EMULATION OF SUN-INDUCED FLUORESCENCE FROM RADIANCE DATA RECORDED BY THE HYPLANT AIRBORNE IMAGING SPECTROMETER

M. Morata1, B. Siegmann2, P. Morcillo-Pallarés1,3, J.P. Rivera-Caicedo4, J. Verrelst1
1Image Processing Laboratory, Univ. València, Spain, 2Institute of Bio- and Geosciences, Plant Sciences (IBG-2), Forschungszentrum Julich GmbH, Germany, 3Institute ITACA, Univ. Politècnica de València, Spain, 4CONACYT-UAN, Tepic, Nayarit, Mexico

ABSTRACT
The retrieval of sun-induced fluorescence (SIF) from hyperspectral radiance data grew to maturity with research activities around the FLuorescence EXplorer satellite mission FLEX, yet the used methods are computationally expensive. To bypass this computational load, this work aims to approximate the currently used spectral fitting method (SFM) by means of statistical learning, i.e. emulation. To do so, we analyzed the possibility of approximating the SFM with an emulator without losing the precision of the original method. In order to enable emulating the hyperspectral radiance spectrum into the multispectral SIF output signal, a double principal component analysis (PCA) dimensionality reduction, i.e. in both input and output, has been implemented. We systematically tested different machine learning regression algorithms, number of principal components (PCs), number of training samples and quality of training samples. The best performing emulator was then applied to a HyPlant flight line containing at sensor radiance information, and the results were compared to the SFM SIF map of the same flight line, which was used as reference. The emulated SIF map was generated quasi-instantaneously and a good agreement with the SFM map could be achieved: $R^2$ of 0.88 and NRMSE of 3.77%. Finally, to evaluate the robustness and transferability, the emulator was applied to other HyPlant flight lines, leading to $R^2$ of 0.97 and NRMSE of 2.56%. Generated emulated SIF maps proved to be consistent while processing time was in the order of 3 minutes. In comparison, by using SFM the SIF processing took approximately 78 minutes.

Index Terms— Emulation, Sun-Induced Fluorescence, SIF, Spectral Fitting Method (SFM), Principal Component Analysis

I. INTRODUCTION
By retrieving sun-induced fluorescence (SIF), essential indicators related to the plant’s actual health status can be derived [1]. Multiple SIF retrieval methods have been developed, most of them focusing on the retrieval in absorption features. More recently, full-spectrum methods such as the Spectral Fitting Method (SFM) allows to obtain a range of values of the spectral signature in the range of the $O_2A$ and $O_2B$ absorption features coming from hyperspectral radiance data [2]. However, the SFM applied to images is time consuming, given the number of per-pixel iterations and a high number of pixels involved. In order to bypass this computational burden, recently a computationally effective technique has been proposed by making use of machine learning regression algorithms (MLRAs) [3]. The core idea is approximating the original input-output relationships by a surrogate statistical learning model which is less computational intensive, also referred to as a meta-model or emulator. Once trained, it is expected that the emulator allows generating SIF outputs quasi-instantaneously while high precision can be maintained. Accordingly, with the purpose of providing a fast alternative to the SFM-based SIF retrieval method, the objective of this work was emulating multispectral SIF outputs in the range of the $O_2A$ absorption feature directly from hyperspectral radiance data. In order to fully exploit this proof of concept, this study systematically evaluated multiple emulation strategies, i.e. analyzing the role of MLRAs, quantity and quality of training data and dimensionality reduction. Eventually, the best performing emulator is applied to radiance images that were recorded from the airborne imaging spectrometer HyPlant for fast generation of SIF maps.

II. MATERIALS AND METHODS
II-A. Spectral Fitting Method (SFM)
Most SIF retrieval methods allow to determine a unique scalar value for each absorption band, which will coincide with the maximum absorption wavelength ($\sim 687$ nm and $\sim 760$ nm). This is not the best solution, since the peak values of SIF are slightly shifted to both peak absorption bands (685 nm and 740 nm), so the maximum SIF value is unknown. Instead, the SFM for HyPlant consists of two main components: (i) atmospheric radiation transfer modeling; and (ii) decoupling reflectance and SIF based on the Spectral Fitting technique [2]. In SFM, the SIF and reflectance are estimated using a numerical optimization technique by comparing the at-sensor radiance computed and HyPlant observations until the best match is found between the two spectra. As an example, Fig. (1) shows how the method is applied in each of the spectral bands. The blue line is the F reference. It can be observed how the adjustment is much more precise in $O_2A$ than in $O_2B$. This is because the interpolation chosen for the B band is not adjusted correctly. SFM applied to HyPlant radiance data using a contemporary desktop PC needs approximately 1.56 seconds to retrieve SIF for 1000 samples.

II-B. Emulation of hyperspectral data
Emulation is a statistical learning technique used to approximate model simulations when the model under investigation is too computationally costly to be run many times [5]. Emulators are based on MLRAs and use a data set made up of input-output pairs for training. In this way, the emulator is able to infer the statistical relationships
in which the original complex model is based and thus imitate the behavior of the original model with a much lower computational cost. These data pairs should ideally cover the maximum multidimensional input space. Once the emulator is built, it is not necessary to perform any additional training of the model.

The challenge of emulating hyperspectral data, however, lies in predicting multiple spectral bands. Only few MLRAs allow generating models that result in multiple outputs that represent the spectral profile. In order to cover the large number of spectral bands, an efficient solution is to take advantage of the Hugens phenomenon, which shows the existence of spectral redundancy [6]. It implies that such data can be converted to a lower-dimensional space through dimensionality reduction (DR) techniques. Accordingly, spectroscopy data can be converted into components, which are only a fraction of the original amount of bands [7]. The methodology for emulator analysis follows our previous work in [7]. Until now DR methods were only applied to reconstruct hyperspectral output data. Here, instead it is proposed to apply DR methods both to input and output. As such, not only the processing is further simplified and accelerated but also it allows to convert hyperspectral input data into hyperspectral output data.

II-C. Used data and experimental setup

The dataset used for the emulation were recorded by HyPlant sensor on board an airplane acquiring flight lines over an agricultural area. HyPlant consists of two sensor modules, DUAL and FLUO [8]. The FLUO module records radiance data from 669.5-781.9 nm with a spectral sampling interval (SSI) of 0.11 nm resulting in 1024 bands. From this data, SIF maps were retrieved by applying the SFM to the O$_2$A (751-777 nm) region with a resolution of 1 nm resulting in 27 bands.

The emulator training data has been collected from a subset of a flight line (700 x 1500 pixels in size) acquired on 26 June 2018 at 13:46 (local time) from 1800 m above ground level. From this subset, a total of 1000 samples from both radiance (input) and SIF (output) were randomly selected with an uniform distribution throughout the image (Fig. 1). Based on this data selection, MLRAs, database size, DR method and number of components were analyzed. Based on earlier experiences, five powerful MLRAs were selected. They were firstly analyzed for the default training database (i.e., 1000-samples) on their predictive power to approximate the O$_2$A SIF spectrum. In order to determine which of the MLRAs obtains most accurate emulation results without compromising execution time, the default training parameters (1000 samples, 20 PCs for input and 5 PCs for output) were established and an emulator was trained for each of the MLRA. The PC used for processing has the following characteristics: Windows 10 Enterprise v.19041.572 64-bits OS, i7-9700K CPI 3.60 GHz, 32 GB RAM. All processing and evaluation steps were conducted within the ARTMO (Automated Radiative Transfer Models Operator) framework.

ARTMO is a scientific software package developed in Matlab that provides tools and toolboxes for running a suite of leaf, canopy and atmosphere RTMs and for post-processing applications such as the emulator toolbox [3]. The toolboxes are freely downloadable at www.artmotoolbox.com.

III. RESULTS

III-A. Analysis optimizing emulation

First the performance of tested MLRAs was evaluated. From the collected samples 70% of the data were used for training while the remaining 30% was used to validate the trained emulators. Starting settings were: 1000 samples, 20 PCs for input and 5 PCs for output. Goodness-of-fit results are provided in Table I. The MLRA that led to the most accurate emulator was the KRR with a NRMSE lower than 6.1%. In addition, the training time of the KRR is 0.57 s, substantially lower than the other models. The spectral NRMSE performance evaluation results for the five MLRAs are shown in Fig. 2. It is shown that the RMSE is stable for the investigated spectral range and does not depend on the wavelength. Given these results, it was decided to continue the study using the KRR model due to its excellent performance in both precision and processing speed (0.08s for 1000 samples).

Table I: Statistics obtained from the performance of the models used. RMSE is in ($x100mWm^{-2}sr^{-1}nm^{-1}$).

| MLRA                                           | RMSE  | NRMSE (%) | Time Train (s) |
|------------------------------------------------|-------|-----------|----------------|
| Kernel ridge Regression                        | 29.72 | 6.09      | 0.57           |
| Gaussian Processes Regression - Matlab         | 30.35 | 6.71      | 10.55          |
| VH. Gaussian Processes Regression              | 31.44 | 6.95      | 80.14          |
| Gaussian Processes Regression                  | 31.47 | 6.96      | 23.80          |
| Multiooutput support vector regression         | 32.03 | 7.08      | 12.33          |

Second, the objective is to obtain a SIF spectrum as output, to which a PCA has been applied. The variation of the NRMSE as a function of the number of PCs was analyzed to determine the optimal number of components that represent the highest amount of information from the original spectra. As expected, by increasing the number of components the development of a more accurate emulator was possible. The NRMSE decreased to a stabilized low level when 20 or more PCs or more were used. Reducing the SIF output which consisted of only 27 bands that are highly correlated, the NRMSE hardly varied depending on the number of components that we used as output, keeping the NRMSE values between 5.622% and 5.626%.

Third, the role of the training database size was analyzed. Proceeding in the same way as before, setting 20 PCs as input and 5 PCs as output and the response of the model is

Fig. 1: Visualization of SFM for O$_2$A and O$_2$B (left)[4]. RGB composite (700.1/754.4/674.4 nm) of HyPlant subset from L12 used to train the models (right).
studied by varying the number of training samples (Fig. 4). Varying the number of samples had the largest effect on the performance of the model. The NRMSE decreased from 7 to approximately 4% and stabilized at this low level when 3000 or more training samples were used. Notably, by adding training data, more processing time will be required, so reaching towards a trade-off between the number of samples and the required training time. A sample size of 3000 was chosen to train the final model.

III-B. Application of the emulator to a HyPlant flight line

Subsequently, the emulation performance of the final model was evaluated, not only with the training subset, but with the entire radiance image of the recorded HyPlant flight lines. The emulator allows generating the $O_2A$ SIF product with all the bands involved from 751-777 nm with a spectral resolution of 1 nm per band. The emulator was applied to the entire flight line and the validation statistics of the results were obtained by comparing them to the reference SIF product. To evaluate the performance of the emulator, the emulated SIF map at 751 nm was compared to the corresponding SIF map retrieved with the SFM (Fig. 5). Although the SIF community is probably more familiar to see the map at 760 nm where the oxygen absorption is at minimum, we chose to show at 751 nm as this band is closest to the SIF maximum located at 740 nm, and thus yielded most pronounced values.

![Fig. 2: Spectral NRMSE (in %) results for the regression algorithms performance assessment as function of the 5 best regression algorithms, using (1000 samples, 20 PCA input and 5 PCA output).](image)

![Fig. 3: NRMSE (in %) (blue axis) and process time (orange axis) results for the KRR emulator performance assessment varying number of PCs in PCA input conversion (1000 samples, 5 PCs output) (a), and PCA output conversion (1000 samples, 20 PCs input) (b).](image)

![Fig. 4: NRMSE (in %) (blue axis) and process time (orange axis) results for the KRR emulator performance assessment varying number of samples (20 PCA input, 5 PCA output).](image)

![Fig. 5: Scatter plot of the SIF map retrieved at 751 nm using the SFM and the corresponding map emulated with the developed KRR model. On the x-axis the values obtained using SFM and on the y-axis using the emulator. The dashed line represents the 1:1-line. Units are in (x100 mWm$^{-2}$sr$^{-1}$nm$^{-1}$).](image)

![Fig. 6: Comparison between the SIF $O_2A$ image at 751 nm obtained using (a) SFM and (b) Emulator. (c) Absolute error between the two images.](image)
III-C. Application and assessment to another flight line

Finally, the emulator has been applied to HyPlant FLUO image data of a different flight line recorded timely close to the investigated one in order to analyze the robustness and spatial transferability of the emulator. Fig. 7 shows the comparison of SIF map at 751 nm derived with the SFM and emulated with the KRR model.

![Image of SIF maps comparison](image)

Fig. 7: Validation of the image obtained by the emulator in comparison with the image obtained by the SFM for the 751 nm band in the full L14 image. On the x-axis the values obtained using SFM and on the y-axis using the emulator. The dashed line represents the 1:1-line. Units are in \((100 \text{ mWm}^{-2}\text{sr}^{-1}\text{nm}^{-1})\)

In this figure some values have been obtained that fit the 1:1-line quite well, except for a small underestimation in the regions with a SIF greater than 10 \(\text{mWm}^{-2}\text{sr}^{-1}\text{nm}^{-1}\). This small underestimation is due to the fact that the model has been trained with the pixels of the L1 flight line subset which contained SIF values between -6 and 6 \(\text{mWm}^{-2}\text{sr}^{-1}\text{nm}^{-1}\). Therefore, the model has not been trained for such high values, giving a slight underestimation for these values. We can also observe a group of pixels where the SFM provided values of around 5 and the emulator estimated values of around 10 \(\text{mWm}^{-2}\text{sr}^{-1}\text{nm}^{-1}\). When inspecting the origin of this data, it appeared to be sludge from the sewage plant, i.e. land cover types that were not included in the training data. Lastly, regarding the statistics against the reference SIF product, a \(R^2 = 0.97\) and a NRMSE of 2.56 have been obtained. These results are satisfactory given that this image is different than the image where pixels were taken to train the model.

IV. DISCUSSION AND CONCLUSIONS

We evaluated the potential of emulators to generate SIF maps from at sensor radiance data of HyPlant. To enable emulation of hyperspectral input data (radiance) towards multispectral output data (SIF), dimensionality reduction was applied to both input and output. The role of various factors affecting the performance of the emulator have been systematically evaluated. Highest accuracies were obtained with a training dataset of 3000 samples, 20 PCs input and 5 PCs output. Several advantages have been identified as opposed to using the original SIF retrieval method, i.e. SFM: (1) above all the reduction of processing time. It implies that emulation can process huge amounts of data in less time. (2) Compared to other SIF retrieval methods, the emulator allows to obtain directly the multispectral SIF output in the \(O_2\) absorption band. (3) In fact, the emulator cannot only approximate SFM-retrieved SIF outputs but can mimic any SIF retrieval method. (4) The emulator is built from image data rather than simulated data, so the trained model is a realistic scenario where noise from images is implicit. Summarizing, the emulator offers an attractive alternative to bypass the computational load of the SFM method for quick generation of SIF maps. The validation of emulated data against reference data demonstrated the potential of the technique. However, it must also be remarked that no perfect approximation was reached throughout the complete image scene. Advances in machine learning and training strategies are expected to further improve the quality of emulation.

V. REFERENCES

[1] G.H. Mohammed, R. Colombo, E.M. Middleton, U. Rascher, C. van der Tol, L. Nedbal, Y. Goulas, O. Pérez-Priego, A. Damm, M. Meroni, J. Joiner, S. Cogliati, W. Verhoef, Z. Malenkovský, J.-P. Gastellu-Etchegorry, J.R. Miller, L. Guanter, J. Moreno, I. Moya, J.A. Berry, C. Frankenberg, and P.J. Zarco-Tejada. “Remote sensing of solar-induced chlorophyll fluorescence (sif) in vegetation: 50 years of progress,” Remote Sensing of Environment, vol. 231, 2019.

[2] B. Siegmann, L. Alonso, M. Celesti, S. Cogliati, R. Colombo, A. Damm, S. Douglas, L. Guanter, J. Hanus, K. Kataja, T. Kraska, M. Matveeva, J. Moreno, O. Muller, M. Piki, F. Pinto, J. Quirós, P. Rademskie, and F. Rodríguez-Moreno, “The high-performance airborne imaging spectrometer hyplant—from raw images to top-of-canopy reflectance and fluorescence products: Introduction of an automatized processing chain,” Remote Sensing, vol. 11, pp. 2760, 11 2019.

[3] J.P. Rivera, J. Verrelst, J. Gómez-Dans, J. Muñoz Mari, J. Moreno, and G. Camps-Valls, “An emulator toolbox to approximate radiative transfer models with statistical learning,” Remote Sensing, vol. 7, no. 7, pp. 9347, 2015.

[4] M.P. Cendero-Mateo, S. Wienieke, A. Damm, L. Alonso, F. Pinto, J. Moreno, L. Guanter, M. Celesti, M. Rossini, N. Sabater, S. Cogliati, T. Julitta, U. Rascher, Y. Goulas, H. Aasen, J. Pacheco-Labrador, and A. Mac Arthur, “Sun-induced chlorophyll fluorescence iii: Benchmarking retrieval methods and sensor characteristics for proximal sensing,” Remote Sensing, vol. 11, no. 8, 2019.

[5] A. O’Hagan, “Bayesian analysis of computer code outputs: A tutorial,” Reliability Engineering and System Safety, vol. 91, no. 10-11, pp. 1290–1300, 2006.

[6] G.F. Hughes, “On the mean accuracy of statistical pattern recognizers,” IEEE Trans. Inf. Theory, vol. 14, no. 1, pp. 55–63, 1968.

[7] J. Verrelst, J.P. Rivera-Caicedo, J. Muñoz Mari, G. Camps-Valls, and J. Moreno, “SCOPE-Based Emulators for Fast Generation of Synthetic Canopy Reflectance and Sun-Induced Fluorescence Spectra,” Remote Sensing, vol. 9, no. 9, 2017.

[8] U. Rascher, L. Alonso, A. Burkart, C. Cilia, S. Cogliati, R. Colombo, A. Damm, M. Drusch, L. Guanter, J. Hanus, et al., “Sun-induced fluorescence—a new probe of photosynthesis: First maps from the imaging spectrometer hyplant,” Global change biology, vol. 21, no. 12, pp. 4673–4684, 2015.