RANKING DRIVERS OF GLOBAL CARBON AND ENERGY FLUXES OVER LAND

Gustau Camps-Valls[1], *Martin Jung*[2], *Kazuhito Ichii*[3], *Dario Papale*[4], *Gianluca Tramontana*[4], *Paul Bodesheim*[5], *Christopher Schwalm*[6], *Jakob Zscheischler*[2], *Miguel Mahecha*[2] and *Markus Reichstein*[2]

[1]Universitat de València, València, Spain; [2] Max Planck Institute for Biogeochemistry, Germany
[3] Japan Agency for Marine-Earth Science and Technology, Japan; [4] University of Tuscia, Italy
[5] Friedrich Schiller University of Jena, Germany; [6] University of Minnesota, USA

ABSTRACT

The accurate estimation of carbon and heat fluxes at global scale is paramount for future policy decisions in the context of global climate change. This paper analyzes the relative relevance of potential remote sensing and meteorological drivers of global carbon and energy fluxes over land. The study is done in an indirect way via upscaling both Gross Primary Production (GPP) and latent energy (LE) using Gaussian Process regression (GPR). In summary, GPR is successfully compared to multivariate linear regression (RMSE gain of +4.17% in GPP and +7.63% in LE) and kernel ridge regression (+2.91% in GPP and +3.07% in LE). The best GP models are then studied in terms of explanatory power based on the analysis of the lengthscales of the anisotropic covariance function, sensitivity maps of the predictive mean, and the robustness to distortions in the input variables. It is concluded that GPP is predominantly mediated by several vegetation indices and land surface temperature (LST), while LE is mostly driven by LST, global radiation and vegetation indices.

Index Terms— Gaussian process, regression, feature ranking, GPP, carbon, energy, global monitoring

1. INTRODUCTION

Estimating biosphere-atmosphere fluxes at continental to global scale based on FLUXNET[1] along with remote sensing and meteorological data has become an emerging and very promising field of active research. In the last decade, global spatial-temporal fields of FLUXNET-derived carbon and energy fluxes are increasingly used for analyzing variations of the global carbon and energy cycles, and to evaluate global land surface models. Model/process-based and data-driven algorithms are the two main approaches to upscale data acquired from flux towers [1, 2]. In the last few years, nevertheless, data-driven statistical learning algorithms have attained outstanding results in the estimation of climate variables and related bio-geo-physical parameters at local and global scales [3]. These algorithms avoid complicated assumptions and provide flexible nonparametric models that fit the observations using massive heterogeneous data. Current operational vegetation products, like leaf area index (LAI), are typically produced with neural networks, Gross Primary Production (GPP) –as the largest global CO_2 flux driving several ecosystem functions– is estimated using ensembles of random forests and neural networks [1,2,4], biomass has been estimated with stepwise multiple regression [5], support vector regression [6] showed high efficiency in modelling LAI, fCOVER and evapotranspiration [7], and kernel methods in general [8, 9] and Gaussian Processes (GPs) in particular [10] recently provided excellent results in chlorophyll content estimation [11–13].

In this work, we focus on the properties of GPs to tackle the problem of carbon and energy fluxes modeling. Gaussian processes are used here: 1) to *estimate* global flux products derived from upscaling FLUXNET eddy covariance observations; and more importantly, 2) to *assess* the relative relevance of the used explanatory remote sensing and meteorological variables. In particular, we will focus on key carbon and energy fluxes only. We evaluate three different techniques to unveil the knowledge learned by the GP models, rooted on either permutation, sensitivity or automatic relevance determination priors.

The remainder of the paper is organized as follows. Section 2 reviews the theory underlying GP regression (GPR) and the techniques used to infer feature rankings from trained GPR models. Section 3 details the data used in this paper. Section 4 gives the experimental results. Finally, Section 5 concludes the paper and outlines the further work.

2. RANKING DRIVERS WITH GP MODELS

2.1. Gaussian Process Regression (GPR)

Standard regression approximates observations (often referred to as *outputs*) $\{y_n\}_{n=1}^N$ as the sum of some unknown latent function $f(\mathbf{x})$ of the inputs $\{\mathbf{x}_n \in \mathbb{R}^D\}_{n=1}^N$ plus *constant power* Gaussian noise, i.e. $y_n = f(\mathbf{x}_n) + \varepsilon_n$, $\varepsilon_n \sim \mathcal{N}(0, \sigma_n^2)$. Instead of proposing a parametric form for $f(\mathbf{x})$ and learning its parameters in order to fit observed data well, GP regression (GPR) proceeds in a Bayesian, non-parametric way. A zero mean[2] GP prior is placed on the latent function $f(\mathbf{x})$ and a Gaussian prior is used for each latent noise term ε_n, $f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}'))$, where $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ is a covariance function parameterized by $\boldsymbol{\theta}$, and σ_n^2 is a hyperparameter that

---

[1]http://www.fluxdata.org

[2]It is customary to subtract the sample mean to data $\{y_n\}_{n=1}^N$, and then to assume a zero mean model.

specifies the noise power. Essentially, a Gaussian process is a stochastic process whose marginals are distributed as a multivariate Gaussian. In particular, given the priors $\mathcal{GP}$, samples drawn from $f(\mathbf{x})$ at the set of locations $\{\mathbf{x}_n\}_{n=1}^N$ follow a joint multivariate Gaussian with zero mean and covariance matrix $\mathbf{K_{ff}}$ with $[\mathbf{K_{ff}}]_{ij} = k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)$, which is parametrized by a set of hyperparameters $\boldsymbol{\theta}$. GPR is intimately related to standard kriging in geostatistics: $\mathbf{x}$ is an arbitrary feature vector rather than just the geographical coordinates of a sample.

If we consider a test example $\mathbf{x}_*$ with corresponding output value y_*, the $\mathcal{GP}$ defines a joint prior distribution between the observations $\mathbf{y} \equiv \{y_n\}_{n=1}^N$ and y_*. Collecting a training data set in $\mathcal{D} \equiv \{\mathbf{x}_n, y_n | n = 1, \ldots N\}$, it is possible to analytically compute the posterior distribution over the unknown output y_*:

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \mathcal{N}(y_* | \mu_{\mathrm{GP}*}, \sigma_{\mathrm{GP}*}^2)$$
$$\mu_{\mathrm{GP}*} = \mathbf{k}_{\mathbf{f}*}^\top (\mathbf{K_{ff}} + \sigma_n^2 \mathbf{I}_n)^{-1} \mathbf{y} = \mathbf{k}_{\mathbf{f}*}^\top \boldsymbol{\alpha}$$
$$\sigma_{\mathrm{GP}*}^2 = \sigma_n^2 + k_{**} - \mathbf{k}_{\mathbf{f}*}^\top (\mathbf{K_{ff}} + \sigma_n^2 \mathbf{I}_n)^{-1} \mathbf{k}_{\mathbf{f}*},$$

where $\mathbf{k}_{\mathbf{f}*}$ is a vector of similarities between the test point and all training points.

GPs offer some advantages over other regression methods. Since they yield a full posterior predictive distribution over y_*, it is possible to obtain not only mean predictions for test data, $\mu_{\mathrm{GP}*}$, but also the so-called "error-bars", $\sigma_{\mathrm{GP}*}^2$, assessing the uncertainty of the mean prediction. The whole procedure only depends on a very small set of hyperparameters $\boldsymbol{\theta}$, which combats overfitting efficiently. Also, inference of the hyper-parameters and the weights $\boldsymbol{\alpha}$ can be performed using continuous optimization of the evidence so there is no need to resort to cross-validation procedures. Note, however, that the bottleneck of the algorithm is the definition of the covariance (kernel or Gram) function $k_{\boldsymbol{\theta}}$: this function should capture the similarity between data instances. A standard, widely used covariance function is the *isotropic* squared exponential (SE), $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / (2\sigma^2))$.

2.2. GPR for feature ranking

In this work, we report results following three different ways to rank the input drivers (features, explanatory variables) used by the GPR models. While rooted in different principles, all of them evaluate how robust is the fitted model to distortions or information content of the features.

2.2.1. Permutation analysis

The first approach is simple and general for any regression algorithm, and basically consists of a greedy algorithm in which the impact of the inputs on the prediction error is evaluated in the context or absence of the other predictors. Essentially, for each feature j, the algorithm permutes its values for all training samples, and evaluates the prediction RMSE. The process is repeated for a number of permutations, $p = 1, \ldots, n_p$, and results are averaged. The relevance of feature j is $r_j = 1/(\frac{1}{n_p} \sum_p \mathrm{RMSE}(p, j))$. In [12, 13], we illustrated the usefulness of this procedure for the identification of the most relevant spectral channels for the retrieval of vegetation parame-

ters (chlorophyll content, LAI and FVC) from hyperspectral data.

2.2.2. Anisotropic Gaussian kernel

An interesting possibility to evaluate the relative relevance of the inputs identified by a trained GP is to resort to an *anisotropic* Gaussian covariance function, as an alternative generalization of the isotropic SE prior. The so-called automatic relevance determination (ARD) prior takes the form

$$k(\mathbf{x}_p, \mathbf{x}_q) = \nu^2 \exp\left(-\sum_{j=1}^D \frac{(\mathbf{x}_p^j - \mathbf{x}_q^j)^2}{2\sigma_j^2}\right),$$

where $\mathbf{x}_p^j$ is the feature j of input vector $\mathbf{x}_p$, σ_j contains a lengthscale per dimension, and ν is a positive scale factor. ARD effectively prunes input feature j for large values of σ_j. Therefore, the relevance of feature j is thus directly given by the inverse of the infered lengthscale, σ_j. In our previous works [12, 14], ARD was used to assess the relevance of the D spectral channels in vegetation chlorophyll content estimation.

2.2.3. Sensitivity analysis over the GP

In this work we exploit the sensitivity analysis over the predictive mean of the GP model, $\mu_{\mathrm{GP}*}$. Sensitivity of feature j is here defined as $s_j = \int (\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}_j})^2 p(\mathbf{x}) d\mathbf{x}$, where $p(\mathbf{x})$ is the probability density function over the input $\mathbf{x}$, and $\phi(\mathbf{x})$ represents the predictive mean, $\mu_{\mathrm{GP}*}$. Intuitively, the objective of the sensitivity map is to measure the changes of the derivative of the function $\phi(\mathbf{x})$ in the jth direction. In order to avoid the possibility of cancellation of the terms due to its signs, the derivatives are squared. Therefore, the resulting sensitivity map will be positive $s_j \geq 0$ for all features. The empirical estimate of the sensitivity for the jth feature can be written as $s_j = \frac{1}{N} \sum_{n=1}^N (\frac{\partial \phi(\mathbf{x}_n)}{\partial \mathbf{x}_j})^2$. It is easy to show that the resulting empirical estimate of the GP mean sensitivity map for the ARD prior is:

$$s_j = \frac{\nu^2}{N} \sum_{q=1}^N \left(\sum_{p=1}^N \alpha_p (\mathbf{x}_p^j - \mathbf{x}_q^j) k(\mathbf{x}_p, \mathbf{x}_q)/\sigma_j^2\right)^2. \quad (1)$$

Note that s_j is computed in closed-form using only training points and the inferred weight vector $\boldsymbol{\alpha}$.

3. DATA COLLECTION AND PREPROCESSING

In this work, we exploit two complementary sets of products with enhanced spatial and temporal resolution in comparison to existing products [2]: a 5 min spatially and 8 day temporally resolved product driven solely by remote sensing based variables, and a daily and vegetation type specific product at 0.5^o driven by meteo and mean seasonal cycle remote sensing based variables.

We used the global La Thuile FLUXNET synthesis data set[1] which is composed of half-hourly FLUXNET eddy covariance measurements processed using standardized proce-

dures of gap-filling and quality control [15, 16]. The fluxes were subsequently aggregated into 8-daily means to conform to the temporal resolution of MODIS products. We excluded data where more than 20% of the data of the monthly mean was based on gap filling with low confidence [17], and applied additional semi-automated screening for bad quality data. Estimates of GPP were based on the flux partitioning method in [17].

We collected and processed MODIS land data at the locations of the flux towers: MOD11A2 Land Surface Temperature (LST) [18], MOD13A2 Vegetation Index (VI) [19], MOD15A2 Leaf Area Index and FPAR (LAI/FPAR) [20], and MCD43A2 and MCD43B4 BRDF-corrected surface reflectances [21]. Two vegetation index, LSWI and NDWI, were created further using BRDF-corrected surface reflectance data. Additional variables were created as the product of vegetation indices and LST, or global radiation (Rg). We used the data from 3 by 3 km regions centered on the flux towers and applied QA/QC to screen bad quality data. Details are given in Table 1.

Table 1: Remote sensing products used to derive the drivers.

Scale	Parameter	Product	Spatial Res.	Temp. Res.	Ref.
Site	LST	MOD11A2	1km	8 day	[18]
	VI	MOD13Q1	250m	16 day	[19]
	LAI/FPAR	MOD15A2	1km	8 day	[20]
	Reflectance	MCD43A2	500m	8 day	[21]
	Reflectance	MCD43A4	500m	8 day	[21]
Global	LST	MOD11A2	1km	8 day	[18]
	VI	MOD13A2	1km	16 day	[19]
	LAI/FPAR	MOD15A2	1km	8 day	[20]
	Reflectance	MCD43B4	1km	8 day	[21]
	Solar Rad	JAXA /JASMES	5km	1 day	[22]

A suite of additional explanatory variables was derived by computing the mean seasonal cycle (MSC) and metrics thereof like minimum, maximum, and mean for all 'raw' variables, which yielded a total set of 216 potentially explanatory variables. The feature selection algorithm presented in [23] was used to identify suitable feature subsets for carbon and energy fluxes respectively. We here focus on the best eight features, $D = 8$: For GPP we used Normalized Difference Water Index (NDWI), Land Surface Temperature during day (LST-Day) and night (LST-Night), the maximum of the LST mean seasonal cycle (MSC-Day), NDVI times the global radiation (Rg) map, Enhanced Vegetation Index (EVI), medium-infrared (MIR) region, and the leaf area index (LAI). For LE, the input features were EVI×LST-Day, the MSC of the Fraction of Photosynthetically Active Radiation (fAPAR) × Rg, EVI×LST, LST-Day and LST-Night, Rg, along with the potential radiation (Rg-Pot) and its minimum MSC (Rg-MSC-Min). All these features are then used to fit machine learning models of relevant carbon and energy fluxes, such as the Gross Primary Production (GPP) and latent energy (LE).

4. EXPERIMENTAL RESULTS

This section shows the numerical results obtained on the GPP and LE modeling, and evaluates the three techniques for ranking drivers of global carbon and energy fluxes over land.

4.1. Experimental setup

In this work we compare numerically five regression algorithms: a multivariate linear regression (LR), multilayer perceptron neural network (MLP) [4], support vector regression (SVR) [6], the kernel ridge regression (KRR) [8], and the standard Gaussian process regression (GPR) [10]. We split the available data ($N = 19797$ points) into 10 independent folds, and show the 10-fold cross-validation results. A one-hidden layer MLP was trained with the Levenberg-Marquardt algorithm to minimize the squared loss. Both SVR and KRR used the SE kernel and employed 2/3 of the training data in each fold for training and 1/3 for validation. GPR used the ARD prior covariance and hyperparameters θ were selected by maximizing the marginal likelihood of the observations. The ARD kernel endorses some more flexibility in the modeling and allows to study the inferred lengthscale for feature ranking. The MATLAB simpleR toolbox[3] was used to fit the models.

4.2. Numerical comparison

Table 2 shows the obtained cross-validation results in terms of bias (mean error, ME), accuracy (RMSE, MAE), and goodness-of-fit (Pearson's correlation coefficient ρ). It can be noticed that GP models perform better that the rest in RMSE, MAE and ρ, while they provide slightly higher bias for GPP. In summary, GPR is successfully compared to multivariate LR in RMSE terms (gain of +4.17% in GPP and +7.63% in LE), as well as nonlinear models (maximum gain of +2.91% in GPP and +3.07% in LE). GPR reveals similar accuracy results to KRR, yet less biased for the latent energy flux. In addition, as we will study in the next section, GPR allows to derive feature ranking under solid Bayesian foundations.

Table 2: Numerical results for GPP and LE and all methods.

GPP	ME	RMSE	MAE	ρ
LR	-0.01	1.83	1.30	0.78
MLP [4]	+0.04	1.92	1.39	0.73
SVR [6]	+0.01	1.80	1.23	0.78
KRR [8]	+0.00	1.81	1.24	0.78
GPR [10]	+0.03	1.76	1.16	0.80
LE	ME	RMSE	MAE	ρ
LR	-0.00	1.70	1.26	0.79
MLP [4]	+0.14	1.68	1.25	0.80
SVR [6]	+0.01	1.55	1.11	0.83
KRR [8]	+0.12	1.53	1.10	0.84
GPR [10]	-0.01	1.52	1.06	0.84

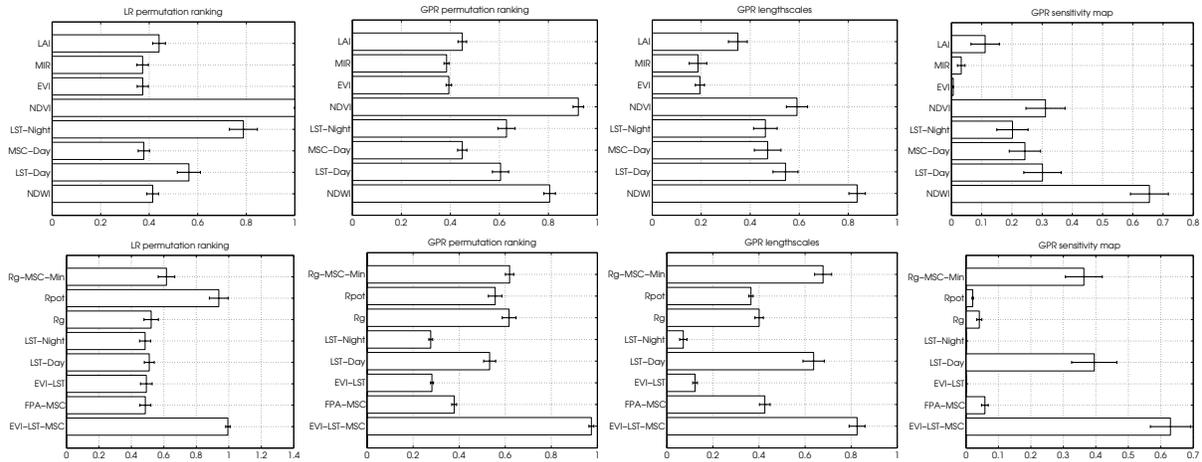[3]http://www.uv.es/gcamps/code/simpleR.html

Fig. 1: Normalized feature rankings (mean±std.dev.) using LR and GPR models for GPP (top) and LE (bottom).

4.3. Ranking features

Figure 1 shows four feature rankings: the permutation analysis of both LR and GPR using $n_p = 100$, sensitivity analysis of GP predictive mean function, and the lengthscales for the best GP models. Several conclusions can be derived: First, LR and GPR lead to somewhat similar results in the permutation analysis: both assign high relevance to the NDVI for modeling GPP and EVI-LST-MSC for modeling LE. Actually, vegetation indices and LST were found to be informative for GPP and LE modelling in several previous studies. Second, unlike LR, GPR assigns a very high relevance to NDWI for GPP, as well as higher relevance to LST-Day and lower relevance to Rg for LE estimation. These might be due to the expected non-linear relationship between moisture availability and GPP, and between LST and LE, respectively. Finally, note that the permutation ranking and the analysis of lengthscales report similar results for GPs, while the sensitivity analysis typically identifies very few features as relevant: NDWI for GPP modeling, while LST-Day and EVI-LST-MSC for LE modeling.

5. CONCLUSIONS

This paper summarized some of our on-going activities to provide high quality global maps for carbon and energy fluxes. In particular, we introduced Gaussian processes as a flexible nonparametric algorithm for parameter retrieval and upscaling. GPs are here exploited to provide high accuracy estimates and feature rankings in different levels of sophistication. It was shown that GP models outperformed linear regression and kernel ridge regression for GPP and LE modeling. In addition, three techniques for model analysis were studied to reveal some knowledge about trained GP models. The most relevant drivers according to GP were several vegetation indices and land surface temperature (LST) for GPP modeling, while LE is mostly driven by LST, global radiation and vegetation indices. Further work will consider bigger sets of features, and modern techniques for visualization of the feature maps.

6. REFERENCES

[1] C. Beer, *et al.*, *Science* **329** (2010).

[2] M. Jung, *et al.*, *Journal of Geophysical Research: Biogeosciences* **116**, 1 (2011).

[3] G. Camps-Valls, D. Tuia, L. Gómez-Chova, J. Malo, eds., *Remote Sensing Image Processing* (Morgan & Claypool, 2011).

[4] S. Haykin, *Neural Networks – A Comprehensive Foundation* (Prentice Hall, 1999), second edn.

[5] L. R. Sarker, J. E. Nichol, *Rem. Sens. Env.* **115**, 968 (2011).

[6] A. J. Smola, B. Schölkopf, *Statistics and Computing* **14**, 199 (2004).

[7] S. Durbha, R. King, N. Younan, *Rem. Sens. Env.* **107**, 348 (2007).

[8] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis* (Cambridge University Press, 2004).

[9] G. Camps-Valls, L. Bruzzone, eds., *Kernel methods for Remote Sensing Data Analysis* (Wiley & Sons, UK, 2009).

[10] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, New York, 2006).

[11] R. Furfaro, R. D. Morris, A. Kottas, M. Taddy, B. D. Ganapol, *AGU Fall Meeting Abstracts* pp. A261+ (2006).

[12] J. Verrelst, *et al.*, *Rem. Sens. Env.* **118**, 127 (2012).

[13] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, J. Moreno, *IEEE Trans. Geosc. Rem. Sens.* **50**, 1832 (2012).

[14] J. Verrelst, L. Alonso, J. P. Rivera, J. Moreno, G. Camps-Valls, *IEEE JSTARS* **6**, 867 (2013).

[15] A. M. Moffat, *et al.*, *Agricultural and Forest Meteorology* **147**, 209 (2007).

[16] D. Papale, *et al.*, *Biogeosciences* **3**, 571 (2006).

[17] M. Reichstein, *et al.*, *Global Change Biology* **11**, 1424 (2005).

[18] Z. Wan, Y. Zhang, Q. Zhang, Z. liang Li, *Rem. Sens. Env.* **83**, 163 (2002).

[19] A. Huete, *et al.*, *Rem. Sens. Env.* **83**, 195 (2002).

[20] R. Myneni, *et al.*, *Rem. Sens. Env.* **83**, 214 (2002).

[21] C. Schaaf, *et al.*, *Rem. Sens. Env.* **83**, 135 (2002).

[22] R. Frouin, H. Murakami, *Jour. Oceanogr.* **63**, 493 (2007).

[23] M. Jung, J. Zscheischler, *Procedia Computer Science* **18**, 2337 (2013). Intnl. Conf. Comp. Sci.