Horizon 2020 Societal challenge 5:
Climate action, environment, resource efficiency and raw materials

VERIFY

Observation-based system for monitoring and verification of greenhouse gases

GA number 776810, RIA

Deliverable number (relative in WP)	**D6.8**
Deliverable name:	First version of the VERIFY database and data-management infrastructure
WP / WP number:	6
Delivery due date:	Month 19 (31/08/2019) shifted to Month 27 (30/04/2020)
Actual date of submission:	Month 27 (14/04/2020)
Dissemination level:	Public
Lead beneficiary:	CEA
Responsible	Philippe Peylin
Contributor(s):	Matthew McGrath
Internal reviewer:	/

Changes with respect to the DoA
This deliverable was delayed. Original submission date was month 19. Given the project review meeting, which took place on Oct. 4, 2019, and feedback from the panel about communication activities and data management within the project, several extra months were taken (in agreement with the project officer) to update the data and information infrastructure within the project to adequately respond to feedback.
Dissemination and uptake **(Who will/could use this deliverable, within the project or outside the project?)**
This deliverable describes how data is managed and stored both internally and made available to external users. It is thus an invaluable tool to data access for both internal and external users. Web addresses for data access are given in the core of the document.
Short Summary of results (<250 words)
The VERIFY project currently stores 1800 GB of data, including 71 GB of output fluxes. This is expected to grow over the final two years of the project, as more groups take advantage of high-resolution forcing data provided by the project. In order to share this data among partners, a series of mechanisms were put in place, focused on a dedicated data server that allows users to download any dataset. The data is separated by input and output data and by primary chemical species to allow easier access. Additional services were placed on top of this server, permitting users to browse and visualize the various datasets without downloading it. **Extensive work has gone into harmonizing both the input and the output datasets in terms of file formats, grids, variable names, variable units, and aggregating spatial products into time series by country/region. This work enabled the creation of three distinct tools to manipulate the data. The first is for advanced scientific users, and helps them personalize plots. The second is for regular scientific users, and helps them explore time series in a user-friendly manner. The third allows policy-makers to view a set of already-completed plots summarizing the results of all countries and multiple regions; these plots were created in coordination with WP5 for use in the annual synthesis paper and with WP1 for use in the country summary Fact Sheets.**

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

Evidence of accomplishment (report, manuscript, web-link, other)
The data management system can be accessed through the main VERIFY website: https://verify.lsce.ipsl.fr/index.php/products **The bulk of the data remains password protected.**

Version	Date	Description	Author (Organisation)
V0	06/09/2019	Creation/Writing	Matthew McGrath, Patrick Brockman (CEA)
V0.1	03/19/2020	Writing/Formatting/Delivery	Matthew McGrath, Philippe Peylin (CEA)
V1	14/04/2020	Formatting/Delivery on the Participant Portal	Philippe Peylin and Aurelie Paquirissamy

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

1. Glossary

Abbreviation / Acronym	Description/meaning
TDS	THREDDS Data Server
THREDDS	Thematic Real-time Environmental Distributed Data Services
LSCE	Laboratoire des Sciences du Climat et de l'Environnement
IT	Information technology
NetCDF	A file data format for storing scientific information

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

2. Executive Summary

This deliverable describes the first version of a state-of-the-art database and data infrastructure developed on the CEA-LSCE premises. After a first year of flux estimates from the different work packages the current data volume excedes 1800 GB of data. In order to share this data among partners, a series of mechanisms were put in place, focused on a dedicated data server that allows users to download any dataset. The data is separated by input and output data and by primary chemical species to allow easier access. Additional services were placed on top of this server, permitting users to browse and visualize the various datasets without downloading them. The database can be access through the "Products" page of the main VERIFY web site: http://verify.lsce.ipsl.fr/index.php/products

Extensive work has gone into harmonizing both the input and the output datasets in terms of file formats, grids, variable names, variable units, and aggregating spatial products into time series by country/region. This work enabled the creation of the following different data services :

i) A searchable catalog that displays all datasets gathered in the database. This facility was developed for expert users who need to upload the raw data and access ancillary associated files (information files associated with each output of VERIFY). In addition, specific "Jupyter notebooks" are provided to the expert users in order to share knowledge and computational programs for specific data analysis. These notebooks are stored in a GitHub repository.

ii) A user-friendly time series visualization tool that allows the user to build his own time series of different GHG fluxes from different data sets and for different pre-defined regions (individual EU countries or groups of countries). Additional tools, such as a mapping facility, will be added to this section of the data visualization infrastructure.

iii) A policy-makers tool to view a set of already-completed plots summarizing the GHG fluxes for all countries and multiple regions; these plots were created in coordination with WP5 for use in the annual synthesis paper and with WP1 for use in the country summary Fact Sheets. The figure below provides an illustration of the policy-maker oriented visualization tool. It provides the main interface on the left for selecting the GHG and regions and one example of such synthesis for the anthropogenic CO2 emissions.

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

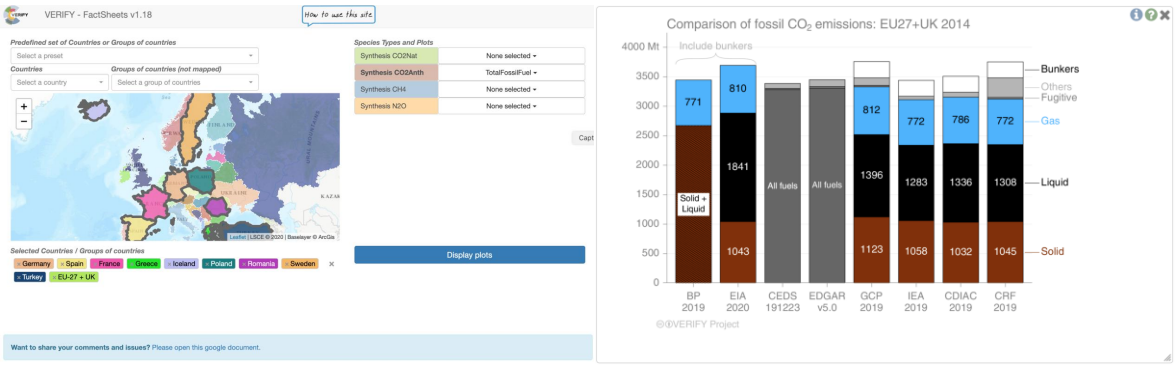

Figure 1 : Illustration of the GHG synthesis plot facility: left: web interface to select the specific synthesis plot; right: example of such synthesis for the anthropogenic CO2 emissions for the EU27+UK.

3. Introduction

Data management ensures the proper dissemination of products from projects, as well as facilitating harmonization across different work packages and research groups. Sharing large data files presents numerous challenges. Email is a poor method to share large files, to the point that every email provider caps the maximum file size to prevent email traffic from clogging network connections. In addition, the lack of coordination among groups can quickly lead to certain groups using different versions of the same file, sabotaging attempts at harmonization. To overcome these two major challenges to data management in VERIFY, we selected a centralized system known as the THREDDS Data Server (TDS) which allows the project coordination team to upload data files to a dedicated data server and then users to download them to their local machines. THREDDS catalogs enable one to mark datasets with meta-data, making browsing more convenient.

TDS was designed to work with the NetCDF data format of files, offering a variety of services. These services include direct download of the whole dataset (via HTTPS), or download, access and display part of the data (Web Coverage Service, WCS; NetCDF Subset Service, NCSS; WMS; OpeNDAP; DAP4). NetCDF files are the preferred file format for representing time-varying spatial data in the climate community, and many tools have been developed over the years to construct, modify, and manipulate these files. NetCDF provides an abstraction layer that allows users to access individual data elements without knowing the details of how they are stored, which makes them faster to access and easier to transfer across different computing systems. This is a major advantage in fields like climate science where large amounts of data are generated on different supercomputer architectures. In addition, metadata can be stored with the variables for more transparent documentation. Finally, the NetCDF format offers the same compression level as classical binary files.

We recognized that many people in the policy community have no knowledge of NetCDF file formats. Therefore, we worked to put in place structures which take advantage of the file format but hide the complexity of the underlaying data layer. This permits us to disseminate our results as widely as possible. This report explains these structures in more detail, including relevant steps in the data processing chain to make results from research groups in VERIFY available to both internal and external users.

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

4. Setup

The data management infrastructure in VERIFY is housed at the Laboratoire des Sciences du Climat et de l'Environnement (LSCE) outside of Paris, France. The main data storage takes place on a hard drive attached to a large computer cluster maintained by the local information technology (IT) team, which ensures professional care and maintenance. The local IT team has also configured the publically-accessible webpage that permits users to access the resources, in conjunction with a specialist in data visualization at the LSCE and members of the VERIFY coordination team. The computing cluster also ensures sufficient computational power for harmonizing (processing) the datasets as described in Section 5.

The data management infrastructure is accessible through the main VERIFY website on the "Products" page: **https://verify.lsce.ipsl.fr/index.php/products.** The layout of this page is illustrated in the figure below.

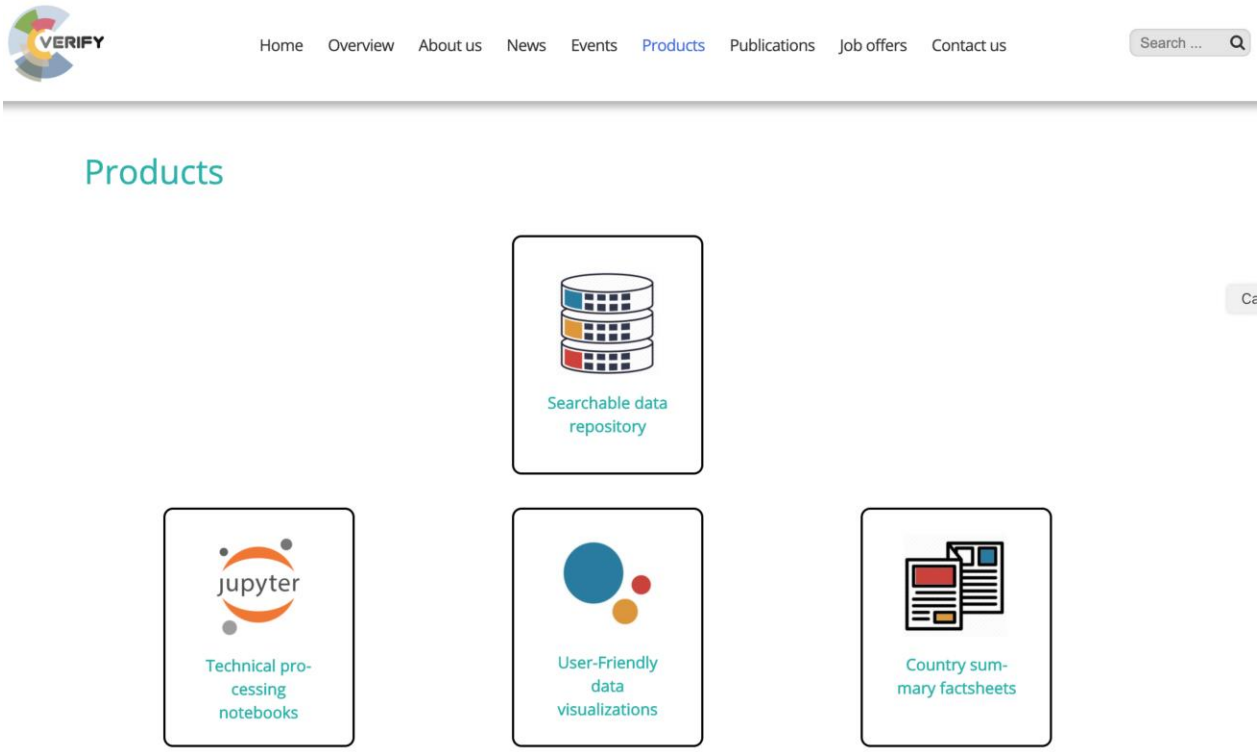

Figure 2: Main display from the "Products" page of the VERIFY web site.

From this page, the user has the choice of accessing several services:

1. A searchable catalog of all the data products (icon on the top of Figure 2);
2. A selection of scripts in the form of Jupyter notebooks for creating customizable plots (lower left icon);

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

3. A tool for real-time visualization, manipulation, and comparison of time-series at the country and regional levels (lower central icon);
4. And a collection pre-made synthesis graphics for each country and region that illustrate the main results of the project (lower right icon).

All of these options are regularly updated with at least annual updates for the pre-made synthesis as the project completes an operational cycle. The different data exploration tools correspond to different types of usage: from expert scientific users who need to access the raw data (service 1) to policy makers who need to be informed by the key results of the project (service 4). All these services are further described in the remainder of the document.

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

5. Data processing

The act of making data available to internal and external users begins with receiving the data from research groups, both inside and outside the project. For research groups kind enough to provide us with data who are outside VERIFY, we inform them of our data management system and obtain explicit consent before we share their products through our infrastructure. Research groups internal to VERIFY are already aware of their obligations under the project.

Research groups use a variety of different formats for their data, which makes developing generally-applicable tools a challenge. Our first step was to design a simulation protocol for two of the major work packages (WP3 and WP4) which specified a number of features to make analysis easier (Milestone 13). We have also been working with the other work packages to spread this same information, largely concerning how time is represented in the data, the spatial grid, the units of the variables, and full transparency and documentation about what the data includes.

Data used for input into models is typically loaded immediately onto the server after making sure documentation is available. As input data is not visualized by the tools described below, and as each group has different needs for the data driving their individual models, there is no reason to drastically change the formatting.

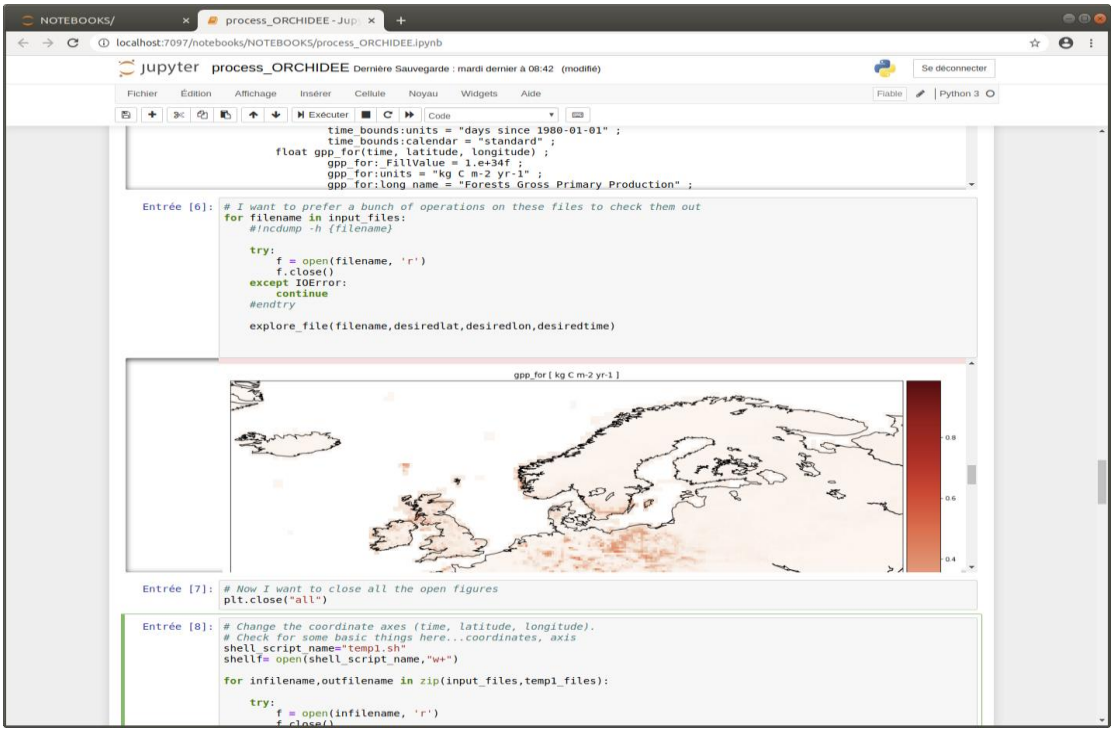

Figure 3: An example of a Jupyter notebook used to process data files before placement on the TDS.

However, we have set in place a specific "quality check" procedure for all output data that will be distributed under the VERIFY database. Computational tools are notoriously sensitive to minor modifications and typographical errors, refusing to work with simple spelling mistakes. Despite our efforts to promote a single file format, therefore, our first action upon receiving output data from groups is therefore to examine it with a script known as a Jupyter notebook. Figure 3, above, illustrates the content of a Jupyter notebook which is an open-source web application that permits one to combine coding language, mark-up text, equations, and visualizations all in the same package script. The main objectives were to:

- One notebook, based on the Python scripting language, is created for each output dataset received due to the uniqueness of every data file, though the overall flow and some of the individual components can remain the same.
- The notebook opens the file, checks the header information, plots some of the variables, and calculates some simple statistics.
- It then runs through a series of actions to harmonize the time axis, variable names, and variable units, before copying the processed file to the TDS.

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

6. Data access

The THREDDS Data Server (TDS) permits access to all VERIFY harmonized data products via the "Searchable data repository" button on the official VERIFY website: https://verify.lsce.ipsl.fr/index.php/products. The content of this page is illustrated on Figure 4.

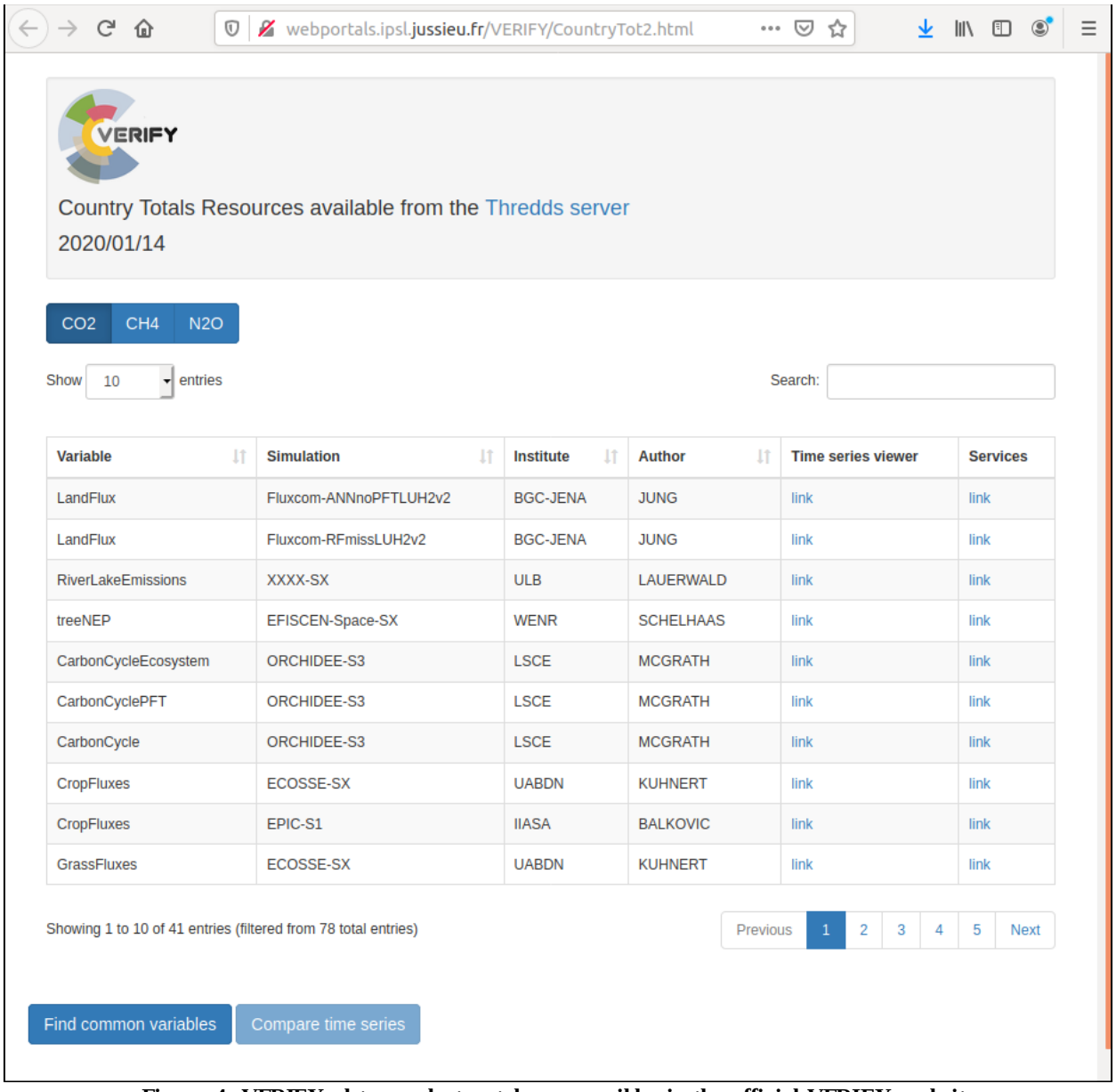

Figure 4: VERIFY data product catalog accessible via the official VERIFY website

Several facilities are proposed:

1. A direct access to the TDS web catalog by clicking on the link "Thredds server" at the top of the page, close to the VERIFY logo.

2. A catalog / table that uses metadata about all the files to allow them to be sorted based on the chemical species (CO2, CH4, or N20, via the buttons at the top), variable (a descriptor of the kinds of variables included in the file), simulation (describes the model or technique used), the institute that produced the dataset, and the researcher (author) the dataset is associated with. Several actions are possible:
    a. Access to the raw data themselves by selecting the link "Services" of product.
    b. Access to a specific "time series" visualization facility for selected products. Several products can be selected.

Direct access to all data on the TDS:
The TDS web catalog can be also access directly with the following link: https://verifydb.lsce.ipsl.fr/thredds/verify/catalog.html. It provides the raw version of the catalog, which is less presentable to the ordinary viewer than the catalog. However, it is convenient for expert users to navigate through the different sub-directories of the TDS in order to find all data files. Figure 5 below illustrates the main directories of the TDS server.

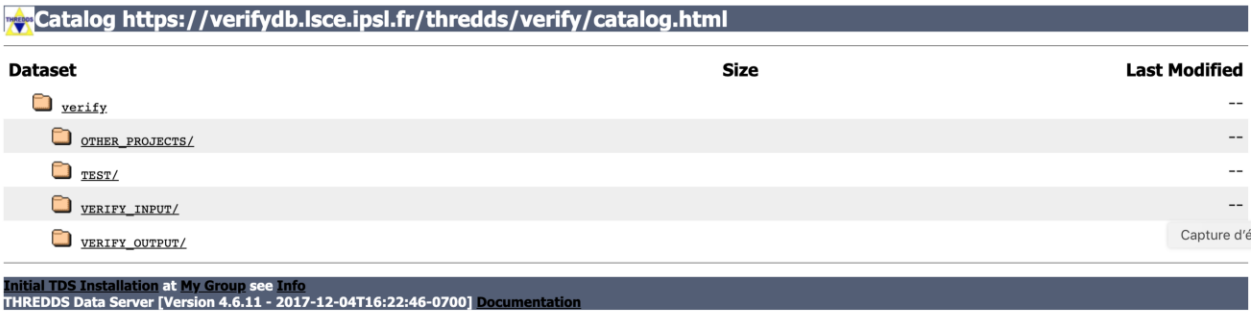

Figure 5: Illustration of the Thredds Data Server structure.

The TDS contains more information on each file than displayed in the catalog, including access to the processing scripts (Python scripts and Jupyter notebooks) used to process many of the VERIFY output fluxes. These scripts used to process a given FILENAME are contained in the FILENAME.process folder under the different VERIFY output sub-directories.

The TDS server contains different data sets including not only those produced by the VERIFY project but also numerous data sets gathered for the GHG flux synthesis of VERIFY. These data sets are grouped under the "OTHER_PROJECTS" sub-directory and their use follow specific data policy / data access.

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

The main VERIFY output data sets can be found under the "VERIFY_OUTPUT" directory and subsequent sub-directories (one per GHG). Different files can found corresponding to:

- Spatially explicit 2D NetCDF files containing the raw output data from model simulations defined over a global / regional spatial grids. These files end in _2D.nc.
- Country total time series created from the 2D files that correspond to the aggregated fluxes (annual or monthly time steps) per regions. The ensemble of regions includes all European countries and a pre-selected set of groups of countries. Two types of country total files were created depending on the masks that were used:
    - files ending with _CountryTotWithEEZ.nc, that correspond to country masks including the Extended Economic Zone (EEZ) for each country, i.e. the coastal ocean under that country's influence.
    - files ending with _CountryTotWithOutEEZ, that correspond to the country masks but without the EEZ.

These files are of less interest to the average user, but very useful to the advanced user, depending on the analysis, the type of GHG and the sector that is envisaged.

Access to country totals for each data product:
Given the main objectives of the VERIFY project, a specific emphasis is given to files containing the country total fluxes.
The link found in the "Services" column of the catalog (see Figure 4) enables the user to download all or part of the "country-total data", using the following techniques: HTTPS, Web Coverage Service (WCS), NetCDF Subset Service (NCSS), WMS, OpeNDAP, and DAP4. Figure 6 illustrates the proposed services associated with these country total files on the TDS.

Figure 6: Illustration of the proposed services associated with a "country total" file, accessible through the "Services" link on the data product catalog (see Figure 4)

7. Data visualization tools

This section describes the first version of an ensemble of tools to interactively visualize the data that are produced by the VERIFY project (including also non-VERIFY data). The objective is to

provide advance users with some tools to create specific plots for the analysis of the GHG fluxes produced in VERIFY. As detailed in section 4 (Setup) two major facilities are provided: i) a technical processing toolbox (based on notebooks) for expert users and ii) a user-friendly data visualization tool for advanced users. Currently the visualization tools consist of a time series viewer; next steps will include a "map visualization" facility.

Toolbox of data processing notebooks:
The page is accessible from the main "Products" page by selecting the "Technical processing notebooks" icon: http://verify.lsce.ipsl.fr/index.php/products
It displays an ensemble of notebooks (Jupyter notebooks) archived under a worldwide software development platform, GitHub. The GitHub repository contains currently (March 2020) a rather limited ensemble of notebooks to i) process the data (see for instance, section 5) and ii) plot CO2 fluxes.
Such advance user facilities will be enriched during the second half of the project with new plotting scripts that will be provided both by the coordinator (CEA-LSCE) and by all partners of VERIFY following specific investigation of all VERIFY results
The time series viewer link opens up a visualization tool described in more detail below.

Time series visualization tool:
The second visualization tool available to the user is the time series tool, reachable from two entry points on the "Products" page of the portal (see Figure 2 and Section 4): both the "Searchable data repository" and "Users-friendly data visualizations" icons. These two icons open up the same **catalog** with all data datasets. From this catalog the "Time series viewer" link (second-to-last column) opens up a window allowing the user to explore time series for different countries and regions in an interactive away.

Selecting products:
The first step consists of selecting an ensemble of products from the catalog. Standard selection mechanisms are available and the selected data sets are highlighted in blue.

Analysis of common variables:
Users can then check which variables are contained in the different data sets and which ones are common to all files, using the button labeled "Find common variables" below the catalog. After clicking this button, a list of variables appears at the bottom of the screen, either in green or in blue. Those variables in green are variables which appear in all the selected datasets. If variables appear in green, users can click the "Compare time series" button to display the time series for any of the common variables for any country/region with labels indicating which dataset they come from.

Inter-active time series plot:
Figure 7 provides an illustration of the interactive time series viewer. Several features are available to the user:

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

- A window in the upper right allows the user to select countries and regions to plot, which are then displayed in the bottom window.
- Multiple variables can be chosen, which will open more windows at the bottom: one variable per window, with all the variables in the file accessible.
- The viewing windows can be dynamically resized, zoomed in, and closed down according to the user's whims.
- The viewing windows can also be saved as a "png" file.
- Rolling means can be imposed to make the data more smooth (upper right box)
- A scale factors can be applied to change units (which are displayed in the title of each window along with the variable).

All features are "dynamic" and the plots are thus updated automatically depending on the selected options. Additional features for this "time series viewer" will be implemented in the second half of the project, following requests from the users (mainly the VERIFY partners).

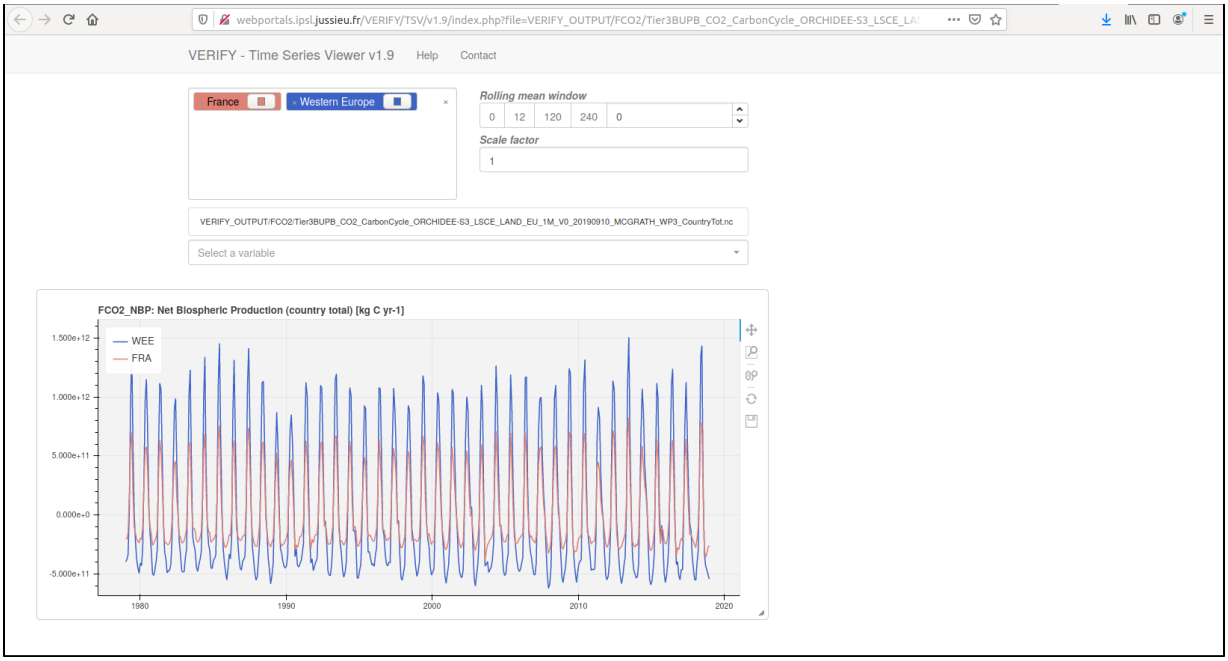

Figure 7: An example of the time series viewer available through the "Searchable data repository" button on the "Products" page on the VERIFY website

8. GHG synthesis and summary fact-sheets

This section describes a key feature of the VERIFY project, linked to the production of the GHG synthesis at country scale. Theses synthetic plots summarize the core work of several work packages, WP1-2-3-4 with a comparison of the UNFCCC flux estimates (WP1) to the observation-based fluxes derived in WP2-3-4.

These plots, also used in the annual scientific synthesis report, correspond to the "Country summary factsheets" icon of the main products page (see Figure 2) or can also be access directly with the link: http://webportals.ipsl.jussieu.fr/VERIFY/FactSheets/. The tool presents several options to choose and display an ensemble of "pre-made" plots. Figure 8 illustrates the main page for selecting the different synthetic plots.

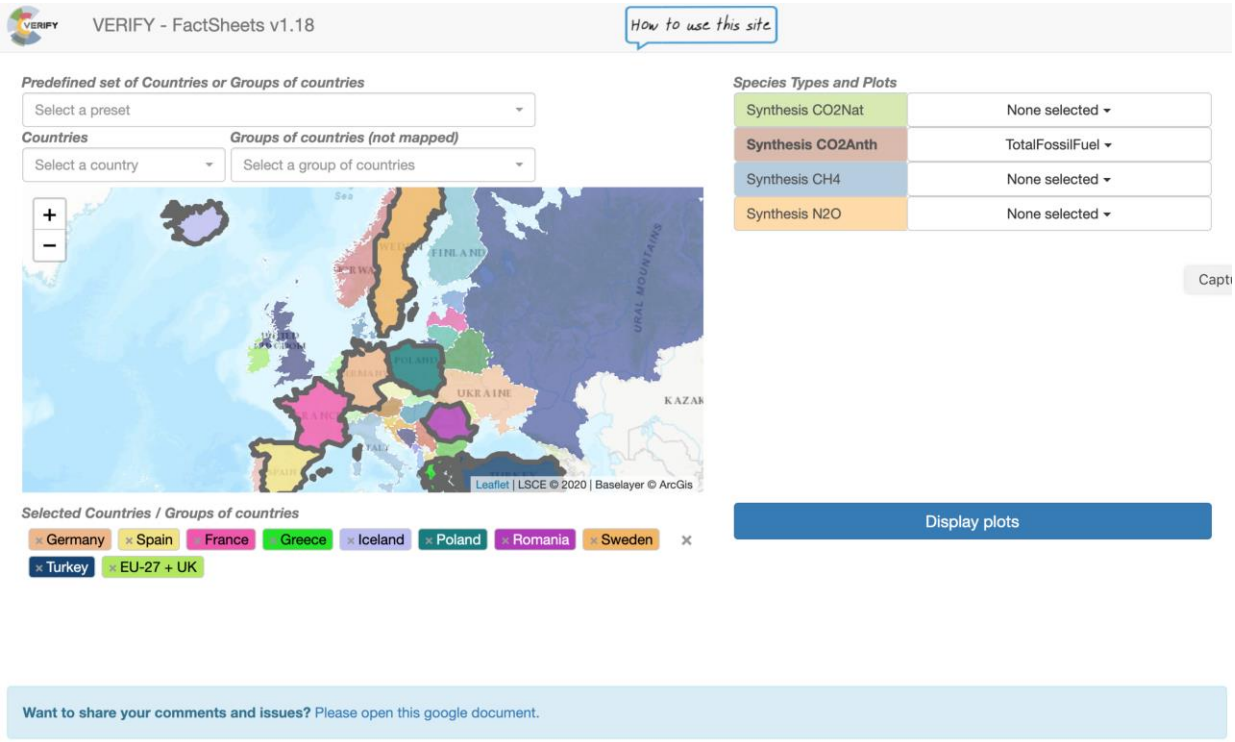

Figure 8: An example of the web tool that selects and displays synthetic plots of greenhouse gas fluxes.

The main page of this web-site offers the following features:

- A map with possibilities of selecting a country by clicking on it.
- Two additional drop-down menus, the top one to select a pre-defined set of regions and the bottom to select from a list different European countries or groups of countries.
- A menu on the top right to select the type of synthesis plots the user wishes to view: one for every species, except for CO_2, which is additionally broken down into "Natural" and "Anthropogenic" emissions. The different plots illustrate key concepts related to greenhouse gas fluxes, such as comparing bottom-up and top-down approaches, or comparing top-down approaches to national inventory methods.
- A button "Display plots" allow the user to view the selected synthesis plots.
- A specific button at the top of the page, "How to use this site", provides a short video tutorial on how to use the different features of the site
- A link at the bottom of the page opens a Google document to share comments and question around the different synthesis plots.

When the "Display plots" button is selected a new page opens to display all plots. Figure 9 provides an example for one region (EU27+UK) of the type of plots that can be obtained for each GHG. Each plot has several icons on the top in order to:

- View shrot description (caption) of each type of plot (clicking on the small "I" icon)
- View the legend that explains the acronyms used across all the plots (clicking on the small "?" icon)
- Access the raw data used to create the plot (clicking on the small "grid" icon)
- Close the specific plot (clicking on the small "X" icon)

Additionally, the plots can be rearranged, resized, and deleted in order to facilitate viewing. By "double clicking" on a plot you open a larger view of the selected plot.

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

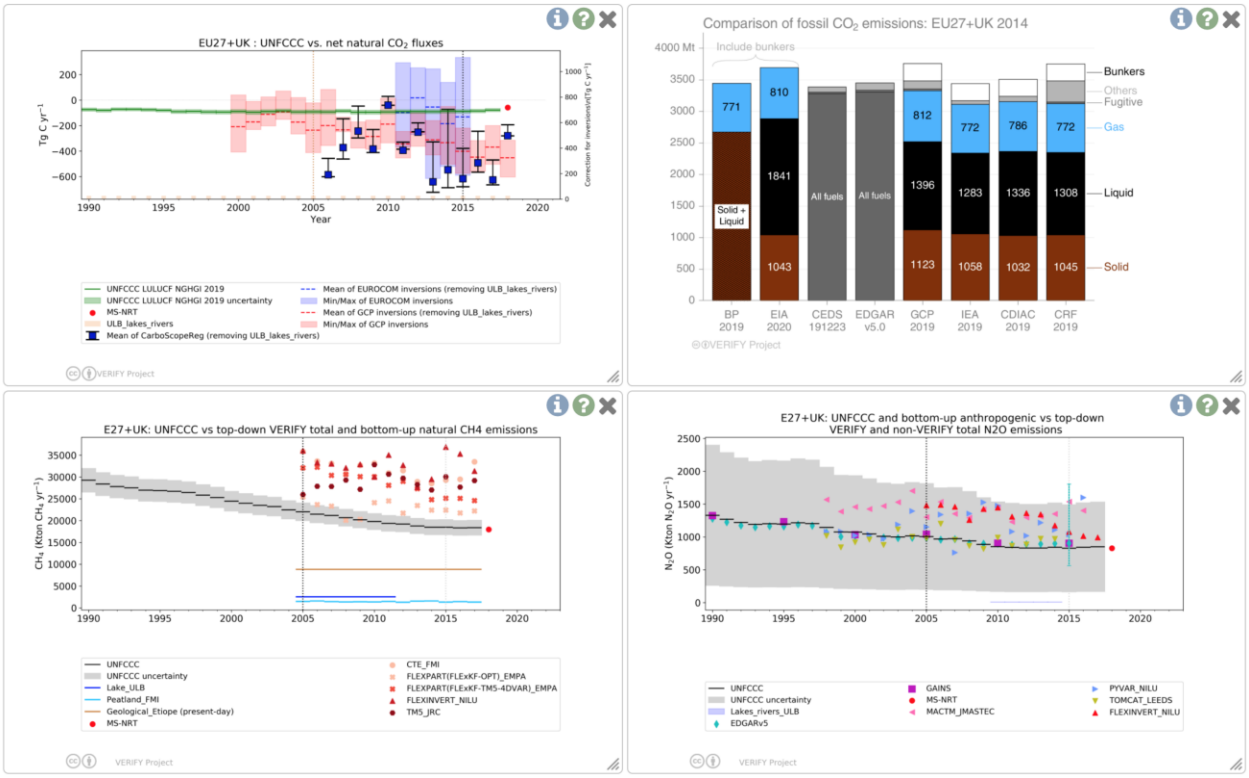

Figure 9: Illustration for one region (EU27 + UK) of typical synthesis plot for CO_2 natural (upper left), CO_2 anthropogenic (upper right), CH_4 (lower left) and N_2O (lower right). Each plot provides a synthesis combining UNFCCC flux estimates and other data-driven flux estimates.

9. Conclusion and perspectives

This deliverable provides a first description of the database and data management infrastructure, as implemented on the CEA-LSCE servers. The main infrastructure relies on a Thredds Data Server (TDS) and offers not only the access to the main data sets but also some visualization facilities. Our infrastructure permits us to meet most requirements of the FAIR policy with this first implementation, although additional capabilities will be added in order to be fully compliant with the FAIR policy:

- **Findable:** Our dataset is directly findable through a dedicated web portal that offers a catalog with searching capabilities. The next step is to provide a DOI (Digital Object Identifier) for all key data sets of VERIFY (i.e. the main flux estimates). These step will be done in collaboration with ICOS.

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.

- **Accessible:** All datasets are accessible through a standard protocol widely used to share data across the climate scientific community: a Thredds Data Server (TDS, managed by unidata) designed to work with the NetCDF data format of files and offering a variety of services (WCS, NCSS, WMS, etc).

- **Interoperable:** Most output from the VERIFY project tries to follow international standards associated with NetCDF files. The metadata output tries to follow the standard NetCDF Climate and Forecast (CF) Metadata convention, using the standard variable names, units, dimensions, axis, required 'coordinates' attribute, etc. Although not compliant yet the following updates will have the objective to be CF compliant.

- **Re-usable:** The main data sets that will be produced by the VERIFY project will be kept available and accessible after the duration of the project through other long-term data portals (such as the ICOS carbon portal).

It should be noted that some VERIFY products are password-protected to ensure that VERIFY research groups have the first attempt at using them. For the moment, access to all raw data is protected behind a username and password made available to all partners and some selected external folks. The visualization tools, on the other hand, are not password protected. Some technical challenges remain to installing password protection on the FactSheets, though we are pursuing this.

Updates will be provided in the next version of this deliverable with respect to key features linked to:

- The management of the data set access rights with variable options depending on the type of data.
- The addition of new user-friendly visualization tools, like mapping facilities to compare the spatial gradient of the GHG fluxes between data sets.
- The addition of specific "Jupyter notebooks", especially notebooks that can be used to create the synthesis plots.
- The transfer of the VERIFY data infrastructure to an existing long-term infrastructure such as the ICOS carbon portal.

VERIFY is a research project funded by the European Commission under the H2020 program. Grant Agreement number 776810.