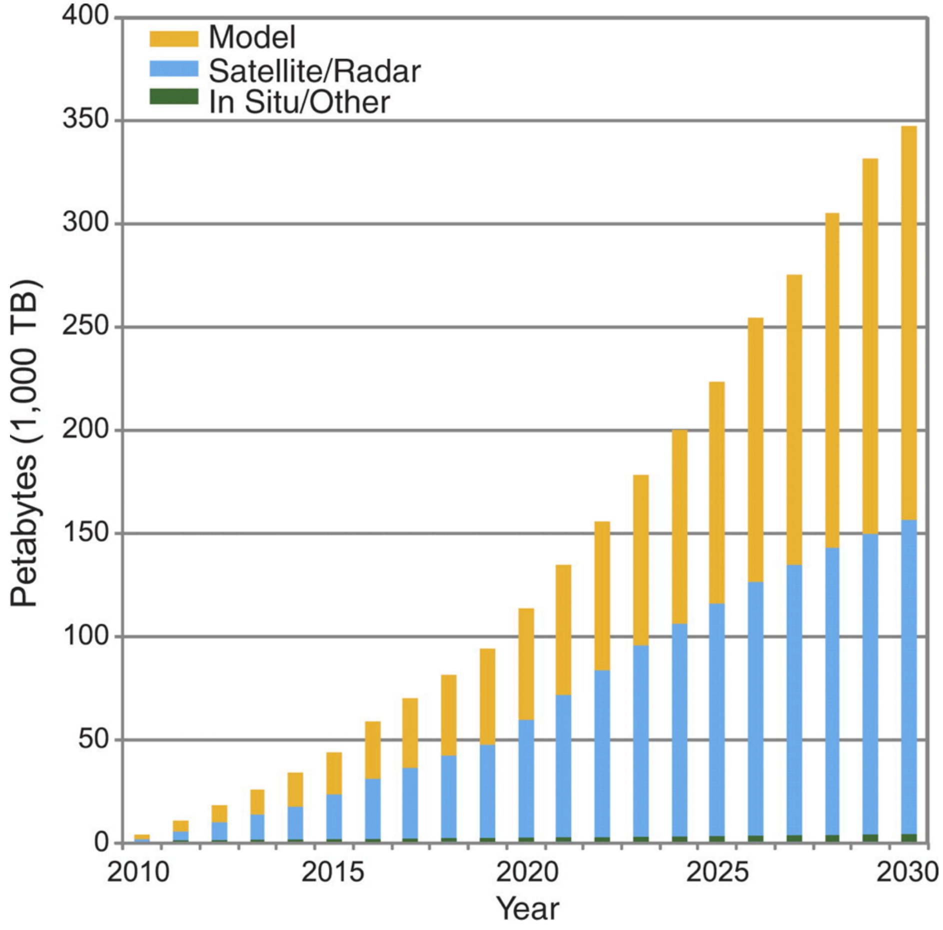
Disagreement with Observations (Reichler and Kim 2009)

From CMIP3 to CMIP5, and Beyond

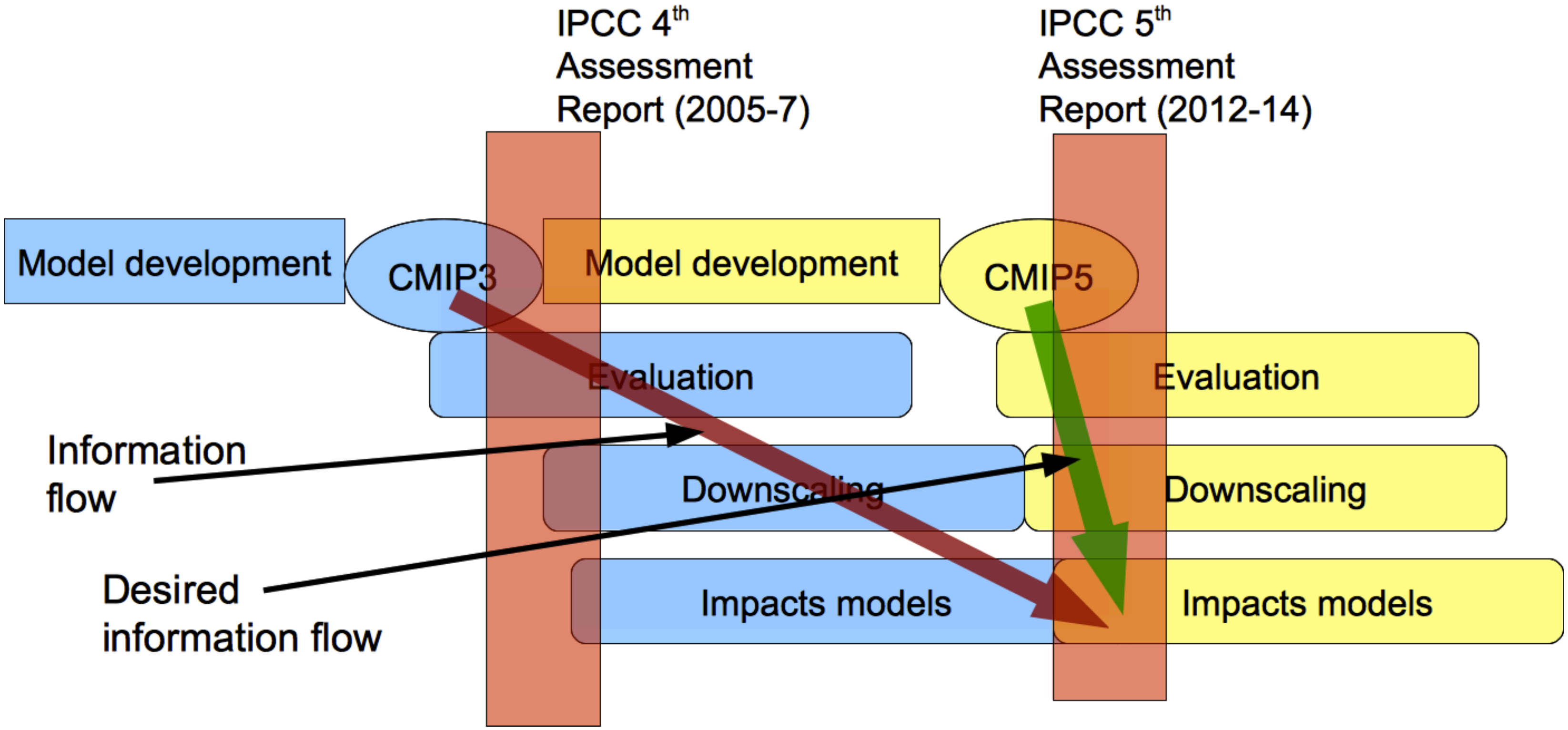
- **CMIP3** (IPCC AR4, 2007):
  - About **18** models.
  - Experiments: historical, future projections.
  - **40 TB** of data.
  - Stored at single site (PCMDI/LLNL).
  - Users download from there.
- **CMIP5** (IPCC AR5, 2013/2014):
  - About **58** models.
  - Experiments: historical, future projections, near-term forecasts, regional experiments (**CORDEX**).
  - Projected **10 PB** of data.

➡ ***How do we deal with all this data?***

Worldwide Climate Data Volume



Need to Accelerate Analysis Cycle



- Data flow outstripping available human resources for analysis.
- To understand local impacts of climate change, downscaled *impacts* models of hydrology and water resources, ecology, agriculture, etc. are used.
- But CMIP3 impacts modelling has lagged CMIP5 development and production cycle.

➡ ***How can we reduce this lag and make climate analysis easier?***

Distributed Climate Analytics: Overarching Goals

- Provide scaleable, fast, open, user friendly access to international ESM archives.
- Minimize data movement across networks.
- Aim for efficient, reproducible, well documented, easily updated, and easily archived climate and impacts analysis.
- The U.S. NRC Panel Report "A National Strategy for Climate Modelling" identifies this effort as an urgent technical and science priority. (View is shared by DOE, NOAA, NASA, NSF, UK CEDA, EC FP7, WMO, and other agencies.)

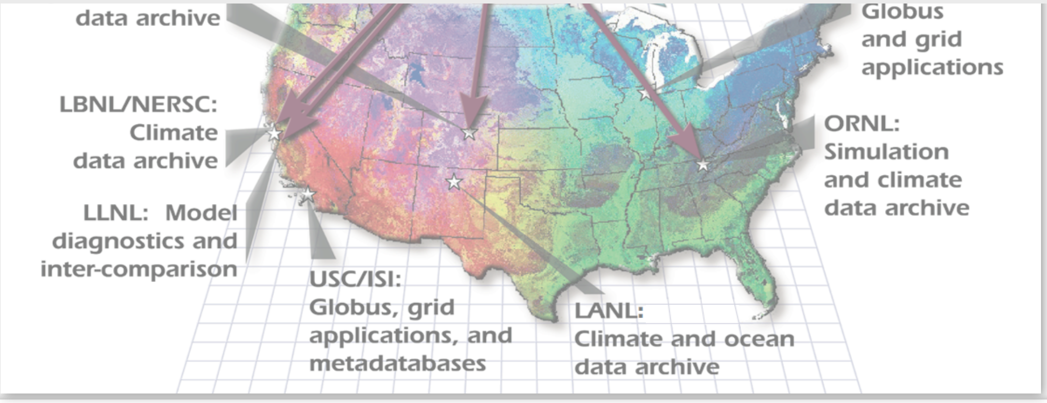
ExArch Context

- Earth System Grid Federation (**ESGF**) is a group of archive service providers and software developers, developing and deploying a robust, distributed data and computation platform using open source software. ESGF has no core funding. It is a federated international effort coordinated by multiple agencies.

The Earth System Grid Federation (ESGF) is the world’s source for climate science data

- ESGF is a free, open consortium of institutions, laboratories and centers around the world that are dedicated to supporting research of Climate Change, and its environmental and societal impact
- Historically originated from Earth System Grid (ESG) project, expanded beyond its constituency and mission to include many other partners in the U.S., Europe, Asia, and Australia
- Groups working at many projects: ESG, Earth System Curator, Metafor, Global Interoperability Program, Infrastructure for

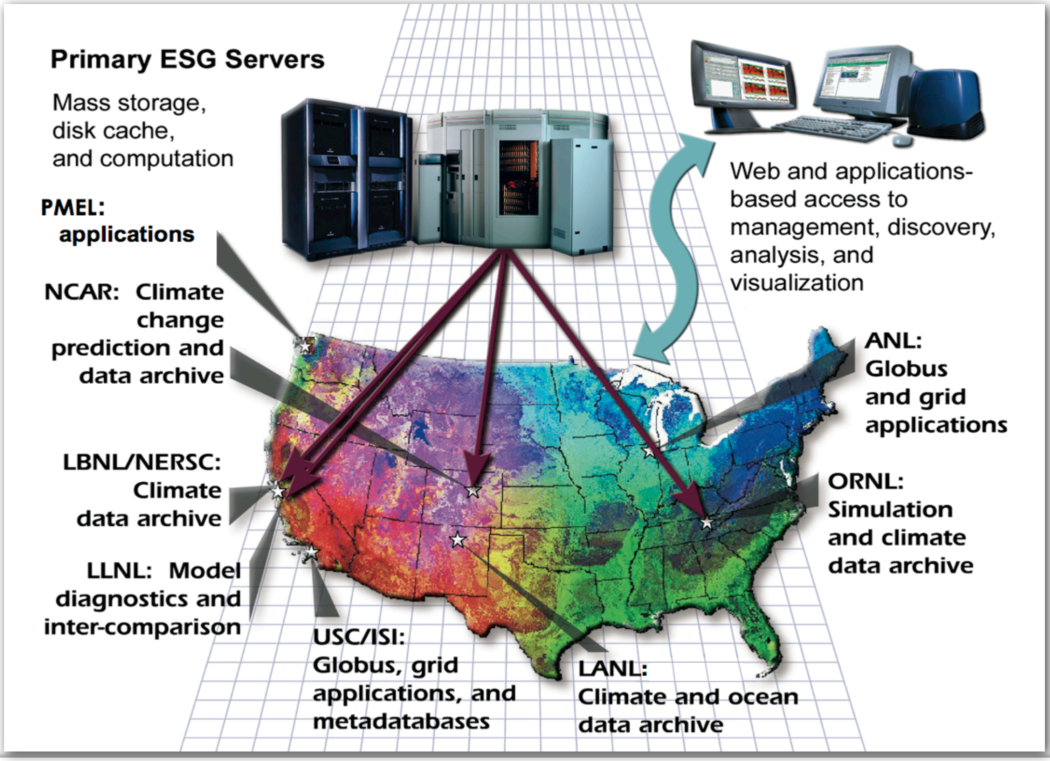
From ESGF overview presented at e-infrastructure 2011


data archive
LBNL/NERSC: Climate data archive
LLNL: Model diagnostics and inter-comparison
USC/ISI: Globus, grid applications, and metadatabases
LANL: Climate and ocean data archive
ORNL: Simulation and climate data archive
Globus and grid applications

- Australia: ANU, Australian Research Collaboration Service, Government Department of Climate Change, Victorian Partnership for Advanced Computing, Australian Environment and Resource Management
- ... and many more ...

The Earth System Grid Federation (ESGF) is the world's source for climate science data

- ESGF is a free, open consortium of institutions, laboratories and centers around the world that are dedicated to supporting research of Climate Change, and its environmental and societal impact
- Historically originated from Earth System Grid (ESG) project, expanded beyond its constituency and mission to include many other partners in the U.S., Europe, Asia, and Australia
- Groups working at many projects: ESG, Earth System Curator, Metafor, Global Interoperability Program, Infrastructure for the European Network for Earth System Modeling, and many more
- U.S. funding from DOE, NASA, NOAA, NSF


Primary ESG Servers
Mass storage, disk cache, and computation
PMEL: applications
NCAR: Climate change prediction and data archive
LBNL/NERSC: Climate data archive
LLNL: Model diagnostics and inter-comparison
USC/ISI: Globus, grid applications, and metadatabases
LANL: Climate and ocean data archive
ORNL: Simulation and climate data archive
ANL: Globus and grid applications
Web and applications-based access to management, discovery, analysis, and visualization

ESG Federation

- U.S.: ANL, ESRL, GFDL, JPL, LANL, LBNL, LLNL/PCMDI, NCAR, ORNL, PMEL, USC/ISI
- Europe: BADC, UK-MetOffice, DKRZ, MPIM, IPSL, LSCE
- Asia: Univ. of Tokyo, Japanese Centre for Global Environmental Research, Jamstec, Korea Meteorological Administration
- Australia: ANU, Australian Research Collaboration Service, Government Department of Climate Change, Victorian Partnership for Advanced Computing, Australian Environment and Resource Management
- ... and many more ...

ExArch Context

- Earth System Grid Federation (**ESGF**) is a group of archive service providers and software developers, developing and deploying a robust, distributed data and computation platform using open source software. ESGF has no core funding. It is a federated international effort coordinated by multiple agencies.
- **IS-ENES1** (2009 – 2013) and **IS-ENES2** (2014 – 2018) are European Union FP7 projects providing integrative research and service activities for the European Earth System Modelling community, including support for distribution of the CMIP5 archive through ESGF.
- The G8 Initiative's support of **ExArch** has provided a critical opportunity to develop a long-range, internationally coordinated strategy to address the goals and requirements of exascale climate analysis.

ExArch: Key Goals

- Exploit CMIP5 experience at petabyte scale so we can use exabyte archives in the next decade.
- Leverage ongoing software engineering efforts in this area.
- Deliver real solutions to research challenges in climate analytics encountered by climate science community.
- Bridge IS-ENES and ESGF communities.
- Develop infrastructure to provide timely, efficient, scalable, resilient, transparent access to geographically distributed archives on heterogeneous platforms.
- Deal with both CMIP5 global Earth System Model output as well as CORDEX regional model output.

Research Team

- ExArch team has been centrally involved in building climate archives at the petascale.
- Our team includes key partners in IS-ENES (Juckes, Lawrence, Denvil, Lautenschlager)
- It also includes on the team and Advisory Board key architects of ESGF and GO-ESSP (Balaji, Williams).
- ExArch makes link to Earth Observations through Waliser (NASA/JPL) and to climate research analytics through Kushner.
- Advisory board includes leaders in WMO World Climate Research Program (WCRP).

Approach and Activities

ExArch Organization

- **Work Package 1 (Juckes): Management and strategy development.** To develop strategy for global infrastructure to support exascale climate analytics.
- **Work Package 2 (Balaji): Informatics Research.** Development of software to support:
    - Server side resource management
    - Collection and distribution of metadata.
    - Complex scientific queries across distributed archive that can be re-used and chained.
    - Efficient distribution of data.
    - Interfaces ensuring security, transparency, and interoperability with Earth Observation archives.
- **Work Package 3 (Kushner): Climate science and scientific quality control.** Prototype real solutions with real scientific benefit, with reference to CMIP5, CORDEX, and available Earth Observations.

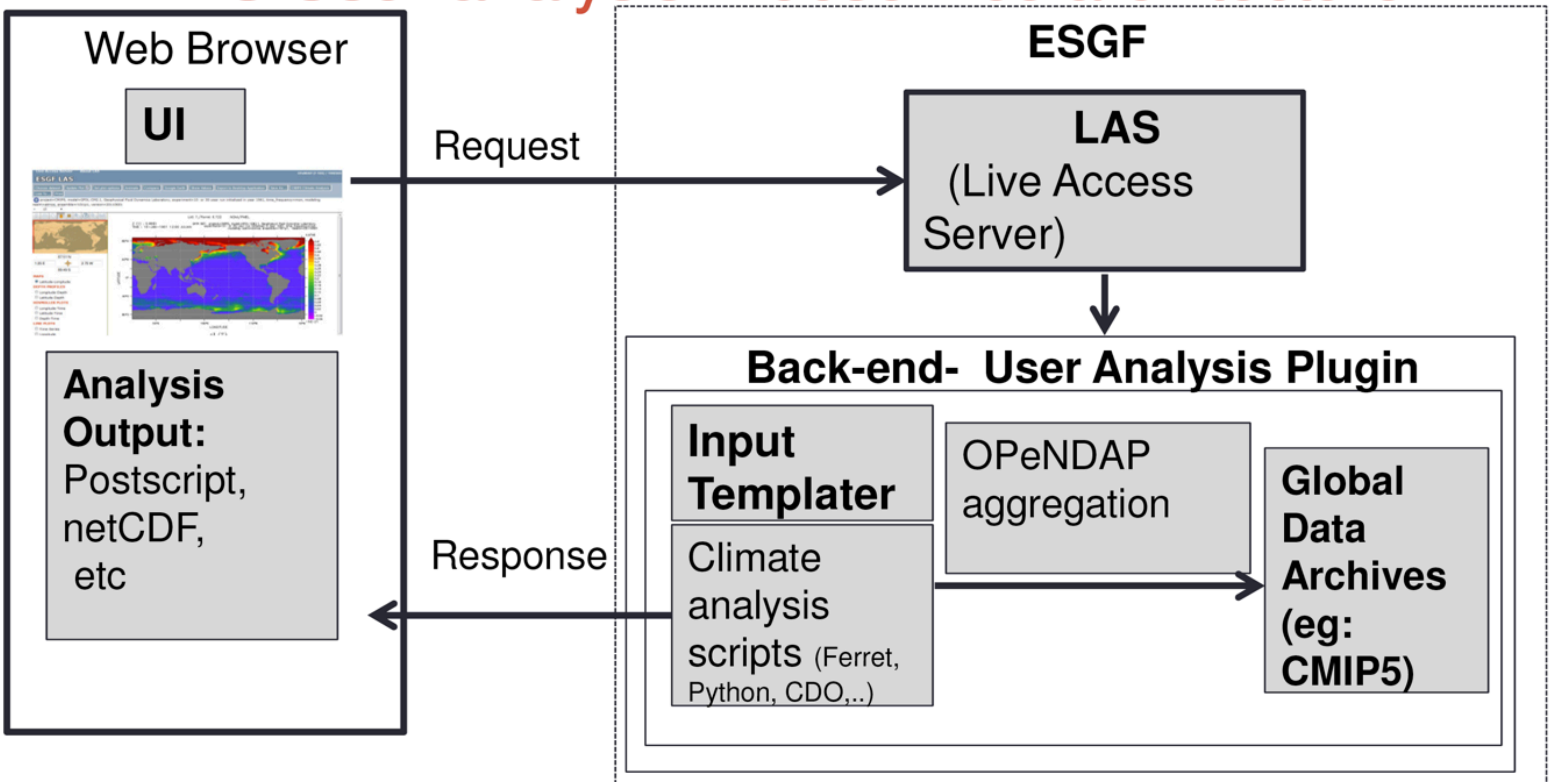
ExArch Components

- Web processing services, query syntax
- Common information model/structured metadata
- Processing operators, quality control
- Scientific diagnostics, Earth Observation data for model evaluation

Server Side Processing

- Minimize data movement by carrying out data reduction and processing near the data server.
- ExArch seeks to leverage work of the Open Geospatial Consortium (**OGC**).
- Builds upon experience with UK Climate Projections Portal, and with FP7 IS-ENES User Interfaces (Climate4Impact)
- ExArch aims to extend UK CEDA OGC Web Service (**COWS**; Stephens and Kershaw).
- ExArch plans:
    - Extend COWS to enable exascale climate data workflow for remote or local data.
    - Develop request syntax that specifies data files, spatio-temporal domain and operators (e.g. MathML).

LAS User-analysis Webservice architecture



Web Browser

UI

Analysis Output: Postscript, netCDF, etc

Request

Response

ESGF

LAS (Live Access Server)

Back-end- User Analysis Plugin

Input Templater

Climate analysis scripts (Ferret, Python, CDO,..)

OPeNDAP aggregation

Global Data Archives (eg: CMIP5)

A. Radhakrishnan

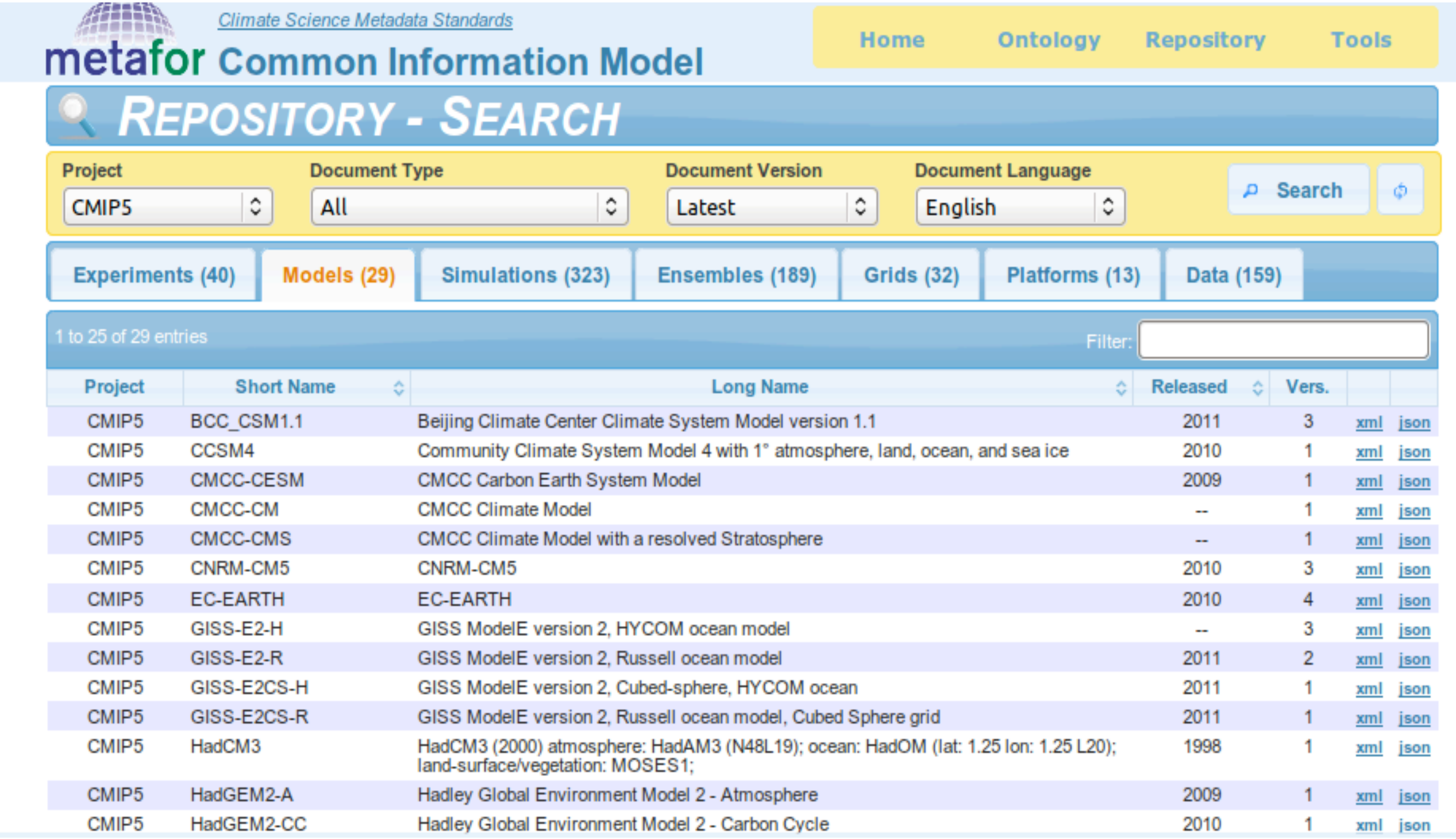
ExArch has also developed WPS link between ESGF and the NOAA Live Access Server (LAS) working with OPeNDAP.

Structured Metadata

CMIP5 with ESGF has pioneered use of structured metadata in ESMs:

- CMIP5 requires data and metadata standards: NetCDF, CF conventions, THREDDS XML catalog based descriptions of data, etc..
- ESM information entered through an on-line questionnaire, over 800 questions (METAFOR). This information is used as part of a three level quality control process.

es-doc.org :: interface to the metadata repository



Project: CMIP5
Document Type: All
Document Version: Latest
Document Language: English

1 to 25 of 29 entries

Project	Short Name	Long Name	Released	Vers.		
CMIP5	BCC_CSM1.1	Beijing Climate Center Climate System Model version 1.1	2011	3	xml	json
CMIP5	CCSM4	Community Climate System Model 4 with 1° atmosphere, land, ocean, and sea ice	2010	1	xml	json
CMIP5	CMCC-CESM	CMCC Carbon Earth System Model	2009	1	xml	json
CMIP5	CMCC-CM	CMCC Climate Model	--	1	xml	json
CMIP5	CMCC-CMS	CMCC Climate Model with a resolved Stratosphere	--	1	xml	json
CMIP5	CNRM-CM5	CNRM-CM5	2010	3	xml	json
CMIP5	EC-EARTH	EC-EARTH	2010	4	xml	json
CMIP5	GISS-E2-H	GISS ModelE version 2, HYCOM ocean model	--	3	xml	json
CMIP5	GISS-E2-R	GISS ModelE version 2, Russell ocean model	2011	2	xml	json
CMIP5	GISS-E2CS-H	GISS ModelE version 2, Cubed-sphere, HYCOM ocean	2011	1	xml	json
CMIP5	GISS-E2CS-R	GISS ModelE version 2, Russell ocean model, Cubed Sphere grid	2011	1	xml	json
CMIP5	HadCM3	HadCM3 (2000) atmosphere: HadAM3 (N48L19); ocean: HadOM (lat: 1.25 lon: 1.25 L20); land-surface/vegetation: MOSES1;	1998	1	xml	json
CMIP5	HadGEM2-A	Hadley Global Environment Model 2 - Atmosphere	2009	1	xml	json
CMIP5	HadGEM2-CC	Hadley Global Environment Model 2 - Carbon Cycle	2010	1	xml	json

Model categories, based on CIM metadata

Atmosphere-Ocean Models:	
Atmosphere, Land, Ocean, Sea ice, Aerosol;	CCSM4, HadCM3, GFDL-CM2p1
Atmosphere, Land, Ocean, Sea ice;	CMCC-CM, EC-Earth
Atmosphere, Ocean, Sea ice;	CMCC-CMS
Coupled-chemistry models:	
Sea ice, Aerosol, Atmospheric Chemistry;	GISS-E2H/E2CS-R
Atmosphere, Land, Ocean, Sea ice, Aerosol, Atmospheric Chemistry;	GFDL-CM3
Earth System Models:	
Atmosphere, Land, Ocean, Sea ice, Ocean Bio-geochemistry;	IPSL-CM5A-LR/MR, MPI-ESM-LR/MR/P, GFDL-ESM2G/M
Atmosphere, Land, Ocean, Sea ice, Aerosol, Atmospheric Chemistry, Ocean Bio-geochemistry;	HadGEM2-ES/CC

Structured Metadata

CMIP5 with ESGF has pioneered use of structured metadata in ESMs:

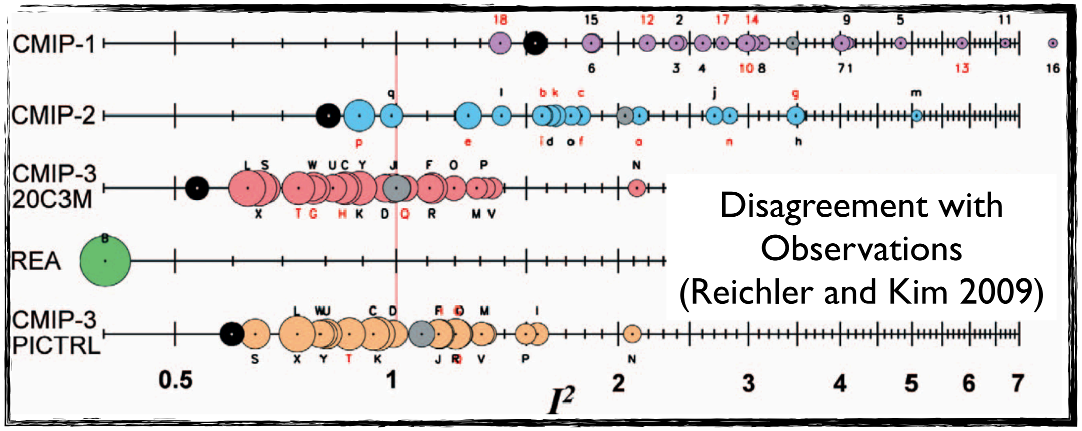
- CMIP5 enforces data and metadata standards: NetCDF, CF conventions, THREDDS XML catalog based descriptions of data, etc..
- ESM information entered through an on-line questionnaire, over 800 questions (METAFOR). This information is used as part of a three level quality control process.

ExArch is leveraging this work to develop:

- Quality control to meet user and software client requirements, and structured descriptions of this QC at multiple levels.
- Extension to regional models in CORDEX.
- Direct generation of metadata from climate models, and transformation from metadata to model configuration files and back.
- Extensions and interoperability with Earth Observation metadata.

Scientific Diagnostics

- Climate science community in CMIP3 has found enormous value in model intercomparison.

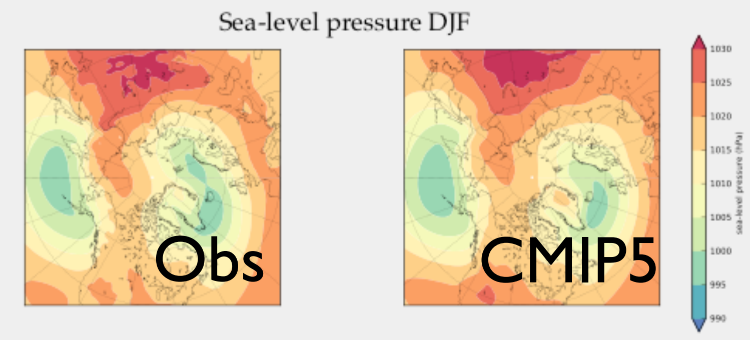


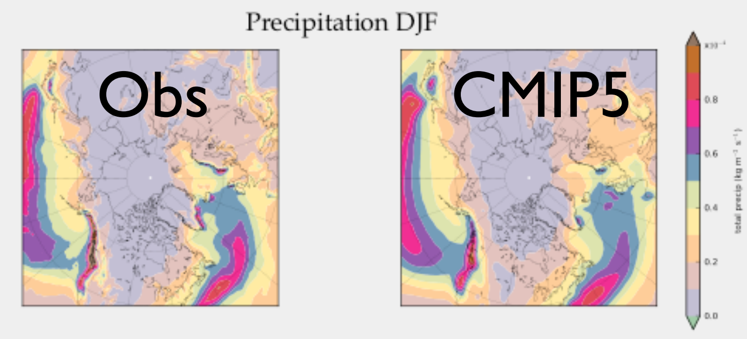
Disagreement with Observations (Reichler and Kim 2009)

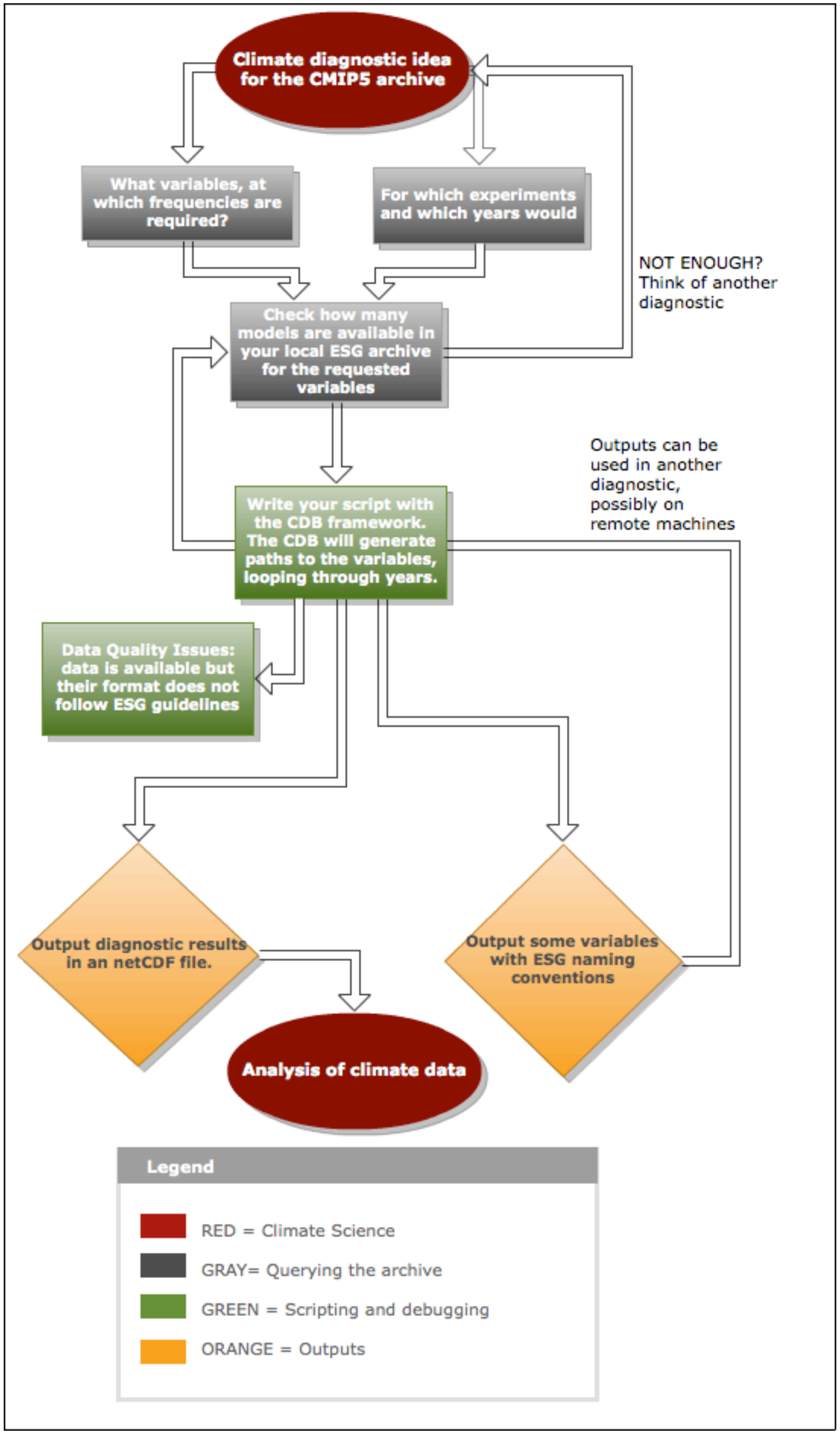
- But downloading even a small fraction of CMIP5 data is impractical, and need for distributed processing is immediate.
- ExArch is developing benchmarks --- target workflows -- which software engineers can optimize and use to test WPS, query syntax, quality control, etc..

Climate Diagnostics Benchmark (**CDB**)

- v0.1, bash based, released this year (PDF F. Laliberté)
- New python based release coming soon.
- Applied to extratropical cyclones, snow albedo feedback analysis.
- Coordinated with WCRP WGNE Quick Metrics (Gleckler).
- We are a couple of steps away from WPS.


Sea-level pressure DJF


Precipitation DJF


Climate diagnostic idea for the CMIP5 archive

What variables, at which frequencies are required?

For which experiments and which years would

Check how many models are available in your local ESG archive for the requested variables

NOT ENOUGH? Think of another diagnostic

Write your script with the CDB framework. The CDB will generate paths to the variables, looping through years.

Outputs can be used in another diagnostic, possibly on remote machines

Data Quality Issues: data is available but their format does not follow ESG guidelines

Output diagnostic results in an netCDF file.

Output some variables with ESG naming conventions

Analysis of climate data

Legend

🟥 RED = Climate Science

⬛ GRAY= Querying the archive

🟩 GREEN = Scripting and debugging

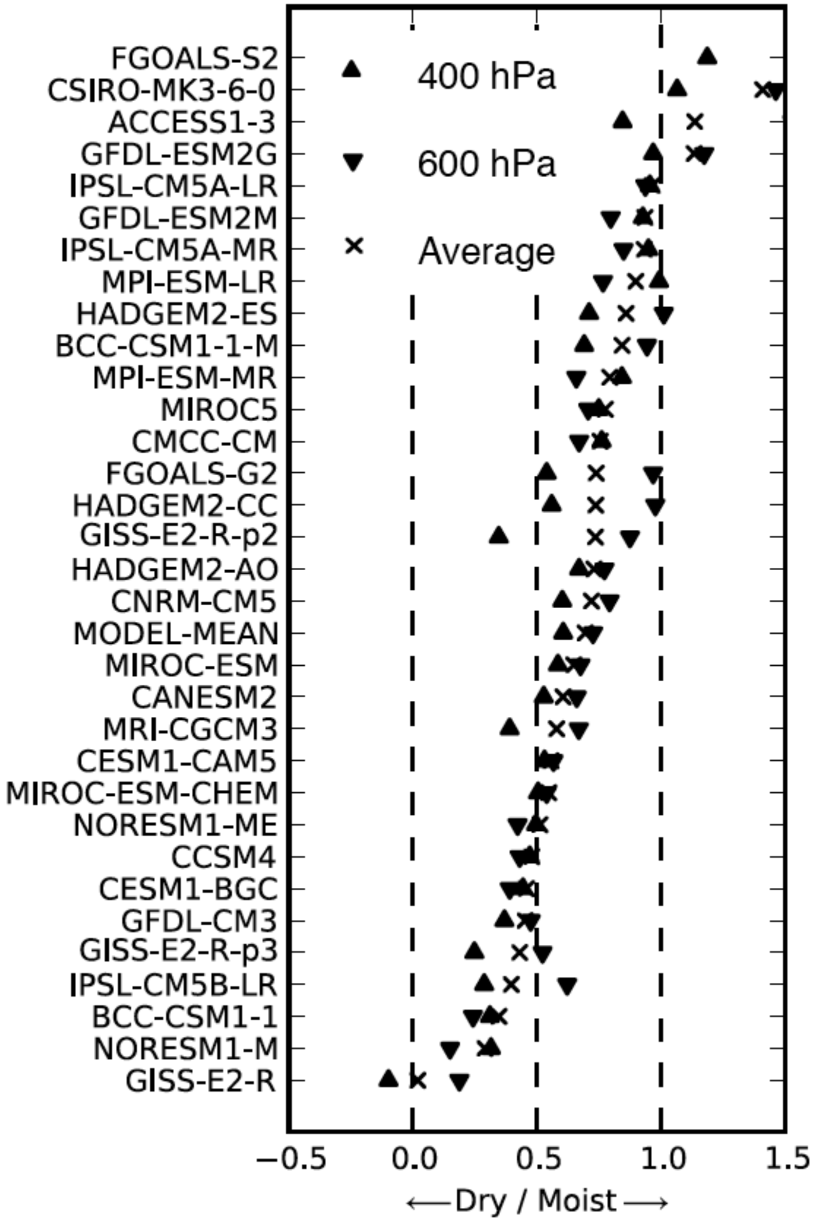
🟧 ORANGE = Outputs

Early G8 ExArch Science Applications

- Laliberté (PDF) and Kushner, (submitted to GRL) explained tropospheric warming in the Arctic using dry and moist theories of midlatitude surface warming.

- G8 ExArch support accelerated this research!

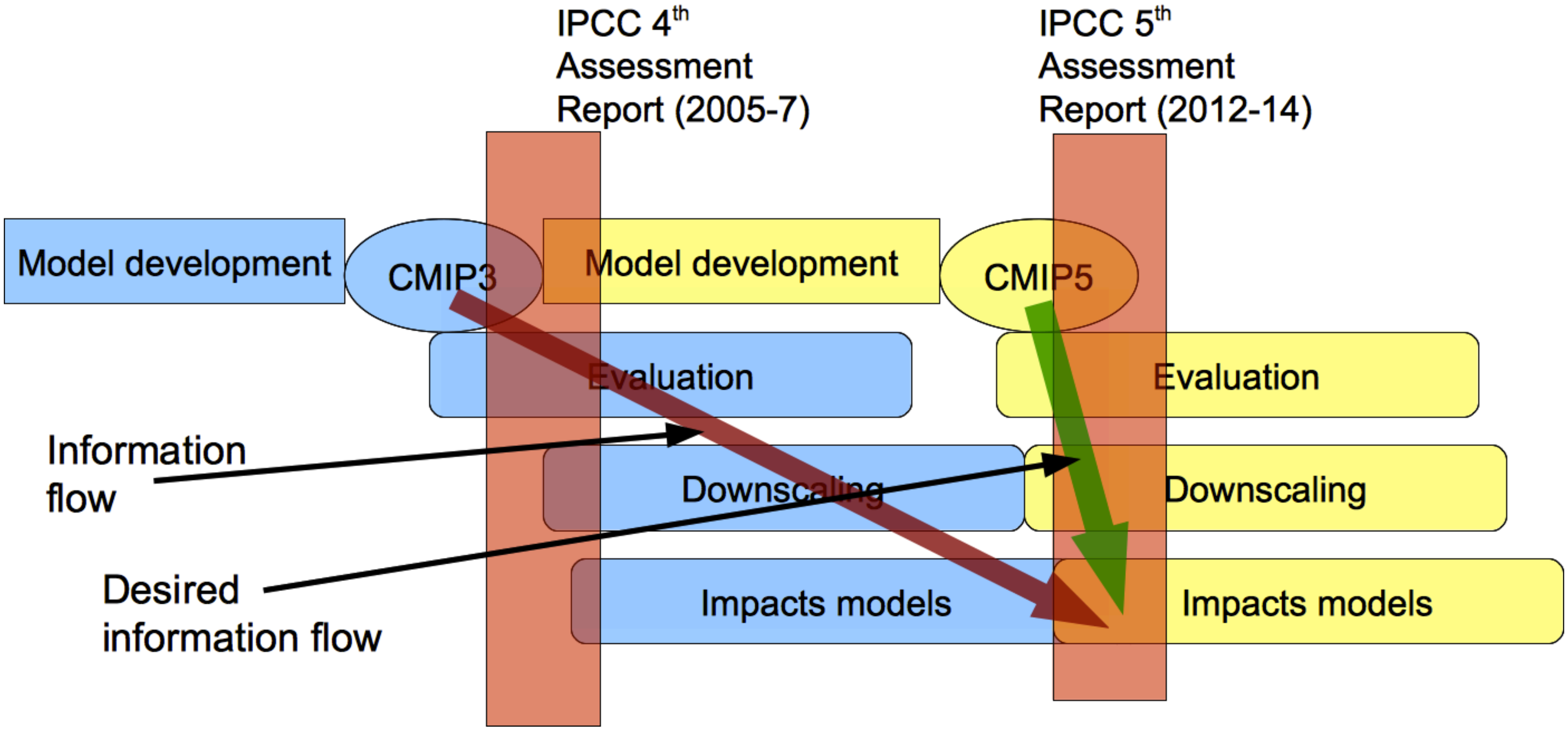
- Analysis scripts from CDB and access at IPSL and BADC. This workflow provides one of our benchmark calculations (monthly three dimensional data). Another benchmark case is based on cyclone statistics (daily surface data).



Scientific Diagnostics Tasks

- Quality assurance for CMIP5 implemented at DKRZ. CORDEX QA is underway.
    - Technical QA is more straightforward and easily automated than scientific QA.
- Implementation of WPS
    - WPS has been implemented in CDO analysis package to a limited extent, but operability across remote servers is still being developed.
    - CDB will be a testbed for this implementation.

Need to Accelerate Impacts Analysis



- CMIP3 impacts modelling being carried out now, while CMIP5 global data is being generated.

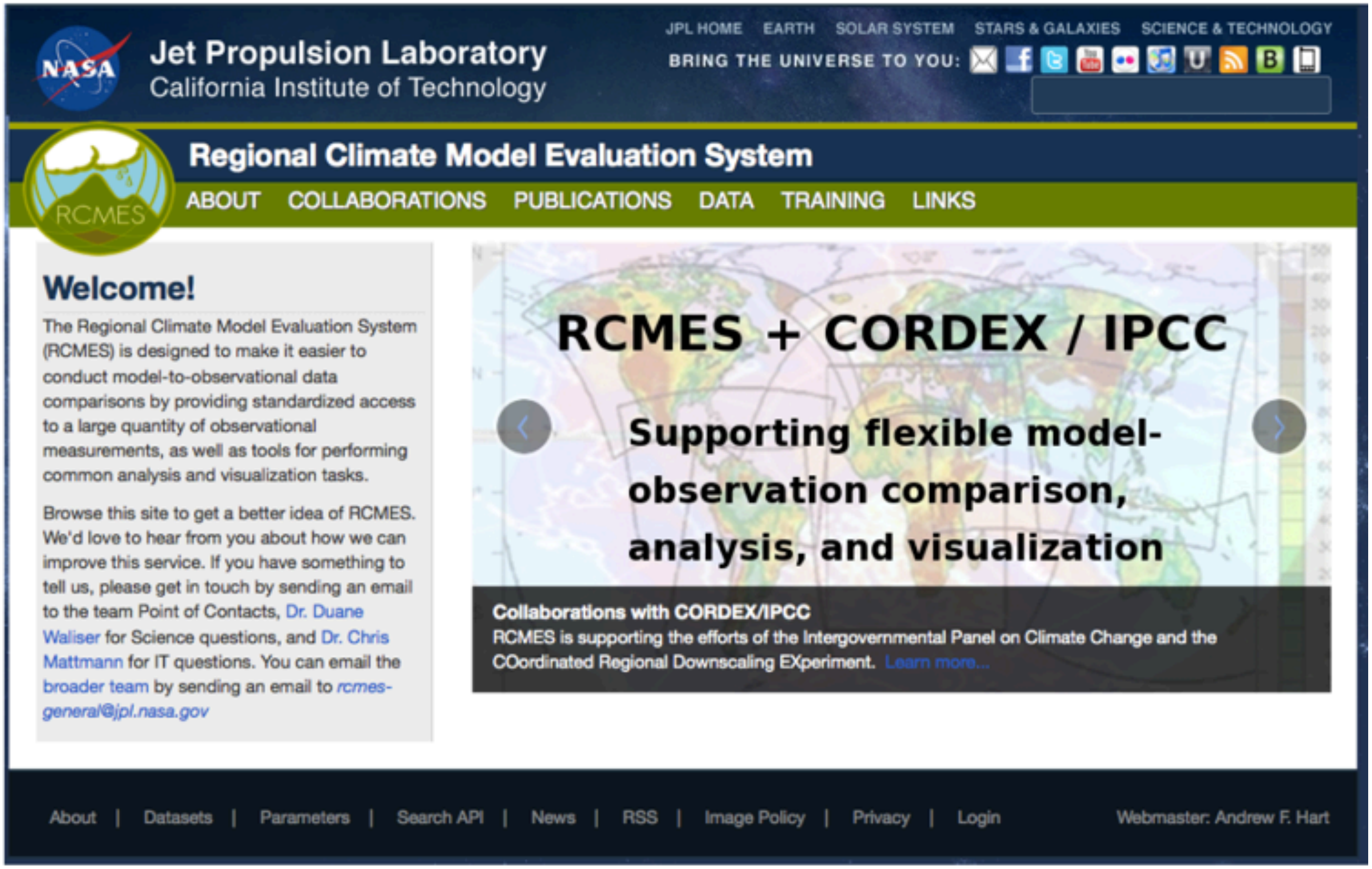
➡ ***How can we reduce this lag?***

Model/Earth Obs Comparison and Impacts Analysis

- Activity centred on UCLA/JPL Regional Climate Modelling Evaluation System (RCMES) for handling massive observational/modelling datasets.
- Aims for decision support for climate impacts, using model & RCMs.
- RCMES now released at rcmes.jpl.nasa.gov.
- Collaborative application of RCMES (and related publications) for CORDEX - Africa, North America, Arctic, East & South Asia
- Collaborations with Cape Town, NCAR, Toronto, Korean Met., Indian Met.

Our new website

- New Public Facing Web site
- Links to Publications, data available from RCMED
- Software/API Specifications
- Information about Collaborators



JPL HOME EARTH SOLAR SYSTEM STARS & GALAXIES SCIENCE & TECHNOLOGY
BRING THE UNIVERSE TO YOU:
Jet Propulsion Laboratory
California Institute of Technology
RCMES
Regional Climate Model Evaluation System
ABOUT COLLABORATIONS PUBLICATIONS DATA TRAINING LINKS

Welcome!

The Regional Climate Model Evaluation System (RCMES) is designed to make it easier to conduct model-to-observational data comparisons by providing standardized access to a large quantity of observational measurements, as well as tools for performing common analysis and visualization tasks.

Browse this site to get a better idea of RCMES. We'd love to hear from you about how we can improve this service. If you have something to tell us, please get in touch by sending an email to the team Point of Contacts, Dr. Duane Waliser for Science questions, and Dr. Chris Mattmann for IT questions. You can email the broader team by sending an email to *rcmes-general@jpl.nasa.gov*

RCMES + CORDEX / IPCC

Supporting flexible model-observation comparison, analysis, and visualization

Collaborations with CORDEX/IPCC

RCMES is supporting the efforts of the Intergovernmental Panel on Climate Change and the COordinated Regional Downscaling EXperiment. Learn more...

About | Datasets | Parameters | Search API | News | RSS | Image Policy | Privacy | Login
Webmaster: Andrew F. Hart

http://rcmes.jpl.nasa.gov/

Conclusion

Summary, Referring to G8 Agencies' Review Points

1. *Scientific objectives, scientific quality, and innovativeness*

- ExArch seeks to address the data challenges of computational climate science, which will be exacerbated as the field moves quickly into the exascale. This challenge presents short-, medium-, and long-term bottlenecks for basic and appplied climate research.
- The U.S. NRC Panel Report "A National Strategy for Climate Modelling" identifies this data challenge as an urgent priority. This view is shared by DOE, NOAA, NASA, UK CEDA, EC FP7, WMO, and other agencies.
- The recent innovation of a federated distribution system (ESGF) requires development, optimization, and end-to-end testing with real climate science problems. ExArch is contributing to all these.

2. *Methodology, resources, management*

- ExArch covers a wide range of topics, with a focus on leverage existing work. ExArch will provide software to enable the ESGF to meet exa-scale demands. This, as we proposed, was the best use of time and resources in these efforts, which are not typically supported with core funding.
- To manage this effort we have assembled a research team and Advisory Board dominated by key architects of existing Earth System Grid partners, as well as researchers in climate who wish to test and develop tools exploiting the new architectures. We have convened two meetings (GO-ESSP 2011 and Windsor, UK, 2012) and have actively participated in related AGU and EGU Earth System Informatics sessions.

Summary, Referring to G8 Agencies' Review Points

3. *Benefits of exascale computing, benefits to exascale computing.*

- By addressing an emerging bottleneck in exascale climate science, the ExArch project's activities have the potential to greatly benefit the future of the field.

4. *Benefits of international co-operation.*

- This effort requires full international co-operation at all stages. ExArch is needed to bridge the gap between existing European and American efforts in this area. This has also been a great opportunity for the NSERC funded Canadian partners, who have obtained access to critical resources in this effort.

5. *Scientific perspectives beyond the project.*

- The availability of intercomparisons in Earth System modelling has lead to some of the most important recent advances in the field. The efforts of this kind of project provide the fundamental technical underpinnings for further progress in the field.
- Distributed climate analytics provide open worldwide access to climate analysis with only modest local resources required. This serves to democratize our science.
- Our activity will benefit broader effort in exascale simulation, wherever model complexity and resolution are outstripping available human resources for analysis.

Larger Issues in Exascale Climate Analytics

- *Governance*: How will modelling centres be required to share, document, and archive their data in the future? Within the global climate community, a vigorous discussion of options for determining standards, protocols, and implementation of federated data analysis is ongoing. It is clear that a federated solution is desireable, but details of decision making and policy implementation need to be worked out.

- *Infrastructure funding*: Few of the activities described here have been supported by core funding (in the way that, e.g. ESM development has been supported). The community urgently requires such support if it is to continue to provide the benefits in exascale climate analytics obtained from CMIP3 at the sub-petascale.

ExArch Early Career Researchers

- Damasio da Costa (STFC)
- Mattman (UCLA)
- Laliberté, Cooke, Karczewska (Toronto)
- Nikonov and Ansari (Princeton)