Technical Notes
Santander Meteorology Group (CSIC-UC)
http://www.meteo.unican.es
SMG:01.2013

The SPECS-EUPORIAS Data Portal: THREDDS Data Server and R Interface

J. Bedia[1], M.E. Magariño[2], S. Herrera[2], R. Manzanas[1], J. Fernández[2], A.S. Cofiño[2] & J.M. Gutiérrez[1]
[1] *Instituto de Física de Cantabria, CSIC-Universidad de Cantabria, Spain.*
[2] *Dpto. de Matemática Aplicada y C.C. Universidad de Cantabria, Spain*
correspondence: joaquin.bedia@unican.es

version:v2.0–*27 May 2013*

Abstract

Different sector-specific impact activities to be undertaken in SPECS (http://www.specs-fp7.eu) and EUPORIAS (http://www.euporias.eu) projects require a reduced number of variables (typically at surface) from different data sources (mainly seasonal forecasts, reanalysis, and observations). The *SPECS-EUPORIAS Data Portal* has been established by the Santander Meteorology Group (UC-CSIC) as part of the data management activities in these projects to provide a unique access for these impact-relevant variables, gathered from existing datasets. This document briefly describes the current state and future plans of the data portal, which is based on a THREDDS data server providing metadata and data access using OPeNDAP and other remote data access protocols. Moreover, a R package[a] has also been developed for exploring and remotely accessing subsets of data, thus reducing the burden of data access in these activities.

User's Guide: https://www.meteo.unican.es/trac/meteo/wiki/SpecsEuporias

[a] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

1 Introduction and Motivation

The impact activities on seasonal timescales involved in the SPECS (http://www.specs-fp7.eu) and EUPORIAS (http://www.euporias.eu) projects require the use of different data sources (mainly seasonal forecasts, reanalysis, and observations). These activities include the calibration, downscaling, and modelling of sector-specific indices in agriculture, energy, health, etc., building on meteorological information. Typically, only a reduced subset of surface variables (precipitation, temperatures, mean sea level pressure, etc.) or in a reduced number of vertical levels (circulation and termodynamic drivers at, e.g., 850, 500, 200 hPa) is required for these activities. The *SPECS-EUPORIAS Data Portal* has been established by the Santander Meteorology Group (UC-CSIC) to gather the relevant information from existing datasets in order to provide a unique homogenized access to data for the SPECS and EUPORIAS partners —in particular for impact-users—.

The *SPECS-EUPORIAS Data Portal* is based on a THREDDS data server providing metadata and data access using OPeNDAP and other remote data access protocols. Moreover, since the R language (http://www.r-project.org) has been adopted for some key tasks in these projects —including the development of comprehensive validation and statistical-downscaling packages,— a user-friendly R package has been developed to explore and access the data portal. This package can be used in R programs to remotely access subsets of data, thus reducing the burden of data access (versions for Python and Matlab are also available under request). This package will be continuously updated (keep you informed at the user's guide URL above) as part of the data management activities to build a data bridge for impact users and for the R developments to be done in these projects.

This document briefly describes the initial state of the data portal, focusing on data from the ECMWF System4 seasonal model, as agreed in the downscaling parallel session of the kick-off meeting.

2 The THREDDS Data Server

The *SPECS-EUPORIAS Data Portal* is based on a password-protected THREDDS data server, providing metadata and data access to various datasets using OPeNDAP and other protocols. The variable names, units and additional metadata follow the CF convention[1]. The variables are spatial grids based on multidimensional arrays of indexed values, following Unidata's _Coordinate convention[2,3].

Typically the data portal includes information at a daily resolution, but mothly-aggregated values could be also provided in some cases due to data limitations —in particular, Météo-France and Met Office have agreeed to provide monthly mean hindcasts for their use by the SPECS and EUPORIAS partners.— In general, the data available will be typical surface variables (e.g. precipitation and near-surface temperature), although several other variables (e.g. geopotential and temperature) on pressure levels will also be stored for the statistical downscaling activities.

The data gathering activities have initially focused on the ECMWF System4 seasonal model. The Meteorological Archival and Retrieval System (MARS) is the main repository of meteorological data at the ECMWF (European Centre for Medium-Range Weather Forecasts). It contains terabytes of operational and research data as well as data from special projects[4]. The large amount of information stored and the inherent complexities of data access, download and post-processing is a first shortcoming for a flexible use of these datasets by a large number of partners. To overcome this issue, a reduced subset of surface variables[5] (precipitation, temperatures and mean sea level pressure) have been downloaded from MARS (a colection of GRIB-1 files) at 0.75° spatial resolution and made available throught the *SPECS-EUPORIAS data portal*. The downloaded data has been exposed as three different virtual datasets using THREDDS:

- **System4 seasonal range (15 members)**: There are twelve initializations (hereafter called "runtimes") per year (the first of January, February, ...) running for 7 months (hereafter called simply "times"). An ensemble of 15 members is available for the whole 1981-2010 period.
- **System4 seasonal range (51 members)**: There are only four runtimes per year (the first of February, May, August and November) and the forecasts run for 7 months. An ensemble of 51 members is available for the whole 1981-2010 period.
- **System4 annual range (15 members)**: As in the previous case, there are four runtimes per year, but the forecasts run for 13 months. An ensemble of 15 members is available for the whole 1981-2010 period.

Data gathering activities will next move to the CFS (`http://cfs.ncep.noaa.gov`) version 2 hindcast, developed at the Environmental Modeling Center at NCEP and also to reanalysis and observational datasets.

Although the THREDDS server provides a web interface to explore and access the datasets, it is strongly recommented the use of OPeNDAP (DODS) client libraries for remote data access from scientific computing environments (`R`, Matlab, Python, etc.). For instance, the `meteoR R` package developed as a companion of the data portal is based on the Unidata's Common Data Model (CDM), which merges the netCDF, OPeNDAP, and HDF5 data models to create a common API for many types of scientific data[6] (a similar approach has been also followed for the Matlab implementation). Alternatively, the most recent NetCDF library versions provide access to OPenDAP datasets (this has been the choice for the Python implementation). In Sec. 3 we show a simple example of data access using the `meteoR` package. In particular different forecast datasets can by directly accessed using the `loadSeasonalForecast` function, allowing the retrieval of slices for a particular variable in any of the dataset dimensions (member/space/runtime/time). Further data access and processing examples are given in the **user's guide examples section**[7]. In addition, the web interface for the OPeNDAP service is described in Sec. 4.

[1] `http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.4/cf-conventions.html`
[2] `http://www.unidata.ucar.edu/software/netcdf-java/reference/CoordinateAttributes.html`
[3] `http://www.unidata.ucar.edu/software/netcdf-java/tutorial/GridDatatype.html`
[4] `http://www.ecmwf.int/services/archive`
[5] `http://www.ecmwf.int/products/changes/system4/technical_description.html#description`
[6] `http://www.unidata.ucar.edu/software/netcdf-java/documentation.htm`
[7] `https://www.meteo.unican.es/trac/meteo/wiki/SpecsEuporias/RPackage/Examples`

3 Accesing the Data Portal via R

The *SPECS-EUPORIAS Data Portal* can be remotely accessed from R using the `loadSeasonalForecast` function included in the `meteoR` package[8]. This function automatically handles the different variable dimensions, given a few simple arguments for subset definition. In addition, instead of retrieving a NetCDF file that needs to be opened and read, the requested data is directly loaded into the current R working session, according to a particular structure described below, allowing for efficient data analysis and/or visualization.

Some preliminary steps are required before starting to use the `loadSeasonalForecast` function (see the user's guide for further information):

1. The **meteoR_v1_0.zip** file containing the required data need to be downloaded in a convenient directory and unzipped, resulting in a new directory called **meteoR**. Then, within the current R session, it is necessary to set this folder as the working directory.
2. The `rJava` package must be installed (e.g. from R-CRAN).
3. The init.R script must be sourced:

```
> source("init/init.R")
```

4. Before accessing the data, authentication is required at the *SPECS-EUPORIAS Data Portal*. This can be done directly from R with the following calls to `rJava` commands, as illustrated below.

```
> username <- "myUsername"
> password <- "myPassword"
> aux <- .jnew("ucar/nc2/util/net/HTTPBasicProvider", username, password)
> J("ucar.nc2.util.net.HTTPSession")$setGlobalCredentialsProvider(aux)
```

Once these requirements are fulfilled, the function can be used for the rest of the session. In the next lines we describe an illustrative example loading data from the dataset "System4 seasonal range (15 members)" for one-month lead time hindcast (1981-2000) of minimum surface temperature for January over a window centered in Europe (-10°W – 30°E and 35°S – 65°N). In order to provide an homogeneized variable definition for the different datasets, `meteoR` is based on a standard pre-defined "vocabulary" which is mapped to each particular dataset by means of variable-to-variable conversion "dictionaries". More information regarding the use of the dictionary and standard variables is provided in the corresponding user's guide section[9].

```
> ds <- "http://www.meteo.unican.es/tds5/dodsC/system4/System4_Seasonal_15Members.ncml"
> dic <- "./datasets/forecasts/System4/System4_Seasonal_15Members.dic"
> openDAP.query <- loadSeasonalForecast(dataset = ds, dictionary = dic,
+      var = "tasmin",
+      lonLim = c(-10,30), latLim = c(35,65),
+      season = 1, years = 1981:2000, leadMonth = 1,
+      members = 1)
```

The result is a list containing the requested data and all the necessary metadata information:

```
> str(openDAP.query)
List of 5
 $ VarName      : chr "tasmin"
 $ MemberData   :List of 1
  ..$ : num [1:589, 1:2120] 3.41 5.48 6.09 4.14 2.94 ...
 $ LatLonCoords : num [1:2120, 1:2] 64.5 63.7 63 62.2 61.5 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:2] "lat" "lon"
 $ RunDates     : chr [1:589] "1981-12-01" "1981-12-01" "1981-12-01" "1981-12-01" ...
 $ ForecastDates:List of 2
```

[8] https://www.meteo.unican.es/trac/meteo/wiki/SpecsEuporias
[9] https://www.meteo.unican.es/trac/meteo/wiki/SpecsEuporias/RPackage

```
..$ Start: POSIXlt[1:589], format: "1982-01-01" "1982-01-02" "1982-01-03" "1982-01-04" ...
..$ End  : POSIXlt[1:589], format: "1982-01-02" "1982-01-03" "1982-01-04" "1982-01-05" ...
```

The argument `MemberData` is a list of matrices (one for each member) containing the data; in this case the dimension of the single matrix (only one member is requested) is 589 (times) for 2120 (gridpoints, covering the specified region with the original 0.75° resolution of the dataset). The times correspond to the daily data requested for January (`season=1`) for the 20 years (`years = 1981:2000`). This is a total of 620 dates; however, since no one-month lead time forecast (`leadMonth = 1`) is available for january 1981 —since the corresponding run time would be the 1st of December 1980, not available for this dataset—, the `ForecastDates` vector start in 1982, leading to a total length of 589.

The rest of elements provide geolocation (`LatLonCoords`, for the 2120 gridpoints) and time (`RunDates` and `ForecastDates`) metadata. Thus, for each of the 589 times of the data matrix, both the corresponding run time `RunDates` (date when the model was initialized) and verification bounding time `ForecastDates` (period corresponding to the data value) are provided. Note that for instantaneous variables, the `ForecastDates$Start` and `ForecastDates$End` vectors are identical.

Finally, it is important to remark that the `loadSeasonalForecast` function has been implemented to access seasonal slices (defined by the `season` argument[10]) for a given period (defined by the `years` period, e.g. `years = 1981:2000`) with a homogeneous forecast lead time (as given by the `leadMonth` argument, e.g. `leadMonth = 1` for one-month lead time) related to the first month of the season. Thus, `season=c(1,2,3)` for `years = 1995:2000` and `leadMonth = 1` will return the following series: JFM 1995 (from the December 1994 runtime forecast), ..., JFM 2000 (from the December 1999 runtime forecast). Note that it is also possible to work with two-year seasons, such as DJF. In this case, `season=c(12,1,2)` for `years = 1995:2000` and `leadMonth = 1` will return the following series: DJF 1994/1995 (from the November 1994 runtime forecast), ..., DJF 1999/2000 (from the November 1999 runtime forecast).

4 Accesing the Data Portal via Web

The *SPECS-EUPORIAS Data Portal* can be accessed through the **Data Portal URL** provided in the abstract. First of all, an authentication dialog will request a valid user name and password.

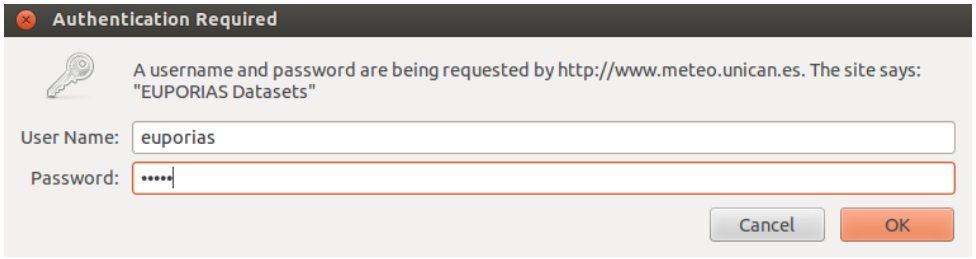
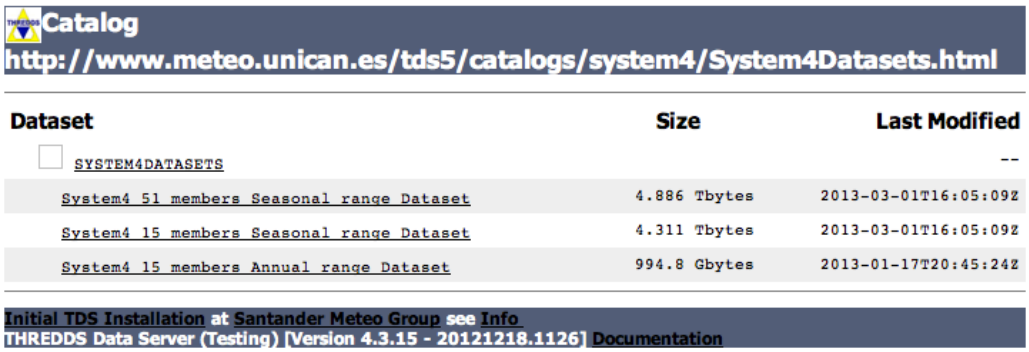


Figure 1: Authentication dialog

Afterwards, the different datasets described in Sec. 2 are listed as links in the web browser window (Fig. 2). By clicking in any of the datasets, a new window will appear providing information on the variables and geospatial and time coverages, and offering different options for data access and/or visualization (Fig. 3). Currently, only the OPeNDAP access service is fully operative in the portal. Therefore, in this example, we will illustrate the use of this service, which allows selecting time/spatial data slices from the OPeNDAP data access form shown in Fig. 4 and downloading the resulting data in both ASCII and Binary formats.

[10]Seasons can be defined in several ways: A single month (`season = 1` for January, as in the example above), a standard season (e.g. `season=c(1,2,3)` for JFM, or `season=c(12,1,2)` for DJF), or any period of consecutive months (e.g. `season=c(1,2,3,4,5,6)`, for the first half of the year).

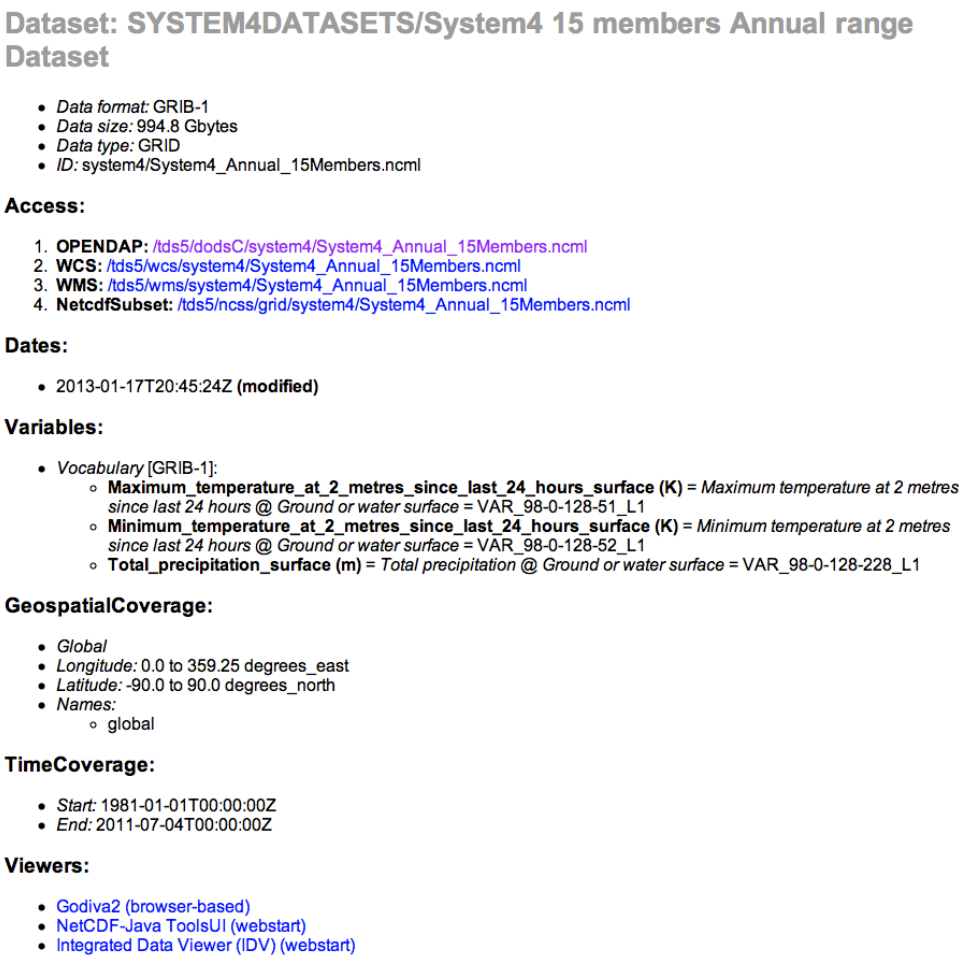


Catalog
http://www.meteo.unican.es/tds5/catalogs/system4/System4Datasets.html

Dataset	Size	Last Modified
SYSTEM4DATASETS		--
System4 51 members Seasonal range Dataset	4.886 Tbytes	2013-03-01T16:05:09Z
System4 15 members Seasonal range Dataset	4.311 Tbytes	2013-03-01T16:05:09Z
System4 15 members Annual range Dataset	994.8 Gbytes	2013-01-17T20:45:24Z

Initial TDS Installation at Santander Meteo Group see Info
THREDDS Data Server (Testing) [Version 4.3.15 - 20121218.1126] Documentation

Figure 2: Catalog of the EUPORIAS-SPECS System4 datasets. Note that although they only include a few variables, their size range from one to four Terabytes.



Dataset: SYSTEM4DATASETS/System4 15 members Annual range Dataset

- *Data format:* GRIB-1
- *Data size:* 994.8 Gbytes
- *Data type:* GRID
- *ID:* system4/System4_Annual_15Members.ncml

Access:

1. **OPENDAP:** /tds5/dodsC/system4/System4_Annual_15Members.ncml
2. **WCS:** /tds5/wcs/system4/System4_Annual_15Members.ncml
3. **WMS:** /tds5/wms/system4/System4_Annual_15Members.ncml
4. **NetcdfSubset:** /tds5/ncss/grid/system4/System4_Annual_15Members.ncml

Dates:

- 2013-01-17T20:45:24Z **(modified)**

Variables:

- *Vocabulary* [GRIB-1]:
  - **Maximum_temperature_at_2_metres_since_last_24_hours_surface (K)** = *Maximum temperature at 2 metres since last 24 hours @ Ground or water surface* = VAR_98-0-128-51_L1
  - **Minimum_temperature_at_2_metres_since_last_24_hours_surface (K)** = *Minimum temperature at 2 metres since last 24 hours @ Ground or water surface* = VAR_98-0-128-52_L1
  - **Total_precipitation_surface (m)** = *Total precipitation @ Ground or water surface* = VAR_98-0-128-228_L1

GeospatialCoverage:

- *Global*
- *Longitude:* 0.0 to 359.25 degrees_east
- *Latitude:* -90.0 to 90.0 degrees_north
- *Names:*
  - global

TimeCoverage:

- *Start:* 1981-01-01T00:00:00Z
- *End:* 2011-07-04T00:00:00Z

Viewers:

- Godiva2 (browser-based)
- NetCDF-Java ToolsUI (webstart)
- Integrated Data Viewer (IDV) (webstart)

Figure 3: Detail of a particular dataset with information on the included variables and geospatial and time coverages. The different options for data access and visualization are also shown.

Note that, as explained before, the variables provided by the data portal (e.g. minimum temperature) are stored as gridsets. Thus, in addition to these variables, also auxiliary coordinate variables (lat, lon, run, time, member) should be handled for geo-temporal data referencing (see Fig. 4). Moreover, three time coordinates are included as referece for different grid variables because they are defined for different forecast times (one extra time for precipitation and different temporal resolution for mean sea level pressure). Note that this highly complicates the direct analysis of the data and, hence, this options is only recommend for data exploration. In the following we show how to use this service to explore the structure of the datasets and to obtain simple pieces of information in ASCII format.

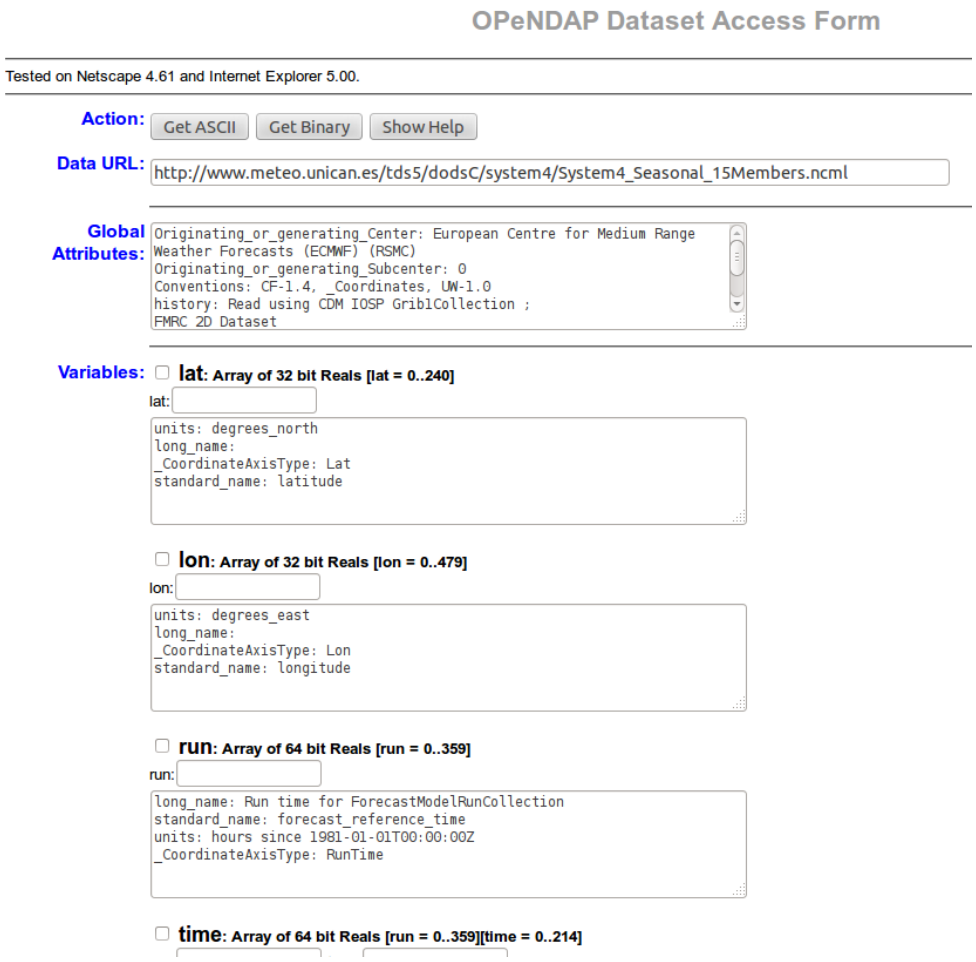


Figure 4: Detail of the OPeNDAP dataset access form for a particular dataset.

By default, if no specifications are given in the different subsetting boxes of the OpenDAP form, the whole data on the whole spatio/temporal and member ranges of the dataset would be accessed. However, this option will raise an error due to the large size of the request (the maximum size of a single request has been set to 100 Mbytes in the `SPECS-EUPORIAS data portal` for the sake of multi-connection efficiency). The basic steps to retrieve subsets of data are the following:

- To select a variable click on the checkbox to its left.
- To constrain the variable, edit the information that appears in the text boxes below the variable. This is a vector of integers indicating index positions of length three, with the following order: `[start:stride:end]`.
- To get ASCII or binary values for the selected variables, click on the Get ASCII or Get Binary buttons of the Action field. Note that the URL displayed in the Data URL field is updated as you select and/or constrain variables. The URL in this field can be cut and pasted in various DODS clients.

The main disadvantage of the OpenDAP service from the end-user point of view is that the specifications for subsetting dimensions are not given in their original magnitudes (i.e., latitudes and longitudes are not given in decimal degrees), but by the indexes of their position along their respective axes (note that first index value is always 0). Thus, to find out the indexes for the desired selection, we need to dump and analyze the particular values defined in the coordinate variable. For instance, Fig. 5 shows the 241 values defined for the `lat` (latitude) coordinate, as provided by the Get ASCII option (checking the corresponding check-box).

```
Dataset {
    Float32 lat[lat = 241];
    Float32 lon[lon = 480];
} system4/System4_Annual_15Members.ncml;
---------------------------------------------
lat[241]
90.0, 89.25, 88.5, 87.75, 87.0, 86.25, 85.5, 84.75, 84.0, 83.25, 82.5, 81.75, 81.0, 80.25, 79.5, 78.75,
78.0, 77.25, 76.5, 75.75, 75.0, 74.25, 73.5, 72.75, 72.0, 71.25, 70.5, 69.75, 69.0, 68.25, 67.5, 66.75,
66.0, 65.25, 64.5, 63.749996, 62.999996, 62.249996, 61.499996, 60.749996, 59.999996, 59.249996, 58.499996,
57.749996, 56.999996, 56.249996, 55.499996, 54.749996, 53.999996, 53.249996, 52.499996, 51.749996,
50.999996, 50.249996, 49.499996, 48.749996, 47.999996, 47.249996, 46.499996, 45.749996, 44.999996,
44.249996, 43.499996, 42.749996, 41.999996, 41.249996, 40.499996, 39.749996, 38.999996, 38.249996,
37.499996, 36.749996, 35.999996, 35.249996, 34.499996, 33.749996, 32.999996, 32.249996, 31.499996,
30.749996, 29.999996, 29.249994, 28.499994, 27.749994, 26.999994, 26.249994, 25.499994, 24.749994,
23.999994, 23.249994, 22.499994, 21.749994, 20.999994, 20.249994, 19.499994, 18.749994, 17.999994,
17.249994, 16.499994, 15.749994, 14.999994, 14.249994, 13.499994, 12.749994, 11.999994, 11.249993,
10.499993, 9.749993, 8.999993, 8.249993, 7.4999933, 6.7499933, 5.9999933, 5.2499933, 4.4999933, 3.749993,
2.999993, 2.249993, 1.499993, 0.7499929, -7.1525574E-6, -0.7500072, -1.5000073, -2.2500074, -3.0000074,
-3.7500074, -4.5000076, -5.2500076, -6.0000076, -6.7500076, -7.5000076, -8.250008, -9.000008, -9.750008,
-10.500008, -11.250008, -12.000008, -12.750009, -13.500009, -14.250009, -15.000009, -15.750009, -16.500008,
-17.250008, -18.000008, -18.75001, -19.50001, -20.25001, -21.00001, -21.75001, -22.50001, -23.25001,
-24.00001, -24.75001, -25.50001, -26.25001, -27.00001, -27.75001, -28.50001, -29.25001, -30.00001,
-30.75001, -31.50001, -32.25001, -33.00001, -33.75001, -34.50001, -35.25001, -36.00001, -36.75001,
-37.50001, -38.25001, -39.00001, -39.75001, -40.50001, -41.25001, -42.00001, -42.75001, -43.50001,
-44.25001, -45.00001, -45.75001, -46.50001, -47.25001, -48.00001, -48.75001, -49.50001, -50.25001,
-51.00001, -51.75001, -52.50001, -53.25001, -54.00001, -54.75001, -55.50001, -56.25001, -57.00001,
-57.75001, -58.50001, -59.25001, -60.00001, -60.75001, -61.50001, -62.25001, -63.00001, -63.75001,
-64.500015, -65.250015, -66.000015, -66.750015, -67.500015, -68.250015, -69.000015, -69.750015, -70.500015,
-71.250015, -72.000015, -72.750015, -73.500015, -74.250015, -75.000015, -75.750015, -76.500015, -77.250015,
-78.000015, -78.750015, -79.500015, -80.250015, -81.000015, -81.750015, -82.500015, -83.250015, -84.000015,
-84.750015, -85.500015, -86.250015, -87.000015, -87.750015, -88.500015, -89.250015, -90.000015
```

Figure 5: Text file displaying the values for the `lat` (latitude) coordinate variable.

Using these facilities it can be obtained after some calculations that the closest `lat` and `lon` coordinates for a particular location of interest (e.g. Madrid) are 66 and 475, respectively. Thus, the time series for Madrid corresponding to the example described in the previous section (minimum temperature forecasts for January with one-month lead time, i.e. from the simulations started the first of December) could be requested as shown in Fig. 6. Note that the indices selected for the run coordinate correspond to the December initilizations (index positions 11, 23,...; note that indexes start in 0) and for the time coordinate correspond to January (positions, 31 to 62, in days after the run time). Note that the proper use of this service requires a full understanding of the data structure and, therefore, it is only advised for data exploration.

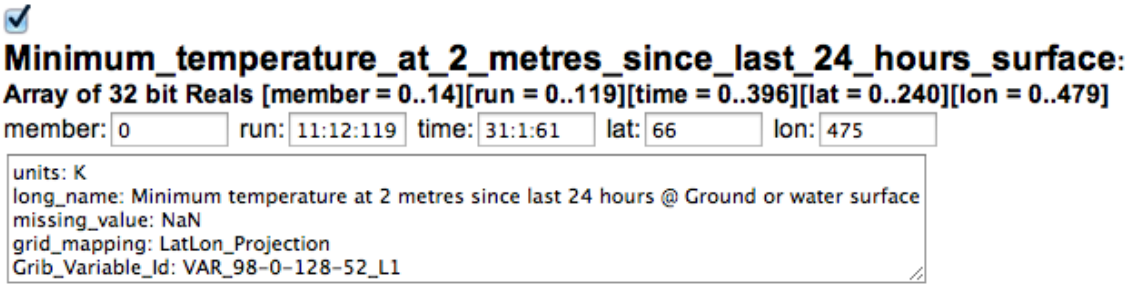


Figure 6: Detail of the query from the OPeNDAP dataset access form to retrieve a subset (a time series for a single gridbox) of minimum temperature.