Exploring the potential of crop specific green area index time series to improve yield estimation at regional scale

Grégory Duveiller[1], Allard de Wit[2], Louis Kouadio[3], Bakary Djaby[3], Yannick Curnel[1,4], Bernard Tychon[3] and Pierre Defourny[1]

[1] *Earth and Life Institute, Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium.*
[2] *Centre for Geo-Information of Alterra (Alterra-CGI), Wageningen UR, Wageningen, The Netherlands.*
[3] *Department of Environmental Sciences and Management, Université de Liege (ULg), Arlon, Belgium*
[4] *Walloon Agricultural Research Centre (CRA-W), Gembloux, Belgium.*

gregory.duveiller@uclouvain.be

ABSTRACT – *Crop status, such as the Green Area Index (GAI), can be retrieved from satellite observations by modelling and inverting the radiative transfer within the canopy. Providing such information along the growing season can potentially improve crop growth modelling and yield estimation. However, such approaches have proven difficult to apply on coarse resolution satellite data due to the fragmented land cover in many parts of the World. Advances in operational crop mapping will sooner or later allow the production of crop maps relatively early in the crop growth season, thereby providing an opportunity to sample pixels from medium/coarse spatial resolution data with relatively high cover fraction of a particular crop type to derive crop specific GAI time series. This research explores how to use such time series derived from MODIS to produce indicators of crop yield using two approaches over part of Belgium. The first method consists in looking at metrics of the decreasing part of the GAI curves when senescence occurs. Such metrics, like the position of the inflexion point, have been shown to be significantly correlated to yield. The second approach is to optimize the WOFOST model used in the European Crop Growth Monitoring System (CGMS) based on the GAI time series. Results show that, although the optimized model shows considerably better performance than the model running on the default parameter, the model is sometimes outperformed by the simpler metric approach. In all cases, indicators including remote sensing information provide better estimates that the average yield of previous years.*

1 INTRODUCTION

Earth observation can bring valuable information for monitoring crop growth and thereby improve crop yield forecasting. Currently operational systems working over large geographic extents, such as the MARS Crop Yield Forecasting System (MCYFS), only rely on remote sensing to complement their analysis based on agro-meteorological crop growth simulations when unexpected circumstances are encountered (e.g. extreme weather conditions, unexpected agricultural practices, uncertain soil conditions, etc.). To provide a diagnostic of the deviation from normal conditions, the NDVI profile of the current year is compared to the average profile over previous years.

A finer description of the crop status can be retrieved from satellite observations in the form of biophysical variables by modelling and inverting the radiative transfer within the canopy. Providing crop specific biophysical variables such as Green Area index (GAI) along the growing season at relevant spatial and temporal resolutions can help improve crop growth modelling either by forcing the model, by recalibrating it or by updating its temporal trajectory using assimilation techniques (Moulin et al. 1998, Dorigo et al. 2007). At field level, such approaches have long been used based on high spatial resolution imagery such as Landsat or SPOT/HRV (e.g. Bouman, 1992; Launay & Guerif, 2005; Hadria et al., 2010). To apply these techniques in an operational crop growth monitoring context, the satellite observations need to be acquired with high temporal frequency and over large geographic extents, conditions currently satisfied only by instruments with coarse pixels (e.g. MODIS or MERIS). In order to keep the information within a pixel crop specific, studies working on assimilation of such observations into crop models generally focus on landscapes with relatively homogeneous land cover (e.g. Bastiaanssen, 2003; Doraiswamy et al., 2005; Patel et al., 2006). However, spatial patterns in

agricultural landscapes are generally much more fragmented with variable requirements in terms of spatial resolution (Duveiller & Defourny 2010).

The current diversity of EO instruments, with wide swath instruments is bound to stimulate the development of techniques to produce crop maps in the current growing season. This will provide an opportunity of sampling pixels from medium/coarse spatial resolution data with a relatively high cover fraction of a particular crop type for deriving crop specific GAI time series (Duveiller et al. 2010, this issue). The aim of the research presented here is to explore two alternatives for using such time series to derive information about crop yield. The first approach consists in looking at metrics of the decreasing part of the GAI curves when senescence occurs. Such metrics, like the position (in degree-days) of the inflexion point, have been shown to be indicators of amount of grain-filling and hence relate to yield (Gooding et al. 2000). The second approach explores the compatibility with the European Crop Growth Monitoring System (CGMS) used in MCYFS. In this study, within-season updates of relevant crop parameters in the WOFOST crop model are applied to improve the crop simulations and yield forecasts.

2 STUDY SITE AND DATA

A critical aspect in relating remote sensing observations with crop specific yield is to know where the target crop has been sown in a given year. In Belgium, such information is available with the vector database of the SIGEC (*Système Intégré de Gestion et de Contrôle*) built by the government of the Walloon region. The extent of the area covered by the SIGEC database is shown on figure 1 and consists of 5 NUTS2 European administrative units for which official yield statistics are available from the EUROSTAT database of the European Commission (http://epp.eurostat.ec.europa.eu). All the fields covered by winter wheat were selected and rasterized to create crop masks for years 2003 to 2007.

The Earth Observation data used in this study consists of daily MODIS reflectance data from both Terra and Aqua platforms downloaded from the NASA Distributed Active Archive Center (DAAC) (https://wist.echo.nasa.gov/api/). Collection 5 products are used, for which atmospherically corrected reflectance is available at 250m in the red and near-infrared spectral domains. Only pixels whose observation footprint overlaps winter wheat crop masks by over 75% are retained. The methodology used to estimate this overlap, or crop specific pixel purity, is described in another study within this book (Duveiller et al. 2010, this issue). The range of selected time series ranges from 3839 to 5017 depending on the year studied.

The WOFOST crop model was implemented over the test area on a 10x10 km grid. Soil maps, weather data and crop parameters were derived from the operational MCYFS and mapped onto the model grid.

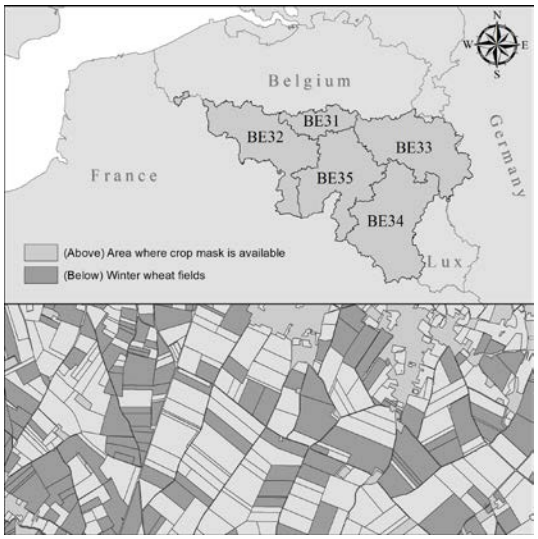

Figure 1. The study area (above) covers 5 NUTS2 administrative regions for which a winter specific crop mask can be built using the SIGEC dataset (below).

3 METHODOLOGY

3.1 Generating crop specific GAI time series

GAI is retrieved from multispectral reflectance using a neural network technique (NNT) trained over canopy radiative transfer simulations. This hybrid approach combines advantages of statistical and physical approaches in biophysical variable retrieval (Dorigo et al. 2007). The approach is based on the algorithm conceived by Baret et al. (2007) to derive the global LAI product developed within the CYCLOPES (Carbon cYcle and Change in Land Observational Products from an Ensemble of Satellites) project from SPOT/VEGETATION data. The radiative transfer model used for the simulations is PROSAIL (Baret et al. 1992), a coupling of the canopy reflectance model SAIL (Verhoef 1984) to the leaf optical properties model PROSPECT (Jacquemoud & Baret 1990). The input of the NNT is red and NIR MODIS reflectance, plus the angles describing the acquisition geometry (view and sun zenith angles and the relative azimuth angle between the imaging instrument and the sun). Once punctual GAI estimations are obtained from the individual observations of both MODIS instruments (Terra and Aqua), a temporal interpolation is applied to combine all the information together for a given spatial point. This interpolation is based on a semi-mechanistic canopy structure dynamic model (CSDM)

which relates GAI to thermal time by way of a simple mathematical relationship representing the combined effect of growth and senescence, taking the form of:

$$GAI\,(tt) = k\,.\,[1/(1+\exp(-a(tt-T_0-T_a)))^c - \exp(b(tt-T_0-T_b)] \quad (1)$$

where a and b define the rates of growth and senescence, c is a parameter allowing some plasticity to the shape of the curve, k is a scaling coefficient and T_0, T_a and T_b are the thermal times of plant emergence, mid-growth and mid-senescence. The driving variable, thermal time (tt), is simply the cumulated daily average temperatures above 0°C (the base temperature of winter wheat below which its growth does not progress).

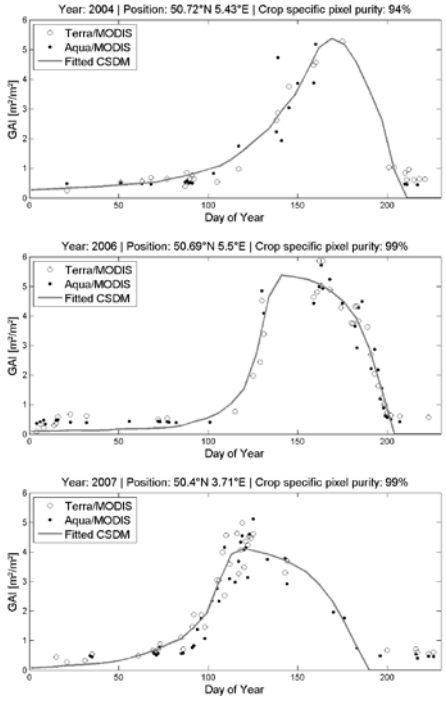

Figure 2. Examples of punctual GAI estimates from both MODIS instruments and the resulting CSDM fit for 3 different time series in different years.

The punctual GAI estimates and the CSDM fit are calculated for all selected pixel time series over the study area and for every year considered. Figure 2 illustrates the results for 3 different GAI time series selected from 3 different years, reflecting the inter-annual variability in the GAI dynamic and in MODIS data availability due to cloud cover.

3.2 Filtering out inadequate time series

Working with MODIS pixels over a fragmented agricultural landscape such a Belgium can result in noisy time series due to signal contamination from adjacent land covers. To ensure that the following processing steps relate only to GAI time series that make agronomic sense, those that do not satisfy the following criteria were discarded:

- Number of observations in growing season >=9 (roughly one every 10 days)
- Maximum GAI value reached by the CSDM > 3.5
- Day of year when maximum GAI is reached by the CSDM must be between 120 and 180
- RRMSE betwee

For finding the optimum model fit, the Weighted Mean Absolute Error (WMAE) between the modelled *GAI* (GAI_m) and the *n* MODIS observed *GAI* (GAI_o) was minimized:

$$\text{WMAE} = \frac{\sum_{1}^{n} w \cdot \left| GAI_o - GAI_m \right|}{\sum_{1}^{n} w} \quad (4)$$

The weight of a *GAI* observation is inversely proportional to the number of observations in a temporal window (size 11) around each observation.

The optimisation was implemented through an exhaustive search algorithm which tested all combinations of TDWI/SPAN within a specified search domain. The range of TDWI was set between 50 and 500 kg/ha in steps of 10 kg/ha while the range of SPAN was set between 20 and 50 days in steps of 1 day.

The optimization procedure was applied on all selected GAI time series and over all available years. This approach yields the joint distribution of the SPAN/TDWI parameters for the Walloon region for each year.

3.5 Aggregating towards regional level

The GAI metrics were aggregated towards NUTS regions by taking the mean value of each metric for each NUTS region for each year. Although this approach is simple, there are questions about the representativeness of this approach given that the available GAI time-series differ strongly from year to year and region to region.

Therefore, the WOFOST simulated biomass values resulting from the optimization described in section 3.4 were not directly averaged to NUTS regions. Instead, we assumed that the joint distribution of TDWI/SPAN values was representative for the entire Walloon region. Next, for each 10x10 km grid, WOFOST was applied in ensemble-mode where the TDWI/SPAN values used were sampled from the TDWI/SPAN joint distribution of each year. From the ensemble of simulated biomass values (total biomass and yield) the average values per grid were calculated and spatially aggregated to regional level using the area of wheat per 10x10 grid as weight.

This procedure was repeated for the baseline run where a single WOFOST run was executed using the default values for TDWI/SPAN.

3.6 Comparison with EUROSTAT statistics

A direct comparison with EUROSTAT yield statistics is not useful as there are substantial biases between EUROSTAT and WOFOST modelled yields. Moreover, the GAI metrics cannot be compared directly as their units deviate from the EUROSTAT yields. All outputs are therefore treated as indicators which are used in a regression model between EUROSTAT yield as the variable to be predicted and the various indicators as explanatory variables. Table 1 lists the indicators used in the regression models with the EUROSTAT yield statistics. Note that two different WOFOST outputs (yield and total biomass) in either default or ensemble (i.e. optimized) mode are used as indicators. Note further that multiple linear regressions were not used to avoid possible collinearity problems between indicators.

Table 1. List of indicator codes and names.

Code	**Indicator name**
01_BE	WOFOST Total biomass – *Ensemble*
02_YE	WOFOST Yield – *Ensemble*
03_BD	WOFOST Total biomass – *Default*
04_YD	WOFOST Yield – *Default*
05_MG	*M* parameter in the Gompertz model
06_KG	*K* parameter in the Gompertz model
07_GM	GAI_{max}
08_ML	*M* parameter in the logistic model
09_KL	*K* parameter in the logistic model

4 RESULTS

Regression analysis between indicators and EUROSTAT reported yields is done separately for each NUTS region as there may be large differences between NUTS regions as a result of socio-economic conditions. Table 2 lists, for each NUTS region, the statistical properties of the 4 best predictors and the average yield as the baseline predictor (the *None* model). Note that region BE34 was due to a lack of sufficient time series in the area. The reason for this is that winter wheat is seldom cultivated in this region, and when it is, the fields used are often too small for MODIS pixels.

For region BE31, the K parameters of the Gompertz and Logistic model are the best predictors with nearly 60% of variance explained. GAI_{max} explains nearly 50%, while the Ensemble total biomass is the next best predictor with only 21%. Only the two best predictors improve beyond the average as their Leave-One-Out (LOO) error is lower than the one of the *None* model. However, none of the predictors is significant.

For region BE32, the GAI_{max} is by far the best predictor with 87% variance explained, a LOO error clearly smaller then the *None* model and a highly significant T-value. The remaining models, (K parameters of the Gompertz and Logistic model, Ensemble total biomass) show similar R^2 values but strongly varying LOO values.

For region BE33, the M parameters of the Gompertz and Logistic model are the best predictors with 69% and 74% variance explained. The latter is significant at alpha=0.1 and has a LOO lower than the *None* model. The remaining two models (Ensemble total biomass and GAI_{max}) show similar R^2 (~50%) but do not improve the LOO. Moreover the Ensemble total biomass has a negative T-value indicating that the relationship with EUROSTAT yield is negative.

For region BE35, the Ensemble total biomass is the best predictor with 75% of variance explained, an LOO clearly lower then the *None* model and a T-value significant at alpha=0.1. The remaining models (M parameters of the Gompertz and Logistic model, GAI_{max}) show R^2 values ranging from 30% to 50%, but none of these models have a LOO value lower than the *None* model.

Table 2. Results from regression with EUROSTAT reported yields. SDr=Std. Dev. of the model residuals. LOO=leave-one-out error. Stud.T= Student-T statistic.

Brabant Wallon (BE31)	*R^2*	*SDr*	*LOO*	*Stud.T*
09_KL	59.5	0.08	0.10	2.10
06_KG	57.7	0.09	0.11	2.02
07_GM	50.1	0.09	0.16	1.74
02_YE	21.1	0.12	0.19	-0.90
None	-	0.12	0.13	-
Hainaut (BE32)				
07_GM	87.2	0.23	0.40	4.52**
09_KL	51.3	0.45	0.51	1.78
06_KG	45.9	0.48	1.22	1.60
01_BE	42.4	0.49	0.71	1.49
None	-	0.56	0.63	-
Liege (BE33)				
08_ML	74.1	0.25	0.40	2.93*
05_MG	68.8	0.27	0.91	2.57
01_BE	48.9	0.35	0.73	-1.69
07_GM	46.1	0.36	0.65	1.60
None	-	0.42	0.47	-
Namur (BE35)				
01_BE	74.8	0.26	0.36	2.98*
07_GM	49.1	0.37	1.04	1.70
05_MG	33.9	0.42	0.80	1.24
08_ML	30.5	0.43	0.75	1.15
None	-	0.45	0.50	-

** significant at alpha=0.05
* significant at alpha=0.1

5 DISCUSSIONS

The objective of our approach was to evaluate if crop specific MODIS-derived GAI time-series can be used to derive crop yield indicators that better characterize the inter-annual variability in wheat yields as reported by EUROSTAT. This objective was tested through the use of simple metrics derived directly from the GAI time-series and through assimilation of these in a biophysical model.

The results demonstrate that in all four regions indicators can be found that improve the LOO error beyond the *None* model (the baseline predictor based on the average). Region BE31 is the only region where none of the indicators has a significant T-value. However, region BE31 has very low variability in crop yield with a standard deviation of only 0.12 ton/ha.

When looking at distribution of the four best performing indicators over the regions, it is the GAI_{max} which is consistently listed among the best four. The ensemble total biomass is present in 3 out of 4 regions, with the remaining region (BE31) listing the ensemble yield.

The K parameters of the Gompertz and Logistic models are selected for regions BE31 and BE32, while the M parameters of the Gompertz and Logistic model are selected for BE33 and BE35. Conversely, the M parameters have no performance at all in regions BE33 and BE35, and similarly the K parameters not in regions BE31 and BE32. Although these indicators are among the best performing in some regions, the inconsequent behaviour of these indicators needs further investigation using a larger dataset.

Finally, the indicators which do not include any remote sensing data ('03 Total biomass – default' and '04 Yield – default') are not listed for any region showing that they are not correlated with the reported crop yield at all. This is a strong indication that the MODIS-derived GAI time-series are improving the prediction of crop yield at regional level.

6 CONCLUSIONS

The main conclusion from this work is that, for all regions studied, indicators derived from GAI time-series estimated from MODIS are generally better predictors of the EUROSTAT reported crop yield than average yield. It also shows that the WOFOST model optimized on the GAI time-series (WOFOST ensemble results) shows considerably better performance than the model running on the default parameters. Nevertheless, the WOFOST ensemble results are outperformed by more simple indicators in 3 out of 4 regions, albeit different indicators for each region.

The current analysis only spans 5 years of EUROSTAT reported yields (2003-2007) over 5

regions in Belgium. Longer time-series and more regions will be needed to confirm those results and obtain a better insight in the stability of the different indicators.

ACKNOWLEDGEMENTS

This research was funded by the Belgian Fond de la Recherche Scientifique-FNRS by way of a PhD grant to the first author. The research also falls in the framework of the GLOBAM project which is financed by the Belgian Scientific Policy (BELSPO) with the STEREO II programme. The authors also thank the Ministry of the Walloon Region (Belgium) for providing the SIGEC database.

REFERENCES

Baret, F., Hagolle, O., Geiger, B., Bicheron, P., Miras, B., Huc, M., Berthelot, B., Nino, F., Weiss, M., Samain, O., Roujean, J. L. & Leroy, M. (2007), 'LAI, fAPAR and fCOVER cyclopes global products derived from vegetation: Part 1: Principles of the algorithm', *Remote Sensing of Environment* **110**(3), 275–286.

Baret, F., Jacquemoud, S., Guyot, G. & Leprieur, C. (1992), 'Modeled analysis of the biophysical nature of spectral shifts and comparison with information-content of broad bands', *Remote Sensing of Environment* **41**(2-3), 133–142.

Bastiaanssen, W. & Ali, S., 2003. 'A new crop yield forecasting model based on satellite measurements applied across the Indus Basin, Pakistan'. *Agriculture, Ecosystems & Environment*, **94**(3): 321-340.

Bouman, B.A.M., (1992). 'Linking Physical Remote-Sensing Models with Crop Growth Simulation-Models, Applied for Sugar-Beet'. *International Journal of Remote Sensing*, **13**(14): 2565-2581.

Doraiswamy, P., Sinclair, T.R., Hollinger, S., Akhmedov, B., Stern, A., & Prueger, J., (2005). 'Application of MODIS derived parameters for regional crop yield assessment'. *Remote Sensing of Environment*, **97**(2): 192-202.

Dorigo, W., Zurita-Milla, R., de Wit, A., Brazile, J., Singh, R. & Schaepman, M. (2007), 'A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling', *International Journal of Applied Earth Observation and Geoinformation* **9**, 165–193.

Duveiller, G., Weiss, M., Baret, F., de Wit, A., & Defourny, P. (2010), Retrieving crop specific green area index from remote sensing data when the spatial resolution is close to the target field size, *Proceedings of the 3rd International Symposium on Recent Advances in Quantitative Remote Sensing (RAQRS'III)*, Torrent (Valencia), Spain, 2010, edited by J. Sobrino.

Duveiller, G. & Defourny, P. (2010), 'A conceptual framework to define the spatial resolution requirements for agricultural monitoring using remote sensing', Remote Sensing of Environment 114(11), 2637–2650. Jacquemoud, S. & Baret, F. (1990), 'Prospect: A model of leaf optical properties spectra', *Remote Sensing of Environment* **34**(2), 75–91.

Gooding, M.J., Dimmock,J.P.R.E., France, J., & Jones, S.A, (2000) 'Green leaf area decline of wheat flag leaves: the influence of fungicides and relationships with mean grain weight and grain yield', *Annals of Applied Biology* **136**, 77–84

Hadria, R., Duchemin, B, Jarlan, L.; Dedieu, G., Baup, F., Khabba, S., Olioso, A. & Le Toan, T., (2010). 'Potentiality of optical and radar satellite data at high spatio-temporal resolutions for the monitoring of irrigated wheat crops in Morocco'. *International Journal of Applied Earth Observation and Geoinformation*, **12**: S32-S37.

Kouadio, L., Duveiller, G., Djaby, B., Defourny, P., & Tychon, B. (2010) Wheat Yield Estimates at NUTS-3 level using MODIS data: an approach based on the decreasing curves of green area index temporal profiles. *Proceedings of RSPSoc2010 Annual Conference*, 1st-3rd September 2010, Cork, Ireland (Nottingham: RSPSoc), pp. 214-221.

Launay, M. and Guerif, M., (2005). 'Assimilating remote sensing data into a crop model to improve predictive performance for spatial applications'. *Agriculture Ecosystems & Environment*, **111**(1-4): 321-339.

Moulin, S., Bondeau, A. & Delecolle, R. (1998), 'Combining agricultural crop models and satellite observations: from field to regional scales', *International Journal of Remote Sensing* **19**(6), 1021–1036.

Patel, N.R., Bhattacharjee, B., Mohammed, A.J., Tanupriya, B. and Saha, S.K., (2006). 'Remote sensing of regional yield assessment of wheat in Haryana, India'. *International Journal of Remote Sensing*, **27**(19): 4071-4090.

Verhoef, W. (1984), 'Light scattering by leaf layers with application to canopy reflectance modeling: The sail model', *Remote Sensing of Environment* **16**(2), 125–141.