REPLACING RADIATIVE TRANSFER MODELS BY SURROGATE APPROXIMATIONS THROUGH MACHINE LEARNING

Jochem Verrelst[1]*, Juan Pablo Rivera*[1]*, Jose Gómez-Dans*[2]*, Gustau Camps-Valls*[1] *and Jose Moreno*[1]

[1]Image Processing Lab (IPL), University of Valencia, Spain; Jochem.verrelst@uv.es
[2]University College London, UK

ABSTRACT

Physically-based radiative transfer models (RTMs) help in understanding the processes occurring on the Earth's surface and their interactions with vegetation and atmosphere. However, advanced RTMs can take a long computational time, which makes them unfeasible in many real applications. To overcome this problem, it has been proposed to substitute RTMs through so-called *emulators*. Emulators are statistical models that approximate the functioning of RTMs. They are advantageous in real practice because of the computational efficiency and excellent accuracy and flexibility for extrapolation. We here present an 'Emulator toolbox' that enables analyzing three multi-output machine learning regression algorithms (MO-MLRAs) on their ability to approximate an RTM. As a proof of concept, a case study on emulating sun-induced fluorescence (SIF) is presented. The toolbox is foreseen to open new opportunities in the use of advanced RTMs, in which both consistent physical assumptions and data-driven machine learning algorithms live together.

Index Terms— Emulator, ARTMO, Fluorescence, FLEX, multi-output regression algorithms

1. INTRODUCTION

With the forthcoming superspectral satellite missions dedicated to land monitoring, as well as the planned imaging spectrometers, an unprecedented data stream will soon become available. The requirements for such a large data stream involve processing techniques enabling the spatio-temporally explicit quantification of vegetation properties. These must be retrieved with accurate, robust and fast methods. Physically-based model inversion methodologies are based on physical laws and established cause-effect relationships. Typically, radiative transfer models (RTMs) are inverted against remote sensing images to retrieve state variables. Nevertheless, these approaches, although considered as physically-sound are not straightforward. Various choices have to be made. Firstly, an RTM has to be selected whereby a trade-off between the realism and inversion possibility of the RTM has to be made. Secondly, the limited information content in the EO data and the strong non-linearities of the RTM results in non-unique solutions, that is usually dealt with a Bayesian treatment with the imposition of prior knowledge. These approaches are numerically costly, and a number of inversion strategies have been proposed. Most practical implementations are based on look-up tables (LUTs).

In a LUT approach the RTM generates spectral outputs for a large range of combinations of variable values. As such, the inversion problem is reduced to the identification of the modeled reflectance set that resembles most closely the measured one. This process is based on querying the LUT and applying a cost function on a pixel-by-pixel basis. In order to produce accurate mappings, LUTs need to have fine sampling in parameter space, which results in a very large number of RTM runs, which is a computationally demanding task. One approach to mitigate this has been to use surrogate models of the RTM, often called *emulators* [1]. Emulators are statistical constructs that are able to approximate the RTM, although at a fraction of the computational cost, and in some cases, providing an estimation of uncertainty, [1].

Over the last few years, various RTMs have been brought together and standardized within a toolbox called ARTMO (Automated Radiative Transfer Models Operator) [2]. These RTMs can be operated in a semi-automated fashion for any kind of optical sensor operating in the visible, near-infrared and shortwave infrared range (400-2500 nm). Having multiple RTMs available, this platform would serve perfectly to the development of an emulator toolbox. This brings us to the following main objective: to develop an emulator toolbox that is able to reproduce the outputs of any of the RTMs available within ARTMO. A second objective is to demonstrate the utility of the toolbox.

2. ARTMO

The ARTMO Graphic User Interface (GUI) is a software package that provides essential tools for running and inverting a suite of optical RTMs, both at the leaf and at the canopy level. ARTMO facilitates consistent and intuitive user interaction, thereby streamlining model setup, running, storing and output plotting. Essentially, ARTMO allows: (1)

to configure and run leaf and canopy RTMs, independently or combined, in an intuitive way through various GUIs with input options to insert single values, value ranges, or imported external datasets; (2) to simulate and store a massive quantity of spectral output based on LUT approach in a relational database; (3) to plot groups of simulated spectra in the same plotting window with color gradients as a function of input parameters; (4) to export simulated spectra and associated meta-data to a text file for further processing; (5) to analyze and apply retrieval techniques in order generate maps of biophysical parameters from optical remote sensing imagery. Currently the following RTMs are implemented: (1) at the leaf scale: PROSPECT-4, PROSPECT-5, DLM, LIBERTY; (2) at the canopy scale: SAIL, FLIGHT, INFORM; and combined: SCOPE. The toolbox is freely downloadable at http://ipl.uv.es/artmo/.

3. ARTMO'S EMULATOR TOOLBOX

The emulator tool was developed based on ARTMO's machine learning regression algorithm (MRLA) toolbox [3]. From this toolbox the following three MLRAs possess multi-output (MO) predictive capabilities and are therefore used as basis for the emulator: (1) partial least squares regression (PLSR), (2) neural networks (NN) and (3) kernel ridge regression (KRR). Similarly as in the MLRA toolbox, the MO-MLRAs need to go through a training phase, which can be based on RTM or on field data. Whereas the MLRA toolbox trains nonparametric regression models to retrieve biophysical parameters, the emulator toolbox is built the other way around and instead generates output spectra. Of importance hereby is the more closely the emulated spectra resembles the RTM-generated spectra, the better the approximation can function as an emulator of the RTM. Therefore, an important step is to validate the MLRA model. In short, the emulator toolbox allows the user to control the following steps:

1. **Input**: selection of ARTMO-generated LUT or external data. Input variables can be selected.
2. **Settings**: selection of a MO-MLRA, splitting of the data into training and validation. Cross-validation sampling options are provided and multiple MO-MLRAs can be selected.
3. **Validation**: This step validates the configured MO-MLRAs though the root mean square error (RMSE) difference between validation spectra and emulated spectra. The GUI provides an overview table, which allows selecting the best-performing MO-MLRA model.
4. **Emulator test**: The chosen surrogate model can be tested against the actual RTM for user-defined input. As such, both outputs are visualized and the accuracy and gain in processing speed are calculated.
5. **Output**: Finally, the chosen model can be applied either to generate a LUT or even to be applied in ARTMO's scene generator to generated simulated scenes.

4. CASE STUDY: EMULATING SIF PROFILES

As a proof of concept, we subsequently applied it for evaluating the performance of the three MO-MLRAs on their capability to emulate the SVAT model SCOPE. SCOPE is essentially an energy budget model that calculates the whole energy budget of a canopy, with sun-induced chlorophyll fluorescence (SIF) as one of their outputs. These simulations are used within applications of ESA's candidate EE8 FLuorescence Explorer (FLEX) mission, e.g. for the development of artificial scenes as observed by FLEX and for sensitivity studies. Here we will evaluate whether the emulator reaches acceptable accuracies and how much processing speed is gained. SCOPE is first briefly outlined, followed by the experimental setup of the LUT generation. Emulating results are then presented and discussed.

4.1. Simulated data: SCOPE v1.60

SCOPE is a vertical (1-D) integrated radiative transfer and energy balance SVAT model [4]. It calculates radiation transfer in a multilayer canopy, in order to obtain reflectance and SIF in the observation direction as a function of the solar zenith angle and leaf inclination distribution. The distribution of absorbed radiation within the canopy is calculated with the SAIL model. The distribution of absorbed radiation is further used in a micro-meteorological representation of the canopy for the calculation of photosynthesis, fluorescence, latent and sensible heat. The fluorescence and thermal radiation emitted by individual leaves is finally propagated though the canopy.

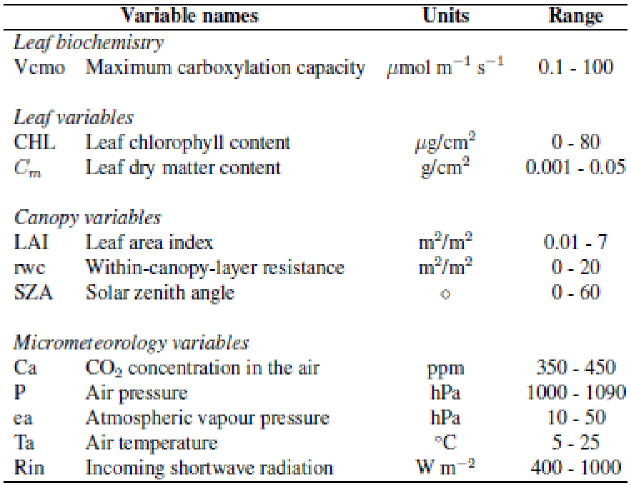
Compared to an earlier release, various improvements have been included in the new SCOPE v1.60, such as processing speed-up through parallel computing routines. Nevertheless, SCOPE v1.60 still takes about one second to finalize a single simulation. Because SCOPE v1.60 is equipped with over 30 input variables and offers a wide range of output products, all types of input-output sensitivity studies can be conducted. However, this comes at a computational cost. In view of FLEX, we are mostly interested in SIF outputs. We will therefore examine the capability of the MO-MLRAs to emulate SCOPE SIF profiles.

4.2. Experimental setup

Although SCOPE is equipped with over 30 input variables, not all of them play a role in the generation of SIF outputs. To find out their relative importance, in an earlier work a global sensitivity analysis (GSA) was employed [5]. It was found that 11 key variables explained 95.5% of the variance of total SIF (integral of the fluorescence broadband signal). These variables, listed in Table 1, were therefore used to generate SCOPE LUTs.

A fully random LUT within the variable space with min-max boundaries as given in Table 1 and a uniform distribution was generated using SCOPE v1.60 for 100, 500 and 1000 samples. Their processing time is given in Table 2. These LUTs were then entered into the Emulator toolbox, with the SIF variable as selected output. Within the 'Settings' window, for each LUT all three MO-MLRAs were selected. In order to speed up the model development, prior to train the MO-MLRAs a PCA was applied and the first 5 components were retained and used for data projection. Further, in order to generate more robust validation results, a 10-fold cross-validation sub-sampling procedure was applied. The generated RMSE statistics are then averaged over the multiple training and test subsets.

Table 1. SCOPE input variables that drive canopy-leaving fluorescence and their ranges.

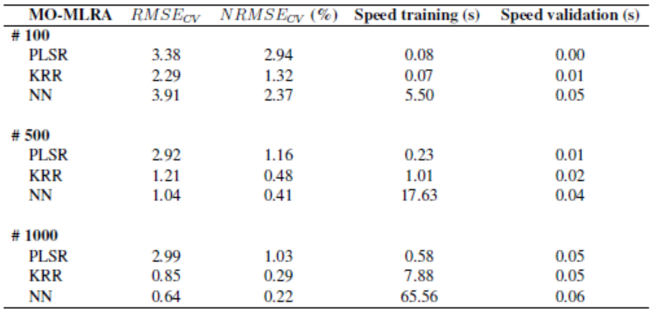


Variable names		Units	Range
Leaf biochemistry			
Vcmo	Maximum carboxylation capacity	μmol m^{-1} s^{-1}	0.1 - 100
Leaf variables			
CHL	Leaf chlorophyll content	μg/cm^2	0 - 80
C_m	Leaf dry matter content	g/cm^2	0.001 - 0.05
Canopy variables			
LAI	Leaf area index	m^2/m^2	0.01 - 7
rwc	Within-canopy-layer resistance	m^2/m^2	0 - 20
SZA	Solar zenith angle	°	0 - 60
Micrometeorology variables			
Ca	CO$_2$ concentration in the air	ppm	350 - 450
P	Air pressure	hPa	1000 - 1090
ea	Atmospheric vapour pressure	hPa	10 - 50
Ta	Air temperature	°C	5 - 25
Rin	Incoming shortwave radiation	W m^{-2}	400 - 1000

5. RESULTS SIF EMULATION

Table 2 displays the RMSE$_{CV}$ goodness-of-fit statistics of the validation dataset and the training and validation processing computational cost of the three MO-MLRAs for the 100, 500 and 1000 random samples datasets. The normalized RMSE (NRMSE$_{CV}$) indicates that relative errors fall below 3%, but significant differences across the three MO-MLRAs can be observed. For the three exercises PLSR performed poorest in accuracy. NN was validated as best performing for the datasets of 500 and 1000 samples, closely followed by KRR. Relative errors fell below 0.5%. Therefore, NN and KRR's predictive accuracy improved when more samples are given to the model. However, NN needed significantly more time to train the model. Note that the PCA transformation considerably improved computational efficiency the training phase; without PCA it took up to a few hours to develop the NN model. In turn, KRR, although being slightly less accurate than NN, needed only a few seconds to train a model. While these goodness-of-fit statistics provided a general indication of the model performance. To visualize the ability of these models to emulate SCOPE SIF outputs, the best and worst matching emulation are plotted for the three MO-MLRAs for the case of the 1000 samples (see Fig. 1). It can be observed that for each MO-MRLA the best validated SIF profile perfectly matched the original SCOPE profile. More interesting is to inspect the worst emulated SIF profile. Large differences can be observed in case of PLSR; it completely missed the close-to-zero SIF profile. Also KRR overestimated a weak SIF profile, but considerably less pronounced. Interestingly, for NN a similar weak SIF profile was encountered as best matching. Here, as worst match, a slight overestimation occurred for a pronounced SIF profile. Considering the close approximation of the SCOPE SIF profile, it shows the powerful potential of NN to approximate the physical RT model SCOPE.

Table 1. MO-MLRAs goodness-of-fit results and processing speed for 100, 500 and 1000 SCOPE samples.



MO-MLRA	$RMSE_{CV}$	$NRMSE_{CV}$ (%)	Speed training (s)	Speed validation (s)
# 100				
PLSR	3.38	2.94	0.08	0.00
KRR	2.29	1.32	0.07	0.01
NN	3.91	2.37	5.50	0.05
# 500				
PLSR	2.92	1.16	0.23	0.01
KRR	1.21	0.48	1.01	0.02
NN	1.04	0.41	17.63	0.04
# 1000				
PLSR	2.99	1.03	0.58	0.05
KRR	0.85	0.29	7.88	0.05
NN	0.64	0.22	65.56	0.06

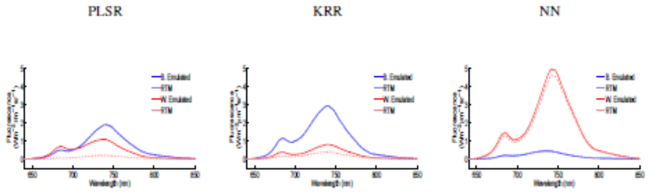


Fig. 1. Best (B) [blue] and worst (W) [red] emulated [solid line] vs. reference RTM SCOPE [dashed line] fluorescence spectra in case of 1000 samples (10-k CV).

The MO-MLRA models were used to generate emulated SIF profiles for the input data of the 100, 500 and 1000 random samples within the Table 1-defined input boundaries. As such the gain in processing speed can be compared to the original simulations be derived. The processing time of SCOPE and the emulator were recorded and the gain in processing times was calculated. It can be viewed (see Table 3) that the emulator reconstructs the SIF profiles much faster than the original SCOPE RTM. Approximately, NN delivers SIF about 50 times faster, PLSR about 400 times faster and KRR even about 800 times faster than the SCOPE model. Hence, given that KRR is also fast in training, it is a promising MO-MLRA to function as emulator; it is about 16 times faster than NN and almost as accurate.

Table 3. SCOPE and MO-MLRA processing speed and gain in speed for 100, 500 and 1000 samples.

SCOPE (s)	MO-MLRA	Emulator (s)	gain in speed (x)
# 100			
	PLRS	0.24	312
75	KRR	0.08	937
	NN	1.58	47
# 500			
	PLRS	0.88	439
386	KRR	0.47	821
	NN	7.69	50
# 1000			
	PLRS	1.83	423
774	KRR	0.98	790
	NN	15.27	51

Finally, to illustrate the performance of the MO-MLRA on their ability to reconstruct SIF profiles, they are visualized for the 1000 samples in Fig. 2. The top-left panel displays the original SCOPE SIF profiles, the other panels display the 1000 emulated profiles for the three MO-MLRA models. Although these profiles were generated by 11 randomly varying variables, the profiles were color-plotted as a function of Cab and LAI. With these graphs it can be observed that PLSR cannot be considered as an accurate emulator; PLSR does not reach the same magnitude as the original SCOPE profiles and, more problematic, leads to negative SIF profiles. This effects were actually expected because PLS, even being a supervised regression algorithm, can only find orthogonal transforms (rotations) and apply a linear regression model. In turn, KRR and NN delivered much more realistic profiles and can cope with the nonlinearities of the problem; they are within the same magnitudes as the original SCOPE profiles; and only a few profiles turn out to fall slightly below zero. For the large majority of samples, KRR and NN reconstructed the 1000 SIF spectra with precision.

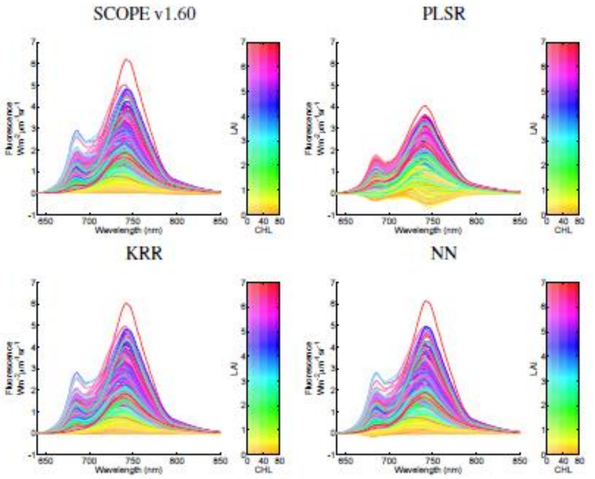


Fig. 2. Original 1000 SCOPE-generated SIF spectra (top-left) and emulated 1000 spectra with three MO-MLRA models. The SIF spectra are color-scaled against LAI and CHL.

6. CONCLUSIONS

Emulators are statistical constructs that approximate the functioning of a physically-based RTM. They provide great savings in memory and tremendous gains in processing speed while yielding similar accuracies. This emulating approach opens many new research and operational remote sensing opportunities. To facilitate the use of emulators, ARTMO's new 'Emulator toolbox' enables analyzing three multi-output machine learning regression algorithms (MO-MLRAs), both linear (PLSR and nonlinear (KRR, NN). The toolbox enables the user to train the MO-MLRA models with data coming from RTMs that are available within ARTMO. Various options are provided that can optimize the training phase, such a PCA pre-processing step, ranging training/validation distributions or through cross-validation sub-sampling procedures. Performance and processing speed of the MO-MLRAs are then calculated. A successfully validated MO-MLRA can then function as emulator.

We analyzed the ability of the implemented MO-MLRAs to substitute the SVAT model SCOPE in the generation of sun-induced fluorescence (SIF) outputs. NN and KRR emulated SIF profiles with great precision (relative errors below 0.5% when trained with 500 or more samples), and this with a gain in processing speed of about 50 (NN) up to about 800 (KRR) times faster than SCOPE v1.60. It is foreseen that the emulator toolbox will open up a diverse range of new applications using advanced RTMs, such as improved inversion strategies, and rendering of simulated scenes in preparation for new satellite missions.

7. REFERENCES

[1] J. L. Gomez-Dans, and P. E. Lewis, "Efficient emulation of radiative transfer models using Gaussian processes." (*In prep)*

[2] J. Verrelst, E. Romijn, and L. Kooistra, "Mapping Vegetation Density in a Heterogeneous River Floodplain Ecosystem Using Pointable CHRIS/PROBA Data," *Remote Sensing*, 4, (9), 2866–2889. 2012.

[3] J. P. R. Caicedo, J. Verrelst, J. Munoz-Mari, J. Moreno, and G. Camps-Valls, "Toward a Semiautomatic Machine Learning Retrieval of Biophysical Parameters," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, 7(4), 1249–1259. 2014.

[4] C. Van der Tol, J.A. Berry, P.K.E. Campbell, and U. Rascher, "Models of fluorescence and photosynthesis for interpreting measurements of solar induced chlorophyll fluorescence," *Journal of Geophysical Research*. 2014.

[5] J. Verrelst, J.P. Rivera, C. Van Der Tol, F. Magnani, G. Mohammed, and J. Moreno, "Global sensitivity analysis of SCOPE v1.53 model: what drives canopy-leaving sun-induced fluorescence*?" Rem Sens of Env*.. Submitted. 2015.