Annex B

Guidelines for describing EO-based information

This Annex is updated version of the guidelines for describing EO-based information given in HELCOM (2015). 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.

EO data is collected in combination with *in situ* monitoring which is used to validate the EO-based chlorophyll values (see section 'Validation details' below).

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.

[1] http://qa4eo.org

Format description

EO data and related metadata will be provided primarily via open and standard interfaces (INSIPIRE compatible format).
Data format, in case individual observation data is provided, 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).

Collection and processing details

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

- Processing responsible: country, organization, institute, PI
- Satellite instrument(s)
- Products derived from satellite data
- Details of the collection sensor
- Resolution of original data
- Algorithm and data 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.)

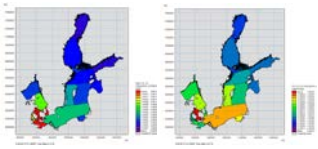
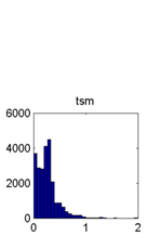
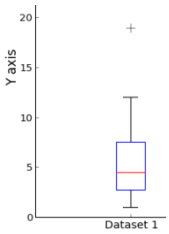
Any additional information of use to secondary users which may have affected the data or have a bearing on its subsequent use. For additional information on quality control procedures, metadata requirements for particular parameters and collection instrumentation, see CF Convention (http://cfconventions.org).

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 validation protocols). 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 should be given or referred to accepted publication (e.g. Table 1, number of match-ups, scatter and systematic difference or bias between the distributions, rms).

Table 1. Table of statistical measures used to describe EO validation. Notations: n = number of observations, ¯X = mean of variable X, σX = standard deviation of variable X, X = independent (*in situ*) data, Y = dependent (EO) data, E = Y - X = Error. References L09 = Lehmann et al. 2009, A07 = Allen et al. 2007. Table continued on next page. Adopted from EU/FP7-project CoBiOS deliverable 5.3&5.7.

Statistical measure	**Formula**	**Ref.**
Descriptive statistics		
Maps of dependent and independent data or and/or time series plots		
Frequency distributions or boxplots	 or 	
Scale	Linear	
Geometric mean (as a tribute to log-normal distributions)	$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}$	
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	
Regression and correlation		
Regr. and corr. results	A, b, r, r², n, p (single sided)	
Error statistics		
Mean absolute error	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert Y_i - X_i \rvert$	L09
Bias	$Bias = \frac{1}{n}\sum_{i=1}^{n} (Y_i - X_i) = \bar{Y} - \bar{X}$	
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{\sum (y_i - x_i)^2}{n}}$	L09
Ratio of standard deviations	$\frac{\sigma_y}{\sigma_x}$	A07[2]

Statistical measure	**Formula**	**Ref.**
Percentage model bias *i.e.* model – data)/data	$Pbias = \frac{\sum_{i=1}^{n}(Y_i - X_i)}{\sum_{i=1}^{n} X_i}$	A07
Median error	50th percentile of the error distribution.	
Model efficiency (Nash Sutcliffe Model Efficiency)	$ME = 1 - \frac{\sum_{i=1}^{n}(Y_i - X_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$	A07
Skewness of error distribution	$s_0 = \frac{\sqrt{n(n-1)}}{n-2} \frac{\frac{1}{n}\sum_{i=1}^{n}(E_i - \bar{E})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i - \bar{E})^2}\right)^3}$	
Cost function (Normalized bias)	$CF = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Y_i - X_i \rvert}{\sigma_x}$	A07
Target diagram		
1.1 Standardized Bias	$Bias^* \frac{1}{n}\sum (y_i - x_i) \,/\, \sigma_x$	
1.2 Standardized unbiased RMSE	$RMSE^* = \sqrt{\frac{\sum((Y - \bar{Y}) - (X - \bar{X}))^2}{n}} / \sigma_x$	
2.1 Standardized Bias (Median)	$Bias^* \frac{1}{n}\sum (y_i - x_i) \,/\, Median_x$	
2.2 Standardized unbiased RMSE (Median)	$RMSE^* = \sqrt{\frac{\sum((Y - \bar{Y}) - (X - \bar{X}))^2}{n}} / Median_x$	

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.