

Report on coordinated NRT in-situ data supply and recommendations for AQ flagging systems

*Aasmund Fahre Vik and Leonor Tarrasón,
December 6th 2011*

This deliverable report is divided in two parts. The first part summarizes recommendations for a coordinated NRT in-situ air quality data supply according to the needs identified by the MACC project. The first part of this deliverable was completed as a separate note report already in August 2010 and has formed the basis for interaction and cooperation with the European Environment Agency (EEA) on GMES in-situ coordination. The note is attached here in Appendix I for the sake of completeness. The second part describes the need a harmonized system for quality control and flagging of air quality (AQ) measurement data, summarizes current practices and provides a set of recommendations on data flagging as seen from the MACC perspective and air quality modeling viewpoint.

1. Coordinated NRT in-situ data supply

Access to NRT in-situ data is essential to the services and products developed by MACC as pre-operational GMES atmospheric service. MACC supports the work of the EEA as GMES in-situ coordination and has contributed to the design of an operational long-term sustainable system for in-situ NRT data and exchange by identifying data providers and specifying the general requirements on these data. The requirements on a coordinated NRT in-situ data supply by MACC have been specified in terms of timeliness, operationality, harmonization, data parameters, coverage, data characterization and data quality. These are better described in Appendix I.

Air Quality observational data are used a) for model validation to provide a quantitative measure of model performance, b) for data assimilation in order to improve analyses and forecasts and c) for process studies to better understand and improve the model systems. For these purposes, the observational data are needed in:

- Near Real Time, NRT, (preliminary data with no or only automatic quality control that are made available only hours after an observation is complete)
- Delayed mode (preliminary data with some level of manual quality control applied that are made available days or weeks after an observation is complete)
- Offline mode (validated data that are typically made available more than a year after an observation is complete)

In order to use these data effectively, it is necessary to know the quality and the level of applied quality control of the observations. This knowledge will enable MACC developers to associate

uncertainties or probability distributions to the data and thereby know to which degree an observational value can be trusted and how large weight it should be given in a calculation chain. MACC, in return, can provide an evaluation of the uncertainty associated to different observation data that can help other modeler users of the same data. A coordinated and harmonized process to define metadata standards and quality control flagging procedures useful to the operational GMES atmospheric service needs to be put in place.

The current characterization of NRT in-situ data is insufficient for the purposes of data assimilation. The main reason is that current metadata standards do not include a relevant representation of the uncertainty associated to the observation. MACC has initiated a process to identify the necessary extension of existing metadata standards to make it useful for the purposes of an operational GMES Atmospheric Service.

There are currently metadata standards for in-situ data as under the INSPIRE directive, but these are far from complete. MACC envisages the distinction of three different levels of in-situ data characterization (metadata) that may also involve different ways of transmission and storage.

- The first type of metadata describes the position of the observation site. In most cases, except for aircraft observations, this is a static representation of the site and could be stored and transmitted separately from the actual data.
- The second type of metadata indicates the representativeness of the observation. This is presently insufficiently characterized in most metadata standards. It should involve in particular the representation of nearby sources, the topography around the site and meteorological information. Such characterization would allow the creation of error covariance matrices necessary for data assimilation purposes, and also would enable better use to be made of the data for validation. This information may change in time, but probably not from hour to hour. Therefore, storage and transmission of this information may occur at relevant intervals together with the actual data. The MACC INSITU team has carried out a study for the characterization of station representativeness for the different components that can be used at European level. The feedback from MACC users on the uncertainty of the observation could be included as part of this second type of metadata information.
- The third type of metadata characterizes the instrument and method used to carry out the actual observation. This includes for instance the factors used for PM mass observations. This information may change in time and therefore should be stored and transmitted together with the actual data.

It is recognized that the definition of new metadata standards is a process that will demand time. MACC and MACC-II are envisaged as pilot programs to test the new metadata standards, however, the adoption and implementation of such standards for in-situ data demands coordinated action beyond the GMES atmospheric service.

In addition to the metadata description the actual observations, either NRT or in delayed mode, need to be characterized in terms of their quality. The current report deals with quality control/assurance and flagging of actual observation data. Current practices are summarized below and further recommendations on the air quality flagging are given at the end of the note.

2. Current practices for AQ data flagging

Current data quality flagging systems have been developed by the observational community, as their significance relies on detailed knowledge of data, instruments and methods used for the observation. However there is limited coordination on the data flagging procedures and the result is a large variety of quality flag systems.

The experience gathered during the MACC project in the INSITU subproject indicates that air quality data are typically submitted to EEA/Airbase with no other flags than a valid/invalid notification. At the national level, the situation is different, but not harmonized throughout Europe. More information about measurement quality is normally available for the data owners, even for NRT data, but this is normally not transmitted to the central European database.

Through the GEMS and MACC (I) project periods, it has been observed that the use of validity flags varies from country to country. When comparing the reported NRT data with the validated dataset from the same station (comparing the data submitted to ECMWF in NRT with data from Airbase) the data capture sometimes appears to be different for the same time series. There seem to be different practices from one country to the other on how to define the data capture level – some simply remove erroneous data from the time series (and thereby reduce the data capture level) while others replace the measurement with a corrected value. A new flagging system should reflect what type of practice is followed and identify the reason for the corrected values.

Another issue related to data capture is the use of modeled or interpolated data to fill gaps in observational time series. Experience in the MACC INSITU project indicates that some data providers report interpolated data as validated measurements, although this information is not available for the data user. For use in model validation or data assimilation it is important to know if data are true observations or if it is not. In any case, the use of interpolated data is not recommended by MACC and such data needs to be identified and flagged for caution in further use.

Some countries submit NRT data that are pre-calibrated, e.g. contain a calibration value from the previous instrument calibration while some other submit un-calibrated values. The same goes for online measurements of PM mass which sometimes contains a correction factor and sometimes not. Again, a new flagging system should reflect this practice and identify if data has been corrected/pre-calibrated or not.

Procedure for submission of NRT data to EEA - P7141a, version 2.7

EEA provides a description of how data providers should submit NRT Air Quality data to the Agency through their P7141a document. It is available here:

http://eea.eionet.europa.eu/Public/irc/eionet-circle/airclimate/library?!=/public/real-time_operational/real-time_maintenance/p7141a_additional/EN_2.5_&a=d.

The document provides guidance for Air Quality data exchange of any component to EEA, but is targeted towards reporting of the primary components O₃, SO₂, NO₂, PM₁₀, PM_{2.5}, etc (these five are specifically mentioned). The reporting of particulate matter is specifically treated and it is emphasized that data must be submitted using a correction factor to make data comparable to the standard measurement method (gravimetric method).

The procedure describes the required data submission formats (CSV or XML formatting accepted) and submission methods. There are dedicated fields for QA/QC information to describe if the data are valid or invalid (QA) and if the data have been validated/pre-validated or not (QC). Both fields are optional. For submission of data in CSV format the two flags can only be set for a given day, while for XML files the flags may be set for each hourly value.

The system that is suggested in the EEA document/procedure P7141a seems reasonable, but possibly not sufficient. The EEA method allows data providers to specify whether a measurement is valid or not (QA) and if the data is fully preliminary or not (QC). These flags are, however, voluntary and to an unknown degree used by the data providers.

Implementing Provisions

The implementing provisions of the Air Quality Directive 2008/50/EC (AQD) constitute the starting point for a transition to an updated and modernised system of electronic reporting and data exchange for air quality (e-Reporting). A mechanism will be established and made operational for sharing the information to be provided by the countries, in accordance with the SEIS principles and the INSPIRE directive. Existing requirements to supply GMES in-situ data are expected to be taken in to account and considered in the design of the system. Therefore it is important that MACC defines its requirements for the AQ quality flagging at this stage.

The AQD implementing provisions (IPR) will apply at the end of the 2-year transitional period, commencing after adoption of the IPR which has been expected at the end of 2011. It is currently (mid December 2011) not known if the IPR will be accepted or not.

The IPR relates not only to reporting of NRT AQ data, but to all AQ data flows and a main objective is to harmonize them better. The main changes to the AQD reporting business logic under the emergent IPR relates to;

1. A new specification of the reporting format, data now to be conveyed in XML (rather than ASCII, DEM, ISO, NASA Ames, spreadsheet based etc)
2. A reorganisation of the management of data flows
3. A reorganisation of the frequency of reporting for some data flows
4. Additional data flows now exist for information items for which there were previously no structure within EU reporting, but for which there were successful voluntary agreements for exchange of NRT ozone (OzoneWeb) and other AQ parameters.

Importantly, the new IPR suggest an expansion of the flagging system. In the new IPR the flagging system will probably be extended to:

- valid
- valid, but value is below detection limit, number not replaced
- valid, but value is below detection limit , number replaced by $0.5 \times$ detection limit
- not valid due to station maintenance or calibration
- not valid due to other reasons or missing

This will be an improvement compared to the current reporting system, but an even more elaborated system could have been useful for GMES purposes.

EMEP quality assurance system

Data submitted to EMEP are generally treated differently than those submitted to EEA or AIRBASE (even though the same measurement may be submitted both places). The EMEP quality assurance procedures feature a comprehensive flagging system that enables data providers (and EMEP data centre personnel) to describe in details the quality of the data. The procedures are described here: <http://www.nilu.no/projects/ccc/qa/index.htm>. The complete list of available flags is provided here: <http://www.nilu.no/projects/ccc/flags/index.html>. This flagging system is probably too complex for practical use in model validation, but it is possible to use it in a simplified way. All flags have a valid/invalid interpretation and a user could easily convert the numbers into a simpler quality assurance flag (valid/invalid).

The EMEP quality assurance system is primarily developed for validated data, but is also used for the EMEP NRT data.

3. MACC Recommendations for a harmonized AQ flagging system

The operational GMES atmospheric service will require an accurate, transparent and traceable system to establish the quality of the NRT in-situ data. Current data quality flagging systems will have to be revised to allow the identification of deviations and routines on AQ data quality control.

The main recommendation from the MACC team is to adopt an air quality flagging system that is applicable to near real time in-situ data, data supplied in delayed mode and validated off-line data. Ideally in the future the supply and storage of NRT in-situ data will follow the same data flow systems as validated data to avoid unnecessary duplication. The envisaged AQ quality flagging system would need to be applied all these types of data.

Data may be flagged during any step in the validation chain: from raw data via NRT data to delayed mode and final validated data in offline mode, and different criteria for setting flags may apply in the different steps in the QA/QC process. It is therefore important to identify at which step in the process a quality flag is provided. This recommendation is particularly relevant in data flows like the EEAs NRT AQ reporting system, where NRT data are sometimes mixed with delayed mode data. At present

it is difficult to differentiate between data of different quality and the new flagging system should take into account at which step of the QA/QC system the specific flag applies.

The proposed air quality flagging system from the IPR Implementing Provisions of the Air Quality Directive 2008/50/EC (AQD) will probably introduce useful additional flags for Air Quality reporting. These new flags describe issues related to measurements close to instrument detection limit and calibration measurements. While these flags are not originating from the needs described by MACC, they are still useful for model validation/assimilation to specify the validity of the measurement. MACC welcomes the recommended flagging system from IPR, with some additional items.

MACC recommends the introduction of at least two additional data quality flags to the proposed IPR flagging system. The first one is to identify whether a correction or pre-calibration value has been added to the NRT data. The second one is to apply for validated data in order to identify whether the data series is the result of the use of interpolated or modeled data. With respect to this second issue, it should be mentioned that MACC INSITU team advises against the use of interpolations in the observation data series. Should such interpolation anyway take place, it is advisable to flag the data clearly as interpolated.

Additional QA/QC flags may be necessary and this possibility should be analyzed further in cooperation with EMEP for validated research data. The recommendation from MACC is to simplify the existing EMEP flagging system to enable a harmonized system also for NRT data.

The simple AQ flagging system described above would enable to identify where in the QA/QC chain the flag has been set. Consequently, the modeler/user of the data would be enabled to assign the specific uncertainty to the observation. MACC recommends that the uncertainty of the observation is set by the data user in a case-by-case approach adapted to the model and the specific use of the data. MACC recommends that the data provider restrains from providing lumped data uncertainty values. A harmonized transparent data flagging system as the one broadly described above will be the best way to characterize the quality of the measurement data.

Appendix I

MACC REQUIREMENTS FOR NEAR REAL TIME AIR QUALITY DATA EXCHANGE

Leonor Tarrasón
(document delivered to EEA on August 2010)

This note presents an overview of the general requirements for near real time (NRT) air quality data exchange from the pre-operational GMES Atmospheric Service, the FP7 funded MACC project, to the European Environment Agency, EEA, on its role as GMES in-situ data coordinator. The requirements are specified in terms of timeliness, operability, harmonization, data parameters, coverage, data characterization and data quality.

The MACC project recognizes and supports the work of the EEA as GMES in-situ coordinator. In cooperation with EEA, the MACC consortium seeks to contribute to the design of an operational long-term sustainable system for in-situ NRT data and exchange, useful for the GMES Atmospheric Service (GAS). The general requirements in this note are a first step towards the design of such sustainable system. MACC will continue to work to further specify these general requirements, identify additional data and links to networks that may be relevant for the GMES atmospheric service at regional and global level.

I. The use of NRT data in MACC

MACC (Monitoring Atmospheric Chemistry and Climate) is the current pre-operational atmospheric component of the European GMES program. MACC combines state-of-the-art atmospheric modeling with Earth observation and in-situ data to forecast and provide information on European Air Quality, Global Atmospheric Composition, Climate, and UV and Solar Energy.

Access to NRT in-situ data is essential to the services and products developed by MACC. The in-situ NRT atmospheric composition data is used in MACC for three main purposes. First, to produce better forecasts of environmental risk situations through data assimilation, second to validate the forecasts and the information provided and used by the service and third, to gain better understanding of the coupling between air pollution and weather flows, thus providing new insights into the atmospheric processes involved in climate change.

The pre-operational GMES atmospheric service MACC does not directly compile NRT in-situ data itself but relies on existing infrastructure and data exchange by national and international organizations. The data used by MACC are commonly acquired for other purposes, mainly for atmospheric monitoring or for research purposes. Therefore data reporting and exchange follow standards defined by the different monitoring and research networks. This is not always ideal for the use of the data in MACC. The following chapter describes specific requirements for in-situ data provision that are seen as essential for enabling use of the data within the project/service.

II. MACC requirements to NRT data

The general requirements for near real time data that follow below are based on the experience gathered under the GEMS, PROMOTE and MACC projects. These are in our current understanding the general requirements for the NRT in-situ data to serve the needs of an operational GMES Atmospheric Service.

1. Timeliness

In-situ NRT data in the operational GMES Atmospheric Service (GAS) should be delivered within an hour after the end of an observation.

In the initial stages of the service, “near-real time data” means data delivered preferably within 3 hours of observations, but possibly up to 24 hours behind time. MACC’s regional forecasts for Europe are currently run once daily, but the aim is to run using a synoptic cycle (at 00 GMT, 06 GMT, 12 GMT and 18 GMT) by the end of the project. In a possible continuation of the service (MACC-II) the aim would be for a 3-hourly cycle of data deliveries. The target is to reach hourly deliveries of NRT data by the time the GMES Atmospheric Service becomes operational. MACC’s global data assimilation system runs with a 12-hourly cycling, and is more tolerant of receiving data that do not meet the one-hour delivery time, although one-hour remains the ideal for the global system also.

2. Operationality (Operational data access)

The operational GMES Atmospheric service will require constant access to NRT in-situ data. This means that the data should be accessible to the GAS/MACC 24 hours per day and 7 days per week. This operational requirement for NRT in-situ data imposes serious constraints on the databases and networks that can be used as basis for the decentralized system of NRT data exchange envisaged by MACC. Not all the existing infrastructure compiling NRT atmospheric composition data can comply at present with this requirement of continuous data accessibility. In addition, the long-term sustainability of the GMES Atmospheric Service imposes also a requirement on the continuity of the databases and data repositories to be used by the operational service. This further limits the number of actual candidates to be in-situ data providers to the operational GMES atmospheric service and strengthens the need for coordinated action to support operational long-term NRT in-situ data compilation and exchange.

3. Harmonisation (Harmonised formats)

The transmission of in-situ NRT data in the operational GMES Atmospheric Service should be facilitated by the use of harmonized data formats.

GEMS and MACC have requested the transmission of in-situ NRT data in the BUFR format. This request was intended to make the NRT atmospheric composition data transmission in GMES compatible with the meteorological data transmission and thus facilitate the coupling of atmospheric composition with meteorological information. As the need for NRT data evolves in MACC and the operational GMES atmospheric service towards more sophisticated data types other than surface air

quality data, a limited number of additional formats should be allowed. The additional data formats should be international standards, with defined conventions. The selection of additional formats should be related to specific data types and justified by necessary constraints in the representation of the data. Possible additional formats that are presently under consideration are HDF, netCDF and NASA-AMES. Furthermore, data must use internationally acknowledged naming standards such as GEOMS (currently being defined for HDF data), netCDF-CF and the EMEP/EUSAAR standard for NASA-AMES 1001 formatted files.

4. Data parameters

The following table includes the requested data parameters and data types identified by the regional, global and climate change components of MACC as significant for the operational GMES atmospheric service. Recognizing that the degree of significance may change in time, the table below indicates our current understanding of the usefulness of the identified in-situ data for data assimilation, validation, calibration of satellite data and the process understanding purposes in MACC.

The identification of the actual data providers goes beyond the purpose of this note. Following the same classification as the data audit presently carried out by the ETC/ACC for the EEA :

- “ES” means that the in-situ data is essential for the creation of MACC products.
- “DS” means that the in-situ data is needed to enhance the quality of MACC products.
- ”UF” means that the in-situ data is useful for the creation of MACC products.

	Data assimilation	Validation Calibration	Process understanding
Surface data			
Ozone	ES	ES	UF
NO	ES	ES	UF
NO2	ES	ES	UF
SO2	ES	UF	UF
PM10 (mass)	ES	ES	UF
PM2.5 (mass)	ES	ES	UF
PM (speciation)	ES	ES	UF
NMVOG (speciated)	UF	DS	ES
CO	DS	ES	UF
CO2 (conc. and fluxes)	ES	ES	UF
CH4 (conc. and fluxes)	ES	ES	UF
Vertical Profiles			
Ozone	DS	ES	ES
NO2	DS	ES	ES
PM10 (mass)	DS	ES	ES
PM2.5 (mass)	DS	ES	ES
PM (speciation)	DS	ES	ES
CO2 (total column)	DS	DS	DS
Aircraft data			
Ozone	ES	ES	ES
NOx	DS	ES	ES
PM2.5	DS	ES	ES
CO2	DS	DS	DS
CH4	DS	DS	DS
Total column data			
AOD	ES	DS	DS
NO2	DS	ES	ES

CO2	DS	DS	DS
CH4	DS	DS	DS
Additional data			
PAN	UF	DS	DS
Isoprene	UF	DS	DS
Benzene	UF	UF	DS
UV index	UF	UF	DS
Halogenated species	UF	UF	DS
Stratospheric data	UF	UF	DS

NOTE: This table may be further updated!

5. Coverage

The geographical coverage of the in-situ data for the GMES atmospheric service is global. Therefore, NRT in-situ data collection should be carried out linking to existing infrastructure all over the globe. Coordinated co-operation efforts are necessary to publicize the in-situ data requirements and link to existing networks elsewhere in the globe.

6. Data characterization (Metadata)

The current characterization of NRT in-situ data is insufficient for the purposes of data assimilation. The main reason is that current metadata standards do not include a relevant representation of the uncertainty associated to the observation. MACC has initiated a process to identify the necessary extension of existing metadata standards to make it useful for the purposes of an operational GMES Atmospheric Service.

There are currently metadata standards for in-situ data as under the INSPIRE directive, but these are far from complete. MACC envisages the distinction of three different levels of in-situ data characterization that may also involve different ways of transmission and storage.

- The first type of metadata describes the position of the observation. In most cases, except for aircraft observations, this is a static representation of the site and could be stored and transmitted separately from the actual data.
- The second type of metadata indicates the representativeness of the observation. This is presently insufficiently characterized in most metadata standards. It should involve in particular the representation of nearby sources, the topography around the site and meteorological information. Such characterization would allow the creation of error covariance matrices necessary for data assimilation purposes, and also would enable better use to be made of the data for validation. This information may change in time, but probably not from hour to hour. Therefore, storage and transmission of this information may occur at relevant intervals together with the actual data.
- The third type of metadata characterizes the instrument and method used in the actual observation. This includes for instance the factors used for PM mass observations. This information may change in time and therefore should be stored and transmitted together with the actual data.

MACC recognizes that the definition of new metadata standards is a process that will demand time. Although MACC and a possible continuation (MACC-II) are envisaged as pilot programs to test the

new metadata standards, the adoption and implementation of such standards for in-situ data will demand coordinated action beyond the atmospheric service.

7. Data quality

The operational GMES atmospheric service will require an accurate, transparent and traceable system to establish the quality of the NRT in-situ data. Current data quality flagging systems will have to be revised to include routines that compile feedback on data quality from the GMES atmospheric service and screening systems to distinguish between NRT non-validated and validated data.

Current data quality flagging systems are developed by the observational community, as their significance relies on detailed knowledge of data, instruments and methods used for the observation. However there is limited coordination on the data flagging procedures and the result is a large variety of quality flag systems. These data flagging systems are often too complicated for the modeling community to make use of them. As the information on data quality is essential to MACC, the project will identify ways to increase the friendliness of data quality flagging systems for use by the modeling community.

The requirement on the accuracy and traceability of the quality of the data constrains the operational data storage and exchange system able to support the operational GMES atmospheric service. In line with the requirements under the INSPIRE Directive and the SEIS communication, MACC envisages a decentralized data exchange system, where the responsibility for storage and quality control of the data is as close as possible to the data providers. As indicated under point 2, the long-term sustainability of the operational service will determine the actual structure of the data exchange system.

A process is necessary at this stage to identify relevant data centers across the globe that can act as storage and transmission centers for the NRT in-situ data.

III. Expectation from EEA as GMES in-situ coordinator

The requirements specified above involve a series of processes where coordinated action is necessary. MACC expects that the EEA on its role as GMES in-situ coordinator will enable and support these processes by allocating resources, publicizing their results and eventually identifying the mechanisms and means for its implementation. In particular, MACC expects that EEA will enable and support:

- A process to supply European regulatory AQ NRT data to the operational GMES atmospheric service
- A process to identify a global network of relevant operative data centers that can supply NRT data other than regulatory to the operational GMES atmospheric service
- A co-ordination process to define metadata standards useful to the operational GMES atmospheric service
- A harmonization process to define quality flagging procedures useful to the operational GMES atmospheric service