Reproducible science at large scale within a continuous delivery pipeline: the BSC vision

Miguel Castrillo
BSC-ES Computational Earth Sciences

ECMWF workshop: Building reproducible workflows for Earth Sciences
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-CNS objectives

- Supercomputing services to Spanish and EU researchers
- R&D in Computer, Life, Earth and Engineering Sciences
- PhD programme, technology transfer, public engagement

BSC-CNS is a consortium that includes

- Spanish Government 60%
- Catalan Government 30%
- Univ. Politècnica de Catalunya (UPC) 10%
### MareNostrum 4

**Total peak performance:** 13,7 Pflops

<table>
<thead>
<tr>
<th>Cluster Type</th>
<th>Peak Performance</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Purpose</td>
<td>11.15 Pflops</td>
<td>1.07.2017</td>
</tr>
<tr>
<td>CTE1-P9+Volta</td>
<td>1.57 Pflops</td>
<td>1.03.2018</td>
</tr>
<tr>
<td>CTE2-AMD</td>
<td>0.52 Pflops</td>
<td>1.11.2019</td>
</tr>
<tr>
<td>CTE3-Arm V8</td>
<td>0.5 Pflops</td>
<td>(????)</td>
</tr>
</tbody>
</table>

**MareNostrum 1**
- 2004 – 42,3 Tflops
- 1st Europe / 4th World
- New technologies

**MareNostrum 2**
- 2006 – 94,2 Tflops
- 1st Europe / 5th World
- New technologies

**MareNostrum 3**
- 2012 – 1,1 Pflops
- 12th Europe / 36th World

**MareNostrum 4**
- 2017 – 11,1 Pflops
- 2nd Europe / 13th World
- New technologies

Access: prace-ri.eu/hpc_acces
Access: bsc.es/res-intranet
MareNostrum 5. A European pre-exascale supercomputer

- **200 Petaflops** peak performance \((200 \times 10^{15})\)

- **Experimental platform** to create supercomputing technologies “made in Europe”

- **223 M€** of investment

Hosting Consortium:
- Spain
- Portugal
- Turkey
- Croatia
Mission of BSC Scientific Departments

**Computer Sciences**
To influence the way machines are built, programmed and used: programming models, performance tools, Big Data, computer architecture, energy efficiency

**Earth Sciences**
To develop and implement global and regional state-of-the-art models for short-term air quality forecast and long-term climate applications

**Life Sciences**
To understand living organisms by means of theoretical and computational methods (molecular modeling, genomics, proteomics)

**CASE**
To develop scientific and engineering software to efficiently exploit super-computing capabilities (biomedical, geophysics, atmospheric, energy, social and economic simulations)
Earth Sciences

Environmental modelling and forecasting, with a particular focus on weather, climate and air quality

- Modeling air quality and Sand and Dust Storms Processes from urban to Global and the impacts on weather, health and ecosystems
- Climate predictions system from subseasonal-to-decadal forecasts

Service Users Sectors:
- Infrastructures
- Solar Energy
- Urban development
- Transport
- Wind Energy
- Agriculture
- Insurance
Design, development, and deployment of Earth science models in close collaboration with the scientific groups aiming to understand and better predict the behaviour of Earth systems.

Research and development of methodologies and tools that allow the running of scientific models in production and taking advantage of the increasing availability and variety of computing resources.
MWT synergies

Models & Workflows

Performance
- Climate Prediction
- Atmospheric Composition
- HPC usage
- User support
- Workflow mgmt.
- Workflows efficiency
- Deployment on new systems
- Reproducibility

Data & diags
- Model development
- Model tools
- HPC resources
- Data processing tools
- High availability for operational systems

Data & diags
- Model development
- Model tools
- HPC resources
- Data processing tools
- High availability for operational systems

Performance
- Climate Prediction
- Atmospheric Composition
- HPC usage
- User support
- Workflow mgmt.
- Workflows efficiency
- Deployment on new systems
- Reproducibility
High Performance Computing in Earth Sciences

• Earth System Models (ESMs) are sophisticated tools with continuously increasing complexity:
  
  • More components of Earth System are included
  
  • Finer Spatial and Temporal resolutions

• This increase in complexity could be developed thanks to the important parallel advances in HPC
Workflows at BSC-ES

Integration tests
Validation
Unit tests
Containers
CVS
......
Workflow managers
Perf. metrics
Data
Our workflow

Develop

Checkout

Run experiment

Continuous Delivery Testing – Reproducibility

Archive & publish

Performance metrics
A workflow for ESM experiments

Model setup
- Configure experiment
- Set up platform-specific files
- Prepare run scripts

Deployment on remote HPC
- Tar and send model files / Synchronize updated files
- Compile the model

Initial data preprocessing
- Choose and deploy the initial data
- Pre-process initial conditions / restarts

Simulation
- Find a proper domain decomposition
- Run simulation
- Assimilation / synchronize jobs

Output data post-processing
- Post-process / normalize output data
- Archive restarts / output files
- Transfer results to local

Verification & diagnostics
- Generate diagnostics
- Generate plots
Workflow managers are **essential** to carry out production experiments in an **efficient** way

- Deal with workflow **complexity**
- Ensure **robustness & portability**
- **Usability** → Scientists more productive
A **versatile** tool to manage Weather and Climate Experiments in diverse Supercomputing Environments:

https://pypi.python.org/pypi/autosubmit
Autosubmit wrapper

Reducing queueing times by wrapping different jobs together
Autosubmit wrappers

Hybrid wrappers
Autosubmit: lessons learned

• Workflows are getting more and more complex

• Workflow managers are required to improve in order to deal with this complexity
Autosubmit visualization

**Grouping** jobs by date, member, chunk, split; or automatically.

**Automatic behavior:**
Collapsing jobs sharing status.

**Hide groups:**
Showing the more relevant information.
Auto-Models

Autosubmit + Model source + Job templates + Auxiliary functions + Utilities
The EC-Earth ESM

EC-EARTH components

Atmosphere GCM: IFS
Land: IFS H-lassel
Vegetation: LPJ
Atmospheric Chemistry and aerosols: TM5
Ocean GCM: NEMO
Joint EC-Earth and ECMWF weather forecast components
Sea-ice: LIM2/3
New EC-Earth components
Marine ecosystem: PISCES
Planned EC-Earth components

Coupler:

Ocean+ICE:

IFS
ECMWF
Auto-EC-Earth branching

EC-Earth SVN

v3.2.2

trunk

v3.2.3

EC-Earth model

Auto-EC-Earth

Auto-EC-Earth Git

EC-Earth model

dev_feature

rXXX-feature

3.2.2_production

dev_feature

trunk

Auto-EC-Earth

BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Auto-Models development

GitLab

Jenkins

feature_branch

master

bugfix

merge request and code review

apply code review suggestions

not approved

reintegrate into master

ok

not ok

master update

run copy of testing suite

approved

every Friday: testing suite running on master

...
Auto-Models development

GitLab board – Agile “Kanban” methodology

Merge requests
**Auto-EC-Earth testing: release tests**

For every version release: run a complete set of tests. Every week: run a smaller set of tests.

<table>
<thead>
<tr>
<th>nord3</th>
<th>CCA</th>
<th>MN4</th>
<th>resolution</th>
<th>type</th>
<th>details</th>
</tr>
</thead>
<tbody>
<tr>
<td>t02c</td>
<td>t00u</td>
<td>t00q</td>
<td>T255L91-ORCA1L75-LIM3</td>
<td>coupled</td>
<td>start from restart</td>
</tr>
<tr>
<td></td>
<td>t00v</td>
<td></td>
<td>T255L91-ORCA1L75-LIM3</td>
<td>coupled</td>
<td>atmos. nudging</td>
</tr>
<tr>
<td></td>
<td>t011</td>
<td></td>
<td>T255L91-ORCA1L75-LIM3</td>
<td>coupled</td>
<td>sppt</td>
</tr>
<tr>
<td>t00s</td>
<td>t00o</td>
<td></td>
<td>T255L91-ORCA1L75-LIM3</td>
<td>coupled</td>
<td>cold start</td>
</tr>
<tr>
<td>t01d</td>
<td>t00z</td>
<td></td>
<td>ORCA1L75-LIM3</td>
<td>nemo</td>
<td>cold start</td>
</tr>
<tr>
<td></td>
<td>t01j</td>
<td></td>
<td>ORCA1L75-LIM3</td>
<td>nemo</td>
<td>cold start ocean nudging</td>
</tr>
<tr>
<td>t01e</td>
<td>t00r</td>
<td></td>
<td>T511L91-ORCA025L75-LIM3</td>
<td>coupled</td>
<td>start from restart</td>
</tr>
<tr>
<td></td>
<td>t01o</td>
<td></td>
<td>ORCA025L75-LIM3</td>
<td>nemo</td>
<td>cold start</td>
</tr>
<tr>
<td>t01b</td>
<td>t00y</td>
<td></td>
<td>T511L91</td>
<td>ifs</td>
<td>cold start</td>
</tr>
<tr>
<td>t00t</td>
<td>t00p</td>
<td></td>
<td>T511L91-ORCA025L75-LIM3</td>
<td>coupled</td>
<td>cold start</td>
</tr>
</tbody>
</table>
Integration tests: testing suite

<table>
<thead>
<tr>
<th>n02c</th>
<th>CCA</th>
<th>Mi4</th>
<th>resolution</th>
<th>type</th>
<th>details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100u</td>
<td>100g</td>
<td>T25SL91-ORCA1L7S-LIM3</td>
<td>coupled</td>
<td>start from restart</td>
</tr>
<tr>
<td></td>
<td>100v</td>
<td></td>
<td>T25SL91-ORCA1L7S-LIM3</td>
<td>coupled</td>
<td>ATM nudging</td>
</tr>
<tr>
<td></td>
<td>1011</td>
<td></td>
<td>T25SL91-ORCA1L7S-LIM3</td>
<td>coupled</td>
<td>spt</td>
</tr>
<tr>
<td></td>
<td>100s</td>
<td>100o</td>
<td>T25SL91-ORCA1L7S-LIM3</td>
<td>coupled</td>
<td>cold start</td>
</tr>
<tr>
<td></td>
<td>101d</td>
<td>100z</td>
<td>ORCA1L7S-LIM3</td>
<td>nemo</td>
<td>cold start</td>
</tr>
<tr>
<td></td>
<td>103j</td>
<td></td>
<td>ORCA1L7S-LIM3</td>
<td>nemo</td>
<td>cold start</td>
</tr>
<tr>
<td></td>
<td>101e</td>
<td>100v</td>
<td>TS11L91-ORCA02SL7S-LIM3</td>
<td>coupled</td>
<td>start from restart</td>
</tr>
<tr>
<td></td>
<td>101o</td>
<td></td>
<td>ORCA02SL7S-LIM3</td>
<td>coupled</td>
<td>cold start</td>
</tr>
<tr>
<td></td>
<td>101b</td>
<td>100y</td>
<td>TS11L91</td>
<td>nemo</td>
<td>cold start</td>
</tr>
<tr>
<td></td>
<td>101x</td>
<td>100p</td>
<td>TS11L91-ORCA02SL7S-LIM3</td>
<td>coupled</td>
<td>cold start</td>
</tr>
</tbody>
</table>
Reproducibility

- Earth models: variety of spatial resolutions, configurations and running environments.
- Scientific codes are often in a near-constant state of development as new science capabilities are added and requirements change.
- If we want to study the model response under some scientific changes, we have to ensure that computational changes do not affect the results.
- We need a method to evaluate the computational efficiency of our models:
  - When the hardware changes, the number of resources or the configuration changes.
Reproducibility

Accuracy
(be close to a reference)

Performance
Accuracy & Reproducibility

Speed
FP Precision

Performance \rightarrow Improved using optimization methods but reduce Reproducibility and Accuracy.
Reproducibility & Accuracy \rightarrow Improve using control methods but reduce Performance.

Reproducibility
(be similar across configurations)

Performance
(use resources efficiently)

Find options to control the tradeoffs among accuracy, reproducibility and performance.
Methodology: EC-Earth CMIP6 case

In comparison:
- Platforms
- Compilers
- Domain Decomposition
- Model options
Reproducibility: Kolmogorov-Smirnov
Methodology

Performance
- (no-prec)-O3 -xHost -r8 -ipo -prof-use -no-prec-div -no-prec-sqrt
- (prof-use)-O3 -xHost -r8 -ipo -prof-gen → -O3 -xHost -r8 -ipo -prof-use
- (ipo)-O3 -xHost -r8 -ipo
- (O3)-O3 -xHost -r8
- (O2)-O2 -xHost -r8
- (no-fma_fz)-O2 -no-fma -ftz -r8
- (fp-except)-O2 -fp-model except -no-fma -ftz -fpe0 -r8
- (fp-precise)-O2 -fp-model precise -fimf-arch-consistency=true -no-fma -fpe0 -r8
- (fp-strict)-O2 -fp-model strict -fimf-arch-consistency=true -no-fma -fpe0 -r8
- (O1)-O1 -fp-model strict -fimf-arch-consistency=true -no-fma -fpe0 -r8

Accuracy
Conclusions: EC-Earth model

• “Standard” flag configurations for performance and FP precision obtain the best results.
  • -fp-model precise -fpe -no-fma -O2 -xHost -r8
    • Good performance, good precision, avoid fp errors, catch fpe exceptions.
  • -O2 -xHost -r8
    • Better performance (6%), good precision.

• Aggressive optimizations (O3, ipo, prof-use) do not improve the performance.
  • Some issues may avoid additional optimizations (loop dependences, non vectorization, MPI overhead ...).

• Strict FP control does not improve the precision and reduce the performance up to 6%-12%.

• Using approximations for FP operations (no-prec-div/sqrt) do not improve the performance and reduce the precision and reproducibility dramatically.
Containerisation

- **Common recurrent problem in scientific environments:** reproducibility between different environments.

- **Containers:** make possible the virtualization at the operating system level.
• **Proposed technologies:** Singularity and Shifter, both designed to be used in HPC environments:

![Singularity and Shifter logos]

• **Main questions:**
  - What to containerize? The whole workflow, just the model?
  - Analyze the impact of containers within the HPC environment. ¿Computational performance?
MPMD case with singularity - Challenge: isolation level

- **Isolate** the complete environment is **complex** (network setup, node visibility, compiler license, etc). Then:

  ```bash
  ./mpirun -np N singularity exec container.img exe1 :
  -np M singularity exec container.img exe2
  ```

- The executables need **binding** dynamic libraries folder and other system folders, so the call should be like:

  ```bash
  mpirun -np N singularity --bind /apps:/apps /opt:/opt /lib64:/lib64 exec container.img exe1 :
  -np M singularity --bind /apps:/apps /opt:/opt /lib64:/lib64 exec container.img exe2...
  ```
Performance metrics & performance reproducibility

Climate predictions have complex workflows. New metrics are needed to evaluate the computational efficiency.

- The simulation does not only involve the execution of a model during a sequence of time steps (represented by the sim job).
- The experiment adds complexity in the horizontal for ensembles (members with perturbations in the initial conditions).
- The experiment adds complexity in the vertical, running long simulations divided into chunks and including pre- and post-processing.

<table>
<thead>
<tr>
<th>Metric</th>
<th>used to evaluate…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallellization</td>
<td>Number of parallel resources allocated</td>
</tr>
<tr>
<td>Simulation Year Per Day (SYPD)</td>
<td>How efficient is your run job per each year of the simulation</td>
</tr>
<tr>
<td>Core-hours Per Year (CHPY)</td>
<td>How efficient is your run job with respect to the number of parallel resources used</td>
</tr>
<tr>
<td>Run Time (Average and Total)</td>
<td>How much time are per job and all the jobs in the critical path running</td>
</tr>
<tr>
<td>Queue Time (Average and Total)</td>
<td>How much time are per job and all the jobs in the critical path waiting in the queue</td>
</tr>
<tr>
<td>Post Run T. (Average and Total)</td>
<td>How much time are per job and all the jobs in the critical path running the post</td>
</tr>
<tr>
<td>Post Queue T. (Average and Total)</td>
<td>How much time are per job and all the jobs in the critical path waiting in the queue the post</td>
</tr>
<tr>
<td>Actual SYPD</td>
<td>How affect queue time to the complete experiment, from the first to the last run job</td>
</tr>
<tr>
<td>Real SYPD</td>
<td>How affect other issues (quota, jobs failing, rerun…) the time to complete an experiment</td>
</tr>
<tr>
<td>Energy Cost Per Year (JCPY)</td>
<td>How much energy is needed per each year of simulation</td>
</tr>
</tbody>
</table>

- The execution of a coupled model is complex. Different components run in parallel, exchanging some information and adding some extra overhead (Communication and Interpolation Time). Load imbalances produce extra waiting time as well.
- The execution of each component could involve some irregular extra overhead, such as output processing or calculations occasionally done, such as radiation.
The Earth System Grid Federation (ESGF): international collaboration serving the World Climate Research Programme's (WCRP) Coupled Model Intercomparison Project (CMIP) and supporting climate and environmental science in general.

- Web interface to the majority of experimental datasets used in model intercomparison projects.
- Data published have PIDs and DOI, and are heavily quality controlled.
- Data can be discovered and accessed through THREDDS.
- Tenths of PB of “Earth data” available.
Some current issues

• Initial **data version control**: Need of a centralised and accessible system to share model initial data. Facilitate reproducibility and save space by incremental diff.

• Containers: Difficult to **isolate HPC environment**. Is it needed? Is it enough with statically compiled binaries?

• Model workflows are strongly based on data movement. Not so straightforward to implement **unit tests**.

• **Science reproducibility** (not b2b) still expert **human** intervention.
Thank you

The ESIWACE project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 675191

miguel.castrillo@bsc.es
Methodology: Reproducibility Test

• Accuracy and reproducibility
  • Reichler-Kim normalized index
    • Calculate for each variable a normalized error variance $e^2$ by squaring the grid-point differences between simulated and observed climate
      $$e^2_{vm} = \sum_n \left( w_n \left( \bar{x}_{vmn} - \bar{o}_{vn} \right)^2 / \sigma^2_{vn} \right)$$
    • Ensure that different climate variables receive similar weights when combining their errors
      $$I^2_{vm} = e^2_{vm} / \bar{e}^2_{vm}$$
    • Mean over all climate variables
      $$I^2_m = \overline{I^2_v}$$
  • Kolmogorov-Smirnov (KS) test
    • Compare two five-member ensembles and determine whether the two ensembles are statistically indistinguishable from one another.
    • Since no prior assumption can be made on the underlying statistical distribution of the samples, the non-parametric KS test is suitable for the evaluation.
Methodology: Reproducibility Test

• About the reproducibility methodology
  • F. Massonnet, M. Ménégoz, M.C. Acosta, X. Yepes-Arbós, E. Exarchou, F.J. Doblas-Reyes

• About the reproducibility test for EC-Earth
  • Adapted by Philippe Le Sager for EC-Earth community
  • https://github.com/plesager/ece3-postproc#reproducibility-test

• About the reproducibility results and how to collaborate
  • Contact Mario (mario.acosta@bsc.es) and Philippe (sager@knmi.nl)
  • Provide feedback: https://dev.ec-earth.org/issues/533
  • People collaborating: Uwe Fladrich (SMHI), Jose M. Rodríguez (AEMET), Paul Nolan (ICHEC)
Methodology: CPMIP Metrics

• About the CPMIP Metrics
  • V. Balaji et al. 2017

• About the adaptation for EC-Earth
  • Adapted by Uwe Fladrich and Domingo Manubens for EC-Earth
  • Adapted by Philippe Le Sager for common CCA (ECMWF) experiments
  • Adapted and extended by Mario Acosta and Domingo Manubens for MN4 (BSC) using a workflow manager

• About future plans
  • Improvement and Portability of CPMIP metrics for ESMs in IS-ENES3 (Mario Acosta, Eric Maisonnaive...).
  • Work in collaboration with the community (V. Balaji, Pier-Luigi...)

• About the CPMIP results and how to collaborate for CMIP6 runs
  • Contact Mario (mario.acosta@bsc.es) and Philippe (sager@knmi.nl)
  • Provide feedback: https://dev.ec-earth.org/issues/532
  • People collaborating: Uwe Fladrich (SMHI), Jose M. Rodríguez (AEMET), Paul Nolan (ICHEC)
• What it is:
  • https://esgf-node.llnl.gov/search/esgf-llnl/
  • web portal to download climate (CMIP5-6, SPECS, obs4MIPS, CORDEX,...) data
  • ingests very standardized data (CMOR-$proj)

• There are data nodes and index nodes:
  https://www.google.com/maps/d/u/0/viewer?mid=1qgItlbd11j6wzirf6vo2Zq_uYqaP4baQ&ll=43.83402373129155%2C14.074149978563128&z=4

  • Data nodes **host the data** and have a thredds server displaying the data hosted locally
  • https://esgf.bsc.es/thredds/catalog/catalog.html
  • Indexes have a landing page allowing to search