Document title	QA/QC guidelines for EO data (draft)
Code	3-1
Category	CMNT
Agenda Item	3 – Progress of EUTRO-OPER during work phase 4
Submission date	24.8.2015
Submitted by	Finland, JRC, ICES, Secretariat
Reference	EUTRO-OPER 4-2015, 5-2015

Background

The EUTRO-OPER project has been tasked to prepare guidelines for QA/QC management of the assessment data, to be included as part of the 'EUTRO-OPER manual'. EUTRO-OPER 4-2015 agreed, that while the guidelines for discrete water sample data, prepared by ICES, are already operational, guidelines for the new data types should be prepared. The first version of QA/QC guidelines for EO-data was presented at EUTRO-OPER 5-2015.

Action required

The meeting is invited to review the draft and guide the work forward.

Guidelines for EO-based information

In this guideline, EO-based information is considered to be any information derived from satellite images for the use of updating indicators. The observations are in validated form, and may be aggregated temporally and spatially to a specific level.

The data must include a distinct time and position information. In the case of aggregated information, these may represent average values.

Receiving data

The Data Centres require the following information to be supplied by the data supplier together with the data. When receiving data, the Data Centres shall strive to meet the following guidelines.

Data standard

All satellite-derived products must be clearly specified and described. If product codes are to be used, then the source data dictionary consistency must be specified (e.g. CF Metadata Convention) . Product units must be clearly stated, and the algorithms used in the computations should be stated. The data should be fully checked for quality and pre-edited or flagged for erroneous values. An explicit statement should be made of the checks and edits applied to the data. A brief description, or a reference to the data collection and processing methods (e.g. reference to a specific technique or specific project protocols) must be included and should contain information regarding:

- Methods and procedures applied to the analysis of original raw data
- Methods / protocols and dataset(s) used for validation, or refer to their original source
- Internal or external quality assurance procedures (e.g. NASA, ESA protocols, QA4EO guidance[1])

A brief description of the data processing procedures must be included and should contain information regarding:

- editing/quality control methods
- how are trace values (values below the detection limit) identified
- how are missing values handled (null vs. zero, or “blanks”)
- what is the precision of the methods (e.g. number of significant figures)
- what units are used
- describe what quality flags are used if any
- supply a validation document

If a report is available describing the data collection and processing, this can be referenced. If possible a copy should be supplied with the data.

Format description

EO data and related metadata will be provided primarily via open and standard interfaces (INSIPIRE compatible format). Data format should be documented for example NetCDF-4 or INSPIRE compliant format. If in doubt about the suitability of any particular format, advice from the Data Centre should be sought. Individual fields, units, etc. should be clearly defined and time zone stated. Time reported in UTC is used. The contents of the data and ancillary information should adhere to the convention for CF (Climate and Forecast) metadata (http://cfconventions.org) or equivalent (e.g. Copernicus Marine Service).

[1] http://qa4eo.org

Collection and processing details

Pertinent information to be included in the data transfer to the Data Centre includes:

- Processing responsible: country, organisation, institute, PI
- Satellite instrument(s)
- Products derived from satellite data
- Details of the collection sensor
- Resolution of original data
- Algorithm and processing used for deriving product
- Atmospheric correction scheme and cloud masking
- Level of temporal and spatial aggregation used
    - spatial: either HELCOM assessment area or HELCOM 20 km grid
    - temporal: daily or annual assessment period
- Uncertainties on product estimates
- Date and time of the start and end of the sampling (UTC)
- Position estimate (latitude and longitude degrees and minutes or decimal degrees can be used. Explicitly state which format is being used. It is recommended that N, S, E and W labels are used instead of plus and minus signs.)

Any additional information of use to secondary users which may have affected the data or have a bearing on its subsequent use. For additional information on quality control procedures, metadata requirements for particular parameters and collection instrumentation, see CF Convention (http://cfconventions.org) .

Validation details

Validation is prerequisite to ensure the distribution of quantitative data products and their subsequent application by the user community. Information on the validation process of the provided data should be able to prove the reliability and consistency of satellite-derived products. Pertinent information includes:

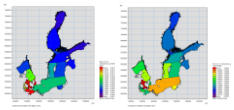
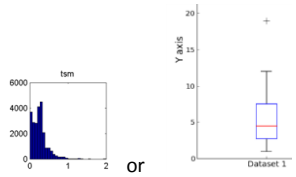
- Well documented validation protocol used (as an example see e.g. for ocean colour Mélin and Franz 2014 and MarCoast/CoBiOS validation protocols The validation protocol must be defined at later stage, as its requirements must be realistic for the infrastructure existing in the Baltic Sea..
- Detailed characteristics of the validation data, i.e. match-up data sets (in case of direct comparison between satellite product and contemporaneous and co-located in-situ measurements of the same quantity)
- Use of existing database (e.g. AERONET, Zibordi et al. 2006) and ICES for the chlorophyll-a.
- Uncertainties associated with field observations in case these are given (e.g. ICES).
- The data used for validation, its temporal and spatial coverage must be described and the validation procedure must be described. The validation must concern the Baltic Sea region.
- Validation metrics/statistics (e.g. number of match-ups, scatter and systematic difference or bias between the distributions, rms..) A table 1 below gives an example of validation metrics (to be defined in the final version of this document). Table 1 is adopted from previous EU/FP7-project CoBiOS deliverable 5.3&5.7.

Table 1. Table of statistical measures used to describe EO validation. Notations: n = number of observations, $\bar{X}$ = mean of variable X, σX = standard deviation of variable X, X = independent (in situ) data, Y = dependent (EO)

Commented [JA1]: The validation protocol must be defined at later stage, as its requirements must be realistic for the infrastructure existing in the Baltic Sea.

Commented [JA2]: these must be agreed, but the list in table is quite comprehensive. In the final version, the level of Optional and mandatory should be determined for each measure.

data, E = Y - X = Error. References L09 = Lehmann et al. 2009, A07 = Allen et al. 2007. Table continued on next page.

Statistical measure	Formula	Scope	Ref.
Descriptive statistics			
Maps of dependent and independent data or and/or time series plots		**Mandatory**	
Frequency distributions or boxplots	 or	Mandatory	
Scale	Linear	**Mandatory**	
Geometric mean (as a tribute to log-normal distributions)	$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}$	**As needed**	
Outlier detection			
Statistical outlier detection (alternatively by expert judgement)	Outlier range: < P25 – 3 * (P75 – P25) or > P75 + 3 * (P75 – P25), where P25 and P75 are the 25th and 75th percentile respectively	As needed	
Regression and correlation			
Regr. and corr. results	A, b, r, r^2, n, p (single sided)	**Mandatory**	
Error statistics			
Mean absolute error	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert Y_i - X_i \rvert$	Mandatory	L09
Bias	$Bias = \frac{1}{n}\sum_{i=1}^{n} (Y_i - X_i) = \bar{Y} - \bar{X}$	**Mandatory**	
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{\sum (y_i - x_i)^2}{n}}$	Mandatory	L09
Ratio of standard deviations	$\frac{\sigma_y}{\sigma_x}$	Optional	A07[2]
Percentage model bias *i.e.* model – data)/data	$Pbias = \frac{\sum_{i=1}^{n}(Y_i - X_i)}{\sum_{i=1}^{n} X_i}$	**Optional**	A07
Median error	50th percentile of the error distribution.	Optional	

Statistical measure	Formula	Scope	Ref.
Model efficiency (Nash Sutcliffe Model Efficiency)	$ME = 1 - \frac{\sum_{i=1}^{n}(Y_i - X_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$	**Mandatory**	A07
Skewness of error distribution	$s_0 = \frac{\sqrt{n(n-1)}}{n-2} \frac{\frac{1}{n}\sum_{i=1}^{n}(E_i - \bar{E})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i - \bar{E})^2}\right)^3}$	**Mandatory**	
Cost function (Normalized bias)	$CF = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Y_i - X_i \rvert}{\sigma_x}$	Optional	A07
Target diagram			
1.1 Standardized Bias	$Bias^* \frac{1}{n}\sum(y_i - x_i) \,/\, \sigma_x$	**Mandatory**	
1.2 Standardized unbiased RMSE	$RMSE^* = \sqrt{\frac{\sum((Y - \bar{Y}) - (X - \bar{X}))^2}{n}} / \sigma_x$	**Mandatory**	
2.1 Standardized Bias (Median)	$Bias^* \frac{1}{n}\sum(y_i - x_i) \,/\, Median_x$	Mandatory	
2.2 Standardized unbiased RMSE (Median)	$RMSE^* = \sqrt{\frac{\sum((Y - \bar{Y}) - (X - \bar{X}))^2}{n}} / Median_x$	Mandatory	

Value added service

When processing and quality controlling data, the Data Centres of the ICES community shall strive to meet the following guidelines.

Quality control

A range of checks are carried out on the data to ensure that they have been imported into the Data Centre's format correctly and without any loss of information. For discrete water sample data, these should include:

- Check header details / metadata (vessel, cruise number, station numbers, date/time, latitude/longitude
  (start and end), instrument number and type, station depth, cast (up and down) data type /no. of data points, platform identifier)
- Plot station positions to check not on land
- Automatic range checking of each parameter (e.g. WOD 1998, Maillard 2000)
- Check units of parameters supplied
- Check for spikes
- Flag suspicious data or correct after consultation with Principal Investigator (PI)

Problem resolution

The quality control procedures followed by the Data Centres will typically identify problems with the data and/or metadata. The Data Centre will resolve these problems through consultation with the originating PI or data supplier. Other experts in the field or other Data Centres may also be consulted.

History documentation

All quality control procedures applied to a dataset are fully documented by the Data Centre. As well, all quality control applied to a dataset should accompany that dataset. All problems and resulting resolutions will also be documented with the aim to help all parties involved; the Collectors, Data Centre, and Users. A history record will be produced detailing any data changes (including dates of the changes) that the Data Centre may make.

Request for support

When addressing a request for information and/or data from the User Community, the Data Centres shall strive to provide well-defined data and products. To meet this objective, the Data Centres will follow these guidelines.

Data description

The Data Centre shall aim to provide to its clients well-defined data or products. If digital data are provided, the Data Centre will provide sufficient self-explanatory series header information and documentation to accompany the data so that they are adequately qualified and can be used with confidence by scientists/engineers other than those responsible for their original collection, processing and quality control. This is described in more detail below:

- A data format description fully detailing the format in which the data will be supplied
- Parameter and unit definitions, and scales of reference
- Definition of additional quality control
- Flagging scheme, if flags are used
- Data history document (as described below)
- Accompanying data

Data history

A data history document will be supplied with the data to include the following:

- A description of data collection and processing procedures as supplied by the data
- collector (as specified earlier)
- Quality control procedures used to check the data (as specified earlier)
- Any problems encountered with the data and their resolution and modification date
- Any changes made to the data and dates of these changes

Any additional information of use to secondary users which may have affected the data or have a bearing on its subsequent use should also be included.

References

Allen J. I., Holt J. T., Blackford J., Proctor R. (2007a): Error quantification of a high-resolution coupled hydrodynamic ecosystem coastal-ocean model: Part 2. Chlorophyll-a, nutrients and SPM. Journal of Marine Systems 68, 381–404.

Allen J. I., Somerfield P. J., Gilbert F. J. (2007b): Quantifying uncertainty in high-resolution coupled hydrodynamic-ecosystem models. Journal of Marine Systems 64, 3–14.

Lehmann M. K., Fennel K., He R. (2009): Statistical validation of a 3-D bio-physical model of the western North Atlantic. Biogeosciences, 6, 1961–1974.

Mélin F., and B.A. Franz (2014). Assessment of satellite ocean colour radiometry and derived geophysical products. In G. Zibordi, G.J. Donlon, and A.C. Parr (eds.) Optical Radiometry for Ocean Climate Measurements. Chap. 6.1 Vol. 47 Experimental Methods in the Physical Sciences. Elsevier Inc.

Zibordi G. Et al. (2006). A network for standardized ocean colour validation measurements. EOS Trans. AGU 87: 293-297.