NILU: TR 7/2006

NILU: TR 7/2006
REFERENCE: U-103032
DATE: OCTOBER 2006
ISBN: 82-425-1798-3

Report on the RETRO database for observations and recommendations for optimal use of the data (D2-2)

Aasmund Fahre Vik, Sjur Bjørndalsæter, Leif Backman, Johannes Staehelin, Folkard Wittrock

REanalysis of the TROpospheric chemical composition over the past 40 years
A long-term global modeling study of tropospheric chemistry funded under the 5th EU framework programme
EU-Contract No. EVK2-CT-2002-00170

NILU: TR 7/2006
REFERENCE: U-103032
DATE: OCTOBER 2006
ISBN: 82-425-1798-3

Report on the RETRO database for observations and recommendations for optimal use of the data (D2-2)

Aasmund Fahre Vik[1], Sjur Bjørndalsæter[1], Leif Backman[2], Johannes Staehelin[3], Folkard Wittrock[4]

1 Norwegian Institute for Air Research, NILU, P.O. Box 100, NO-2027 Kjeller, Norway
2 Finnish Meteorological Institute, FMI, P.O. Box 503, FI-00101 Helsinki, Finland
3 ETH Zürich, HG, Rämistrasse 101, CH-8092 Zürich, Switzerland
4 Institute of Environmental Physics, University of Bremen, Otto-Hahn-Allee 1, DE-28359 Bremen, Germany

REanalysis of the TROpospheric chemical composition over the past 40 years
A long-term global modeling study of tropospheric chemistry funded under the 5th EU framework programme
EU-Contract No. EVK2-CT-2002-00170

Contents

Report on the RETRO database for observations and recommendations for optimal use of the data (D2-2)

1 Introduction

The main objectives of the use and collection of observations are twofold within RETRO. Firstly, the observations were expected to be used to evaluate models and to define the stratospheric boundary conditions. Secondly, it was planned to calculate long-term trends directly from observations. For these two purposes, measurements from satellites, aircrafts, ozone sondes and the surface were used.

A preliminary web-portal, providing observational data for RETRO, was set up at NILU during the first year of the project. This is no longer maintained, but is still available at http://nadir.nilu.no/retro_aasmund. This site was both a feasibility study for testing out methods of providing data and also the first delivery of data in an easy way. Different web-designs were tested and ozone sondes from the NADIR data centre were made available through password-protected pages – all in accordance with the rules for data usage and project policies. The data were not processed nor the data format adapted to RETRO needs at this stage. Original data files were simply made accessible through a web interface. This had not been done for these data before. The data or data service were not directly applicable to the RETRO consortium and the need for conversion of data into a more structured format, preferably netCDF, was discussed at an early project meeting. The existing service has, however, been discovered by external scientists, and NILU has received a few emails from non-European students that have shown an interest in the product.

The next step for the database was to make it more useful for RETRO and to implement functionality to store data in a more structured and searchable way. A new database was therefore set up and made operational through http://nadir.nilu.no/retro. The database is not only able to store observations in a structured way, but also model-output, emission data etc. As long as everything follows the rules for how data are to be stored, it is possible to treat these different data sets in an integrated manner. In essence, they all have to follow a common set of metadata guidelines. The basic concept and functionality of the system is described in the following chapter. A description of the observational data currently residing in the database is provided in chapter 3 while references to other observational collections are given in chapter 4. The 5th chapter is dedicated to future strategies for observations and further developments of operational data centres, while some final recommendations are given in chapter 6.

2 Technical description of the RETRO database of observations

In order to avoid redoing and rethinking through issues that have been solved before, it was decided that the RETRO data on observations should use the ESA ENVISAT Cal/Val metadata guidelines. In this way, it was also possible to reuse the Cal/Val database technology for the RETRO project. The ENVISAT Cal/Val database was developed and implemented at NILU for archiving of correlative data during the calibration and validation effort for the ENVISAT instruments AATSR, MERIS, GOMOS, SCIAMACHY and MIPAS. In order to address the various needs of the project, a clone of the original system had to undergo several changes and matured through continuous developments into a RETRO database. The system is therefore not only able to store data, but is also able to export metadata and large amounts of data in a structured way. This was to aid the modellers with the validation of their simulations. In addition, a graphical user interface was developed to visualise data location, thus making it easier to find data from specific areas.

The remaining parts of this chapter is dedicated to a detailed description on how metadata is defined in the project, how the database architecture is designed, an overview of key functionality and finally a recommendation for how to use the database. The latter is of technical nature. In chapter 4, the recommendations for use of observational data and databases are discussed in a more strategic manner with special emphasis on ongoing international programmes and activities.

2.1 Description of metadata

Metadata are in fact data about data. They provide the information that the data-user needs in order to understand the actual data. For an atmospheric observation, the data can be a series of numbers that does not make sense unless you provide the metadata on what the numbers represent. Typical metadata in this case would be time and location of the measurement, what parameter is measured, what is the uncertainty in the measurement, who did the measurement, what unit is used, etc. For RETRO purposes, where trends in atmospheric composition changes are of main importance, the calibration history and its documentation is an important metadata element.

In the ESA project on calibration and validation of ENVISAT (Cal/Val), a comprehensive effort was put down into developing a structure for defining such metadata. The structure was based on previous developments at NILU, mainly through the experiences gained from the EMEP database that has been operative since 1979. The structure for the Cal/Val database specifies all the metadata parameters that are needed for each data file. Table 1 shows the complete list of metadata parameters used in HDF files at the ENVISAT Cal/Val data centre. The structure is very flexible and is designed to store most types of measurements. The first entry in the table are 12 different parameters that are used to identify the owners of the file. The *Variable Description* and *Visualisation Attributes* must be separately declared for each parameter in the file, while the other attributes only occur once in the file. A document describing the metadata system in detail is available at the RETRO database for observations.

In addition to this structure, most metadata parameters are associated with a separate list of legal values that must be used to describe the observation. This makes it easier to store similar or related types of observations in a comparable manner. As an example, a variable containing ozone should be named O3.CONCENTRATION, and not ozone, ozone_concentration, etc. Only the legal values of metadata will be accepted by the database.

Table 1: Metadata parameters used in the ENVISAT Cal/Val database. The table indicates which parameters that need additional entries of legal values. Blank fields indicate that no changes or additions are needed.

Originator Attributes
PI, DO, DS with NAME, AFFILATION, ADDRESS and EMAIL
Dataset Attributes
DATA_DESCRIPTION
DATA_DISCIPLINE
DATA_GROUP
DATA_LOCATION
DATA_SOURCE
DATA_TYPE
DATA_VARIABLES
DATA_START_DATE
DATA_FILE_VERSION
DATA_MODIFICATIONS
DATA_CAVEATS
DATA_RULES_OF_USE
DATA_ACKNOWLEDGEMENT
File Attributes
FILE_NAME
FILE_GENERATION_DATE
FILE_ACCESS
FILE_PROJECT_ID
FILE_ASSOCIATION
FILE_META_VERSION
Variable Description Attributes
VAR_NAME
VAR_DESCRIPTION
VAR_NOTES
VAR_DIMENSION
VAR_SIZE
VAR_DEPEND
VAR_DATA_TYPE
VAR_UNITS
VAR_SI_CONVERSION
VAR_VALID_MIN
VAR_VALID_MAX
VAR_MONOTONE
VAR_AVG_TYPE
VAR_FILL_VALUE
Variable Visualisation Attributes
VIS_LABEL
VIS_FORMAT
VIS_PLOT_TYPE
VIS_SCALE_TYPE
VIS_SCALE_MIN
VIS_SCALE_MAX

The RETRO database builds on the efforts laid down in the Cal/Val project and uses the same lists of legal parameters. However, new entries to these lists had to be provided in order to cope with the different requirements and the scope of RETRO. This work included continuous updates throughout the project as new observations were added to the database. Because of this, a section of up-to-date values of the various metadata parameters is provided on the database web-interface.

As for the ENVISAT Cal/Val effort, the HDF4.1r3 file format was chosen in order to maintain compatibility with the different database systems. The HDF format is furthermore easy to convert to netCDF which is extensively used by the RETRO modellers. Software tools were provided to the project partners to facilitate use of HDF data.

2.2 Description of database architecture

The RETRO database for observations is built around a MySQL relational database with an automatic file-processor to handle all incoming files and with a dynamic web-portal to allow easy, yet secure access for data users. The system architecture of the RETRO database for observations is shown in Figure 1.

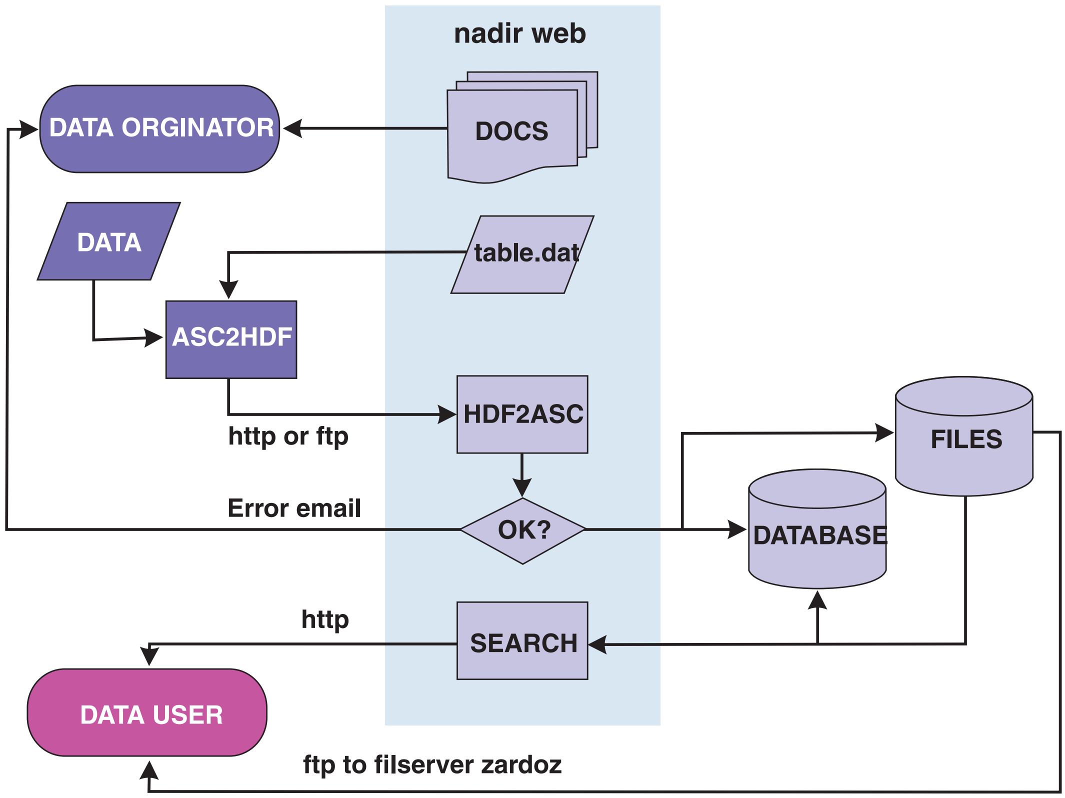

Figure 1: System architecture of the RETRO database.

After registration and provision of access rights, any owner of data may upload his/her data to the database. The person may then access the database through the web interface to download documentations on how to format data files and how to submit them to the data centre. The same web pages also have file conversion tools (ASC2HDF) for download. The ASC2HDF program is developed in Fortran 90 and is made available for the Microsoft Windows, SUN Solaris, HP UX and the Linux platforms. The data originator then formats his/her data with this tool according to the given metadata rules. The files are then submitted to the data centre, either via ftp or through the database web interface. The files are checked for consistency with the current RETRO metadata guidelines, and are either rejected (error-message automatically sent to the data submitter) or inserted into the database. The data files are archived in a hierarchical directory structure while the metadata are stored in a relational database. The file-processing task is handled by system scripts and a program called HDF2ASC (see Figure 1) which runs every five minutes to check for new incoming files. The whole procedure runs fully automated. When files have passed the tests and have been inserted into the system, users of the data may search for and retrieve files through the database web interface. The relational database is used for this purpose.

The RETRO database web interface was made operational in July 2004 and made available to the RETRO consortium through personal user-accounts. The dynamic web-pages are connected directly to the relational database and the interface is programmed in Macromedia Cold Fusion MX. A snap-shot of a web-page is shown in Figure 2.

Figure 2: An example of a web-page from the RETRO database for observations.

The RETRO database web interface has sections for: Documentation, metadata information on legal parameters and values (up-to-date metadata retrieved from database), file conversion tools, plotting tools, data upload, file search, on-line browsing of metadata, on-line plotting of data files and some other RETRO specific topics. Except for some ESA specific topics, all functionality has been kept when the RETRO database was cloned from the previous ENVISAT Cal/Val system. In addition some new functionality has been added. This includes the possibility for a structured export of metadata (to be used by modellers) and a graphical interface to search for data and the ability to upload data-files not formatted according to the given rules. The latter enables storage of e.g. satellite raster images and all the needed metadata are then inserted manually through the web interface. The database functionality as seen from the data users side is described in the next section.

2.3 Description of database functionality

The index database contains the official list of allowed metadata values in the CDB HDF data files, in addition to logs of uploaded/downloaded files, an overview of metadata contents, and the variable list of all accepted HDF data files. All this information is available to dynamic web pages at the web site. The main end user tool on this site is the "Search Data" page, which allows the user to sort through the data files with advanced criteria selections. Filtering by data supplier, project, location, data source, data type, component and other metadata elements is supported. Data files may also be filtered by a "4-D box algorithm" (any file with data relevant for a given geographical location and time). Furthermore, files can be filtered by submission date and update status. All data files that match the search criteria are listed in a new web page, with links to HDF data files download, to comments, and to a variable list. In the variable list page the user may select the different parameters in the file and generate an on-line plot. Plotting is supported for up to 4D data arrays. An example of plotting of a 2D data array is shown in Figure 3. In the file list the user may also select multiple files for download as a tar-ball. The user may furthermore save the search criteria in the index database for convenient re-use at a later time.

In addition to the described search interface a graphic interface showing the geo-graphical location of the data may be used for searching. This is a webpage with a Macromedia Flash MX application showing a world map with information on geo-location. The map works in two modes, the Station and the Trajectory mode, and the user may zoom into the map to study details. The station mode displays data where geo-location is constant within a data file and the trajectory mode displays data where geo-location varies over time, but is constant for each time step in the file. The trajectory mode furthermore displays the altitude of the location by a colour scheme. A user may retrieve data by clicking on a dot (in the station mode) or on a trajectory. Within the flash-application a filter-tool with drop-down menus similar to those of the already described text-based search interface is available, and the user may chose one or several parameters. The map in the background will automatically be updated when the user chooses a value for one of the parameters and dots or trajectories not matching the chosen values will vanish. This allows the user to differentiate between all available data files and the user will see the location of all data files before he/she press the "Get file(s)" button at the bottom of the filter. It is also possible to click on any of the remaining dots/trajectories to retrieve data from only that parcel or location. A screen shot of the mapping tool is shown in Figure 3.

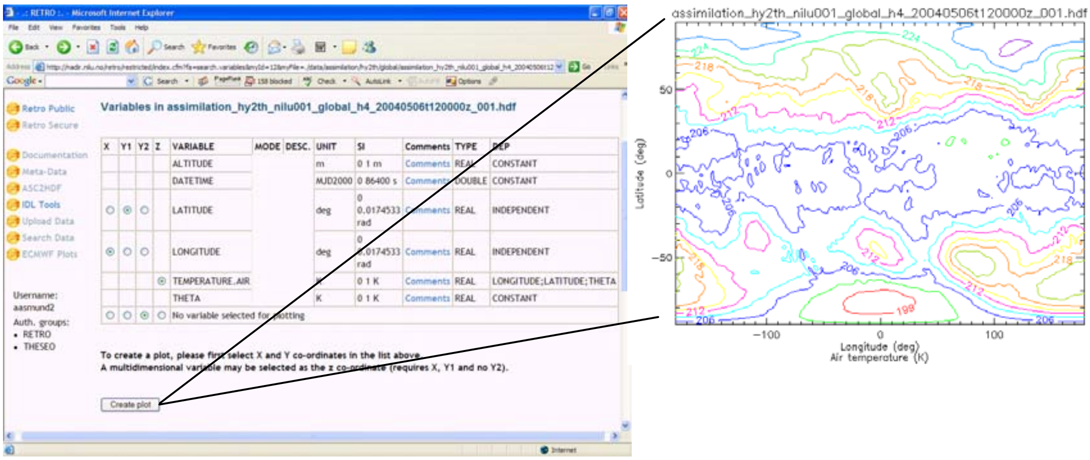

Figure 3: Screenshot showing listing of variables/parameters in a file and on-line visualisation of data content.

Users that have an IDL license may download IDL scripts for HDF data file formatting (excluding the detailed error checking available in ASC2HDF) and for plotting of data sets from HDF files.

3 Description of the observational data available in the database

Early in the project, it was decided that observational data have to be available in a structured and standardised form in order to allow extensive validation against the various model runs. The HDF4 format with metadata definition from the ENVISAT Cal/Val activity was chosen and a large effort on conversion of files from historical observations was done. Regarding the choice of datasets we mainly concentrated on groundbased and balloon borne observations since the ETH-Z database already contained data from several large-scale aircraft campaigns. In this way, we got two complementary data collections. We collected measurements from rural, regional or global sites that were representative for larger geographical areas, and it was important to focus on datasets covering the time period of the RETRO project as a whole. For this reason, newer types of data such as VOC measurements were not considered for being archived in the database.

Another aspect that needed to be considered, was the amount of data available for a certain type of measurement, and efficiency in data format conversion was a major concern. It was necessary to develop specially adapted conversion software to read in the historical data. Data were generally stored in common formats for one data type, but each measurement programme typically used different archive procedures and file standards. We therefore focused on data records with significant size and importance in order to avoid too much technical programming work for reading data. For these reasons, data from the EMEP program, the WOUDC database and some of the CMDL measurements were converted to RETRO HDF-files and uploaded to the database. In addition, text-files from the WDCGG (World Data Centre for Green-house Gases) were archived in the system and HCHO data from the GOME satellite instrument were converted and uploaded. More than 42500 data files have been processed for the RETRO database. The various datasets are described in the following and discussed in terms of measurement methods, data quality issues and known problems.

3.1 EMEP data

The EMEP monitoring programme was originally focused on long-range transport of acidic compounds and measurements of acid rain. The EMEP database therefore contains large amounts of data from chemical analyses of precipitation and from particles collected on filters. These data are in many ways the core data product of the EMEP program, but are they not of specific interest for the RETRO project, since the global models do not calculate the chemical composition of aerosols nor produce accurate estimates for chemical concentration in precipitation. However, the EMEP monitoring program has evolved since it started in 1979 and now includes data from surface measurements of reactive gases such as SO_2, NO_2 and O_3. These data sets are interesting for the RETRO project and all available data sets of these three compounds have been extracted from the EMEP database, converted into HDF-files and uploaded to the project database. This includes a total of 1950 files covering the period of 1975 to 2004 and the region of (mainly western) Europe. Data are structured in one file per station, component, and year.

The quality of the data is assured as part of the yearly EMEP report to the Convention on Long-range Transboundary Air Pollution of UN-ECE. The detailed flagging system used by the EMEP programme is not preserved in the converted files, but only data values marked as "valid" are converted as they are. Other values are declared *missing*. Other general information about the quality of EMEP data may be found at http://www.nilu.no/projects/ccc/qa/. Error estimates are never reported for EMEP data, but statistical methods are used at the EMEP data centre (at NILU) to check data consistency as part of the QA/QC routines. The specific data quality objectives for each component are given at the EMEP-CCC web pages: http://www.nilu.no/projects/ccc/qa/dqo.html. Results from field intercomparison of European NO_2 and SO_2 measurements are furthermore given in http://www.nilu.no/projects/ccc/qa/summary_air.htm. These provide a good overview of the quality of data at the various sites. EMEP measurements are furthermore described in detail in yearly data reports. All of them are available from NILU at the EMEP-CCC web pages, but only the most recent are available in electronic format. A complete overview of all reports are available at: http://www.nilu.no/projects/ccc/reports.html.

3.1.1 SO_2 measurements

A total of 815 data files are available from the stations shown in Figure 4 for the period from 1975 to 2003. Most of the data are daily averages, but some of the newest measurements are hourly means. All data are reported to the EMEP data centre as μg Sulphur per cubic meter, but the values are converted to μg SO_2 / m^3 to allow for easier comparison with model-data.

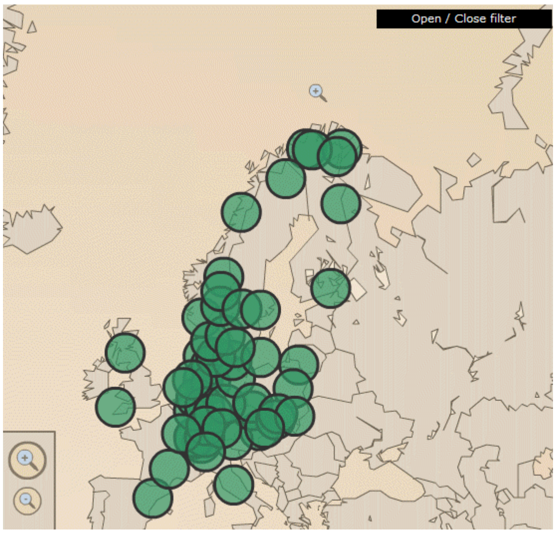

Figure 4: Geospatial availability of EMEP SO_2 data in the RETRO database.

A variety of measurement methods have been applied throughout the years. In particular the oldest data were measured using air samplers like filters, absorption tubes and absorption solutions. The samples were then analysed in a laboratory and data were manually reported to EMEP. The quality of these data may be as high as that of more modern methods, but such data collection is considered to be too costly nowadays. Newer data are typically acquired by using air monitors that automatically collect the air sample and perform the chemical analysis. For SO_2 measurements, UV fluorescence and DOAS (Differential Optical Absorption Spectroscopy) techniques are used for detecting the trace gas concentration. The conversion from manual to automatic sampling occurred gradually at the different stations.

3.1.2 NO_2 measurements

A total of 442 files are available from the stations shown in Figure 5 for the period from 1980 to 2003. As for SO_2 data, most of the measurements are daily averages, but hourly sampling is also used for some stations. All data are reported to the EMEP data centre as µg Nitrogen per cubic meter, but the values are converted to µg NO_2 / m^3 to allow for easier comparison with model-data.

As for SO_2 measurements, the NO_2 data have been acquired using a variety of more or less standardised methods. Chemiluminescence is used for detecting the trace gas in automatic air monitors and measurements are aggregated into daily averaged values before delivery to the EMEP data centre. The method simultaneously also determines the amount of NO, but these data are usually not reported to EMEP. Manual methods using air samplers, with subsequent analysis in a laboratory, are commonly performed by absorbing the gas in a solution or on a filter or glass sinter.

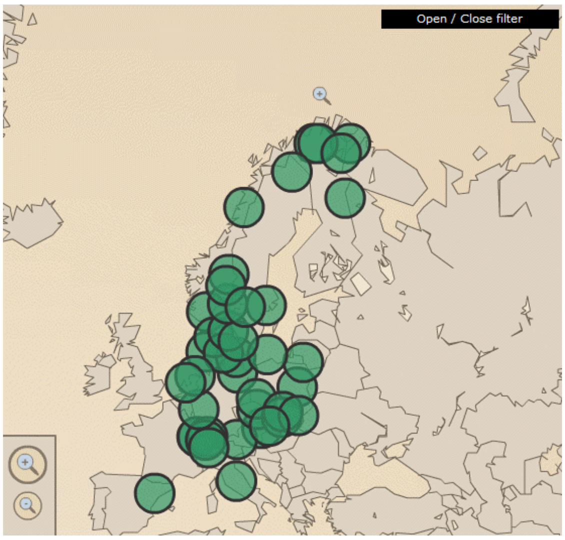

Figure 5: Geospatial availability of EMEP NO2 data in the RETRO database.

3.1.3 O_3 measurements

A total of 692 files are available from the stations shown in Figure 6 for the period from 1984 throughout 2003, all provided as hourly averages. Data are reported as $\mu g\ O_3 / m^3$.

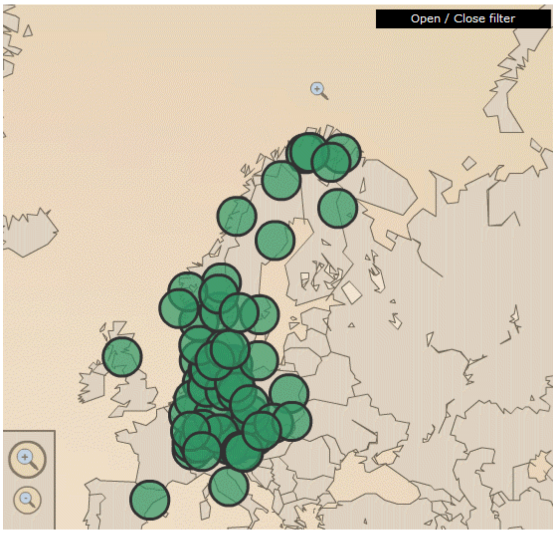

Figure 6: Geospatial availability of EMEP O3 data in the RETRO database.

All surface ozone data from the EMEP network have been acquired with the same method: Air monitors perform automatic analyses of the ozone concentration by absorption of UV light.

3.1.4 File names and metadata

The RETRO metadata definitions require the name of the data source (e.g. instrument or model) to be provided with the measurements. The description tag consists of three elements: 1) the name of the instrument (e.g. spectrometer, O_3

sonde, barometer, etc.) 2) the acronym of the institute that owns the instrument and 3) a number to differentiate between instruments if an institute operates several ones of the same type. For EMEP data, it is not trivial to extract the institute acronym information from the database, and “NILU” was used as acronym for all data sources instead. Providing the correct institute acronym was not considered essential for the RETRO project. The HDF file name consists of 6 underscore-separated metadata parameters, and one of them is the data source, all EMEP data files will contain the entry NILU001. As explained above, this does not mean that NILU has operated all the instruments performing the measurements.

Regarding the information on Principal Investigators (PI), Data Originators (DO) and Data submitters (DS), these were all set to point at the EMEP data centre and the people running it. The PI is therefore always Kjetil Tørseth, DO is simply EMEP and DS is Aasmund Fahre Vik. This procedure made data conversion much simpler since data contact personnel is sometimes stored unsystematically in the EMEP database. Correct names of the current PIs and contact persons for each participating country may be obtained at the EMEP-CCC data centre, http://www.nilu.no/projects/ccc/onlinedata/common/main_ozone_contacts.html. Information on calibration history of the various instruments are available upon request from NILU. Measurements and calibrations are done according the requirements of EMEP Standard Operating Procedures (SOPs).

3.2 WOUDC ozone sonde data

Most long-term tropospheric ozone profile information is available from measurements from light balloons reaching altitudes of approximately 30 km before they burst. These measurements are archived in the World Ozone and Ultraviolet Data Center (WOUDC) at Toronto (Canada) (http://www.woudc.org). The longest series including the used sensors and some additional information are listed in Table 1. The bigger part of long-term ozone sonde measurements is available in the Northern Hemisphere. In the Southern Hemisphere, measurements are strongly restricted to mid-latitudes, where long-term ozone sonde measurements started at only one single station in 1986 (Lauder, New Zealand, see Table 1). The temporal resolution of measurements varies considerably from station to station. For instance, whereas only very few measurements, ranging from no to around 20 ascents, are available from the Japanese stations in the second half of the 1970s, nowadays many stations record profiles at an interval of 2-3 days.

3.2.1 Description of the dataset

Section 3.2.1. is from a manuscript in preparation (Staehelin et al., 2006). The presently used sensors can be classified into two main types: In the Brewer Mast (BM) sonde two electrodes are located in the same cell consisting of a silver anode and a plantinum cathode immersed in a solution of potassium iodide. In the former GDR another sensor similar to the design of BM was used. However, the data quality of these measurements was poor, and therefore they are not listed in Table 2. In the electrochemical (ECC) and carbon-iodine sondes, the electrochemical cell contains two half cells both containing a platinum mesh serving as electrode with the two chambers linked together by an ion bridge.

Table 2: Long-term ozone sonde records. BM: Brewer Mast type; ECC: Electro chemical sonde; KC denotes a sensor similar to ECC sondes (adapted from Staehelin, 2002).

Name of the station	Lat./Long.	Type	Period	averaged number of ascents per month
North America				
Resolute, Canada	75°N/95°W	BM	Jan. 1966-Nov. 1979	Jan.70-Dec.96: 3.3
		ECC	since Dec. 1979	
Churchill, Canada	59°N/94°W	BM	Oct.1973-Aug. 1979	Jan.70-Dec.96: 3.3
		ECC	since Sept. 1979	
Goose Bay, Canada	53°N/60°W	BM	June 1969-Aug. 1980	Jan.70-Dec.96: 3.7
		ECC	since Sept. 1980	
Alert, Canada	83°N/62°W	ECC	start: 1987	
Edmonton, Canada	53°N/114°W	BM	Oct. 1972-Aug.1979	Jan.73-Dec.96: 3.5
		ECC	since Sept. 1979	
Boulder, USA	40°N/105°W	BM	August 1963-July 1966	
		ECC	since March 1979	Jan. 80-Dec.84: 1.5 Jan. 85-Dec-96: 3.8
Wallops Island, USA	38°N/76°W	ECC	May 1970	May 70-Apr. 95: 2.5
Europe				
Uccle, Belgium	51°N/4°E	BM	Jan. 1969-March 1997	Jan.70-June 89: 9.1 July 89-Dec. 96: 11.0
		ECC	since Apr. 1997	
Hohenpeissenberg, Germany	48°N/11°E	BM	Nov. 1966	Jan.70-Dec.77: 3. Jan. 78-Dec.96: 9.9
Payerne, Switzerland	47°N/7°E	BM	Nov. 1966-2002 **(?)**	Jan.70-Dec.75: 8.2
		ECC	since 2003 ?	Jan. 76-Dec.96: 10.9
Biscarosse, France	44°N/1°W	BM	March 1976-Dec. 1982	
Haute Provence, France	44°N/6°W	ECC	1989-present	1 per week
Asia				
Sapporo, Japan	43°N/141°E	KC	Dec. 1968	Jan.70-Dec.74: 2.6 Jan.75-June 89: 1.0 July 89-Dec. 96: 3.2
Tskuba (Tateno), Japan	36°N/140°E	KC	Nov.1968	Jan.70-Dec.74: 2.4 Jan.75-June 89: 1.7 July 89-Dec. 96: 4.2
Kagoshima, Japan	32°N/131°E	KC	Jan.1969	Jan.70-Dec.74: 2.3 Jan.75-June 89: 1.0 July 89-Dec. 96: 3.0
Tropics:				
Naha, Japan	26°N/127°E	KC	start: Sept. 1989	
Hilo, USA	20°N/155°W	ECC	Dec. 1982	Sept. 82-Dec.96: 3.3
Southern midlatitudes				
Lauder, New Zealand	45°S/170°E	ECC	Aug. 1986	Aug. 86-Dec.96: 5.7
Antarctica				
Syowa	69°S/40°E	KC	1966	

The sensors used in Japan (KC) are a modification of the carbon-iodine sonde. Experience shows that the Brewer Mast type sensors provide more scattered results and they need careful and time consuming pretreatment before launch to gain reliable results. Due to these problems, the BM sondes were replaced by

ECC sensors over the past decades except at one station (Meteorological Observatory Hohenpeissenberg, Germany).

The switch from BM to ECC sensors can lead to serious inhomogeneities in tropospheric ozone time series (e.g. Tarasick et al., 2002), as has been demonstrated for the Canadian stations (see Table 2, see Tarasick et al., 2005). Various comparative campaigns that were carried out over the last three decades investigated the difference between the response of BM and ECC sensors in simultaneous ascents. Most of the earlier comparisons showed that BM sensors measure approximately 10 to 25% less ozone in the trosposphere than ECC-sondes ((e.g. Hilsenrath et al., 1986; Beekmann et al., 1994) which contradicts more recent results showing 7-8% more ozone for BM sensors (Calisesi et al., 2003). The inhomogeneities in a long-term series related to the change in the sonde type can be minimized evaluating simultaneous measurements of the two sonde types over a sufficiently long period at the respective site as was done for the series of Uccle (Belgium) (De Backer et al., 1998; Lemoine & De Backer, 2001) and Payerne (Switzerland) (Stuebi et al., .manuscript in prep.)

Today, two companies produce ECC sondes (using different pumps), and two types of solute concentrations are currently used. Large progress in characterizing the differences between ECC ozone measurements has been made by recent laboratory experiments. In the Jülich Ozonesonde Intercomparison experiment (JOSIE), ozone is produced by an ozone generator in the environmental simulation chamber at the World Ozonesonde Calibration Facility at the Research Center Jülich. The different ozone sonde types were tested under a variety of realistic pressure and temperature conditions simulating entire atmospheric profiles, and the results were compared with those of a UV-photometer which served as a reference of the experiment (e.g., Smit and Sträter, 2004). The results of JOSIE indicated that both, the type of pump and used solute concentrations can have a substantial effect on the measured ozone concentration in the troposphere were concentrations are low compared to the stratosphere. In summer 2004, the results of JOSIE measurements were confirmed by field measurements in the campaign BESOS, in which a large gondola was equipped by ECC sondes manufactured by the two companies and each of them with two solute concentrations together with an UV instrument. Laboratory experiments have shown that after the fast response of the sensor attributable to reaction (1) a slow signal appears in ECC sondes which also can affect the data quality of the sonde (Johnson et al., 2002). Thus, changes in ozone sonde practice can introduce artificial breaks, which is particularly important when looking at long-term trends.

The quality of a single ozone sonde profile can be crudely measured using the correction factor (CF). The CF compares the amount of ozone integrated over the profile plus the amount of ozone above the burst level with the column ozone measured by a collocated ozone spectrophotometer such as a Dobson instrument. The ozone amount above burst level can be extrapolated or obtained from satellite climatology. In BM sondes, the CF is used to linearly scale the profiles. Ideally, the CF is expected to be one, and deviations from one can be used as measure of the data quality of the ascents. However, the CF is only an approximate measure of data quality because of potentially compensating errors associated with over-estimation of concentrations in one segment of the profile and underestimation in another. Data quality and the CF are also affected by a changing pump efficiency with altitude, which decreases with decreasing ambient pressure (Komhyr and Harris, 1965). The standard practice to correct for this effect is the use of pump correction profiles (E.g. Steinbrecht et al. 1998 and references therein).

3.2.2 Data conversion

All the ozone sondes available at WOUDC were downloaded to NILU and a program was written to convert the data into regular HDF files. This program was a further development of a program that is running operationally at NILU, processing ozone sonde data in NRT and converting it into the CREX format for ECMWF. These data are used for off-line validation of their assimilation system and an automatic quality control procedure is therefore applied to the sonde data by the conversion program. The data converted for RETRO are therefore also screened for obvious errors such as negative pressures or below freezing point temperatures in the ECC electrolyte. Files containing such errors were rejected by the conversion program. A total of 40484 sondes from 109 stations were successfully converted and uploaded to the database. Spatial coverage is indicated in Figure 7, and it should be mentioned that the sondes from the Atlantic ocean have been launched from ships. As for the EMEP data, some simplifications regarding extraction of PI and data originator (DO) information from the files were applied and PI was always set to “Aasmund Fahre Vik” and DO were set to WOUDC. The proper information was sometimes missing in the original files.

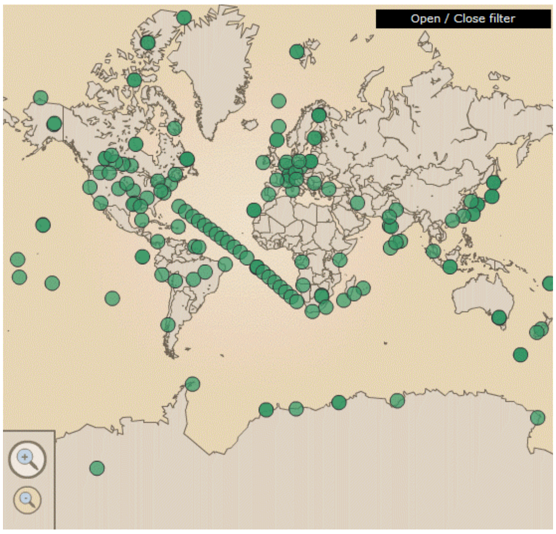

Figure 7: Geospatial availability of WOUDC ozone sonde data in the RETRO database.

3.3 CMDL CO data

As of October 1, 2005, the Climate Monitoring and Diagnostics Laboratory (CMDL) has merged into the Earth System Research Laboratory (ESRL) as part of its Global Monitoring Division (GMD).

GMD's mission is to observe and understand, through accurate, long-term records of atmospheric gases, aerosol particles, and solar radiation, the Earth's atmospheric system controlling climate forcing, ozone depletion and baseline air quality, in order to develop products that will advance global and regional environmental information and services.

A full description of the CO measurement programme at CMDL/GMD is available at ftp://140.172.192.211/ccg/co/flask/README_co.html.

Measurements are available from 1988 and onwards, and data are stored in one file per station. CO data are recorded by using flasks to sample air that are subsequently analysed by a gas chromatograph in a laboratory. Air samples are collected both from fixed sites and from mobile platforms (ships), but only data from the fixed sites have been converted and made available in the RETRO database. A total of 38 files are available from the stations shown in Figure 8.

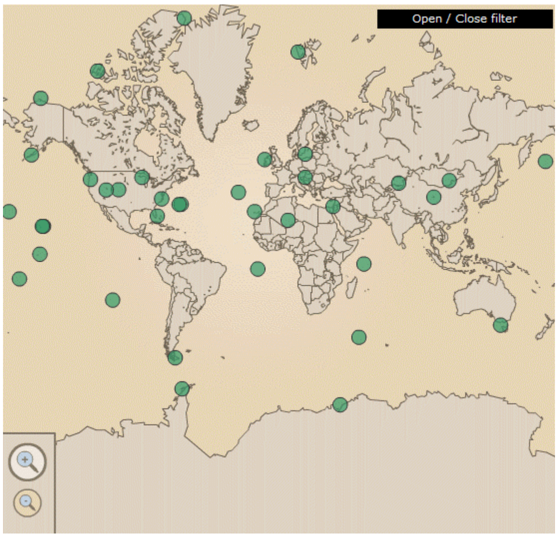

Figure 8: Geospatial availability of GMD (CMDL) CO data in the RETRO database.

3.3.1 Measurement history

The text below is adapted from the GMD (CMDL) web-pages.
Through most of the period 1988-1991, one flask of a sample pair was analyzed for CO, and when there was suitable pressure remaining in the flask, two or more aliquots were analyzed. The difference in CO mixing ratios between the two aliquots, were then used as an indication of the precision of the measurement. Details of the analytical procedures through 1990 are described in Novelli et al. (1992). From the beginning of CO measurements in flask samples (mid 1987) to December 1990 a single point calibration sequence were used, as the response characteristics of the CO instrument had been shown to be linear over a range of 0 to 1000 ppb CO (Novelli et al., 1991).

In January 1991, the instruments used for the analysis of CO in flask air samples were changed. The new instrument exhibited a non-linear response over the range of 0 to 250 ppb CO. It was therefore necessary to change from a single-point calibration routine to a multiple standard calibration scheme (a multi-point calibration procedure based upon that used for calibration of standards, as described in Novelli et al. (1994; 1998) was also used for flask analysis).

In October 1991 CMDL began analyzing a single aliquot from both members of a flask pair (rather than 2 or more aliqouts from a single flask). The principle reasons for the change were to simplify flask-handling procedures (the CMDL carbon dioxide and methane projects also measure both flasks of the sample pair) and to have flask pair agreement (the difference in mixing ratio between the two flasks collected simultaneously) as an additional diagnostic to use in evaluating the quality of the data.

In 2002 a correction to the CO data were applied. Based upon several sets of gravimetric standards, a time dependant correction was applied to all air samples analyzed since 1991. Smallest changes are found in the earlier data. A full description of the revision is available in Novelli et al. (2003).

3.3.2 General comments on data

The text below is adapted from the GMD (CMDL) web-pages and is a description of how measurements are currently taken.

Ambient and standard air samples are injected into the gas chromatograph (GC). Carbon monoxide (CO) and molecular hydrogen (H_2) are separated from other sample constituents using dual columns. CO and H_2 are reacted with hot HgO bed to produce mercury (Hg). Hg is then determined photometrically. The non-linear detector requires a multipoint calibration (6 standards in the atmospheric range are used). This process is highly automated for field and laboratory operations. Measurements are reported in units of nanomol/mol (10^{-9} mol CO per mol of dry air (nmol/mol) or parts per billion (ppb)) relative to the WMO CO scale (Novelli et al., 1991; Novelli et al., 1994; Novelli et al., 1998). Reproducibility of our measurements, based on repeated analysis of air from a high-pressure cylinder, is 1 nmol/mol at 50 nmol/mol and 2 nmol/mol at 200 nmol/mol over the period of our measurements. The absolute accuracy of our CO scale is unknown.

The Pacific Ocean Cruise (POC, travelling between the US west coast and New Zealand or Australia) data have been merged and grouped into 5 degree latitude bins. For the South China Sea cruises (SCS) the data are grouped in 3 degree latitude bins.

Sampling frequencies are approximately weekly for the fixed sites and average one sample every 3 weeks per latitude zone for POC and about one sample every week per latitude for SCS.

The air samples are collected by two general methods: flushing and then pressurizing glass flasks with a pump, or opening a stopcock on an evacuated glass flask. During each sampling event, a pair of flasks is filled.

3.3.3 Conversion of original files into standard HDF format

In order to make conversion of the original text-files into RETRO HDF files effective, a perl-program that automatically reads and re-formats data was written.

The full QC-flag information is not preserved in this process, but data marked as missing in the original files are also set as missing in the HDF files. Specific information on sampling methods for individual data points and instrument identification are also not preserved in the conversion routine as these data are not easy to extract from the raw data.

3.4 GOME data

In the following, the GOME formaldehyde and tropospheric nitrogen dioxide data products developed within the frame of the RETRO project are described. The emphasis is on the technical aspects and not on the general retrieval or science questions involved which have been discussed in a number of recent publications (see reference section). Also, the details of the implementation and the results of the many sensitivity studies performed for individual parameters of the retrieval are not included and can be found in Wittrock (2006) and Nüß (2005).

The Global Ozone Monitoring Experiment (GOME) aboard the second European Remote Sensing satellite (ERS-2) is successfully monitoring the earth atmospheric composition since its deployment in space in April 1995 (Burrows et al., 1997 and references therein). GOME is the first European passive remote sensing instrument operating in the ultraviolet, visible, and near infrared wavelength regions. Primary data products are vertical columns of ozone and nitrogen dioxide. Advanced products include vertical profiles of ozone and columns of bromine oxide, chlorine dioxide, and other minor trace gases as well as tropospheric NO_2 columns (e.g. Burrows and Chance, 1992; Eisinger and Burrows, 1998; Hegels et al., 1998; Richter et al., 1998; Wagner and Platt, 1998; Hoogen et al., 1999; Wittrock et al, 1999; Chance et al., 2000).

3.4.1 The GOME instrument

GOME is a 4 channel UV/visible grating spectrometer covering the spectral range from 240–790 nm with a spectral resolution of 0.2–0.4 nm. GOME was launched on the ERS-2 satellite, and atmospheric data is available since July 1995. The satellite is in a near polar sun-synchronous orbit with 14 orbits per day and a descending node of 10:30 LT. This is an important constraint on the GOME measurements as the local time of measurement will always be the same for a given latitude. In low and middle latitudes, GOME will probe the atmosphere in the late morning which has to be taken into account when interpreting the measurement results.

GOME observes the solar radiation scattered in the atmosphere or reflected on the ground in near nadir viewing geometry. With an across-orbit swath width of 960 km, global coverage is reached at the equator within three days, and within one day at 65°. Each scan is divided into 4 sub-scans, three in forward direction with a spatial resolution of 40 x 320 km^2 and one fast back-scan covering 40 x 960 km^2. On some days, the swath width is reduced to 240 km, increasing the spatial resolution to 40 x 80 km^2 at the cost of spatial coverage.

Once per day, GOME takes a measurement of the direct solar irradiance, and this measurement is used as an absorption free background spectrum in the analysis. Unfortunately, the measurements have to be taken using a reflecting diffuser, and this optical component introduces an offset in the analysis that varies with time (Richter and Wagner, 2001). Therefore, a compensation measurement has to be used, usually over an unpolluted region. This is discussed in more detail below.

Since June 2003, the last tape recorder on ERS-2 failed and only about 30 per cent of the data can be retrieved on ground although the instrument is still fully functional.

3.4.2 NO_2 measurements

The GOME tropospheric NO_2 analysis is performed in a series of steps. Briefly, the first step is the retrieval of the total amount of NO_2 along the effective line of sight (Slant Column SC). This is determined using the well known Differential Optical Absorption Spectroscopy (DOAS) technique. The slant column depends on the viewing geometry, the solar zenith angle and the amount and vertical distribution of the absorber in the atmosphere. The second step of the analysis is the correction for stratospheric absorption which is in principle achieved by subtracting a simulated stratospheric slant column based on a SLIMCAT model run. The third step is the conversion of the remaining tropospheric slant column to a tropospheric vertical column using an airmass factor (AMF) based on radiative transfer calculations. As the airmass factor is strongly dependent on a priory assumptions, a large effort went into improving the reliability of the input used.

A simplified formula describing the overall method used to derive the vertical tropospheric columns is

$$VC_{trop} = \frac{SC - SC_{strat}}{AMF_{trop}} = \frac{SC - VC_{SLIM} * AMF_{strat}}{AMF_{trop}}$$

where SC is the slant column, VC the vertical column, AMF the airmass factor and the index strat and trop stand for the stratosphere and troposphere, respectively and SLIM denotes SLIMCAT data.

Tropospheric columns of NO_2 have been derived and uploaded as monthly averages to the NILU database for all GOME data available since January 1996. The spatial resolution of these data sets is 0.5° both in latitude and longitude.

Some information on parameterisation and conditions of the data sets is given below:

1. Use of daily SLIMCAT data to account for the variability in stratospheric NO_2
2. Use of the surface albedo climatology of Koelemeijer et al.
3. Use of a variable tropospheric aerosol loading
4. Correction for topography effects
5. Use of airmass factors that are based on vertical NO_2 profiles predicted by the MOZART model for 1997
6. Use of a 0.2 cloud cover threshold based on the FRESCO algorithm

3.4.3 HCHO measurements

The GOME HCHO product is derived from GOME spectra in a similar procedure than that for NO_2. First a DOAS algorithm to derive HCHO slant columns from earth irradiance spectra measured by GOME has been developed: A spectral fitting window of 337.5–357 nm was selected here. The next step is the conversion of the slant columns to tropospheric vertical columns using appropriate airmass factors (AMF) based on calculations with the full spherical radiative transfer model SCIATRAN. Since the AMF for satellite observations depends strongly on several meteorological boundary conditions it is therefore necessary to account for all the parameters in the radiative transfer calculations. Compared to the standard AMF used in previous HCHO data sets (Wittrock et al., 2000), deviations by a factor of up to four are possible. This is particularly true for the source regions e.g. during biomass burning events as they not only have the largest values close to the surface but in addition, are often correlated with regions of enhanced aerosol burden. In order to obtain a realistic global picture of HCHO, the most practical way to do this is to create pre-calculated tables of AMF for a set of SZAs which depend on time and location and are interpolated to the conditions of an individual measurement. In order to be able to investigate the satellite measurements even on a regional scale, the spatial resolution for the AMF tables was selected to 0.5° x 0.5° in latitude and longitude. With that tropospheric columns of HCHO have been retrieved and uploaded as monthly averages to the NILU database for all GOME data available since April 1996. The spatial resolution of these data sets is the same as for NO_2: 0.5° both in latitude and longitude.

Again some information on parameterisation and conditions of the data sets is given below:

1. Use of the surface albedo climatology of Tanskanen et al.
2. Use of a variable tropospheric aerosol loading
3. Correction for topography effects (TerrainBase, Row et al.)
4. Use of four different HCHO profiles depending on origin (anthropogenic, background, biogenic, biomass burning)
5. Use of a 0.2 cloud cover threshold based on the FRESCO algorithm

3.5 WDCGG surface ozone data

The world Data Centre for Surface Ozone (WDCSO) was originally hosted by NILU, but the task was handed over to the World Data Centre for Greenhouse Gases in 2002. WDCGG is hosted by the Japanese Meteorological Agency and established under the Global Atmosphere Watch (GAW) programme to collect, archive and provide data for greenhouse (CO_2, CH_4, CFCs, N_2O, etc.) and related (CO, NO_X, SO_2, VOC, etc.) gases and surface ozone in the atmosphere and ocean, measured under GAW and other programmes. From their web site, one may obtain information including WDCGG's publications and measurement data that have been contributed by organizations and individual researchers in the world: http://gaw.kishou.go.jp/wdcgg.html.

Data from a total of 26 stations (one file per station) has been upload to the RETRO database. The data have not been reformatted into HDF-files, but are stored in the original text format. Data that are part of the EMEP monitoring network have been archived as HDF files as described in section 3.1.

4 Identification of other observational data sets useful for model evaluation

RETRO focuses on gas-phase chemistry and especially oxidising capacity in the troposphere and to a certain degree regional air pollution trends. This brief summary of other known datasets are therefore focused on this and not on climate related compounds or aerosol species.

The ETH-MEG database is specifically designed to support CTM and Chemistry-GCM modelling groups in evaluating their (global) models by comparing simulated with observed distributions of the most relevant tracers related to ozone photochemistry. The database contains data from numerous measurement campaigns making use of both scientific and commercial aircraft as observation platforms, and ozone soundings. These data were processed in a way that supports an easy and direct comparison with output from a model. Currently, the measurements span the period from 1 Jan 1995 until 31 Dec 1998. The ETH-MEG database has been actively used in the RETRO project and has a special focus on the UTLS region.

The EMEP data made available for RETRO is only a small part of the data collection currently available at NILU. The database contains several hundred components from analyses of precipitation, snow, air and particles. Some of these may be of interest for further studies of past tropospheric composition. Other long-term field campaign data set are those from FZ-Jülich, the University of Michigan, the TOR-1 and the NADIR databases. The two latter are both hosted by NILU and cover tropospheric and stratospheric campaigns respectively.

Two major aircraft monitoring activities are currently operational in Europe – the MOZAIC and the CARIBIC projects. Old MOZAIC data are made available through the ETH-MEG database (only after signature of the MOZAIC data protocol), while a complete set of new and old data are available from the MOZAIC database web portal at http://www.aero.obs-mip.fr/mozaic/. The programme is currently being continued through the IAGOS project and data are in the future to be found through http://www.fz-juelich.de/icg/icg-ii/iagos.

Satellite data are of major importance for model validation, making assessments and for data assimilation purposes. Only formaldehyde records from the GOME sensor have been considered for the RETRO database, but there are numerous data products available in addition to this. Of main interest are the NO_2 and tropospheric O_3 products (the ozone record are not suitable for direct comparison since its quality is too poorly assessed) from GOME and the continued production of similar products from SCHIAMACHY and OMI. Most of these data are currently available through the PROMOTE pages (www.gse-promote.org) or through the web portals of KNMI (http://www.temis.nl), University of Bremen (http://www.iup.uni-bremen.de/sciamachy) or University of Heidelberg (http://giger.iup.uni-heidelberg.de/). In addition to these, a number aerosol products are also available through the PROMOTE web pages. Time-series of American satellite data are also available from several sensors such as MOPPIT (tropospheric O_3 and CO) and MODIS (aerosols). Data from more recent American instruments onboard the AURA satellite are available through the validation data centre (http://avdc.gsfc.nasa.gov).

Groundbased aerosol records are available from the PHOTONS and AERONET networks (Groundbased AOD), the EUSAAR project (Groundbased in-situ aerosol properties) and EARLINET-ASOS (Groundbased remote-sensed profiles of aerosol properties).

The Network of Excellence ACCENT has an activity focusing on making field data available for research purposes. This consists of a meta-database with references to existing databases, monitoring networks and campaign data archives. The metadata is available through the accent web portal (http://www.accent-network.org), and will be continuously updated in the future. The work in RETRO contributed to the set-up and development of the ACCENT metadata base.

5 Future atmospheric data centres

This chapter describes the current status of atmospheric data centres and how they are likely to evolve in the future. General monitoring strategies for atmospheric composition are not presented, but focus is given to issues related to data management. Strategies related to future atmospheric monitoring are discussed in deliverable D2-5.

5.1 Current observational data centres

The developments of atmospheric data centres in the past may be summarised in the following bulleted list:

- Early monitoring data recorded manually starting some 100+ years ago. Large increase in monitoring activity around 1957 and Int. Geophys. Year. There is generally a lack of metadata and the knowledge about data quality is limited. Also, the oldest measurements (available since mid 1920s) concern only ozone (total atmospheric columns). For other tracers historical time series are much shorter.
- First digital data archives appearing in 1960s and 1970s (e.g. EMEP database established in 1970ies). There is limited metadata and the databases were generally very discipline specific. QA/QC references are partly missing and no formal SOP's (Standard Operating Procedures) had been agreed upon.
- New data centres were established through the 1980s and 1990s. Generally broader themes and proper documentation of data and methods. Relational databases allowing searchable systems. (Current EMEP database, NDSC)
- Emerging technologies such as internet and communication protocols allow for a wider and much faster distribution of data, products and information after about 1995. Different types of data centres containing data from

campaigns, programmes, projects, etc. are established by 2006. Initiatives are working toward an integrated way of accessing data.

The developments have led to a number of well identified problems:

- There is a general fragmentation between data sources and networks are partly overlapping. There is no location where all available data sets are listed.
- Data sometimes (often) have unknown quality and unknown origin/version status.
- Data are not used to its full potential, partly due to lack of knowledge about their existence, but also due to data access restriction.
- Multiple reporting procedures from one data source to several databases occurs. Data in the different databases are often not updated at the same time.
- Many different data formats. Metadata (data about data) contents are not standardised. Data formats are not always rigorously applied so that errors occur during data treatment.
- There is a lack of metadata in general, and in particular about instrument changes, calibration, changes in location etc. Such data are essential to ultimately establish homogeneous time series.
- The fact that users normally need data from several data sources have led to the creation of data collections in so-called snap-shot databases. This is done in order to make searching for and use of data easier and the task is commonly coordinated through a time limited project. The RETRO, TRADEOFF and AEROCOM databases are examples of such snap-shot databases, where data have been copied from its original source once and there is a chance that some datasets become outdated after e.g. a major reanalysis of the original dataset. This leads to the coexistence of several different versions of one dataset. There is furthermore a risk that snap-shot databases provide limited acknowledgement to the program through which funding of the measurements are provided.
- There is often a lack of long-term commitment by the database provider as the initial establishment of a data collecting system (i.e. a database with associated data submitters) is often accomplished through a research project with limited duration and funding. It is important that data services seeks foundation in monitoring conventions to become sustainable in the long run.

5.2 Future operational atmospheric data centres

It is essential that the future development of atmospheric data centres are to be user driven in order to maximise the output of all the effort laid down in monitoring, archiving and dissemination of data. Having that said, it is not directly obvious who the users are and what data centres may be useful for the different user categories. There is a question whether a given data centre should advance through input from all potential users or just from selected groups (e.g. only through input from research scientists and not from policy makers).

It is not intended to provide a full overview of all possible categories of atmospheric data users, but the current document focuses on the needs of those of relevance to RETRO. In addition to a general overview of what future data centre should be able to provide, separate sections are provided for users involved in off-line model validation, NRT model or satellite validation and data assimilation. The data submitters and how to make their part of the work as easy as possible, is also mentioned. A brief overview of other European and international studies that discuss management of atmospheric data is also given.

5.2.1 General aspects

Some providers of ground based data currently have to submit their data to many different data centres and they commonly have to provide their data in several formats. This is e.g. the case for some of the measurements done by the GMD (former CMDL) lab and it is clearly not a fruitful solution. The data centres are typically validation centres such as the AURA and the Envisat Cal/Val databases, real-time systems like the Nadir centre at NILU, the database of the NDSC (now called NDACC) network, the WMO-GAW databases, etc. In the case of an ozone sonde, the data could be submitted to all these data centres and in addition end up in the RETRO and the ETH-MEG databases. One solution to this problem would be to harmonize and connect the different catalogues. The data could then be submitted to just one data centre and then be distributed further through an interconnected network of databases. Such a solution is technically possible today and the system could also keep track of different data versions and reanalyses. It is, however, a very complex task to accomplish and is probably only reasonable to implement for a limited number of parameters. Some problems related to visibility of the individual networks and data providers furthermore need to be solved. In addition, there are commonly different requirements for data quality in the different data centres or network. Access to data through e.g. WOUDC is furthermore unrestricted while data from the Envisat Cal/Val and MOZAIC databases is available only after the signature of a data protocol.

Another option is to keep all data files in only one database and instead dynamically share information on data availability between the various data centres. Such information (purely metadata) is commonly not restricted and could be freely distributed while access to the actual data could remain restricted and regulated by e.g. a data protocol. The data could then only be reported to the network or project it belongs to and all other data centres would point to this network whenever a user requests the dataset. This solution would work fine for data acquired as part of an official monitoring program or research project. Other data, like e.g. European ozone sonde measurements do, however, not belong to a single network or monitoring program and it is not clear which data centre they should be delivered to if only one is to be chosen. Such data are likely to be reported to many databases as long as the data submitters are willing to do it (some data submitters may also prefer direct and personal submission to the users to attract scientific involvement like co-authorship, etc.). Another issue for this approach is the harmonisation of data protocols so that users do not have to deal with different forms and procedures when they acquire multiple data sets.

One data centre is rarely only serving one user-group and one user-group is rarely accessing only one data centre to get the data they need. Data quality routines and data formats are, however, agreed upon within a network or project, and a given data centre is organising its data systems primarily to support project/network objectives. Collaboration between the different networks to harmonise their methods have so far occurred only within some disciplines. One example where such collaboration has been going on for more than two decades concerns measurement and reporting procedures for acid rain related parameters. This is now harmonised fairly well between regional networks and the WMO-GAW system and similar processes are currently ongoing for e.g. harmonisation of aerosol monitoring methods. For other parameters and especially between networks monitoring different parameters or on different scales (e.g. regional vs. urban scale), there is less harmonisation. In these cases, each data centre uses its own standard data format and has its own rules for quality control. As explained above, this is not practical for the data submitters, but it is not practical for the users of the data either. While modellers would prefer to retrieve their validation data in ASCII or netCDF format, those involved in satellite validation would prefer HDF and a meteorological agency would probably require CREX, BUFR or even GRIB files. Common for all user groups is that they would prefer data in one format, but this format is not the same from one group to the other. Their requirements for metadata and auxiliary information is also not uniform. A solution to this problem would be that all data centres support a number of formats and provide routines to convert or export data to/import data from several standardised formats even though the data is archived in a single format.

5.2.2 Demands from users doing off-line numerical model validation:

The model groups involved in RETRO are typical representatives for users running atmospheric models of some kind who want to validate their results with independent observational data or other independent model output. A common approach is to download all these data from a data centre and store them locally. Reformatting and re-gridding of data is often done in order to use it in automatic calculation of e.g. model-bias. The ability to extract data from a centre in a user specified format and grid (spatial and temporal) would therefore be one main interest for this user group. The treatment of measurement errors and uncertainties is generally not sophisticated. Modellers are rarely interested in performing quality analyses or assurance of independent data. This is also made difficult due to the many possible error sources and the lack of standards for describing these errors (e.g. as metadata attribute). It is therefore important that data made available through data centres also contain links to such QA/QC studies that describe the data quality (work is currently ongoing in the ACCENT project to collect and organise information on QA/QC routines) or that the data files contain direct estimates of errors. Provision of characteristic instrument information or averaging kernels (for remote sensed profiling data) is not enough since many of these properties require expert knowledge to interpret. Concise information must instead be made easily available.

In addition to purely technical re-gridding and homogenisation of data, a significant effort is needed to generate properly harmonised time series, climatologies and aggregated data sets. Such efforts involve connection of calibration scales of different instrumental techniques and takes into account the variety in sensitivity and representativenes. This is not done by simply writing a clever computer code, but should be performed by an experienced scientist. The output would be highly useful for model validation purposes.

5.2.3 Additional demands from users doing NRT model or satellite validation

Those users involved in NRT model or satellite validation have most of the same requirements as just described, but timely delivery of data is an additional major concern. This normally means delivery of quality assured data within 2 to 3 hours and it is necessary to include automatic screening routines in the process chain that delivers the data. Data access or data distribution must also be automatic, and this has commonly been handled by cron-jobs and ftp routines in the past. In the future, it will be necessary to also use web-services, broadcasting or GRID technologies in order to handle the increasing amounts of data. A wider implementation of the OpenDAP technology is also an option for the future. OpenDAP is a standard protocol for (distributed) data exchange that can be employed to facilitate access to, and exchange of geospatial and non-geospatial data. It is commonly used within the meteorological community.

Since most of the quality control routines need to be implemented as part of an automated operational routine, it is important that the data quality is well characterised. For networks of groundbased stations providing in-situ or remote sensed data, it is for example important that the data are harmonised and that all stations provide data of similar quality. This could otherwise lead to a geospatial bias in the validation results.

5.2.4 Demands from users doing data assimilation of NRT data

In order to make use of data for NRT assimilation into a CTM or GCM, it is even more important to have a timely and reliable delivery of data. Most data assimilation routines have some built in routines for screening of data in order to avoid large fluctuation in the model fields due to assimilation of erroneous measurements. It is, however, important to have some kind of automatic quality control of the observations and detection of constant biases in the observation is especially valuable in order to avoid drift in the model fields. An estimate of the measurement errors is useful since the assimilation process is able to include this in the calculations.

5.2.5 References to European and international strategies and programmes that includes databases

A number of strategy studies and programmes have been conducted and initiated during the last few years and a brief overview of some of them are given in the following. It is not intended to give a complete overview of all existing programmes/studies, but focus is given to those of main relevance to RETRO and whenever the issue of atmospheric data centre is mentioned. A general discussion about atmospheric monitoring or Earth observations is not included in this document.

GMES

Global Monitoring for Environment and Security (GMES) is a joint initiative between the European Union and the European Space Agency and was in practice initiated in 2002 when the first GMES projects were announced. Several partners in RETRO participated in these initial projects that were trying to define what GMES should contain, how it should be implemented, by whom and when. GMES-GATO was a NILU coordinated project trying to define a monitoring strategy for global atmospheric observation through GMES and provided a detailed strategy report to the EU in 2003. The report is available from the project web pages (http://www.nilu.no/gmes-gato) and discusses all aspects of measurements and assessments of the atmospheric state. A section on regional air quality (AQ) states the following about databases:

> *The effective use of data is a critical part of the assessment of AQ and its impacts. There is a requirement for effective databases of AQ measurements and model data. This could be achieved using metadata databases or even through the integration of various data sources. The effectiveness of any database strategy is inherently linked to QA and QC issues. Databases may not only provide repositories for primary data but should also be used to disseminate results and analyses through appropriate web portals.*

In addition to GMES-GATO, the EU funded several other thematic projects on atmospheric research in relevance to GMES. These were CREATE (Creation on an European Aerosol Database), DAEDALUS (Delivery of AErosol proDucts for Assimilation and environmentaL Use), METHMONITEUR (Methane monitoring in the European region) and APMoSPHERE (Air Pollution Modelling for Support to Policy on Health and Environmental Risk in Europe). None of them were of direct relevance to the objectives of RETRO, but important work on creation of databases and integration of existing monitoring capabilities were initiated in the EU projects.

In addition to the research projects on atmospheric issues, the EU also funded other thematic project on topics like environmental stress, land cover in Europe, etc (all in all 32 projects). Furthermore, there were also funded a number of cross-cutting projects that tried to link together and synthesise the output from the different thematic projects. The BICEPS project (Building an Information Capacity for Environmental Protection and Security) provided a report that describes several issues that are of interest to both the RETRO project and to atmospheric data centres in general. Issues like data acquisition, archiving and services are discussed. A separate chapter is dedicated to the issue on interoperability between different data sources and information services. One user will need access to many data sources and it is stated that it might be difficult to get an overview of all information available and permissions to access the different sets are always different. The report is still sceptic towards yet another open ended data cataloguing due to numerous previous attempts that are not maintained and finalised. With the current capabilities of modern search engines it is necessary to do better than Google (which most users in search for data use today) if the effort should be justified. Metadata cataloguing and enrichment of existing data sets and services is encouraged, but also this needs to be balanced against cost. The NASA Global Master Change Directory is acknowledged as an important collection of environmental metadata, but it is questioned whether it is sufficiently detailed to enable effective searching for data.

The European Space Agency has also funded a series of projects since 2002 and with their GMES Service Element (GSE) they have tried to define and provide data services to designated users. PROMOTE (Protocol Monitoring) was the atmospheric component of GSE and focused on delivering data and information products related to UV radiation, climate and air quality. The project was finished in the spring of 2006, but a PROMOTE stage II is expected to start up later in the year. The project does not focus on issues like data centres or management of atmospheric data, but the project has made an important contribution in making more data and products available. Several potential users of Earth Observation data have furthermore for the first time been invited to and included in the definition of data products. Many of these products or data services will be continued also in the future.

IGACO

IGACO (Integrated Global Atmospheric Chemistry Observations) is a component of the IGOS (Integrated Global Observing Strategy) partnership and is an initiative by WMO and ESA which should be implemented through the global Atmosphere Watch (GAW) of WMO. Within the IGACO a strategy is defined on how to combine ground-based, aircraft and satellite observations with data archives and global models with a purpose to provide information about key atmospheric chemical species. Recommendations related to the implementation of IGACO regarding Data Bases include:

- *Distribution of data: universally recognised distribution protocols for exchange of data on atmospheric chemical constituents should be established (General Recommendation 8, section 5.2.2, IGACO, 2004).*

Data distribution protocols are needed for fast, easy and effective use of the data and quality assurance information, e.g. for assimilation into forecast models.

- *Multi-stake holder World Integrated Data Archive Centres (WIDAC) should be established for the targeted chemical variables (General Recommendation 9, section 5.2.2, IGACO, 2004).*

- *Storage for raw data should be established so that they can be re-interpreted as models and understanding improve (General Recommendation 10, section 5.2.2, IGACO, 2004).*

The recommendation is that observations of key components should be made available through World Integrated Data Archive Centres, which may be centralized or distributed. A WIDAC should have periodic updates of observations and reanalysis in order to be up to date with the improvement in e.g. satellite retrieval algorithms. Access to data should be easy and free of charge.

GEOSS

Global Earth Observation System of Systems (GEOSS) is the most general and high-level one of the three programmes and strategies mentioned here. GEOSS is multidisciplinary and international and GMES is thought to be the European component in the system of systems. GEOSS is still in an early start-phase and a slow start has been initiated in 2006. A 10-year implementation plan has been produced (GEO, 2005) that describes the first details on how the architecture could be developed, but goes quite far in suggesting how data access and interoperability should be handled. If successful, GEOSS is likely to shape how most data centres, data providers and data users are to operate in the future and the programme has a very high level of support from politicians. Especially the acceptance of the programme in the U.S. by most of their major agencies and departments is seen as a sign of strength and sustainability.

Even though GEOSS is general in its approach, the 10-year implementation plan is specific in its goals for system architecture, interoperability and data sharing. As described in the plan, as developed by the ad hoc Group on Earth Observations (GEO) and adopted by the Third Earth Observation Summit held in Brussels Belgium, on February 16, 2005:

> *The success of GEOSS will depend on data and information providers accepting and implementing a set of interoperability arrangements, including technical specifications for collecting, processing, storing, and disseminating shared data, metadata and products. GEOSS interoperability will be based on non-proprietary standards, with preference given to formal international standards. Interoperability will be focused on interfaces, defining only how system components interface with each other and thereby minimizing any impact on affected systems other than where such systems have interfaces to the shared architecture.*

The focus on standardising interfaces rather than trying to create a system that uses only one common standard is a central principle in GEOSS. This allows current system and networks to continue using their specialised reporting routines, formats and metadata definitions. Attention is drawn to the use of standards and international standards organisations and institutes in the definition of inter-operability procedures, and the use of “open-source” software is encouraged. The societal benefits of Earth Observation cannot be achieved without data sharing and the following GEOSS principles are to be used in the implementation:

- *There will be full and open exchange of data, metadata and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation.*
- *All shared data, metadata and products will be made available with minimum time delay and at minimum cost.*
- *All shared data, metadata and products being free of charge or no more than cost of reproduction will be encouraged for research and education.*

The data sharing principles may cause a problem with implementing some European data archives into the system of systems since access to data is sometimes highly restricted. How to handle Intellectual Property Rights (IPR) and at the same time maximise the use of data is therefore an important problem yet to be solved in the GEOSS implementation.

Since the GMES programme was transferred within the European Union from DG Research to DG Enterprise, the research community has lost most of their influence (and possibly also some interest) in the programme. Instead, the DG Research has embraced GEOSS as the major programme for developing capabilities for Earth Observation and managing of data, services and information. An integrated project called GEOMon (Global Earth Observations and Monitoring) is currently in the negotiation phase with the EU and is aiming at providing the European component of GEOSS on atmospheric chemistry. The implementation of IGACO is therefore also of main interest for the project. GEOMon is expected to start up early 2007 and will last for four years. A separate activity on data management is aiming at setting up a central data centre. The data centre will aim at connecting and linking all available European data centres containing atmospheric data together. Providing access to NRT data is also a major issue in this activity.

5.3 Conclusions

As described in Chapters 3, 4 and 5.1, a number of observation and monitoring capabilities have been developed and as described in chapter 5.2, and number of processes have been initiated to improve and further develop the Earth Observation systems for the benefit of the different users. There is a clear tendency towards integrations of data, i.e. using data from multiple platforms and instrument in an integrated way and it is proposed to connect existing systems together into a system of systems. There are a number of well identified problems with the current systems and in order to allow full integration of data and maximise use of monitoring systems, there is a need to:

- Solve the problem of multiple reporting of data to several data centres to avoid excessive work for data submitters and to avoid co-existence of multiple versions of data sets. One solution could be automated sharing of data or metadata between data centres, but this may impose standardisation of data formats and QA/QC routines
- Standardise the use of data protocols so that users do not have to deal with different forms and procedures when acquiring data from multiple sources and originators
- Standardise the data formats used by the different data centres and allow for submission and download of data in several formats even though the data is archived in just on format
- Allow users (especially those involved in “off-line” model or satellite validation) to extract data in a user-specified format and grid
- Increase the amount of supporting metadata that describes the quality of the data. This could be descriptions of calibration history and routines and/or description of QA/QC procedures. Direct estimation of errors should be given as far as possible. Provision of instrument characteristics or averaging kernels (for remote sensed data) is not enough since this requires expert knowledge to interpret
- Generate value-added data sets such as data sets harmonised across changes in instrumentation and/or calibration procedure, climatologies created through

integration of multiple data sources and/or models, etc. This requires an experienced scientist and would be very valuable for model validation purposes

- Provide access to timely delivered and quality assured (screening) data through automatic distribution systems, to allow more efficient NRT (Near Real Time) model or satellite validation
- Harmonise the data quality across networks in order to avoid geospatial bias in model or satellite validation results
- Provide an automatic estimation of errors in NRT data that may be used in on-line data assimilation processes

A number of strategic programmes have been initiated to cope with future evolution of atmospheric monitoring and the issue of data management is dealt with by most of them. The GMES-GATO study states that “Databases may not only provide repositories for primary data but should also be used to disseminate results and analyses through appropriate web portals”. The idea of a one-stop shop with access to a multitude of data from one access point is mentioned by several studies. The BICEPS study is, however, sceptic towards yet another open ended data cataloguing due to numerous previous attempts that are not maintained and finalised. The access point is pointless unless it is able to do better than Google. Apart from the two studies, a number of additional atmospheric research project were funded by EU (DG Research) during the definition phase of GMES. The programme is now handled by DG Enterprise and there is currently a risk that the research community is being decoupled from the continuation of the process and that the suggestions of the initial studies is not being implemented by the new directorate. It is therefore uncertain if GMES will be able to implement the changes and improvements suggested above.

The IGAC has a strong position in Europe and DG Research is currently enabling its implementation through the funding of the GEOMon Integrated Project, which is furthermore providing a European component to GEOSS on atmospheric chemistry (DG Research is also funding a secretariat for implementation of GEOSS at WMO). IGACO (IGACO, 2004) includes data management as part of the monitoring strategy and recommends the following:

- Standardised protocols for exchange of data on atmospheric chemical species should be established (GR 8)
- World Integrated Data Archive Centres should be established for targeted chemical components. These should have periodic updates as e.g. satellite retrieval algorithms is being improved. Access to data should be easy and free of charge (GR 9)
- Raw-data should be permanently archived to allow for future reanalysis as knowledge and models improve (GR 10)

In order to solve the current problems with atmospheric data centres and to implement the recommendations previously described, it is probably necessary that the research community takes an active part in current and future strategic programmes and projects. Significant resources are also required to support developments. The role of GMES is currently uncertain, but it will most likely give rise to a number of new and improved data services and products. To increase the usefulness of these, it is essentials that users (e.g. modellers involved in RETRO) of data get involved in the product definition tasks and tell the service providers what they need. The ESA GSE project PROMOTE Stage II may provide an ideal forum for this. GEOSS is still in an early phase, but its 10 years implementation plan looks very promising and is an ambitious plan for how current monitoring systems could be developed into an integrated Earth Observation system of systems. The focus on links and interfaces rather than the creation of a uniform system seems reasonable. The GEOSS process currently seems to be the most important and promising one regarding development of future atmospheric data centres and data products.

6 Recommendations for optimal use of observations

The RETRO database for observations was developed as a snap-shot database (and therefore contains data copied from their original sources) in order to provide modellers with easy access to large amounts of uniformly formatted data. There is therefore a risk that the data becomes outdated if new revisions are being made by the originators of the data sets. The RETRO project ended in June 2006 and there is currently no funding available for making continuous updates of data. There is, however, plans to make use of the database in the “Task Force on Hemispheric Transport of Air Pollution” which is being organised by the Convention on Long Range Transport of Air Pollution under UN-ECE (United Nations Economic Commission for Europe). This will not only secure some funding (either externally through the task force or through internal NILU funds) for maintenance of the database (and data), but also dictate further developments of the system and inclusion of more data. RETRO was about historic observation and the data were therefore already some years old when they where copied from their host data centres. Most measurements were therefore in a “final” state and frequent revisions of data are not likely to occur in the future. In addition to this, all the conversion of data into HDF files were done with the help of programs that could easily be run again if a dataset is to be updated. The chance of data in the database being outdated is therefore small even though the RETRO database is a snap-shot system (with the weaknesses explained in chapter 5). It is, however, recommended that users of the database keep in mind that new versions of datasets may appear and that there is always a chance that the RETRO data becomes outdated.

The RETRO database was mainly developed to support modellers. The main recommendation to this user group is to use the observational data with care and not treat them as the absolute truth. The instrument accuracies for the different observation types is given for the different species in Chapter 3 whenever available, but it is important to be aware of potential problems with certain measurement series. Ozone-sondes are normally tuned for measurements in the stratosphere and the values in the troposphere (where the ozone partial pressure is lower) may be inaccurate. It is important to be aware of potential errors, especially in the oldest measurements. The same goes for EMEP measurements where the first data sets (from the seventies) are done before the EMEP program was formally initiated (the data are from the OECD project on Long Range Transport of Air Pollution) and SOPs were defined. The RETRO data are furthermore sometimes an extraction of the most essential parts of the dataset, and e.g. the flags in the EMEP data are not fully preserved in the conversion process. Data marked as erroneous are not converted, but there might be flags describing unusual (but still valid) data. An example is flag no. 147: “Below theoretical detection limit or formal Q/A limit, but a value has been measured and reported and is considered valid”. It may sometimes be necessary to use the original dataset in studies that require detailed comparison of time series since the RETRO data does not include all available metadata.

In addition to instrument error and information about measurement accuracy (including biases, drift, calibration methods, etc.) there is always a large uncertainty regarding representativeness when comparing model output with observational data. While the model output normally is representative for a volume of the atmosphere the size of one grid-cell, the observation may be representative for a completely different air mass. The EMEP measurement network was designed to be representative for regional back-ground air masses or rural areas and is not expected to be representative for major cities or industrial districts. A comparison between EMEP data and e.g. a global model may therefore not be relevant for all the stations since the global model rarely is able to reproduce the complex emission and transformation patterns of all European regions. It is therefore necessary to choose the groundbased measurements that are comparable to the model and it may furthermore be necessary to spatially or temporally aggregate the datasets. Similar problems or challenges are present with the use of the other datasets in the RETRO database and it is necessary to understand what the measurements represent and what the model simulate when comparisons are being made. The situation may be different from one component to another or from one site to another.

The use of satellite data in model validation impose a problem related to that of representativeness. Even though the satellite data represent a volume of air of comparable size as the model, the measurements are commonly based on a series of assumptions that may cause problems. Nadir viewing instruments like GOME and SCIAMACHY use e.g. a priori profiles in the retrieval process for trace gases. This means guessing how the trace gas is distributed in the atmosphere and the estimate is commonly based on climatologies and/or a few measurements from other instruments. This normally works out fine, but there is a risk that the a priori profile is incorrect and will cause an error in the total column values. The satellite has usually lower sensitivity in the lower troposphere than in the free troposphere or above, and the retrieval process is tuned towards this based on the a priori information. The issue of sensitivity is a problem common to all remote sensing instruments and is a major contribution to the uncertainty in ground-based, airborne and satellite-borne measurements. Such measurements must therefore be used with care when comparing the data with simulated values. In the case of tropospheric measurements, it is important to understand how the data are being processed (there are e.g. several alternatives for subtracting stratospheric contributions) and to realise what the limitations are.

As was discussed in chapter 5.2, there is a need for expert scientists on observations to provide harmonised time-series and aggregated datasets for model validation purposes. Use of observations today require some degree of understanding of measurement techniques, limitations, biases, inaccuracies, representativeness and sensitivity. A natural continuation of the work on observations in RETRO would be that modellers increase their knowledge on measurements while those involved in measurements work on collation of data into value added supersets from existing data.

7 References

Beekmann, M., Ancellet, G., Megie, G., Smit, H.G.J. and Kley, D. (1994) Intercomparison campaign of vertical ozone profiles including electrochemical sondes of ECC and brewer-mast type-a ground-based UV-differential absorption Lidar. *J. Atmos. Chem.*, *19*, 259-288.

Burrows, J.P., Buchwitz, M., Eisinger, M., Rozanov, V., Weber, M., Richter, A. and Ladstaetter-Weissenmayer, A. (1997) The Global Ozone Monitoring Experiment (GOME): Mission, instrument concept, and first results (O_3 and NO_2). Paper presented at 3rd ERS Scientific Symposium, 17-21 March 1997, Space at the service of our Environment, ESA, Florence, Italy. URL: http://earth.esa.int/workshops/ers97/papers/burrows1/index.html

Burrows, J.P. and Chance, K.V. (1993) SCIAMACHY and GOME: The scientific objectives. In: *Optical Methods in Atmospheric Chemistry*, *22-24 June 1992, Berlin.* Ed. By H.I. Schiff and U. Platt. Bellingham, Wash., SPIE (EurOpto series Proceedings, vol. 1715) pp. 502-511.

Calisesi, Y., Stubi, R., Kampfer, N. and Viatte, P. (2003) Investigation of systematic uncertainties in Brewer-Mast ozone soundings using observations from a ground-based microwave radiometer. *J. Atmos. Oceanic Technol.*, *20*, 1543-1551.

Chance, K., Palmer, P.I., Spurr, R.J.D., Martin, R.V., Kurosu, T.P. and Jacob, D.J. (2000) Satellite observations of formaldehyde over North America from GOME. *Geophys. Res. Lett.*, *27*, 3461-3464.

De Backer, H., De Muer, D. and De Sadelaer, G. (1998) Comparison of ozone profiles obtained with Brewer-Mast and Z-ECC sensors during simultaneous ascents. *J. Geophys. Res.- Atmos.*, *103D*, 19641-19648.

Eisinger, M., and J. P. Burrows (1998), Tropospheric sulfur dioxide observed by the ERS-2 GOME instrument. *Geophys. Res. Lett.*, *25*, 4177-4180.

GEO (2005) Global Earth Observation System of Systems GEOSS. 10 year implementation plan reference document. Noordwijk, Group on earth observations (ESA SP-1284) (GEO 1000R).
URL: http://www.earthobservations.org/docs/10-Year%20Plan%20Reference%20Document%20(GEO%201000R).pdf

Hegels, E., Crutzen, P.J., Klupfel, T., Perner, D. and Burrows, J.P. (1998) Global distribution of atmospheric bromine-monoxide from GOME on earth observing satellite ERS-2. *Geophys. Res. Lett., 25*, 3127-3130.

Hilsenrath, E., Attmannspacher, W., Bass, A., Evans, W., Hagemeyer, R., Barnes, R.A., Komhyr, W., Mauersberger, K., Mentall, J., Proffitt, M., Robbins, D., Taylor, S., Torres, A. and Weinstock, E. (1986) Results from the balloon ozone intercomparison campaign (BOIC). *J. Geophys. Res.- Atmos., 91D,* 3137-3152.

Hoogen, R. Rozanov, V.V. and Burrows, J.P. (1999) Ozone profiles from GOME satellite data: Algorithm description and first validation. *J. Geophys. Res.- Atmos., 104D,* 8263-8280.

IGACO (2004) The changing atmosphere. An integrated global atmospheric chemistry observation theme for the IGOS partnership. Report of the Integrated Global Atmospheric Chemistry Observation Theme Team. Noordwijk, European Space Agency (ESA SP-1282; GAW No. 159; WMO TD No. 1235). URL: http://ioc.unesco.org/igospartners/docs/theme_reports/IGACO-ThemeReport-Sep2004.pdf

Johnson, B.J., Oltmans, S.J., Vomel, H., Smit, H.G.J., Deshler, T. and Kroger, C. (2002) Electrochemical concentration cell (ECC) ozonesonde pump efficiency measurements and tests on the sensitivity to ozone of buffered and unbuffered ECC sensor cathode solutions. *J. Geophys. Res.- Atmos., 107D,* 4393.

Komhyr, W.D. and Harris, T.B. (1965) Note on flow rate measurements made on Mast–Brewer ozone sensor pumps. *Mon. Wea. Rev., 93,* 267-268.

Lang, P.M., Steele, L.P., Martin, R.C. and Masarie, K.A. (1990a) Atmospheric methane data for the period 1983-1985 from the NOAA/GMCC global cooperative flask sampling network. Boulder, CO. (NOAA Technical Memorandum ERL CMDL-1).

Lang, P.M., Steele, L.P. and Martin, R.C. (1990b) Atmospheric methane data for the period 1986-1988 from the NOAA/CMDL global cooperative flask sampling network. Boulder, CO. (NOAA Technical Memorandum ERL CMDL-2).

Lemoine, R. and De Backer, H. (2001) Assessment of the Uccle ozone sounding time series quality using SAGE II data. *J. Geophys. Res.- Atmos., 106D,* 14515-14523.

Novelli, P.C., Collins, J.E., Myers, R.C., Sachse, G.W. and Scheel, H.E. (1994) Reevaluation of the NOAA/CMDL carbon monoxide reference scale and comparisons to CO reference gases at NASA-Langley and the Fraunhofer Institut. *J. Geophys. Res.- Atmos., 99D*, 12,833-12,839.

Novelli, P.C., Elkins, J.W. and Steele, L.P. (1991) The development and evaluation of a gravimetric reference scale for measurements of atmospheric carbon monoxide. *J. Geophys. Res.- Atmos., 96D*, 13,109-13,121.

Novelli, P.C., Masarie, K.A. and Lang, P.M. (1998) Distributions and recent changes in carbon monoxide in the lower troposphere. *J. Geophys. Res.- Atmos., 103D*, 19,015-19,033.

Novelli, P.C., Masarie, K.A., Lang, P.M., Hall, B.D., Myers, R.C. and Elkins, J.W. (2003) Re-analysis of tropospheric CO trends: Effects of the 1997-1998 wildfires. *J. Geophys. Res.- Atmos., 108D*, 4464. doi: 10.1029/2002JD003031.

Novelli, P.C., Steele, L.P. and Tans, P.P. (1992) Mixing ratios of carbon monoxide in the troposphere. *J. Geophys. Res.- Atmos., 97*, 20,731-20,750.

Nüß, H. (2005) Verbesserung des troposphärischen NO_2-Retrievals aus GOME und SCIAMACHY-Daten. Bremen, Bremen Universität. (In German.)

Richter, A., Wittrock, F., Eisinger, M. and Burrows, J.P. (1998) GOME observations of tropospheric BrO in northern hemispheric spring and summer 1997. *Geophys. Res. Lett., 25*, 2683-2686.

Smit, H.G.J. and Straeter, W. (2004) JOSIE-2000, Jülich Ozone Sonde Intercomparison Experiment 2000, The 2000 WMO international intercomparison of operating procedures for ECC-ozone sondes at the environmental simulation facility at Jülich. Geneva, World Meteorological Organization (WMO Global Atmosphere Watch report series, No. 158 (Technical Document No. 1225)).
URL: http://www.fz-juelich.de/icg/icg-ii/datapool/page/337/JOSIE-2000-GAW-Report-Nr-158.pdf

Steinbrecht, W., Schwarz, R. and Claude, H. (1998) New pump correction for the Brewer–Mast ozone sonde: Determination from experiment and instrument intercomparisons. *J. Atmos. Oceanic Technol., 15,* 144-156.

Tarasick, D.W., Davies, J., Anlauf, K., Watt, M., Steinbrecht, W. and Claude, H.J. (2002) Laboratory investigations of the response of Brewer-Mast ozonesondes to tropospheric ozone. *J. Geophys. Res.- Atmos., 107D, 4308*.

Tarasick, D.W., Fioletov, V.E., Wardle, D.I., Kerr, J.B. and Davies, J. (2005) Changes in the vertical distribution of ozone over Canada from ozonesondes: 1980-2001. *J. Geophys. Res.- Atmos., 110D,* D02304.

Wagner, T. and Platt, U. (1998) Satellite mapping of enhanced BrO concentrations in the troposphere. *Nature, 395,* 486-490.

Wittrock, F. (2006) The retrieval of oxygenated volatile organic compounds by remote sensing techniques. Bremen, University of Bremen.

Wittrock, F., Richter, A. and Burrows, J.P. (1999) Validation of GOME BrO and OClO measurements in the northern hemisphere. In: *Proceedings of the European Symposium on Atmospheric Measurements from Space (ESAMS 99), Jan. 18-22, ESTEC, Noordwijk, Netherlands.* Noordwijk, European Space Agency (ESA-WPP-161) pp. 735-738.

Wittrock, F., Richter, A., Ladstätter-Weissenmayer, A. and Burrows, J.P. (2000) Global observations of formaldehyde. In: *Proceedings of the ERS-ENVISAT symposium, Gothenburg October 2000.* Noordwijk, European Space Agency (ESA publication SP-461).

Appendix A

Databases mentioned in the report

The current report discuss various databases containing atmospheric chemistry observations. A brief overview of these archives are given in the table below. The report have not aimed for completeness and not all know data centres are described in the document or in the table. Focus is furthermore given to global scale tropospheric chemistry and data that are of direct relevance to RETRO. A more comprehensive overview of data archives and on-line monitoring networks is available through the ACCENT access to field data portal, http://www.accent-network.org. Other portals of interest is the ESA PROMOTE web pages, http://www.gse-promote.org, that provides access to satellite data products and the NASA Global Change Master Directory, http://gcmd.nasa.gov/.

Name	Address	Species	Availability	Focus	Section
RETRO	http://nadir.nilu.no/retro	O_3, SO_2, CO, HCHO	Open	Tropospheric chemistry	2 & 3
WOUDC	http://www.woudc.org	O3	Open	Stratospheric chemistry	3.2
MOZAIC	http://www.aero.obs-mip.fr/mozaic/	O_3, NO_x, CO, ++	Data protocol	Tropospheric and UTLS chemistry	4
IAGOS	http://www.fz-juelich.de/icg/icg-ii/iagos	O_3, NO_x, CO, ++	Data protocol	Tropospheric and UTLS chemistry (MOZAIC continuation)	4
Envisat Cal/Val	http://nadir.nilu.no/calval	O_3, NO_x, ++	Data protocol	Satellite validation	2, 5.2.1
EMEP	http://www.nilu.no/projects/ccc	O_3, NO_x, SO_2, PM, ++	Open	Transport of regional-scale air pollution	3.1
ETH-MEG	http://www.megdb.ethz.ch/	O_3, NO_x, ++	Data protocol	UTLS model validation. (TRADEOFF database)	4, 5.2.1
CMDL (GMD)	http://www.cmdl.noaa.gov/	O_3, CO, ++	Open	Global Atmospheric Monitoring	3.3
WDCGG	http://gaw.kishou.go.jp/wdcgg.html	O_3,CH_4, CO_2, ++	Open	Climate gas monitoring	3.5
CARIBIC	http://www.caribic-atmospheric.com/	O_3, CO, NO_x, HCHO,++	Data protocol	Global Atmospheric Composition	4
AURA Cal/Val	http://avdc.gsfc.nasa.gov	O_3, NO_x, ++	Data protocol	Satellite validation	4
AERONET	http://aeronet.gsfc.nasa.gov/nasa_menu.html	AOD	Open	Aerosol climatology	4
EARLINET-ASOS	http://www.earlinet.org	Aerosol properties	Data protocol	Vertical Aerosol climatology	4
EUSAAR	http://www.eusaar.org	Aerosol properties	Open	Groundbased in-situ aerosol monitoring and research	4
NDSC	http://www.ndsc.ncep.noaa.gov/	O_3, NO_2, Aerosols, ++	Open	Stratospheric chemistry	5.2.1
NADIR	www.nilu.no/datacentre	O_3, NO_x, ++	Data protocol	Atmospheric chemistry	4, 5.2.1
CREATE	http://www.nilu.no/projects/ccc/create/database.htm	Aerosol properties	Open	Groundbased in-situ aerosol monitoring and research	5.2.5
DAEDALUS	http://www-loa.univ-lille1.fr/Daedalus/data.htm	Aerosol properties	Open	Aerosol climatology	5.2.5
TOR-1	http://www.nilu.no/projects/tor/	O_3,++	Open	Tropospheric photo-oxidants	4
AEROCOM	http://nansen.ipsl.jussieu.fr/AEROCOM/	Aerosol properties	Open	Global aerosol distribution and Climate	5.1

Norwegian Institute for Air Research (NILU)
P.O. Box 100, N-2027 Kjeller, Norway

REPORT SERIES TECHNICAL REPORT	REPORT NO. TR 7/2006	ISBN 82-425-1798-3 ISSN 0807-7185	
DATE	SIGN.	NO. OF PAGES 43	PRICE NOK 150.-
TITLE Report on the RETRO database for observations and recommendations for optimal use of the data (D2-2)		PROJECT LEADER Aasmund Fahre Vik	
		NILU PROJECT NO. U-103032	
AUTHOR(S) Aasmund Fahre Vik, Sjur Bjørndalsæter, Leif Backman, Johannes Staehelin, Folkard Wittrock		CLASSIFICATION * A	
		CONTRACT REF. EVK2-CT-2002-00170	

REPORT PREPARED FOR
RETRO, EU project

ABSTRACT
This report describes the RETRO database for observational data. The system architecture and functionality is explained and a description of the data content is given. The report includes reference to other data centres and sources and a chapter on strategies for future atmospheric data centres.

NORWEGIAN TITLE

Beskrivelse av databasen for observasjoner i EU prosjektet RETRO og veiledning om optimal bruk av data

KEYWORDS		
Observations	Era40	Database

ABSTRACT (in Norwegian)
Denne rapporten beskriver RETRO databasen for observasjoner. Systemets arkitektur og funksjonalitet forklares og en oversikt over innholdet av data blir gitt. Rapporten inneholder også en veiledning om optimal bruk av databasen.

** Classification*	*A*	*Unclassified (can be ordered from NILU)*
	B	*Restricted distribution*
	C	*Classified (not to be distributed)*