Data Infrastructure in Climate Sciences

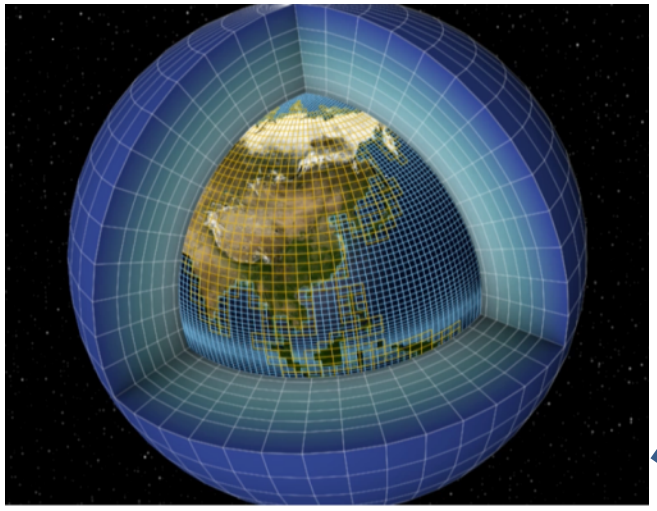
Subtitles:

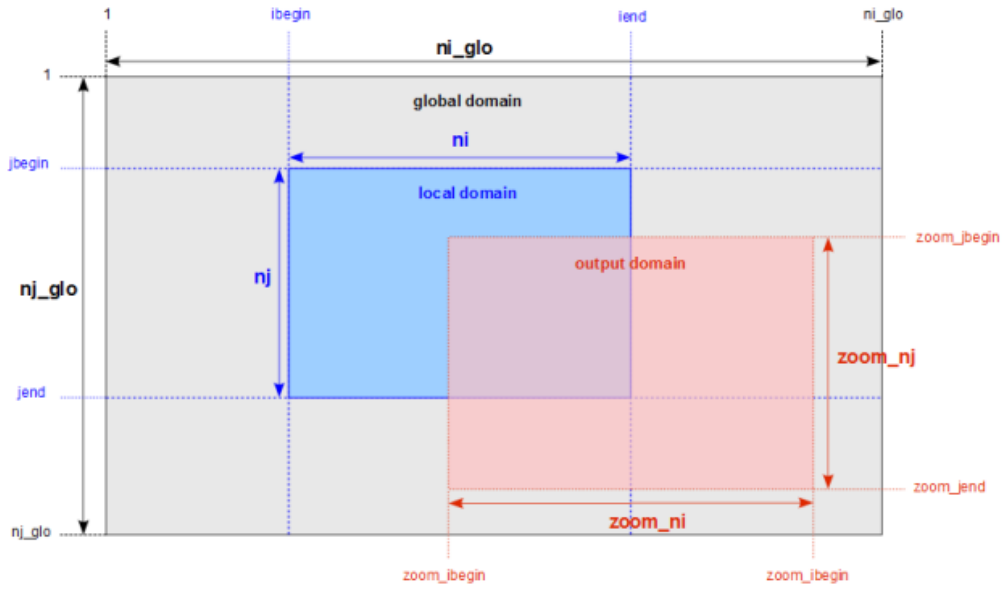
"Challenges" in terms of accessibility and usage

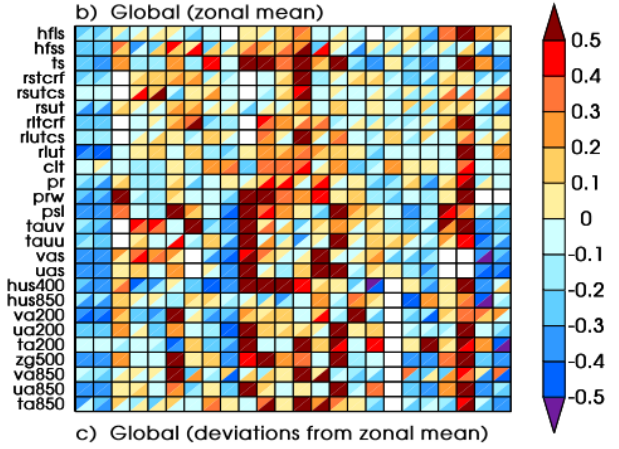
Sébastien Denvil
(IPSL/CNRS)

IPSL climate modelling centre (ICMC)

http://icmc.ipsl.fr



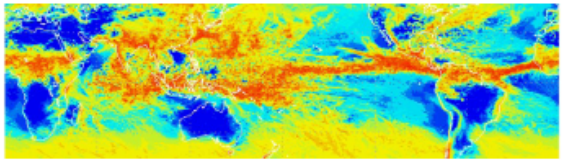


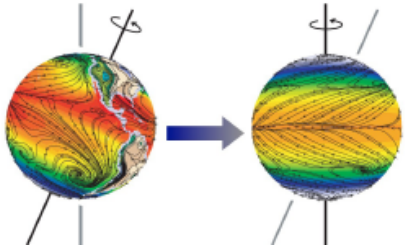


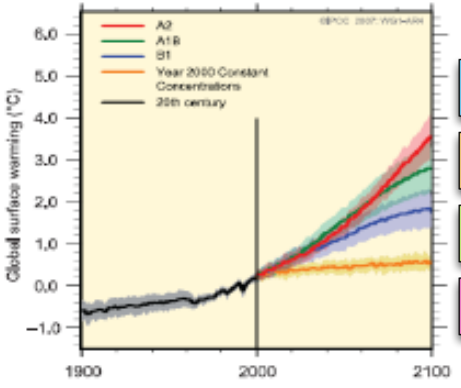
1. Simulation

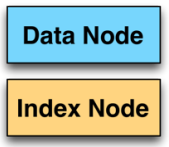
2. Archive and distribute

3. Analysis

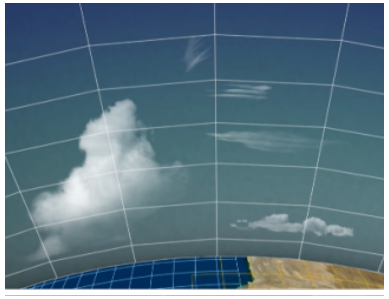
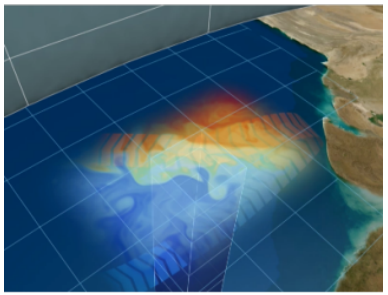
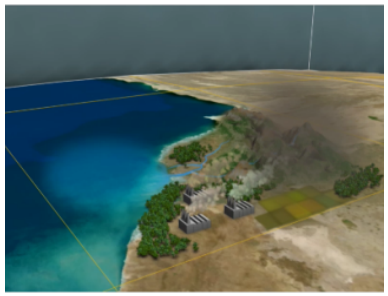
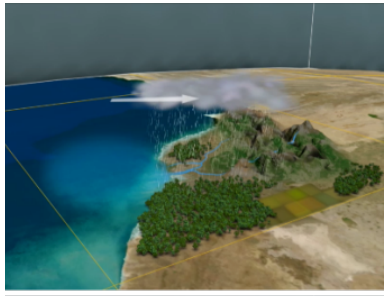


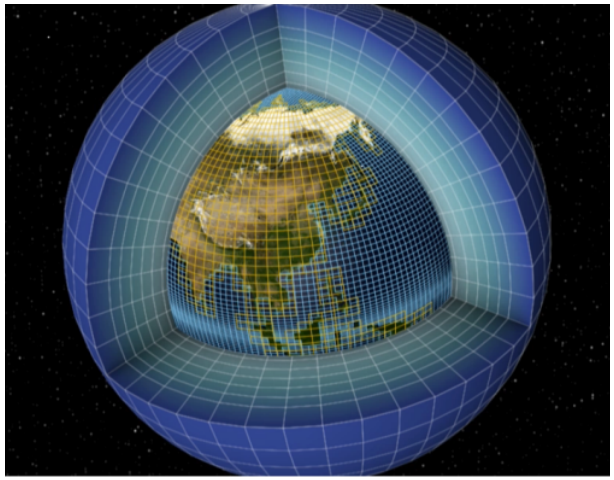


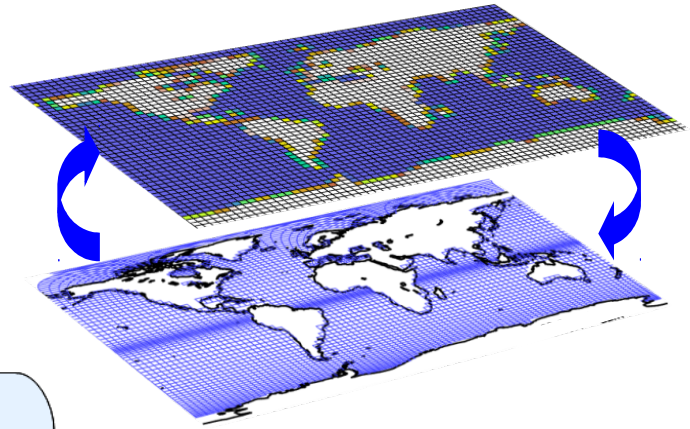




IPSL climate model

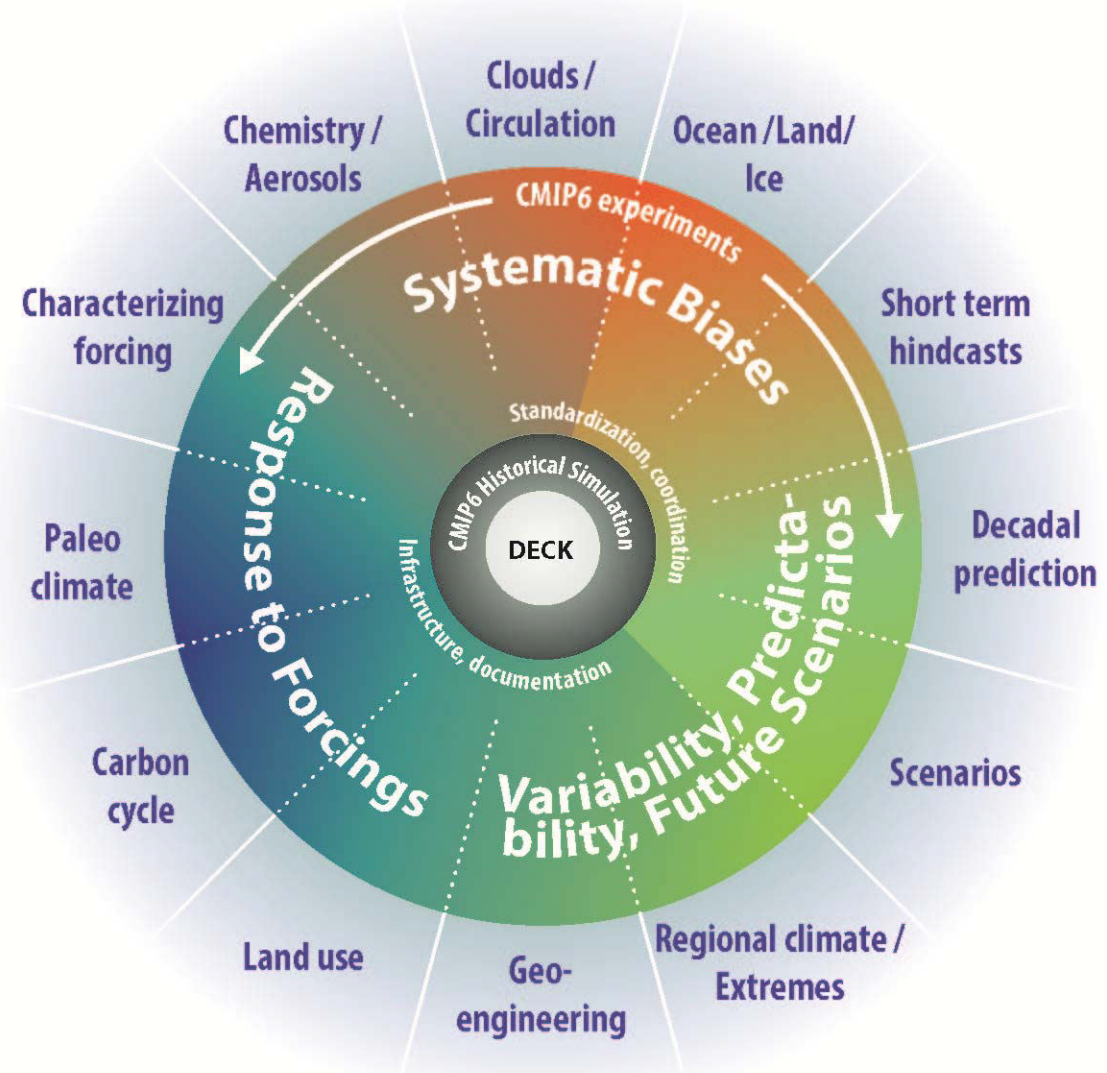


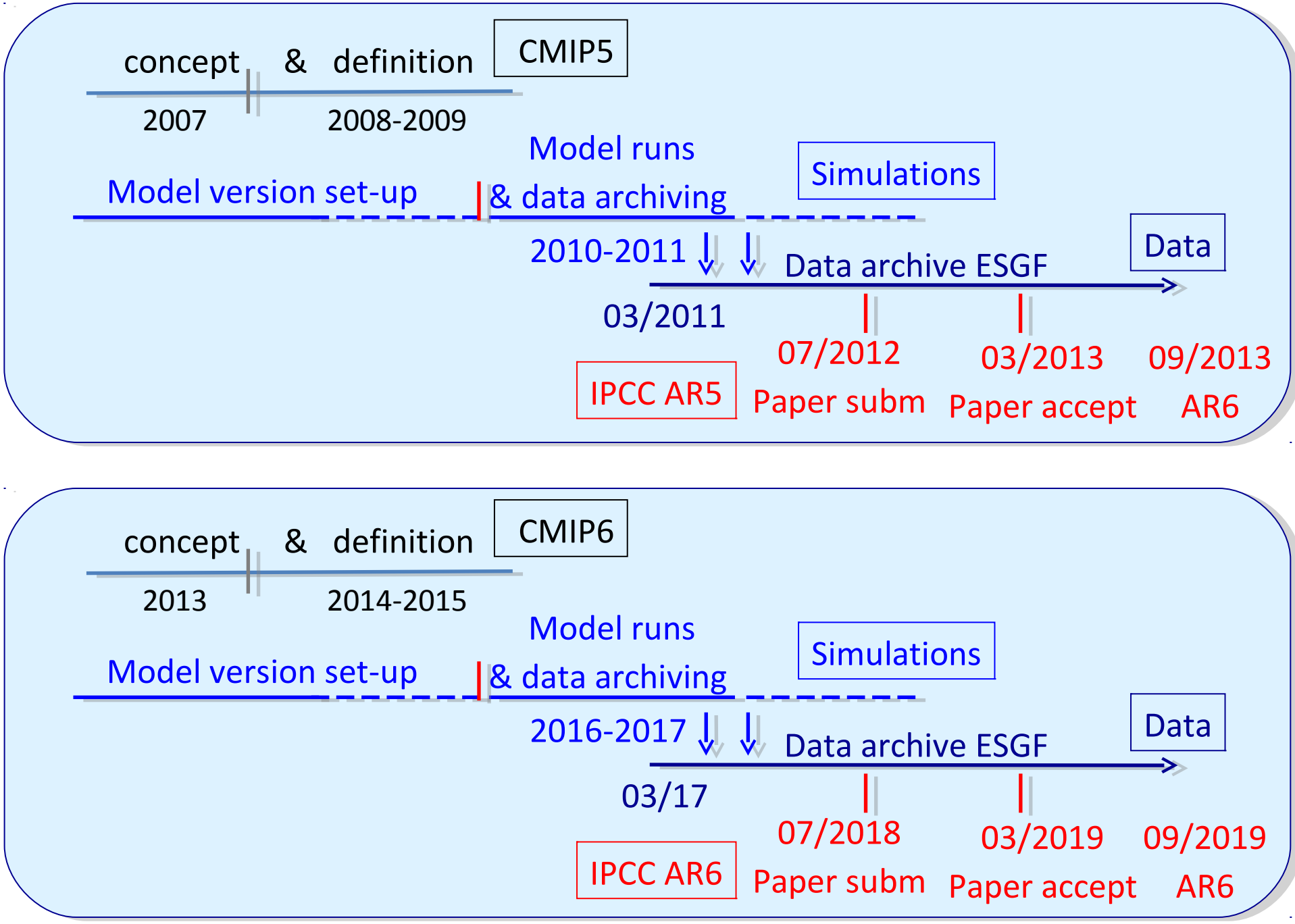


Climate Simulations

CMIP6 (Coupled Model Intercomparison Projet phase 6), préparation du

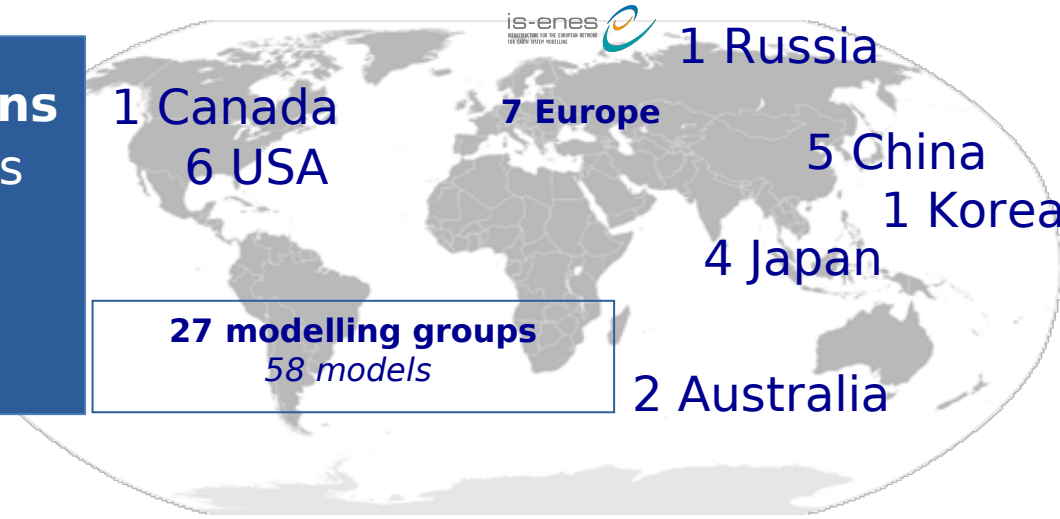
IPCC AR6 (Intergovernmental Panel Climate Change, Assesment Report 6)



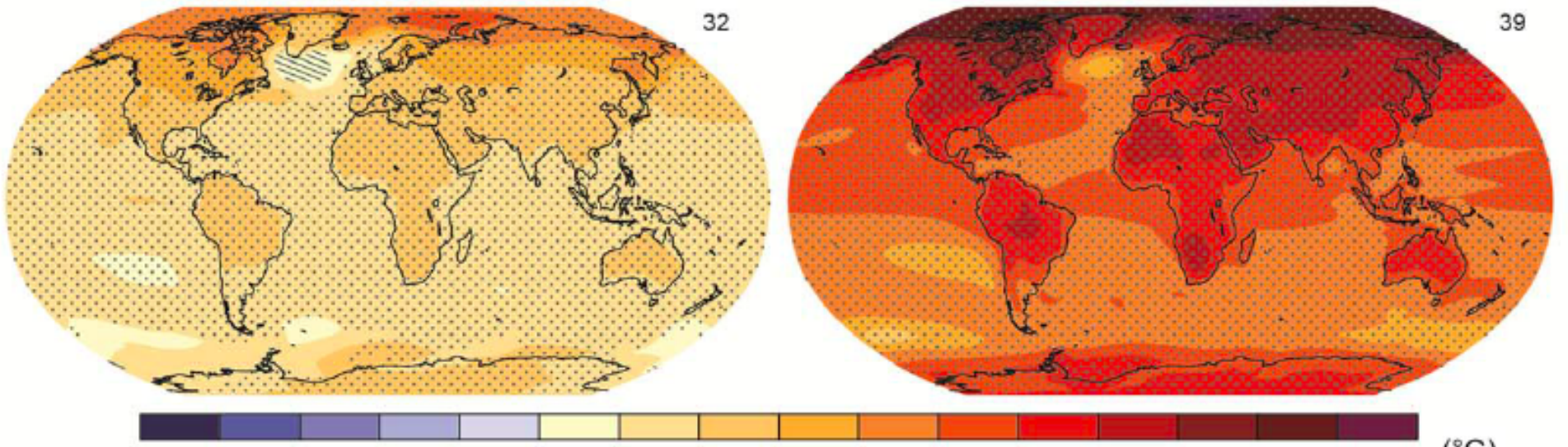

concept & definition
CMIP5
2007
2008-2009
Model version set-up
Model runs
& data archiving
Simulations
2010-2011
Data archive ESGF
Data
03/2011
07/2012
03/2013
09/2013
IPCC AR5
Paper subm
Paper accept
AR6
concept & definition
CMIP6
2013
2014-2015
Model version set-up
Model runs
& data archiving
Simulations
2016-2017
Data archive ESGF
Data
03/17
07/2018
03/2019
09/2019
IPCC AR6
Paper subm
Paper accept
AR6

Support to international coordinated experiments : CMIP5

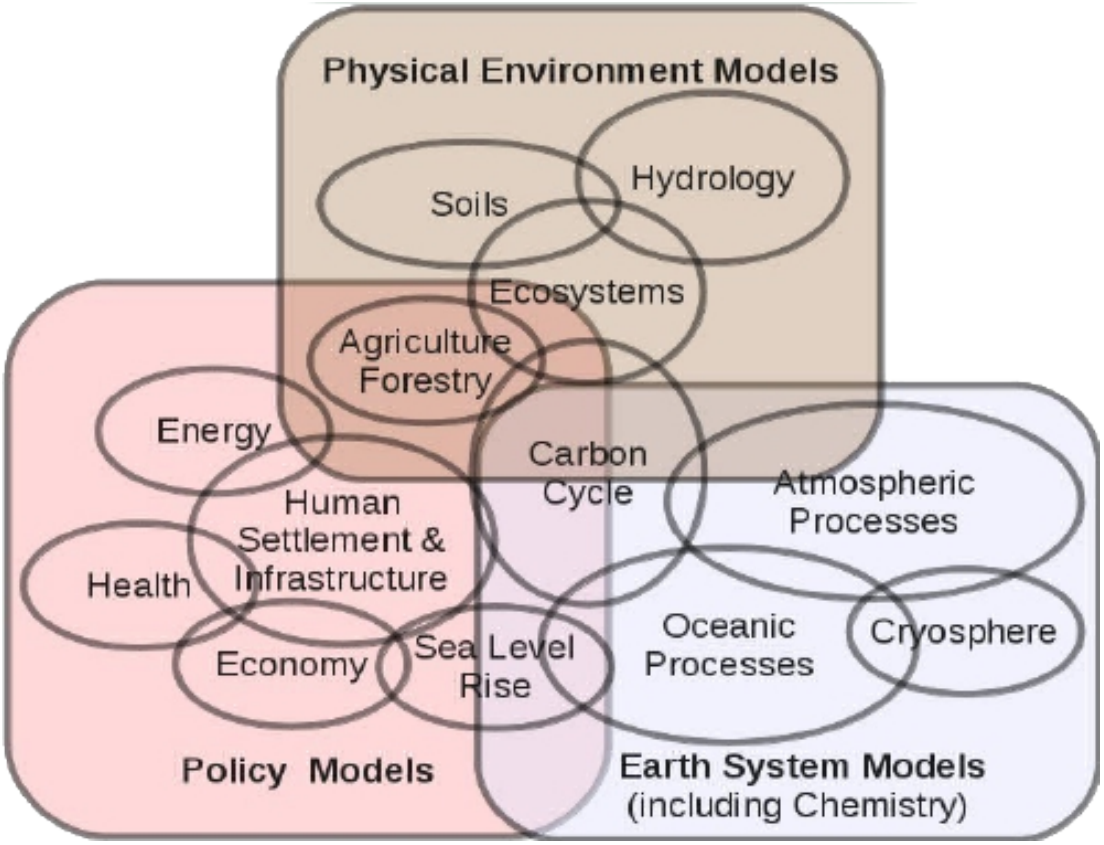
Evaluate/Understand/Projections
3400 simul. yrs up to > 12000 yrs
50 expts up to > 160 expts
CMIP5: 1000 – 3000 Tbytes
(CMIP3: 36)



IPCC AR5 SPM (2013)



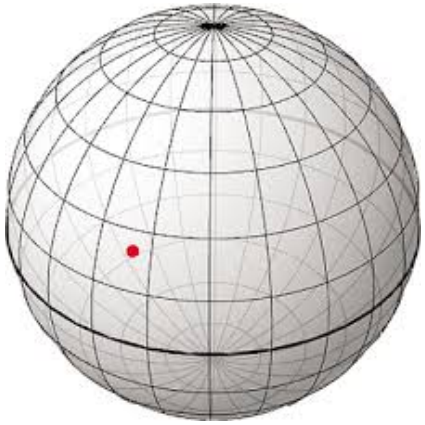
Many, many processes, many, many communities !

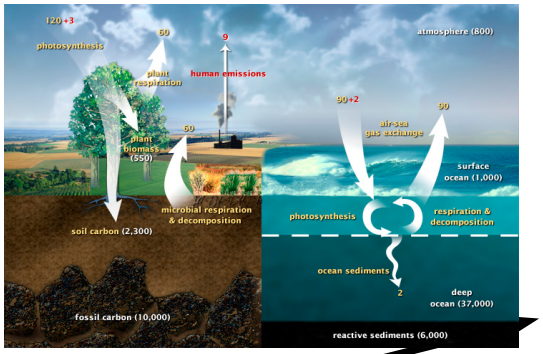


Interconnected communities, all needing access to (some of) the data!

IPCC AR5 variable counts

	1 hour	3 hour	6 hour	daily	month	annual	**totals**
aerosol	0	0	0	0	81	0	**81**
atmosphere	75	101	9	86	184	0	**455**
land	0	3	0	2	59	0	**64**
land ice	0	0	0	2	13	0	**15**
ocean	0	1	0	3	116	0	**120**
biogeochemistry	0	0	0	0	88	71	**159**
sea ice	0	0	0	4	47	0	**51**
totals	**75**	**105**	**9**	**97**	**588**	**71**	**945**



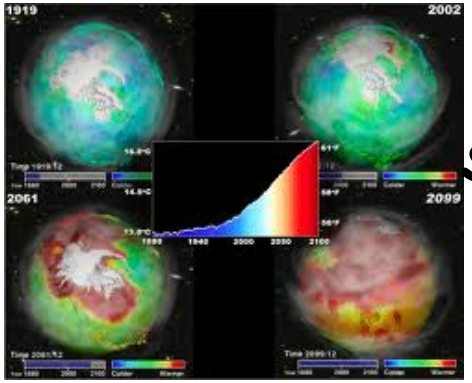


Earth Observations

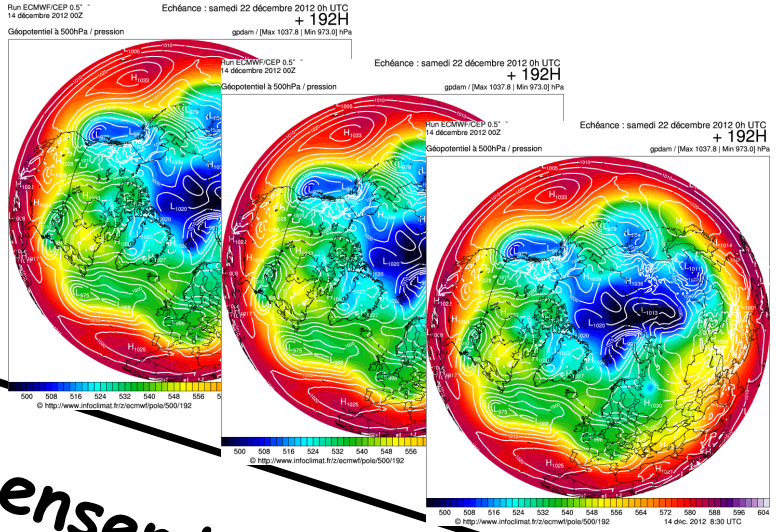
Complexity

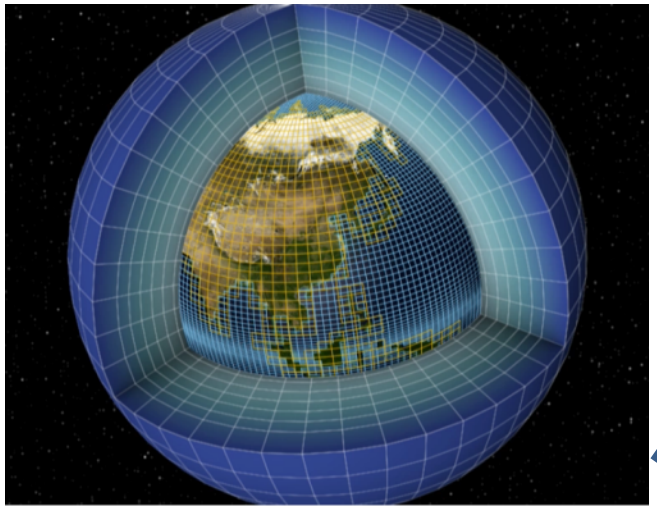
Ehanced computing resources produce MORE DATA

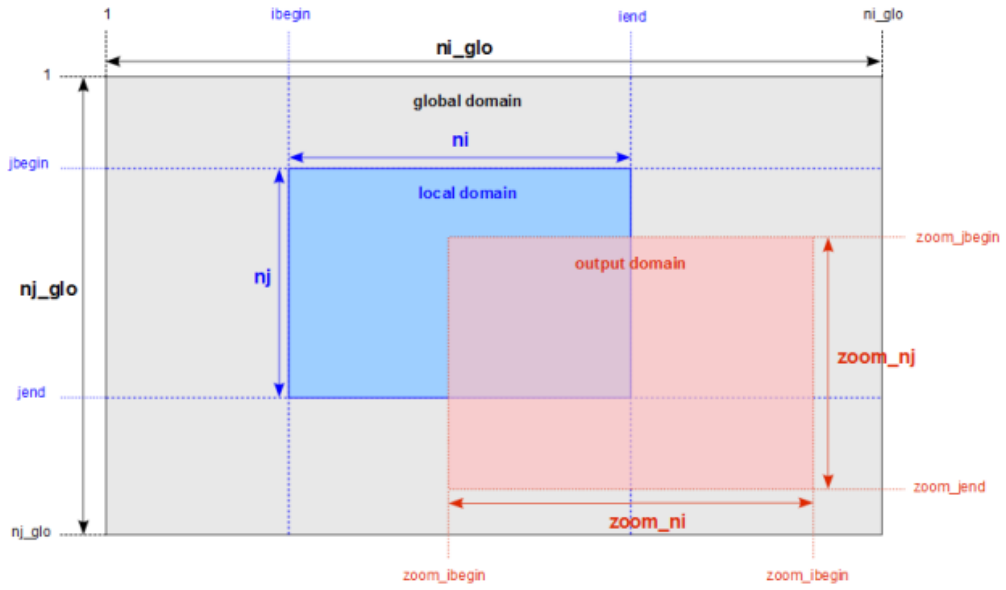
Resolution

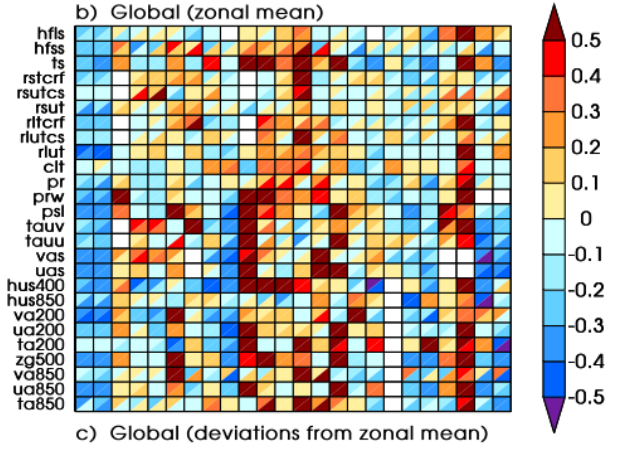


Duration and ensemble size





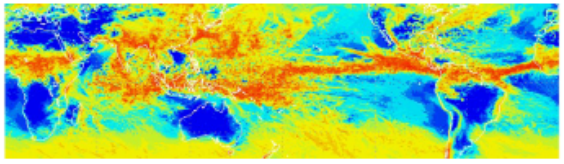


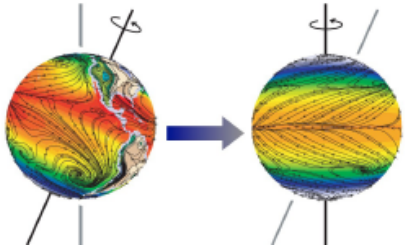


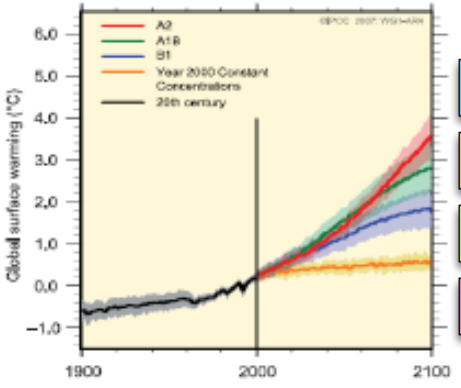
1. Simulation

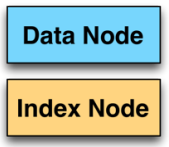
2. Archive and distribute

3. Analysis









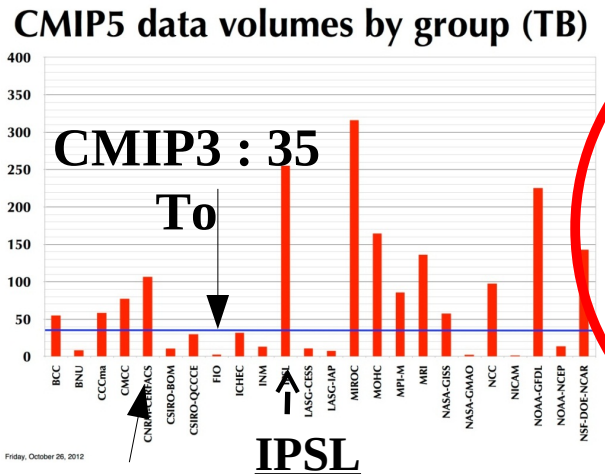
A one slide guide to CMIP5 from a data perspective

Fifth Climate Model Intercomparison Project (CMIP5)

World Climate Research Programme WCRP- WGCM Involves all the major climate modelling centres.

Original Timing: o(2) PB of requested output from 20+ modelling centres finished early 2010!
Actual Timing? Years late.

101 experiments
61 model variants
59,000 datasets!
4.5 million files
2 PB in global archive.
Unknown PB locally!



Community developed s/w infrastructure for data delivery:
Earth System Grid Federation

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

prodiguer

Worldwide distributed system

The **Earth System Grid Federation (ESGF)** is a multi-agency, international collaboration of persons and institutions working together to build an open source software infrastructure for the management and analysis of Earth Science data on a global scale

- Software development and project management: ANL, ANU, **BADC**, **CMCC**, **DKRZ**, ESRL, GFDL, GSFC, JPL, **IPSL**, ORNL, LLNL (lead), PMEL, …
- Operations: tens of data centers across Asia, Australia, Europe and North America

Worldwide distributed system

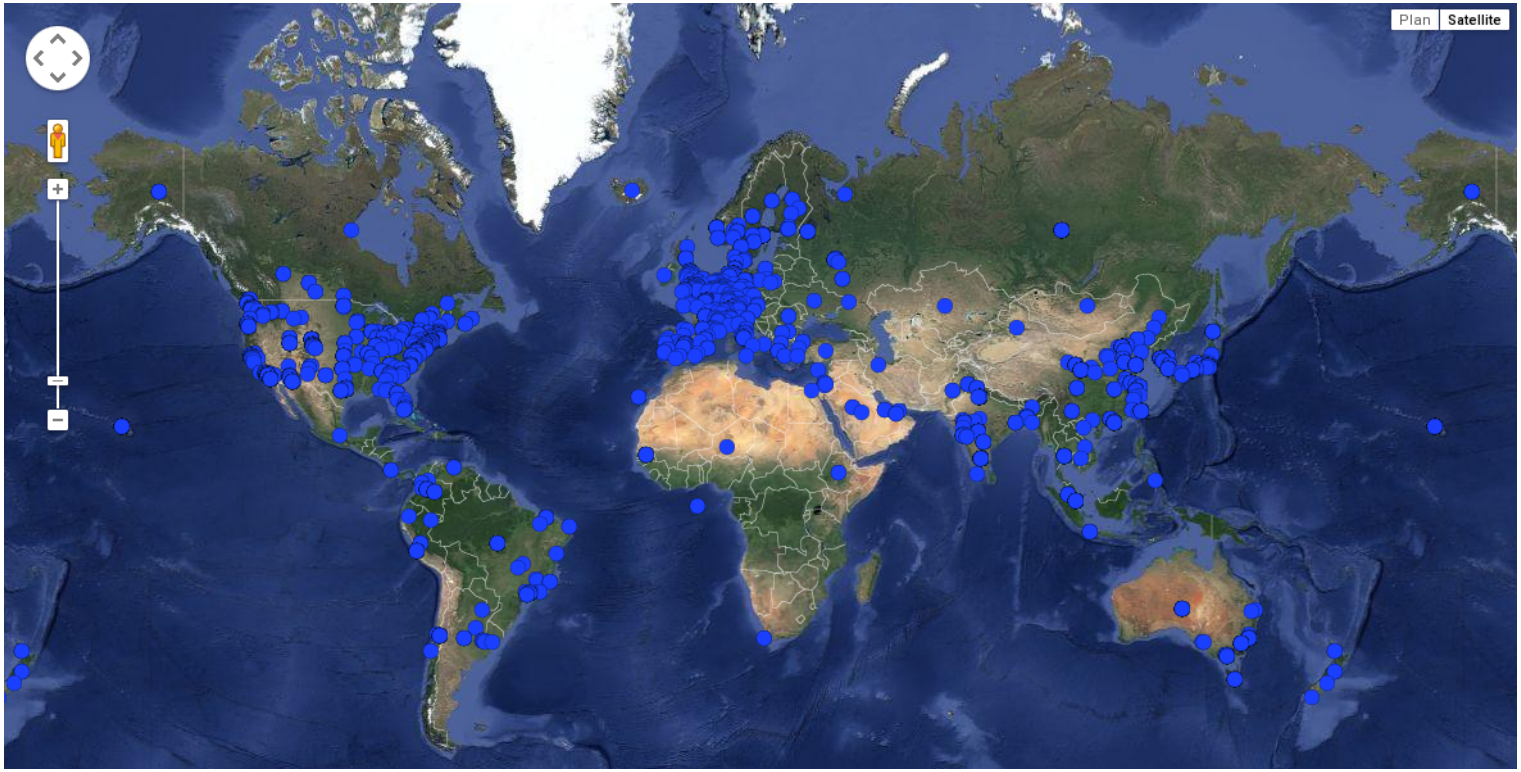
Storage evolution in 6 years time (from CMIP3 to CMIP5) : a factor x30

- Operational since 2011
- Hundreds of users per month
- Hundreds of To per month
- About 10 000 registered users

CMIP3: centralized

CMIP5: distributed system

- 60 climate models
- 2 PB of data



System Architecture

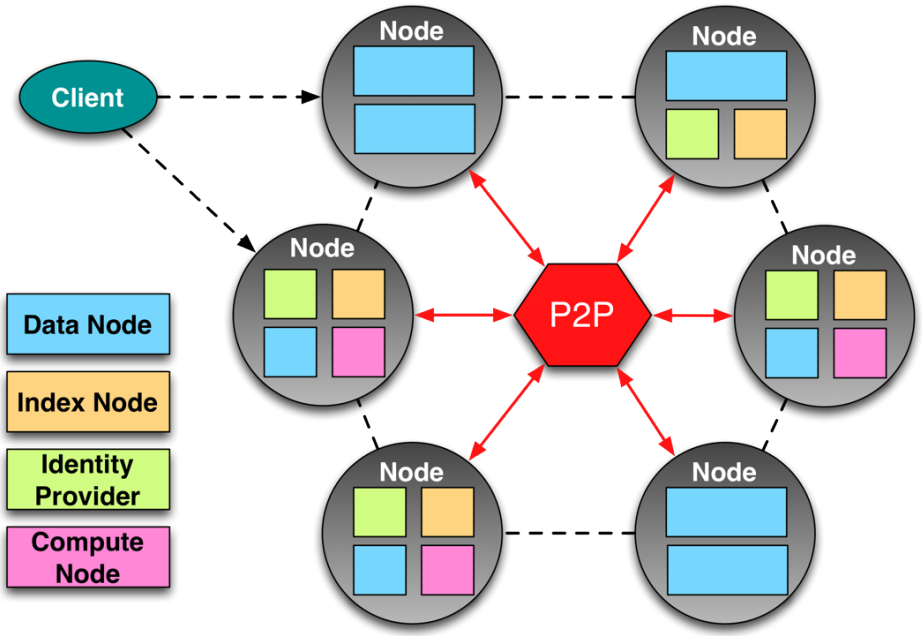
ESGF is a system of distributed and federated Nodes

that interact dynamically through a Peer-To-Peer (P2P) paradigm

Distributed: data and metadata are published, stored and served from multiple centers ("Nodes")

Federated: Nodes interoperate because of the adoption of common services, protocols and APIs, and establishment of mutual trust relationships

Dynamic: Nodes can join/leave the federation dynamically - global data and services will change accordingly

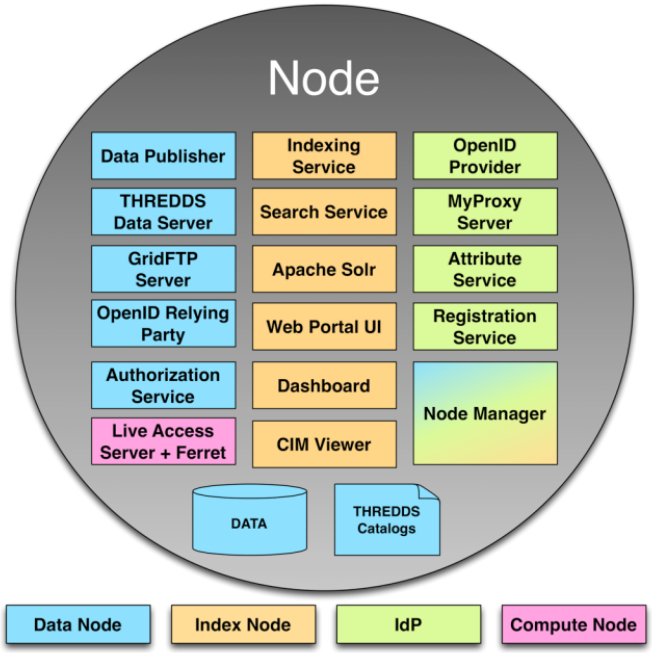


A client (browser or program) can start from any Node in the federation and discover, download and analyze data from multiple locations as if they were stored in a single central archive.

Software Stack

Internally, each ESGF Node is composed of services and applications that collectively enable metadata discovery, data access, and user management. Software components are grouped into 4 areas of functionality (aka “flavors”):

- Data Node : secure data publication and access
- Index Node :
  - metadata indexing and searching (w/ Solr)
  - web portal UI to drive human interaction
- Identity Provider : user authentication and group membership
- Compute Node : analysis and visualization



- Open Source with BSD License
- Source code: http://esgf.org/gitweb → https://github.com/ESGF
- Documentation: https://esgf.llnl.org

European coordination

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

http://is.enes.org/

CAPACITIES

InfraStructure for Earth System modelling

IS-ENES & IS-ENES2 EU projects

1rst phase – 03/2009 to 02/2013
18 partners

2nd phase – 04/2013 to 03/2017
23 partners

+

Infrastructure strategy for ENES for the next 10 years

Global & Regional climate models

Key role of infrastructure : models, data & computing

Recommandations:

1)Access to world-class HPC for climate at least «tailored » for climate up to « dedicated »

2)Develop the next generation of climate models

3)Set up data infrastructure (global and regional models) for large range of users from impact community

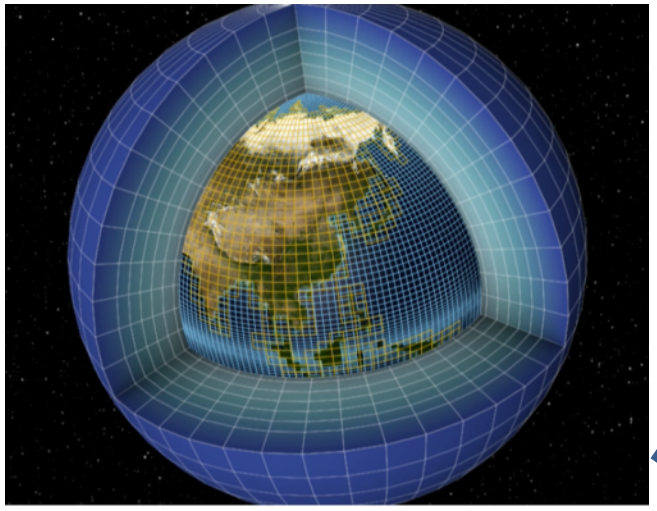
4)Improve physical network (e.g. link national archives)

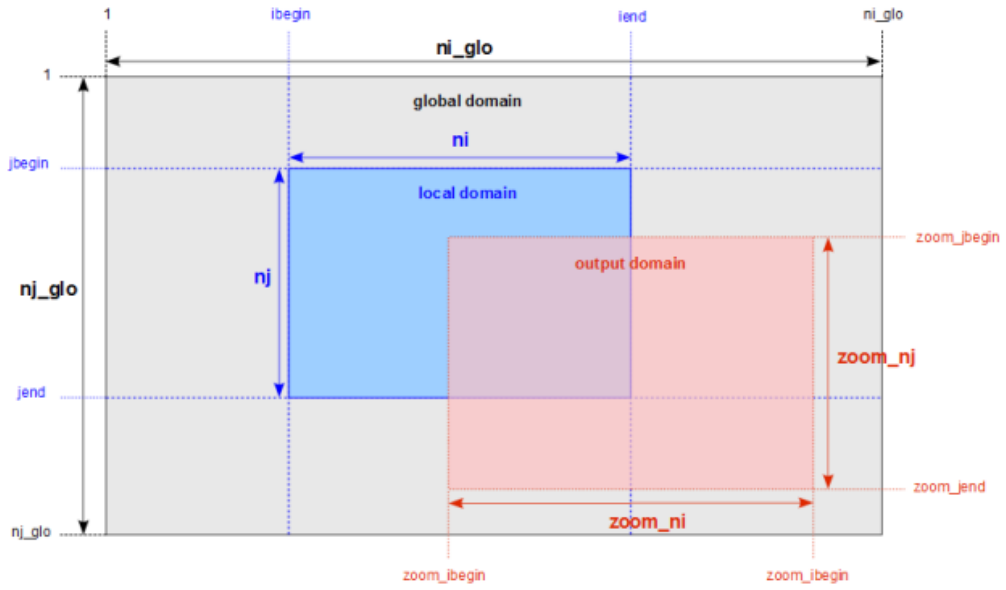
5)Strengthen European expertise and networking

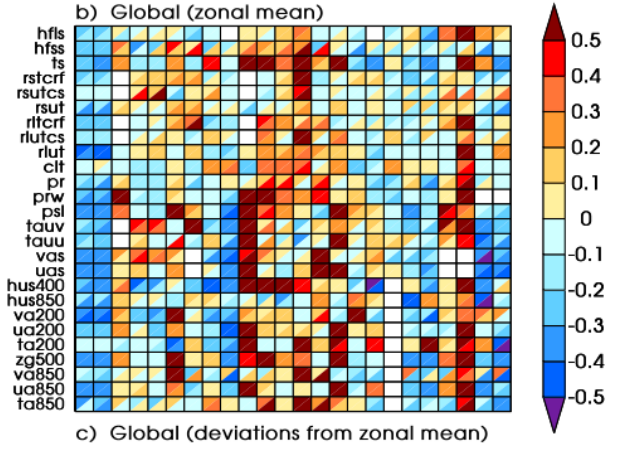
Input to IS-ENES2

ENES

Towards an European Climate Infrastructure Initiative : a sustainable virtual laboratory



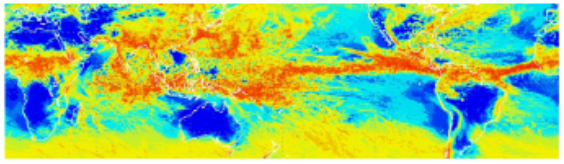


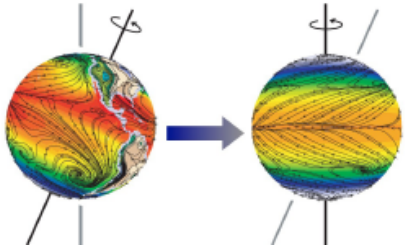


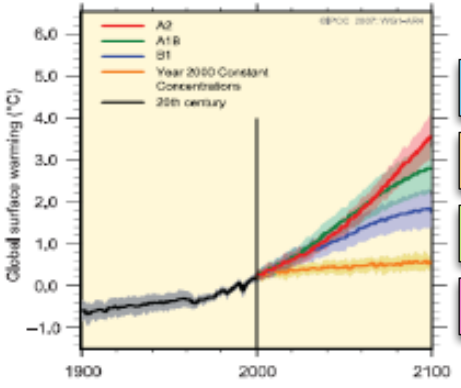
1. Simulation

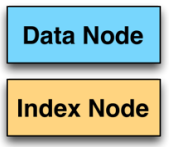
2. Archive and distribute

3. Analysis









To keep in mind

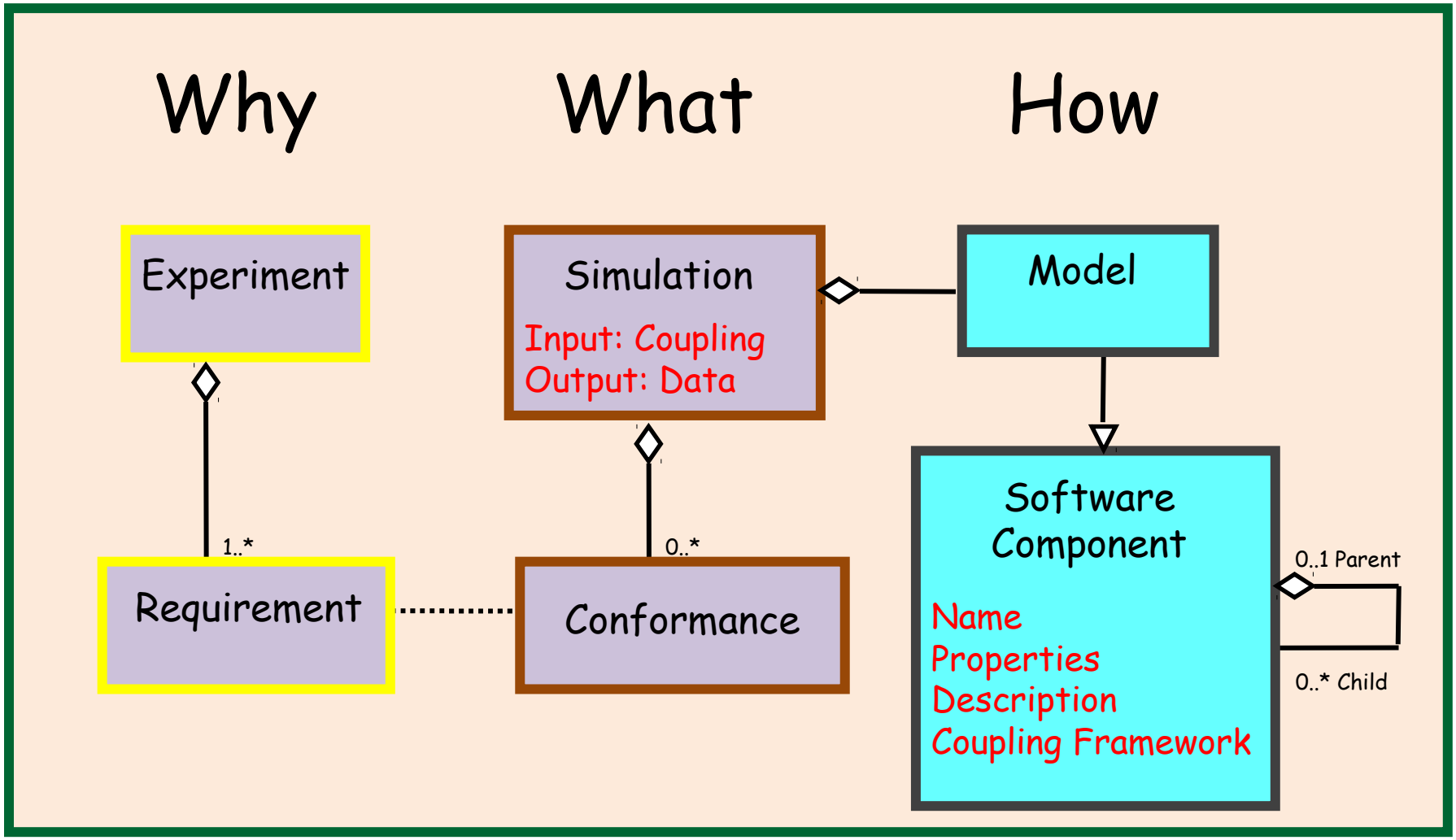
"the potential to interpret, compare and reuse climate information results is strongly related to the quality of their description"

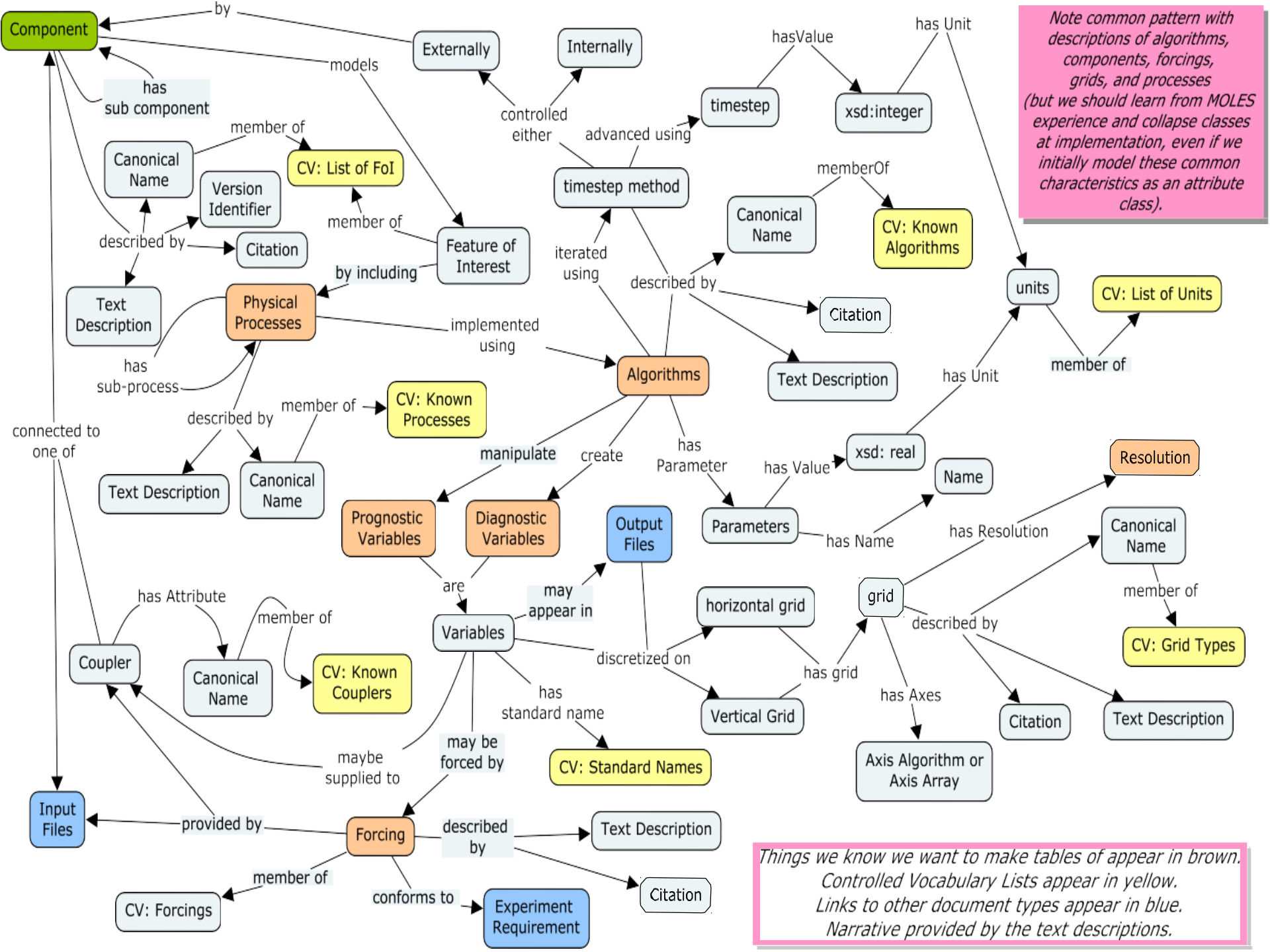
Computation useless if results cannot be stored/distributed/read

es-doc
Earth System Documentation

A climate simulation

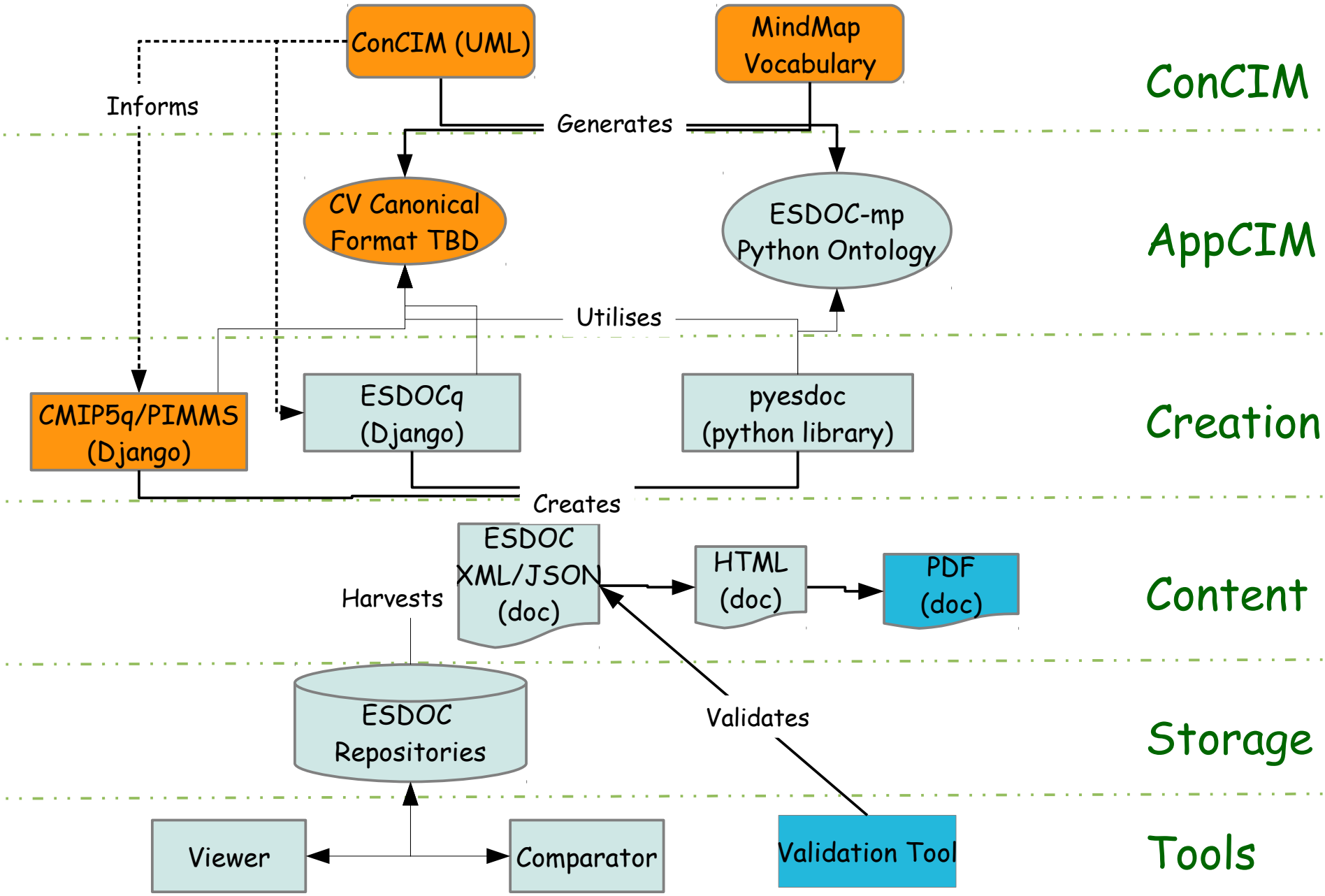
http://earthsystemcog.org/projects/es-doc-models/




Component
by
Externally
Internally
models
has sub component
controlled either
advanced using
timestep
hasValue
has Unit
xsd:integer
Note common pattern with descriptions of algorithms, components, forcings, grids, and processes (but we should learn from MOLES experience and collapse classes at implementation, even if we initially model these common characteristics as an attribute class).
Canonical Name
member of
CV: List of FoI
Version Identifier
described by
Citation
member of
Feature of Interest
timestep method
memberOf
Canonical Name
CV: Known Algorithms
iterated using
described by
Citation
Text Description
Physical Processes
by including
implemented using
Algorithms
Text Description
units
CV: List of Units
has Unit
member of
has sub-process
connected to one of
described by
member of
CV: Known Processes
manipulate
create
has Parameter
has Value
xsd: real
Text Description
Canonical Name
Prognostic Variables
Diagnostic Variables
Parameters
has Name
Name
Resolution
has Resolution
Canonical Name
Output Files
are
may appear in
Variables
discretized on
horizontal grid
Vertical Grid
has grid
grid
described by
member of
CV: Grid Types
has Attribute
member of
Coupler
Canonical Name
CV: Known Couplers
has standard name
CV: Standard Names
has Axes
Axis Algorithm or Axis Array
Citation
Text Description
maybe supplied to
may be forced by
Input Files
provided by
Forcing
described by
Text Description
Citation
member of
CV: Forcings
conforms to
Experiment Requirement
Things we know we want to make tables of appear in brown.
Controlled Vocabulary Lists appear in yellow.
Links to other document types appear in blue.
Narrative provided by the text descriptions.

CIM

- The CIM is **intentionally very general**. It can be customized for particular user communities through the addition of specific Controlled Vocabularies.
- A Controlled Vocabulary **defines the content** that can be used within a CIM document. For example, in the case of climate models, the CIM schema (structure) allows a ModelComponent to have a child ModelComponent. And each of those components can have “types.”
- A CV **is required** to list the permitted types. For example, the CMIP5 CV allows an “atmosphere” model to have a child “advection” model, but not a child “ocean” model. Thus, in order to be valid a CIM document must conform both to the CIM schema and to a particular set of CVs.


ConCIM (UML)
MindMap Vocabulary
ConCIM
Informs
Generates
CV Canonical Format TBD
ESDOC-mp Python Ontology
AppCIM
Utilises
CMIP5q/PIMMS (Django)
ESDOCq (Django)
pyesdoc (python library)
Creation
Creates
ESDOC XML/JSON (doc)
HTML (doc)
PDF (doc)
Content
Harvests
ESDOC Repositories
Validates
Storage
Viewer
Comparator
Validation Tool
Tools

ES-DOC tools
Guided Tour

ES-DOC

Tools - Document Viewer

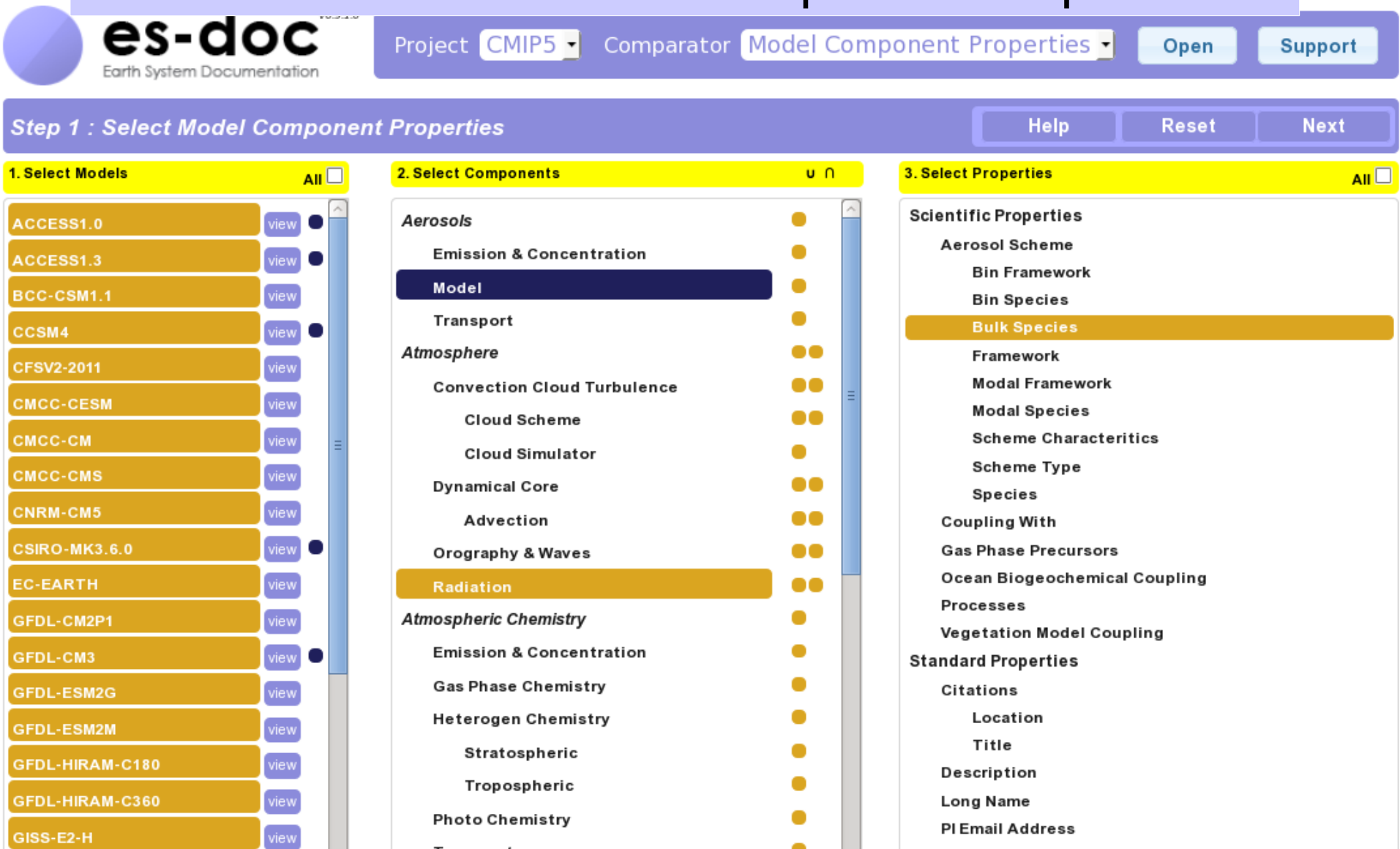
CMIP5 Model - HadGEM2-ES | Model | Simulation | Experiment | Platform

Overview | Citations | Contacts | Components | Grids

Project	CMIP5
Short Name	HadGEM2-ES
Long Name	Hadley Global Environment Model 2 - Earth System
Institute	UK Met Office Hadley Centre
Funder	UK Met Office Hadley Centre
Principal Investigator	Chris Jones
Release Date	2009-11-26 00:00:00
Language	--
Description	The HadGEM2-ES model was a two stage development from HadGEM1, representing improvements in the physical model (leading to HadGEM2-AO) and the addition of earth system components and coupling (leading to HadGEM2-ES). [1] The HadGEM2-AO project targeted two key features of performance: ENSO and northern continent land-surface temperature biases. The latter had a particularly high priority in order for the model to be able to adequately model continental vegetation. Through focussed working groups a number of mechanisms that improved the performance were identified. Some known systematic errors in HadGEM1, such as the Indian monsoon, were not targeted for attention in HadGEM2-AO. HadGEM2-AO substantially improved mean SSTs and wind stress and improved tropical SST variability compared to HadGEM1. The northern continental warm bias in HadGEM1 has been significantly reduced. The power spectrum of El Nino is made worse, but other aspects of ENSO are improved. Overall there is a noticeable improvement from HadGEM1 to HadGEM2-AO when comparing global climate indices. [2] In HadGEM2-ES the vegetation cover is better than in the previous HadCM3LC model especially for trees, and the productivity is better than in the non-interactive HadGEM2-AO model. The presence of too much bare soil in Australia though may cause problems for the dust emissions scheme. The simulation of global soil and biomass carbon stores are good and agree well with observed estimates except in regions of errors in the vegetation cover. HadGEM2-ES compares well with the C4MIP ensemble of models. The distribution of NPP is much improved relative to HadCM3LC. At a site level the component carbon fluxes validate better against observations and in particular the timing of the growth season is significantly improved. The ocean biology (HadOCC) allows the completion of the carbon cycle and the provision of di-methyl sulphide (DMS) emissions from phytoplankton. DMS is a significant source of sulphate aerosol over the oceans. The diat-HadOCC scheme is an improvement over the standard HadOCC scheme as it differentiates between diatom and non-diatom plankton. These have different processes for removing carbon from the surface to the deep ocean, and respond differently to iron nutrients. The HadOCC scheme performs well with very reasonable plankton distributions, rates of productivity and emissions of DMS. The diat-HadOCC scheme has slightly too low levels of productivity which requires further tuning to overcome. The additions of a tropospheric chemistry scheme, new aerosol species (organic carbon and dust) and coupling between the chemistry and sulphate aerosols have significantly enhanced the earth system capabilities of the model. This has improved the tropospheric ozone distribution and the distributions of aerosol species compared to observations, both of which are important for climate forcing. Including interactive earth system components has not significantly affected the physical performance of the model.

ES-DOC

Tools – Document Comparator step 1



ES-DOC

Tools - Document Comparator step 2

es-doc
Earth System Documentation

Project CMIP5 Comparator Model Component Properties Open Support

Step 2 : View report table Help CSV Back

??? = Incomplete documentation. N/A = Not applicable (model did not realize component). **Models = 42, Components = 2, Properties = 3.**

Component	Aerosols > Model	Atmosphere > Radiation	Atmosphere > Radiation
Property	Scientific Properties > Aerosol Scheme > Bulk Species	Scientific Properties > Aerosol Types	Scientific Properties > Longwave > Number Of Spectral Intervals
ACCESS1.0	BC (black carbon / soot) | Dust | POM (particulate organic matter) | SOA (secondary organic aerosols) | Sea salt | Sulphate	BC (black carbon / soot) | Dust | POM (particulate organic matter) | SOA (secondary organic aerosols) | Sea salt | Sulphate	9
ACCESS1.3	BC (black carbon / soot) | Dust | POM (particulate organic matter) | SOA (secondary organic aerosols) | Sea salt | Sulphate	BC (black carbon / soot) | Dust | Nitrate | POM (particulate organic matter) | SOA (secondary organic aerosols) | Sea salt | Sulphate	9
BCC-CSM1.1	N/A	BC (black carbon / soot) | Dust | Organic | POM (particulate organic matter) | Sea salt | Sulphate	8
CCSM4	BC (black carbon / soot) | Dust | Organic | Sea salt | Sulphate	BC (black carbon / soot) | Dust | Organic | Sea salt | Sulphate	8
CFSV2-2011	N/A	BC (black carbon / soot) | Dust | Organic | Sea salt | Sulphate	16
CMCC-CESM	N/A	Sulphate	16
CMCC-CM	N/A	Sulphate	16
CMCC-CMS	N/A	Sulphate	16
CNRM-CM5	N/A	BC (black carbon / soot) | Dust | Ice | Organic | POM (particulate organic matter) | Sea salt | Sulphate	16
CSIRO-MK3.6.0	BC (black carbon / soot) | Dust | Organic | POM (particulate organic matter) | SOA (secondary organic aerosols) | Sea salt | Sulphate	BC (black carbon / soot) | Dust | Ice | POM (particulate organic matter) | Sea salt | Sulphate	10
EC-EARTH	N/A	BC (black carbon / soot) | Dust | Organic | Sulphate	6
GFDL-CM2P1	N/A	BC (black carbon / soot) | Dust | Organic | Sea salt | Sulphate	10
GFDL-CM3	BC (black carbon / soot) | Organic | POM (particulate organic matter) | SOA (secondary organic aerosols) | Sulphate	BC (black carbon / soot) | Dust | Organic | Sea salt | Sulphate	10
GFDL-ESM2G	N/A	BC (black carbon / soot) | Dust | Organic | Sea salt | Sulphate	10
GFDL-ESM2M	N/A	BC (black carbon / soot) | Dust | Organic | Sea salt | Sulphate	10
GFDL-HIRAM-C180	N/A	BC (black carbon / soot) | Dust | Organic | Sea salt |	10

High Performance Data (HPD) Analysis Environment

Mutualized

Jointly *delivered* by
→IPSL laboratories.
Joint *users* (initially):
→IPSL community
Joint *users* (target):
→French Academic community

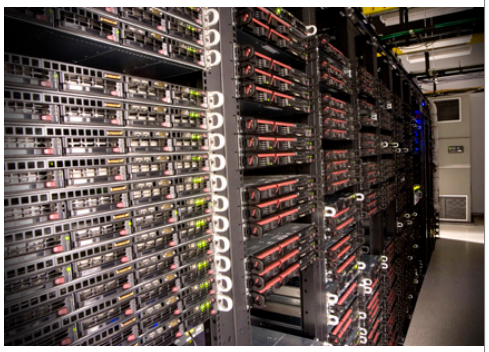
Analysis capabilities

Environmental Data
Compute Service
Web Service Provision for :
→Climate Science
→Earth Observation
→Environmental studies

Access services to ESGF System

Users don't have to find, download, and keep up to date the data they need

CMIP5,
CORDEX
Reanalysis,
Obs4MIPs
...



Big DATA Platform

Collaboration Environment
→ Access to Curated Archive.
→ Large shared "Group Work Spaces"
→ climate analysis enabled system
→ + 1 PB of high performance disk coupled to hundreds of cores configured for analysis

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

prodiguer

Thank you for your attention