Climate Modelling User Group

Deliverable 3.5

Technical note on
Status of ECMWF Climate Monitoring Database Facility

Centres providing input: ECMWF

Version nr.	Date	Status
0.1	31 Dec 2012	Initial Draft for internal ECMWF review
1.0	25 Jan 2013	Updates from internal review, submitted to ESA
1.1	03 April 2013	Incorporating comments from CMUG and ESA

Deliverable 3.5

Technical note

Status of ECMWF Climate Monitoring Database Facility

Table of Contents

Status of ECMWF Climate Monitoring Database Facility

1. Purpose and scope of this Technical Note

ECMWF has been developing an interactive environment to visualize and facilitate model-observation confrontation for Level-3 data products - with a focus on assessing low-frequency (multi-year) variability of statistical averages (typically monthly/regional means). We refer to the environment as the ECMWF Climate Monitoring (Database) Facility because activities to support climate monitoring -- including examination of CCI products -- are envisaged as a significant application, and a key element of the environment is a flexible relational database. The Facility does not remain static but rather changes as capabilities and needs evolve, and thus should be considered as work-in-progress.

This document describes the status of the Climate Monitoring Facility (CMF), and constitutes a first version of the CMF documentation. Section 2 gives some context for the Facility, explaining its relevance for examination of CCI products from the perspective of the reanalysis community (a subset of the wider community of observational dataset users). Section 3 describes the current status of the Facility, covering not just the database design and current contents, but also the associated user interfaces and implementation of visualization tools. Section 4 contains a summary and outline of future developments. CMUG use of the CMF for a pre-final version of the HadISST2 dataset (treated as a precursor for future CCI products) is described elsewhere (CMUG, 2012), and further CMF use for specific CCI-generated datasets is the subject of another document (CMUG, 2013).

2. Context of the Climate Monitoring Database Facility

Why is Model-Observation Confrontation important?

Advances in our understanding of the Earth's climate-system are to a large extent underpinned by Model-Observation Confrontation. On the one hand, confronting Models with Observations provides an important mechanism for developing/improving the conceptual and computational aspects of such models. Conversely, confronting Observations with Models is increasingly a key component of establishing the quality of the observations. It is worth noting that the observational datasets used for climate purposes, and especially those derived from satellite-based remote-sensing instruments, are typically the result of retrieval models, i.e. computational schemes that depend, implicitly and/or explicitly, on conceptual models of the physical processes involved. There is a growing awareness that Model-Observation Confrontation is effective in establishing the robustness of the underlying retrieval models and hence the quality of the resulting observational datasets. The level of robustness in turn influences the level of confidence that data users have in a particular dataset, and hence their willingness to use the dataset in their applications.

Model-Observation Confrontation assists in anticipating the adverse impacts arising from inconsistency and inhomogeneity between/within datasets, and thus is expected to play a significant role in the decision process that determines whether a dataset is deemed suitable

(from the user's perspective) for other important applications (providing Boundary Conditions, providing Initial Conditions, and providing Assimilable Observations). CMUG therefore sees Model-Observation Confrontation as a key component of the CCI's efforts to develop climate-quality ECV Products.

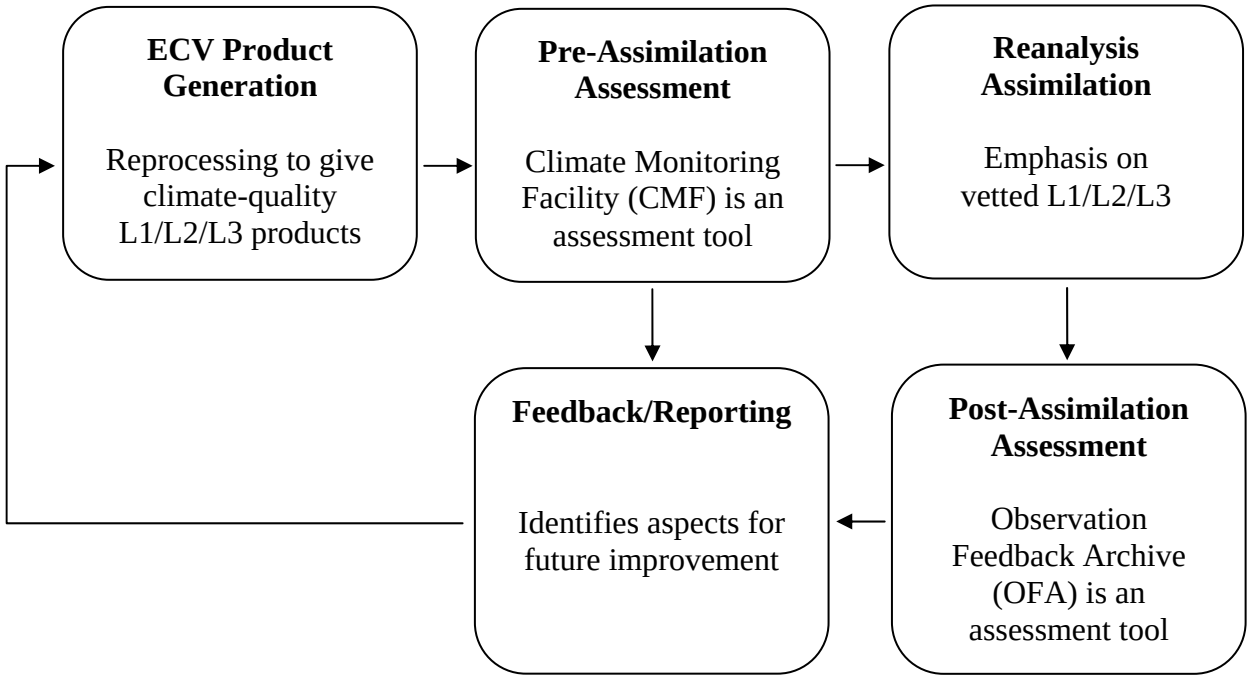

Figure 1: ECV Product Development viewed as an iterative process of Generation-Assessment-Feedback. In this example the Assessment steps involve 3 ECMWF systems, namely the Climate Monitoring Facility (CMF), the Reanalysis Assimilation System, and the Observation Feedback Archive (OFA).

Why does the CMF link ECV Product Generation and Reanalysis Applications?

Figure 1 shows ECV Product Development viewed as an iterative process of Generation-Assessment-Feedback incorporating Reanalysis Assimilation as one (but not the only) application. In this context, the ECMWF Climate Monitoring Facility (CMF) plays a prominent role in assessing ECV products prior to reanalysis assimilation applications, motivating the Facility documentation provided in this Technical Note. Use of ECV products by the Reanalysis Assimilation System, and the product feedback available from ECMWF's Observation Feedback Archive (OFA), are beyond the scope of the current document and are described elsewhere.

For the reanalysis community, consistency and homogeneity of Level-1 and Level-2 products are of prime importance for such products to warrant serious consideration as assimilation input. From the perspective of the reanalysis community, Level-3 products derived from

Level-1 and Level-2 represent useful downstream diagnostics facilitating a partial assessment of consistency/homogeneity and contributing to the vetting of Level-1 and Level-2. The CMF provides an environment to conduct such an assessment and thus is an important link between ECV product generation and reanalysis applications.

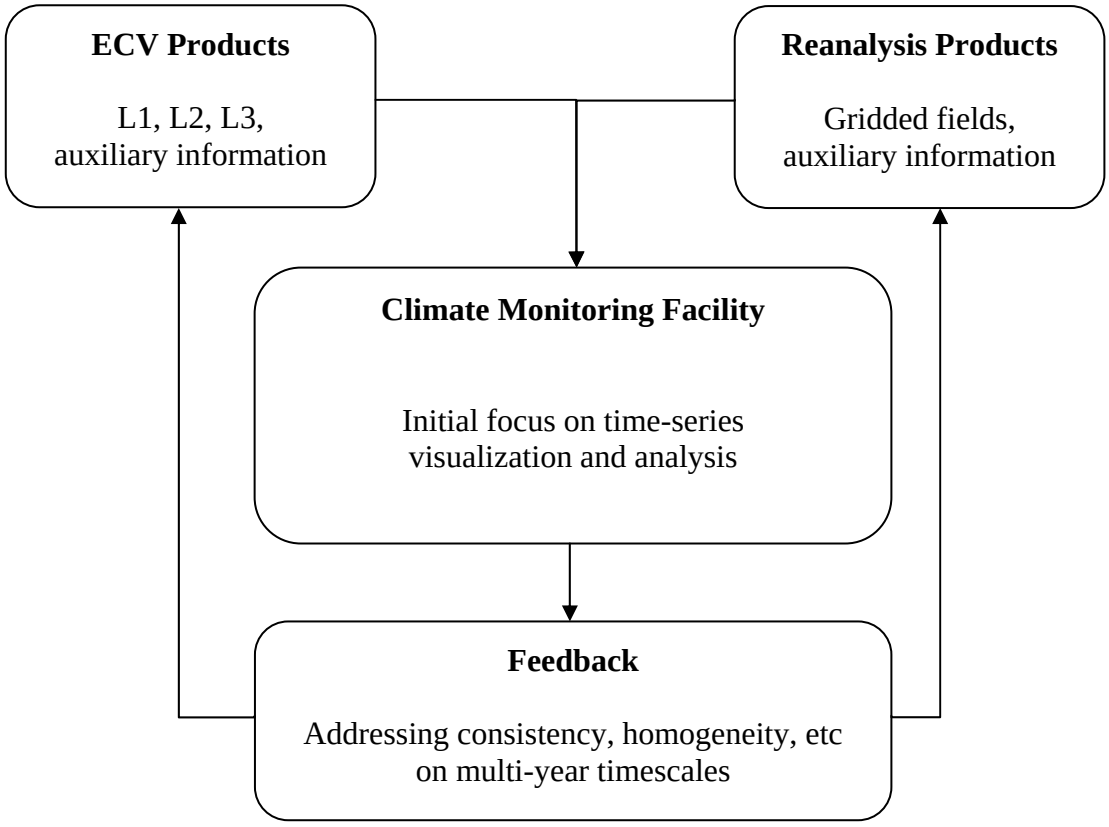

Figure 2: Model-Observation Confrontation: Role of the ECMWF Climate Monitoring Facility in Pre-Assimilation Assessment of ECV Products. The time-series analysis is primarily based on confronting regional means of Level-3 ECV products with related means of Reanalysis fields. Nonetheless, the Feedback may actually address lack of consistency and homogeneity in upstream quantities such as Level-2 ECV products and Reanalysis inputs.

Figure 2 elaborates on the use of the Climate Monitoring Facility to conduct pre-assimilation assessment of ECV products. An important emphasis is to enable assessment of consistency and homogeneity amongst various ECV products. From the reanalysis perspective, the requirements for consistency and homogeneity span different product levels, different ECVs, and a wide range of spatial and temporal scales. The requirements are increasingly driven by efforts to understand the Earth's climate as an integrated system incorporating coupling between dynamical, chemical and biological processes.

At the time of writing, the initial focus has been directed at consistency and homogeneity over multi-decadal timescales. This has guided the CMF development to have an emphasis on time-series visualization and analysis. Given the wide-ranging and exploratory nature of dataset assessment, it was recognized that flexibility and interactivity would be important features of the CMF.

To summarize, ECMWF's development of a Climate Monitoring Facility is motivated by the need for an interactive environment to visualize and facilitate model-observation confrontation for Level-3 data products. It permits activities to support climate monitoring including examination of CCI products, with an initial emphasis on assessing consistency and homogeneity of ECV products spanning multi-decadal timescales. The need for flexibility led to the decision to build the Facility in the form of an environment based on a relational database with supporting tools. The Facility does not remain static but rather changes as capabilities and needs evolve, and thus should be considered as work-in-progress. More detailed technical descriptions are given in the remainder of this document.

3. Currrent Status of the CMF

3.1 Overview of implementation

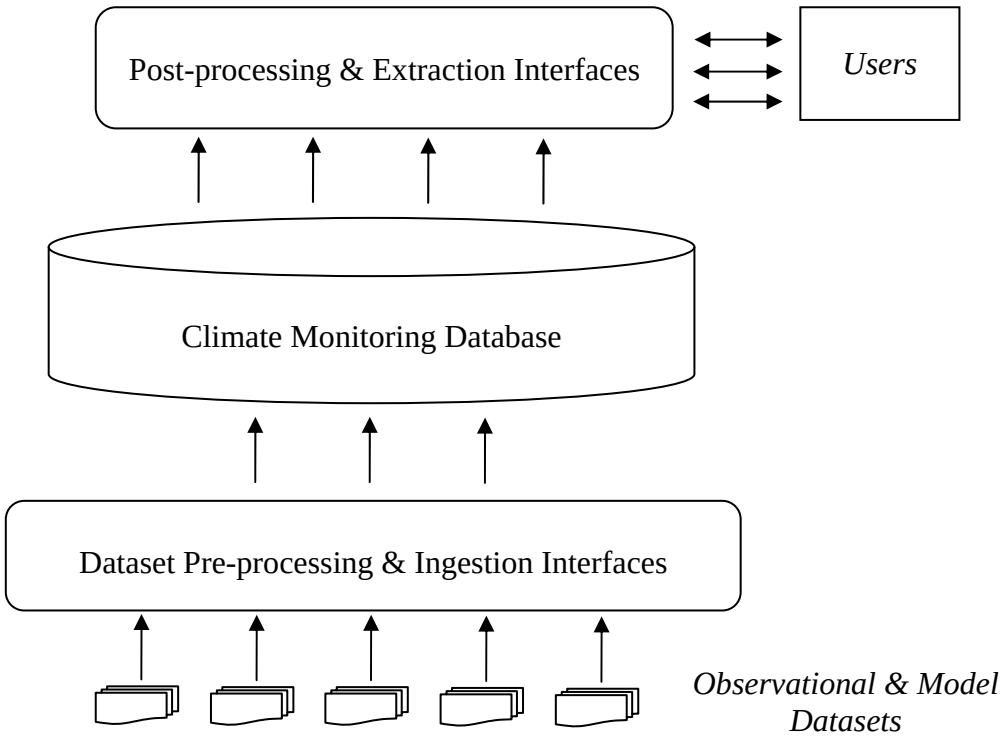

Figure 3: Schematic overview of the main components of the Climate Monitoring Facility

Figure 3 shows the three main components of the Climate Monitoring Facility. From the perspective of data flow, the components are:

- The Dataset Pre-processing & Ingestion Interfaces. This component of the facility is responsible for depositing dataset time-series into the database. The observational and model datasets of interest span a wide diversity in terms of content and format, so pre-processing and ingestion comprise an array of tools customized for specific datasets as described below.
- The Climate Monitoring Database. This component is the relational database designed to hold a wide range of data relevant for model-observation confrontation. The database is flexible and concisely described by a Schema, given below.
- The Post-processing & Extraction Interfaces. This component comprises the User interfaces that permit data extraction, manipulation and visualization, described in more detail below.

This document now describes the three components in separate sub-sections. To make the description more relevant to Users, we dispense with the data flow sequence and describe next the Database Schema and Post-processing/Extraction Interfaces. Pre-processing and Ingestion Interfaces represent a substantial technical development effort, but these are relatively more remote from the User and so their description is given last.

We note in passing that the components are implemented in a modular way and distributed across different ECMWF computing platforms. This imparts flexibility for different aspects to evolve in a self-contained manner, as well as the potential for interfacing to other applications and implementations in future.

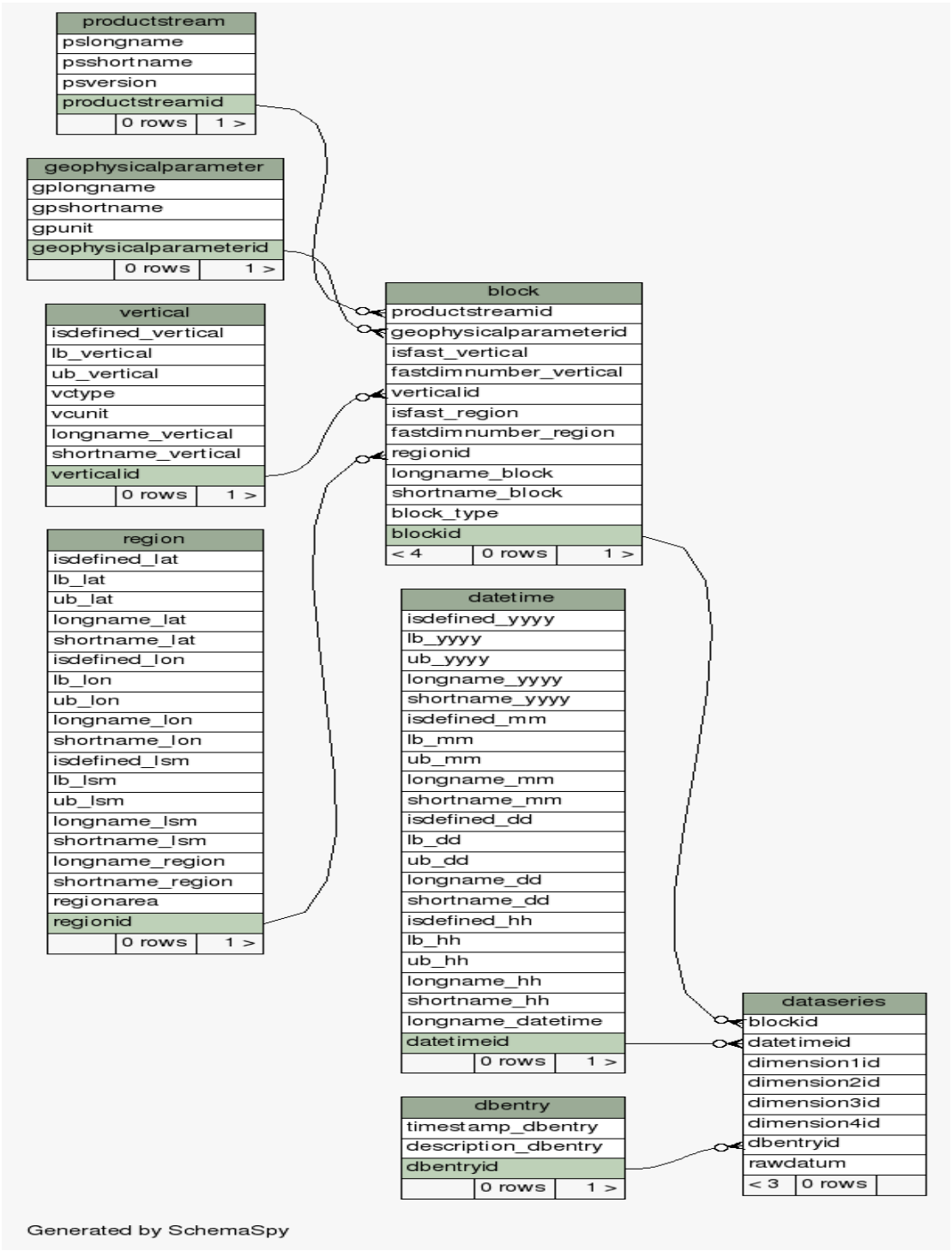

Figure 4: Current implementation of the CMF Database Schema. A time-series corresponds to all instances of RawDatum sharing a common BlockId but different values for DateTimeId

3.2 Climate Monitoring Database: Schema Definition and Time-series Representation

Figure 4 shows the current implementation of the schema that defines the structure of the Climate Monitoring Database. Within the DataSeries table, the rawdatum element stores the numerical values from the ingested datasets, irrespective of whether they come from models or observations. The remaining elements and tables store auxilliary information about the attributes of the dataset values, to facilitate logical grouping of the numerical values and efficient storage/access.

For example, the monthly mean for June 1979 of 2m Temperature over Land from ERA-Interim corresponds in the database to the following attributes:

> gpshortname="T2m", shortname_region="Land", psshortname="ERA-Int", lb_yyyy=1979, ub_yyyy=1979, lb_mm=6, ub_mm=6

The analogous monthly mean for July 1979 changes only the lb_mm and ub_mm attributes, while the mean for June 1980 changes only the lb_yyyy and ub_yyyy attributes.

Different year-month combinations are recognized within the database by unique values of the DateTimeId element (key). Similarly, different combinations of geophysical parameter, region, and product stream and vertical coordinate are recognized by unique values of the BlockId element (key).

The design is such that a single time-series corresponds to a convenient subset of the DataSeries table, given by a unique BlockId value but spanning a range of DateTimeId values. All the attributes of the time-series are identified by the unique BlockId value, apart from variation in the time-dimension which is identified by different entries in the DateTime table. Continuing the previous example, the time-series of 2m Temperature over Land from ERA-Interim is identified by a specific value of the BlockId but this value is usually not explicitly referred to by the User. Instead the User will typically access the time-series by specifying (through the pull-down menus described later) the following attributes:

> gpshortname="T2m", shortname_region="Land", psshortname="ERA-Int"

For the sake of clarity, a Table summarizing the current contents of the database is postponed until Section 3.4; this being the natural place to describe the contents in terms of their dataset sources.

3.3 Tools for Extraction, Visualization and Time-series Analysis

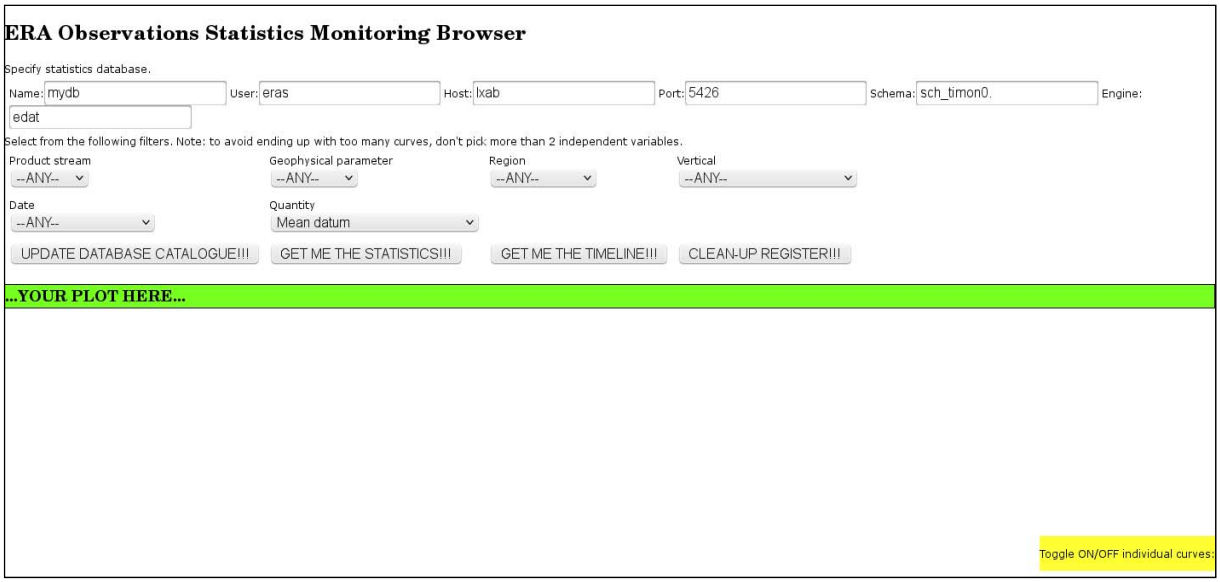

Figure 5: The User's Web-browser Interface to the Climate Monitoring Facility on startup

The CMF's main tools for extraction, visualization and time-series analysis are presented to the User through a Web-browser interface (see Figure 5 for the startup display). Via a webpage written in HTML, the interface presents the user with a combination of pull-down menus and buttons to control data selection, extraction, analysis and plotting. The corresponding actions are implemented using a mixture of JavaScript® (https://developer.mozilla.org/en-US/docs/JavaScript) and Python (http://www.python.org/), together with the Matplotlib plotting library (http://matplotlib.org/).

The Primary Functionalities are:

> Data selection: this is controlled via 6 pull-down menus through which the user chooses the data to be retrieved from the database. The pull-down menus reflect the structure of the database, providing 4 independent options for the elements defining the database Block (i.e. the ProductStream, GeophysicalParameter, Region, and Vertical definitions), and also options for the Date range definition and the statistical Quantity to be computed/plotted.
>
> "Get Me The Statistics": clicking on this button triggers extraction of the selected data and a preliminary plot. The selection is translated into a database query which is sent to the database server. The data are returned by the database server to the HTML layer and added incrementally to a Data Register. The HTML layer sends the data to a Plotting Engine and renders the resulting plot on the screen. It also provides an option to save the returned data in the form of a JSON file (see Supporting Functionalities).

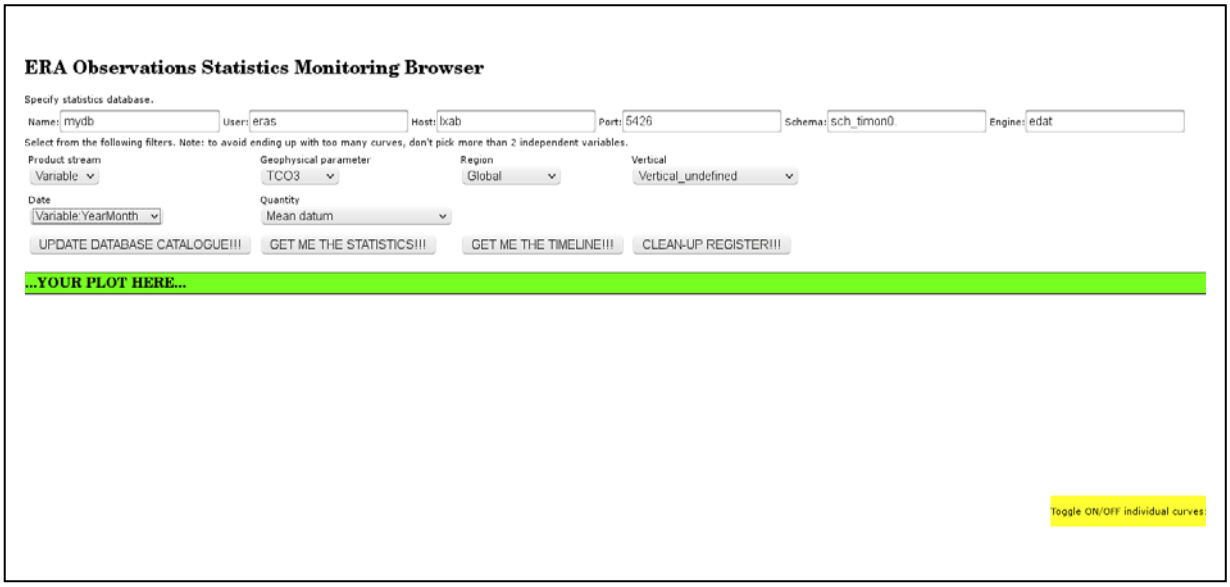

Figure 6: Data Selection for Total Column Ozone, Global Average and Monthly Means. Clicking on "Get Me The Statistics" triggers retrieval and plotting (see next Figure)

As an illustrative example, Figure 6 shows the settings to select Total Column Ozone, Global Average, Monthly Means, from all Products available in the Database, prior to data extraction and plotting. The "Geophysical parameter" uses the abbreviated name "TCO3" and the "Region" is "Global". Because Total Column Ozone is a vertically integrated quantity, the "Vertical" setting is "Vertical_undefined", in contrast to other options such as "pressure=100hPa". The settings for "Product stream" and "Date" both include the term "Variable". This indicates to the Selection interface that all matching instances should be returned as individual values without being averaged together - this is the mechanism that makes the values available for grouping into different time-series.

Supporting Actions and Functionalities (see Figure 7) include:

"Download JSON output": clicking on this button triggers dialog boxes enabling the user to save the data returned by the "Get Me The Statistics" action in the form of a JSON file.

Plot customization: a series of check-boxes (on/off toggles) enable the user to determine whether each individual time-series is displayed or not.

"Update Database Catalog": clicking on this button triggers an update of the Data Selection pull-down menus. The menu updates are driven by a query of the database to determine the currently available parameters. Ingestion tasks are typically executing while users are accessing the CMF, so Catalog Updating provides a mechanism for

having up-to-date access to any new parameters that have been added since the last update of the pull-down menus.

"Clean Up Register": clicking on this button clears the Data Register so that the next rendering by the Plotting Engine begins with an empty graph. When this action is not invoked, successive requests via "Get Me The Statistics" result in over-plotting on the same graph. This permits visualization of multiple ECVs and facilitates assessment of cross-ECV consistency.

The Plotting Engine: takes data from the Data Register and plots it. The plotting engine renders all register data as multiple curves in a single graph. Until invocation of the "Clean Up Register" action, successive requests via "Get Me The Statistics" result in over-plotting on the same graph. This permits visualization of multiple ECVs and facilitates assessment of cross-ECV consistency.

The options to "Download JSON output" and the toggles (check-boxes) for plotting individual time-series are only visible after "Get Me The Statistics" has been invoked, as depicted in Figure 7 for the Data Selection of Figure 6. In the current implementation, plot attributes such as the axis ranges, line styles and colours, are all selected automatically by the Plotting Engine.

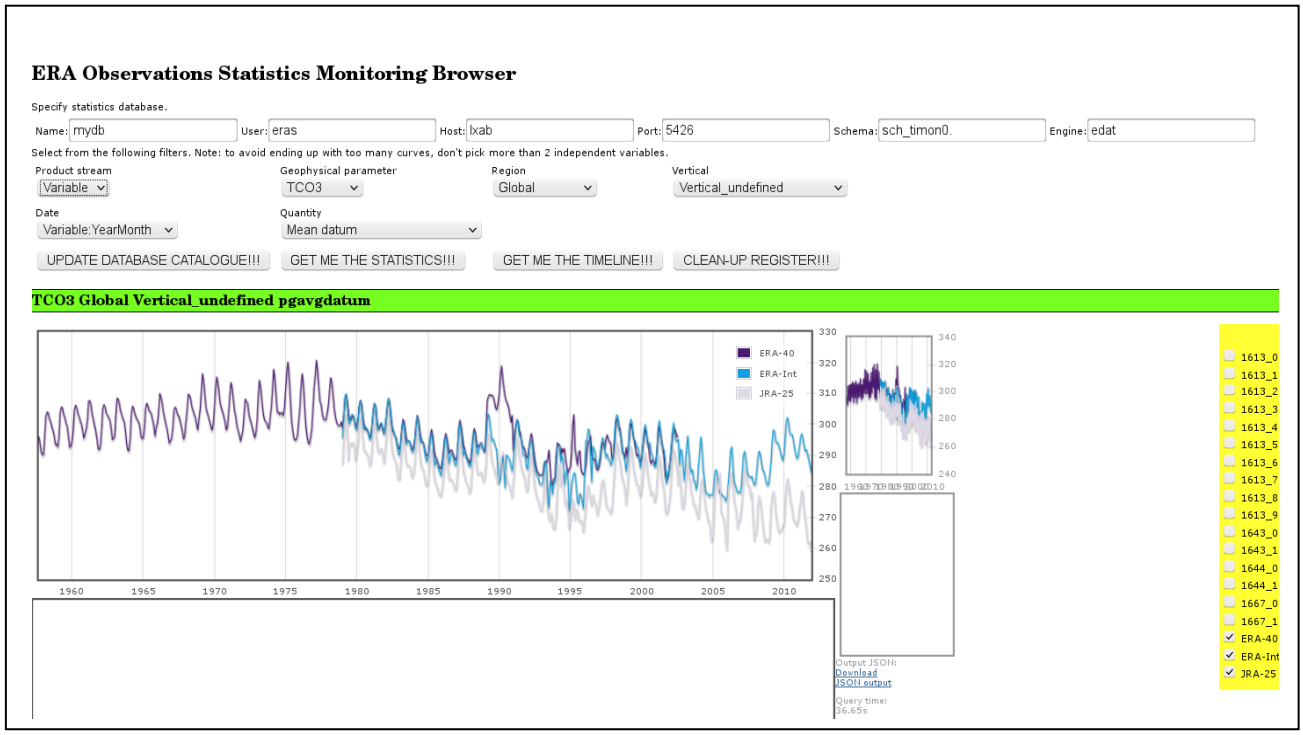

Figure 7: Result of invoking "Get Me The Statistics" for the data selection from the previous Figure. Toggles presented by the User Interface have been set to display only the time-series from ERA-40, ERA-Interim and JRA-25.

The focus of the tool is for time-series analysis, although the design of the database means that maps or other spatial representations offer potential for future development. In the future it is anticipated that a set of complementary tools will be developed, some of which will be suited to maps.

The main focus of the Facility is to address homogeneity and consistency, and the tool allows consistency to be examined through comparison of timeseries from different sources and production processes. Both for a common variable but also amongst variables correlated through physical processes. The assessment of quantitative uncertainty information through intercomparison of the timeseries and their differences provides a means for assessing/validating provided uncertainty estimates.

3.4 Tools for Dataset Pre-processing and Ingestion

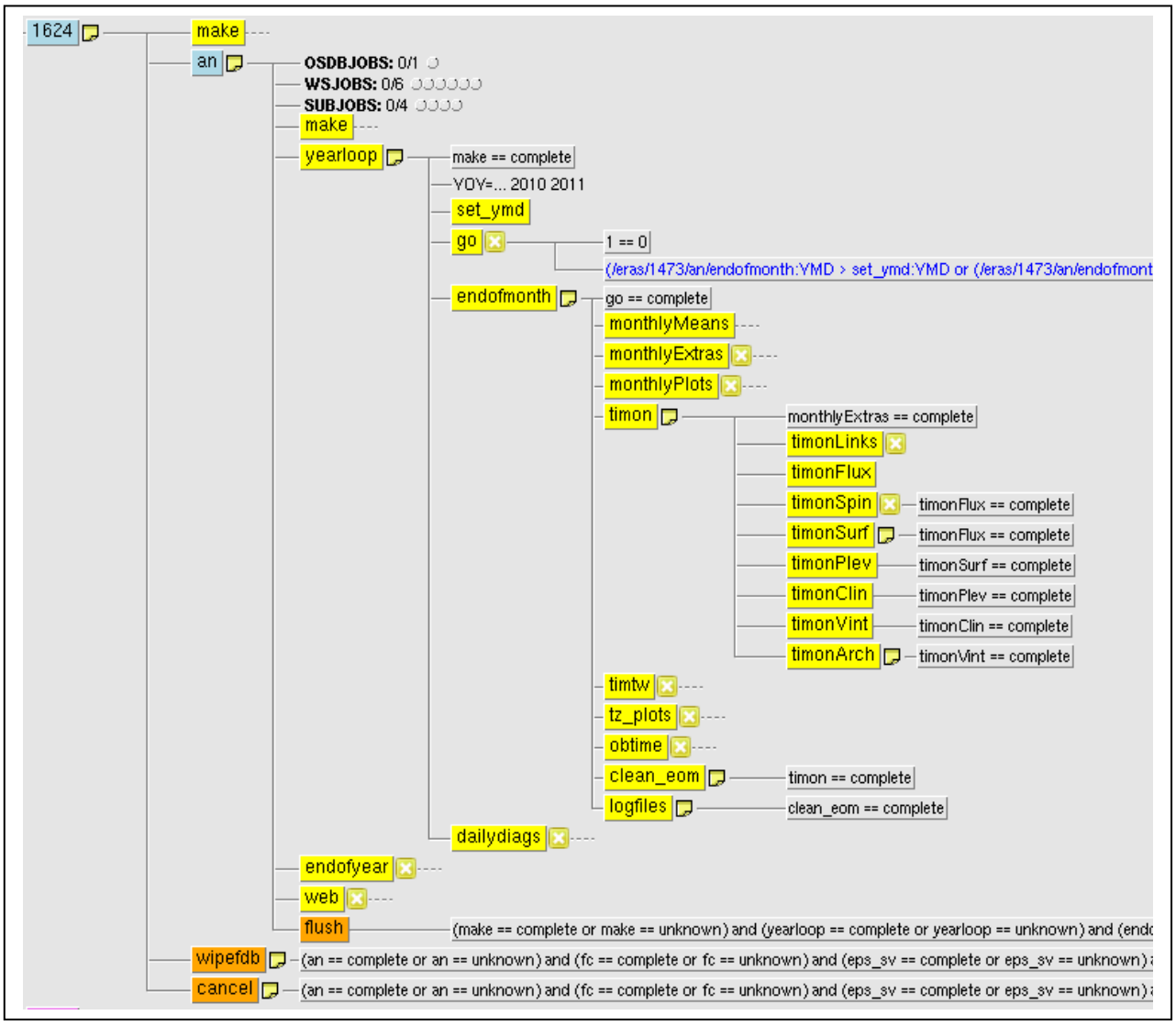

Figure 8: Ingestion suite for ERA-Interim, ERA-40, JRA-25 and NRA-2

The tools required for Dataset Pre-processing are specific to the dataset of interest.

A number of standard reanalysis datasets (ERA-Interim, ERA-40, JRA-25 and NRA-2) shared sufficient commonality that their pre-processing and ingestion could be accomplished by common tasks:

1. computation of regional means from monthly-mean gridded fields (Fortran code with input in GRIB format and output in simple readable text format)
2. addition of database ingestion directives (Python scripts producing intermediate file in JSON format)
3. ingestion of the intermediate JSON file by the database (more Python scripts)

These tasks are run in the ECMWF SMS environment and are shown in Figure 8 as the "timon" family of tasks. The higher levels of the hierarchy control execution - in this case in Backlog mode, looping over the years from 1957 (the start of the ERA-40 dataset) up to 2011, with the capability to extend to 2012 and beyond.

It is worth noting that, even for these four standard reanalysis datasets, a preparatory step of obtaining GRIB-format monthly means from daily/sub-daily gridded fields required dataset-specific customization. In the case of ERA-Interim and ERA-40, Fortran code with input and output in GRIB format was used in upstream SMS suites. For NRA-2, GRIB-format monthly means were computed by the data producers so these were transferred to ECMWF. For JRA-25, monthly means were computed by the data producers but in a different binary format, so these were transferred to ECMWF and converted to GRIB. For reasons of efficiency and practicality, it is important to limit the number of input formats for database ingestion. So we are conscious of the need to arrive at more standardization of product formats, especially for observational datasets.

For other datasets such as HadISST2, computation of monthly and regional means had already been performed in the frame of the ERA-CLIM project, but with a different output format. Consequently, some customization of the Python scripts was needed to achieve the addition of database ingestion directives and subsequent ingestion.

For model datasets from other projects, namely ERA-CLIM and CERA, other variations to the reanalysis ingestion suite were required - in particular treatment of multiple ensemble members. Because the timelines for generating these datasets was different, it was natural to have a separate ingestion suite. In this instance, Backlog mode effectively became Tracking mode with a yearly update frequency, i.e. ingestion of each year as soon as it is produced. For future datasets such as ERA-SAT, we anticipate Tracking mode with a monthly update frequency.

Dataset	SMS Ingestion	Time Period ingested	Comments

	Suite	into CMF	
Reanalyses			
ERA-40	1624	1957 -- 2001	
ERA-Interim	1624	1979 -- 2011	2012 -- present pending
JRA-25	1624	1979 -- 2011	
NRA-2	1624	1979 -- 2011	
MACC-II	Pending		For atmospheric composition (initially aerosol, later greenhouse gases and ozone)
Model datasets			
E20CM	1613	1899 -- 2011	The ERA-CLIM AMIP (10-member ensemble, IFS Cycle Cy37r3)
CERA prototypes	1667	1899 -- 2009	2-member ensembles, experimental, may be superseded/discontinued
ERA-SAT	Pending	Pending	Replacement for ERA-Interim. Ingestion via Monthly Tracking mode anticipated
Observational datasets			
p-HadISST2	Offline	1899 -- 2011	Pre-final version (received December 2011) in earlier version of database.
HadISST2	Offline	1899 -- 2007/2011	Ingestion of version received October 2012 pending.

Table 1: Datasets ingested into the CMF, current and planned. Other observational datasets of sufficient quality/duration will be accepted after vetting.

As promised in Section 3.2, Table 1 summarizes the datasets already ingested into the CMF, along with others that are planned. Note that the shortest time-period is currently 1979 to 2011, i.e. 33 years. Datasets of duration shorter than 5 years will have only limited value in this context.

4. Summary and Future Developments

Summary

Consistent with its involvement in CMUG and the CCI, ECMWF has been developing an interactive environment to visualize and facilitate model-observation confrontation for Level-3 data products - with a focus on assessing low-frequency (multi-year) variability of statistical averages (typically monthly/regional means). The environment is referred to as the ECMWF Climate Monitoring (Database) Facility because activities to support climate monitoring, including examination of CCI products, are envisaged as a significant application, and a key element of the environment is a flexible relational database. In addition, examination of Level-3 products is an integral part of vetting their upstream Level-1/2 products as potential reanalysis input.

This document has described the status of the Climate Monitoring Facility (CMF), and constitutes a first version of the CMF documentation. The context for the Facility, given in Section 2, explains its relevance for examination of CCI products from the perspective of the reanalysis community (a subset of the wider community of observational dataset users). The current status of the Facility is described in Section 3, covering not just the database design and current contents, but also the associated user interfaces and implementation of visualization tools. CMUG use of the CMF for a pre-final version of the HadISST2 dataset (treated as a precursor for future CCI products) is described elsewhere (CMUG, 2012), and further CMF use for specific CCI-generated datasets is the subject of another document (CMUG, 2013).

The Climate Monitoring Facility does not remain static but rather changes as capabilities and needs evolve, and thus should be considered as work-in-progress. We conclude this Section with an outline of future development.

Future development of the Climate Monitoring Facility

The future development of the Climate Monitoring Facility is guided by the intention to make the Facility available to a wide range of users, noting that such availability is dependent on an Operational framework that includes adequate user and technical support.

Current development is conducted within a Research & Development framework. Keeping in mind the long-term development, a number of aspects have been identified as priorities for the short-/medium-term and are given in Table 2.

Development item	Rationale	Nature of the development work	Further comments
Ingestion of new Reanalysis product streams and parameters	Enable more comprehensive model-observation confrontation	Extension of the tools for dataset pre-processing and ingestion	Priority product streams include MACC-II (for Aerosol and other atmospheric composition ECVs), priority parameters

			include soil moisture (from ERA-Interim initially)
Evolution of schema	Improve internal representation and user post-processing of ensembles	Impacts all aspects from pre-processing and ingestion to the user interfaces	
Functionality for User-upload	Permit ECV product generators to make rapid self-evaluations	Definition and handling of a suitable data format	Self-evaluations are an essential precursor to product vetting & ingestion by CMUG/ECMWF
Co-ordination with other developments	Efficient use of developments in other projects	Evolution of the Facility as required to make best use of developments in other projects	Interfacing to GUIs with a wider range of functionalities would save on in-house development
Beta-testing	Pre-cursor to operational roll-out	Training and support for a restricted set of users	Based on ECMWF and CMUG colleagues in the first instance.

Table 2: Future Development of the Climate Monitoring Facility

5. References

CMUG, 2012: CMUG ECV Quality Assessment Report, Technical Note/Deliverable 3.1_A, version 1.2, August 2012.

CMUG, 2013: Demonstration of CMF Functionality for the Assessment of Ozone, Aerosol and Soil Moisture. Technical Note/Deliverable 3.1_1D, due June 2013.

6. Abbreviations

CCI	Climate Change Initiative
CERA	Coupled ECMWF Reanalysis
CMF	Climate Monitoring Facility
CMUG	Climate Modelling Users Group
DOE	Department of Energy
ECMWF	European Centre for Medium-Range Weather Forecasting
ECV	Essential Climate Variable
ERA-40	The ECMWF 40-year Reanalysis
ERA-Interim	The ECMWF Interim Reanalysis
IFS	ECMWF's Integrated Forecast System

JMA	Japanese Meteorological Agency
JRA-25	The JMA 25-year Reanalysis
L1/L2/L3	Level-1/2/3
MACC-II	Monitoring Atmospheric Composition and Climate - Interim Implementation
NCEP	National Center for Environmental Prediction
NRA-2	NCEP-DOE Reanalysis 2
OFA	Observation Feedback Archive
SMS	Scheduling and Monitoring System