DAM: Datasets for atmospheric modelling

Stefano Galmarini[(1)], Giovanni Graziani[(1)], Roberto Bellasio[(2)] and Roberto Bianconi[(2)]
[(1)]European Commission Joint Research Centre, Environment Institute, TP321, I-21020 Ispra (VA), Italy
[(2)]Enviroware srl, Centro Direzionale Colleoni, Palazzo Andromeda 1,
I-20041 Agrate Brianza (MI), Italy

Keywords: Datasets for model validation, Internet, WWW, DAM

1 Introduction

The rapid growth of the Internet and the World Wide Web (WWW) in particular has led to an enormous increase of the information available to the scientific community in almost any field. While its owners in general make this information broadly available, difficulties still remain in finding it. This is due to the dimension of the WWW and the time and effort it takes to make people aware of the existence of such information.
For what concerns the atmospheric modelling (and in particular dispersion and air quality models), the validation of existing and newly developed models with datasets collected during experiments is of fundamental importance. To date the typical approach to model validation was to rely on few well-known and used datasets, and it could take years before a new dataset could become known to the wide research community and used. Another limitation of the old days was the lack of control on the use of the data, that was distributed sometimes in incomplete format and in some cases without any reference on the scientific literature.
The WWW offers now the possibility to distribute the data world-wide in a controlled fashion and at practically no cost. Even if the data are not directly accessible through the Internet, the WWW is the medium to communicate to the scientific community more widely and in shortest time the existence of collected data.
An obstacle is still represented by the distribution of the information. The presence of a dataset on the web can in general communicated by word of mouth, publication of the information on scientific journals or it can be "searched" using web facilities. All these aspects are not always efficient and reduce the potential of the data availability on the web.

In order to give an answer to this twofold need, a website named DAM (Datasets for Atmospheric Modelling**, http://rtmod.ei.jrc.it/dam**) has been created, where a large quantity and variety of information on datasets for validating atmospheric models is provided. In this paper an overview of the resources to be used to build a collection of datasets is given, and the structure of the updateable database interfaced by the website is described. Being the information collected necessarily incomplete, an overview of the references to resources currently available is then presented. Finally, future directions and possible developments for the indexing of existing information are suggested.

2 Resources for building a collection of datasets

There are several resources available on the WWW to set up a collection of links to atmospheric modelling datasets. They include search engines, web directories, and links from subject-related web-sites.
Search engines are search forms into web pages that search through databases of HTML documents gathered by a robot. A "robot" (or "spider") is a program that crawls the WWW to collect the contents by accessing one web page, generally the index, and then recursively retrieving all the pages that are there referenced. These systems can collect huge and useful quantities of links, and frequently updates it frequently. However they have the disadvantage that, even if they are based on

advanced heuristic rules for cataloguing the contents, the result of the search might be imprecise whenever the query terms are common to other fields of science and technology.
On the contrary, there are several web-sites that offer pre-defined and specialised *directories*. A directory contains selected and well defined information organised in an indisputable fashion (as it can be the case for the words "model", "dataset", "validation", "dispersion", and so on). However this organisation requires a specific experience and knowledge that is not always possessed by people managing these web-sites. Another disadvantage is that the information is not updated as a "robot" can automatically do for search engines, and the ageing of the content is often evident. It is also frequently observed that the contents and links of these directories are duplicated from one web site to another.
While these two resources can be used to create a basic dataset, it is then necessary to refer to pages with more specific contents, such as for example the web page of a given tracer experiment, that are generally maintained by the performer of the experiment. The problem connected with searching these pages is that the owner (universities, research institutes, etc.) often limits or denies the access to robots that attempt to index their sites. This implies that the knowledge on the existence of specific information relating to a specialised subject has to be retrieved from non-internet sources, (publications or private contacts).

3 Development of a website for distributing the information available on the World Wide Web

In order to overcome the drawback of the methods described above, a database for cataloguing the available internet resources on model validation (mainly tracer experiment datasets and measuring campaigns) was built. It was decided to structure the catalogue as a relational database, developed under MySQL, that includes the following fields for each record:

- acronym of the experiment or dataset
- geographical location of the experiment
- organisation managing the experiment/campaign
- date or period of the experiment/campaign
- one or more keywords defining the dataset content
- a brief description, generally the one provided by the website where the information is available.
- bibliography
- internet references (http://)
- contact person
- on-line availability of data

The DAM database is accessible from an HTML interface at the address **http://rtmod.ei.jrc.it/dam**.The database records can be visualised through an alphabetical menu, or they can be displayed as the result of a query to the database, based on the keywords set for each record and on the acronyms of the experiments. The software also includes an administration interface to insert new records, to update or delete existing ones, as well as an e-mail form to submit new datasets.
Up to now there are more than 100 datasets catalogued, most of them also linked to internet resources, or at least providing the contact information to access or request the datasets.

4 Future developments and directions

While so far the internet software has been built and installed, it is evident that much effort has to be put in keeping the database updated and enlarged. Towards this end, it is envisaged that the scientific community of atmospheric modellers and developers makes use and contributes to the inclusion of new information and on the revision of the available content. Opinions and suggestions on the choice of keywords made to describe the datasets are also mostly welcome.
In the future, this web-site might expand to other branches of atmospheric modelling, including meteorology.

5 Conclusions

There are currently several datasets available to the Scientific Community for model validation. The research groups that collected them generally make these data available through WWW sites or, in many cases, the information on them is available and the data can be obtained upon request.
DAM's objective is to facilitate the access to such valuable information to any model developer or user that intends to validate his/her modelling tool. DAM must be intended as an interface between modellers and the information available through existing web sites or contact points. Although DAM can obviously not be considered capable of answering any possible type of request from model users or developer, it aims at being as complete as possible. For this reason a continuous update is necessary for the inclusion of new information about existing datasets which are not mentioned in the site or new ones made available to the Scientific Community. To pursue this intention the contribution of the modelling and experimental community is considered essential.

Aknowledgements

Enviroware srl has contributed to this project as partial fulfilment of contract 16220-2000-06 F1SC ISP IT with the JRC-EI of the European Commission.