Monitoring of Climate Model Ensembles

Project description TNM090

Background

Modern climate models (CMs) are used to understand how the Earth's climate evolves over time and how it responds to changed conditions such as the increasing amount of CO_2 in the atmosphere. In order to account for uncertainties in both the underlying mathematical model as well as the CM implementation, scientific experiments usually consist of so called ensemble runs. This means that the CM is run several times with different data sets and the results are analysed with statistical methods.

As CMs are computationally expensive applications, scientific experiments need large computing resources, as provided by high-performance computing (HPC) systems, or supercomputers. This need is amplified by ensemble runs, which further multiply the resource requirements. Even so, comprehensive climate simulations (of a few hundreds of years) are long running, taking up to a few month to complete.

From a practical perspective, it is important to monitor the progress of a certain scientific experiment while it is running on an HPC system. Monitoring allows to observe the progress of the experiment and to react to erroneous conditions. Such conditions might emerge for technical reasons or because of incorrect physical configuration of the CM.

Project Outline

In order to provide a monitoring system for climate model ensemble running on a HPC system, two components are to be developed in this project: A simple protocol for gathering the monitoring data from the HPC system and an application to visualise it on a remote system. Both parts are described in more detail hereafter:

Monitoring Data Protocol

While the CM ensemble is running on the HPC system, monitoring data can be collected by the runtime environment of the model. The kind and amount of monitoring data can vary a lot, but a few typical scenarios can be defined:

- Simple numbers, such as the number of completed time steps (zero dimensions)
- Time dependent numbers, such as some temperature value that changes over time (one dimension)
- Maps of geographically distributed values, such as a global temperature map (two dimensions)

The monitoring data protocol should define the kind of data that the CM runtime environment collects from the running model and the format that this data is stored in. Note that the implementation of the actual data collection is not part of this project. Regarding the data storage, simple and clear solutions, such as plain file storage, are preferred. It is important to keep in mind that the data is collected not only from a single model run, but from several ensemble members.

The second part of the protocol should define how the monitoring data is transferred from the HPC system to the visualisation location. This needs to take into account the usually limited number of underlying data transfer protocols available on HPC systems (for security reasons). Typically, access is only provided through Secure Shell (ssh), but other options might be available.

Monitoring Data Visualisation

Once the monitoring data is (periodically) collected from the HPC system, the visualisation subsystem should provide a graphical representation of the data. The main objective of this part of the project is wide accessibility of the data displays. Hence, different technical solutions are conceivable, such as a web service or a standalone desktop of mobile application. Again, a relatively simple design based on established components is preferable. Provision of required infrastructure (e.g. for a web service) is not part of the project.

The intention of the monitoring data visualisation is to provide the researchers with easy access to vital data about the progress of CM experiments and allow for an rapid assessment of the status.

Requirements and Priorities

The project should focus on state-of-the-art solutions for providing CM monitoring data to climate scientists, thus creating added value to modelling work flow. The intended outcome can be outlined as

(1A) Data defining part of the monitoring data protocol

(1B) Transfer part of the monitoring data protocol

(2) Design and implementation of the visualisation subsystem

of which (2) and (1A) have the highest priority (in that order). It is understood that (1A) and (1B) require some insight into the work flow of CM experiments and it is proposed to maintain a clear interface between the CM world and the scope of the monitoring tool in order to minimise initial learning effort.

In more general terms, the design and implementation of the monitoring solution should follow most of the following rules:

- Simple design, based (as much as possible) on established solutions
- Open source-type license in order to allow access and contributions
- Flexibility and modularity as much as needed
- Implementation based on Linux/Unix
  (with the possible exception of a potential mobile application)
- Good integration into the climate scientist's work flow
- Good software engineering practices, particularly adopted to scientific software development