The SPECS-EUPORIAS Data Portal

Different sector-specific impact activities to be undertaken in ?SPECS and ?EUPORIAS 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. The data portal is based on a THREDDS data server providing metadata and data access using OPeNDAP and other remote data access protocols. Moreover, a user-friendly ?R package has also been developed for exploring and remotely accessing subsets of data, thus reducing the burden of data access in these activities. This package will be also a key component for other tasks of the projects based on R, including the validation and downscaling packages to be developed within SPECS and sector-specific calibration and modeling tools to be developed in EUPORIAS.

This trac/wiki page provides an up-to-date description of the SPECS-EUPORIAS Data Portal, including information of the available datasets and the documentation and code of the R package. This page is currently under construction, but both a first tutorial describing the basic functioning and a first version of the R package (a R function) are already available::

Code: ?loadSystem4.R

Tutorial: PDF file?

Contents (under development):

Table of Contents

Introduction and Motivation

The impact activities on seasonal timescales involved in ?SPECS and ?EUPORIAS 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 informed at the documentation 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 current state of the data portal, which has initially focused on data from the *ECMWF's System4 seasonal model*, as agreed in the downscaling parallel session of the kick-off meeting.

The THREDDS Data Server

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

Typically the data portal will include information at a daily resolution, but monthly-aggregated values could be also provided in some cases due to data limitations (in particular, *Mètèo-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 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[3]. 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[4] (precipitation, temperatures and mean sea level pressure) have been downloaded from `MARS` (a colection of `GRIB-1` files) at 0.75º spatial resolution and made available throught the *SPECS-EUPORIAS data portal*. The downloaded data has been exposed as three different virtual datasets using `TDS`:

- **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.

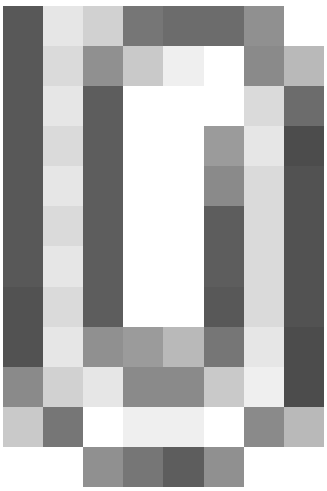
Data gathering activities will next move to the CFS (?http://cfs.ncep.noaa.gov) version 2 hindcast, developed at the *Environmental Modeling Center at NCEP* and also to reanalysis and observational datasets.

Although the `TDS` provides a web interface to explore and access the datasets (shown in web access section), it is strongly recommented the use of `OPeNDAP` (a.k.a. `DODS`) client libraries to remotely access the data from scientific computing environments (`R`, `Matlab`, `Python`, etc.). For instance, the `R` function provided in this tutorial is based on the *NetCDF Java* OPeNDAP client[5], using the `rJava` `R` package (a similar approach is been also made for the `Matlab` implementation). Alternatively, the most recent *NetCDF library* versions provide access to `OPeNDAP` datasets (this is the solution for the `Python` implementation). In the following, we show a simple example of data access using the `R` package developed as part of the data portal. In particular the *System4* datasets can by directly accessed using the `loadSystem4` function, allowing the retrieval of slices for a particular variable in any of the dataset dimensions (`member`/space/`runtime`/`time`). Note that a more ellaborated worked example using `R` is shown in the R example section. Moreover, for a better understanding of the datasets structure, the use of the web interface for the `OPeNDAP` service is also illustrated web access section.

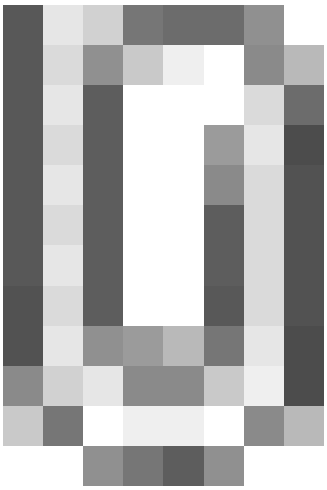
Accesing the Data Portal via R

Accesing the Data Portal via Web

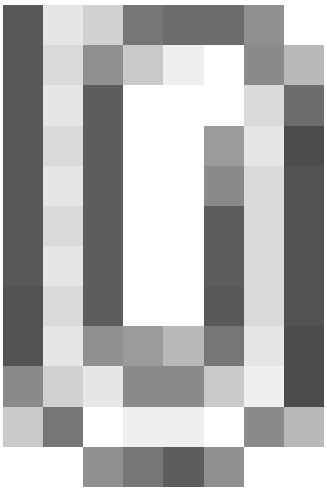
The *SPECS-EUPORIAS Data Portal* can be accessed through the **Data Portal URL** provided in the abstract. First of all, an authentication dialog will request a valid user name and password.



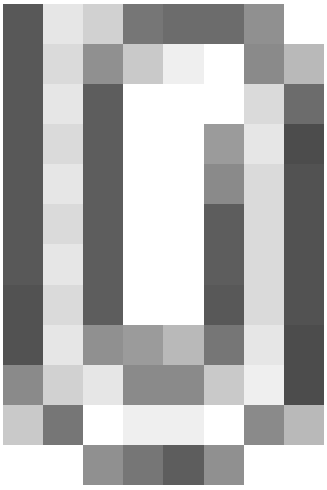
Afterwards, the different datasets described in TDS section are listed as links in the web browser window.



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.



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. \ref{fig:opendapwin} and downloading the resulting data in both *ASCII* and *Binary* formats.

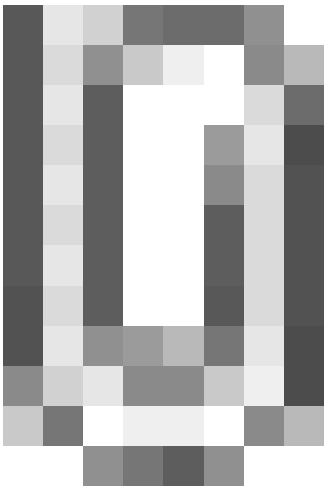


Note that, as explained before, the variables provided by the data portal (e.g. minimum temperature) are stored as gridsets. Thus, in addition to these variables, also auxiliary coordinate variables (lat, lon, run, time, member) should be handled for geo-temporal data referencing (see Figure). Moreover, 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`.

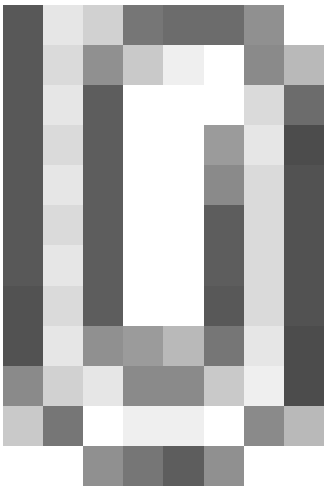
By default, if no specifications are given in the different subsetting boxes of the OpenDAP form, the whole data on the whole spatio/temporal and member ranges of the dataset would be accessed. However, this option will raise an error due to the large size of the request (the maximum size of a single 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:

1. To select a variable click on the checkbox to its left.
2. 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: \texttt{[start:stride:end]}.
3. 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 OPeNDAP 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.~\ref{fig:latlonDump} shows the 241 values defined for the `lat` (latitude) coordinate, as provided by the *Get ASCII* option (selecting the corresponding check-box).



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 Figure



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.

Example of Data Analysis with R

References

1. 1. ?http://www.unidata.ucar.edu/software/netcdf-java/reference/CoordinateAttributes.html
2. 2. ?http://www.unidata.ucar.edu/software/netcdf-java/tutorial/GridDatatype.html
3. 3. ?http://www.ecmwf.int/services/archive/
4. 4. ?http://www.ecmwf.int/products/changes/system4/technical_description.html#description
5. 5. ?http://www.unidata.ucar.edu/software/netcdf-java/documentation.htm