Big Data for climate and air quality

4th BDEC Workshop, 16-17 June 2016, Frankfurt

Francesco Benincasa[i], Pierre-Antoine Bretonnière[i], Francisco Doblas-Reyes[i,ii], Kim Serradell[i]

[(i)] Earth Sciences Department, Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS)[1]
[(ii)] Institució Catalana de Recerca i Estudis Avançats (ICREA)[2]

Introduction

Weather, climate and air quality data are dramatically increasing in volume and complexity, just as their users rapidly increase in number and diversity[3]. This suggests a new paradigm of more user-friendly and fast access to these data to ensure that society can reduce vulnerability to weather extremes, air quality events, and climate variability and change, while at the same time exploiting opportunities for a better management of sensitive systems. This community offers a perfect example of the famous four Vs in Big Data: volume, velocity, veracity and variety. The bulk of weather, climate and air quality data is created by process-based models of the Earth system and satellite instruments. Improvements in these models take into account an increasing number of physical phenomena, leading to codes that require larger computers, with increased resolution and more processes that imply more variables to both analyse and disseminate. As a consequence, data volumes have become an increasingly large problem that can only become worse in the exascale era. Supercomputers are taking care of providing enough fuel for these simulations to be produced, provided that the codes can run efficiently. At the same time, care has to be taken of the way the resulting outputs are handled, analysed and reduced, both during the simulation and once the data is stored to be analysed and/or disseminated. Critical steps such as online post-processing, the offline data mining and dissemination of the simulated datasets, often with different formats and standards, and the observational data used to validate them have Big Data requirements. Multiple downloads (for instance, by a range of users with slightly different interests) and redundant transfers take a substantial amount of time during these steps when using conventional methods. Hence, methods to reduce the data traffic have become essential.

Towards common solutions

The BSC-CNS Earth Sciences department is tackling some of the challenges described above at different levels, both from the HPC and Big Data sides.

The *Autosubmit*[4] workflow manager has been developed to handle multi-model simulations on a range of supercomputers. As a climate prediction experiment is made of a number of jobs (individual start dates and ensemble members) running independently and, if possible, simultaneously by wrapping together ensemble members for different start dates, such a tool is required. The simultaneous running of an ensemble allows a selection of the relevant ensemble members in such a way that those considered redundant according to specific metrics could be stopped, saving a substantial amount of resources. However, this is done at the expense of implementing these metrics in parallel, contrary to what has been done up

[1] http://www.bsc.es/earth-sciences
[2] https://www.icrea.cat/en
[3] Overpeck et al. (2011), Science
[4] https://earth.bsc.es/wiki/doku.php?id=tools:autosubmit

to now by the community. Autosubmit also allows a user to run the same experiment on a range of HPC platforms in a transparent way, bringing back to the storage chosen by the user without manual intervention. With this configuration the load of very expensive climate prediction experiments (for instance, high-resolution decadal predictions[5]) can be distributed across several HPC platforms, which has opened important questions about experiment reproducibility across platforms and running environments.

Metrics and diagnostics have been implemented using the R package *s2dverification*[6], which performs analytics via performance indicators of climate models. This package propagates the metadata during the data processing so that the final visualization, when required, is appropriately referenced, ensuring again the reproducibility of the results. S2dverification can analyse data available either locally or remotely (for instance, from ESGF[7]) and can also be used online as the model runs.

An example of the increasingly tight link between Big Data and Extreme Computing in Earth sciences is the current diagnostics workflow of an usual EC-Earth[8] climate model execution run by Autosubmit on a supercomputer (2000 cores per member). The model integrates an open source C++ I/O server named XIOS[9], which is in charge of moving outputs to the storage (approx. 140 Gb/year simulated). Diagnostics currently are computed completely offline on fat-node cluster, retrieving data from the archive and providing results for user analysis. This workflow implies a large data traffic, the need of a cluster for diagnostic and metric computation and delays to provide relevant data to users. Our group aims at re-designing the workflow to produce as many diagnostics and metrics as possible while the experiment is running. The need for such an approach increases when model resolution increases as in the example above and needs to be addressed well before the advent of exascale machines. This challenge will be addressed by adding an Analytics as a Service (AaaS) layer based on the PyCOMPSs/COMPSs[10] system developed at the BSC-CNS, the whole experiment still being managed by Autosubmit. This layer should be built on top of the XIOS I/O server on the HPC side (computing nodes). The result chain will be model-I/O server-AaaS all in-place. This approach will be very convenient in PRIMAVERA[11] project where BSC will run a decadal prediction simulation using high-resolution global climate model producing petabytes of output.

As well as the design and development of technical solutions, another key point is the need to foster the collaboration in the weather, climate and air quality communities in both Big Data and HPC, as well as with other scientific communities to define standards in terms of formats, protocols, ontologies etc. through the participation in initiatives like the Centre of Excellence in Simulation of Weather and Climate in Europe (ESiWACE)[12] or the Research Data Alliance (RDA)[13]. In this sense, the department is leading the creation of an RDA Interest Group on Weather, Climate and Air Quality[14].

[5] http://www.geosci-model-dev-discuss.net/gmd-2016-78/

[6] https://earth.bsc.es/wiki/doku.php?id=tools:s2dverification

[7] http://esgf.llnl.gov/

[8] http://www.ec-earth.org/

[9] http://forge.ipsl.jussieu.fr/ioserver/wiki

[10] https://www.bsc.es/computer-sciences/grid-computing/comp-superscalar

[11] https://www.primavera-h2020.eu/

[12] https://www.esiwace.eu/

[13] https://rd-alliance.org/

[14] https://rd-alliance.org/bof-weather-climate-and-air-quality-ig.html