

CMIP6 Publication Workflow at GFDL

Sergey Nikonov, V.Balaji, Erik Mason,
Aparna Radhakrishnan



ENGILITY
Your Mission. Our Commitment.

Feb 2015 GO-ESSP Workshop, UK

Outline

- Current GFDL Publishing Workflow
- CMIP5 lessons
- CMIP6 projections
- Addressing shortcomings
- New Elements in GFDL Data Service
- Conclusions

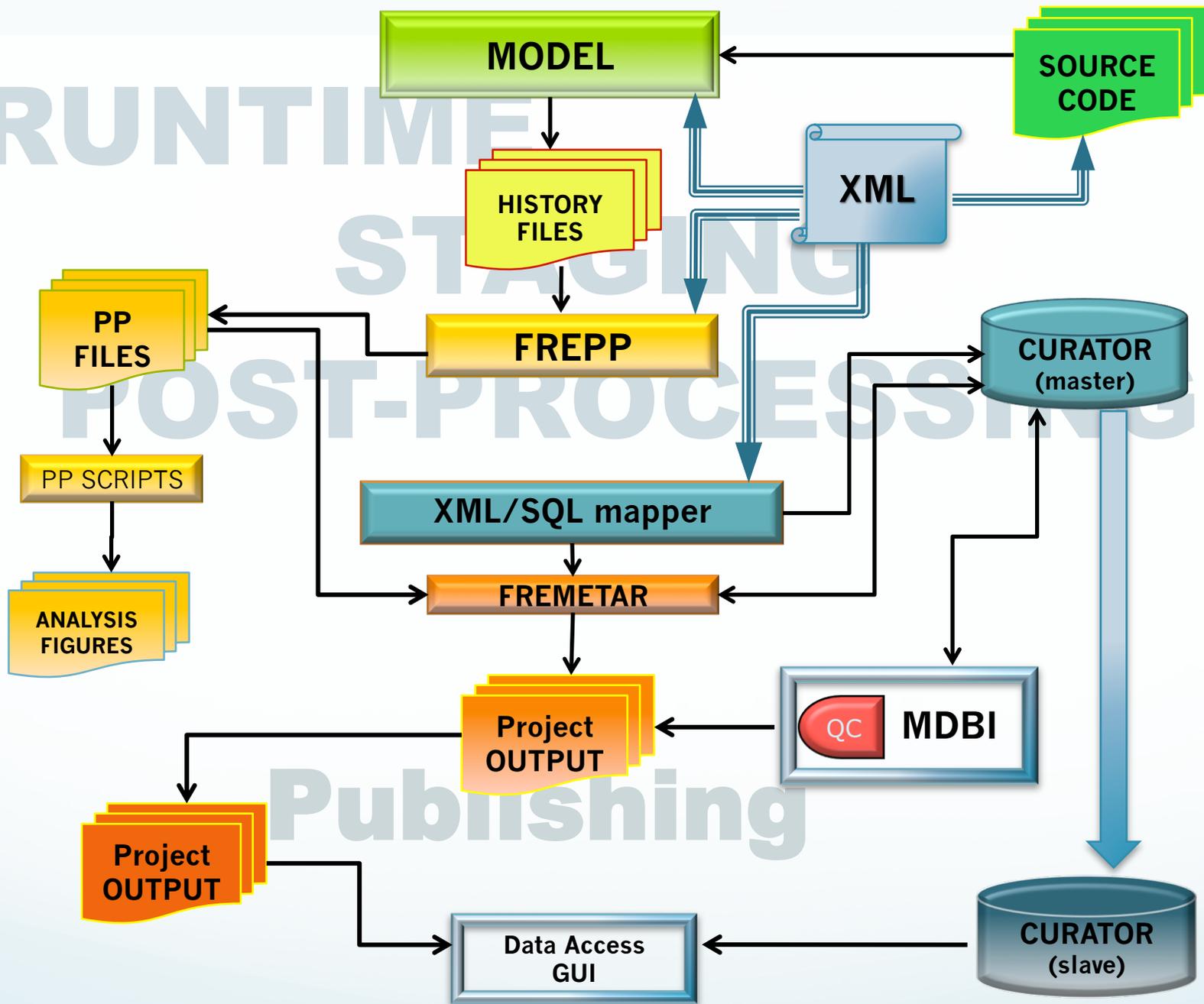
Current GFDL Publishing Workflow

- Data publishing workflow is the part of FRE Curator subsystem which plays a role of the main metadata integrator. It collects, analyses, stores and supports users and applications with comprehensive information about simulations conducted in GFDL. Publishing data is the another important role of FRE Curator.
- GFDL modeling workflow is organized and carried out by FRE (FMS Runtime Environment). It acts under instructions written in XML format and comprises all stages of model simulation from source checkout to postprocessing simulation data output (get source->compile->get input data -> configuring -> build run scripts-> running -> postprocessing)
- FRE Curator makes use of this lucky chance to grab all metadata for simulation from XML, parses it and stores in DB
- CMORizing, QCing, publishing data, Data Portal web apps depend on curator database which serves as a metadata storage.
- Nowadays, with Model Development DB Interface (MDBI) enhancements it became a powerful instrument for researchers allowing real-time runs monitoring, navigation, discovery, comparison, sharing, and analysis of experiments

Current Publishing Workflow (cont.)

- FRE Curator consists of metadata database and set of tools:
 - XML/SQL mapper server
 - Web Interface with set of services
 - fremetar - CMOR-analogue metadata rewriter
 - publishing, maintenance tools
 - MySQL database storing metadata
- Publishing process is implemented in semi-automated manner. Preliminary preparation should be made in DB which includes:
 - setting up variables mapping into project adopted names
 - describing variable bundles (aka CMIP tables)
 - setting up project specific metadata standards descriptors (for Netcdf headers, file names, and directory structure) in specific DB tables
- The only human intervention is needed on QC stage. Even then scientists get this procedure accelerated having integral characteristics for each variable and file (missed values, averages, min/max, variances) which are calculated by fremetar and stored in DB.

RUNTIME STAGING POST-PROCESSING

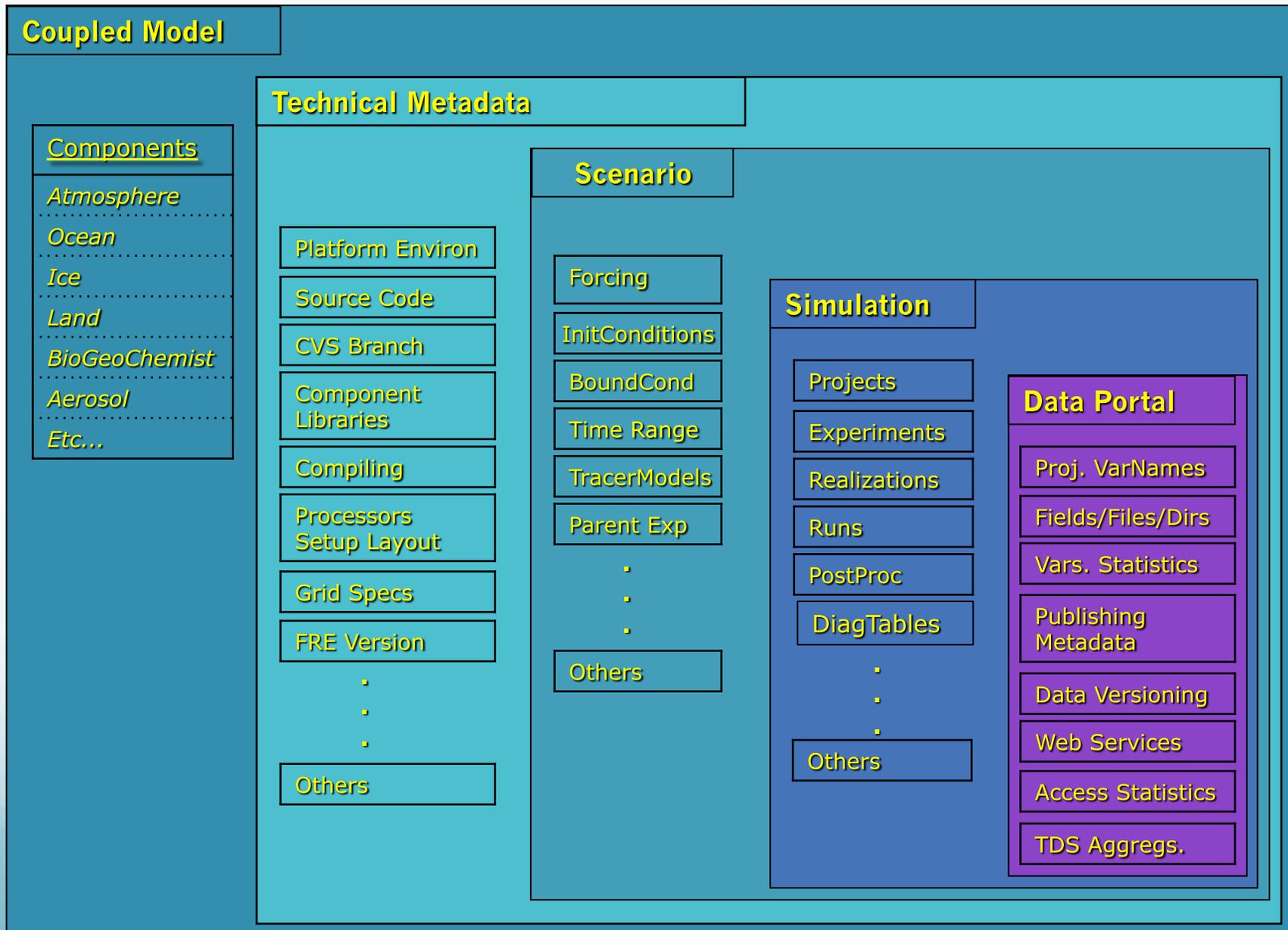


Database curator capability

Database Sections:

- **Model Metadata** section contains models' descriptions
- **Experiment Metadata** Section contains scenarios, experiments, projects, etc.
- **Workflow** section - all about setting up simulation run on HPCS
- **Postprocessing** section defines model diagnostic output and postprocessing type (averaging and output time chunks and years to output)
- **Data Portal** section contains project specific metadata standards descriptors, experiment published on Data Portal, THREDDS aggregations, download statistics

Metadata in curator database



Economical XML/SQL Mapping

- Experiment in DB is represented as a tree of reusable by all experiments entities bound by primary keys
- Every DB entity can be referenced by many different experiments decreasing the size of DB
- Entity is written in DB only if it does not already exist in DB; it eliminates redundancy in DB
- This approach results to
 - small total size of all tables populated from XML: it's 30 MB vs. 600 MB of size of all XMLs mapped into DB (compression factor – 20!)
 - fast and simple experiments comparison.
- Tree structure representing experiments implemented in SQL schema allows to pull from DB all metadata about given experiment by it's triple ids: exper_id, realiz_id, run_id

Model Development DB Interface (MDBI)

Transparent view of the curator database in a user-friendly manner allows:

- Experiment overview (technical, scientific, output metadata)
- Experiment Search (by tripleID, name, owner)
- Experiments comparison (input files, components, output variables)
- Easy, direct access to analysis figures
- HPCS job monitoring years simulated and data readiness
- Online real-time climate monitoring in a course of simulation
- XML generation for given experiment for sharing within community
- QC of variables intended for publishing and control and accounting of publishing process
- Secure/Restricted access to the interface/experiment metadata according to user privacy setup

Model Development DB Interface (cont.)

Navigation

- ACCMP
- AeroCOM IND3
- AeroCom
- AMIP
- APPOSITE
- AWG
- CMIP5 *Project*
- AM3p9
- CM3Z-TEST
- GFDL-AM3-COURSE
- GFDL-CM2p1
- GFDL-CM2p5
- GFDL-CM3 *Model*
 - CM3Z_Control-1860_D1
 - historicalMisc
 - historical *Experiment Realization*
 - CM3Z_D1_1860-2005_AiIForc_H1
 - CM3Z_D1_1860-2005_AiIForc_H2
 - CM3Z_D1_1860-2005_AiIForc_H3
 - CM3Z_D1_1860-2005_AiIForc_H4
 - CM3Z_D1_1860-2005_AiIForc_H5
 - historical_old
 - historicalNat
 - historicalGHG
 - CM3Z_D1_1PctTo4X_I1
 - abrupt4xCO2
 - CM3Z_H1_2006-2100_RCP3_W1
 - CM3Z_H1_2006-2100_RCP6_Y1
 - CM3Z_H1_2006-2100_RCP85_Z1
 - rcp45
 - c48L48_am3p9_1860SST_K1
 - c48L48_am3p9_2000aero_1860SST
 - c48L48_am3p9_2000sulf_1860SST
 - c48L48_am3p9_4XCO2_1860SST_F
 - c48L48_am3p9_H1_4XCO2_M1
 - c48L48_am3p9_H1_SSTplus4K_U1
- amip
- GFDL-ESM2G
- GFDL-ESM2M
- GFDL-HIRAM-C180
- GFDL-HIRAM-C360
- HIRAM-C180
- COOKIE
- Curator-test
- CWG

Experiment Info | Monitoring | Filter | Compare experiments | Show XML | Login | Help

Description | Administration | Platform_Envir | Component_Cc | Checkout_Proc | Compile_Proce | Input_Files | Post_Processir | Run_Descriptio | PP_Directory | Publishing

MODEL: GFDL-CM3
EXPERIMENT: CM3Z_D1_1860-2005_AiIForc_H1

CMIP5/CFMIP5 Tables	Information/Forms	Status for Quality Control
Amon	Quality Control Form FREMetarized Data Published Data	Published available files
3hr	Quality Control Form FREMetarized Data Published Data	Published available files
Omon	Quality Control Form FREMetarized Data Published Data	Not QCed yet
aero	Quality Control Form FREMetarized Data Published Data	Not QCed yet
Oclim	Quality Control Form FREMetarized Data Published Data	Published available files
fx	Quality Control Form FREMetarized Data Published Data	Completely QC'ed and Partially published
Lmon	Quality Control Form FREMetarized Data Published Data	Not QCed yet
Llmon	Quality Control Form FREMetarized Data Published Data	Published available files
dav	Quality Control Form FREMetarized Data	Published available

Model Development DB Interface (cont.)

Quality Control Form

(Variable level)

(Please check the appropriate checkboxes against the files that were quality controlled, and hit the Submit button below this form)

GFDL Model: GFDL-CM3 / Experiment: CM3Z_D1_1860-2005_AllForc_H1 (CMIP Table: Amon)								
S.No.	CF NAME	IPCC VARIABLE NAME	GFDL VARIABLE NAME	MIN	MAX	No. of QC'ed files/Total no. of FREtmetarized files	Input For Quality Control Select/Deselect all <input checked="" type="checkbox"/>	Comments
1	air_pressure_at_convective_cloud_base	ccb	conv_cl_d_base	43173.4	105025	30/30	<input checked="" type="checkbox"/>	<ul style="list-style-type: none"> all files were checked by lwh at 12-6-2011 13:50:33 File ccb_Amon_GFDL-CM3_ami4xCO2_r1i1p1_197901-198312.nc(ccb) was checked by lwh at 7-12-2013 12:45:57
2	air_pressure_at_convective_cloud_top	cct	conv_cl_d_top	6519.17	103503	30/30	<input checked="" type="checkbox"/>	<ul style="list-style-type: none"> all files were checked by lwh at 12-6-2011 13:50:33 File cct_Amon_GFDL-CM3_sstClim4xCO2_r1i1p1_000101-000512.nc(cct) was checked by lwh at 7-12-2013 12:47:2
3	mole_fraction_of_cfc113_in_air	cfc113global	mvf113	0	83.8338	30/30	<input checked="" type="checkbox"/>	<ul style="list-style-type: none"> all files were checked by lwh at 12-21-2011 14:2:58 File cfc113global_Amon_GFDL-CM3_ami4xCO2_r1i1p1_197901-198312.nc(cfc113glob
4	mole_fraction_of_cfc11_in_air	cfc11global	mvf11	0	269.5	30/30	<input checked="" type="checkbox"/>	<ul style="list-style-type: none"> all files were checked by lwh at 12-21-2011 14:2:58
5	mole_fraction_of_cfc12_in_air	cfc12global	mvf12	0	540.045	30/30	<input checked="" type="checkbox"/>	<ul style="list-style-type: none"> all files were checked by lwh at 12-21-2011 14:2:58 File cfc12global_Amon_GFDL-CM3_ami4xCO2_r1i1p1_197901-198312.nc(cfc12globa
6	mole_fraction_of_methane_in_air	ch4global	mvch4	804.563	1755.05	30/30	<input checked="" type="checkbox"/>	<ul style="list-style-type: none"> all files were checked by lwh at 12-21-2011 14:2:59
7	convection_time_fraction	ci	conv_freq	0	1	30/30	<input checked="" type="checkbox"/>	<ul style="list-style-type: none"> all files were checked by lwh at 12-21-2011 13:58:31 File ci_Amon_GFDL-CM3_ami4xCO2_r1i1p1_197901-198312.nc(ci) was

Model Development DB Interface (cont.)

PCMDI Table: Amon

Experiment: CM3Z_D1_1860-2005_AllForc_H1 (exper_id_1XhP9uft45)

Model: GFDL-CM3

Clicking on file name under "Source And Fremetarized Files" will display the Source File location and FREmetarized file locations for quality control.

Source File/Model Output:

Fremetarized File:

Back

Quality Control Form (File level)

(Please check the appropriate checkboxes against the file that was quality controlled, and hit the Submit button below this form)

IPCC Variable Name: ccb / GFDL Variable Name: conv_cld_base (CF name:air_pressure_at_convective_cloud_base)							
IPCC Units: Pa (CMOR Min:-1e+20/ CMOR Max:1e+20)							
(If CMOR configs do not have a min/max specified, the defaults will point to -1e+20/1e+20 respectively)							
S.No.	Source And Fremetarized Files <small>(Please click on the filenames to list the fremetarized and source file locations)</small>	Min	Max	Avg	StdDev	Number of missing_values	Input For Quality Control Select/Deselect all <input type="checkbox"/>
1	ccb_Amon_GFDL-CM3_historical_r1i1p1_186001-186412.nc	45805.8	104735.0	95000.0	6299.48	71732	<input checked="" type="checkbox"/>
2	ccb_Amon_GFDL-CM3_historical_r1i1p1_186501-186912.nc	46421.0	104650.0	94992.9	6293.67	70774	<input checked="" type="checkbox"/>
3	ccb_Amon_GFDL-CM3_historical_r1i1p1_187001-187412.nc	45269.3	104162.0	94977.4	6310.5	72103	<input checked="" type="checkbox"/>
4	ccb_Amon_GFDL-CM3_historical_r1i1p1_187501-187912.nc	44730.7	104507.0	94991.6	6326.3	70580	<input checked="" type="checkbox"/>
5	ccb_Amon_GFDL-CM3_historical_r1i1p1_188001-188412.nc	46374.9	104642.0	94979.0	6288.61	69658	<input checked="" type="checkbox"/>
6	ccb_Amon_GFDL-CM3_historical_r1i1p1_188501-188912.nc	45357.0	104151.0	94982.8	6280.77	71433	<input checked="" type="checkbox"/>
7	ccb_Amon_GFDL-CM3_historical_r1i1p1_189001-189412.nc	45200.5	104479.0	94981.0	6287.55	75084	<input checked="" type="checkbox"/>
8	ccb_Amon_GFDL-CM3_historical_r1i1p1_189501-189912.nc	45597.8	104002.0	94991.5	6277.98	72736	<input checked="" type="checkbox"/>
9	ccb_Amon_GFDL-CM3_historical_r1i1p1_190001-190412.nc	45950.7	104171.0	94985.8	6294.55	73508	<input checked="" type="checkbox"/>
10	ccb_Amon_GFDL-CM3_historical_r1i1p1_190501-190912.nc	46122.0	104220.0	94980.0	6282.0	71067	<input checked="" type="checkbox"/>

CMIP5 Lessons

- We have experienced problems with curator system which was launched just before CMIP5.
- There were some difficulties to convince scientists to use it as they considered it unjustified complicated for IPCC purposes. Later it fully justified itself when they were aware how huge this project.
- It happened mapping variables names into CF and grouping into bundles was a considerable human effort, also setting up metadata standards (DRS) directives in DB required serious manual job (600 variables totally). Bundles were not organized strictly by realm and time frequency as it was in IPCC AR4, that invoked difficulties with variable categorizing in DB.
- Version control needs to elaborate thoroughly especially customers notification about erroneous datasets.

CMIP5 Lessons (cont.)

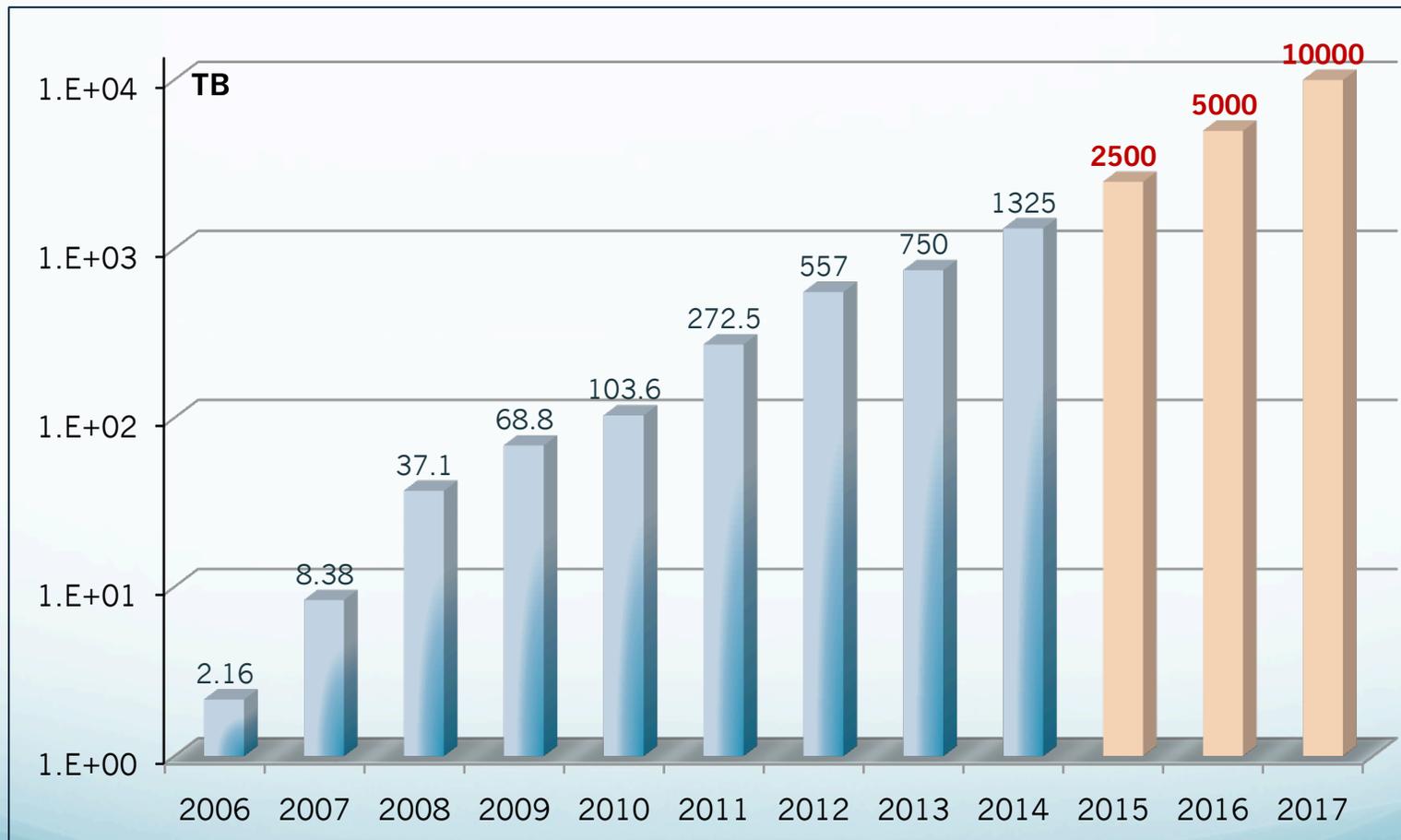
- We produced all files with 5yrs chunks data and as result we have about $8e+5$ files. Some users complained about difficulties to handle such enormous numbers. Also, ESGF does not allow to create wget script with >1000 files (at least in our installed version).
- There were problems with first versions of ESGF which was released then (especially with support). It was not so critical for us as we are running our grown data portal.
- QC was not so easy as scientists assumed. Very time consuming. Considerable part of data was not published due to lack of resources for that.
- GFDL Climate Help Desk receives many similar scientific questions which can be categorized, generalized and posted as FAQ. It will relief the burden on administrator.
- Download was unfair sometimes when one IP occupies whole bandwidth.

Global Download Statistics

- Total volume downloaded for whole GFDL DP history: ~1.3 PB, 250 TB from them via ESGF Node
- Total number of successful requests: ~6.3e+6; ESGF: ~1.3e+6)
- Distinct files: 8e+5, distinct hosts: 6K
- Roughly, every GFDL file is downloaded 6 times and it means GFDL has 6 complete copies of published data over the world
- The IPCC data is ~95% of this numbers
- Very uneven demand - from 1 TB/month to 400 TB/month (1 Gbps). We expect at least quadrupled rate for CMIP6

Cumulative Download

2006-2014 (real) and 2015-2017 (projection)



Preparation to CMIP6

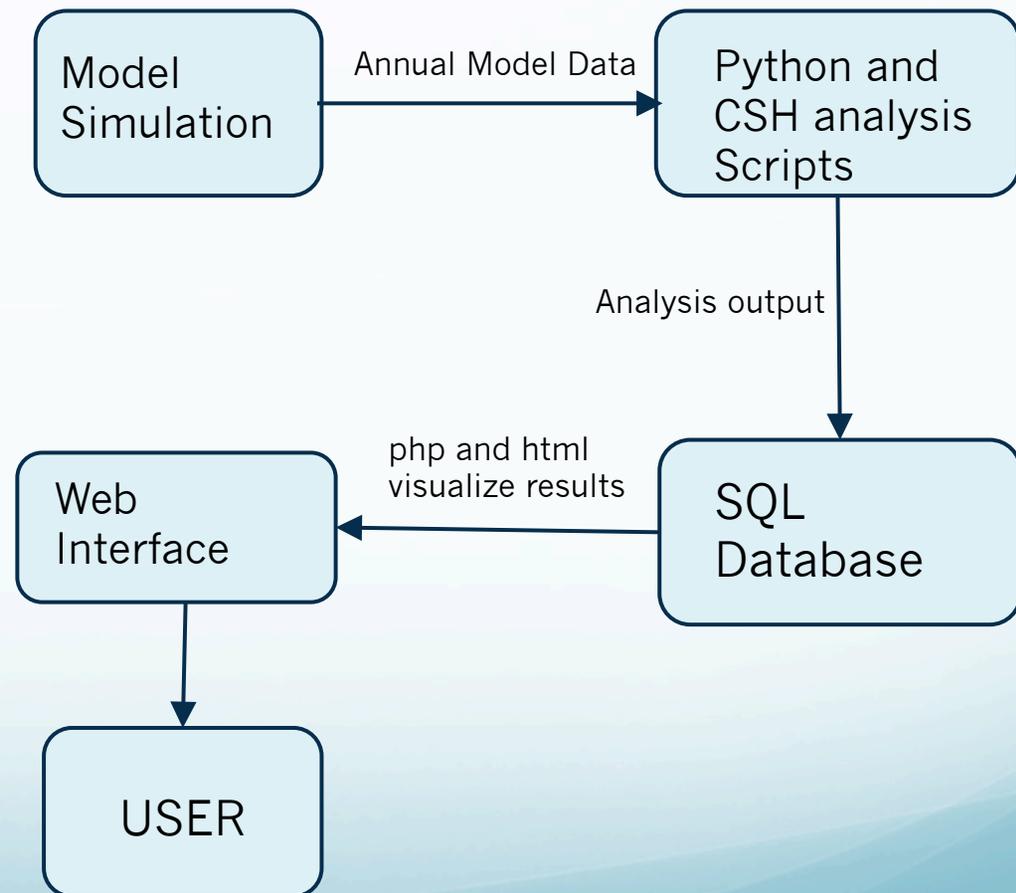
- All components of publishing workflow were revised to make sure they can address serious challenges imposed by CMIP6 immense plans.
- As a result the roadmap of measures for improving was compiled and subdivided into priority importance/urgency categories.
- It comprises more than 50 tasks in different areas of publishing workflow:
 - Transition in models to CF variables names
 - Curator database populating with new variables, clearing bottlenecks
 - SQL/XML mapper – enhancing interaction with users
 - MDBI – new features in search/filtering experiments, comparison capabilities, privacy, online experiment annotation, REST interface elements
 - Additional means for real time runs & climate monitoring
 - Converging publishing processes on ESGF Node and GFDL Data Portal
 - Feeding Metafor XMLs with metadata from curator DB (at least where metadata exists)
 - fremetar (GFDL analog of CMOR)
 - Quality Control procedure
 - New GFDL Data Portal web interface

Realtime Climate Simulation Monitoring

➤ Goal:

Provide a workflow of analysis scripts and tools to help scientists fine-tune their experiments and reduce the CPU cycles and time to achieve their research goals.

➤ Current workflow:

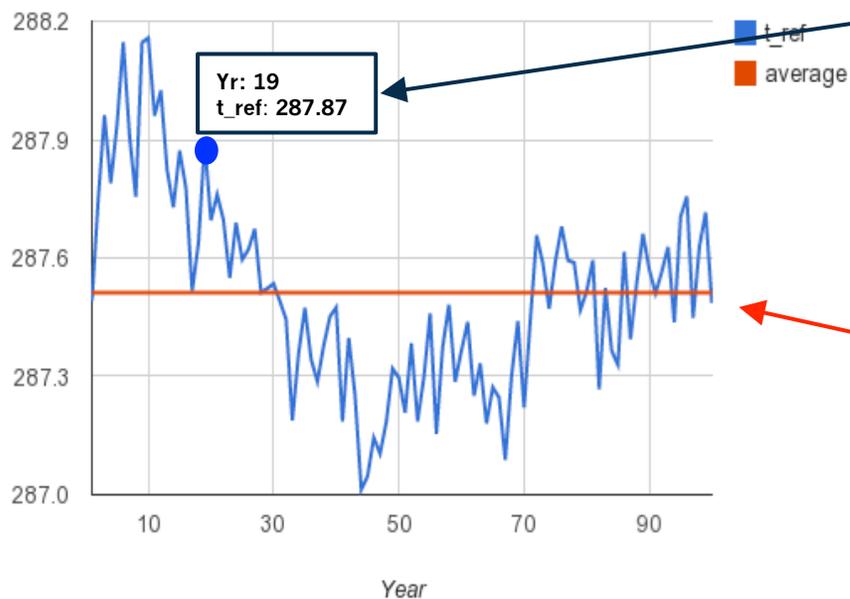


A closer look at the Web Interface

[Global](#) | [Tropics](#) | [NH Extratropics](#) | [SH](#)

Atmosphere

Near Surface Temperature (K)



- Interactive Google Chart API allows users to hover over values and see data points.
- Average of value of the chosen field to date.

Conclusions

- Coupling GFDL publishing process with ESGF and Metafor to prevent double job is very important.
- Have a tool with user interface for configuring project specific metadata in DB (variables, bundles, DRS)
- Implement some rudimental automation in QC (checking variable values limits, variants, averages)
- Intercomparison analysis tools
- Bandwidth increasing, it should be 10 Gbps at least to meet CMIP6 GFDL data volume projection – 700 TB
- Moderator tool for harmonization download streams
- Wiki page with FAQ to reduce burden on GFDL Climate Help Desk and may be non relational DB with textual search



Thanks!

Questions?