Greenhouse gas management in European land use systems

FP7 Project GHG-Europe
Grant agreement No 244122

Deliverable D5.1			
Title	Protocols for site scale simulations of fluxes with data oriented models		
Delivery date from Annex I (project month)	6		
Actual delivery date	30/06/2010 (month 6)		
Lead participant	WP	Nature	Dissemination level
CEA (2)	5	R	RE

Deliverable description

The simulation protocol of fluxes with data-oriented model comprises 1) the processing, QA/QC of the training data (fluxes to be estimated), 2) the processing of the driving data, 3) the model evaluation. Due to specific requirement of different data-oriented methods (e.g. time step) not a full standardisation is possible, but a minimum protocol to be followed was established as follows:

1) Training data
   - Usage of half-hourly eddy covariance measurements processed using standardised procedures of gap-filling and quality control [Reichstein et al. 2005, Moffat *et al*., 2007; Papale *et al*., 2006].
   - Aggregation into daily or monthly means.
   - Exclusion of data where the mean was based on more than 20 % data gap-filled with low confidence (Reichstein *et al*. 2005).
   - Usage of 2 estimates of GPP and TER from the flux data, based on flux-partioning and estimation of TER using night-time or day-time data (Reichstein *et al*. 2005, Lasslop *et al*. 2010). Choose GPP based on Lasslop *et al*. 2010 and TER based on Reichstein *et al*. 2005, because they are largely independent estimates.
   - Corrections of monthly LE and H measurements according to Twine *et al*. 2000, which preserves the Bowen ratio: $F_{corrected} = A \times F_{uncorrected} = (Rn - G) / (H_{uncorrected} + LE_{uncorrected}) \times F_{uncorrected}$.
   - Filter out data points that exhibit strong inconsistencies of the 2 flux-partitioning methods using an outlier test which indicates possible problems due to uncertain night-time data and therefore possible systematic errors: (1) $GPP_{Reichstein} - GPP_{Lasslop}$, (2) $TER_{Reichstein} - TER_{Lasslop}$, (3) $NEE_{measured} - NEE_{Lasslop}$. If differences were larger than range given by the median ± 1.5 times the inter-quartile range, data was discarded.

- Filtering uncertainties which originate from determining the u* threshold. The u* uncertainty in the carbon fluxes is inferred by bootstrapping based on Papale *et al.* 2006. The 5 % most uncertain data points of each carbon flux are removed.

2) Driver data
    - Driver data as listed in Tab. 1 can be used by the data-driven models: This includes (1) monthly fAPAR from the SeaWiFS sensor, precipitation, and temperature (both in-situ measured); (2) annual changes of the fAPAR that describe properties of vegetation structure such as minimum, maximum, mean, and amplitude; (3) mean annual climate such as mean annual temperature, precipitation, sunshine hours, relative humidity, potential evapotranspiration, climatic water balance (precipitation – potential evaporation), and their seasonal dynamics; and (4) the vegetation type according to the IGBP classification plus a flag regarding the photosynthetic pathway (C3, C4, C3/C4) (in-situ information). These driving data are prepared in a standardised way and are available to project partners employing data-driven models.

3) Evaluation of data-driven models
    - Use 5-fold cross-validations: A 5-fold cross-validation implies that the dataset is stratified into 5 parts with approximately equal number of samples. The target values for each of these 5 parts are predicted based on the training using the remaining 4 parts.
    - Wwo experiments: Where (1) entire sites were removed from the training (~20 %), and (2) consecutive parts of the time series of the sites were removed. Hence, the first cross-validation experiment corresponds to the uncertainty of predicting the flux time series for unknown sites, while the second experiment assesses the uncertainty of filling long gaps in the time series based on the information from all flux sites.
    - Evaluation performed along three categories of temporal variability: (1) among-site variability, (2) seasonal variation, and (3) anomalies. First calculate the mean seasonal cycle (FMSC) per site, i.e. averaging the values for a month across all available years (at least 2 values (i.e. years) for a month must be available). Calculated a mean value for each site (FSITE) given as the mean of FMSC if at least 6 out of 12 values of FMSC are present. Calculate the seasonal variation FSEAS by removing FSITE from FMSC. Calculate anomalies as the deviation of a flux value from FMSC.
    - Calculate performance measures: Pearson's correlation (Cor), Nash-Sutcliff's modelling efficiency (MEf, [Nash and Sutcliffe, 1970]), root mean squared error (RMS), median absolute deviation (MAD), and ratio of variances (RoV) which is the variance of the predicted values divided by the variance of the observed values.
    - Remove extreme outliers from the computation of performance measures to avoid biased statistics. Identify as outlier data points outside the range defined by the median of the residuals ± 7 times the inter-quartile range of the residuals.

Table 1: Driving data for data-driven models in WP5.

	Variable	Type of variability
Climate (for data stratification)	Mean annual temperature	Static
	Mean annual precipitation sum	Static
	Mean annual climatic water balance	Static
	Mean annual potential evaporation	Static
	Mean annual sunshine hours	Static
	Mean annual number of wet days	Static
	Mean annual relative humidity	Static
	Mean monthly temperature	Monthly but static over years
	Mean monthly precipitation sum	Monthly but static over years
	Mean monthly climatic water balance	Monthly but static over years
	Mean monthly potential evaporation	Monthly but static over years
	Mean monthly sunshine hours	Monthly but static over years
	Mean monthly number of wet days	Monthly but static over years
	Mean monthly relative humidity	Monthly but static over years
Vegetation structure	Maximum fAPAR of year	Yearly
	Minimum fAPAR of year	Yearly
	Maximum – Minimum fAPAR	Yearly
	Mean annual fAPAR	Yearly
	Sum of fAPAR over the growing season	Yearly
	Mean fAPAR of the growing season	Yearly
	Growing season length derived from fAPAR	Yearly
	Sum of fAPAR × potential radiation of year	Yearly
	Maximum of fAPAR × potential radiation of year	Yearly
	IGBP vegetation type	Static
Meteorology	Temperature	Monthly/daily
	Precipitation	Monthly/daily
	Potential radiation	Monthly but static over years
	Short-wave incoming radiation	Monthly/daily
	Vapour pressure deficit	Monthly/daily
Vegetation status	fAPAR	Monthly/10daily
	fAPAR x potential radiation	Monthly/10daily