Project no. GOCE-CT-2003-505539

Project acronym: ENSEMBLES

Project title: ENSEMBLE-based Predictions of Climate Changes and their Impacts

Instrument: Integrated Project

Thematic Priority: Global Change and Ecosystems

D2B.17 GCM-based predictor data set for statistical downscaling (Task 2B.2.5)

Due date of deliverable: Month 30 (30-Oct-2007)
Actual submission date: 2 May 2008

Start date of project: 1 September 2004   Duration: 60 Months

Organisation name of lead contractor for this deliverable:
Partner University of Cantabria (UC)
          University of East Anglia (UEA)

Revision [15-Oct-2007
11-Jan-2008
6-Apr-2008]

1. Introduction

The Statistical Downscaling (SD) approach requires working with different sources of information: observations, reanalysis datasets, and GCM outputs (either from seasonal forecast models or from ACC (anthropogenic cliamte change) models). Unfortunately, these datasets are distributed in a number of different servers with different formats and coding tables, thus complicating the work of the SD partners. In order to overcome this bottleneck, it was planned to develop a "climate data access portal" within ENSEMBLES to gather and centrally store observations (predictands) together with a common subset of GCM variables which are commonly used as predictors in statistical downscaling tasks. The goal is that ENSEMBLES partners can obtain this information directly from this server using a Web Portal (see www.meteo.unican.es/ensembles). Furthermore, it was planned that this data would be also available for the ensembles SD Portal so end-users can run their downscaling jobs on the web, using these datasets and a set of pre-defined SD algorithms. A first prototype of the portal was already described in the month 18 deliverable D2B.4 and a major extension was described in D2B.19. Moreover, an initial list of ERA-40 based predictor data set for SD was described in D2B.13.

In the present deliverable, we describe the GCM-based predictors which are already available at the Climate Data Access portal (to be downloaded as text files) and at the SD Portal (to be used for downscaling purposes). First of all, the initial list of predictors described in D2B.13 has been extended to cover other interesting parameters (this list will be extended along the project duration and the current version can be consulted in www.meteo.unican.es/ensembles/listofvariables.htm):

Variable (Code)	Levels	Time
Geopotential (Z)	300-500-700-850-925-1000	00-12
V velocity (V)	300-500-700-850-925-1000	00-12
U velocity (U)	300-500-700-850-925-1000	00-12
Temperature (T)	300-500-700-850-925-1000	00-12
Relative humidity (R)	300-500-700-850-925-1000	00-12
Specific humidity (Q)	300-500-700-850-925-1000	00-12
Potential Vorticity (PV)	300-500-700-850-925-1000	00-12
Relative vorticity (VO)	300-500-700-850-925-1000	00-12
Divergence (D)	300-500-700-850-925-1000	00-12
Total Column Water Vapour (TCWV)	0	00-12
Total Column Water (TCW)	0	00-12
MSLP (MSL)	0	00-12
2m Temperature (2T)	0	00-12
2m Dew Point (2D)	0	00-12

The above list of variables is the minimum dataset to be stored for all reanalysis models and GCM simulations (in some cases, due to the lack of data only a subset of the above list is included; in that case we mark the dataset as "incomplete").

As of the end of March 2008, the data access portal is operative and 30 users from 20 ENSEMBLES partners are registered. Moreover, an ENSEMBLES Public Report (not committed in the updated DoW), has been produced to provide users with instructions for using the portal, since it is much complex (and powerful) than expected: A.S. Cofiño, D. San-Martín, J.M. Gutiérrez, B. Orfila (2008) "Guidelines for the use of the ENSEMBLES RT2B Climate Data Access Web Portal". This report is included as an

Anex to this deliverable. Moreover, an email list "[ENSEMBLES Downscaling Portal]: News" has been set up to periodically inform users of news (new datasets, etc.) and any problems identified.

The current data contained in the portal is listed below (all datasets are restricted to the European domain, unless a wider region is explicitly mentioned):

OBSERVATIONS:

> • **ECA** (European Climate Assessment & Dataset project). Daily datasets of precipitation, temperature, pressure, humidity, cloud cover, sunshine and snow depth since 1900 over networks of 100-1000 stations (from eca.knmi.nl).
> • **Ensembles 50km** gridded daily observation records of precipitation and surface temperature. 1950-2006.

REANALYSIS:

> • **NCEP/NCAR** Reanalysis1. 1948-2007. Incomplete.
> • **ERA40** ECMWF: 1957-2002.
> • **JRA25** Japanese Reanalysis: 1979-2004. Incomplete.

SESONAL TO DECADAL:

> • **DEMETER**. Multi-model seasonal prediction experiment including seven models run for six months four times a year using 9 different perturbed initial conditions (i.e., 9 members).
> • **ENSEMBLES** Stream 1 (in this case only for Spain currently). Incomplete.

CLIMATE CHANGE SCENARIOS:

In this case, the datasets have been obtained from different sources. On the one hand the IPCC data centre provides access to the models contributing to the AR4; these datasets are referred to as *PCMDI_Model*. On the other hand, the ENSEMBLES project is storing a common list of parameters for different ACC experiments (www.mad.zmaw.de/projects-at-md/ensembles); these datasets are referred to as *CERA_Model*. Finally some datasets are obtained directly from local providers; these datasets are referred to as *Model*. So far, the following data from ENSEMBLES STREAM1 is available, but the portal provides updated information of new datasets included as they become available. All this data covers the whole globe:

> • **PCMDI_CGCM3.** Canadian Centre for Climate Modelling and Analysis (www.cccma.bc.ec.gc.ca/data/data.shtml), including 20th century (from 1951 to 2000) and scenarios A1B, B1 (daily data in pressure levels is only included for the periods 2046-2065 and 2081-2100).
> • **CERA_MPI-ECHAM5,** including 20th century data (1961-2000) and scenarios A1B, B1, and A2 (2001-2100).
> • **CNRM-CM3,** including 20th century (1961-2000) and scenarios A1B, B1, and A2 (2001-2100).

The main reason for the lack of more ENSEMBLES data in the portal is the delay suffered by the central archiving process in the CERA database (this has also caused a delay of this D2B.17 deliverable). The state of the STREAM1 experiment as reported by the CERA database on March 15 2008 is the following:

- **CNRM-CM3**: T and Z for are still only available above 300mb. These are two important variables for downscaling (in particular Z500). We downloaded this model directly from the local provider and it is available at the portal (thanks to CNRM).
- **MPI-ECHAM5-MPI-OM**. Available in the SD Portal.
- **DMI-ECHAM5-MPI-OM**. No information about this model is given in CERA (no data and no metadata).
- **NERSC, bccr_bcm2**. Only 20cm3 and SRESA2 data. We will wait until SRESA1B is included in CERA.
- **FUB-EGMAM**. Complete data: we are downloading it and it will be available as soon as possible.
- **INGV SINTEX-G**. Only metadata provided (i.e., the description of the variables which are planned to be included in the future).
- **IPSL-CM4**. Incomplete data for 20cm3 and SRESA1B (no data for 850, 925 and 1000mb); only metadata for B1 and A2.
- **METO-HC**. Only metadata provided, and only for 20CM3.

ANNEX 1. ENSEMBLES Public Report

"Guidelines for the use of the ENSEMBLES RT2B Climate Data Access Web Portal"

A.S. Cofiño[*], D. San-Martín[*]
J.M. Gutiérrez[*], B. Orfila[**]
* Applied Meteorology Group, University of Cantabria, Spain
** Instituto Nacional de Meteorología, INM, Spain

Version 2.0 (31/1/2008)

The *Data Access Web Portal* (DAP) has been developed as a part of the ENSEMBLES project with the aim of maximizing the exploitation of the multi-model *seasonal* and *climate change* predictions produced by different modelling centres (for more details see www.meteo.unican.es/ensembles). These predictions are used by a wide range of end-users from different socio-economic sectors (agriculture, energy, etc.) to produce different impact studies. Nevertheless, they need to interpolate the low-resolution predictions (typically ~100 km) to particular locations, or high-resolution grids, to run their impact and/or analysis models. In practise, this post-processing step may be a time-consuming problem for some end-users, who are not familiar with standard meteorological formats, such as GRIB or NetCDF, and with interpolation procedures. Therefore, the apparently simple task of obtaining a simulated data field (e.g. total precipitation) at the required temporal aggregation (e.g. monthly) interpolated to a specific lon-lat point (e.g. to Madrid, Spain) may result very time-consuming for those users. The DAP has been designed to facilitate this task through a user-friendly Web portal, so end-users can interactively define the local points or lattice of interest and obtain their interpolated data in simple formats (e.g. a text file) not worrying about the details of the interpolation techniques (e.g., bilinear interpolation) nor the computing technologies and data formats involved (e.g., GRIB, NetCDF, OPeNDAP, etc.).

NOTE. In cases where historical records of the variables of interest are available at the interpolation points, there are more efficient methods to statistically adapt the model outputs to the local climatology. The Statistical Downscaling Portal (SDP) can be used as an advanced "interpolation" tool for this task (see www.meteo.unican.es/downscaling).

1. Login and General Overview of the Portal

A registered *user* is required in order to log into the portal with full access to data and computational resources (register in www.meteo.unican.es/ensembles/registration.html). However, an anonymous account (user: *anonymous*, pass: *anonymous*) has been created to test the portal with limited functionality.

The DAP has been designed to facilitate the access to **historical observations** and also to Global Circulation Model (**GCM**) **simulations** with different lead times (*reanalysis* projects, *seasonal* forecasts, and *climate change* simulations). These datasets have been organized in different tabs (see *1* in the figure below). The tab ***My home*** is the user's main window and provides information about the status of the

requested data (*Jobs panel*) and the user profile (*Profile panel*). Each data request is handled by the portal as an independent job, so several requests can be handled simultaneously. The ***jobs panel*** (*2*) allows monitoring the status (running, finished, etc.) and type (observations, etc.) of the jobs, and downloading the text files produced by the request when the job finishes (*3*). The button allows refreshing the status of the running jobs (the execution timeline of each job is also provided).

The ***profile panel*** (*4*) displays the databases available for the current user, as well as the account information, restrictions (maximum number of simultaneous jobs, maximum size of data request, etc.) and areas of study allowed.
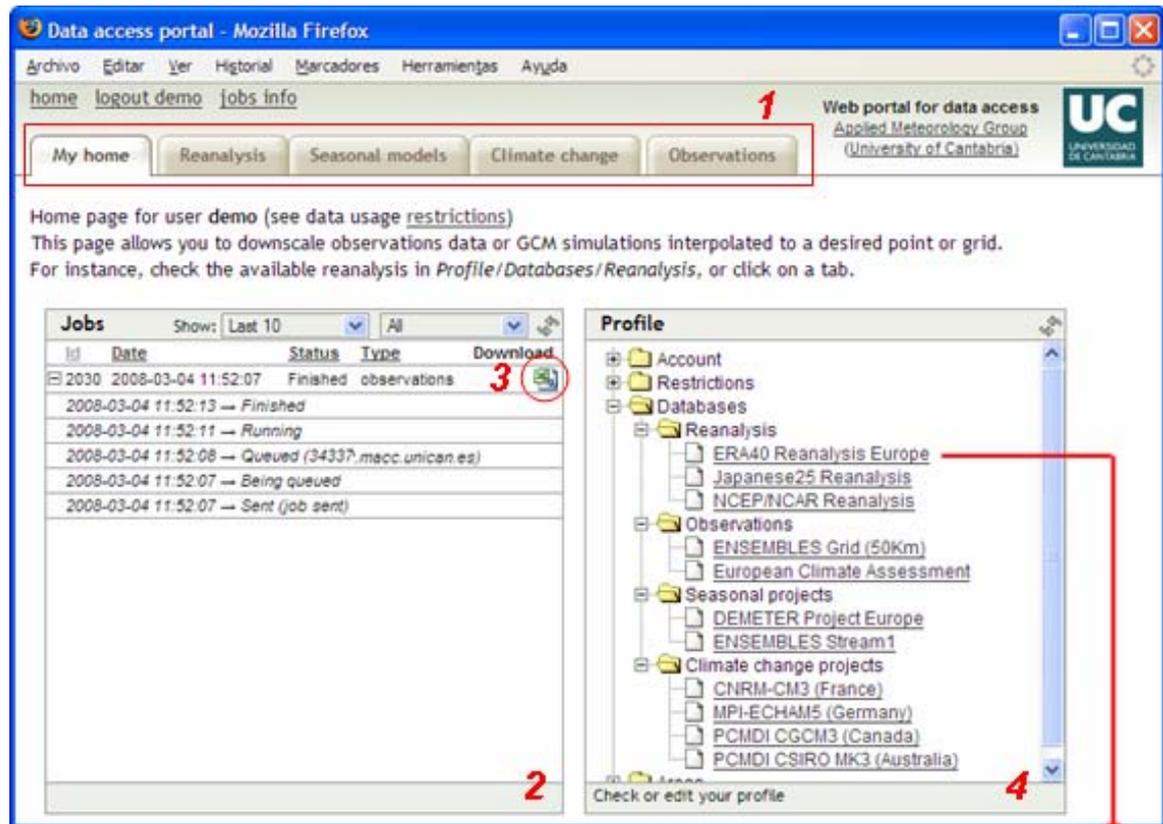
The remaining four tabs provide access to historical *observation* records and also to *reanalysis* data, *seasonal predictions*, and *climate change* simulations.

The datasets available for the logged user can be consulted in the ***profile panel*** by clicking on the corresponding section. For instance, the available variables from the ERA40 reanalysis can be checked by clicking on "ERA40 Reanalysis Europe". Note that this dataset is geographically limited to the European domain, as indicated in the longitude and latitude information (*5*). This information is given for each particular dataset.
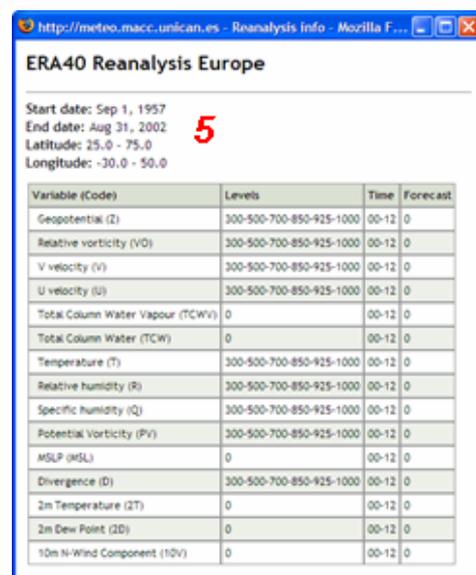
2. Accessing local or gridded observation records

The DAP provides access to different daily historical records, including both raw observations from local stations and interpolated high-resolution grids:

- **ECA** (European Climate Assessment & Dataset project). Daily datasets of precipitation, temperature, pressure, humidity, cloud cover, sunshine and snow depth since 1900 over networks of 100-1000 stations (http://eca.knmi.nl).
- **Ensembles** 50km gridded daily observation records of precipitation and surface temperature. 1950-2006 (http://eca.knmi.nl).

In order to get (or draw) some data from the above datasets follow the steps:

1. Selected the dataset of interest from the database popup menu (*1* in the figure below); the button [?] provides information about the geographical domains and the variables available for the selected dataset.
2. Select the variable and the period of interest (*2*); each variable has a proper temporal *resolution* (e.g. 24 hours, or daily), but a larger time scale can be defined selecting an *aggregation* step (5-day, monthly, yearly) and an aggregation function (the mean, maximum or minimum, or accumulated values).
3. Select the locations of interest using the mouse tools [+] or [−] to add or remove stations to the current selection list (initially empty); click and drag to define a region including the desired stations, the selected locations are marked in red colour (*3*). There are also magnifying tools and checkboxes that can be activated to display city names, topography, borderlines, rivers and lakes, etc., for a better visualization of the region of interest (some of these tools, such as the *cities* and *hydro* are zoom-adaptive, showing more or less information depending on the zoom scale). Alternatively, the desired stations can be selected by including a list of IDs (*4*).
4. Finally, click on the *plot* or *get data* buttons (*5*).

See the figure below for a particular **example** from the ECA network; in this case, mean yearly averages of maximum daily temperature in Madrid (Spain) for the period 1950-2000 are plotted, exhibiting a clear temperature trend during the last 25 years. The figure below shows the values selected before clicking the *Plot data* button and the resulting graph.
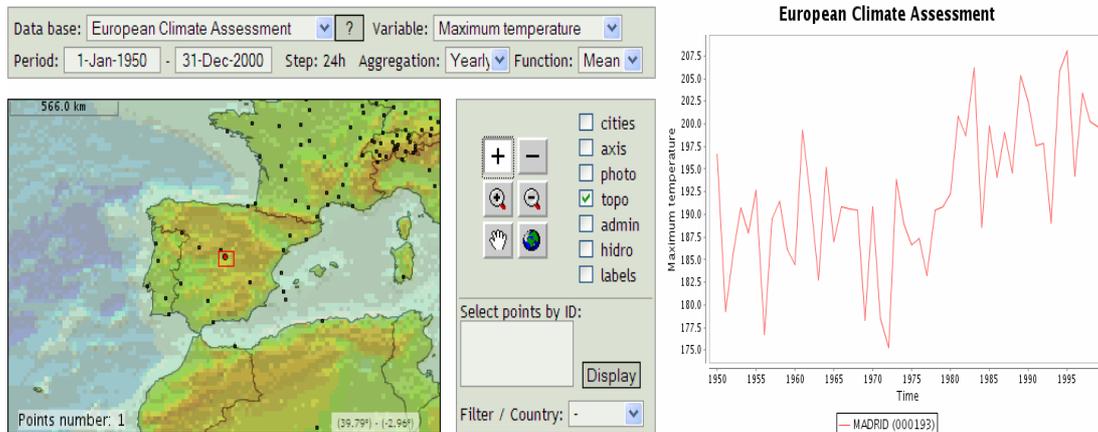
The above data can be also downloaded by clicking in the option *Get data*. In this case a "data access" job is submitted to the portal and processed by a local cluster in order to prepare the requested data as a simple text file; the status of the current jobs can be monitored in the *Jobs* panel of the *My home* tab (see Sec. 1). Initially, the job status is labelled as "submitted" (it is waiting to be processed in the local cluster); then, the status changes to "running" when the job is allocated in some computing node

and finally it is labelled as "finished" when the job ends and the text file (.csv, comma-separated-value) is ready to be downloaded (by clicking in the SpreadSheet icon); note that spread-sheets can directly read *csv* text files.


The file resulting from this process has a simple format with some header lines including the information of the selected database, the variable and temporal resolution, the units, the aggregation step and aggregation function used (if any). Then, some additional lines are displayed with the IDs (labels) of the selected stations (together with their longitude, latitude and height); finally, the data for the different stations is displayed in columns with the dates appearing in rows (note that the spreadsheet may need to be configured in order to automatically separate the data in different columns using "," as a separator character).

The figure below illustrates a typical file obtained from the portal and opened in Excel, with some calculations done in order to plot the series.
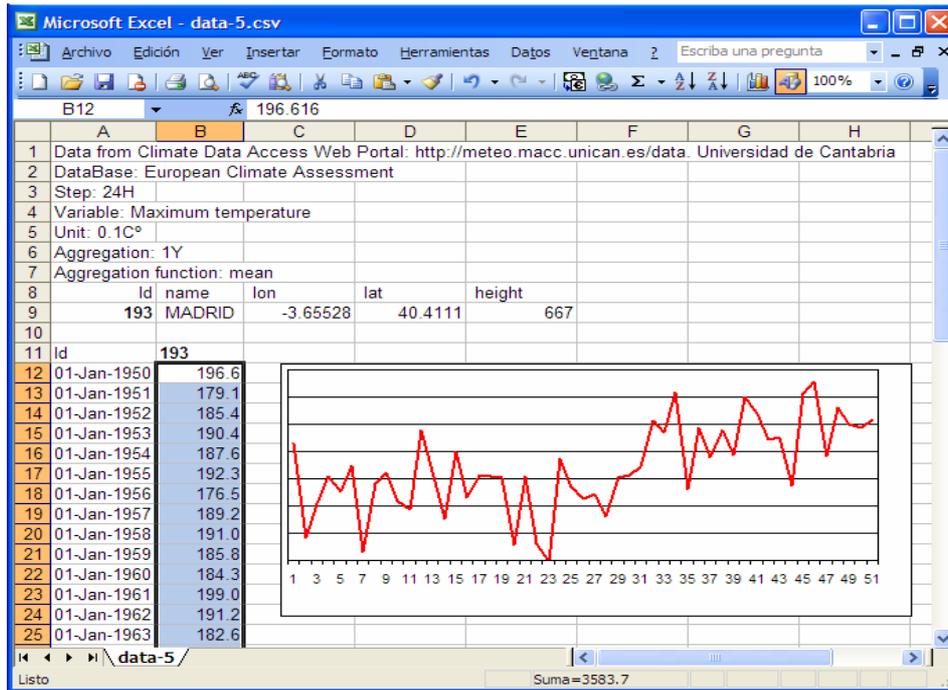
NOTE. The SD portal has not been designed for massive data downloading. Thus, there are some restrictions for the maximum number of selected stations (or grid points) and the maximum size of the resulting file. Moreover, the number of simultaneous jobs allowed for a user account is also limited (see the details of your account in the **Profile** *panel of* **My home***). For instance, the figure below shows the restrictions of a particular account with a maximum of 10 stations, or grid points per query, two simultaneous jobs and a maximum file size of 10MB.*
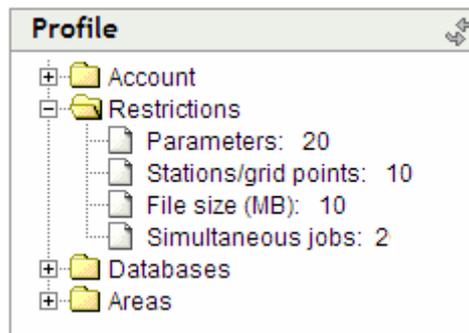
NOTE (DATA USAGE TERMS AND CONDITIONS). The portal is not proprietary of any data served. Therefore, the user must be aware of the terms and conditions that apply to the usage of the data and to the redistribution to third parties. The portal includes information about the data restrictions holding for any particular dataset; they can be checked at any time by clicking on the help button [?] *for any dataset. When registering into the portal, the user is informed of these restrictions and explicitly accepts the usage terms and conditions. For instance, the ENSEMBLES grid of observation records is freely available for research purposes (check the dataset information for the proper acknowledgement in publications using this data), but commercial applications of the data are forbidden.*

3. Accessing Global Circulation Model (GCM) Outputs

The portal also provides access to daily global model outputs from reanalysis projects, seasonal predictions, and climate change simulations:
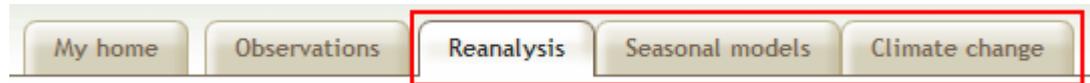

3.1. Reanalysis Data

The following reanalysis datasets are available in the portal (all with daily data):

- **NCEP/NCAR** Reanalysis1. 1948-2007 (http://www.cdc.noaa.gov/).
- **ERA40** ECMWF: 1957-2002 (http://www.ecmwf.int/research/era/).
- **JRA25** Japanese Reanalysis: 1979-2004 (http://jra.kishou.go.jp/index_en.html).

In order to get data from the above datasets follow the steps:

1. Selected the dataset of interest from the database popup menu (**1** in the figure below); the button ? provides information (geographical domain, variables and usage conditions, if any) for the selected reanalysis. All the variables are stored with a common temporal resolution (e.g. 12 hours). Select also the temporal period and aggregation step desired for the data (note that different users may have different aggregation possibilities: daily, 5-day, monthly, and/or yearly, depending on the dataset restrictions and the account permissions).
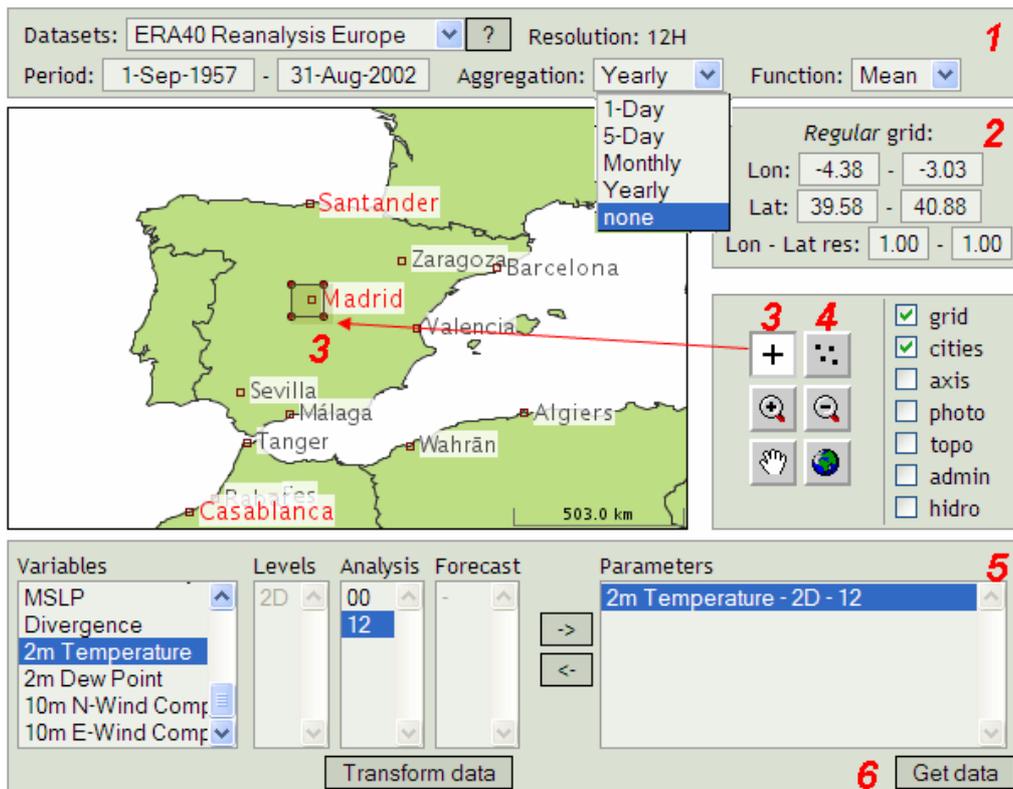
2. Select a rectangular grid by entering longitude and latitude limits and resolution (**2**); initially, the values correspond to the default region and the original resolution of the reanalysis project; the resulting grid can be visualized by checking on the

grid checkbox below. The regular grid can be also defined visually (*3*) by clicking and dragging the area of interest (+).

Alternatively, an irregular grid can be defined by specifying the longitude and latitude pairs for each point (*4*) and clicking the display button (for instance, the figure below shows a data request for a single grid point close to Madrid, Spain).
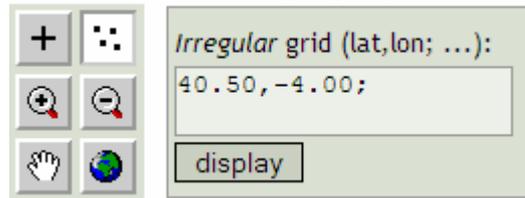

3. Select the parameters of interest (*5*). Each parameter is the combination of a variable, a level (geopotential height, e.g. 500 mb) and an analysis time. The right and left arrows allow including new parameters to the list (in some cases, holding the Ctrl key multiple choices can be made).
   Note that surface variables do not require a level value (it is set to *2D* by default); moreover, some variables have also associated a forecast time (e.g. large scale precipitation), since they are not assimilated in the analysis process.
4. Finally, click on the **get data** button (*6*) to get the reanalysis data interpolated to the desired grid points (the DAP uses a bilinear interpolation scheme to interpolate the data fields to a regular or irregular grid defined by the user). The status of the resulting data request job can be monitored in the **My Home** tab (see Sec. 1).

The figure below shows the file obtained from the portal for the yearly averaged values of the 2m temperature interpolated in four grid points around Madrid (see the previous figure). The four series exhibit a bias in the mean due to the North-South gradient.
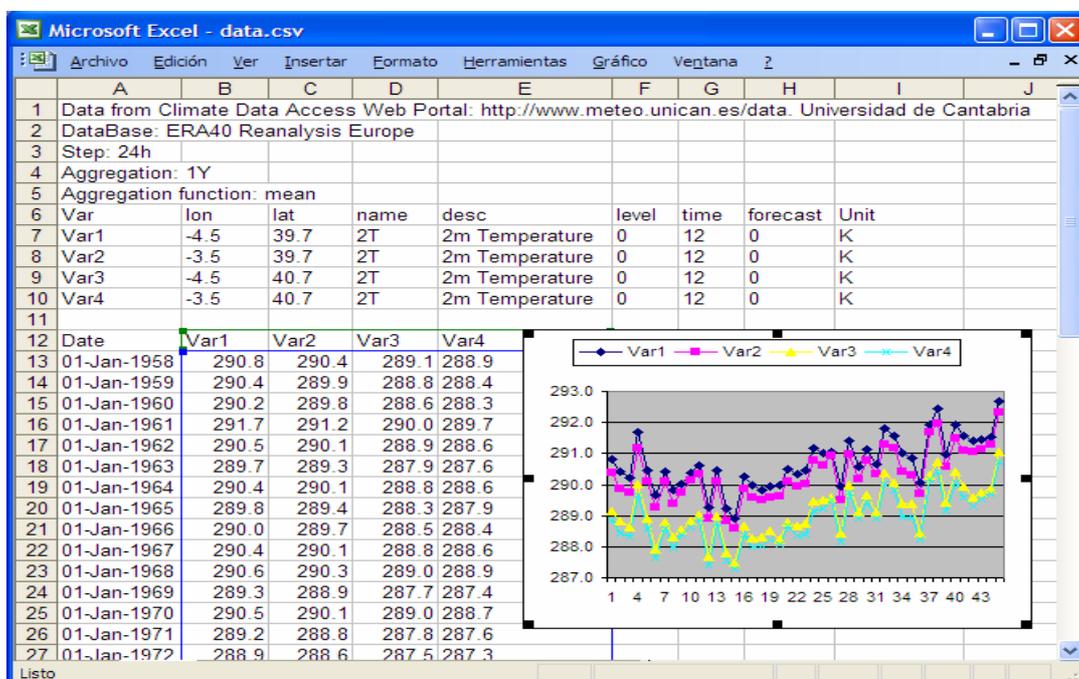
Note that in this case there is a total of four time parameters involved in a data request. First, the original **temporal resolution** of the data, which is daily or intra-daily in all the cases (12 hours in the above example), then the **analysis time** (which is the available model initialization hours within the day, 0H and 12H in the above example); then the **forecast time**, which is the lead time for the model "forecast"

variables (and has a zero value by default for most of the variables) and, finally the **aggregation period** (daily, 5-day, monthly, yearly). The possible analysis times are 0, 6, 12, and 18H for a 6H resolution dataset, and are a subset of this list for other resolution values (e.g. 0H and 12H in the example above for a 12H resolution dataset). Note that by selecting a daily aggregation, the data for the different analysis times for a particular parameter selected in the query will be aggregated (using the mean, maximum, etc.) in the resulting file. Most of the seasonal and climate change model outputs described below store data only for a particular time (most of the times for 0H base hour); therefore, the analysis time of the reanalysis data should be chosen carefully to match the times of the forecast data in intercomparison studies.

NOTE. Due to the interpolation process, although the resolution of the grid can be set to any arbitrary value (e.g., a high-resolution grid), the real information contained on the gridded data is limited by the original resolution of the reanalysis dataset.

*NOTE. The amount of data that can be interpolated online in a single process is limited, since **the portal has not been designed for massive data access**. A normal request is limited to 10 MB which corresponds to the daily series for 10 parameters in a 10 grid points' lattice for the whole 40-years period (or 1 parameter in a grid with 10x10 points).*

3.2. Seasonal Prediction Data

The following seasonal predictions datasets are available in the portal:

- **DEMETER**. Multi-model seasonal prediction experiment including seven models ran for six months four times a year using 9 different perturbed initial conditions (9 members), http://www.ecmwf.int/research/demeter:
    - cnrm (from 1958 to 2001)
    - crfc (from 1980 to 2001)
    - lody (from 1974 to 2001)
    - scnr (from 1973 to 2001)
    - scwf (from 1958 to 2001)
    - smpi (from 1969 to 2001)
    - ukmo (from 1959 to 2001)
- **ENSEMBLES** Stream 1 (http://www.ecmwf.int/research/eu_projects_ensembles/). Check the help in the portal for updated information about this dataset.

The procedure to access seasonal forecast data is similar to the reanalysis case (Sec. 3.1) with some particularities. In this case, since the datasets are multi-model ensemble predictions, a single **model** and one or several **members** must be selected in order to obtain the data of interest.

For instance, the DEMETER dataset has seven models and nine members (the *All* option in the members "*Mem*" menu allows obtaining all the members of the prediction). The selection of the grid of interest works as in the reanalysis case (see Sec. 3.1).
Finally, the parameters' selection panel requires the **analysis and forecast months** (for instance, in the DEMETER dataset the simulations are initialized the first of

February, May, November and August, so there are four options for the analysis). The forecast months are refreshed automatically when changing the analysis time to show the next five forecast months (lead time 1 to 5 months). Note that although the analysis and forecast times are given in months, the data will be obtained in a daily basis, unless the monthly aggregation is selected. Note that each dataset has a **base hour** label indicating the day hour of the data (e.g. DEMETER parameters correspond to 0 hours' values).

3.3. Anthropogenic Climate Change (ACC) Simulations

In this case the datasets have been obtained from different sources. On the one hand the IPCC data centre (www.ipcc-data.org/ and www-pcmdi.llnl.gov/ipcc/about_ipcc.php) provides access to the models contributing the AR4; these datasets are referred to as *PCMDI_Model*. On the other hand, the ENSEMBLES project is storing a common list of parameters for different ACC experiments (www.mad.zmaw.de/projects-at-md/ensembles); these datasets are referred to as *CERA_Model*. Finally some datasets are obtained directly from local providers; these datasets are referred to as *Model*.

So far, the following data is available, but the portal provides updated information of new datasets included as they become available:

- **PCMDI_CGCM3.** Canadian Centre for Climate Modelling and Analysis (www.cccma.bc.ec.gc.ca/data/data.shtml), including 20th century (from 1951 to 2000) and scenarios A1B, B1 (daily data in pressure levels is only included for the periods 2046-2065 y 2081-2100).
- **CERA_MPI-ECHAM5,** including 20th century data (1961-2000) and scenarios A1B, B1, and A2 (2001-2100).
- **CNRM-CM3,** including 20th century (1961-2000) and scenarios A1B, B1, and A2 (2001-2100).

The procedure to access data in this case is similar to the reanalysis case (Sec. 3.1) with some particularities. In this case, since the models are ran under different **future scenarios**, the scenario of interest should be initially chosen (note that the time period changes from scenario to scenario). The example below shows a selection of data (year means obtained from the daily 12H fields) for the CGCM3 Canadian model for the 20th century run covering the period 1961-2000.

The panel for parameters selection includes the **analysis** and **forecast months**. Note that in this case the models are initialized just once, at the beginning of the

simulation period, and thus the analysis time does not hold in this case (it is always disabled, but it has been kept in the panel to keep consistency with other datasets). This panel has been designed to allow the selection of data for a single month or season during the simulation period. For instance to analyze the summer temperature increase for the 20th century, the only forecast months needed are June, July and August; this allows reducing the size of the data requests for seasonal studies. In the example below, only surface temperature for August is selected.
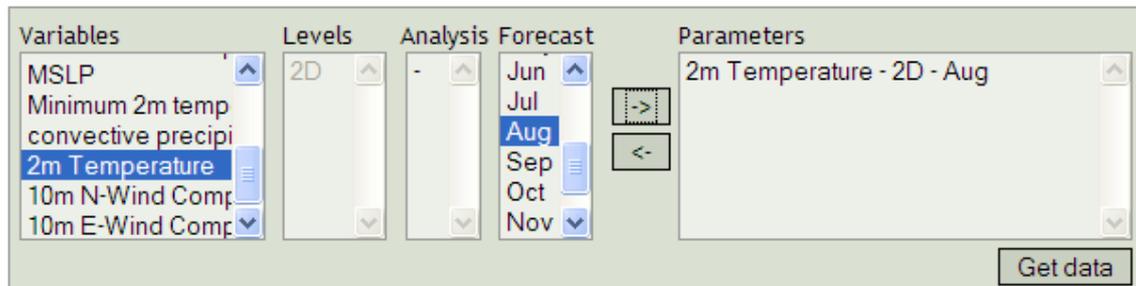
For instance, selecting only a grid point close to Madrid (see below) we can obtain the 20th century August surface temperature forecasted by the CGCM3 model.
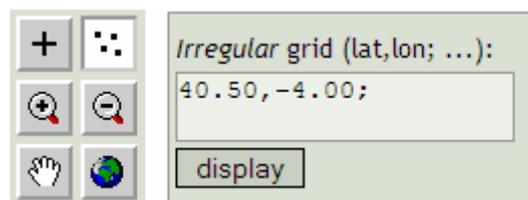
The resulting file (opened using a Spread-Sheet) is shown in the figure below; this example exhibits a trend similar to the observation and the reanalysis data shown in previous examples (Sec. 2 and 3.1). However, in this case, the data does not correspond year by year with the observed values.