A Survey on Gaussian Processes for Earth Observation Data Analysis

Gustau Camps-Valls, *IEEE Senior member*, Jochem Verrelst, Jordi Muñoz-Marí, Valero Laparra, Fernando Mateo-Jiménez, and José Gómez-Dans

***Abstract*— Gaussian Process (GP) has experienced tremendous success in bio-geophysical parameter retrieval in the last years. It goes without saying that GPs constitute a solid Bayesian framework to formulate many function approximation problems consistently. This paper reviews the main theoretical GP developments in the field. We review new algorithms that respect the signal and noise characteristics, that extract knowledge via automatic relevance kernels to yield feature rankings automatically, that allow applicability of associated uncertainty intervals to transport GP models in space and time, that can be used to uncover causal relations between variables, and that can encode physically-meaningful prior knowledge via radiative transfer model emulation. We will treat the important issue of computational efficiency as well. All these developments are illustrated in the field of geosciences and remote sensing at a local and global scales through a set of illustrative examples. In particular, we treat important problems for land, ocean and atmosphere monitoring: from accurate estimation of oceanic chlorophyll content and pigments, to vegetation properties retrieval from multi- and hyperspectral sensors, as well as the estimation of atmospheric parameters (such as temperature, moisture and ozone) from infrared sounders. We conclude the survey with a discussion on the upcoming challenges and research directions.**

***Index Terms*— Kernel methods, Gaussian Process Regression (GPR), Bio-geophysical parameter estimation.**

I. Introduction

Spatio-temporally explicit, quantitative retrieval methods for Earth surface and atmosphere characteristics are a requirement in a variety of Earth system applications. Optical Earth observing satellites, endowed with a high temporal resolution, enable the retrieval and hence monitoring of climate and bio-geophysical variables [1], [2]. With forthcoming super-spectral Copernicus Sentinel-2 (S2) [3] and Sentinel-3 missions [4], as well as the planned EnMAP [5], HyspIRI [6], PRISMA [7] and ESA's candidate FLEX [8], an unprecedented data stream for land, ocean and atmosphere monitoring will soon become available to a diverse user community. This vast data streams require enhanced processing techniques that are accurate, robust and fast. But, in addition, statistical models should capture plausible physical relations and explain the problem at hand.

Manuscript received July 2015;

GCV, JV, JMM, VL, and FMJ are with the Image Processing Laboratory (IPL). Universitat de València. C/ Catedrático Escardino, Paterna (València) Spain. Web: http://isp.uv.es. E-mail: {gcamps,jverrelst,jordi,lapeva,fmateo}@uv.es.

JGD is with Department of Geography, University College London, UK. E-mail: j.gomez-dans@ucl.ac.uk

This paper has been partially supported by the Spanish Ministry of Economy and Competitiveness under project TIN2012-38102-C03-01.

Over the last few decades a wide diversity of bio-geophysical retrieval methods have been developed, but only a few of them made it into operational processing chains, and many of them are still in its infancy [9]. Essentially, we may find two main approaches to the inverse problem of estimating biophysical parameters from spectra. On the one hand, *parametric physically-based models* constitute a common choice to model the biological processes and climate variables involved in Earth monitoring. These models rely on established physical relations and implement complex combinations of scientific hypotheses. Unfortunately they do not exploit empirical data to constrain the simulation outcomes and thus, despite their solid physical foundation, they are becoming more obscure as more complex processes, parametrizations and priors need to be included. These issues give rise to too rigid solutions and large model discrepancies (see [10] and references therein). Alternatively, the framework of *non-parametric statistical models* is typically only concerned about developing *data-driven models*, paying little attention to the physical rules governing the system. The field has proven successful in many disciplines of Science and Engineering [11] and, in general, nonlinear and nonparametric model instantiations typically lead to more flexible and improved performance over physically-based approximations [12].

"Non-parametric machine learning algorithms are mature enough to undertake the complex problems of biophysical parameter estimation, and Gaussian processes provide a powerful framework to this end."

In the last decade, machine learning has attained outstanding results in the estimation of climate variables and related bio-geophysical parameters at local and global scales [13]. For example, current operational vegetation products, like leaf area index (LAI), are typically produced with neural networks [14], [15], 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 [16], [17], biomass has been estimated with stepwise multiple regression [18], PCA and piecewise linear regression for sun-induced fluorescence (SIF) estimation [19], support vector regression showed high efficiency in modelling LAI, fractional vegetation cover (fCOVER), evapotranspiration [20], [21], relevance vector machines were successful in ocean chlorophyll estimation [22], and recently, Gaussian Processes (GPs) [23] provided excellent results in vegetation properties estimation [24]–[27].

The family of Bayesian non-parametrics[1], and of Gaussian processes in particular [23], have been payed wide attention in the last years in remote sensing data analysis, because they are endorsed with some important properties, which are of relevance to common problems in our field. First, GPs can not only provide good accuracy estimations but also error bars (i.e. uncertainties) for the predictions. Also, and very importantly, they can accommodate very easily different data sources (multimodal data, multiple sensors, multitemporal acquisitions, etc.), and they can be designed to deal with different noise sources. The use of GPs in problems involving large data has been a traditional problem, but recently advanced sparse, variational and distributed computing techniques allow training models in almost linear cost. We will study modern approaches to tackle all these issues in the present work.

Beyond these interesting features of GPs, we should stress that statistical inference methods should not only be able to fit data well, i.e. focus only on data *exploitation*, but also learn something about the physical rules governing the problem, i.e. *data exploration*. Therefore, these (too) flexible models should be constrained to provide with physically plausible predictions. This is why, in recent years, the combination of machine learning and physical models seem to be a very promising direction to take, either via data assimilation, hybrid approaches or emulation of radiative transfer models. We will review some of these approaches in this paper too. In this respect, GPs can be used to learn about the relevance of the features in the problem, as (1) they can adapt to anisotropic data distributions, (2) the derivatives of the predictive mean and variance can be computed in closed-form, and (3) they are ideal to be used in empirical (not interventional) causal inference. On top of that, a remarkable fact is that GPs have been the first choice in the process of emulating radiative transfer models to endorse these statistical models with physically-meaningful constraints [28]. This survey reviews all these very exciting issues as well.

"If you want better Physics, take machine learning out; if you want to approximate reality better, put physics in machine learning models."

The remainder of the paper is organized in two main parts: the first three sections present the state of the art in GP under quantitative terms, while the latter sections are more focused on the use of GP regression models to learn about the problem in quantitative terms. Section II reviews the main notation and theory of GP. Section III presents some of the most recent advances of GP models applied to remote sensing data processing. Section IV is concerned on some recent advances in the field of GPs to cope with large scale datasets so that GP models can be effectively used in geosciences. This closes the first part of the paper. Section V pays attention to the techniques to analyze the relative relevance of input features, and Section VI focuses on the use of GP models as efficient emulators of radiative transfer models. We conclude in Section VII the survey with a discussion about the upcoming challenges and research directions.

[1]Excellent online lectures are available at:
http://videolectures.net/mlss09uk_teh_nbm/
http://videolectures.net/mlss09uk_orbanz_fnbm/

II. Gaussian Process Regression

Regression, function approximation and function emulation are old, largely studied problems in statistics and machine learning. The problem boils down to optimize a loss (cost, energy) function over a class of functions. A large class of regression problems in particular are defined as the joint minimization of a loss function accounting for errors of the function $f \in \mathcal{H}$ to be learned, and a regularization term, $\Omega(\|f\|^2_{\mathcal{H}})$, that controls its capacity (excess of flexibility).

A. Gaussian processes: a gentle introduction

Gaussian processes (GPs) are Bayesian state-of-the-art tools for discriminative machine learning, i.e., regression [29], classification [30] and dimensionality reduction [31]. GPs were first proposed in statistics by Tony O'Hagan [32] and they are well-known to the geostatistics community as *kriging*. However, due to their high computational complexity they did not become widely applied tools in machine learning until the early XXI century [23]. GPs can be interpreted as a family of kernel methods with the additional advantage of providing a full conditional statistical description for the predicted variable, which can be primarily used to establish confidence intervals and to set hyper-parameters. In a nutshell, Gaussian processes assume that a Gaussian process prior governs the set of possible latent functions (which are unobserved), and the likelihood (of the latent function) and observations shape this prior to produce posterior probabilistic estimates. Consequently, the joint distribution of training and test data is a multidimensional Gaussian and the predicted distribution is estimated by conditioning on the training data.

"Gaussian processes provide a solid Bayesian framework to deal with uncertainties, noise sources, high dimensional data, and knowledge discovery."

This paper focuses on the recent success of GPs to deal with regression problems in biophysical parameter retrieval and generic model inversion in geosciences. 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 (homoscedastic)* Gaussian noise, i.e.

$$y_n = f(\mathbf{x}_n) + \varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

Instead of proposing a parametric form for $f(\mathbf{x})$ and learning its parameters in order to fit observed data well, GP regression 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 parametrized by $\boldsymbol{\theta}$ and σ^2 is a hyperparameter that specifies the noise power. Essentially, a GP is a stochastic process whose marginals are distributed as a multivariate

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

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 $[\mathbf{K_{ff}}]_{ij} = k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)$.

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

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \mathcal{N}(y_*|\mu_{\text{GP}*}, \sigma^2_{\text{GP}*}), \tag{2}$$

which is a Gaussian with the following mean and variance:

$$\mu_{\text{GP}*} = \mathbf{k}_{\mathbf{f}*}^\top (\mathbf{K_{ff}} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y} \tag{3}$$

$$\sigma^2_{\text{GP}*} = \sigma^2 + k_{**} - \mathbf{k}_{\mathbf{f}*}^\top (\mathbf{K_{ff}} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{k}_{\mathbf{f}*}, \tag{4}$$

where $\mathbf{k}_{\mathbf{f}*} \in \mathbb{R}^{N \times 1}$ contains the kernel similarities of the test point $\mathbf{x}_*$ to all training points in $\mathcal{D}$, $\mathbf{K_{ff}}$ is a $N \times N$ kernel (covariance) matrix whose entries contain the similarities between all training points, $\mathbf{y} = [y_1, \dots, y_N]^\top \in \mathbb{R}^{N \times 1}$, σ^2 is a hyperparameter accounting for the variance of the noise, k_{**} is a scalar with the self-similarity of $\mathbf{x}_*$, and $\mathbf{I}_n$ is the identity matrix of size n. Note that both the predictive mean and the variance can be computed in closed-form, that the predictive variance $\sigma^2_{\text{GP}*}$ do not depend on the outputs/target variable.

which is computable in $\mathcal{O}(n^3)$ time (this cost arises from the inversion of the $n \times n$ matrix $(\mathbf{K_{ff}} + \sigma^2 \mathbf{I})$, see [23]. In addition to the computational cost, GPs require large memory since in naive implementations one has to store the training kernel matrix, which amounts to $\mathcal{O}(n^2)$. Recent improvements on efficiency will be reviewed in §4.

B. On the model selection

The corresponding hyperparameters $\{\boldsymbol{\theta}, \sigma_n\}$ are typically selected by Type-II Maximum Likelihood, using the marginal likelihood (also called evidence) of the observations, which is also analytical (explicitly conditioning on $\boldsymbol{\theta}$ and σ_n):

$$\log p(\mathbf{y}|\boldsymbol{\theta}, \sigma_n) = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K_{ff}} + \sigma_n^2 \mathbf{I}). \tag{5}$$

When the derivatives of (5) are also analytical, which is often the case, conjugated gradient ascend is typically used for optimization. Therefore, the whole procedure for learning a GP model only depends on a very small set of hyper-parameters that combats overfitting efficiently. Finally, inference of the hyper-parameters and the weights for doing predictions, $\boldsymbol{\alpha}$, can be performed using this continuous optimization of the evidence.

C. On the covariance function

The core of any kernel method in general, and of GPs in particular, is the appropriate definition of the covariance (or kernel) function. A standard, widely used covariance function is the squared exponential,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2)),$$

which captures sample similarity well in most of the (unstructured) problems, and only one hyperparameter σ needs to be tuned. Table I summarizes the most common kernel functions in standard applications with kernel methods.

TABLE I
SOME KERNEL FUNCTIONS USED IN THE LITERATURE.

Kernel function	Expression
Linear	$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}' + c$
Polynomial	$k(\mathbf{x}, \mathbf{x}') = (\alpha \mathbf{x}^\top \mathbf{x}' + c)^d$
Gaussian	$k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/(2\sigma^2))$
Exponential	$k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|/(2\sigma^2))$
Rational Quadratic	$k(\mathbf{x}, \mathbf{x}') = 1 - (\|\mathbf{x} - \mathbf{x}'\|^2)/(\|\mathbf{x} - \mathbf{x}'\|^2 + c)$
Multiquadric	$k(\mathbf{x}, \mathbf{x}') = \sqrt{\|\mathbf{x} - \mathbf{x}'\|^2 + c^2}$
Inv. Multiquad.	$k(\mathbf{x}, \mathbf{x}') = 1/(\sqrt{\|\mathbf{x} - \mathbf{x}'\|^2 + \theta^2})$
Power	$k(\mathbf{x}, \mathbf{x}') = -\|\mathbf{x} - \mathbf{x}'\|^d$
Log	$k(\mathbf{x}, \mathbf{x}') = -\log(\|\mathbf{x} - \mathbf{x}'\|^d + 1)$

In the context of GPs, kernels with more hyperparameters can be efficiently inferred as we have seen before. This is an opportunity to exploit assymetries in the feature space by including a parameter per feature, as in the very common anisotropic squared exponential (SE) kernel function:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \nu \exp\left(-\sum_{f=1}^{F} \frac{(x_i^f - x_j^f)^2}{2\sigma_f^2} \right) + \sigma_n^2 \delta_{ij},$$

where x_i^f represents the feature f of the input vector $\mathbf{x}_i$, ν is a scaling factor, σ_n is the standard deviation of the (estimated) noise, and a σ_f is the length-scale per input features, $f = 1, \dots, F$. This is a very flexible covariance function that typically suffices to tackle most of the problems. However, note that a SE typically can approximate smoothly-varying functions, which may not be the case in particular problems. Also, note that when the data is *structured*, i.e. data reveals a particular (e.g. time, spatial) structure, the design of the covariance is of paramount relevance, and many approaches have exploited standard properties of functional analysis to do so [33]. We will advance in this discussion in the next section.

D. Gaussian processes exemplified

Let us illustrate the solution of GPR in a toy example. In Fig. 1 we include an illustrative example with 6 training points in the range between -2 and $+2$. We firstly depict several random functions drawn from the GP prior and then we include functions drawn from the posterior. We have chosen an isotropic Gaussian kernel and $\sigma_\nu = 0.1$. We have plotted the mean function plus/minus two standard deviations (corresponding to a 95% confidence interval). Typically, the hyperparameters are unknown, as well as the mean, covariance and likelihood functions. We assumed a Squared Exponential (SE) covariance funtion and learned the optimal hyperparameters by minimizing the negative log marginal likelihood (NLML) w.r.t. the hyperparameters. We observe three different regions in the figure. Below $x = -1.5$, we do not have samples and the GPR provides the solution given by the prior (zero mean and ± 2). At the center, where most of the data points lie, we have a very accurate view of the latent function with small error bars (close to $\pm 2\sigma_\nu$). For $x > 0$, we do not have training samples neither so we have same behaviour. GPs

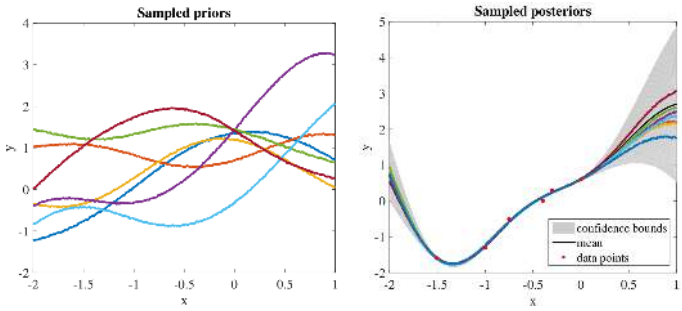

Fig. 1. Example of a Gaussian process. Left: some functions drawn at random from the GP prior. Right: some random functions drawn from the posterior, i.e. the prior conditioned on 6 noise-free observations indicated in red dots. The shaded area represents the pointwise mean plus and minus two times the standard deviation for each input value (corresponding to the 95% confidence region). It can be noted that the confidence intervals become large for regions far from the observations.

typically provide an accurate solution where the data lies and high error bars where we do not have available information and, consequently, we presume that the prediction in that area is not accurate. This is why in regions of the input space without points the confindence intervals are wide resembling the prior distribution.

E. Source code and toolboxes

The most widely known sites to obtain free source code on GP modeling are GPML[3] and GPstuff[4]. The former website centralizes the main activities in GP modeling and provides up-to-date resources concerned with probabilistic modeling, inference and learning based on GPs, while the latter is a versatile collection of GP models and computational tools required for inference, sparse approximations and model assessment methods. Both sites are highly useful for the reader interested in learning the main aspects of GP modeling, as they provide free code, demos, and pointers to relevant tutorials and books.

We also recommend to the interested reader in regression in general, our MATLAB SimpleR[5] toolbox that contains many regression tools organized in families: tree-based, bagging and boosting, neural nets, kernel regression methods, and several Bayesian nonparametric models like GPs. The toolbox is intended for practitioners with little expertise in machine learning, and that may want to assess advanced methods in their problems easily.

III. Advances in Gaussian Process Regression

In this section, we review some recent advances in GPR especially suited for remote sensing data analysis. We will review the main aspects to design covariance functions that capture non-stationarities and multiscale time relations, GPs that can learn arbitrary transformations of the observed variable and noise models, as well as to tackle the problem of multitask and multioutput problems, very common in our field.

[3] http://www.gaussianprocess.org/
[4] http://becs.aalto.fi/en/research/bayes/gpstuff/
[5] http://www.uv.es/gcamps/code/simpleR.html

A. Structured, non-stationary and multiscale GPR

Commonly used kernels families include the squared exponential (SE), periodic (Per), linear (Lin), and rational quadratic (RQ), cf. Table I. Illustration of the base kernel and drawings from the GP prior is shown in Fig. 2. These base kernels can be actually combined following simple operations: summation, multiplication or convolution. This way one may build sophisticated covariances from simpler ones. Note that the same essential property of kernel methods apply here: a valid covariance function must be positive semidefinite. In general, the design of the kernel should rely on the information that we have for each estimation problem and should be designed to get the most accurate solution with the least amount of samples.

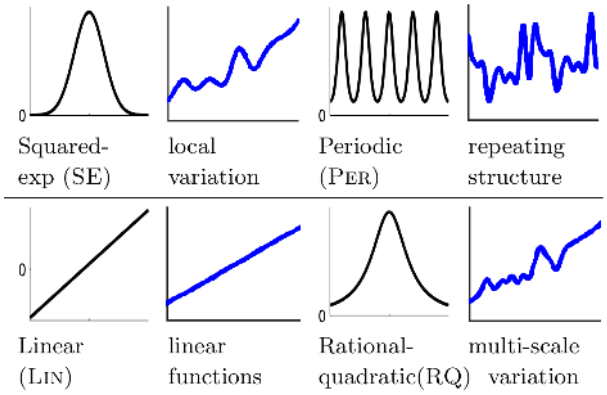

Fig. 2. Left and third columns: base kernels $k(\cdot, 0)$. Second and fourth columns: draws from a GP with each repective kernel. The x-axis has the same range on all plots.

"Advanced GP models are now able to cope with structured domains in time and space, a wide diversity of noise sources, multi-output problems, and skewed distributions of the observed variables."

In Fig. 2, all the base kernels are one-dimensional. Nevertheless, kernels over multidimensional inputs can be actually constructed by adding and multiplying kernels over individual dimensions. By summing kernels, we can model the data as a superposition of independent functions, possibly representing different structures in the data. For example, in multitemporal image analysis, one could for instance dedicate a kernel for the time domain (perharps trying to capture trends and seasonal effects) and another kernel function for the spatial domain (equivalently capturing spatial patterns and auto-correlations). In time series models, sums of kernels can express superposition of different processes, possibly operating at different scales: very often changes in geophysical variables through time occur at different temporal resolutions (hours, days, etc.), and this can be incorporated in the prior covariance with those simple operations. In multiple dimensions, summing kernels gives additive structure over different dimensions, similar to generalized additive models [11]. Alternatively, multiplying kernels allows us to account for interactions between different input dimensions or different

notions of similarity. In the following section, we show how to design kernels that incorporate particular time resolutions, trends and periodicities.

B. Time-based covariance for GPR

As already stated before, *time* is an additional and very important variable to be considered in many remote sensing applications. Signals to be processed typically show particular characteristic, with time-dependent cycles and trends. One could of course include time t_i as an additional feature in the definition of the input samples. This *stacked approach* [34] essentially relies on a covariance function $k(\mathbf{z}_i, \mathbf{z}_j)$, where $\mathbf{z}_i = [t_i, \mathbf{x}_i]^\top$. This is very convenient as it does not include additional hyper-parameters to learn, but has an important shortcoming: the time relations are naively left to the nonlinear regression algorithm, and hence no explicit time structure model is assumed. In order to cope with such temporal behavior of the observed signal in a more consistent way, one can use a linear combination (or composite) of different kernels: one dedicated to capture the different temporal characteristics, and the other to the feature-based relations. A simple strategy, quite common in statistics and signal processing is to rely on a tensor kernel

$$k(\mathbf{z}_i, \mathbf{z}_j) = k(\mathbf{x}_i, \mathbf{x}_j) \times k(t_i, t_j),$$

but more sophisticated structures can be adopted. The issue here is how to design kernels capable to deal with non-stationary processes.

A possible approach is to use a *stationary* covariance operating on the variable of interest after being mapped with a nonlinear function engineered to discount such undesired variations. This approach was used in [35] to model *spatial patterns* of solar radiation with GPR. It is also possible to adopt a squared exponential (SE) as stationary covariance acting on the *time* variable mapped to a two-dimensional *periodic space* $\mathbf{z}(t) = [\cos(t), \sin(t))]^\top$, as explained in [23],

$$k(t_i, t_j) = \exp\left(-\frac{\|\mathbf{z}(t_i) - \mathbf{z}(t_j)\|^2}{2\sigma_t^2}\right), \quad (6)$$

which gives rise to the following periodic covariance function

$$k(t_i, t_j) = \exp\left(-\frac{2\sin^2[(t_i - t_j)/2]}{\sigma_t^2}\right), \quad (7)$$

where σ_t is a hyper-parameter characterizing the periodic scale and needs to be inferred. It is not clear, though, that the seasonal trend is exactly periodic, so we modify this equation by taking the product with a squared exponential component, to allow a decay away from exact periodicity:

$$k_2(t_i, t_j) = \gamma \exp\left(-\frac{2\sin^2[\pi(t_i - t_j)]}{\sigma_t^2} - \frac{(t_i - t_j)^2}{2\sigma_d^2}\right), \quad (8)$$

where the time variable t is measured in years, γ gives the magnitude of the kernel function, σ_t the smoothness of the periodic component, σ_d represents the *decay-time* for the periodic component, and the period has been fixed to one year. Therefore, our final covariance is expressed as

$$k([\mathbf{x}_i, t_i], [\mathbf{x}_j, t_j]) = k_1(\mathbf{x}_i, \mathbf{x}_j) + k_2(t_i, t_j), \quad (9)$$

where $k_1(\mathbf{x}_i, \mathbf{x}_j)$ and $k_2(t_i, t_j)$ are two kernel functions working with the input and the time variable, respectively. The kernel k is then parameterized by only three more hyperparameters collected in $\boldsymbol{\theta} = \{\nu, \sigma_1, \ldots, \sigma_F, \sigma_n, \sigma_t, \sigma_d, \gamma\}$.

TABLE II
VARIABLES AND THEIR SOURCE CONSIDERED IN THIS PROBLEM OF GLOBAL SOLAR IRRADIATION PREDICTION.

Source	Data	Units	min-max
Cimel sunphotometer	Aerosol Optical Depth	-	0.01-1.38
Brewer spectrophotometer	Total Ozone	Dobson	242.50-443.50
Atmospheric sounding	Total Precip. Water	mm	1.33-41.53
GFS	Cloud amount	%	2-79.2
Pyranometer	Measured global solar irradiation	kJ/m^2	4.38-31.15

We show the advantage of encoding such prior knowledge and structure in the relevant problem of solar irradiation prediction, which has direct applications in renewable energy. Solar irradiation prediction is a very important and challenging problem with direct applications in renewable energy. Solar is one of the most important green sources of energy, that is currently under expansion in many countries of the world, especially in those with more solar potential, such as mid-east and southern Europe countries [36], [37]. An accurate estimation of the energy production in solar energy systems involves the accurate prediction of solar irradiation, depending on different atmospheric variables [38]–[40].

Recently, a high number of machine learning techniques have been introduced to tackle this problem, mostly based on neural networks and support vector machines. We evaluate GPR for the estimation of solar irradiation. Noting the nonstationary temporal behavior of the signal, we develop a particular time-based composite covariance to account for the relevant seasonal signal variations. We use a unique meteorological dataset acquired at a radiometric station, that include both measurements, radiosondes, and numerical weather prediction models. The target variable is the real global solar irradiation that reaches the ground. Data from the AEMET radiometric observatory of Murcia (Southern Spain, 38.0° N, 1.2° W) were used. Specifically, global daily mean values from the measurements of a pyranometer have been considered[6]. These data range from January 1st, 2010, to December 31st, 2011. We removed data with missing values: the dataset finally contains 512 examples, and 10 input features (cf. Table II).

Table III reports the obtained results with GPR models and several statistical regression methods: regularized linear regression (RLR), support vector regression (SVR), relevance vector macine (RVM) and GPR. All methods were run with and without using two additional dummy time features containing the year and day-of-year (DOY). We will indicate the former case with a subscript, like e.g. SVR$_t$. First, including time information improves all baseline models. Second, the

[6]Brewer and Cimel networks as well as the pyranometer used are managed under a Quality Management System certified to ISO 9001:2008.

TABLE III
RESULTS FOR THE ESTIMATION OF THE DAILY SOLAR IRRADIATION OF LINEAR AND NONLINEAR REGRESSION MODELS. SUBSCRIPT METHOD$_t$ INDICATES THAT THE METHOD INCLUDES TIME AS INPUT VARIABLE. BEST RESULTS ARE HIGHLIGHTED IN BOLD, THE SECOND BEST IN ITALICS.

Method	ME	RMSE	MAE	R
RLR	0.27	4.42	3.51	0.76
RLR$_t$	0.25	4.33	3.42	0.78
SVR [41]	0.54	4.40	3.35	0.77
SVR$_t$	0.42	4.23	3.12	0.79
RVM [42]	0.19	4.06	3.25	0.80
RVM$_t$	0.14	3.71	3.11	0.81
GPR [23]	0.14	3.22	2.47	*0.88*
GPR$_t$	0.13	*3.15*	*2.27*	*0.88*
TGPR	*0.11*	**3.14**	**2.19**	**0.90**

best overall results are obtained by the GPR models, when including time information or not. Third, in particular, the proposed TGPR outperforms the rest in accuracy (RMSE, MAE) and goodness-of-fit (R), and closely follows the elastic net in bias (ME). TGPR performs better than GPR and GPR$_t$ in all quality measures.

C. Heteroscedastic GPR: Learning the noise model

The standard GPR is essentially homoscedastic, i.e., assumes constant noise power σ^2 for all observations. This assumption can be too restrictive for some problems. Heteroscedastic GPs, on the other hand, let noise power vary smoothly throughout input space, by changing the prior over ε_n to

$$\varepsilon_n \sim \mathcal{N}(0, e^{g(\mathbf{x}_n)})$$

and placing a GP prior over $g(\mathbf{x}) \sim \mathcal{GP}(\mu_0\mathbf{1}, k_{\boldsymbol{\theta}_g}(\mathbf{x}, \mathbf{x}'))$. Note that the exponential is needed[7] in order to describe the non-negative variance.The hyperparameters of the covariance functions of both GPs are collected in $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_g$, accounting for the signal and the noise relations, respectively.

Relaxing the homoscedasticity assumption into heteroscedasticity yields a richer, more flexible model that contains the standard GP as a particular case corresponding to a constant $g(\mathbf{x})$. Unfortunately, this also hampers analytical tractability, so approximate methods must be used to obtain posterior distributions for $f(\mathbf{x})$ and $g(\mathbf{x})$, which are in turn required to compute the predictive distribution over y_*. Next we summarize previous approaches to deal with the problem and the proposed variational alternative.

The heteroscedastic GP model was first described in [43], where an expensive Markov chain Monte Carlo (MCMC) procedure was used in order to implement full Bayesian inference. A faster but more limited method is presented in [44] in order to perform maximum a posteriori (MAP) estimation. These approaches have certain limitations: MCMC is hundreds of times slower, whereas MAP estimation does not integrate out all latent variables and is prone to overfitting. As an alternative to these costly previous approaches, variational techniques allow to approximate intractable integrals arising in Bayesian inference and machine learning in general. They are typically used to 1) provide analytical approximations to the posterior probability of the unobserved variables and hence do statistical inference over these variables; and 2) derive a lower bound for the marginal likelihood (or "evidence") of the observed data, which allows model selection because higher marginal likelihoods relate to greater probabilities of a model generating the data.

In order to overcome the aforementioned problems, a sophisticated variational approximation called *Marginalized Variational (MV)* approximation was introduced in [45]. The MV approximation renders (approximate) Bayesian inference in the heteroscedastic GP model both fast and accurate. In [45], an analytical expression for the Kullback-Leibler divergence between a proposal distribution and the true posterior distribution of $f(\mathbf{x})$ and $g(\mathbf{x})$ (up to a constant) was provided. Minimizing this quantity with regard both the proposal distribution and the hyper-parameters yields an accurate estimation of the true posterior while simultaneously performing model selection. Furthermore, the expression of the approximate mean and variance of the posterior of y_* (i.e., predictions) can be computed in closed form. We will refer to this variational approximation for heteroscedastic GP regression as VHGPR. A simple comparison between the homoscedastic canonical GP and the VHGPR model is shown in Fig. 3.

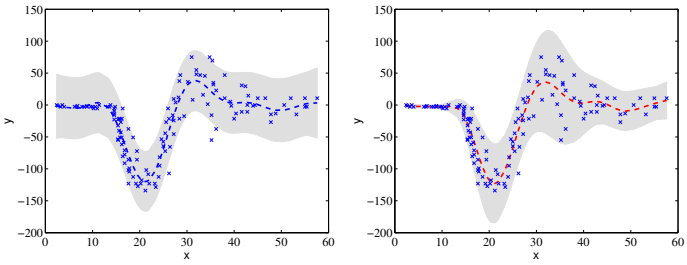

Fig. 3. Predictive mean and variance of the standard GP (left) and the heteroscedastic GP (right). It is noticeable that in the low noise regime the VHGP produces tighter confidence intervals as expected, while high noise variance associated to high signal variance (middle of the observed signal) the predictive variance is more reasonable too.

D. Warped GPR: Learning the output transformation

Very often, in practical applications, one transforms the observed variable to better pose the problem. Actually, it is a standard practice to linearize/uniformize the distribution of the observations (which is commonly skewed due to the sampling strategies in in-situ data collection) by applying non-linear link functions like the logarithmic, the exponential or the logistic functions.

Let us now review a GP model that automatically learns the optimal transformation by warping the observation space. The method is called *warped* GPR [46], and essentially warps observations $\mathbf{y}$ through a nonlinear parametric function g to a latent space:

$$z_i = g(y_i) = g(f(\mathbf{x}_i) + \varepsilon_i),$$

[7]Of course, other transformations are possible, just not as convenient.

where f is a possibly noisy latent function with d inputs, and g is a function with scalar inputs parametrized by $\boldsymbol{\psi}$. The function g must be *monotonic*, otherwise the probability measure will not be conserved in the transformation, and the distribution over the targets may not be valid [46]. It can be shown that replacing y_i by z_i into the standard GP model leads to an extended problem that can be solved by taking derivatives of the negative log likelihood function in (5), but now with respect to both $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ parameter vectors.

For both the GPR and WGPR models we need to define the covariance (kernel, or Gram) function $k(\cdot,)$, which should capture the similarity between samples. We used the standard Automatic Relevance Determination (ARD) covariance [23]. Model hyperparameters are collectively grouped in $\boldsymbol{\theta} = \{\nu, \sigma_n, \sigma_1, \ldots, \sigma_d\}$. In addition, for the WGPR we need to define a parametric smooth and monotonic form for g, which can be defined as:

$$g(y_i; \boldsymbol{\psi}) = \sum_{\ell=1}^{L} a_\ell \tanh(b_\ell \; y_i + c_\ell), \qquad a_\ell, \; b_\ell \geq 0,$$

where $\boldsymbol{\psi} = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$. Even though any other sensible parametrization could be used, this one is quite convenient since it yields a set of *smooth steps* whose size, steepness and position ara controlled by a_ℓ, b_ℓ and c_ℓ parameters, respectively. Recently, flexible non-parametric functions have replaced such parametric forms [47], thus placing another prior for $g(\mathbf{x}) \sim \mathcal{GP}(f, c(f, f'))$, whose model is learned via variational inference.

For illustration purposes, we focus on the estimation of imagesic chlorophyll-a concentrations from remote sensing upwelling radiance just above the images surface. A variety of bio-optical algorithms have been developed to relate measurements of images radiance to *in situ* concentrations of phytoplankton pigments, and ultimately most of these algorithms demonstrate the potential of quantifying chlorophyll-a concentrations accurately from multispectral satellite images color data. In this context, robust and stable non-linear regression methods that provide inverse models are desirable. In addition, we should note that most of the bio-optical models (such as Morel, CalCOFI and OC2/OC4 models) often rely on empirically adjusted nonlinear transformation of the observed variable (which is traditionally a ratio between bands).

Here we used the SeaBAM dataset [48], [49], which gathers 919 *in situ* pigment measurements around the United States and Europe. The dataset contains coincident *in situ* chlorophyll concentration and remote sensing reflectance measurements (Rrs(λ), [sr^{-1}]) at some wavelengths (412, 443, 490, 510 and 555 nm) that are present in the SeaWiFS images color satellite sensor. The chlorophyll concentration values range from 0.019 to 32.79 mg/m^3 (revealing a clear exponential distribution). Actually, even though SeaBAM data originate from various researchers, the variability in the radiometric data is limited. In fact, at high Chl-a concentrations, Ca [mg/m^3], the dispersion of radiance ratios Rrs(490)/Rrs(555) increases, mostly because of the presence of Case II waters. The shape of the scatterplots is approximately sigmoidal in log-log space. At lowest concentrations the highest Rrs(490)/Rrs(555) ratios are slightly lower than the theoretical limit for clear natural waters. See analysis in [22].

TABLE IV
RESULTS USING BOTH RAW AND EMPIRICALLY-TRANSFORMED OBSERVATION VARIABLES.

	ME	RMSE	MAE	R
Raw				
GPR	0.02	1.74	0.33	0.82
VHGPR	0.29	2.51	0.46	0.65
WGPR	0.08	1.71	0.30	0.83
Empirically-based				
GPR	0.15	1.69	0.29	0.86
VHGPR	0.15	1.70	0.29	0.85
WGPR	0.17	1.75	0.30	0.86

Table IV shows different scores –bias (mean error, ME), accuracy (root-mean-square error RMSE and mean absolute error MAE) and goodness-of-fit (Pearson's correlation R)– between the observed and predicted variable when using the raw data (no ad hoc transform at all) and the empirically adjusted transform. Results are shown for three flavours of GPs: the standard GP regression (GPR) [23], the variational heteroscedastic GP (VHGPR) [50], and the proposed warped GP regression (WGPR) [46], [47] for different rates of training samples. Empirically-based warping slightly improves the results over working with raw data for the same number of training samples, but this requires prior knowledge about the problem, time and efforts to fit an appropriate function. On the other hand, WGPR outperforms the rest of GPs in all comparisons over standard GPR and VHGPR ($\sim +1-10\%$). Finally, WGPR nicely compensates the lack of prior knowledge about the (possibly skewed) distribution of the observation variable.

E. Multitask and Multioutput GP models

Very often we deal with problems involving several variables to be estimated. Individual models are typically trained separately. This approach ignores the (potentially) cross-relations among output variables, e.g. between LAI, chlorophyll content and fractional cover. To account for this important relations in the output, some multitask and multioutput GP models are available. A simple form of multi-output GP models the response vector as a linear combination of a set of M latent GPs, thus giving rise to a block-diagonal covariance matrix $[\mathbf{K}_{ij}^m] = k_m(\mathbf{x}_i, \mathbf{x}_j)$, where $m = 1, \ldots, M$. More sophisticated models are now available to account for fixed correlations between output variables[8]. An effective model based on GPs for multitask problem is called the Gaussian process regression networks (GPRN) [51]. The model combines the properties of Bayesian neural networks with the non-parametric flexibility of GPs.

All these approaches, however, suffer when the output dimensionality is very high. In what follows, we show a much simpler approach to deal with this problem. In particular we focus on the estimation of water vapor profiles, which is an

[8] http://gaussianprocess.com/publications/multiple_output.php

important parameter for weather forecasting and atmospheric chemistry studies [52]. Observations from spaceborne high spectral resolution infrared sounding instruments can be used to calculate the profiles of such atmospheric parameters with unprecedented accuracy and vertical resolution [53]. We focus on the data coming from the Infrared Atmospheric Sounding Interferometer (IASI), that provides radiances in 8461 spectral channels, between 3.62 and 15.5 μm with a spectral resolution of 0.5 cm^{-1} after apodization [54]. This huge input data along the high output dimensionality (the variable is sampled at 137 points in the atmospheric column) makes the direct application of the previous methods unbearable. Alternatively, and noting the high vertical correlation of the profiles, we opted for a simpler strategy: develop a unique GP model predicting simultaneously all the PCA-projected state vector onto the top p principal components, and solve

$$\mathbf{\Lambda} = (\mathbf{K_{ff}} + \sigma^2 \mathbf{I}_n)^{-1}\mathbf{Y},$$

where $\mathbf{Y}$ contains in the columns the p scores (projected variables). This approach will be again exploited for RTM emulation in §6.

Results are given in Fig. 4. We trained a linear regression (LR) and a GP model using the first 100 principal components of an IASI orbit (2008-07-17), both using 5000 samples, and tested in several unseen data. Essentially we observe that GPs largely improve the linear regression models with an average gain of +1.5K, which are also statistically significant in all regions.

IV. Efficiency in Gaussian Process Regression

"With current advances in sparse, variational and structured learning, GPs have become extremely competitive in time and memory requirements."

The naive implementation of GPs in equations (3) and (4) grows as $\mathcal{O}(N^3)$, where N is the number of training samples. This make them unfeasible when a large number of training samples are available. In order to reduce the computation complexity of GPs, the general approach is to compute them using approximations[9]. The approximation methods can be broadly classified as in (1) *sparse methods*, (2) *localized regression methods*, and (3) *matrix multiplication methods*. Finally, we highlight some recent developments on GPs efficiency that exploit random features and particular kernel structures.

A. Sparse methods

These methods are also known as *low-rank covariance matrix approximation methods*, and are based on approximating the full posterior by expressions using matrices of lower rank $M \ll N$, where the M samples are typically selected to represent the dataset well, e.g. via clustering or smart sampling. Because the selected M samples represent al others, this methods are considered to be *global*, as opposed to the *local* methods described in next section. These global methods are well suited for modeling smooth-varying functions with high correlations (i.e., long length-scales). They use all the data for predictions like the full GPs. Methods in this family are based on substituting the joint prior with a reduced one using a set of m latent variables $\mathbf{u} = [u_1, \ldots, u_M]^\top$ called *inducing variables* [55]. These latent variables are values of the Gaussian process corresponding to a set of input locations $X_\mathbf{u}$, called *inducing inputs*. By adopting a 'subsets of data' (SoD) approach, the computational complexity drastically reduces to $\mathcal{O}(M^3)$, being $M \ll N$.

Examples of these approximation methods are the Subsets of Regressors (SoR), Deterministic Training Conditional (DTC), Fully Independent Training Conditional (FICT), Partially Independent Training Conditional (PICT) [55], and Partially Independent Conditional (PIC) [56]. All these methods, with some exceptions on PIC, are based on replacing the joint prior of training and test samples by an approximation following the assumption that they are *conditionally independent* given the set of M latent inducing variables. The exact *prior* are substituted by approximations based on the latent variables, which effectively lower the ranks of the covariance matrices. On the other hand, they use the exact *likelihood*. Table V summarizes the predictive distributions for the aforementioned methods, together with their computational complexities for training and test.

Regarding the performance of these methods, SoR obtains approximate predictive means, but unrealistic predictive variances. This is because its approximate prior is so restrictive that, given enough training data, the family of plausible functions under the posterior is very limited, leading to overconfident predictive variances. DTC solves this issue by relaxing the SoR prior and using the exact test conditional. It obtains the same predictive mean, and reliable predictive variances, but on the other hand it cannot be considered a true GP because the training and test covariances are computed in a different way. To partially solve and improve DTC, FITC approximates the training conditional using the exact values of the diagonal training covariance matrix. A further step on this direction comes from PITC [55], which instead of using an diagonal matrix uses a block diagonal matrix, thus preserving more exact values. Finally, PIC [56] improves PITC by relaxing the conditional independence condition between training and test samples, treating them equally according only to their location, which allows to exploit global and local information efficiently.

B. Localized regression methods

All methods described above are based on defining a set of inducing variables of size $M \ll N$ that represent all N points. This is the reason why these methods are classified as *global* methods. They are well suited for modeling smoothly-varying function with high correlations. But if M is too small, then the representation of the whole set is poor and the performance of the associated GP is low. On the other hand, the so called *local* methods are best suited for modeling highly-varying functions with low correlations, but they only use local data

[9]We intentionally omit other forms of efficiency that involve parallelization and hardware-specific approaches, and focus on pure GP algorithms.

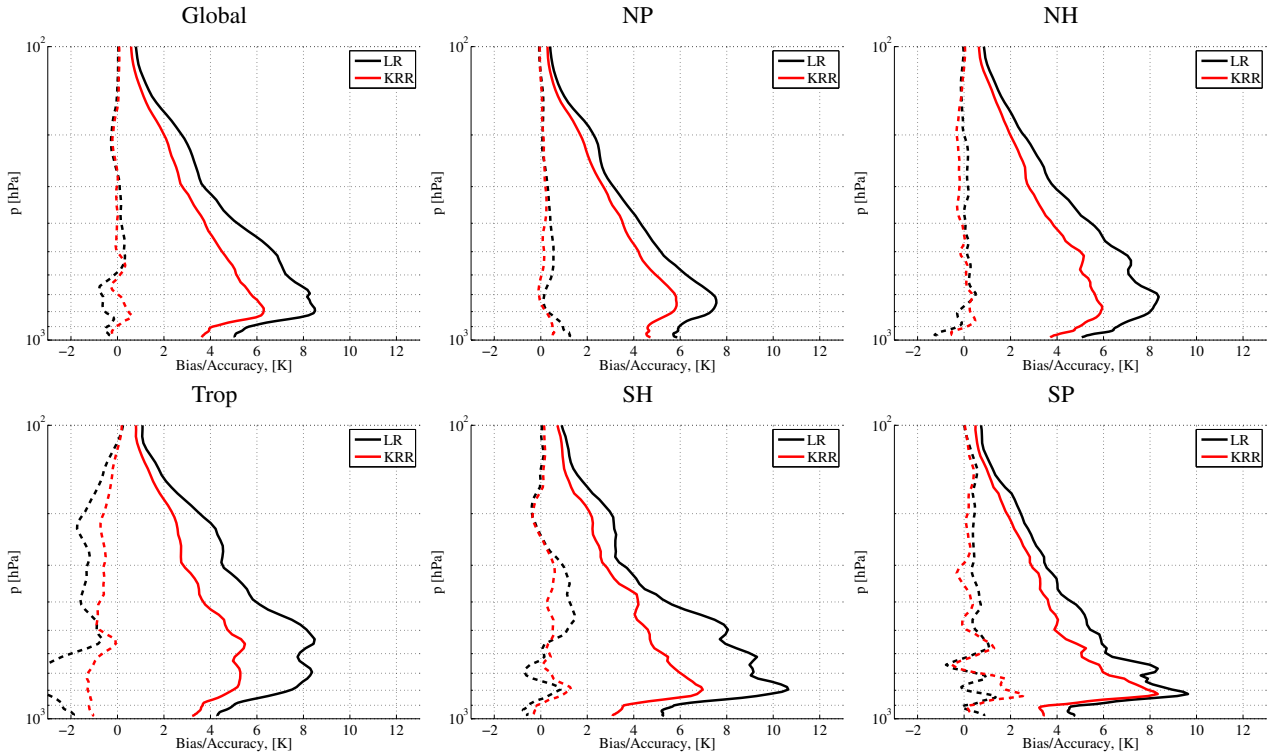

Fig. 4. Mean error (thin dashed lines) and RMSE (solid) throughout the atmospheric column for a linear regression and a GP model. Results are averaged for the whole globe and considered orbits, and for different regions (north/south poles, north/south hemispheres, tropics).

TABLE V

PREDICTIVE DISTRIBUTION FOR THE LOW-RANK APPROXIMATION METHODS DESCRIBED IN SECTION IV. THE LAST COLUMNS REFER TO THE COMPUTATIONAL COMPLEXITY FOR TRAINING, PREDICTIVE MEAN AND PREDICTIVE VARIANCE. N IS THE NUMBER OF SAMPLES, M IS THE NUMBER OF LATENT INDUCING VARIABLES (SEE MAIN TEXT), AND $B = M/N$ IS THE NUMBER OF BLOCKS FOR METHODS THAT USE THEM.

$Q_{\mathbf{a},\mathbf{b}} \equiv K_{\mathbf{a},\mathbf{u}} K_{\mathbf{u},\mathbf{u}}^{-1} K_{\mathbf{u},\mathbf{b}}$.

Method	Predictive mean, $\boldsymbol{\mu}_*$	Predictive variance, $\boldsymbol{\sigma}_*$	Training	Test mean	Test variance
SoR	$Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I})^{-1}\mathbf{y}$	$Q_{*,*} - Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I})^{-1}Q_{\mathbf{f},*}$	$\mathcal{O}(NM^2)$	$\mathcal{O}(M)$	$\mathcal{O}(M^2)$
DTC	$Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I})^{-1}\mathbf{y}$	$K_{*,*} - Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I})^{-1}Q_{\mathbf{f},*}$	$\mathcal{O}(NM^2)$	$\mathcal{O}(M)$	$\mathcal{O}(M^2)$
FITC	$Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \boldsymbol{\Lambda})^{-1}\mathbf{y}$	$K_{*,*} - Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \boldsymbol{\Lambda})^{-1}Q_{\mathbf{f},*}$	$\mathcal{O}(NM^2)$	$\mathcal{O}(M)$	$\mathcal{O}(M^2)$
PITC	As FITC, but $\boldsymbol{\Lambda} \equiv \text{blkdiag}[K_{\mathbf{f},\mathbf{f}} - Q_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I}]$.		$\mathcal{O}(NM^2)$	$\mathcal{O}(M)$	$\mathcal{O}(M^2)$
PIC	$K_{*,\mathbf{f}}^{\text{PIC}}(Q_{\mathbf{f},\mathbf{f}} + \boldsymbol{\Lambda})^{-1}\mathbf{y}$	$K_{*,*} - K_{*,\mathbf{f}}^{\text{PIC}}(Q_{\mathbf{f},\mathbf{f}} + \boldsymbol{\Lambda})^{-1}Q_{\mathbf{f},*}$	$\mathcal{O}(NM^2)$	$\mathcal{O}(M+B)$	$\mathcal{O}((M+B)^2)$

for predictions. Local GPs are obtained by just dividing the region of interest and training a GP in each division. This strategy has two main advantages: i) each local GP performs well on the (small) region it has been trained, and ii) each local GP is trained with a (relatively) small number of training points, thus reducing the computational cost. If dividing in B blocks such as $B = N/M$, the computational complexity goes from $\mathcal{O}(N^3)$ to $\mathcal{O}(NM^2)$. As main disadvantages, they show discontinuities at the limits between local GPs, and they perform poorly when predicting in regions far from their locality. This poses a problem when training data is only available in parts of the input region.

Recently new approximate methods have been presented that take the best from both approaches. One of such methods is PIC [56]. As we stated before, it successfully combines both global and local information by treating the input samples with regard to their location instead of if they are training or test samples. Moreover, the PIC prior covariance is a general case covering full GPs, FITC and local GPs. Actually, using $M = N$ inducing variables and setting them as training samples, then the exact covariance is obtained. On the other hand, if the blocks size is set to one then FITC is obtained, while if the number of inducing variables M is set to zero, then a pure local GP predictor is obtained. See [56] for details.

C. Matrix vector multiplication approximation methods

These methods are based on speeding up the solving of the linear system $(\mathbf{K} + \sigma^2\mathbf{I})\boldsymbol{\alpha} = \mathbf{y}$ using an iterative method, such as the conjugate gradient (CG). Each iteration of the CG method requires a matrix vector multiplication (MVM) which takes $O(N^2)$. The CG method obtains the exact solution if iterated N times, but one can obtain an approximate solution if the method is stopped earlier, so the total cost would be $\mathcal{O}(BN^2)$, being $B < N$ the number of CG iterations. To further speed up the computation ($\mathcal{O}(BN^2)$ is still too slow for large problems), the MVM multiplication needs to be accelerated. In CG, step one has to compute an MVM of the

form $\mathbf{k}_i\mathbf{v}$ for different i and $\mathbf{v}$, which is a sum of N products. This sum can be distributed and computed efficiently using hardware with a large number of cores, as in GPUs.

D. Recent advances

In recent years, we have witnessed a huge improvement in GP runtime and memory demands. Inducing methods became popular but may lack expressive power of the kernel. A very useful approach is the sparse spectrum Gaussian Processes [57], which is somewhat related to random kitchen sinks in [58] that allows to approximate a kernel matrix with a set of random bases sampled from the Fourier domain. On the other hand, there are methods that try to exploit structure in the kernel, either based on Kronecker or Toeplitz methods. The limitations of these methods to deal with data in a grid have been remedied recently with the KISS GP [59], which generalizes inducing point methods for scalable GPs, and scales $\mathcal{O}(N)$ in time and storage for GP inference.

V. Analysis of Gaussian Process Models

An interesting possibility in GP models is to extract knowledge from the trained model. We will show in what follows three different approaches: 1) feature ranking exploiting the automatic relevance determination (ARD) covariance; 2) uncertainty estimation looking at the predictive variance estimates; and 3) the exploitation of the GP models to infer causal relations between biophysical variables under a fully empirical, non-interventional setting. We intentionally relegate to the next section the use of GP models to mimic radiative transfer models, as a way to encode physical knowledge in the statistical models.

A. Ranking features through the ARD covariance

One of the advantages of GPs is that during the development of the GP model the predictive power of each single band is evaluated for the parameter of interest through calculation of the ARD. Specifically, band ranking through σ_b may reveal the bands that contribute most to the development of a GP model. An example of the σ_b's for one GP model trained with field leaf chlorophyll content (Chl) data and with 62 CHRIS bands is shown in Fig. 5 (left). The band with highest σ_b is the least contributing to the model. It can be noted that a relatively few bands (about 8) were evaluated as crucial for Chl estimation, while the majority of bands were evaluated as less contributing.

This is in agreement with earlier works [24], [25] and does not necessarily mean that other bands are obstructing optimized accuracies. For instance, in [25] it was demonstrated using the same CHRIS dataset that accuracies remained constant when removing iteratively the least contributing band. Only when less than 4 bands were left accuracies started to degrade rapidly Fig. 5 (right).

"GP allows us clean inspection of the knowledge encoded in the model: from the relative relevance of drivers to the uncertainty of the estimates."

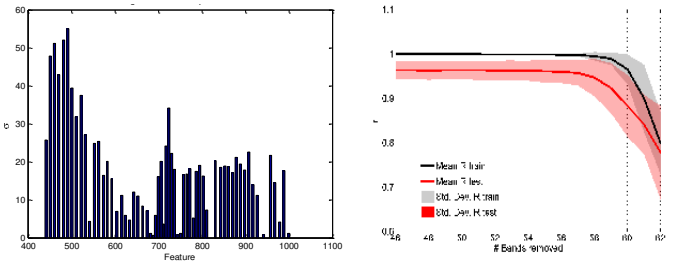

Fig. 5. Estimated σ_b values for one GP model using 62 CHRIS bands (left). The lower the σ_b the more important the band is for regression. Chl r and standard deviation (SD) of training and validation for GP fittings using backward elimination of worst σ_b. (right)

Hence, all CHRIS bands can be used without running the risk of losing accuracy. Of more interest here is identifying where most relevant bands are located. Essentially, the figure suggests that the most relevant spectral region is to be found between 550 and 1000 nm. This means that starting from the green spectral region the full CHRIS spectrum proved to be a valuable Chl detector. Most contributing bands were positioned around the red edge, at 680 and 730 nm respectively, but not all bands within the red edge were evaluated as relevant. This is due to when having a large number of bands available then neighbouring bands do not provide much additional information and can thus be considered as redundant. Remarkably, a few relevant bands fell within the 950-1000 nm region, which is outside the Chl absorption region. A reason for why these bands were evaluated as important is that at canopy scale the measured reflectance is not only related to biochemistry but also governed by variation in structural descriptors and abiotic factors such as variations in soil cover (e.g., due to soil composition and soil moisture). Effectively, the near-infrared (NIR) part of the reflectance is particularly affected by vegetation structure and water content [60]. Consequently, the Chl sensitivity in the NIR may be driven by secondary relationships, as also observed by [61], [62].

Consequently the σ_b proved to be a valuable tool to detect most sensitive bands of a sensor towards a biophysical parameter. A more systematic analysis was applied by sorting the bands on their relevance and counting the band rankings over 50 repetitions. In [24] the four most relevant bands were tracked for Chl, LAI and fCOVER and for different Sentinel-2 settings. It demonstrated the potential of Sentinel-2, with its new band in the red-edge, for vegetation properties estimation. Also in [12] σ_b were used to analyze band sensitivity of Sentinel-2 towards LAI. A similar approach was pursued on analyzing leaf Chl based on tracking the most sensitive spectral regions of sun-induced fluorescence data [63], as displayed in Fig. 6.

B. Uncertainty intervals

In this section, we use GP models for retrieval and portability in space and time. For this, we will exploit the associated predictive variance (i.e. uncertainty interval) provided by GP models. Consequently, retrievals with high uncertainties refer

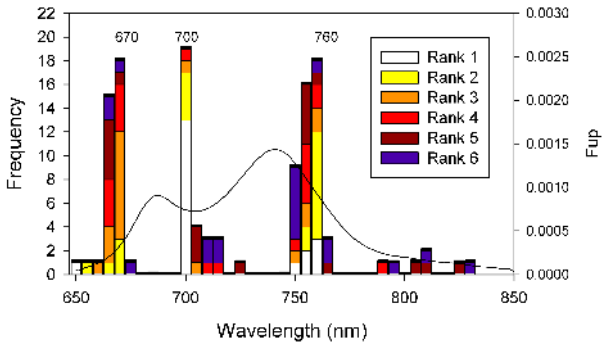

Fig. 6. Frequency plots of the top eight ranked bands with lowest σ_b values in 20 runs of GPR prediction of Chl based on upward fluorescence (F_{up}) emission. An emission curve is given as illustration.

to pixel spectral information that deviates from what has been represented during the training phase. In turn, low uncertainties refer to pixels that were well represented in the training phase. The quantification of variable-associated uncertainties is a strong requirement when remote sensing products are ingested in higher level processing, e.g. to estimate ecosystem respiration, photosynthetic activity, or carbon sequestration [64].

The application of GPs for the estimation of biophysical parameters was initially demonstrated in [25]. A locally collected field dataset called SPARC-2003 at Barrax (Spain) was used for training and validation of GPs for the vegetation parameters of LAI, Chl and fCOVER. Sufficiently high validation accuracies were obtained ($R^2 > 0.86$) for processing a CHRIS image into these parameters, as shown in Fig. 7. While generated maps provide spatially-explicit information about the vegetation status, the associated uncertainty maps can be certainly more revealing. Within these maps, areas with reliable retrievals are clearly distinguished from areas with unreliable retrievals. Low uncertainties were found on irrigated areas and harvested fields. High uncertainties were found on areas with remarkably different spectra, such as bright, whitish calcareous soils, or harvested fields. This does not necessarily mean that the estimates were wrong. Rather, it informs that the input spectrum deviates from what has been presented during the training stage, thereby imposing uncertainties to the retrieval. Hence, a practical implication of the uncertainty maps is that they detect those areas that may benefit from a denser sampling regime.

Nevertheless one has to be careful with its interpretation. Given that $\pm\sigma$ represents the uncertainty interval around the mean predictions, requires that they need to be interpreted in relation to the estimates. For instance, an Chl uncertainty interval of about 5 would be more problematic for a mean estimate of 5 μg/cm^2 than of 50 μg/cm^2. Therefore, calculating the relative uncertainties, i.e. the coefficient of variation, $CV[\%] = 100 \times \sigma/\mu$, may be more meaningful. Relative uncertainties maps can then be evaluated against an uncertainty threshold. For instance Global Climate Observing System (GCOS) proposed a threshold of 20% [65]. Consequently, relative uncertainty intervals can be used as a quality mask, thereby discarding retrievals that are considered of unacceptable quality.

GP models were subsequently applied to the SPARC dataset that was resampled to different Sentinel-2 band settings (4, 8 and 10 bands) and then uncertainties were inspected [24]. On the whole, adding spectral information led to reduction of uncertainties and thus more meaningful biophysical parameter maps. It remains nevertheless to be questioned how robust the locally-trained GP models function when applied to other sites and conditions. In this respect, the delivery of uncertainty estimates may enable to evaluate the portability of the regression model. Specifically, when uncertainty intervals as produced by a locally trained GP model over an arbitrary site are on the same order as those produced over the successfully validated reference site, then it can be reasonably assumed that the retrievals are of the same quality as the retrievals of the reference site. Thus, when successfully validated over a reference imagery then the uncertainty estimates can work as a quality indicator. Note, however, that the previous conclusions should be taken with caution, given that the predictive variance provided by the GP is just an *estimate* of the actual uncertainty.

Accordingly, the locally-trained GP models were applied to simulated Sentinel-2 images in a follow-up study [66]. Time series over the local Barrax site as well images across the world were processed. Also the role of an extended training dataset ($TrEx$; adding spectra of non-vegetated surfaces) were evaluated. Subsequently the uncertainty values were analyzed. By using $TrEx$ not only further improved performances but also allowed a decrease in theoretical uncertainties. This underlines the importance of a broad and diverse training dataset. More importantly, the GP models were successfully applied to simulated Sentinel-2 images covering various sites; associated relative uncertainties were on the same order as those generated by the reference image, i.e., vegetated surfaces were below the 20% requirements. However, typically large uncertainty variation within an image was observed due to surface heterogeneity. Contrary to the common belief that statistical methods are poorly transportable, larger ranges of uncertainties within an image than between images were observed.

As a final example, uncertainty estimates were exploited to assess the robustness of the retrievals at multiple spatial scales. In [26], retrievals from hyperspectral airborne and spaceborne data over the Barrax area were compared. Based on the spareborne SPARC-2003 dataset, GP developed a model that was excellently validated (R^2: 0.96). The SPARC-trained GP model was subsequently applied to airborne CASI flightlines (Barrax, 2009) to generate Chl maps. The accompanying uncertainty maps provided insight in the robustness of the retrievals. In general similar uncertainties were achieved by both sensors, which is encouraging for upscaling estimates from field to landscape scale.

The high spatial resolution of CASI in combination with the uncertainties allows us to observe the spatial patterns of retrievals in more detail. However, uncertainties worsened somewhat when inspecting the CASI airborne maps. Particularly poorer uncertainties were found on recently irrigated agricultural areas, probably due to the spectral mixture between elongated vegetation and wet soil cover. The reason for this decrease is that at the airborne scale a much more detailed

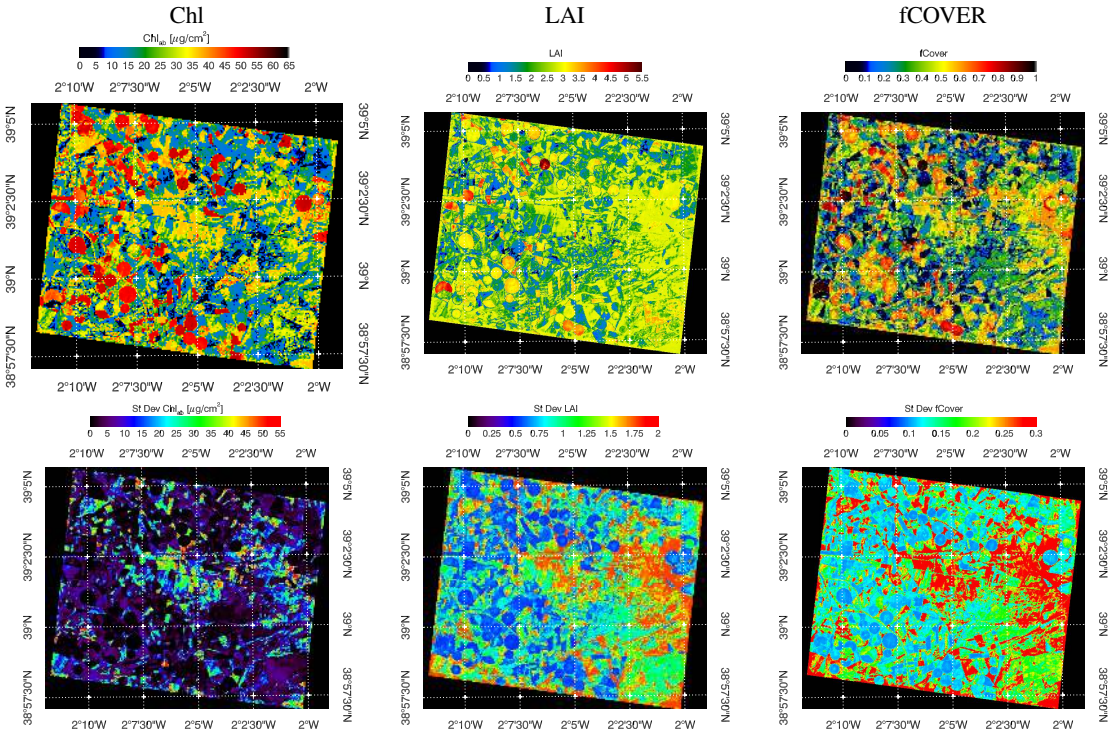

Fig. 7. Prediction maps (top) and associated uncertainty intervals (bottom), generated with GP and four bands of the CHRIS 12-07-2003 nadir image.

variation in land cover types are being observed than at the spaceborne scale of CHRIS. Some examples of mean estimates and associated uncertainties are shown in Fig. 8.

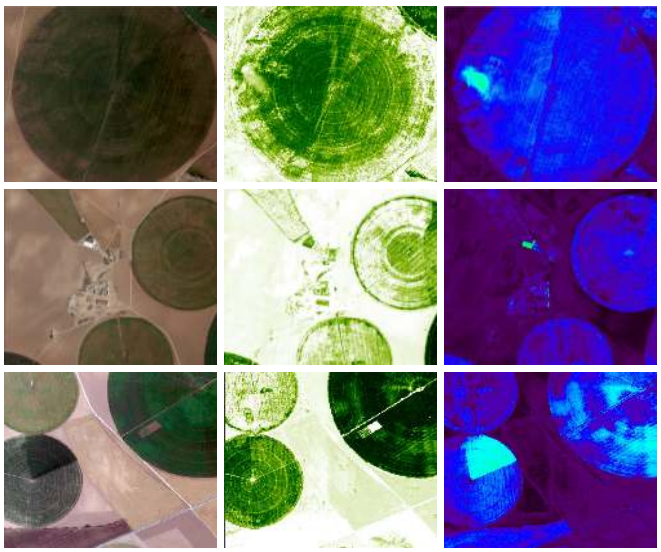

Fig. 8. Three examples [top, middle, bottom] of CASI RGB snapshots [left], Chl estimates [middle], and related uncertainty intervals [right].

C. From correlation to causation

Establishing causal relations between random variables from empirical data is perhaps the most important challenge in today's Science. In this section, we use GP models for causal discovery. To this end, we follow the approach in [67] to discover causal relations between observed variables $\mathbf{x}$ and $\mathbf{y}$. The methodology performs nonlinear regression from $\mathbf{x} \rightarrow \mathbf{y}$ (and vice versa, $\mathbf{y} \rightarrow \mathbf{x}$) and assesses the independence of the forward, $r_f = \mathbf{y} - f(\mathbf{x})$, and backward residuals, $r_b = \mathbf{x} - g(\mathbf{y})$, with the input variable $\mathbf{y}$ (or $\mathbf{x}$). The statistical significance of the independence test tells the right direction of causation. Essentially, the framework exploits nonlinear, non-parametric regression to assess the plausibility of the causal link between two random variables in both directions: statistically significant residuals in just one direction indicate the true data-generating mechanism. The framework was extended in [68] to get rid of the possibly strong assumption about the noise distribution, and proposed maximizing a dependency measure between residuals and regressors.

"GPs permit inferring causal relations from observational data, an important leap towards understanding physics through machine learning."

Note that the estimation of causal relations with this model suffers when the noise is not Gaussian and we use linear models. Both scenarios pose serious identifiability problems, which have led to an increasing interest in nonlinear regression models that consider eventually non-Gaussian noise [69], [70]. The interest here is to assess causality by discounting elusive masking effects due to the noise Gaussianity assumption, as well as possibly skewed distributions of the observation variable. This is why we use for comparison standard GPR, VHGPR and WGPR.

We exemplify the approach in a relevant geoscience problem. The last few hundred years, human activities have precipitated an environmental crisis on Earth, commonly described as 'global climate change.' Since the discovery of fossil carbon as a convenient form of energy, the residues of

past photosynthetic carbon assimilation have been combusted to CO_2 and returned to the Earth's atmosphere. Terrestrial ecosystems absorb approximately 120 Gt of carbon annually from the atmosphere, about half is returned as plant respiration and the remaining 60 Gt yr^{-1} represent the Net Primary Production (NPP). Out of this, about 50 Gt yr^{-1} are returned to the atmosphere as soil/litter respiration or decomposition processes, while about 10 Gt yr^{-1} results in the Net Ecosystem Production (NEP). The problem here deals with estimating the causal relation between the photosynthetic photon flux density (PPFD), which is a measure of light intensity[10], and the NEP, which results from the potential of ecosystems to sequestrate atmospheric carbon. Discovering such relations may be helpful to better understand the carbon fluxes and to establish sinks and sources of carbon through the globe. We use here three data sets taken at a flux tower at site DE-Hai involving PPFD(total), PPFD(diffuse), PPFD(direct) drivers and the NEP consequence variable [71]. Results for all three scenarios are shown in Table VI, which generally confirm the good capabilities of the presented methods, leading to lower p-values for the forward direction, p_f (though similar p-values of the backward direction, p_b) for the GPR methods. We should stress that, as more flexible GP models are deployed, the sharpness in the causal detection becomes more evident. Interestingly, the heteroscedastic GP 'discounts' the noise effects so the dependency estimate becomes slightly more reliable.

TABLE VI
RESULTS IN THE 'PPFD CAUSES NEP' CAUSAL PROBLEM.

Method	p_f	p_b	Conclusion
GPR	3.86×10^{-61}	1.57×10^{-119}	PPFD(tot)→ NEP
WGPR	2.12×10^{-50}	3.33×10^{-115}	PPFD(tot)→ NEP
VHGPR	6.11×10^{-60}	2.50×10^{-109}	PPFD(tot)→ NEP
GPR	1.59×10^{-11}	1.24×10^{-79}	PPFD(diff)→ NEP
WGPR	1.17×10^{-11}	9.40×10^{-77}	PPFD(diff)→ NEP
VHGPR	2.44×10^{-12}	9.16×10^{-75}	PPFD(diff)→ NEP
GPR	2.05×10^{-8}	1.56×10^{-112}	PPFD(dir)→ NEP
WGPR	1.20×10^{-15}	3.67×10^{-110}	PPFD(dir)→ NEP
VHGPR	3.44×10^{-17}	1.01×10^{-115}	PPFD(dir)→ NEP

VI. EMULATING RADIATIVE TRANSFER MODELS THROUGH GAUSSIAN PROCESSES

A slightly different approach to the use GPs in RS is to use them as fast approximations to complex physical models, an approach with a long story in statistics [28], [32], [72]. These surrogate models or metamodels are generally orders of magnitude faster than the original model, and can then be used *in lieu* of it, opening the door to more advanced biophysical parameter estimation methods, using e.g. data assimilation (DA) concepts [73], [74].

[10]The total PPFD was measured here as the number of photons falling on a one square meter area per second, while NEP was calculated by photosynthetic uptake minus the release by respiration, which is known to be driven by either the total, diffuse or direct PPFD.

A. Function approximation, regularization and emulation

A function is a mapping from an input parameter space to an output space. Now consider that for a particular application, we wish to use a particular function, but we are only able to run this function for a limited number of times (this might be because the function is so complicated that it would take too long, for example). For our interest, in what follows, the reader may think of such function as being a radiative transfer model (RTM). One way to get around this limitation is to carry out an inference on the function itself. To do this, we need to place a prior that encodes our belief in the properties of the function (such as smoothness, continuity, finite values), and use the limited pairings of inputs and outputs of the function as our likelihood (e.g. the probability of the outputs given the inputs). A generic prior with the desirable properties mentioned above is a GP (with an associated covariance function, as explained above), and assuming the likelihood is also Gaussian and independent additive noise, we end up with a reparametrisation of the prior GP as the posterior. This means that we can now predict the output of our function for an arbitrary input $\mathbf{x}_*$, conditional on the limited sampling of input/output pairings of the original model. The prediction will provide an estimate of the function value $\mu_{\text{GP}*}$, but importantly, also an estimate of the predictive uncertainty, $\sigma^2_{\text{GP}*}$. If the GP is able to correctly reproduce the function where only a limited number of runs were available (which in this context is called the *simulator*), we can start using the GP in its stead. We term this use of GPs *emulation*, and it is an exploitation of the versatility of GPs to effectively cope with varied mappings (or simulators).

Although emulators might appear like a trivial diversion, they have a number of important advantages. First of all, if the simulator is computationally expensive, an emulator typically provides a very fast approximation to the simulator. Given the ability of the GPs to cope well with fairly non-linear problems, the method can be effective for a large number of complex physical models, and here we focus on radiative transfer (RT) models that describe in some detail the scattering and absorption of photons by the atmosphere, vegetation, etc. The emulator can thus be seen as a drop-in replacement for a complicated physical model. The fact that there is an associated uncertainty with the emulator prediction is of importance: the user can decide whether the emulation is accurate enough for the application at hand, or can propagate this emulation model error through the application. Having fast physical models opens new avenues to the use we can make of them. We will review some of these next.

B. From forward and backward models to statistical emulation

A particular problem often found in remote sensing is the inverse problem, where a physical RT model is used to interpret observations of e.g. surface directional reflectance or microwave backscatter in terms of biophysical parameters such as leaf area index (LAI), soil roughness, etc. The computational complexity of the models at hand usually makes analytic inversions intractable, and thus the inversion method typically results in a least squares problem, where the input parameters of the model are varied until a minimum difference with

the observations is found. EO data are however corrupted by uncertainty (additive noise, imaging artifacts, etc.) that degrade the information content in the data, and observations are typically only available over small spectral or angular regions, giving a partial overview of e.g. the land surface. Additionally, the processes that describe the fate of photons interacting with the scene are non linear. These effects conjure a situation where many possible combinations of input parameters result in an adequate description of the observations, and therefore a large uncertainty in the retrieved parameters. To help circumvent the ill-posed nature of the inverse problem, we would need to either add more prior information or more evidence (observations). The flexibility of the RT models makes the latter strategy possible, as RT models can usually account for different sensor configurations (geometry, spectral sampling, etc.) while keeping a consistent description of the scene. New observations are typically hard to come by, and new observations will again be limited by uncertainty and partial observation of the whole system. Adding prior information is thus a necessary way to better constrain the inverse problem. Prior estimates include parameter distributions (derived e.g. from expert knowledge, or from historical data), expectations of smoothness in time and space, or physiological models of vegetation growth. Ultimately, the calculation of the posterior is a complicated problem that can typically be solved by e.g. Markov Chain Monte Carlo (MCMC) methods, requiring many iterations (and therefore many executions of the RT model), or (under some assumptions) by a non-linear cost function minimisation problem. The latter is typically an iterative procedure, and for efficiency, gradient descent methods are required. Remember that the aim here is to infer the land surface parameters conditioned on the EO data and any other prior knowledge, with an estimate of the uncertainty of the parameters.

C. GP models as efficient emulators

GP emulators can be used in complex inverse problems settings to a great advantage. If MCMC methods are used, the physical model can be emulated directly, resulting in much faster exploration of parameter space. In cost function minimisation, the emulator can be used instead of the full model, but additionally, we can use the GP to approximate the gradient of the emulated model as:

$$\frac{\partial \mu_{\text{GP}*}}{\partial \mathbf{x}_*} = \left(\frac{\partial \mathbf{k}_{\mathbf{f}*}}{\partial \mathbf{x}_*}\right)^\top \left(\mathbf{K}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_n\right)^{-1} \mathbf{y}. \tag{10}$$

From Eq. 10, we see that higher order partial derivatives (e.g. the Hessian matrix of second order derivatives) are straightforward. The Hessian is important because in many cost function minimisation approaches, the inverse of this matrix as the maximum *a posteriori* point is the posterior covariance matrix, and thus an statement on the uncertainty of the retrieved parameters. A further benefit of numerically cheap approximations to the gradient is that local linearisations of the model are now available, allowing the use of efficient linear solvers to invert problems (either directly, or as part of an internal linear loop in the solution to the non linear problem). Ultimately, the ability of having fast surrogate models of the most computationally demanding part of the inversion problem allows us to practically implement inversion strategies that were practically impossible with these models, and to extend them to practical problem sizes.

"Emulators are statistical constructs that are able to approximate the RTM, although at a fraction of the computational cost, providing an estimation of uncertainty and function gradients."

A particular requirement in many RT models is the prediction of e.g. spectral reflectance over the solar reflective domain (broadly from 400 to 2500 nm), so that instrument band pass functions can be applied to the data. In order to emulate full spectra, we can extend the idea of principal component analysis (PCA) of hyperspectral data, where there are large degrees of spectral redundancy. Let our output data set $\mathbf{y}$ be given a stacking of N_t spectra. Each of these spectra can be approximately reconstructed from

$$\mathbf{y}_i \approx \sum_{j=1}^{L} \sigma_j \cdot \mathbf{w}_j, \tag{11}$$

where we only consider the first L principal components, and σ_j is the j-th score associated with the $\mathbf{w}_j$ principal component. In PCA, the principal components are orthogonal over the input set, so a strategy is to emulate the scores $\sigma_1, \ldots, \sigma_L$ with independent emulators, and then use these emulators to reconstruct a full spectrum (uncertainties and gradients can also follow through quite easily due to the linearity of Eq. 11).

D. An illustrative example

As an example, consider a coupled soil-leaf-canopy model over the solar reflective domain, PROSAIL [75]. We will use a simple linear spectral mixture model for the soil (hence assuming the soil properties are isotropic), the PROSPECT leaf optical properties model and the SAIL canopy RT model. Our aim is to map from a state made up of soil, leaf, canopy and parameters such as leaf area index (LAI), chlorophyll concentration, etc to top-of-canopy reflectance. This is an important example, as the coupled model can be used within a DA system to infer the properties of the land surface (vegetation structure and biochemistry) from atmospherically corrected directional surface reflectance. We show a validation of the emulation approach in Fig. 9, where the emulator has been trained with 250 input parameter-reflectance pairs (these have been chosen using a latin hypercube sampling design). Using the approach outlined in the previous Section for multivariate output, L in Eq. 11 was chosen to be 11, so as to encompass 99% of the variance in the training set. We see immediately that the emulator is virtually indistinguishable from the original model, with negligible bias in the validation, and a very small root mean squared error. Although PROSAIL is a fast model, this emulator is some 5000 times faster than the original in a contemporary PC, and in the evaluation of the GP, the gradient of PROSAIL is also calculated.

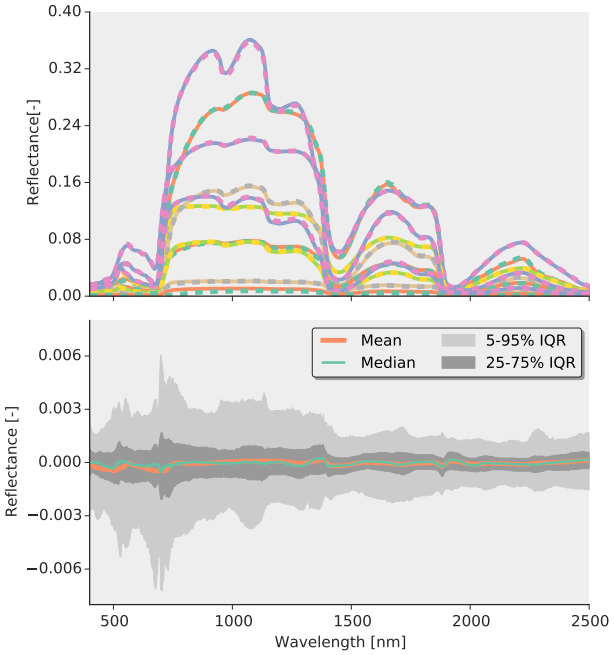

Fig. 9. Example of RT model emulation with GPs. The PROSAIL soil-leaf-canopy model is emulated spectrally. The top panel shows the complete model (full lines) and the emulated reflectance (dashed lines) for ten random input parameter sets. The bottom panel shows the mean, median, 5-95% and 25-75% interquantile ranges for the residuals of the full model minus the emulator. This example assumes a sun zenith angle of 30°, a view zenith angle of 0° and a relative azimuth of 0°, and the validation is done with a set of 1000 uniformly independent samples.

VII. Conclusions and Further Work

This paper provided a comprehensive survey to the field of Gaussian Processes (GPs) in the context of remote sensing data analysis, and in particular for statistical biophysical parameter estimation. We summarized the main properties of GPs and the advantages over other methods for estimation: essentially GPs can provide competitive predictive power, gives error-bars for the estimations, allows to design and optimize sensible kernel functions, and also to analyze the encoded knowledge in the model via automatic relevance determination kernels.

GP models offer as well a solid Bayesian framework to formulate new algorithms well-suited to the signal characteristics. We have seen for example that by incorporating proper priors, we can encompass signal-dependent noise, and infer parametric forms of warping the observations as an alternative to either ad hoc filtering or linearization, respectively. On the downside of GPs we need to mention the scalability issue: essentially, the optimization of GP models require computing determinants and invert matrices of size $n \times n$, which runs cubically in computational time and quadratically in memory storage. In the last years, however, great advances have appeared in machine learning and now it is possible to train GPs with millions of points in (almost) linear time.

All the developments were illustrated at a local and global planetary scales through a full set of illustrative examples in the field of geosciences and remote sensing. In particular, we treated important problems of ocean, land and atmospheric sciences: from accurate estimation of oceanic chlorophyll content and pigments, to vegetation properties (such as LAI or fluorescence) from multi- and hyperspectral sensors, as well as the estimation of atmospheric parameters (such as temperature, moisture and ozone) from infrared sounders.

The step forward we have made in this paper is to introduce and illustrate two relevant usages of the GP technology: first we studied the important issue of passing from regression to causation from empirical data, and also reviewed the field of approximating radiative transfer models with GPs. Both approaches, yet in its infancy, are promising to fully aboard the problem developing flexible statistical models that discover and incorporate physical knowledge about the problem. We envision more exciting developments in the intersection of physics and machine intelligence.

VIII. Acknowledgements

The authors wish to deeply acknowledge the collaboration, comments and fruitful discussions with many researchers during the last decade on GP models for remote sensing and geoscience applications: Miguel Lázaro-Gredilla (Vicarious), Robert Jenssen (Univ. Tromsø, Norway), Martin Jung (MPI, Jena, Germany), and Salcho Salcedo-Saez (Univ. Alcalá, Madrid, Spain).

References

[1] W. A. Dorigo, R. Zurita-Milla, A. J. W. de Wit, J. Brazile, R. Singh, and M. E. Schaepman, "A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling," *International Journal of Applied Earth Observation and Geoinformation*, vol. 9, no. 2, pp. 165–193, 2007.

[2] M. Schaepman, S. Ustin, A. Plaza, T. Painter, J. Verrelst, and S. Liang, "Earth system science related imaging spectroscopy-An assessment," *Rem. Sens. Env.*, vol. 113, no. 1, pp. S123–S137, 2009.

[3] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini, "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services," *Rem. Sens. Env.*, vol. 120, pp. 25–36, 2012.

[4] C. Donlon, B. Berruti, A. Buongiorno, M.-H. Ferreira, P. Féménias, J. Frerick, P. Goryl, U. Klein, H. Laur, C. Mavrocordatos, J. Nieke, H. Rebhan, B. Seitz, J. Stroede, and R. Sciarra, "The Global Monitoring for Environment and Security (GMES) Sentinel-3 mission," *Remote Sensing of Environment*, vol. 120, pp. 37–57, 2012.

[5] T. Stuffler, C. Kaufmann, S. Hofer, K. Farster, G. Schreier, A. Mueller, A. Eckardt, H. Bach, B. Penné, U. Benz, and R. Haydn, "The EnMAP hyperspectral imager-An advanced optical payload for future applications in Earth observation programmes," *Acta Astronautica*, vol. 61, no. 1-6, pp. 115–120, 2007.

[6] D. Roberts, D. Quattrochi, G. Hulley, S. Hook, and R. Green, "Synergies between VSWIR and TIR data for the urban environment: An evaluation of the potential for the Hyperspectral Infrared Imager (HyspIRI) Decadal Survey mission," *Rem. Sens. Env.*, vol. 117, pp. 83–101, 2012.

[7] D. Labate, M. Ceccherini, A. Cisbani, V. De Cosmo, C. Galeazzi, L. Giunti, M. Melozzi, S. Pieraccini, and M. Stagi, "The PRISMA payload optomechanical design, a high performance instrument for a new hyperspectral mission," *Acta Astronautica*, vol. 65, no. 9-10, pp. 1429–1436, 2009.

[8] S. Kraft, U. Del Bello, M. Drusch, A. Gabriele, B. Harnisch, and J. Moreno, "On the demands on imaging spectrometry for the monitoring of global vegetation fluorescence from space," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 8870, 2013.

[9] J. Verrelst, G. Camps-Valls, J. Muñoz Marí, J. Rivera, F. Veroustraete, J. Clevers, and J. Moreno, "Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties - a review," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2015.

[10] M. Berger, J. Moreno, J. A. Johannessen, P. Levelt, and R. Hanssen, "ESA's sentinel missions in support of earth system science," *Rem. Sens. Env.*, vol. 120, pp. 84–90, 2012.

[11] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. New York, USA: Springer-Verlag, 2009.

[12] J. Verrelst, J. Rivera, F. Veroustraete, J. Muñoz Marí, J. Clevers, G. Camps-Valls, and J. Moreno, "Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods - A comparison," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2015.

[13] G. Camps-Valls, D. Tuia, L. Gómez-Chova, and J. Malo, Eds., *Remote Sensing Image Processing*. Morgan & Claypool, Sept 2011.

[14] C. Bacour, F. Baret, D. Béal, M. Weiss, and K. Pavageau, "Neural network estimation of LAI, fAPAR, fCover and LAI×Cab, from top of canopy MERIS reflectance data: Principles and validation," *Rem. Sens. Env.*, vol. 105, no. 4, pp. 313–325, 2006.

[15] F. Baret, M. Weiss, R. Lacaze, F. Camacho, H. Makhmara, P. Pacholcyzk, and B. Smets, "Geov1: LAI and FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. part1: Principles of development and production," *Rem. Sens. Env.*, vol. 137, no. 0, pp. 299 – 309, 2013.

[16] C. Beer, M. Reichstein, E. Tomelleri, P. Ciais, M. Jung, N. Carvalhais, C. Rödenbeck, M. A. Arain, D. Baldocchi, G. B. Bonan, A. Bondeau, A. Cescatti, G. Lasslop, A. Lindroth, M. Lomas, S. Luyssaert, H. Margolis, K. W. Oleson, O. Roupsard, E. Veenendaal, N. Viovy, C. Williams, F. I. Woodward, and D. Papale, "Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate," *Science*, vol. 329, no. 834, 2010.

[17] M. Jung, M. Reichstein, H. A. Margolis, A. Cescatti, A. D. Richardson, M. A. Arain, A. Arneth, C. Bernhofer, D. Bonal, J. Chen, D. Gianelle, N. Gobron, G. Kiely, W. Kutsch, G. Lasslop, B. E. Law, A. Lindroth, L. Merbold, L. Montagnani, E. J. Moors, D. Papale, M. Sottocornola, F. Vaccari, and C. Williams, "Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations," *Journal of Geophysical Research: Biogeosciences*, vol. 116, no. G3, pp. 1–16, 2011.

[18] L. R. Sarker and J. E. Nichol, "Improved forest biomass estimates using ALOS AVNIR-2 texture indices," *Rem. Sens. Env.*, vol. 115, no. 4, pp. 968–977, 2011.

[19] L. Guanter, Y. Zhang, M. Jung, J. Joiner, M. Voigt, J. A. Berry, C. Frankenberg, A. Huete, P. Zarco-Tejada, J.-E. Lee, M. S. Moran, G. Ponce-Campos, C. Beer, G. Camps-Valls, N. Buchmann, D. Gianelle, K. Klumpp, A. Cescatti, J. M. Baker, and T. J. Griffis, "Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence," *Proceedings of the National Academy of Sciences, PNAS*, 2014.

[20] F. Yang, M. White, A. Michaelis, K. Ichii, H. Hashimoto, P. Votava, A.-X. Zhu, and R. Nemani, "Prediction of Continental-Scale Evapotranspiration by Combining MODIS and AmeriFlux Data Through Support Vector Machine," *IEEE Trans. Geosc. Rem. Sens.*, vol. 44, no. 11, pp. 3452–3461, nov. 2006.

[21] S. Durbha, R. King, and N. Younan, "Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer," *Rem. Sens. Env.*, vol. 107, no. 1-2, pp. 348–361, 2007.

[22] G. Camps-Valls, L. Gómez-Chova, J. Vila-Francés, J. Amorós-López, J. Muñoz-Marí, and J. Calpe-Maravilla, "Retrieval of oceanic chlorophyll concentration with relevance vector machines," *Rem. Sens. Env.*, vol. 105, no. 1, pp. 23–33, 2006.

[23] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. New York: The MIT Press, 2006.

[24] J. Verrelst, J. Muñoz, L. Alonso, J. Delegido, J. Rivera, J. Moreno, and G. Camps-Valls, "Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3," *Rem. Sens. Env.*, vol. 118, no. 0, pp. 127–139, 2012.

[25] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, "Retrieval of vegetation biophysical parameters using Gaussian process techniques," *IEEE Trans. Geosc. Rem. Sens.*, vol. 50, no. 5 PART 2, pp. 1832–1843, 2012.

[26] J. Verrelst, L. Alonso, J. Rivera Caicedo, J. Moreno, and G. Camps-Valls, "Gaussian process retrieval of chlorophyll content from imaging spectroscopy data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 2, pp. 867–874, 2013.

[27] H. Roelofsen, L. Kooistra, P. Van Bodegom, J. Verrelst, J. Krol, and J. c. Witte, "Mapping a priori defined plant associations using remotely sensed vegetation characteristics," *Rem. Sens. Env.*, vol. 140, pp. 639–651, 2014.

[28] M. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 63, no. 3, pp. 425–450, 2001.

[29] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Neural Information Processing Systems, NIPS 8*. MIT Press, 1996, pp. 598–604.

[30] M. Kuss and C. Rasmussen, "Assessing approximate inference for binary Gaussian process classification," *Machine learning research*, vol. 6, pp. 1679–1704, Oct 2005.

[31] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Machine learning research*, vol. 6, pp. 1783–1816, Nov 2005.

[32] A. O'Hagan and J. F. C. Kingman, "Curve fitting and optimal design for prediction," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 40, no. 1, pp. 1–42, 1978.

[33] M. Reed and B. Simon, *I: Functional Analysis, Volume 1 (Methods of Modern Mathematical Physics) (vol 1)*, 1st ed. Academic Press, Jan. 1981.

[34] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, 2006.

[35] P. Sampson and P. Guttorp, "Nonparametric estimation of nonstationary spatial covariance structure," *Journal of the American Statistical Association Publication*, vol. 87, no. 417, pp. 108–119, Mar 1992.

[36] M. Wittmann, H. Breitkreuz, M. Schroedter-Homscheidt, and M. Eck, "Case studies on the use of solar irradiance forecast for optimized operation strategies of solar thermal power plants," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,*, vol. 1, no. 1, pp. 18–27, 2008.

[37] S. A. Kalogirou, "Designing and modeling solar energy systems," *Solar Energy Engineering*, pp. 583–699, 2014.

[38] T. Khatib, A. Mohamed, and K. Sopian, "A review of solar energy modeling techniques," *Renewable and Sustainable Energy Reviews*, pp. 2864–2869, 2012.

[39] E. Geraldi, F. Romano, and E. Ricciardelli, "An advanced model for the estimation of the surface solar irradiance under all atmospheric conditions using MSG / SEVIRI data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 8, pp. 2934–2953, 2012.

[40] E. Lorenz, J. Hurka, D. Heinemann, and H. G. Beyer, "Irradiance forecasting for the power prediction of grid-connected photovoltaic systems," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 2, no. 1, pp. 2–10, 2009.

[41] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[42] M. E. Tipping, "The Relevance Vector Machine," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge, Mass: MIT Press, 2000.

[43] P. Goldberg, C. Williams, and C. Bishop, "Regression with input-dependent noise: A Gaussian process treatment," in *Advances in NIPS*, 1998.

[44] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard, "Most likely heteroscedastic Gaussian processes regression," in *Proc. of the ICML*, 2007, pp. 393–400.

[45] M. Lázaro-Gredilla and M. K. Titsias, "Variational heteroscedastic gaussian process regression," in *28th International Conference on Machine Learning, ICML 2011*. Bellevue, WA, USA: ACM, 2011, pp. 841–848.

[46] E. Snelson, C. Rasmussen, and Z. Ghahramani, "Warped gaussian processes," in *Advances in Neural Information Processing Systems, NIPS*. MIT Press, 2004.

[47] M. Lázaro-Gredilla, "Bayesian warped gaussian processes," in *NIPS*, 2012, pp. 1628–1636.

[48] J. E. O'Reilly, S. Maritorena, B. G. Mitchell, D. A. Siegel, K. Carder, S. A. Garver, M. Kahru, and C. McClain, "Ocean color chlorophyll algorithms for SeaWiFS," *Journal of Geophysical Research*, vol. 103, no. C11, pp. 24 937–24 953, Oct 1998.

[49] S. Maritorena and J. O'Reilly, *OC2v2: Update on the initial operational SeaWiFS chlorophyll algorithm*. NASA Goddard Space Flight Center, Greenbelt, Maryland, USA: John Wiley & Sons, 2000, vol. 11, pp. 3–8.

[50] M. Lázaro-Gredilla, M. K. Titsias, J. Verrelst, and G. Camps-Valls, "Retrieval of biophysical parameters with heteroscedastic gaussian processes," *IEEE Geosc. Rem. Sens. Lett.*, vol. 11, no. 4, pp. 838–842, 2014.

[51] A. G. Wilson, D. A. Knowles, and Z. Ghahramani, "Gaussian process regression networks," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, J. Langford and J. Pineau, Eds. Edinburgh: Omnipress, June 2012.

[52] K. N. Liou, *An Introduction to Atmospheric Radiation*, 2nd ed. Hampton, USA: Academic Press, 2002.

[53] H. L. Huang, W. L. Smith, and H. M. Woolf, "Vertical resolution and accuracy of atmospheric infrared sounding spectrometers," *J. Appl. Meteor.*, vol. 31, pp. 265–274, 1992.

[54] D. Siméoni, C. Singer, and G. Chalon, "Infrared atmospheric sounding interferometer," *Acta Astronautica*, vol. 40, pp. 113–118, 1997.

[55] J. Quinonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate gaussian process regression," *Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[56] E. Snelson and Z. Ghahramani, "Local and global sparse gaussian process approximations," in *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, 2007.

[57] M. Lázaro-Gredilla, J. Q. Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse spectrum gaussian process regression," *Journal of Machine Learning Research*, vol. 11, pp. 1865–1881, 2010.

[58] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Neural Information Processing Systems*, 2007.

[59] A. G. Wilson and H. Nickisch, "Kernel interpolation for scalable structured gaussian processes (KISS-GP)," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015, pp. 1775–1784.

[60] D. Haboudane, J. Miller, E. Pattey, P. Zarco-Tejada, and I. Strachan, "Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture," *Rem. Sens. Env.*, vol. 90, no. 3, 2004.

[61] I. Filella and J. Penuelas, "The red edge position and shape as indicators of plant chlorophyll content, biomass and hydric status," *International Journal of Remote Sensing*, vol. 15, no. 7, pp. 1459–1470, 1994.

[62] S. Stagakis, N. Markos, O. Sykioti, and A. Kyparissis, "Monitoring canopy biophysical and biochemical parameters in ecosystem scale using satellite hyperspectral imagery: An application on a phlomis fruticosa mediterranean ecosystem using multiangular chris/proba observations," *Rem. Sens. Env.*, vol. 114, no. 5, pp. 977–994, 2010.

[63] S. Van Wittenberghe, J. Verrelst, J. Rivera, L. Alonso, J. Moreno, and R. Samson, "Gaussian processes retrieval of leaf parameters from a multi-species reflectance, absorbance and fluorescence dataset," *Journal of Photochemistry and Photobiology B: Biology*, vol. 134, pp. 37–48, 2014.

[64] J. Jagermeyr, D. Gerten, W. Lucht, P. Hostert, M. Migliavacca, and R. Nemani, "A high-resolution approach to estimating ecosystem respiration at continental scales using operational satellite data," *Global Change Biology*, vol. 20, no. 4, pp. 1191–1210, 2014, cited By 2.

[65] GCOS, "Systematic observation requirements for satellite-based products for climate, 2011," p. 138, 2011. [Online]. Available: http://www.wmo.int/pages/prog/gcos/Publications/gcos-154.pdf

[66] J. Verrelst, J. Rivera, J. Moreno, and G. Camps-Valls, "Gaussian processes uncertainty estimates in experimental Sentinel-2 LAI and leaf chlorophyll content retrieval," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 86, pp. 157–167, 2013.

[67] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *NIPS*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2008, pp. 689–696.

[68] J. Mooij, D. Janzing, J. Peters, and B. Schölkopf, "Regression by dependence minimization and its application to causal inference in additive noise models," in *Proceedings of the 26th International Conference on Machine Learning*, A. Danyluk, L. Bottou, and M. Littman, Eds. New York, NY, USA: ACM Press, June 2009, pp. 745–752.

[69] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer, "Estimation of a structural vector autoregression model using non-gaussianity," *JMLR*, vol. 11, no. 5, pp. 1709–1731, 2010.

[70] M. Yamada, M. Sugiyama, and J. Sese, "Least-squares independence regression for non-linear causal inference under non-gaussian noise," *Machine Learning*, vol. 96, no. 3, pp. 249–267, 2014.

[71] A. M. Moffat, C. Beckstein, G. Churkina, M. M. Martin, and M. Heinmann, "Characterization of ecosystem responses to climatic controls using artificial neural networks," *Global Change Biology*, vol. 16, no. 1, pp. 2737–2749, 2010.

[72] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and analysis of computer experiments," *Statistical science*, pp. 409–423, 1989.

[73] T. Quaife, P. Lewis, M. De Kauwe, M. Williams, B. E. Law, M. Disney, and P. Bowyer, "Assimilating canopy reflectance data into an ecosystem model with an ensemble kalman filter," *Remote Sensing of Environment*, vol. 112, no. 4, pp. 1347–1364, 2008.

[74] P. Lewis, J. Gómez-Dans, T. Kaminski, J. Settle, T. Quaife, N. Gobron, J. Styles, and M. Berger, "An earth observation land data assimilation system (eo-ldas)," *Remote Sensing of Environment*, vol. 120, pp. 219–235, 2012.

[75] S. Jacquemoud, W. Verhoef, F. Baret, C. Bacour, P. Zarco-Tejada, G. Asner, C. François, and S. Ustin, "PROSPECT + SAIL models: A review of use for vegetation characterization," *Rem. Sens. Env.*, vol. 113, no. Suppl. 1, pp. S56–S66, 2009.