Climate Applications in the EELA Project

R. Abarca[1], C. Baeza[1], M. Carrillo[2], J. Casado[3], A. Cofiño[4], I. Dutra[5], F. Echeverría[1], J.M. Gutiérrez[4], R. Miguel[2] y R. Mayo[3]
(on behalf of the EELA Project)

[1] UdeC (Chile) [2] SENAMHI (Peru) [3] CIEMAT (Spain) [4] UniCan (Spain) [5] UFRJ (Brazil)

The EELA Project (E-infrastructure shared between Europe and Latin America) aims to build a digital bridge between the existing initiatives in Europe and Latin America through the creation of a collaborative network which will share a Grid infrastructure for supporting the deployment and test of applications.
Of main interest in the Project are the Climate Applications and, in particular, the analysis and understanding of the most important patterns and phenomena in the atmospheric and oceanic variability (El Niño). This problem requires the efficient access and analysis of the climate data bases and the simulation of numerical oceanic and atmospheric models, the number and size of which is quite large. It is also necessary doing rightly statistical analysis (data mining) for discovering significance patterns and the relationship among them.
Besides, the long term effects on the Climate Change will be studied in these phenomena. For doing this, already performed global predictions will be combined with the results of new high-resolution predictions for the region of interest just integrating complex numerical Climate models under the conditions fixed by different forced scenarios.
The use of Grid technologies allows sharing storage and computing resources in a transparent way, just solving in a brief time some problems that last for days nowadays

OBJECTIVES

Modern climate science deals with different sources of geographically distributed observational data (surface, atmosphere, ocean, etc.) stored in different platforms and formats
→
Efficient problem-driven statistical analysis tools are required for discovering knowledge, or useful information, within the huge amount of information (Data mining and machine learning techniques)

There is an increasing need of computing power required for climate applications (supercomputers or high-performance computer networks with parallel computing techniques and storing of a huge amount of data)
→
Grid technology offers a solution to these problems
→
But only a limited number of initiatives using grid technology have appeared in the last few years (Earth System Grid ESG)

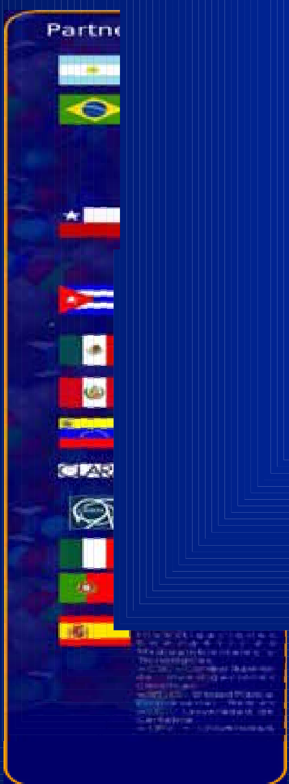
Climate simulation requires running the same model over different initial data sets (production) or running different configurations (parameterizations) of the model over the same initial data
→
Datasets are distributed among different weather services and research laboratories and data mining algorithms are computing intensive, thus requiring parallel processing
→
An efficient development of this category of applications for the GRID environment requires middleware components for application-performance monitoring, efficient distributed data access and specific resource management

Typical climate applications usually require a set of processes to be run in cascade:

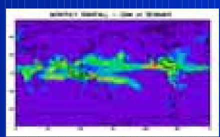
Simulation of atmosphere/ocean models   Efficient data access to observations and previous simulations   Data analysis and mining applications.

We'll do the integration of these components into the EELA testbed.

Contract No: IST-02649
Project Type : SPECIFIC SUPPORT ACTIONS
Start Date: 1/01/2006
Duration: 24 Months
Total Budget: 2,568,320 Euro
Funding from EC: 1.700.000 Euro
Funding from CIEMAT: 400,000 Euro
Total Effort in Person-Month: 1109

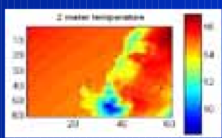


GLOBAL CLIMATE SIMULATIONS (SENAMHI – Peru)



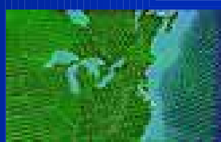
- The Community Atmosphere Model (CAM) is the latest in a series of global atmosphere models developed at NCAR for the weather and climate research communities [R1]
- Also serves as the atmospheric component of the Community Climate System Model (CCSM)
- Is a numerical model that uses governing equations of atmosphere to make climate forecast for long periods of time (centuries)
- Used in a global scale at 300 Km of resolution simulating the climate system of the last fifty years in a global scale by SENAMHI
- An ensemble of different simulations from different initial conditions can be produced as different GRID jobs to characterize the model climatology and can be also run with different forcing emission scenarios to analyze climate change.

REGIONAL CLIMATE APPLICATIONS (UdeC – Chile)



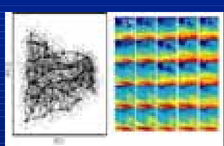
- The PSU/NCAR mesoscale model (known as MM5) and Weather Research and Forecasting (WRF) version are limited-area models designed to simulate or predict regional atmospheric circulation
- Can work with nested domains with different resolutions and require as input the boundary conditions from a global model (e.g., the CAM model)
- Are supported by several pre- and post-processing programs (the MM5 modelling system) and can been used as a computing performance benchmark.
- Regional models are highly dependent on the specific parameterizations chosen for different sub-grid physical phenomena resolved explicitly
- An optimal tuning of the model for a given region requires running an ensemble of simulations with different combinations of model parameters (e.g. running slightly different models with the same initial files).

DISTRIBUTED DATASETS AND ACCESS



- The output of global and regional meteorological models is stored in particular binary meteorological formats (netCDF, GRIB…) and different simulations are geographically distributed among different centres.
- But sometimes partially access to the datasets since only a subset of the original data is used (a certain geographical region) is only needed.
- A distributed inventory system becomes a requirement to allow data producers (weather services, laboratories, etc) to expose metadata information to be harvest by users using any common protocol like HTTP and allow a universal access.
- The THREDDS (Thematic Realtime Environmental Distributed Data Services) [R2] project is developing middleware to bridge the gap between data providers and data users simplifying the discovery and use of scientific data.
- ESG Project, has made an initial attempt to gridify this technology. To this aim, OpenDAP data servers are included within the grid infrastructure and data enters the grid storage elements when they are first requested from OpenDAP servers.

DATA MINING APPLICATIONS: SELF-ORGANIZING MAPS (UniCan – Spain)



- Unsupervised clustering techniques allow partitioning the simulation databases, producing realistic weather or climate models of great variability governing the global dynamics.
- Self-Organizing Maps (SOM) are especially suitable for high dimensional data visualization and modelling.
- It uses unsupervised learning (no domain knowledge is needed and no human intervention is required) for creating a set of prototype vectors representing the data
- A topology preserving projection of the prototypes from the original input space onto a low-dimensional grid is carried out. Thus, the grid can be efficiently used for extracting data features, clustering the data, etc.
- The simplest form for parallelizing the SOM algorithm is splitting up the data between different processors, but after each complete cycle the slaves must send the prototypes to the master, which computes them up, sending the final centres back to the slaves.
- An alternative for distributing computational resources: a replicated (or centralized) prototype vectors with MPI.

REFERENCES

[R1] Collins W.D. et al. Journal of Climate, 2006, **19** 11, 2122–2143.
[R2] THREDDS, http://www.unidata.ucar.edu/projects/THREDDS/),
[R3] Lawrence R.D. et al. Data Mining and Knowledge Discovery, 1999, **3**, 171-195

Contract no. 026409

Project Web Site http://www.eu-eela.org