IL NUOVO CIMENTO Vol. 28 C, N. 2 Marzo-Aprile 2005
DOI 10.1393/ncc/i2005-10189-x

G-AQFS (Grid Air Quality Forecast System): An experimental system based on GRID computing technologies to forecast atmospheric dispersion of pollutants(*)

G. P. Marra([1])([2]), I. Schipa([1]), G. Aloisio([2]), M. Cafaro([2]), D. Conte([1])([2])
C. Elefante([1]), C. Mangia([1]), M. Miglietta([1]), U. Rizza([1])
and A. Tanzarella([1])

([1]) *ISAC-CNR, Sezione di Lecce - Lecce, Italy*
([2]) *Center for Advanced Computational Technologies/ISUFI, Università di Lecce - Lecce, Italy*

(ricevuto il 2 Febbraio 2005; approvato il 30 Maggio 2005; pubblicato online il 23 Settembre 2005)

Summary. — Aim of this work is the implementation of an integrated system for modeling air pollutants dispersion using GRID computing technologies. The system includes two meteorological models, emission pre-processors and two dispersion models for inert and photochemical pollutants. Both meteorological and dispersion models use models which run on the European scale for the initial and boundary conditions. As a test, the modeling system has been applied to the Salento Peninsula, located in the south-east corner of Italy, in the Mediterranean central area with complex meteorological conditions.

PACS 02.70.-c – Computational techniques.
PACS 92.60.Sz – Air quality and air pollution.
PACS 07.05.Tp – Computer modeling and simulation.
PACS 01.30.Cc – Conference proceedings.

1. – Introduction

The management of air quality has become, over the last few decades, a major problem for decision makers. For the concentration of some atmospheric pollutants has drastically decreased due to adequate emission strategies. However, the relation between emission and atmospheric concentrations can be quite complex due to the non-linear character of chemical transformations. This is the case of photochemical pollution, where non-linear chemistry is combined with meteorological processes, that cause great hourly, daily and

(*) Paper presented at CAPI 2004, 8° Workshop sul calcolo ad alte prestazioni in Italia, Milan, November 24-25, 2004.

seasonal variations. Therefore, a combined modeling system, that couples atmospheric flows with dispersion and chemistry is fundamental to our understanding of the physical and chemical mechanisms that lead to the accumulation of tropospheric ozone. The emerging Grid technology provides the necessary resources to perform complex atmospheric and climate simulations. The Computational Grid is a collection of distributed, possibly heterogeneous resources, which can be used as an ensemble to execute large-scale applications. By using these resources, it is possible to access information about the grid components, locate and schedule resources, communicate between nodes, access programs and data sets within data archives, measure and analyze performance and finally authenticate users and resources [1]. We exploit the Globus Toolkit, the *de facto* middleware standard for computational grid, offering the power and security needed to develop atmospheric modeling applications. In this paper, we describe the use of Grid Computing technology, optimising the results obtained from some meteorological and dispersion models coupled in cascade: RAMS [2], CALMET [3], CALPUFF [4] and CALGRID [5]. As a test case, the modeling system has been applied to simulate a summer photochemical smog episode in the Salento Peninsula. This is a narrow flat land in the south-eastern part of Italy with big manufacturing facilities on the opposite coastlines. The geographic position of this area favours the development of complex meteorological circulations and the consequent complex pattern of ground level pollutant concentration. Simulation results are compared with the data measured by environmental stations. The paper is organized as follows. In sect. **2**, we recall some essential background information about Grid Computing technologies, and in sect. **3**, we describe the modeling system and the models used in the simulations. Section **4** describes G-AQFS and its main components, whereas sect. **5** describes meteorological and dispersion simulations and also presents a case study. Finally, the conclusions and the future evolution of our work are given in sect. **6**.

2. – Grid Computing and the Globus Toolkit

Grid computing brings parallel and distributed computing and high-speed networking together; it is an evolution of some of the concepts dating back to 1992 when metacomputing was introduced [6]. The key idea of a virtual supercomputer made of several computing resources connected by high-speed networks promised to provide the solution to problems otherwise too large for a single supercomputer, or whose execution would have benefited from division into several components to be executed on different architectures. Grid technologies need to provide support for: security, resource management, data management, communication, quality of service and adaptation. A security infrastructure is needed to provide users with authentication and authorization services; the aim of the resource management is to hide heterogeneity providing a coherent and uniform interface. Data grids need mechanisms to handle large amounts of data, such as metadata indexing, searching, and parallel data transfers. Communication requires specialized protocols for the wide environmental area, possibly guaranteeing service quality. Finally, the support for adaptation is crucial in grid environments and it is based on the knowledge of static information, which is known in advance, about resources, networks etc. and dynamic information, which is unknown until runtime (*e.g.*, resources load). The Globus Toolkit [7] has gained worldwide acceptance, so that it is deployed by several organizations as the middleware of choice for grid computing and many scientific endeavours rely on it. It is a layered architecture that addresses grid security, remote access and control providing support for PKI [8] single sign-on authentication/authorization, an information-rich en-

vironment based on the LDAP protocol and a standardized interface to heterogeneous computing resources. However, the complexity of this middleware hinders the majority of scientists (who are not computer scientists) from performing their useful work. The Globus Toolkit 4 (GT4) [9] will be the next release of Globus Toolkit and it will represent a major advance in terms of quality, robustness, easiness-of-use and documentation. In particular, GT4 will improve the Web services (WS) components first introduced in prototype form in The Globus Toolkit 3 (GT3) and it will be compliant with Web services standards as the WS-Interoperability (WS-I) Basic Profile [10]. GT4 will support WS-Resource Framework (WSRF) [11] and WS-Notification (WSN) [12] specifications which were submitted to OASIS in May 2004 as a result of the experience with the Open Grid Services Infrastructure (OGSI) specification developed in GGF [13].

3. – Modeling system

The modeling system consists of two meteorological models called RAMS and CALMET, a emission pre-processor and two dispersion models for inert and photochemical pollutants. RAMS (Regional Atmospheric Modeling System) is a prognostic mesoscale model developed at the Colorado State University to simulate and forecast weather system. It contains an atmospheric model, which performs the actual simulation, and a data analysis package which prepares the initial data for the atmospheric model from the observed meteorological data. The atmospheric model is constructed around the full set of primitive dynamical equations, which govern atmospheric motions. The RAMS model in this study was initialised and driven using the European Centre for Medium-Range Weather Forecasts (ECMWF) [14] data, updating fields every six hours. CALMET (CALifornian METeorological model) is a meteorological model which includes a diagnostic wind field generator containing objective analysis and parameterized treatments of slope flows, kinematical terrain effects, terrain blocking effects, a divergence minimization procedure, and a micrometeorological model for overland and over water boundary layers. The input required by CALMET consists of four categories of data: geophysical data file, upper air sounding data, surface meteorological data, over water data, and, optionally, a prognostic gridded wind field. The latter is the option we have chosen. The output of CALMET consists of 3D gridded fields of wind components and air temperature, and 2D fields of turbulent parameters. CALMET is designed to drive the two dispersion models CALPUFF and CALGRID. Emission data for the modeling domain are obtained using two data sources: the CORINAIR national inventory and the industrial database. In order to provide the proper detail in time and space, the yearly Province data level inventory has been subjected to a disaggregation procedure, and the organic compounds have also been subjected to a speciation into the individual species used by the chemical mechanism of the model. The whole procedure of model emission construction follows 3 steps. The first is the spatial disaggregation of yearly Province emissions into the horizontal grid cells using surrogate variables strictly correlated with the emissions. The second step is the hourly disaggregation of yearly cells emissions using local information concerning the emission activity. The third aspect regards the photochemical pollution and consists in the speciation of organic compounds into individual species according to proper speciation profiles and their aggregation in lumped species considered in the chemical mechanism of the photochemical model. The outputs provided by RAMS/CALMET system and emission data are then used by the dispersion models, CALPUFF and CALGRID. CALPUFF (CALifornian PUFF model) is a non–steady-state Gaussian puff model containing modules for complex terrain ef-

fects, over water transport, coastal interactive effects, building downwash, dry and wet pollutant removal, and simple chemical transformation. It is designed to use meteorological fields provided by CALMET and time-dependent source and emission data. It produces one-hour averaged ground concentrations for the simulated species. CALGRID (CALifornian GRID model) is an Eulerian photochemical three-dimensional model which includes accurate modules for horizontal and vertical advection/diffusion. The model is based on the SAPRC-90 chemical mechanism, which contains 54 chemical species and 129 reactions. It requires information about the meteorological and turbulent field (by CALMET) and emission data in the domain, at the boundary and at initial time. It produces a 3D hourly field of concentration of the simulated species. As for initial and boundary conditions, we used measured data from the environmental network at this stage, yet we are going to utilise the CHIMERE model output in the future, which runs at a continental scale [15].

4. – Grid Air Quality Forecast System

G-AQFS is a tool that allows the management of the running of models in cascade, integrated in a Computational Grid. The application flow of this system is defined by the number of models considered and by their logical interconnections. Furthermore, the user can redefine the logical workflow, based on the same models, but with different logical interconnections. The G-AQFS software core must be available on at least one machine, where someone (a power user) has installed and configured all the packages of the system. The computational strategy is based on the Master-Slave model. The Master runs on the machine where the system is installed, whereas the Slaves run on other nodes belonging to the same computational Grid. In particular, the machine where G-AQFS is installed, is called Master Grid Node (MGN). The other computational Grid resources are named Slave Grid Nodes (SGNs). The installation of G-AQFS on MGN includes RAMS, CALMET, CALPUFF, and CALGRID models, the configurations files and the software modules required by the integrated modeling system. G-AQFS requires some data repositories to operate: the territorial data sets, the large-scale synoptic models data coming from the European Centre for Medium-Range Weather Forecasts - ECMWF, the continental chemistry and transport models data coming from the PREVAIR system and other emissions data sets. A Computational Grid is made of platform-independent nodes. Generally, many platforms of calculation are available: super computers, workstations and PCs. Thus, the executables of the G-AQFS packages for several platforms are needed (the source code can be also provided for some packages). In the workflow execution, the output of a model is the input of another model. To include datasets coming from particular data repositories (topography, ECMWF, measures, etc.) as well, we need to manage the presence of various data formats and select the right information to be used as the input for a specific model. Package execution is managed by appropriate drivers called Package Drivers (PDs). The driver is a glue to connect applications with the Grid Computational Resources. In particular, it is responsible for input file preparation, and for package running (on Master Grid Node or Slave Grid Nodes). In the case of the execution on a Slave Grid Node, the PD needs to move the input and the executable files towards the slave machine and to re-collect on the master machine the output generated during the execution. An overview of the G-AQFS general architecture is shown in Figure 1A. The Core System module integrates functional elements needed for the management of the modeling system workflow and for the control of single jobs on a master or slave grid node. The Workflows Repository collects and catalogues all models, the configuration

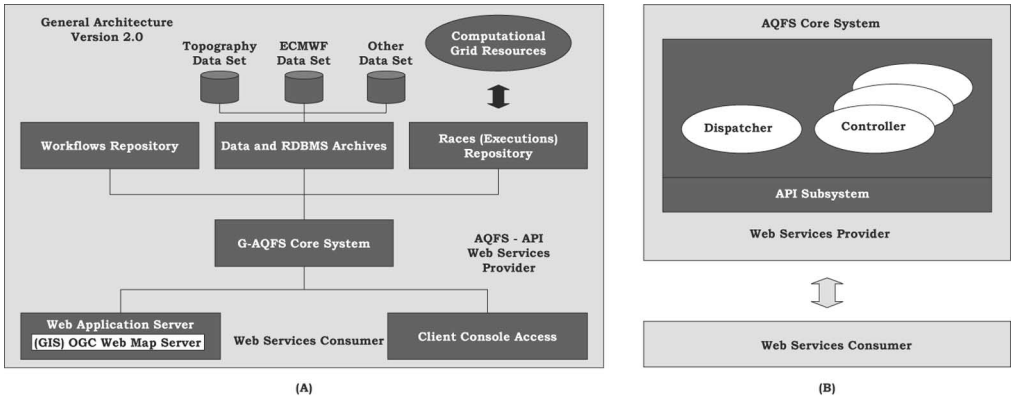

Fig. 1. – (A) G-AQFS general architecture; (B) G-AQFS Core System.

files, the packages drivers and the workflow topologies of the modeling system. The Core System uses the Races Repository module to store and to organize every simulation of the model chain. The Core System is a web services provider related to the users front-end of the system. In particular, the front end can be an Application Server that integrates a Web Map Server, an Open Geospatial Consortium GIS component [16, 17], to georeferenced raster image maps produced by the output of models or a command line client for users console access.

4˙1. *G-AQFS Core System.* – An overview of the G-AQFS Core System is shown in fig. 1B. The main components are the Dispatcher, the Controller and an API Subsystem in order to build the web services for the user front-end, that is the web services consumer of the whole system. The Dispatcher instantiates a Controller module for every user execution of a workflow present in the system. An overview of the Controller is shown in fig. 2. The main components are the Work Flow Scheduler (WFS) and the Round Robin Scheduler with priority (RRSP). The first component uses an algorithm, that we have developed for this system, called Depth-First Search Job with Priority (DFSP), in order to determine the operating sequence of the packages [18, 19]. The second component builds a queue of Grid resources, in order to implement a round robin mechanism. In particular, by querying the Globus MDS (Monitoring and Discovery Service), it is possible to know the dynamic information related to the status of Grid resources. Thus, a queue of available computational resources is built, ordered on the basis of computational features (*e.g.*, cpu type and speed, RAM, workload, etc.). The RRSP module takes a package from the Q1-FIFO queue and runs it on the best computational resource available at the Ti time. The best computational resource available at the Ti time, called $Best\_Grid\_Host(T_i)$, is the highest value obtained by applying the simple metric M shown below, based on MDS attributes.

$$M = \frac{\left(\frac{\sum_{i=1}^{3} a_i}{3}\right) * b * c}{100}, \tag{1}$$

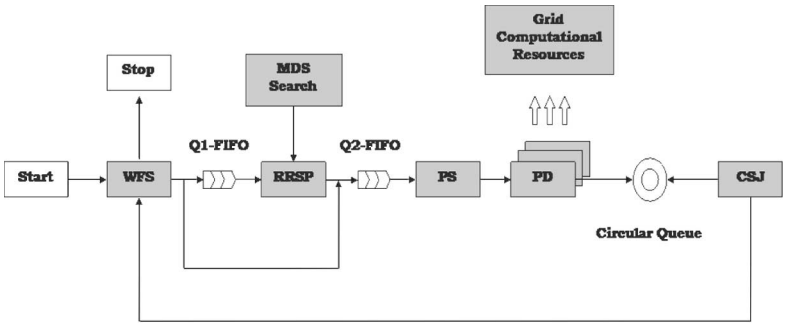

Fig. 2. – Controller overview.

where

$$a_1 = Mds\_Cpu\_total\_free\_1Min, \quad a_2 = Mds\_Cpu\_total\_free\_5Min,$$
$$a_3 = Mds\_Cpu\_total\_free\_15Min, \quad b = Mds\_Cpu\_SpeedMhz \text{ and}$$
$$c = Mds\_Cpu\_Total\_count.$$

The metric uses both static and dynamic information for each host, in a defined period, provided by MDS Globus Toolkit component. For static information we mean: the single CPU speed (Mds-Cpu-speedMHz) and the count of processors (Mds-Cpu-Totalcount). For dynamic information we mean: the rate of available free CPU in a defined period (Mds-Cpu-Total-Free-1Min, Mds-Cpu-Total-Free-5Min and Mds-Cpu-Total-Free-15Min). Other components of the controller are: the Package Starter (PS), the Package Deriver (PD) and the Check Status of Jobs (CJS). The PS constantly controls if there are any elements inserted in the Q2-FIFO queue. In this case, it creates a thread for independent running of the PD, associated with the package. The PD starts the execution by adding a token with the associated package name in a circular queue. The monitoring of job advancements is achieved by PD querying the grid. If the jobs fails, PD deletes the related token from the circular queue. The CSJ monitors the token status inserted in the circular queue. When a package terminates the execution in a normal way, the CSJ notifies it to the WFS module. If any problem occurs during the package running, the CSJ communicates the status error and the type of failure to the WFS module. In this case, the WFS module will reschedule the package execution. The CSJ includes a mechanism, using a timeout, in order to assure that a token does not remain in circular queue for a long time, due to an error of the associated PD. In such a situation, the token is removed and the CSJ module communicates an error message to the WFS; the package will be forwarded to the Q1-FIFO queue for a new running. Once all packages are executed, the Controller stops.

5. – A case study

Aim of this work is the implementation of the first version of the G-AQFS. In order to do this, it is necessary to develop, debug and run the modeling system on a small Grid environment (with few Linux workstations). As a case study, the modeling system has been applied to the Salento Peninsula in Apulia (south-eastern Italy), to simulate photochemical pollution in a typical summer scenario characterised by a weak synoptic

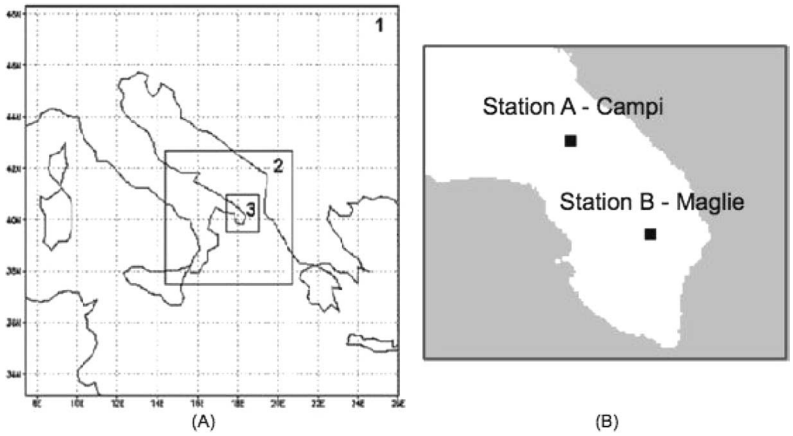

Fig. 3. – The modeling domain and the three nested grids (A). Location of two air quality monitoring stations (B).

forcing with the development of a complex sea breeze system [20]. The chosen period was 1st to 3rd July 2003. At the beginning, the synoptic charts showed the presence of a ridge affecting the south-western Mediterranean Sea. Such a ridge was on the southern side of a wide cyclonic circulation, with its minimum positioned south-west of the British islands. In the following days, as the upper level progressively deepened in the south-north direction, the extension of the ridge was reduced. A south-western slightly cyclonic circulation affected southern Italy at the end of the period. The surface circulation presented a prevailing north-western synoptic component over Apulia in the first 2 days. During the last day, the pressure gradient was significantly reduced, favouring conditions for the development of sea breeze convergence over the region. The simulations with the RAMS model have been performed in a two-way nested grids configuration with three grids (see fig. 3). This allows to resolve meteorological features at different spatial scales. For initial and boundary conditions, the Isentropic Analysis System (ISAN) package (the RAMS module for the generation of data analyses) was used. At initial time, analysed fields were based on the ECMWF (European Centre Medium Weather forecast) gridded datasets. Every six hours, the lateral and the top boundary conditions were updated in the coarsest grid, by using the ECMWF gridded datasets. In the coarsest grid domain, a nudging toward the data was applied in the 3 grid points closest to the lateral boundaries and in the upper 5 grid levels. CALMET and CALGRID were run on the inner grid.

Figures 4A and B, respectively, show the modelled surface wind fields and ground ozone concentration at 12:00 UTC on July 3, 2003. It is possible to notice the development of the sea breezes with the convergence zone over the peninsula with a consequent accumulation of air pollutants in the same area. The modeling domain has two air quality monitoring stations, located in the middle of modeling domain (see fig. 3B), which daily report air quality data; these are Campi station and Maglie station. The above simulation results are thus compared to the hourly ozone data from these two measurement stations, as shown in fig. 5. These plots show that the modeling system realistically reproduces the typical diurnal ozone cycle: the ozone peak typically occurs when sunlight and ambient temperature are at their highest day level. The figure also shows that simulation well agrees with the measurements and that the maximum modelled is of the

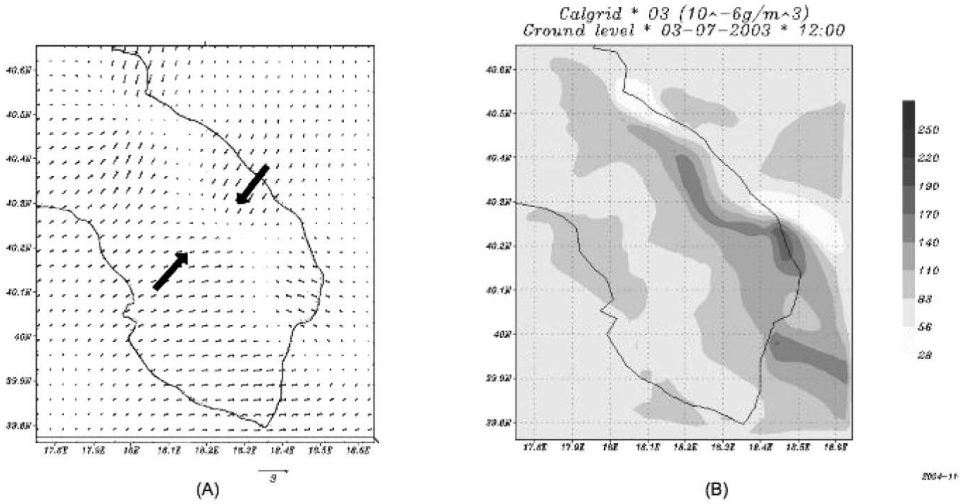

Fig. 4. – Modelled wind field with RAMS/CALMET, the arrows represent the wind direction of ground level (A) and ground level O3 concentration 12:00 UTC of 3 July 2003 (B).

same order than measured.

5˙1. *Performance considerations*. – We have used a small grid environment in this work, created with four Linux PC having the same configuration of the Globus Toolkit (GT) and connected to the same LAN. Our preliminary test of G-AQFS has been performed to assess the results of the simulations. In order to assess the G-AQFS performance on our small grid environment, we have taken the same case study into consideration, running G-AQFS on one machine without the Globus Toolkit, and with the grid components disabled. In this context, by repeating the experiment several times and by making a performance comparison between the modeling system execution time on one machine (without GT) and on our small grid environment, we have obtained a nearly 50% time reduction. In order to obtain a good analysis of the time performance, it is necessary to test G-AQFS on a real Grid infrastructure, where the computational resources are distributed and heterogeneous.

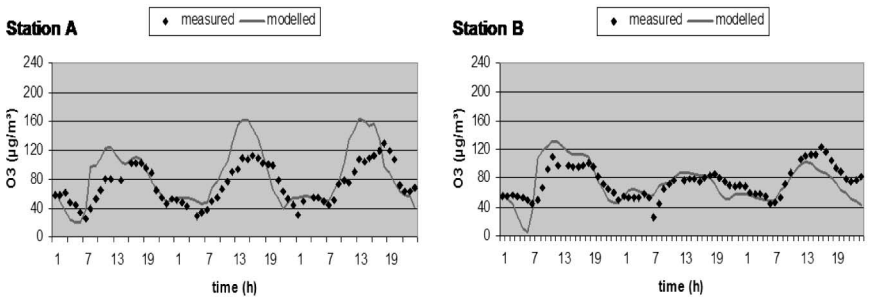

Fig. 5. – Comparison between measured and modelled ozone concentration for the period 3 July 2003.

6. – Conclusions and future work

An integrated G-AQFS to study the transport, diffusion and reaction of air pollutants has been presented. The system includes two meteorological models, emission pre-processors and two dispersion models for inert and reactive pollutants. The integrated grid computing system (G-AQFS) allows the investigation of meteorological and chemical effects on the formation of air pollution. The computational Grid resources are scheduled with a round robin mechanism and by querying the Globus MDS: a simple metric is used to define the best computational Grid resource at Ti time. As a test case, the modeling system on a small Grid environment has been applied to the Salento Peninsula, in a summer period to study photochemical pollution. Comparison between predicted and measured ground level ozone concentrations indicates that the system can realistically simulate the ozone evolution. We plan to test G-AQFS on a real Grid infrastructure. We are going to investigate the scheduling techniques and to compare the various possible metrics based on more Globus MDS attributes. Furthermore, it needs to be mentioned that our work can be extended in many ways, that is by varying the scheduling algorithm or by adding other models for instance.

* * *

This work has been partly supported by the Italian Ministry of Scientific Research and Technology (MIUR) Fund through project no. CNR-ISAC-245-2002. Special thanks to the "Osservatorio dell'inquinamento dell'atmosfera e dello spazio circumterrestre di Campi Salentina (Lecce-Italy)" for supplying Environmental data and to Mr. G. Rispoli and Mr. G. Lella for their technical support.

REFERENCES

[1] Foster I. and Kesselman C., *The Grid: Blueprint for a New Computing Infrastructure* (Morgan Kaufmann) 1998.
[2] Pielke R. A., Cotton W. R., Walko R. L., Tremback C. J., Lyons W. A., Grasso L. D., Nicholls M. E., Moran M. D., Wesley D. A., Lee T. J. and Copeland J. H., *Meteorol. Atmos. Phys.*, **49** (1992) 69.
[3] Scire J. S., Insley E. M. and Yamartino R., *Model Formulation and user's guide for the CALMET meteorological Model* (California Air Resource Board) 1990.
[4] Scire J. S., Stimatis D. G. and Yamartino R., *Model Formulation and user's guide for the CALPUFF dispersion Mode* (California Air Resource Board) 1990.
[5] Yamartino R. J., Scire J., Hana S. R., Carmichael G. R. and Chang Y. S., *CALGRID: A Mesoscale Photochemical Grid Model*, Volume I: Model Formulation Document Sigma Research Report No. A6-215-74. PTSD, CA 94814. September, 1989 (California Air Resources Board. Sacramento).
[6] Foster I. and Kesselman C., *Int. J. Supercomputer Appl.*, **11** (1997) 115.
[7] *Information for the Globus Toolkit Development Community* `http://www.globus.org/developer/`.
[8] Tuecke S., *Grid Security Infrastructure (GSI) Roadmap*, Internet Draft 2001 `http://www.gridforum.org/security/ggf1_200103/ drafts/draft-ggf-gsi-roadmap-02.pdf`.
[9] Globus Toolkit Development Documents `http://www-unix.globus.org/toolkit/docs/development/`.
[10] The Web Services-Interoperability Organization, *WS-I Basic Profile 1.0* `http://www.ws-i.org/archive/Profiles/Basic/2003-08/BasicProfile-1.0a.htm`.
[11] The Globus Alliance, *The WS-Resource Framework*, `http://www.globus.org/wsrf/`.

[12] *WS-Notification* `http://www-106.ibm.com/developerworks/library/specification/ws-notification/`.
[13] *From Open Grid Services Infrastructure to WSResource Framework: Refactoring & Evolution*, `http://www.globus.org/wsrf/specs/ogsi_to_wsrf_1.0.pdf`.
[14] *The European Centre for Medium-Range Weather Forecasts*, `http://www.ecmwf.int/`.
[15] *Prévisions et observations de la qualité de l'air en France et en Europe*, `http://www.prevair.org`.
[16] *The Open Geospatial Consortium, Inc.*, `http://www.opengeospatial.org/`.
[17] *OpenGIS Web Map Server Cookbook of the Open Geospatial Consortium, Inc.*, `http://www.ogcnetwork.org/docs/03-050r1.pdf`.
[18] Aloisio G., Cafaro M., Cesari R., Mangia C., Marra G. P., Miglietta M., Mirto M., Rizza U., Schipa I. and Tanzarella A., *G-AQFS: Grid computing exploitation for the management of air quality in presence of complex meteorological circulations, International Conference on Information Technology 2004*, edited by Pradip k. Srimani (Program Chair), Part II (Las Vegas, Nevada) 2004, pp. 83-87.
[19] Aloisio G., Cafaro M., Cesari R., Mangia C., Marra G. P., Miglietta M., Mirto M., Rizza U., Schipa I. and Tanzarella A., *J. Digital Information Manag.*, **2** (2004) 67.
[20] Mangia C., Martano P., Miglietta M., Morabito A. and Tanzarella A.,, *Meteorol. Appl.*, **11** (2004) 231.