Light-curve modelling of GPP across the ABoVE domain using MODIS MAIAC data and a flux tower network.

G. Hmimina, R. Yu, R. Wang, K. F. Huemmrich, J. Gamon

01/2022

1. Context

The approach adopted involved performing a multi-site analysis centered on a network of 80 flux towers located within the extended ABoVE domain, then an upscaling of the results obtained at site-level using the MODIS MAIAC dataset, spanning the whole extended ABoVE domain from 2002 to 2015, at a 1 kilometer and sub-daily resolution.

The first step involved collecting and processing site-level data (mainly carbon flux data to use as a reference). The second step involved the correction and analysis of MODIS MAIAC data at site-level. The final step was to develop a framework to produce large-scale estimates of phenology and ecosystem sensitivity to climate change.

2. General approach

While the gross primary productivity (GPP) and the light use efficiency (LUE) are highly dynamic at the sub-daily scale and are therefore difficult to upscale from flux towers to satellite scale due to a temporal resolution mismatch, the daily variability in GPP can be described using a light-curve model similar to the one used for partitioning flux-tower NEE measurements into GPP and Ecosystem respiration (Reco). This allows us to deconvolve the GPP variability into 3 components: the absorbed photosynthetically active radiation (aPAR) which varies at the daily scale, the maximum GPP or maximum photosynthetic capacity (GPPmax), and the photosynthetic limitation or amount of light needed to reach maximum capacity (PPFDmax). These two components (GPPmax and PPFDmax) vary seasonally and can therefore be related to satellite reflectance measurements sampled at the daily scale.

This process allows us to generate a light-curve parameter model, taking MAIAC corrected reflectance measurements as an input, which can be used to estimate GPP at a daily resolution. The process involved in the derivation of this model and its use to compute GPP is described in Figure 1.

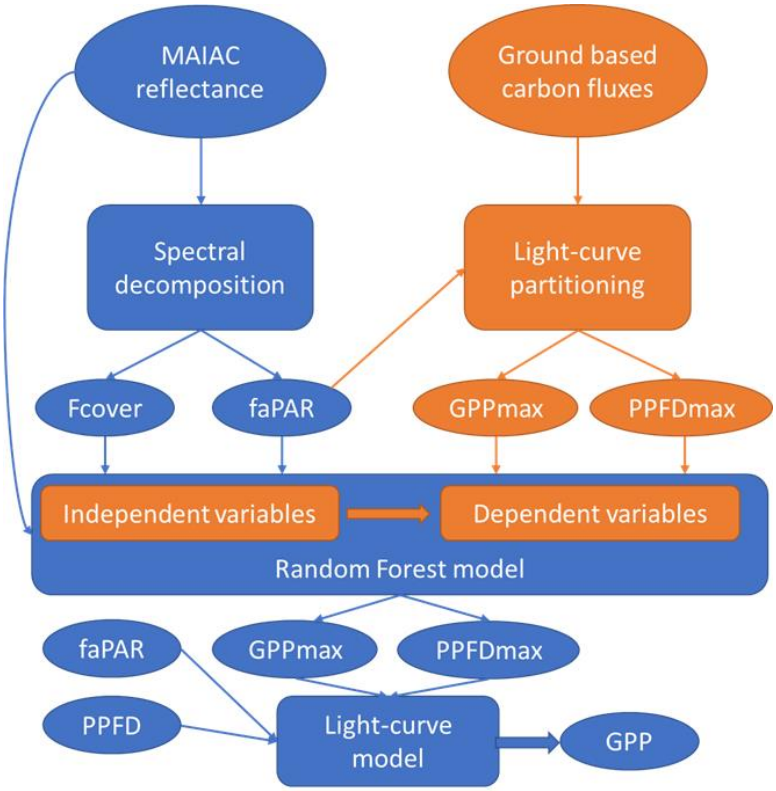

Figure 1: functional diagram of the tile level GPP model. Blue components are dataset or process at tile-level scale, orange components are datasets of process at flux-tower level.

3. Algorithm

3.1. Estimation of $faPAR_{vegetation}$ at tile level

The light curve model used both at the flux tower level to extract GPPmax and PPFDmax, and at tile level to predict GPP from remote-sensing derived GPPmax and PPFDmax estimations uses aPAR rather than PAR and requires an independent estimation of the fraction of absorbed PAR (faPAR). This process is required to decorrelate the GPPmax (dependent on the area-based amount of chlorophyll and therefore faPAR) and PPFDmax which would be dependent on $faPAR_{vegetation}$ if PAR was used as the light-curve dependent variable instead of aPAR.

The estimation of $faPAR_{vegetation}$ at tile level was done through a process of spectral decomposition of the MAIAC reflectance measurements at daily scale. A spectral library was compiled in order to represent the spectral variability of the ABoVE domain through its whole seasonal cycle. Leaf spectral measurements from 400 to 2500 nm were collected from the EcoSis database (https://ecosis.org/), by downloading all datasets located within the coverage of the MAIAC dataset. Leaf reflectance spectra measurements were pulled from the Ecosis database (https://ecosis.org/), measurements corresponding to locations within the MAIAC dataset footprint were selected. Soil and water spectra were pulled from the USGS spectral library (https://crustal.usgs.gov/speclab/QueryAll07a.php).

The resulting compiled database was first screened for artefacts (reflectance values less than 0 or greater than 1, or spectra exhibiting discontinuities bigger than 1% in reflectance). The resulting spectra were then used to estimate a MODIS-equivalent spectral sample using the following equation:

$$R_b = \frac{\int_{300}^{2500} R * T_b}{\int_{200}^{2500} T_b}$$

With Rb the MODIS band-equivalent reflectance, R a cubic spline fitted to the spectrum being analyzed and T the MODIS response curve in the considered MODIS band. This convolution was done analytically using Matlab symbolic math toolbox.

With R a cubic spline fitted to the considered ASD spectrum, and PPFD a cubic spline fitted to a standard PPFD spectrum (ASTM G-173-03 reference spectrum).

A second filter was then applied to remove redundant spectra (correlations between simulated MODIS reflectances superior to 0.999).

The resulting spectral library can be used to estimate the composition of each pixel for each observation through solving the following equation system:

$$\begin{cases} R_{MAIAC} = \sum_{i\,snow} a_i * R_i^{library\ snow} + \sum_{i\,soil} a_i * R_i^{library} + \sum_{i\,vegetation} a_i * R_i^{library} + b \\ 0 \le a_i \le 1 \\ -1 < b \le 0 \\ \sum a_i + b = 1 \end{cases}$$

This equation system was solved using the Interior point algorithm.

The b coefficient stands for shadow effects, shadows being modeled as black bodies.

To avoid confusion due to soil, snow and leaf spectra which happen to be similar to other components (i.e, dead leaf spectra which are similar to soil spectra), a last screening step is needed to obtain the final spectra library. 200 pure soil and snow MODIS observations were selected from the available extracted MAIAC data at the flux tower sites based on NDVI and the minimum daily NEE (NDVI<0 for snow selection, NDVI>0.3 and minimum daily NEE >0 for "soil"). Each spectrum of the library was decomposed using the pure MODIS observations, and the distribution of the coefficients obtained with MAIAC observations of a different class were collected. A confusion metric was defined as follow :

$$C_i = Percentile_{95\%}([a_i^{class \neq class\ i}])$$

All spectra for which the confusion index was higher than 1% were removed from the library. The resulting library counts 28 leaf spectra, 40 soil spectra and 4 water/snow spectra, which are shown in Figure 2.

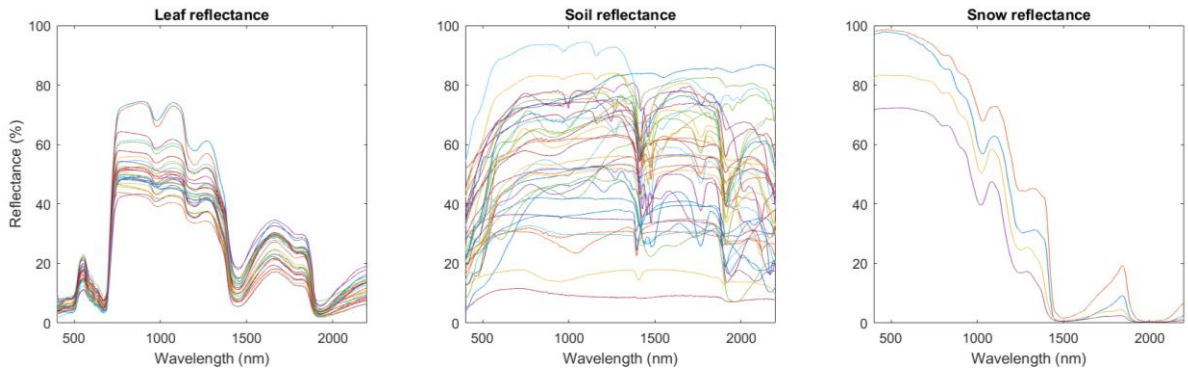

Figure 2: reflectance spectra of the filtered spectral library.

For each selected spectrum, an estimation of faPAR and faSW was computed while neglecting transmittance (assumed to be negligible), using the following equations :

$$faPAR = \frac{\int_{400}^{700}(1-R)*PPFD}{\int_{400}^{700}PPFD}$$

$$faSW = \frac{\int_{400}^{2500}(1-R)*PPFD}{\int_{400}^{2500}PPFD}$$

The simulated percentage of reflected PAR and shortwave radiation was computed for each flux tower site as follow:

$$R_{PAR} = 1 - \sum_{i} a_i * faPAR_i$$

$$R_{SW} = 1 - \sum_{i} a_i * faSW_i$$

The obtained estimations were compared to the percentage of reflected PAR and shortwave radiation computed based on measured daily average downwelling and upwelling flux tower-based measurements (Figure 3).

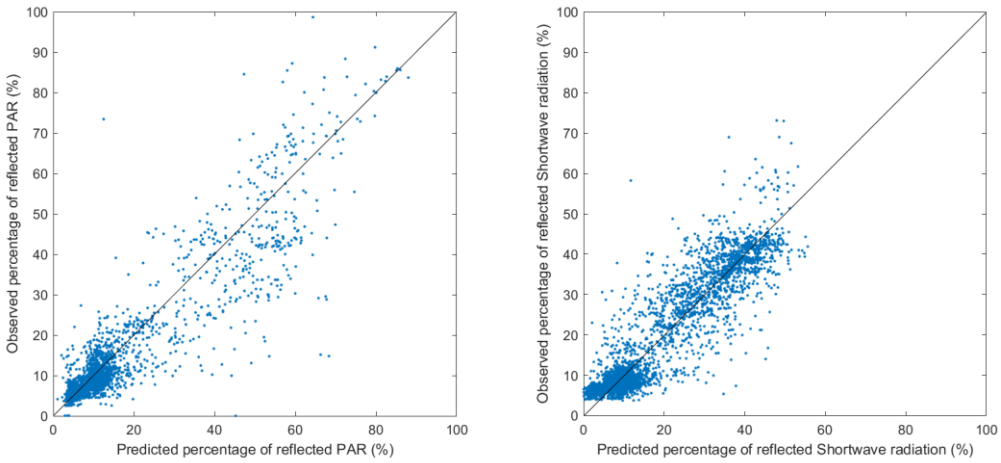

Figure 3: comparison between MAIAC and flux tower derived estimations of the percentage of reflected PAR (left panel, 16 sites, RMSE=6.7%) and shortwave radiation (right panel, 17 sites, RMSE=5.4%).

All MAIAC observations across the extended ABoVE domain, from the year 2000 to 2017 were then averaged for each day. The Fcover of snow, soil and vegetation were computed over a random samle of 300000 observations as the sum of the coefficient a for the corresponding library spectra:

$$F_{cover\ vegetation} = \sum_{i\ vevetation} a_i$$

And $faPAR_{vegetation}$ was computed as:

$$faPAR_{vegetation} = \sum_{i\ vegetation} a_i * faPAR_i$$

As this process is computationally intensive, a 50 trees random forest was then fitted using the MAIAC reflectance data as an input in order to predict the estimated $faPAR_{vegetation}$ across the random sample, and this random forest model was then used to interpolate $faPAR_{vegetation}$ at tile level.

3.2. Carbon flux data processing and light-curve parameters estimation

The availability of a consistently processed dataset of carbon fluxes measurement covering a wide range of biomes and latitudes is a foundation for evaluating Boreal and Arctic ecosystems productivity via the use of remote sensing. This partitioning process has proved to be particularly challenging in ecosystems which exhibit low and highly variable gross productivity and ecosystem respiration such as artic and boreal ecosystems.

Flux-tower level estimations of light-curve parameters are required in order to train the random forest which will be applied to the tile-level MAIAC data, and GPP estimates are required in order to validate the final GPP product.

3.2.1. Dataset acquisition and consolidation

3.2.1.1. Dataset description

This dataset covers 80 flux tower sites. Those sites are located along a SE-NW gradient, spanning from grassland to tundra, including deciduous and evergreen forests (Table. 1).

Site Name	IGBP class	latitude	longitude
CA-ARB	WET	52.694999	-83.945226
CA-ARF	WET	52.70078	-83.955045
CA-CF2	WET	58.6658	-93.83
CA-Ca1	ENF	49.8673	-125.3336
CA-Ca2	ENF	49.8705	-125.2909
CA-Ca3	ENF	49.5346	-124.9004
CA-Cbo	DBF	44.31667	-79.9333
CA-Cha	MF	45.8847	-67.3569
CA-DBB	WET	49.1293	-122.9849
CA-ER1	CRO	43.640458	-80.412303
CA-Gro	MF	48.2167	-82.1556
CA-Let	GRA	49.70928	-112.94017
CA-MA1	CRO	50.164472	-97.876222
CA-MA2	GRA	50.170972	-97.876222
CA-MA3	GRA	50.177417	-97.868639
CA-MR3	GRA	50.8671	-111.9045
CA-MR5	GRA	50.9056	-111.8823
CA-Man	ENF	55.87962	-98.48081
CA-NS1	ENF	55.87917	-98.48389

CA-NS2	ENF	55.90583	-98.52472
CA-NS3	ENF	55.91167	-98.38222
CA-NS4	ENF	55.91437	-98.380645
CA-NS5	ENF	55.86306	-98.485
CA-NS6	OSH	55.91667	-98.96444
CA-NS7	OSH	56.63583	-99.94833
CA-NS8	ENF	55.89806	-98.21611
CA-Na1	ENF	46.4722	-67.1
CA-Oas	DBF	53.62889	-106.19779
CA-Obs	ENF	53.98717	-105.11779
CA-Ojp	ENF	53.91634	-104.69203
CA-Qc2	MF	49.7598	-74.5711
CA-Qcu	ENF	49.26712	-74.0365
CA-Qfo	ENF	49.6925	-74.34206
CA-SCB	WET	61.3089	-121.2984
CA-SCC	ENF	61.3079	-121.2992
CA-SF1	ENF	54.48503	-105.81757
CA-SF2	ENF	54.25392	-105.8775
CA-SF3	OSH	54.09156	-106.00526
CA-SJ1	ENF	53.908	-104.656
CA-SJ2	ENF	53.945	-104.649
CA-SJ3	ENF	53.87581	-104.64529
CA-TP1	ENF	42.66093611	-80.55951944
CA-TP2	ENF	42.77441944	-80.458775
CA-TP3	ENF	42.70681111	-80.34831389
CA-TP4	ENF	42.710161	-80.357376
CA-TPD	DBF	42.635328	-80.557731
CA-WP1	WET	54.95384	-112.46698
CA-WP2	WET	55.5375	-112.3343
CA-WP3	WET	54.47	-113.32
US-A03	BSV	70.495328	-149.882297
US-A10	BSV	71.3242	-156.6149
US-An1	OSH	68.99	-150.28
US-An2	OSH	68.95	-150.21
US-An3	OSH	68.93	-150.27
US-Atq	WET	70.4696	-157.4089
US-Bn1	ENF	63.919813	-145.378178
US-Bn2	DBF	63.919813	-145.378179
US-Bn3	OSH	63.92268	-145.74416
US-Brw	WET	71.3225	-156.60917
US-EML	OSH	63.8784	-149.2536
US-FPe	GRA	48.3077	-105.1019
US-Fcr	OSH	65.3968	-148.9348
US-HVa	WET	69.1423	-148.8412
US-ICh	OSH	68.6068	-149.2958
US-ICs	WET	68.6058	-149.311

US-ICt	OSH	68.6063	-149.3041
US-Ivo	WET	68.4865	-155.7503
US-MSR	CRO	47.47578	-111.7207
US-NGB	SNO	71.280044	-156.609181
US-NGC	GRA	64.8614	-163.7008
US-Prr	ENF	65.12367	-147.48756
US-Rpf	DBF	65.1198	-147.429
US-Uaf	ENF	64.86627	-147.85553
US-Upa	WET	70.28147	-148.88483
US-xBN	ENF	65.15401	-147.50258
US-xDC	GRA	47.16165	-99.10656
US-xDJ	ENF	63.88112	-145.75136
US-xHE	OSH	63.87569	-149.21334
US-xTL	WET	68.66109	-149.37047
US-xWD	GRA	47.12823	-99.24136

Table 1: list of processed flux tower sites

The flux data from these flux tower sites was downloaded from Ameriflux and updated on the 28/07/2020.

These sites are distributed over 9 IGBP classes as show in table 2:

IGBP class	BSV	CRO	DBF	ENF	GRA	MF	OSH	SNO	WET
Description	Barren Sparse Vegetation	Cropland	Deciduous Broadleaf Forests	Evergreen Needleleaf Forests	Grassland	Mixed Forest	Open-Shrubland	Snow and Ice	Permanent Wetlands
Number of sites	2	3	5	30	9	3	12	1	15
Number of years	15	7	61	270	45	18	56	8	79

Table 2: distribution of the sites and site-years across sampled IGBP classes

The following variables were extracted for each site:

FC	PPFD_BC_OUT	RH_PI_F	TA_PI_F
FC_PI_F	PPFD_DIF	SC	TIMESTAMP_START
GPP_PI	PPFD_DIR	SC_PI_F	TS
GPP_PI_F	PPFD_IN	SWC	TS_PI_1
NEE_PI	PPFD_IN_PI_F	SWC_PI_F_1	USTAR
NEE_PI_F	PPFD_OUT	SW_IN	VPD_PI
NETRAD	PPFD_OUT_PI_F	SW_IN_PI_F	VPD_PI_F
NETRAD_PI_F	RECO_PI	SW_OUT	
PPFD_BC_IN	RECO_PI_F	SW_OUT_PI_F	
PPFD_BC_IN_PI_F	RH	TA	

Table 3: list of extracted variables

For variables measured in several locations at the same levels (temperature, incoming PPFD), the replicates were averaged. For variables measured at different levels along a vertical profile, all levels were extracted.

3.2.1.2. PPFD pre-processing

The PPFD measurements suffer from three main issues which generate data gaps or introduce artificial differences between sites. These artefacts would affect the estimated aPAR at the flux tower level, hence the light curve parameters. In order to derive a generic light curve prediction at tile level, these artefacts need to be eliminated. Three different types of measurement error were addressed :

- Sensor failure: sensors failures tend to be relatively frequent in boreal and arctic ecosystems due to environmental constraints (large changes in temperature, intense cold…). A new algorithm was designed to make use of ancillary measurements (Net Radiation, Shortwave radiation, PPFD at different levels) to predict PPFD and gapfill it in case of sensor failure. It uses a regression random forest to predict each variable using all available other variables (temperature, PPFD, SW at different levels) as independent variables. Each random forest is used to predict each variable, validated against the measured data, and used to gapfill the missing data.

- Sensor offset: sudden changes in temperature and moisture can impact the electronics used to process the output of the PPFD sensors and introduce an offset, which may vary with temperature or moisture within the sensor. Such events were automatically detected by fitting a spline to the night-time PPFD data, and the estimated offset was subtracted (Fig. 3)

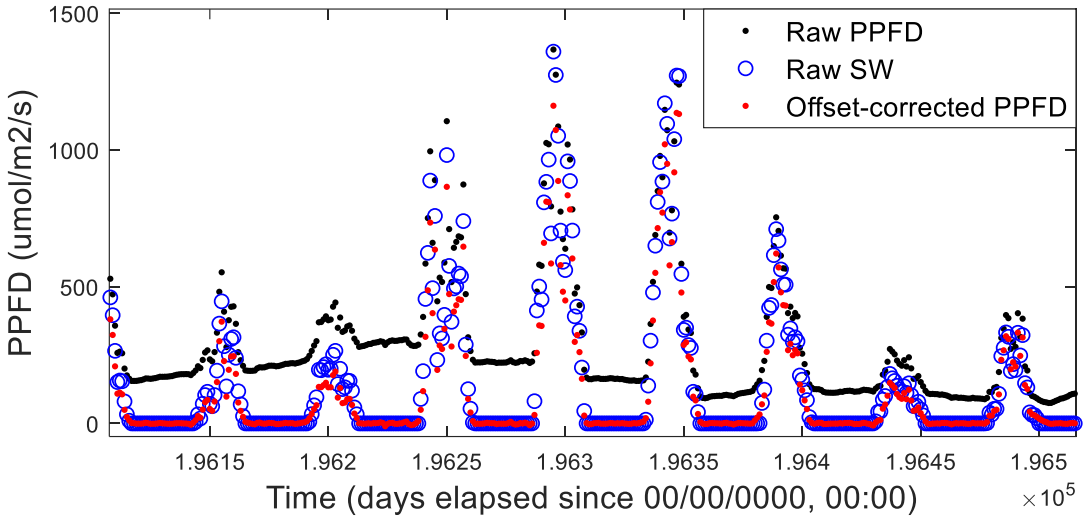

Figure 3: Example of PPFD, rescaled SW and offset-corrected PPFD time series (site CA-Ca1)

- Sensor bias: PPFD sensors over large networks of flux towers are rarely inter-calibrated, and there may be biases which can hamper inter-site comparisons. Moreover, these sensors are exposed to the weather and accumulate dust over time. For sites in remote locations, maintaining the sensors is a challenge, and the accumulation of dust can introduce a significant drift in PPFD measurement. These drifts were estimated by comparing clear sky PPFD measurements against a potential PPFD model (Fig. 4). The bias was estimated using the highest mode of the PPFD versus Potential PPFD ratio over a 1 month moving window, which captures clear sky observations. The resulting ratio is then used to correct the decrease of PPFD sensor sensitivity over time.

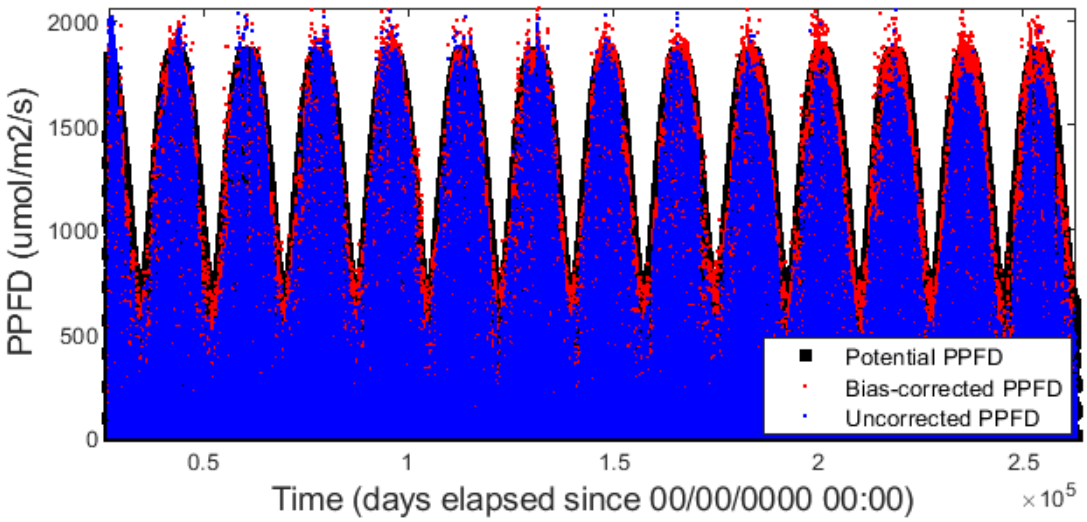

Figure 4: Example of PPFD bias correction (site CA-Ca1)

3.2.1.3. NEE computation

Nee was computed using FC (CO2 flux) and corrected by the storage term SC when available (NEE=FC+SC). As Boreal and Arctic sites can exhibit sudden spikes in NEE, and tend to have low respiration and photosynthesis, hence a low signal versus noise ratio, traditional spike-filtering techniques could not be used. A new filtering technique based on NEE distribution was used to filter out outliers.

3.2.2. NEE U* correction

3.2.2.1. U* damping effect model

Traditional U* filtering (Papale et al. 2006), intended to filter out observations made when the level of air turbulence does not allow accurate carbon fluxes estimations, led to the deletion of up to one third of all observations in arctic sites, particularly during key periods such as the early spring, during the activation of photosynthesis. In low U* conditions, the measured NEE significantly underestimate the real NEE, and this underestimation increases as U* decreases. The relationship between U* and NEE was studied and extracted from the overall NEE variability while accounting for the effect of temperature and PPFD using a Random Forest model, and an analytic model was then formulated in order to describe the damping effect of U* on NEE measurements.

For all site years, the NEE response to U* had a double-sigmoid shape. The following NEE damping ratio model was found to provide the best fit RMSE and the tightest parameters confidence intervals:

$$K_{U^{*}(U^{*})=erf\left(\frac{U^{*}}{a}\right)}$$

With a the fitted parameter. The 95% U* threshold can be retrieved using the following equation:

$$U^{*}_{95\%} = erfinv(0.95) \times a$$

The following respiration model was then fitted to night-time NEE measurements:

$$R_{eco} = K_{U^{*}} \times (b \times faPAR + c) \times \left(0.5 + 0.5 \times erf\left(\frac{T - T_{infl}}{T_{range}}\right)\right)$$

With a, b, c, T_{infl} and T_{range} the fitted parameters.

3.2.2.2. U* correction of carbon fluxes

The obtained a values were then used to compute K_{U^*} for each site-year, and NEE was corrected by dividing it by $K_{U^*}$to account for the damping effect of U*. Examples of uncorrected and corrected NEE are shown in Figure 5.

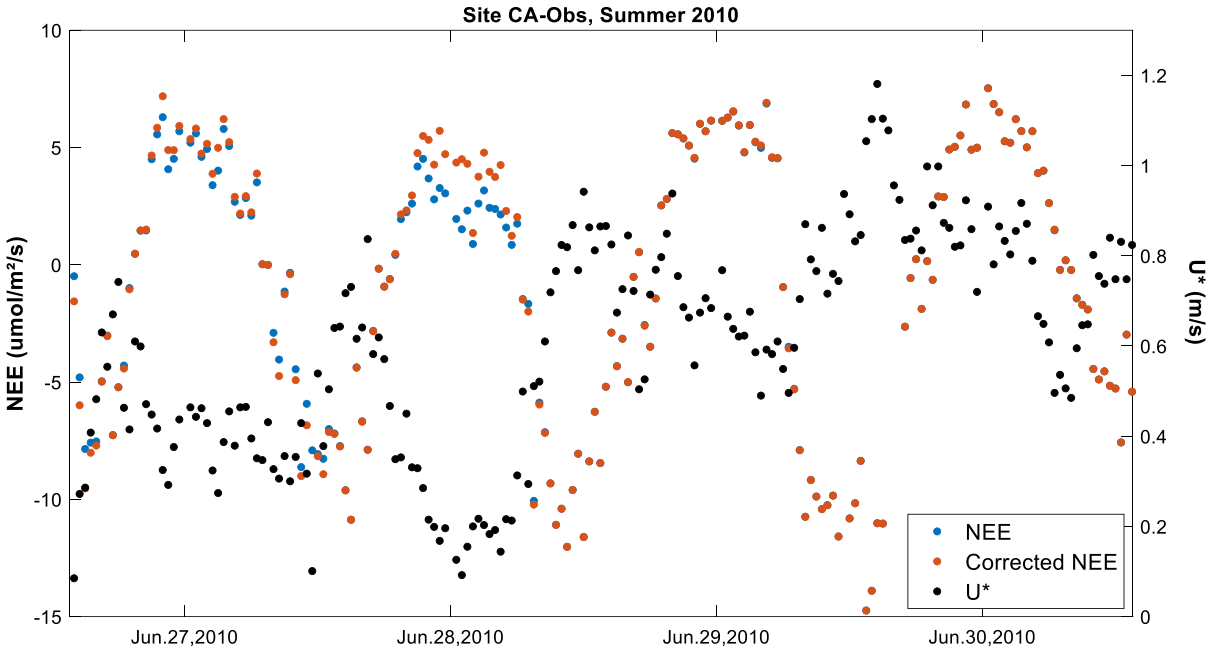

Figure 5: Example of uncorrected and corrected NEE time series.

3.2.3. Partitioning

Carbon fluxes partitioning traditionally rely on fitting a model which describes the relationship between temperature and ecosystem respiration on night-time data and applying it to day-time data to estimate respiration and subtract NEE to obtain GPP. This method is not suited to boreal and arctic ecosystems due to the large differences between night-time and daytime temperature values. Two different light-curve fitting schemes were tested.

3.2.3.1. Light curve models

In order to apply this framework to carbon fluxes partitioning, a temperature/light-curve model was defined. The NEE was defined as:

$$NEE = R_{eco} - GPP \qquad \text{Eq.5}$$

GPP was defined as:

$$GPP(t) = GPP_{max} \times \mathrm{Erf}\left(\frac{PPFD(t) \times faPAR_{vegetation}}{PPFD_{max} \times \frac{1}{erfinv(0.95)}}\right) \qquad \text{Eq.6}$$

With Erf(x) the error function, the integral of a Gaussian function between – infinity and x, GPP_{max} the maximum GPP (GPP at saturation) and $PPFD_{max}$ the PPFD value needed to reach 95% of GPPmax. The $faPAR_{vegetation}$ values used were extracted from the ones computed at tile level in 3.1. This model has an asymptotic exponential shape, and two scaling parameters: one scaling parameter on the GPP axis (GPPmax) and one on the PPFD axis (PPFDmax). The fact that those two

parameters operate on the two different axes of the light-curve limit potential error compensation between them. The theoretical framework behind this model is detailed in Annex 1.

For modelling Reco, the following model was used:

$$Reco(t) = Reco_{max} \times \left(0.5 + 0.5 \times Erf\left(\frac{T_{air}(t) - T_{infl}}{T_{range}}\right)\right) \quad \text{Eq.7}$$

With Recomax the potential respiration (respiration in ideal temperature conditions), Tair the air temperature, Tinfl the temperature at the respiration inflexion point, and Trange the range of temperature within which respiration varies. This model has a sigmoid shape and was chosen to limit error compensation between Recomax and the other two parameters. The light-curve model was fitted on a 15-day moving window, shifting by one day for each fit. For each day, the average and standard deviation for each parameter over the 15 overlapping windows was computed.

The random forest-based partitioning was used as a reference. 3 sites for which the R^2 between the light-curve fit and the random-forest one was lower than 90% were excluded. The light-curve based partitioning was then used to generate light-curve parameters at a daily scale for each site.

3.3. GPP prediction at tile level

3.3.1. Light-curve parameters prediction at tile level

A 200 trees random forest was then fitted using the extracted daily average MAIAC reflectance values, the day of year and $faPAR_{vegetation}$ for the flux tower pixels as independent variables and the obtained light curve parameters GPPmax and PPFD max averaged at daily scale and temporally matched with the MAIAC observations. The obtained Random Forest model was then applied at tile level to derive daily estimations of GPPmax and PPFDmax.

3.3.2. GPP estimation at tile level

The GPP was then estimated at tile level and at hourly resolution by injecting the obtained GPPmax, PPFDmax, $faPAR_{vegetation}$ and ERA-5 Land hourly PPFD estimations.

3.3.3. Error estimation

The obtained simulated GPP were extracted for the flux tower pixels and compared to the random forest derived partitioned GPP. To avoid circularity due to the same light curve model being used to estimate tower-based GPP and to predict the MAIAC-based GPP estimates, a new independent and non-parametric estimate of GPP was derived at flux towers using a random-forest variability partitioning approach.

A random forest was fitted onto night-time and daytime NEE using the variables listed in Table 4. The ecosystem respiration was then estimated as the predicted NEE for a PPFD value forced through zero and U* fixed at its 95% percentile, and ecosystem GPP was estimated as Reco-NEE simulated for a U* fixed at its 95% percentile. This non-parametric estimate of GPP was used to validate the final tile-level GPP model, and to estimate the error. An independent non-parametric estimate of GPP is therefore used instead of GPP estimates derived from the same light-curve model which was used for model training and prediction.

To avoid overestimation of the error due to spatial mismatch between the MODIS MAIAC pixels and the flux tower footprints, a set of sites were selected using the MCD12Q1 yearly land cover product. A total of 17 sites for which a MODIS MAIAC pixel composed of at least 75% of the same land cover

class as the flux tower would be retrieved were selected. The relationship between the daily random-forest based GPP and the predicted MAIAC GPP is shown in figure 6 :

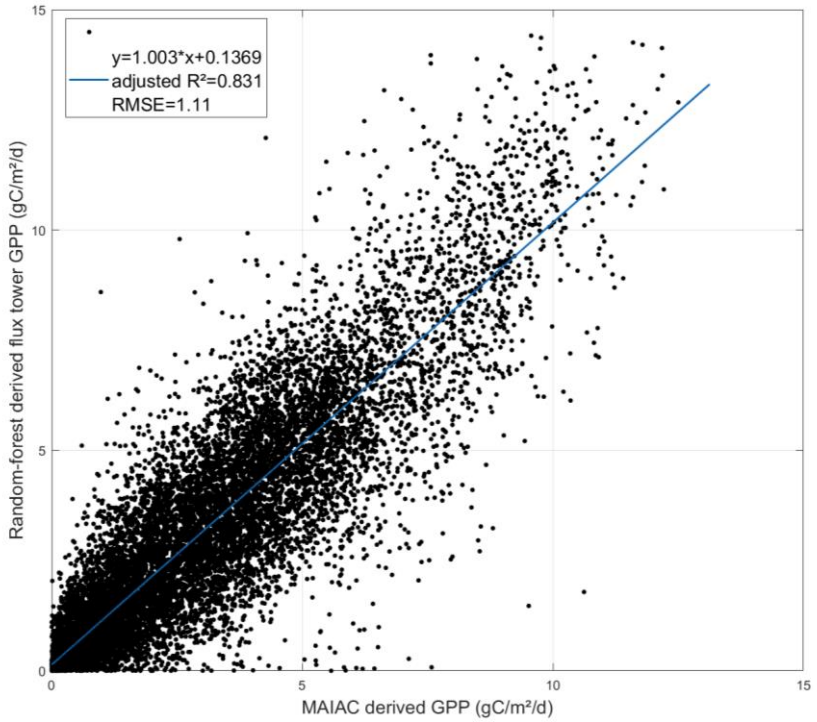

Figure 6: comparison between flux tower derived GPP and MAIAC derived GPP at the daily scale.

The structure of the error was then investigated in order to check for biases and to formulate an error model.

No independent significant trend in RMSE could be found according to IGBP classes, years nor months. The relationship between the predicted GPP and the prediction RMSE is shown in figure 7.

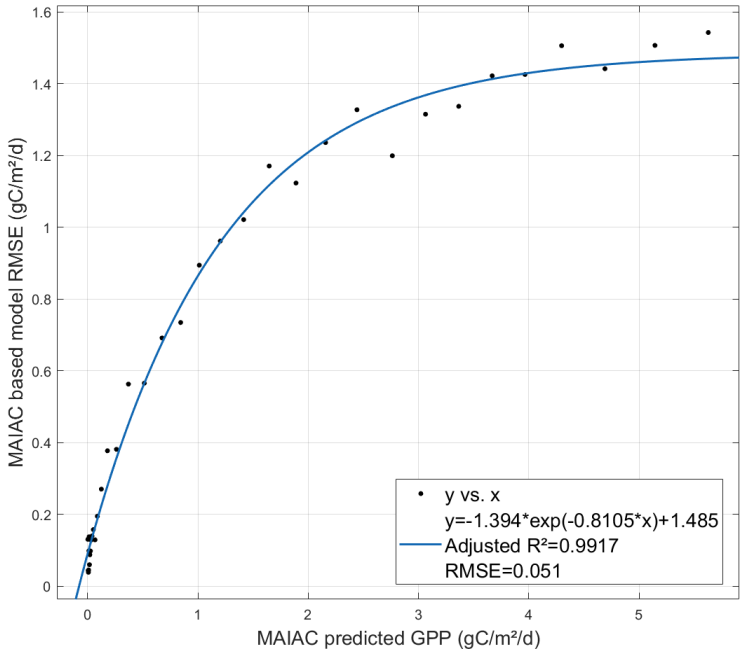

Figure 7: relationship between the predicted GPP and the model root-mean-square deviation.