DUE GlobBiomass

D5

Validation Protocol

Prepared for European Space Agency (ESA-ESRIN)

In response to ESRIN/Contract No. 4000113100/14/I_NB

Wageningen University and Research Centre, Laboratory of Geoinformation Science and Remote Sensing, The Netherland

December 2015

Revision History

Deliverable	D5, VP		
Work Package	2000		
Due date	KO+12		
Authors	Valerio Avitabile, Heiko Balzter, Sytze de Bruin, Joao Carreiras, Nuno Carvalhais, Shaun Quegan, Kevin Tansey, Pedro Rodriguez-Veiga		
Distribution	FSU: Christiane Schmullius, Evelin Matejka		ESA: Frank Martin Seifert; Nathalie Boisard
Reason for change			
Issue			
Revision			
Date			
Release	1		
Version	01		

WAGENINGEN UNIVERSITY
WAGENINGEN UR

University of Leicester

The University Of Sheffield.

Max Planck Institute for Biogeochemistry

Contents

1. Introduction

The Validation Protocol is aimed at providing a common framework for assessing and reporting the accuracy of the GlobBiomass products, namely the regional and global biomass maps, and evaluate the user acceptance of these products. The validation and user assessment are essential for providing high-quality products that are accepted and applied by the user community. While a standardized approach recognized by the international community for validating large area biomass products is not yet available, significant experience regarding validation concepts and procedures have been developed in other communities, and the current validation task will be based on the lessons learned from these previous projects. Ultimately, the accuracy assessment is aimed to build international confidence in the GlobBiomass products and their adoption by the user community, and possibly establish a precursor for permanent and operational biomass validation activities.

The Validation Protocol provides common metrics and procedures to report uncertainty statistics and maps in a standardized way, allowing direct comparison among the products. The Validation protocol follows the 2006 IPCC Good Practice Guidelines (IPCC, 2006) and is in line with concepts and procedures developed in the Land Product Validation (LPV) sub-group of the CEOS Working Group on Calibration and Validation (http://lpvs.gsfc.nasa.gov/).

The GlobBiomass products, which include a global and four regional biomass datasets, are characterized by large differences in terms of data available (both remote sensing and ground observations) and methodological approaches used to estimate biomass density from the input data. For this reason, the Validation Protocol does not indicate only one method but rather presents a framework and a set of methods to report accuracy estimates, so that each case study can adopt the approach for the uncertainty assessment and validation most suited to the data available and model applied, still maintaining consistency with the accuracy reports of other products.

The common framework consists of four main components that jointly lead to the achievement of the validation objectives (Figure 1):

- Uncertainty assessment
- Independent validation
- Maps inter-comparison
- User assessment

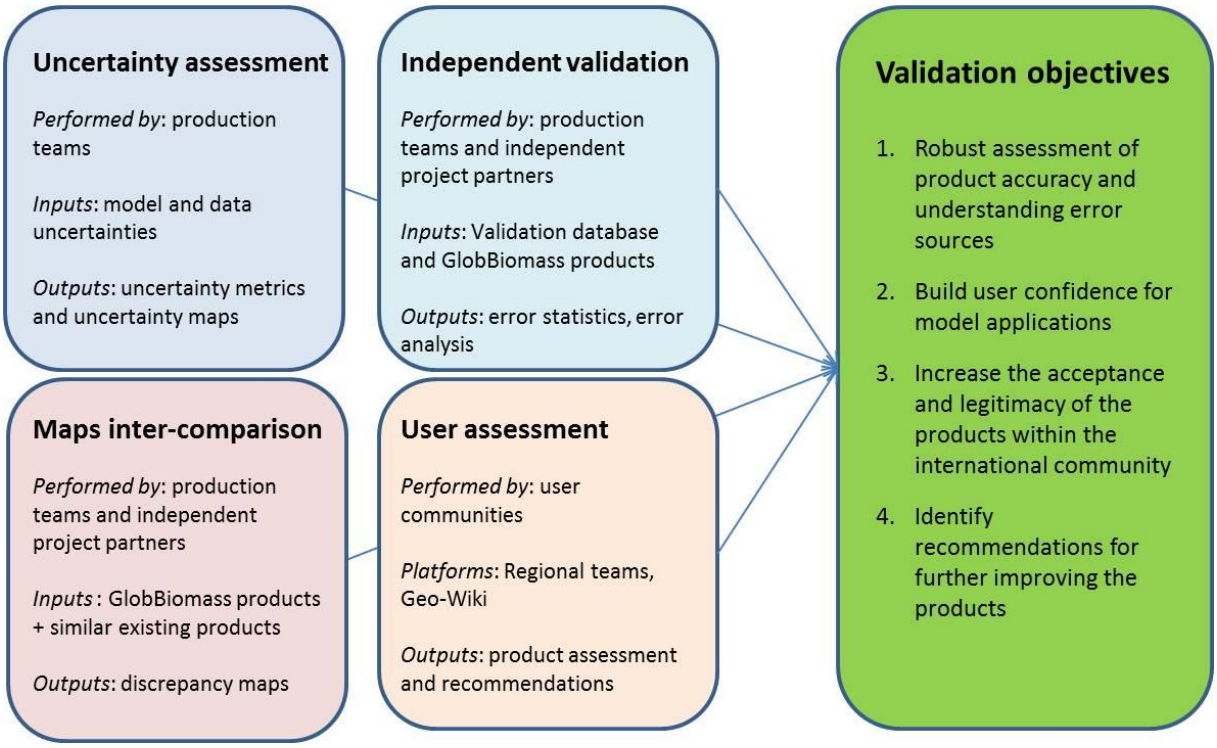

Figure 1: The validation scheme includes four components (left) to achieve the validation objectives (right)

2. Concepts of uncertainty and accuracy

2.1. Definitions

Accuracy: the agreement between the true value and the average of repeated measurements or estimates of a variable (Figure 2). An accurate measurement or prediction lacks systematic error. Bias can occur because of failure to capture all relevant processes involved, because the available data are not representative of all real-world situations, or because of instrument error. According to the IPCC (2006), accuracy is a relative measure of the exactness of an estimate, and accurate estimates are systematically neither over nor under the true values, so far as can be judged.

Precision: the agreement among repeated measurements of the same variable (Figure 2). Precision is independent of accuracy. Larger precision is equivalent to smaller random error, which represents random variation above or below a mean value. Usually, the random error is quantified with respect to a mean value, but the mean could deviate from the true value. Thus, random error is a distinct concept compared to systematic error.

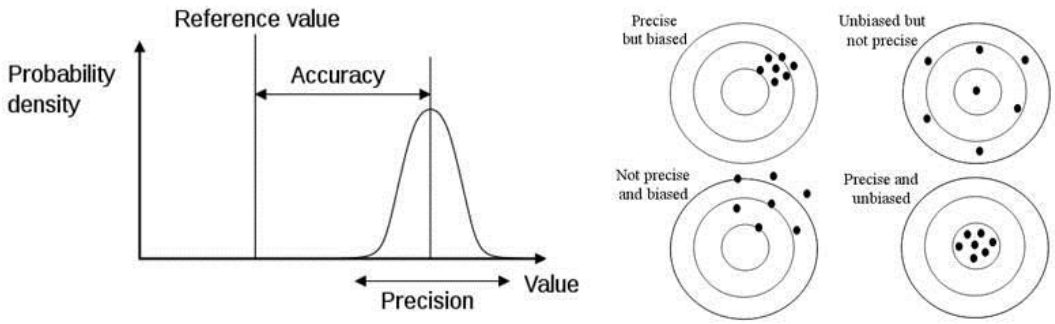

Figure 2: Precision and Bias of a measurement system

Uncertainty: uncertainty can be due to both random errors and/or systematic errors. Random errors are statistical fluctuations in the data due to incomplete knowledge of the variable of interest and/or variability of that variable in space and time. Systematic errors, however, are inaccuracies that are consistently occurring in the same direction. These may be caused by faulty instrumentation, or in the case of modelling, erroneous model structure or co-variates with systematic measurement error or drivers. The IPCC (2006) focuses on the uncertainty that can be described using probability theory and defines uncertainty as the "*lack of knowledge of the true value of a variable that can be described as a probability density function characterizing the range and likelihood of possible values*". Uncertainty depends on "*the quality and quantity of applicable data as well as knowledge of underlying processes and inference methods*".

The main sources of uncertainty are related to conceptualization, models, and input data. Conceptualization uncertainty is related to the failure to properly specify appropriate and relevant model structural assumptions (Cullen and Frey, 1999), and typically results in biased estimation procedures. Model uncertainty is related to imperfections in how the conceptualizations of reality developed from experimental findings as well as theoretical concepts are modelled, and both can lead to imprecision and deviations between observations and model predictions. Conceptualization and model uncertainty can be more difficult to address than the uncertainties in the inputs data and have the potential to produce biased estimation procedures. However, the lack of representativeness of data may also lead to bias, while measurement error can be either random or systematic.

In addition, uncertainty may also be due to vagueness and ambiguity of the class or object represented in the data. If the class or object is semantically poorly defined and different people have different perceptions and interpretations of the labels, then the data are subject to 'vagueness', which can be explicitly treated with fuzzy set theory (Fisher, 1999). The other type of uncertainty is 'ambiguity', which describes the confusion over the definition of a class owing, typically, to differing classification systems.

According to the IPCC (2006), uncertainties due to poor definitions should be eliminated as far as possible before undertaking uncertainty analysis. In the context of the GlobBiomass project, all spatial datasets to be assessed are considered to have well-defined terminology and definitions, and therefore an explicit representation of the degree of vagueness and ambiguity in the datasets is not considered further, and probabilistic methods of uncertainty assessment are appropriate (Fisher, 1999).

Probabilistic uncertainty analysis tends to deal primarily with random errors based on the inherent variability of a system, the finite sample size of available data (uncertainty from sampling), random components of measurement error (uncertainties from measuring), or inferences regarding the random component of uncertainty obtained from expert judgement (uncertainties from expert judgement), if used. In contrast, systematic errors may arise because of imperfections in conceptualization, models, measurement techniques, or other systems for recording or making inferences from data. Systematic errors can be much more difficult to quantify if reference data free of error are not available. Uncertainties due to expert judgement cannot, by definition, be assessed by statistical means since expert judgements are only used where empirical data are sparse or unavailable (IPCC, 2006).

2.2. Uncertainty metrics

2.2.1. Types of variables

Variables can be measured on one of four different scales. Ordinal and nominal scales refer to distinct classes or categories that, respectively, can or cannot be brought into a logical order. Interval scales refer to measurements that can be treated with addition and subtraction but not with division and multiplication because their absolute zero point is not well defined. Metric scales refer to measurements for which addition, subtraction, division and multiplication are allowed because they have a well-defined absolute zero point.

For nominal and ordinal scales (e.g., forest class), uncertainty is best expressed as the probability of an assigned class being correct. These can be estimated from the co-occurrence matrix. For interval and metric scales (e.g., biomass density), the uncertainty is usually expressed as a confidence interval.

In addition, for metric variables the IPCC (2006) recognises a number of basic statistical concepts and terms that are central to the understanding of uncertainty. The recommended uncertainty information includes: the arithmetic mean of the data set; the standard deviation of the data set (the square root of the variance); the standard deviation of the mean (the standard error of the mean); skewness (asymmetry of the distribution); kurtosis (peakedness of the distribution); the probability density function of the data; covariances of the quantity with other input quantities used in the calculations.

2.2.2. The 95% Confidence Interval

According to the IPCC (2006), it is best practice to state the **95% confidence interval** as a definition of the range, given that there is sufficient information to define the underlying probability distribution for conventional statistical analysis. The 95% confidence interval is an interval estimate of a population parameter, calculated from the sample observations, that 95% of the times includes the value of an unobservable parameter of interest if the experiment is repeated. Therefore, if 100 samples were drawn, and a confidence interval were constructed for each sample, 95% of such confidence intervals would include the true value. Importantly, it is not known whether the confidence interval for any particular sample includes the true value, only that 95% of such intervals would include the true value. In the context of an inventory value (such as biomass), the IPCC (2006) considers the confidence intervals in terms of probability, and states that the range of an uncertain quantity should be expressed such that "*there is 95% probability that the actual value of the quantity estimated is within the interval defined by the confidence limits, and it is equally likely that the actual value, should it be outside the range quoted, lies above or below it*".

Two different confidence intervals can be calculated:

- ***The 95% confidence interval for the individual estimates***, based on the standard deviation of the values, includes 95% of the estimated values and excludes 5% of extreme outliers (2.5% on either side of the distribution). For example, if 1000 forest stands were surveyed and their biomass estimated, then 950 of them would have biomass values inside this interval.
- ***The 95% confidence interval of the mean*** is based on the standard error of the mean and describes the range of values in which the true mean is located with 95% certainty. It is narrower than the confidence interval of the individual estimates because the standard error of the mean is divided by the square root of the number of samples used in calculating the mean. For the 1000 forest stands, this confidence interval states that if the survey of all forest stands was repeated, the new estimate of the mean would lie within these bounds with 95% certainty.

The limits of the confidence interval are directly dependent on the probability distribution or the probability density function (PDF) chosen to represent the dataset. The PDF, or density of a continuous random variable, is a function that describes the relative likelihood for this random variable to take on a given value. If the variable under consideration has a normal distribution, the confidence limits of a 95% confidence interval are symmetric about the estimate and approximately equal to 2 standard deviations of the variable, above and below the estimate. More generally, where the PDF is symmetrical the confidence interval can be conveniently **expressed as percentage**, as plus or minus half the confidence interval width divided by the estimated value of the variable (e.g., ± 10%). Where the PDF is not symmetrical, upper and lower limits of the confidence interval need to be specified separately (e.g., -30%, +50%). If a Gaussian distribution is assumed, the shape and symmetry of the probability density function should be examined and its normality should be tested, for example with the Kolmogorov-Smirnov test. As noted by IPCC (2006), "*the confidence intervals estimated from small numbers of observations via a variance (and an assumed probability distribution) have uncertainties associated with them, and in these cases, further observations may either increase or decrease these calculated uncertainty limits. Ultimately, large numbers of observations will decrease the uncertainty limits of the standard deviation*".

The PDF is essential for the determination of the confidence interval of the quantity being considered and for the propagation of uncertainties. PDFs are an approximation to the real data, and identifying which function best fits a dataset can be difficult. The IPCC (2006) provides some guidance on this task and indicates that the knowledge of the underlying physical processes should determine the choice of a probability function. In addition, the IPCC suggests that "*where empirical data are available, the first choice should be to assume a normal distribution of the data, unless the scatter plot of the data suggests a better fit to another distribution*".

3. Uncertainty assessment

3.1. Approach to assess uncertainty in biomass maps

An uncertainty analysis should be seen, first and foremost, as a means to help prioritise efforts to reduce the uncertainty of estimates (e.g., biomass density), and guide decisions on methodological choice. Since uncertainties should be reduced as far as practicable (IPCC, 2006), when focusing efforts to reduce uncertainty, priority should be given to those inputs that have the most impact on the overall uncertainty of the estimate.

In the present document, the uncertainty assessment consists on the evaluation and **quantification of the error sources in the model predictions**. The terms 'uncertainty' and 'error' are used interchangeably in this section.

The uncertainty assessment is aimed at **identifying a confidence interval for each estimate at pixel level).** The error should be computed and reported as relative 95% Confidence Interval (%) and/or as Coefficient of Variation, as appropriate. A different metric for a certain error source may be used if more appropriate. The sources of uncertainty should be quantified separately and then aggregated using an uncertainty propagation method.

The error sources should be directly quantified from the input data used to derive the GlobBiomass product, where possible. If the input data lack such information, approximate values from literature that quantify the error sources for the same or similar type of data may be used. If the error source is considered negligible, it can be not accounted for, given that proper justification is provided. If no reliable information are available from the literature, it will be noted that the error source cannot be accounted for. A **list of all error sources** accounted and non-accounted for the uncertainty analysis **should be provided in a tabular form**, including the respective values and data sources used for their assessment (from data or literature, see Table 1).

Where possible, spatial correlation in the data should be quantified (e.g., using variogram) and used to improve the assessment of the related uncertainty. Failure to account for spatial correlation in the assessment of uncertainty will inevitably lead to under-estimates of the effects of uncertainty, such as a confidence interval that is too short. Once the sources of uncertainty have been determined, their estimates can be combined using an error propagation equation or using probabilistic simulations such as Monte-Carlo Analysis, which can also include auto- and cross-correlations among the components. Both methods have been reported to provide similar results (IPCC, 2006) and are further described below.

Bayesian statistics is another way of quantifying the uncertainty in the input parameters (prior distribution) to the estimated values (posterior distribution). Given two stochastic events

proposition A and evidence B, Bayes' theorem is usually expressed as follows:

$$p(A \mid B) = p(B \mid A)\, p(A) / p(B)$$

where:

- P(A), the prior, is the the probability of A before B is observed.
- P(A | B), the posterior, is the probability of A after B is observed.
- P(B | A) represents the likelihood, or the probability of observing B given A.

In the case of biomass mapping using RS data, when a certain biomass density is estimated at a pixel location from spectral reflectance observations, then P(A) is the distribution of biomass that is expected before taking the spectral measurements B (e.g., derived from an old forest inventory or expert knowledge). The likelihood P(B | A) describes how likely it is to record the observed spectral measurements B for a given biomass A. P(B) is a normalisation constant (or total probability of the signal/data). Bayesian statistics allows to quantify the posterior probability P(A | B), which indicates the probability that the pixel has a certain biomass A given the observed spectral measurements B. The important feature of Bayesian inference is that prior, posterior and likelihood are probability density functions (PDF). This implies that the 95% confidence intervals for the estimated posterior can be explicitly derived from the shape of its PDF by calculating the 2.5% and the 97.5% percentiles.

3.2. Sources of uncertainty in biomass maps

The uncertainty of a biomass map produced using ground observations, remote sensing (RS) data and a (semi)empirical modelling approach is related to the uncertainty of the input data (ground and RS data), uncertainty due to their integration (spatial and temporal mismatch) and uncertainty due to the model assumptions and parametrization (modelling error). In turn, the uncertainty of each of these components depends on various error sources. Measurement error of ground variables can be reduced by exercising greater caution, using better protocols, and using more advance measuring instruments. Sampling variability of ground data cannot be changed, although its effects can be reduced using different sampling designs and estimators. The uncertainty in model predictions due to the uncertainty of model parameters can be reduced by increasing the sample size while the 'structural' uncertainty, or residual uncertainty around model predictions due to the lack of fit of the model, can be reduced using a more representative model and/or better predictor variables.

The main error sources to be considered for each component and the approach to use for the error estimation (e.g., from literature or re-measurement) are described below.

Error sources:

1. **RS data**: errors of satellite images after pre-processing, it includes:

   - spatial error (geolocation accuracy)
   - spectral error (e.g., due to haze, sensor characteristics, etc.)

   An estimate of these errors is usually provided by the data source (Space agency)

2. **Additional datasets**: Accuracy of additional input maps or datasets, if used as input variable (e.g., Forest mask, VCF, etc.). The uncertainty of the additional datasets is usually provided by the data source or available in the literature

3. **Ground data**: errors due to the amount, distribution and quality of the calibration data. It includes:

   - *Tree-level errors*: error of measuring tree variables, as diameter and height, error in the identification of tree species and error in the estimation of related wood density. These error sources can be estimated from plot re-measurement or derived from literature
   - *Allometric model prediction uncertainty*: uncertainty from the use of allometric models, it is usually quantified in the scientific paper or document where the allometric model is published/presented. It has two components:
     - *Parameter uncertainty*, or uncertainty about the true value of the equation parameters
     - *Structural uncertainty*, or residual uncertainty around model predictions, due to the lack of fit of the model.
   - *Sampling variability*: error due to incomplete representativeness of field samples of the 'true' biomass distribution of the area (e.g., some biomass classes are not or under sampled). The sampling error can be derived from literature or it can already be accounted for in the modelling error (e.g., in the MaxEnt model, according to Saatchi et al., 2011)
   - *Geolocation error*: geolocation accuracy of plot data can be estimated from re-measurement of plot coordinates or from literature

4. **Spatial mismatch** between ground and RS data: this error source is originated from two different components:
   - *Geolocation error* of RS and ground data: combination of the geolocation errors of the two input datasets (RS and ground data) described above
   - *Different support size (resolution) and alignment* between ground observations (field plots) and map units (pixels). Plots may be smaller than the pixel footprint and not representative of the mean biomass density at pixel level, especially in areas where biomass presents strong local spatial variation. This error source can be estimated from sample data (i.e., if larger plots covering one or more pixels are available) or from literature

5. **Temporal mismatch** between ground and RS data: the ground data may not be representative of the RS data, as a consequence of two different processes:
   - *Natural growth and mortality*: it is relevant when a considerable time period (>3-5 years) occurred between the acquisition of the ground observations and RS data. This error source can be quantified and corrected using growth rates, if available.
   - *Land change processes* (e.g. deforestation): land change can be identified through visual analysis of recent very high resolution images (e.g., Google Earth) and/or from forest change maps and datasets, if available. This error source can be corrected by discarding the plots not representative of the state of the land at the time of acquisition of the RS data. This error source should be considered and corrected already during the quality screening of the plot data

6. **Model**: prediction error, or uncertainty of the algorithm used to transform RS signals in biomass density. This error is related to the unexplained variance in an algorithm developed from a dataset, and can be due to two different sources:
   - *Parameter* uncertainty: uncertainty about the true value of each parameter estimate in the algorithm (including the exponents)
   - *Structural uncertainty*: uncertainty due to the lack of fit of the model (e.g., using a linear model for a curvilinear phenomenon).

For continuous prediction (e.g., biomass density) the prediction error is summarized and usually reported as R^2, RMSE or relative RMSE. For categorical predictions (e.g., biomass class), the classification methods may allow a direct quantification of uncertainty, as in the case of the Maximum Likelihood classifier that calculates the probability of each object (or pixel) belonging to a certain class.

Table 1: example values and literature sources for some error sources. The values are provided by the GlobBiomass regional case study for Mexico (courtesy of CESBIO).

Source	**Example values (provided by CESBIO):**
RS data	**Spectral error for radar data:** • Radiometric stability = 0.5 dB (Shimada et al., 2010) • Speckle noise = 0.4 dB • Aerosol optical thickness: see Hagolle et al., 2015
Additional datasets	NA
Ground data	Error at tree level: • Diameter breast height: 2.25% (from re-measurement of 5% trees) • Height: 4.47% (from re-measurement of 5% trees) • Wood specific gravity: 10% (Chave et al., 2005) • Allometric equations: 5% (Chave et al., 2005) Sampling variability= b/sqrt(plot area) (Réjou-Méchain et al., 2014)
Spatial mismatch	NA
Temporal mismatch	Natural growth (Gourlet-Fleury et al. 2013): • logged plots: growth rates of 2.58 t/ha/yr • unlogged plots: growth rates of 4.82 t/ha/yr
Model	NA

3.3. Uncertainty propagation

Once the individual error sources listed above and related uncertainties have been determined, they should be combined to provide uncertainty estimates for the model estimate (e.g. biomass density at pixel level). There are three main approaches for the estimation of combined uncertainties: the use of error propagation equations, Monte Carlo techniques or the Taylor series method. Either approach may be used, subject to the assumptions and limitations of each method and availability of resources. Even though the Taylor series method has wide applicability and

similar advantages of the Monte Carlo techniques, currently the methods most widely used in the estimation of biomass density by remote sensing based approaches are the error propagation equations and the Monte Carlo techniques (e.g. Saatchi et al., 2011, Weisbin et al., 2014, Baccini et al., 2012, Gonzalez et al., 2010, Zhang et al., 2014), which are further described below. Both approaches focus on quantifying the random component of the uncertainty of the estimate, and assume that known sources of bias have been removed. Hence, biases should be addressed prior to applying either approach: an assessment of bias and potential disagreements among modelling approaches should be conducted, and any action identified to improve the estimate should be taken.

A comprehensive review of the approaches used to estimate the effects of uncertainty in model predictions on large-area estimates including sampling theory, Taylor series approximations and Monte Carlo simulations is provided by McRoberts and Westfall (2014). Relevant studies related to the remote sensing-based modelling and mapping applications of forest properties as tree volume, biomass and emissions, include the works of Ståhl et al. (2014), Berger et al. (2014) and Breidenbach et al. (2014). For an assessment of the effects of uncertainty in allometric model predictions of individual tree volumes on large-area, sample-based estimates of mean volume per unit area, see McRoberts and Westfall (2015).

3.3.1. Error propagation theory

This approach uses error propagation equations to combine the uncertainties in the model components. The uncertain quantities (usually expressed as coefficient of variation) are combined either by multiplication or by addition (or subtraction), depending on the structure of the model. This approach requires estimates of the mean and the standard deviation for each input, and is based on the assumptions that the uncertainties are statistically independent (uncorrelated, or no significant covariance) and have Gaussian distribution. If correlations exist, then either the correlation can be included explicitly or data can be aggregated to an appropriate level such that correlations become less important (otherwise, the Monte Carlo or the Taylor series approach should be used). In addition, this approach is best suited when the uncertainties are relatively small (i.e., the standard deviation divided by the mean value is < 0.3) and present Gaussian (normal) distributions.

Two convenient rules for combining uncorrelated uncertainties (expressed as coefficient of variation) under addition and multiplication are provided by the IPCC:

- ***Addition***: *"Where uncertain quantities are to be combined by addition, the standard deviation of the sum will be the square root of the sum of the squares of the standard deviations of the quantities that are added with the standard deviations all expressed in absolute terms (this rule is exact for uncorrelated variables). Using this interpretation, a simple equation can be derived for the uncertainty of the sum, that when expressed in percentage terms becomes:*

$$U_{total} = \frac{\sqrt{(U_1 \bullet x_1)^2 + (U_2 \bullet x_2)^2 + \ldots + (U_n \bullet x_n)^2}}{x_1 + x_2 + \ldots + x_n}$$

Where:

U_{total} is the percentage uncertainty in the sum of the quantities (half the 95% confidence interval divided by the total (i.e. mean) and expressed as a percentage);

x_i and U_i are the uncertain quantities and the percentage uncertainties associated with them, respectively." (IPCC, 2006).

It should be noted that with correlated variables, the above formula will under-estimate the uncertainty.

- **Multiplication**: *"Where uncertain quantities are to be combined by multiplication, the same rule applies except that the standard deviations must all be expressed as fractions of the appropriate mean values (this rule is approximate for all random variables). A simple equation can also be derived for the uncertainty of the product, expressed in percentage terms:*

$$U_{total} = \sqrt{U_1^2 + U_2^2 + \ldots + U_n^2}$$

Where:

U_{total} is the percentage uncertainty in the product of the quantities (half the 95% confidence interval divided by the total and expressed as a percentage);

U_i are the percentage uncertainties associated with each of the quantities." (IPCC, 2006)

3.3.2. Monte-Carlo simulations

The Monte Carlo analysis is suitable for detailed category-by-category assessment of uncertainty, particularly where uncertainties are large, distribution is non-normal, the algorithms are complex functions and/or there are correlations between some of the uncertainty sources. This method includes auto- and cross-correlations, and allows the determination of the likely value of the covariance, or the related correlation coefficient, between the various input quantities.

The Monte Carlo approach, described in more details in IPCC (2006) and elsewhere, consists on performing the calculation of the parameter to be estimated several times, each time with the model parameters and input data chosen randomly within the distribution of uncertainties specified initially by the user. This process generates an uncertainty distribution for the estimate that is consistent with the input uncertainty distributions on the input data and model parameters. The method is very intensive in terms of computing time, but is well suited to the problem of propagating and aggregating uncertainties (IPCC, 2006). This method uses the same input data as the method previously described, and both methods have been reported to provide similar results (IPCC, 2006).

3.3.3. Uncertainty estimation with a Bayesian method (MaxEnt)

A Bayesian method, such as MaxEnt, estimates probability density functions from which uncertainty at pixel level can be calculated. MaxEnt calculates biomass probability distribution maps for each biomass *(AGB)* class. The continuous values of *AGB* for each pixel are calculated as the weighted average *AGB* per pixel with the probabilities as weights. The uncertainty of the *AGB* prediction ($\varepsilon_{prediction}$) is calculated from the root mean square error ($\sigma_{\widehat{AGB}}$) obtained per pixel. The following equations are used (Saatchi et al., 2011):

$$\widehat{AGB} = \frac{\sum_{i=1}^{N} P_i^n AGB_i}{\sum_{i=1}^{N} P_i^n}$$

$$\varepsilon_{prediction} = \sigma_{\widehat{AGB}} / \widehat{AGB}\, x100$$

$$\sigma_{\widehat{AGB}} = \sqrt{\frac{\sum_{i=1}^{N}(AGB_i - \widehat{AGB})^2 P_i}{\sum_{i=1}^{N} P_i}}$$

where $\widehat{AGB}$ is the *AGB* prediction per pixel, and $\mathrm{P_i}$ is the probability estimated by MaxEnt for each *AGB* range AGB_i (average value within class i). The power of the probability n is used to weight the predicted value towards the maximum probability closest to the true value when other probabilities are small.

The total uncertainty at pixel level is composed of different sources of error which are assumed to be random and independent. These errors are propagated using the error propagation equation as shown by Saatchi *et al.* (2011):

$$\varepsilon_{AGB} = (\varepsilon^2_{measurement} + \varepsilon^2_{allometry} + \varepsilon^2_{sampling} + \varepsilon^2_{prediction})^{1/2},$$

Weisbin *et al.* (2014) presents a similar approach to estimating uncertainty in biomass maps derived from remote sensing data, which can be used to create uncertainty maps.

4. Independent validation

The independent validation consists on the **comparison of the GlobBiomass estimates with independent reference data** (the validation dataset) not used for product development, allowing to assess the accuracy of the GlobBiomass products. Map accuracy is reflected by the degree to which map predictions correspond to the reference observations at each location.

4.1. Characteristics of the validation dataset

An appropriate biomass validation dataset is essential for the accuracy assessment of the GlobBiomass products. According GFOI (2013) two criteria are key to obtain an appropriate validation dataset: the sampling design and the sample size of the dataset. For accuracy assessment and estimation to be valid using the familiar design- or probability-based framework (McRoberts, 2014), the reference data must be collected using a probability sampling design, regardless of how the training data are collected, and assessed uses an estimation procedure that is compatible with the sampling design. The most common probability sampling designs are simple random, systematic, stratified random (simple random sampling within strata), and stratified systematic sampling (systematic sampling within strata). Given the likely rarity of some strata (e.g., forest types) and the large costs associated with large samples, stratified sampling for which the strata correspond to forest classes is considered a powerful technique to manage and reduce the uncertainty in inventory estimates.

If there are sufficient resources, the IPCC (2006) suggests to undertake a monitoring campaign and design a stratified sample of measurements choosing the most appropriate variables to stratify the sample. The whole set of data can be used to estimate the probability density function and summary statistics. Then statistical tools can be used to estimate the bias of the mean and variance, the confidence intervals and the distribution of the errors. When data are missing at a regional level, it is possible to extrapolate information from existing literature provided that care is taken to choose data coming from sources with similar characteristics to those being estimated. In this case, expert judgement is needed (IPCC, 2006).

The validation dataset presents the following characteristics:

a) **Data source**: biomass data that can be used as a reference are primarily field observations (forest inventory plots) but also remotely sensed data that are of greater quality than the map data are considered acceptable for use as reference data (Stehman, 2009), and these may include biomass estimates derived from airborne Lidar data or extracts from reliable high-resolution biomass maps (Avitabile et al., 2015). Reliable sub-national or national statistics can also be used for additional evaluation of agreement of the model predictions with reference values at regional level.

b) **QA/QC of plot data**: The field plot reference data should be screened according to the following quality criteria:
   - Coordinates acquired with GPS, with approximation at meter or sub-meter level (4-6 decimals are provided for coordinates in decimal degree)
   - Minimum plot area: 0.04 - 0.1 ha
   - Year of acquisition: ± 1-10 years difference from map reference year

- Variable: Aboveground biomass of all living trees with DBH > 0-10 cm (indicate if palms and dead trees are included)
- Allometric model is reliable, i.e., it is appropriate to the forest type to which is applied and use sufficient parameters (dbh and wood density and/or height)

Where possible, the field plots are further screened at individual level for temporal mismatch, i.e., to verify that they are representative of the map cells (pixels) and no change processes (e.g., deforestation) have occurred between the plot measurement and reference year of the biomass map (which is related to the acquisition year of the RS data). The screening is performed by visual analysis with Google Earth images or similar very high resolution images acquired for the reference year of the biomass map

c) **QA/QC of reference maps:** Reference data can be extracted from existing high-quality biomass maps. Maps are considered of high-quality if they fulfil the following quality criteria:
   - High spatial resolution (≤ 100 m)
   - Published in the scientific literature
   - Calibrated with local ground data and (possibly) with airborne lidar data

Selection of the reference pixels: Only the cells with largest confidence (i.e., smallest uncertainty) should be selected from the maps and used as reference data. If the reference maps are based on empirical models, the map cells with greatest confidence are assumed to be those in correspondence of the training data (field plots and/or LiDAR data). Uncertainty maps can also be used to identify the map cells with smallest uncertainty. For maps based only on radar or optical data, whose signals saturate above a certain biomass density value, only pixels below such a threshold should be considered (e.g., excluding biomass values > 150 T/ha for maps based only on Landsat or ALOS data).

Amount of reference data extracted from a reference map: If multiple reference datasets are available (i.e., including both field plots and high-resolution maps), in order to compile a reference database that is representative of the area of interest and well-balanced among the various reference datasets, the amount of reference data extracted from the reference biomass maps should be proportional to their area and not greater than the amount of samples provided by the field datasets representing a similar area.

d) **Data harmonization**: All selected reference data (plots, maps) should be harmonized with the map to be assessed in terms of spatial and thematic (i.e. biomass unit) resolution. If reliable data on growth rates are available, the data should also be harmonized in terms of temporal resolution (i.e., considering annual increment rates to correct temporal differences between the plots and the map to assess)

e) **Sample size** (size of the validation dataset): the complete reference dataset should be divided in Calibration and Validation data. If a large reference dataset is available (> 100 plots), the validation data should consist of 10-30% of the complete reference dataset (calibration + validation). If the reference dataset has smaller size (<100 plots in total, or <20-30 plots per biomass class), a cross-validation approach can be used, setting aside an appropriate (smaller) percentage *p* of the reference dataset as testing set. Multiple rounds of cross-validation are performed using different partitions, and the validation results are then averaged over the rounds. For very small datasets, the leave-one-out cross-validation approach (where $p = 1$) may be appropriate.

f) **Sample selection** (Selection of the validation data from the reference dataset): the validation data are selected from the complete reference data as a random sample. Any probability sampling design can be used as long as the estimation procedure matches the sampling design. A stratified random sampling strategy may be advisable (see above), with random samples stratified by 4-6 biomass classes (suggested: 5 classes) and with biomass classes intervals being proportional to the amounts of validation data or to the biomass distribution in the study area (derived from the map histogram).

4.2. Validation metrics

The validation data are considered as reference data and are compared to the estimates of the GlobBiomass products to assess their error and the following validation metrics:

- Bias (mean error)
- Standard Deviation of the error
- 95% Confidence Interval of the mean error
- RMSE and Relative RMSE (to compare with other products)
- Scatterplots of predicted vs. reference data at validation sites (pixels)
- Scatterplots of errors vs. reference data at validation sites in absolute (Mg/ha) and relative (%) terms

Each metric is computed for all validation data, and then separately by biomass class. As indicated above, the biomass classes intervals are proportional to the amounts of validation data or to the biomass distribution in the study area.

Where possible, uncertainty in the validation data should be quantified (see section 3.2.1. “Sources of uncertainty in biomass maps”) and included in the validation process, which allows to express the reference data as a mean value with an associated 95% Confidence Interval.

4.3. Validation of categorical maps (e.g. forest mask)

The basis for assessing accuracy of categorical maps is the confusion matrix (Aronoff, 1982), and it should be used for the validation of categorical GlobBiomass products, as forest mask, forest type or biomass class.

The error matrix gives the correspondence of the classified map to the reference data. Although it does not directly provide estimates of activity areas or their uncertainties, the information in an error matrix can be used to do so (McRoberts and Walters, 2012, Olofsson et al., 2013). The overall accuracy p_0 is the percentage of correctly classified objects:

$$p_0 = \frac{1}{N}\sum_{j=1}^{n} p_{jj}$$

However, p_0 depends on the number of classes and the expected chance agreement. These properties make it impossible to compare p_0 values from different classifications. To correct for the expected chance agreement p_e, different types of coefficients of agreement have been developed. Chance agreement is defined based on the multinomial distribution (Nishii and Tanaka, 1999). The

range of possible values of these coefficients is –1 to 1, where 0 is pure chance agreement and 1 is perfect agreement. Negative values indicate a classification accuracy that is worse than a random assignment of classes to the objects.

A priori coefficients like τ (Ma and Redmond, 1995) use prior knowledge of the expected class frequencies to estimate the chance agreement between the classification and the ground data. Class frequencies are normally assumed equal if no prior knowledge exists.

A posteriori coefficients of agreement like κ (Cohen, 1960) estimate the chance agreement from the observed marginal distributions of the confusion matrix. κ can be calculated from the confusion matrix by

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

with $p_0 = \frac{1}{N}\sum_{j=1}^{n} p_{jj}$ and $p_e = \frac{1}{N^2}\sum_{j=1}^{n}\sum_{k=1}^{n} p_{j\bullet} p_{\bullet k}$

κ only distinguishes whether an object has been assigned the correct class. For ranked or ordered classes like forest biomass classes for example, a modified coefficient exists that is weighted by the seriousness of the classification error. The further the true class is away from the assigned class label, the more serious the error is. The weighted κ_w coefficient (Cohen, 1968) uses the entire information of the confusion matrix.

5. Maps inter-comparison

This task consists on the comparison of the GlobBiomass maps with other existing products covering the same geographic extent. The comparison does not provide information on product accuracy but is aimed to complement the above tasks (uncertainty assessment and product validation) and to provide the following information: evaluation of relative consistency among different products; identification of areas with larger disagreements and assessment if these disagreement areas are in need of further study; assessment of strengths and weaknesses of different datasets based on the analysis of the data and methods used by each map; build confidence in the user communities.

The maps inter-comparison consists of the following steps. First, the comparable datasets (e.g., regional or global maps) should be identified and acquired. Second, the datasets should be harmonized in terms of spatial, thematic (i.e., biomass unit) and temporal resolutions. Third, the following comparison metrics should be computed at pixel level and at aggregated resolution (e.g., 1 Km):

1. Comparison statistics:
   - Mean absolute difference
   - Root Mean Square Difference
   - Linear correlation
   - Statistics on spatial correlation may be added
2. Comparison maps:
   - Difference maps, in absolute values
   - Difference maps, as percentage difference to the GlobBiomass map
3. Comparison graphs:
   - Scatterplots
   - Compare histograms representing data frequency per biomass class

6. User assessment

The User Assessment is an essential quality control and feedback mechanism, aimed at assessing the users' acceptance of the products, evaluating the quality and limitations from User's perspective, and obtaining recommendations to future improvements. The output of the user assessment consists of a User Survey report.

The user assessment is performed using standard questionnaires, which are produced for each GlobBiomass product. The questionnaires have the following aims:

- Assess user satisfaction
- Assess product usability
- Assess the delivery system (timing, delivery method, naming, format, etc.)
- Assess product quality and limitations related to:
  - spatial resolution
  - temporal resolution
  - variable mapped (e.g. need for other biomass pools)
- Assess applications of the products (e.g., REDD+ implementation, Natural resource management, Etc.)
- Assess need of capacity building (optional)
- Future data and product requirements

References

ARONOFF, S. 1982. The map accuracy report-a users view. Photogrammetric Engineering and Remote Sensing, 48, 1309-1312, ISSN: 0099-1112.

AVITABILE, V., HEROLD, M., HEUVELINK, G. B. M., LEWIS, S. L., PHILLIPS, O. L., ASNER, G. P., et al. 2015. An integrated pan-tropical biomass map using multiple reference datasets. Glob Change Biol. Accepted Author Manuscript. doi:10.1111/gcb.13139

BACCINI, A., GOETZ, S. J., WALKER, W. S., LAPORTE, N. T., SUN, M., SULLA-MENASHE, D., HACKLER, J., et al. 2012. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. Nature Climate Change, 2, 182–185.

BERGER, A., GSCHWANTNER, T., MCROBERTS, R. E., & SCHADAUER, K. 2014. Effects of measurement errors on individual tree stem volume estimates for the Austrian National Forest Inventory. Forest Science, 60(1), 14-24.

BREIDENBACH, J., ANTÓN-FERNÁNDEZ, C., PETERSSON, H., MCROBERTS, R. E., & ASTRUP, R. 2014. Quantifying the model-related variability of biomass stock and change estimates in the Norwegian National Forest Inventory. Forest Science, 60(1), 25-33.

CHAVE, J., ANDALO, C., BROWN, S., CAIRNS, M.A., CHAMBERS, J.Q., EAMUS, D., FÖLSTER, H., FROMARD, F., HIGUCHI, N., KIRA, T. AND LESCURE, J.P., 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. Oecologia, 145(1), pp.87-99.

COHEN, J. 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20, 37-46, ISSN: 0013-1644.

COHEN, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin, 70, 213, ISSN: 1939-1455.

CULLEN, A. C., & FREY, H. C. 1999. Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs. Springer Science & Business Media.

FISHER, P. F. 1999. Models of uncertainty in spatial data. Geogr. information systems, 1, 191-205.

GFOI 2013. Integrating remote-sensing and ground-based observations for estimation of emissions and removals of greenhouse gases in forests: Methods and Guidance from the Global Forest Observations Initiative, Geneva, Switzerland. ISBN: 978-92-990047-4-6.

GONZALEZ, P., ASNER, G. P., BATTLES, J. J., LEFSKY, M. A., WARING, K. M. & PALACE, M. 2010. Forest carbon densities and uncertainties from Lidar, QuickBird, and field measurements in California. Remote Sensing of Environment, 114, 1561-1575, ISSN: 0034-4257.

GOURLET-FLEURY, S., MORTIER, F., FAYOLLE, A., BAYA, F., OUÉDRAOGO, D., BÉNÉDET, F., & PICARD, N. 2013. Tropical forest recovery from logging: a 24 year silvicultural experiment from Central Africa. Philosophical Transactions of the Royal Society, Biological Sciences, 368(1625), 20120302.

HAGOLLE, O., SYLVANDER, S., HUC, M., CLAVERIE, M., CLESSE, D., DECHOZ, C., LONJOU, V., POULAIN, V. 2015. SPOT-4 (Take 5): Simulation of Sentinel-2 Time Series on 45 Large Sites. Remote Sensing, 7(9), 12242-12264.

IPCC 2006. 2006 IPCC Guidelines for National Greenhouse Gas Inventories, Prepared by the National Greenhouse Gas Inventories Programme. Eggleston H.S., Buendia L., Miwa K., Ngara T. and Tanabe K. ed.: IGES, Japan.

MA, Z. & REDMOND, R. L. 1995. Tau coefficients for accuracy assessment of classification of remote sensing data. Photogrammetric Engineering and Remote Sensing, 61, 435-439, ISSN: 0099-1112.

MCROBERTS, R. E. 2014. Post-classification approaches to estimating change in forest area using remotely sensed auxiliary data. Remote Sensing of Environment, 151, 149-156, ISSN: 0034-4257.

MCROBERTS, R. E. & WALTERS, B. F. 2012. Statistical inference for remote sensing-based estimates of net deforestation. Remote Sensing of Environment, 124, 394-401, ISSN: 0034-4257.

MCROBERTS, R. E., & WESTFALL, J. A. 2014. Effects of uncertainty in model predictions of individual tree volume on large area volume estimates. Forest Science, 60(1), 34-42.

MCROBERTS, R. E., & WESTFALL, J. A. (2015). Propagating uncertainty through individual tree volume model predictions to large-area volume estimates. Annals of Forest Science, 1-9.

NISHII, R. & TANAKA, S. 1999. Accuracy and inaccuracy assessments in land-cover classification. Geoscience and Remote Sensing, IEEE Transactions on, 37, 491-498, ISSN: 0196-2892.

OLOFSSON, P., FOODY, G. M., STEHMAN, S. V. & WOODCOCK, C. E. 2013. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. Remote Sensing of Environment, 129, 122-131, ISSN: 0034-4257.

RÉJOU-MÉCHAIN, M., MULLER-LANDAU, H.C., DETTO, M., THOMAS, S.C., TOAN, T.L., SAATCHI, S.S., BARRETO-SILVIA, J.S., BOURG, N.A., BUNYAVEJCHEWIN, S., BUTT, N. AND BROCKELMAN, W.Y., 2014. Local spatial structure of forest biomass and its consequences for remote sensing of carbon stocks. Biogeosciences Discussions, 11, p.5711.

SAATCHI, S. S., HARRIS, N. L., BROWN, S., LEFSKY, M., MITCHARD, E. T. A., SALAS, W., ZUTTA, B. R., et al. 2011. Benchmark map of forest carbon stocks in tropical regions across three continents. Proceedings of the National Academy of Sciences, 108, 9899-9904

SHIMADA, M., & OHTAKI, T. (2010). Generating large-scale high-quality SAR mosaic datasets: Application to PALSAR data for global monitoring. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, 3(4), 637-656.

STÅHL, G., HEIKKINEN, J., PETERSSON, H., REPOLA, J., & HOLM, S. (2014). Sample-based estimation of greenhouse gas emissions from forests—A new approach to account for both sampling and model errors. Forest Science,60(1), 3-13.

STEHMAN, S. V. 2009. Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover change from remote sensing. Remote Sensing of Environment, 113(11), 2455-2462.

WEISBIN, C. R., LINCOLN, W. & SAATCHI, S. 2014. A Systems Engineering Approach to Estimating Uncertainty in Above-Ground Biomass (AGB) Derived from Remote-Sensing Data. Systems Engineering, 17, 361-373, ISSN: 1520-6858.

ZHANG, G., GANGULY, S., NEMANI, R. R., WHITE, M. A., MILESI, C., HASHIMOTO, H., WANG, W., SAATCHI, S., YU, Y. & MYNENI, R. B. 2014. Estimation of forest aboveground biomass in California using canopy height and leaf area index estimated from satellite data. Remote Sensing of Environment, 151, 44-56, ISSN: 0034-4257.

Wageningen (The Netherlands), 16.12.2015

……

Prof. Dr. Martin Herold