A Survey on Gaussian Processes for Earth-Observation Data Analysis

A comprehensive investigation

Advances in Machine Learning for Remote Sensing and Geosciences

GUSTAU CAMPS-VALLS, JOCHEM VERRELST, JORDI MUÑOZ-MARÍ, VALERO LAPARRA, FERNANDO MATEO-JIMÉNEZ, AND JOSÉ GÓMEZ-DAN

Gaussian processes (GPs) have experienced tremendous success in biogeophysical parameter retrieval in the last few years. GPs constitute a solid Bayesian framework to consistently formulate many function-approximation problems. This article reviews the main theoretical GP developments in the field, considering new algorithms that respect signal and noise characteristics, extract knowledge via automatic relevance kernels to yield feature rankings automatically, and 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 can encode physically meaningful prior knowledge via radiative transfer model (RTM) emulation. The important issue of computational efficiency will also be addressed. These developments are illustrated in the field of geosciences and remote sensing at local and global scales through a set of

Digital Object Identifier 10.1109/MGRS.2015.2510084
Date of publication: 13 June 2016

illustrative examples. In particular, important problems for land, ocean, and atmosphere monitoring are considered, from accurately estimating oceanic chlorophyll content and pigments to retrieving vegetation properties from multi- and hyperspectral sensors as well as estimating atmospheric parameters (e.g., temperature, moisture, and ozone) from infrared sounders.

UNPRECEDENTED DATA STREAM FOR LAND, OCEAN, AND ATMOSPHERE MONITORING

The spatiotemporally explicit, quantitative retrieval methods for Earth's surface and atmosphere characteristics are required in a variety of Earth system applications. Optical Earth-observing satellites that are endowed with high temporal resolution enable the retrieval and, hence, the monitoring of climate and biogeophysical variables [1], [2]. With the forthcoming superspectral Copernicus Sentinel-2 (S2) [3] and Sentinel-3 missions [4], as well as the planned EnMAP [5], HyspIRI [6], PRISMA [7], and the European Space Agency'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 stream requires enhanced processing techniques that are accurate, robust, and fast. Additionally, the statistical models should capture plausible physical relationships and explain the problem at hand.

A wide variety of biogeophysical retrieval methods have been developed over the last few decades, but only a few of them have made it into operational processing chains, and many are still in their infancy [9]. Essentially, there are two main approaches to the inverse problem of estimating biophysical parameters from spectra: 1) parametric physically based models and 2) nonparametric statistical models. On one hand, parametric, physically based models are commonly used to model biological processes and climate variables in Earth monitoring. These models rely on established physical relationships and implement complex combinations of scientific hypotheses. Unfortunately, they do not exploit empirical data to constrain simulation outcomes; thus, despite their solid physical foundation, they are becoming more obscure because more complex processes, parameterizations, and priors need to be included. These issues give rise to too-rigid solutions and large-model discrepancies (see [10] and the references therein). Alternatively, nonparametric statistical models are typically only concerned with developing data-driven models, paying little attention to the physical rules governing the system. The field has proven to be successful in many disciplines of science and engineering [11], and, in general, nonlinear and nonparametric model instantiations typically lead to a more flexible and improved performance over physically based approximations [12].

In the last decade, machine learning has attained outstanding results in estimating climate variables and the related biogeophysical parameters on local and global scales [13]. For example, the current operational vegetation products, such as leaf area index (LAI), are typically produced with neural networks [14], [15]; gross primary production, 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]; principal component analysis (PCA) and piecewise linear regression have been used for sun-induced fluorescence (SIF) estimation [19]; support vector regression (SVR) showed high efficiency in modeling LAI; fractional vegetation cover (fCOVER), evapotranspiration [20], [21], and relevance vector machines (RVMs) were successful in estimating ocean chlorophyll [22]; and, recently, GPs [23] provided excellent results in estimating vegetation properties [24]–[27].

The family of Bayesian nonparametrics, and of GPs in particular [23], has been given great consideration in remote sensing data analysis in recent years because they are endorsed with important properties that are relevant to common problems in our field. GPs can provide excellent accuracy estimations as well as error bars (i.e., uncertainties) for the predictions. Also, and very importantly, they can easily accommodate different data sources (e.g., multimodal data, multiple sensors, multitemporal acquisitions) and can be designed to deal with different noise sources. The use of GPs in problems involving large data has traditionally been problematic, but recently advanced sparse, variational, and distributed computing techniques allow training models in almost linear cost. This article studies the modern approaches to tackle these issues.

Beyond these interesting features of GPs, statistical inference methods should be able to fit data well (i.e., focus only on data exploitation) but should also show something about the physical rules governing the problem (i.e., data exploration). Therefore, these too-flexible models should be constrained to provide physically plausible predictions, which is why, in recent years, combining machine learning and physical models seems promising, either via data assimilation, hybrid approaches, or the emulation of physically based RTMs. In this respect, GPs can be used to learn about the relevance of the problem features, as they can adapt to anisotropic data distributions, the derivatives of the predictive mean and variance can be computed in closed form, and they are ideal for use in empirical (i.e., noninterventional) causal inference. Additionally, GPs have been the first choice in emulating RTMs to endorse these statistical models with physically meaningful constraints [28].

GAUSSIAN PROCESS REGRESSION

Regression, function approximation, and function emulation are old, largely studied problems in statistics and machine learning. The problem boils down to optimizing a loss (e.g., cost or energy) function over a class of functions. In particular, a large class of regression problems 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 (i.e., excess of flexibility).

GAUSSIAN PROCESSES: A GENTLE INTRODUCTION

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 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 21st century [23]. GPs can be understood as a family of kernel methods with the additional advantage of providing a full conditional, statistical description for the predicted variable that can primarily be used to establish confidence intervals and set hyperparameters. In a nutshell, GPs assume that a GP prior governs the 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 GP, and the predicted distribution is estimated by conditioning on the training data.

This article focuses on the recent success of GPs in dealing with regression problems in biophysical parameter retrieval and the generic model inversion in geosciences. Standard regression approximates observations (which are 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 to fit observed data, GP regression (GPR) proceeds in a Bayesian, nonparametric way. It is customary to subtract the sample mean to data $\{y_n\}_{n=1}^N$ and then to assume a zero mean model. A zero mean GP prior is placed on the latent function $f(\mathbf{x})$ and a Gaussian prior is used for each latent noise term ε_n, $f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k_\theta(\mathbf{x}, \mathbf{x}'))$, where $k_\theta(\mathbf{x}, \mathbf{x}')$ is a covariance function parameterized by θ 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 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 $\mathbf{K}_{\mathbf{ff}}$ with $[\mathbf{K}_{\mathbf{ff}}]_{ij} = k_\theta(\mathbf{x}_i, \mathbf{x}_j)$.

If considering 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 $\mathcal{D} \equiv \{\mathbf{x}_n, y_n \mid n = 1, \ldots N\}$, 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 $\mathcal{D}$,

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

which is a Gaussian with the following mean and variance:

$$\mu_{\mathrm{GP}*} = \mathbf{k}_{\mathbf{f}*}^\top \left(\mathbf{K}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_N\right)^{-1} \mathbf{y} \tag{3}$$

$$\sigma^2_{\mathrm{GP}*} = \sigma^2 + k_{**} - \mathbf{K}_{\mathbf{f}*}^\top \left(\mathbf{K}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_N\right)^{-1} \mathbf{k}_{\mathbf{f}*}, \tag{4}$$

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 $\mathcal{D}$, $\mathbf{K}_{\mathbf{ff}}$ is an $N \times N$ kernel (covariance) matrix whose entries contain the similarities between all training points, $\mathbf{y} = [y_1, \ldots, y_N]^\top \in \mathbb{R}^{N \times 1}$, 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. It is important to note that both the predictive mean and the variance can be computed in closed form, and the predictive variance $\sigma^2_{\mathrm{GP}*}$ does not depend on the outputs/target variable. Also note that the predictive mean is computable in $\mathcal{O}(N^3)$ time since it involves the inversion of the $N \times N$ matrix $(\mathbf{K}_{\mathbf{ff}} + \sigma^2 \mathbf{I})$ [23]. In addition to the computational cost, GPs require large memory because, in naive implementations, one has to store the training kernel matrix, which amounts to $\mathcal{O}(N^2)$. Recent improvements in efficiency will be reviewed in the "Efficiency in Gaussian Process Regression" section.

TABLE 1. 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	$k(\mathbf{x}, \mathbf{x}') = \exp(-\Vert \mathbf{x} - \mathbf{x}' \Vert^2 / (2\sigma^2))$
Exponential	$k(\mathbf{x}, \mathbf{x}') = \exp(-\Vert \mathbf{x} - \mathbf{x}' \Vert / (2\sigma^2))$
Rational quadratic	$k(\mathbf{x}, \mathbf{x}') = 1 - (\Vert \mathbf{x} - \mathbf{x}' \Vert^2) / (\Vert \mathbf{x} - \mathbf{x}' \Vert^2 + c)$
Multiquadric	$k(\mathbf{x}, \mathbf{x}') = \sqrt{\Vert \mathbf{x} - \mathbf{x}' \Vert^2 + c^2}$
Inverse multiquadratic	$k(\mathbf{x}', \mathbf{x}') = 1 / (\sqrt{\Vert \mathbf{x} - \mathbf{x}' \Vert^2 + \theta^2})$
Power	$k(\mathbf{x}, \mathbf{x}') = -\Vert \mathbf{x} - \mathbf{x}' \Vert^d$
Log	$k(\mathbf{x}, \mathbf{x}') = -\log(\Vert \mathbf{x} - \mathbf{x}' \Vert^d + 1)$

ON MODEL SELECTION

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

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

When the derivatives of (5) are also analytical, which is often the case, conjugated gradient ascent is typically used for optimization. Therefore, the entire procedure of learning a GP model only depends on a very small set of hyperparameters that efficiently combats overfitting. Finally, inference of the hyperparameters and the weights for doing predictions, α, can be performed by continuous evidence optimization.

ON THE COVARIANCE FUNCTION

In general, the core of any kernel method, 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 (SE),

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

which captures the sample similarity well in most of the unstructured problems, and only one hyperparameter σ needs to be tuned. Table 1 summarizes the most common kernel functions in standard applications with kernel methods.

In the context of GPs, kernels with more hyperparameters can be efficiently inferred, which is an opportunity to exploit asymmetries in the feature space by including a parameter per feature, as in the very common anisotropic 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 σ_f is the length scale per input features, $f = 1, \ldots, F$. This is a very flexible covariance function that typically suffices to tackle most of the problems. However, it is important to note that an SE can typically approximate smoothly varying functions, which may not be the case in some particular problems. Also, when the data is structured, i.e., when it reveals a particular structure (e.g., time or spatial), the covariance design is of paramount relevance, and many approaches have exploited the standard properties of functional analysis to do so [33].

GAUSSIAN PROCESSES EXEMPLIFIED

Figure 1 presents an illustrative example with six training points that range between −2 and +2. Depicted are several random functions drawn from the GP prior and functions drawn from the posterior. An isotropic Gaussian kernel and $\sigma_\nu = 0.1$ was chosen. The mean function has been plotted, plus or minus two standard deviations, corresponding to a 95% confidence interval. Typically, the hyperparameters are unknown, as are the mean, covariance, and likelihood functions. An SE covariance function was assumed and the optimal hyperparameters were learned by minimizing the negative log marginal likelihood (NLML) with respect to the hyperparameters. There are no samples below x = −1.5; hence, the GPR simply provides the solution given by the prior (zero mean and ±2). At the center, where most of the data points lie, there is a very accurate view of the latent function with small error bars (i.e., close to $\pm 2\sigma_\nu$). The same behavior is observed since training samples for x > 0 are not available. GPs typically provide an accurate solution where the data exists and high error bars where information is not available; consequently, it is presumed that the prediction in that area is inaccurate. For this reason, in regions of the input space without points, the confidence intervals are wide, resembling the prior distribution.

SOURCE CODE AND TOOLBOXES

The most widely known websites to obtain free source code on GP modeling are GPML, http://www.gaussianprocess.org/, and GPstuff, http://becs.aalto.fi/en/research/bayes/gpstuff/. The GPML website centralizes the main activities in GP modeling and provides up-to-date resources regarding probabilistic modeling, inference, and learning based on GPs, while the GPstuff website 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 readers interested in learning the main aspects of GP modeling, as they provide free code, demonstrations, and recommendations of relevant tutorials and books. For readers interested in

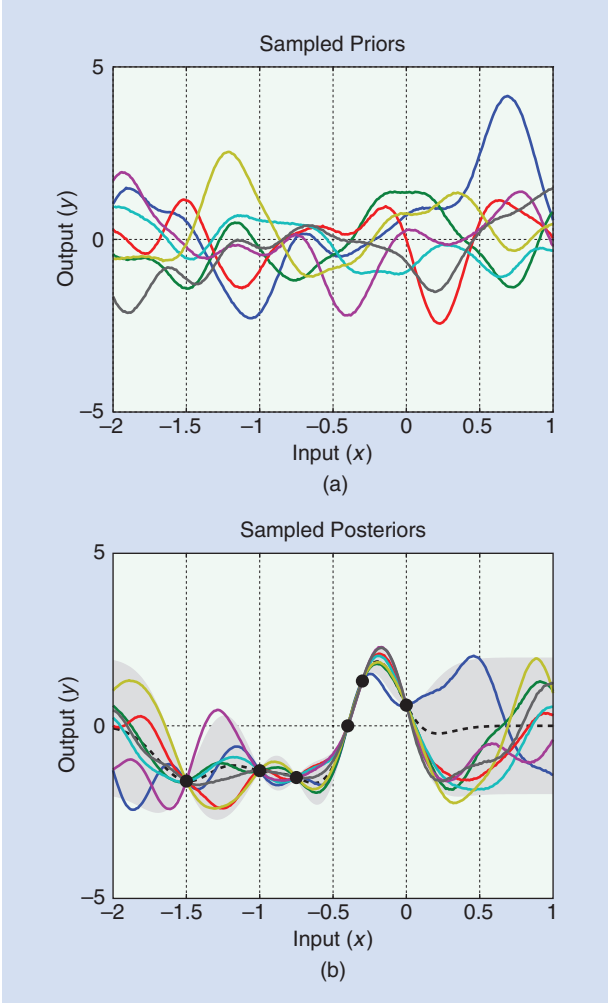

FIGURE 1. An example of a GP. (a) Some functions drawn at random from the GP prior. (b) Some random functions drawn from the posterior, i.e., the prior conditioned on six noise-free observations indicated by the black dots. The shaded area represents the point-wise mean plus and minus two times the standard deviation for each input value (corresponding to the 95% confidence region). The confidence intervals become larger in regions that are farther from the observations. (This is an animated figure that only works when viewing in Adobe Acrobat.)

regression, in general, the MATLAB SimpleR toolbox is recommended, http://www.uv.es/gcamps/software.html, which contains several regression tools that are organized into families (i.e., tree-based, bagging and boosting, neural nets, kernel regression methods, and several Bayesian nonparametric models such as GPs). The toolbox is intended for practitioners with little expertise in machine learning who may want to easily assess advanced methods in their problems.

ADVANCES IN GAUSSIAN PROCESS REGRESSION

This section reviews some of the recent advances in GPR that are especially suited to remote sensing data analysis. Also discussed are the main aspects of design covariance functions that capture nonstationarities and multiscale time relations, as well as GPs that can learn arbitrary transformations of the observed variable and noise models. The multitask and multioutput problems are also discussed.

STRUCTURED, NONSTATIONARY, AND MULTISCALE GAUSSIAN PROCESS REGRESSION

Commonly used kernel families include the SE, periodic (Per), linear (Lin), and rational quadratic (RQ) (see Table 1). Figure 2 shows the base kernel illustrations and drawings from the GP prior. These base kernels can be combined by following simple operations (i.e., summation, multiplication, or convolution) so that one may build sophisticated covariances from simpler ones. It is important to note that the same essential property of kernel methods applies here; therefore, a valid covariance function must be positive semidefinite. In general, the kernel design should rely on the information available for each estimation problem and should strive for the most accurate solution with the fewest number of samples.

In Figure 2, all of the base kernels are one-dimensional, but kernels over multidimensional inputs can be constructed by adding and multiplying kernels over individual dimensions. By summing kernels, the data can be modeled as a superposition of independent functions, possibly representing different structures. For example, in multitemporal image analysis, one could dedicate one kernel for the time domain (perhaps trying to capture trends and seasonal effects) and a kernel function for the spatial domain (equivalently capturing spatial patterns and autocorrelations). In time-series models, the sums of kernels can express the superposition of different processes that are possibly operating at different scales; changes in geophysical variables through time often occur at different temporal resolutions (e.g., hours or days), and this can be incorporated into 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. The following section will explain how to design kernels that incorporate particular time resolutions, trends, and periodicities.

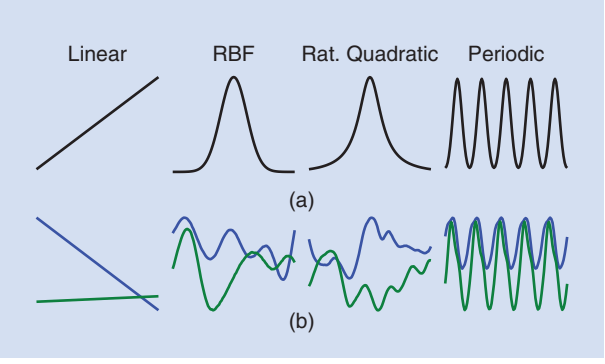

FIGURE 2. (a) The base kernels and two random draws from a GP with each (b) respective kernel. See Table 1 for the explicit functional form of each kernel.

GAUSSIAN PROCESS REGRESSION TIME-BASED COVARIANCE

As previously stated, time is an additional and important variable to consider in many remote sensing applications. Signals to be processed typically show particular characteristics with time-dependent cycles and trends. One could, of course, include time, t_i, as an additional feature in the input sample definition. This stacked approach [34] essentially relies on a covariance function $k(\mathbf{z}_i, \mathbf{z}_j)$, where $\mathbf{z}_i = [t_i, \mathbf{x}_i]^\top$, which is convenient because it does not require learning additional hyperparameters. However, the shortcoming is that the time relationships are naively left to the nonlinear regression algorithm, and, hence, no explicit time-structure model is assumed. To more consistently cope with such temporal behavior of the observed signal, one can use a linear combination (or composite) of different kernels, i.e., one dedicated to capturing the different temporal characteristics and the other to the feature-based relationships. A simple strategy that is quite common in statistics and signal processing is to rely on a tensor kernel, as in

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

but more sophisticated structures can be adopted. The issue here is how to design kernels that are capable of dealing with nonstationary processes.

One 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 the spatial patterns of solar radiation with GPR. It is also possible to adopt an SE as a 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],

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

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 hyperparameter characterizing the periodic scale and needs to be inferred. However, it is not clear that the seasonal trend is exactly periodic, so this equation is modified by taking the product with an SE 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 is 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 $\boldsymbol{\theta} = \{\nu, \sigma_1, \ldots, \sigma_F, \sigma_n, \sigma_t, \sigma_d, \gamma\}$.

The advantage of encoding this prior knowledge and structure in the relevant problem of solar irradiation prediction is shown, which is an important and challenging problem with direct applications in renewable energy. Solar is one of the most important green sources of energy that is currently expanding in many countries, especially in those with more solar potential such as middle eastern and southern European countries [36], [37]. Accurately estimating energy production in solar energy systems involves correctly predicting 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. GPR is evaluated to estimate solar irradiation. Noting the nonstationary temporal behavior of the signal, a particular, time-based composite covariance is developed to account for relevant, seasonal signal variations. A unique meteorological data set is used that was acquired at a radiometric station that includes 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 were considered. Brewer and Cimel networks, as well as the pyranometer used, are managed under a quality management system certified to ISO 9001:2008. These data range from 1 January 2010 to 31 December 2011. Data with missing values was removed, resulting with the final data set containing 512 examples and ten input features (see Table 2).

TABLE 2. THE VARIABLES AND SOURCES CONSIDERED IN THE 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 water precipitation	mm	1.33–41.53
Global forecast system	Cloud amount	%	2–79.2
Pyranometer	Measured global solar irradiation	kJ/m^2	4.38–31.15

TABLE 3. THE ESTIMATIONS OF THE DAILY SOLAR IRRADIATION OF LINEAR AND NONLINEAR MODELS.

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**

Table 3 shows the results obtained with GPR models and several statistical regression methods, i.e., regularized linear regression (RLR), SVR, RVM, and GPR. All of the methods were run with and without using two additional dummy-time features containing the year and day of year (DOY). The former case will be indicated with a subscript (e.g., SVR$_t$). Including the time information can improve all of the baseline models. Additionally, the best overall results are obtained by the GPR models, whether they include time information or not. Also, the proposed TGPR particularly outperforms the rest in accuracy [i.e., the root-mean-squared error (RMSE) and mean absolute error (MAE)] and goodness-of-fit (R), and it closely follows the elastic net in bias (mean error, or ME). TGPR performs better than GPR and GPRt in all quality measures.

HETEROSCEDASTIC GAUSSIAN PROCESS REGRESSION: LEARNING THE NOISE MODEL

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

$$\varepsilon_n \sim \mathcal{N}(0, e^{g(x_n)})$$

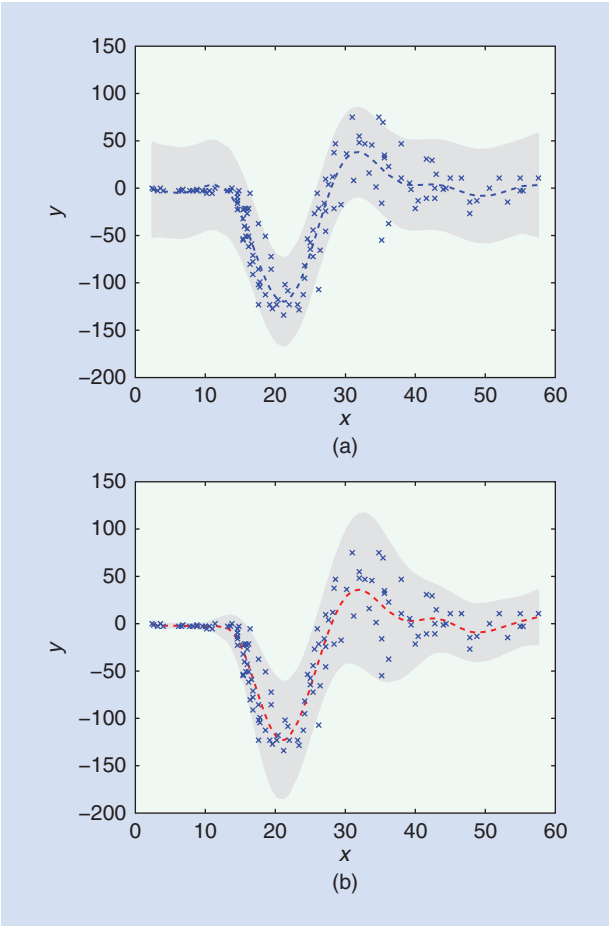

FIGURE 3. The predictive mean and variance of (a) the standard GP and (b) the heteroscedastic GP. Notice that, in the low-noise regime, the VHGP produces tighter confidence intervals as expected, while high noise variance associated with high signal variance (i.e., the middle of the observed signal) provides a more reasonable predictive variance, too.

and placing a GP prior over $g(\mathbf{x}) \sim \mathcal{GP}(\mu_0 \mathbf{1}, k_{\theta_g}(\mathbf{x}, \mathbf{x}'))$. It is important to note that the exponential is needed to describe the nonnegative variance. Of course, other transformations are possible, just not as convenient. 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 γ_*.

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

To overcome the aforementioned problems, the sophisticated marginalized variational (MV) approximation was introduced in [45], which 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 to the proposal distribution and the hyperparameters 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 γ_* (i.e., predictions) can be computed in closed form. A simple comparison between the homoscedastic canonical GP and the variational approximation for heteroscedastic GP regression (VHGPR) model is shown in Figure 3.

WARPED GAUSSIAN PROCESS REGRESSION: LEARNING THE OUTPUT TRANSFORMATION

In practical applications, the observed variable is often transformed to better pose the problem. Actually, it is standard practice to linearize or uniformize observation distribution, which is commonly skewed due to the sampling strategies in in-situ data collection, by applying nonlinear link functions such as logarithmic, exponential, or logistic functions.

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

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

where f is a possibly noisy latent function with d inputs and g is a function with scalar inputs parameterized 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]. Replacing γ_i with z_i in the standard GP model leads to a further 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, the covariance (i.e., kernel or gram) function $k(\cdot,\cdot)$ needs to be defined, which should capture the similarity between samples. The standard automatic relevance determination (ARD) covariance was used [23], and the model hyperparameters are collectively grouped in $\boldsymbol{\theta} = \{\nu, \sigma_n, \sigma_1, \ldots, \sigma_d\}$. In addition, for the WGPR, a parametric smooth and monotonic form needs to be defined for g, which can be defined as

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

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

For illustration purposes, the focus is on the estimation of imagesic chlorophyll-a concentrations from remote sensing upwelling radiance that is just above the image's surface. A variety of bio-optical algorithms have been developed to relate the measurements of image radiance to in-situ concentrations of phytoplankton pigments, and, ultimately, most of these algorithms demonstrate the potential to quantify chlorophyll-a concentrations accurately from multispectral-satellite-image color data. In this context, robust and stable nonlinear regression methods that provide inverse models are desirable. In addition, most of the bio-optical models (e.g., 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, the SeaBAM data set was used [48], [49], which gathers 919 in-situ pigment measurements around the United States and Europe. The data set contains coincident in-situ chlorophyll concentration and remote sensing reflectance measurements (Rrs(λ), [sr^{-1}]) at some wavelengths (i.e., 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. Although 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-to-log space. At their lowest concentrations, the highest Rrs(490)/Rrs(555) ratios are slightly lower than the theoretical limit for clear, natural waters (see the analysis in [22]).

Table 4 shows different scores, i.e., the bias (ME), accuracy (RMSE), MAE, and goodness of fit (Pearson's correlation R), between the observed and predicted variables when using the raw data (i.e., with no ad-hoc transform) and the empirically adjusted transform. The results are shown for three kinds of GPs: standard GPR [23], VHGPR [50], and the proposed WGPR [46], [47] for different rates of training samples. Empirically based warping slightly improves the results of working with raw data for the same number of training samples, but this requires prior knowledge of the problem, time, and effort to fit an appropriate function. On the other hand, WGPR outperforms the other GPs in all comparisons over standard GPR and VHGPR (~ +1−10%). Finally, WGPR nicely compensates for the lack of prior knowledge of the (possibly skewed) observation variable distribution.

MULTITASK AND MULTIOUTPUT GAUSSIAN PROCESS MODELS

Very often problems are dealt with that involve several variables that must be estimated. Individual models are typically trained separately, which ignores the potential cross-relations among output variables (e.g., between LAI, chlorophyll content, and fractional cover). Some multitask and multioutput GP models are available to account for this in the output. A simple, multioutput GP can model 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; see http://gaussianprocess.com/publications/multiple_output.php. An effective model based on GPs for the multitask problem, the GPR networks (GPRNs) [51], combines the properties of Bayesian neural networks with the nonparametric flexibility of GPs.

All these approaches, however, suffer when output dimensionality is very high. The following sections will show a much simpler approach to dealing with this problem, particularly focusing on estimating water vapor profiles, which are an 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

TABLE 4. THE 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

vertical resolution [53]. Focus is placed on the data coming from the Infrared Atmospheric Sounding Interferometer (IASI), which provides radiances in 8,461 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, noting the high vertical correlation of the profiles, a simpler strategy is developing a unique GP model that simultaneously predicts all of the PCA-projected state vectors onto the top p principal components, and solving

$$\mathbf{\Lambda} = (\mathbf{K}_{\mathrm{ff}} + \sigma^2 \mathbf{I}_N)^{-1}\mathbf{Y},$$

where $\mathbf{Y}$ columns contain the p scores (i.e., the projected variables). This approach will be exploited again for RTM emulation, as described in the "Emulating Radiative Transfer Models Through Gaussian Processes" section.

Figure 4 shows results of applying this strategy using IASI data to predict (multioutput, 137 dimensions) dew point temperature profiles. A linear regression (LR) and a GP model were trained using the first 100 principal components of an IASI orbit (2008-07-17), both using 5,000 samples and tested in several unseen data. Essentially, it was observed that GPs largely improve the LR models, with an average gain of +1.5 K, which is also statistically significant in all regions.

EFFICIENCY IN GAUSSIAN PROCESS REGRESSION

The naive implementation of GPs in (3) and (4) grows as $O(N^3)$, where N is the number of training samples, which makes them unfeasible when a large number of training samples are available. To reduce the GPs' computation complexity, they are generally computed using approximations. (Other forms of efficiency that involve parallelization and hardware-specific approaches and focus on pure GP algorithms are intentionally omitted here.) The approximation methods can be broadly classified as sparse, localized regression, and matrix multiplication. Finally, some recent developments are highlighted in GP efficiency that exploit random features and particular kernel structures.

SPARSE METHODS

Sparse 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 well-represent the data set (e.g., via clustering or smart sampling). Since the selected M samples represent all others, these methods are considered global, as opposed to the local methods described in the next section. These global methods are well suited to model smooth-varying functions with high correlations (i.e., long length scales), and they use all the predictions data, such as full GPs. The methods in this family are based on substituting the joint prior with a reduced

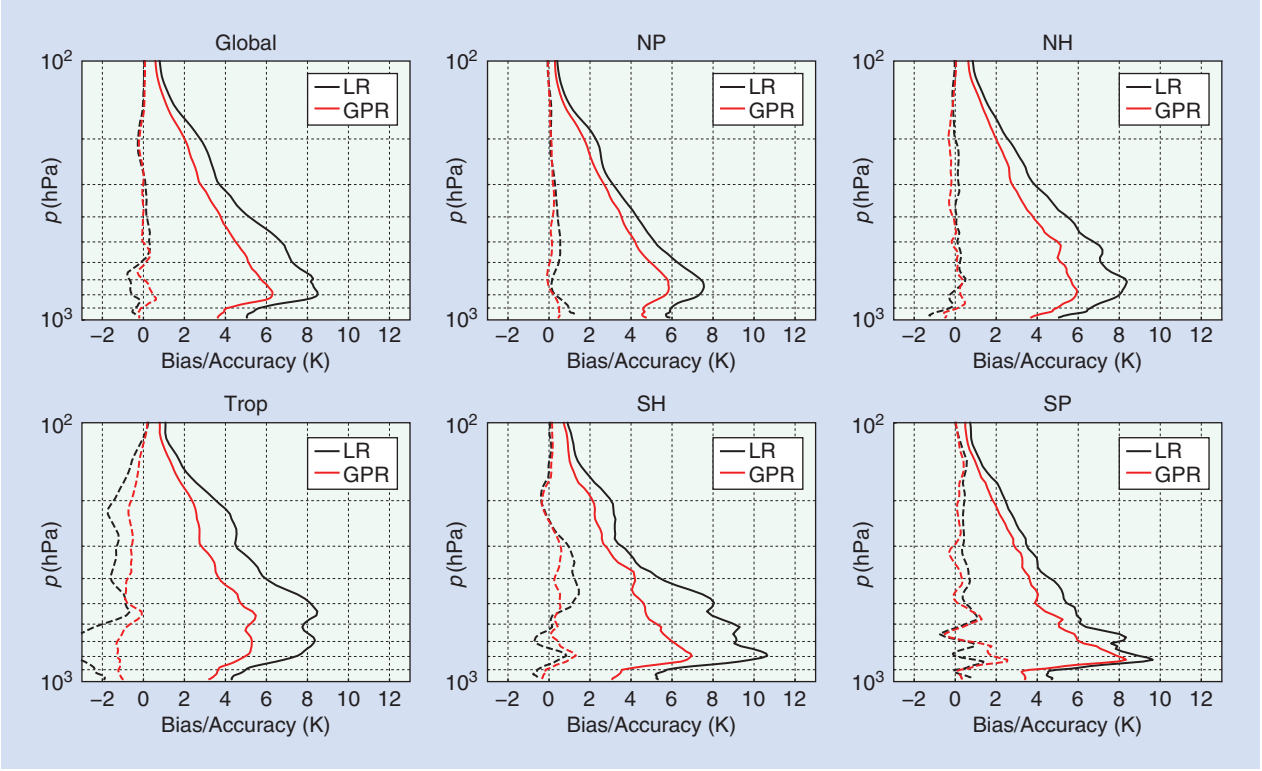

FIGURE 4. The ME (thin dashed lines) and RMSE (solid lines) throughout the atmospheric column for a linear regression and a GP model predicting dew point temperature profiles. The results are averaged for the whole globe and considered orbits, as well as for different regions (i.e., north/south poles, north/south hemispheres, and tropics).

one by 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 GP that correspond to a set of input locations called *inducing inputs*. By adopting a subsets-of-data approach, the computational complexity drastically reduces to $O(M^3)$, being $M = N$.

Some examples of these approximation methods are the subsets of regressors (SoRs), deterministic training conditional (DTC), fully independent training conditional (FICT), partially independent training conditional (PICT) [55], and partially independent conditional (PIC) [56]. All of these methods, with some exceptions for PIC, are based on replacing the joint prior of training and test samples by an approximation, assuming that they are conditionally independent given the set of M-latent-inducing variables. The exact prior is substituted with approximations based on the latent variables, which effectively lowers the ranks of the covariance matrices. On the other hand, these approaches use the exact likelihood. Table 5 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 has unrealistic predictive variances 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 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 in this direction comes from PITC [55] that uses a block diagonal matrix instead of using a diagonal matrix, thus preserving more exact values. Finally, the PIC [56] improves the PITC by relaxing the conditional independence condition between the training and test samples, treating them equally according only to their location, which allows one to efficiently exploit global and local information.

LOCALIZED REGRESSION METHODS

All of the previously described methods are based on defining a set of inducing variables of size $M \ll N$ that represent all N points, which is why these methods are classified as *global* methods as they are well suited to model smoothly varying function with high correlations. However, if M is too small, representation of the whole set is poor and the associated GP's performance is low. On the other hand, the so-called *local* methods are best suited to model highly varying functions with low correlations, but they only use local data for predictions. Local GPs are obtained by dividing the region of interest and training a GP in each division, which has two main advantages: 1) each local GP performs well in the small region on which it has been trained and 2) 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 $O(N^3)$ to $O(NM^2)$. The main disadvantages are that they show discontinuities at the limits between local GPs, and they perform poorly when predicting in regions far from their locality, which poses a problem when the training data is only available in parts of the input region.

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

MATRIX VECTOR MULTIPLICATION APPROXIMATION METHODS

Matrix vector multiplication (MVM) approximation methods are based on speeding up the process of solving the linear system $(\mathbf{K} + \sigma^2\mathbf{I})\boldsymbol{\alpha} = \mathbf{y}$ using an iterative method, such as the

TABLE 5. THE PREDICTIVE DISTRIBUTIONS FOR THE LOW-RANK APPROXIMATION METHODS DESCRIBED IN THE "EFFICIENCY IN GAUSSIAN PROCESS REGRESSION" SECTION.

METHOD	PREDICTIVE MEAN, μ_*	PREDICTIVE VARIANCE, σ_*	TRAINING	TEST MEAN	TEST VARIANCE
SoR	$Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I})^{-1}\mathbf{y}$	$Q_{*,*} - Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I})^{-1}Q_{\mathbf{f},*}$	O (NM^2)	O (M)	O (M^2)
DTC	$Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I})^{-1}\mathbf{y}$	$K_{*,*} - Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I})^{-1}Q_{\mathbf{f},*}$	O (NM^2)	O (M)	O (M^2)
FITC	$Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \Lambda)^{-1}\mathbf{y}$	$K_{*,*} - Q_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \Lambda)^{-1}Q_{\mathbf{f},*}$	O (NM^2)	O (M)	O (M^2)
PITC	As FITC, but $\Lambda \equiv$ blkdiag $[K_{\mathbf{f},\mathbf{f}} - Q_{\mathbf{f},\mathbf{f}} + \sigma^2\mathbf{I}]$.		O (NM^2)	O (M)	O (M^2)
PIC	$K^{\text{PIC}}_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \Lambda)^{-1}\mathbf{y}$	$K_{*,*} - K^{\text{PIC}}_{*,\mathbf{f}}(Q_{\mathbf{f},\mathbf{f}} + \Lambda)^{-1}Q_{\mathbf{f},*}$	O (NM^2)	O ($M + B$)	O ($M + B$)

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

conjugate gradient (CG). Each iteration of the CG method requires an 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 $O(BN^2)$, being $B < N$ the number of CG iterations. To further speed up the computation, as $O(BN^2)$ is still too slow for large problems, MVM must be accelerated. In CG, step one needs 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.

RECENT ADVANCES

There has been a huge improvement in GP runtime and memory demands in the recent years. Inducing methods has become popular but may lack the expressive power of the kernel. A useful approach is the sparse spectrum GP [57], which is somewhat related to random kitchen sinks in [58] that allow an approximation of a kernel matrix with a set of random bases sampled from the Fourier domain. On the other hand, some methods try to exploit structure in the kernel, either based on Kronecker or Toeplitz methods. The limitations of these methods in dealing with data in a grid have recently been remedied with the kernel interpolation for scalable structured GP [59], which generalizes inducing-point methods for scalable GPs and scales $O(N)$ in time and storage for GP inference.

ANALYSIS OF GAUSSIAN PROCESS MODELS

One possibility in using GP models is to extract knowledge from the trained model, for which there are three different approaches: 1) feature ranking that exploits the ARD covariance, 2) uncertainty estimation looking at predictive variance estimates, and 3) the exploitation of the GP models to infer causal relations between biophysical variables under a fully empirical, noninterventional setting. The next section will discuss the use of GP models to mimic RTMs as a way to encode physical knowledge in the statistical models.

RANKING FEATURES THROUGH THE AUTOMATIC RELEVANCE DETERMINATION COVARIANCE

One of the advantages of GPs is that, during GP-model development, the predictive power of each band is evaluated for the parameter of interest by calculating the ARD. Specifically, band ranking through σ_b may reveal the bands that contribute most to GP-model development. An example of the σ_b's for one GP model trained with field leaf chlorophyll content (*Chl*) data and with 62 compact high resolution imaging spectrometer (CHRIS) bands is shown in Figure 5(a). The band with the highest σ_b contributes the least to the model. Relatively few bands (i.e., approximately eight) were evaluated as crucial for *Chl* estimation, while the majority of bands were observed as contributing less. 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], using the same CHRIS data set, it was demonstrated that accuracies remained constant when iteratively removing the least contributing band. Only when fewer than four bands were left did accuracies start to rapidly degrade, as seen in Figure 5(b).

Hence, all CHRIS bands can be used without running the risk of losing accuracy. Of more interest here is identifying where the most relevant bands are located. Essentially, Figure 5 suggests that the most relevant spectral region is between 550 and 1,000 nm, meaning 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 because, when there is a large number of bands available,

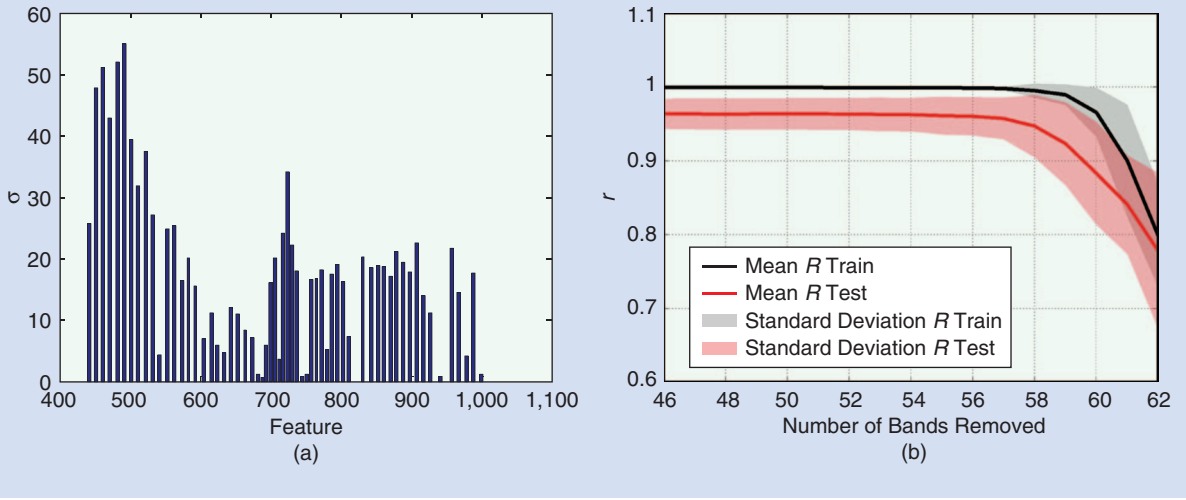

FIGURE 5. (a) The estimated σ_b values for one GP model using 62 CHRIS bands. The lower the σ_b, the more important the band is for regression. (b) The mean and standard deviation of the correlation coefficient r for training and validation for GP fittings using backward elimination of worst σ_b.

neighboring bands do not provide much additional information and can thus be considered redundant. Remarkably, a few relevant bands fell within the 950–1,000 nm region, which is outside the *Chl* absorption region. One reason these bands were seen as important is that, at the canopy scale, the measured reflectance is not only related to biochemistry but is 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 the vegetation structure and water content [60]. Consequently, the *Chl* sensitivity in the NIR may be driven by secondary relationships, as also observed in [61] and [62].

Consequently, the σ_b proved to be a valuable tool to detect most of the sensitive bands of a sensor toward a biophysical parameter. A more systematic analysis was applied by sorting the bands according to relevance and counting the band rankings over 50 repetitions. In [24], the four most relevant bands were tracked for *Chl*, LAI, fCOVER, and different S2 settings, demonstrating the potential of S2 with its new band in the red edge for estimating vegetation properties. Also in [12], σ_b was used to analyze the band sensitivity of S2 toward LAI, and a similar approach was pursued when analyzing leaf *Chl* by tracking the most sensitive spectral regions of SIF data [63], as displayed in Figure 6.

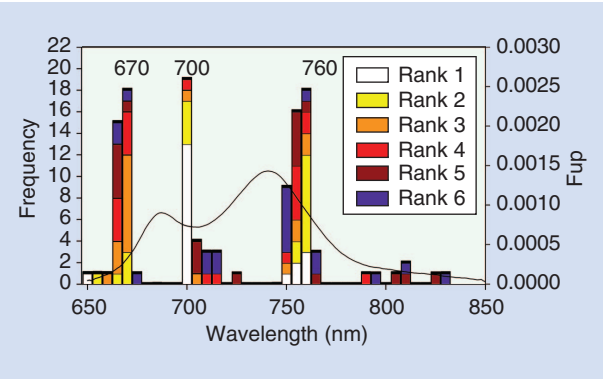

FIGURE 6. The frequency plots of the top-eight-ranked bands with the lowest σ_b values in 20 runs of the GPR prediction of *Chl*, based on upward fluorescence (*Fup*) emission. An emission curve is given as illustrated.

UNCERTAINTY INTERVALS

In this section, GP models for retrieval and portability in space and time are used. For this, the associated predictive variance (i.e. uncertainty interval) provided by GP models is exploited. Consequently, retrievals with high uncertainties refer to pixel spectral information that deviates from what has been represented during the training phase. In turn, low uncertainties can 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].

Applying GPs to estimate biophysical parameters was initially demonstrated in [25]. A locally collected field data set, called *scalable processor architecture* (*SPARC*) *2003*, in Barrax, Spain, was used to train and validate 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 Figure 7. Although generated maps can provide spatially explicit information about vegetation status, the associated uncertainty maps can be 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, and 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 shows that the input spectrum deviates from what was presented during the training stage, thereby imposing retrieval uncertainties. Hence, a practical implication of uncertainty maps is that they detect areas that may benefit from a denser sampling regime.

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

GP models were subsequently applied to the SPARC data set that was resampled to different S2 band settings (i.e., four, eight, and ten bands) and then the uncertainties were inspected [24]. On the whole, adding spectral information led to reduced uncertainties and, thus, more meaningful biophysical parameter maps. Nevertheless, it remains to be seen how robust the locally trained GP models function when applied to other sites and conditions. In this respect, uncertainty estimates may enable the portability of the regression model to be evaluated. 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, it can be reasonably assumed that the retrievals are of the same quality. Thus, when successfully validated over a reference imagery, the uncertainty estimates can work as a quality indicator. Note, however, that the previous conclusions should be taken with caution, as the GP-provided predictive variance is only an estimate of the actual uncertainty.

Accordingly, locally trained GP models were applied to simulated S2 images in a follow-up study [66]. A time series

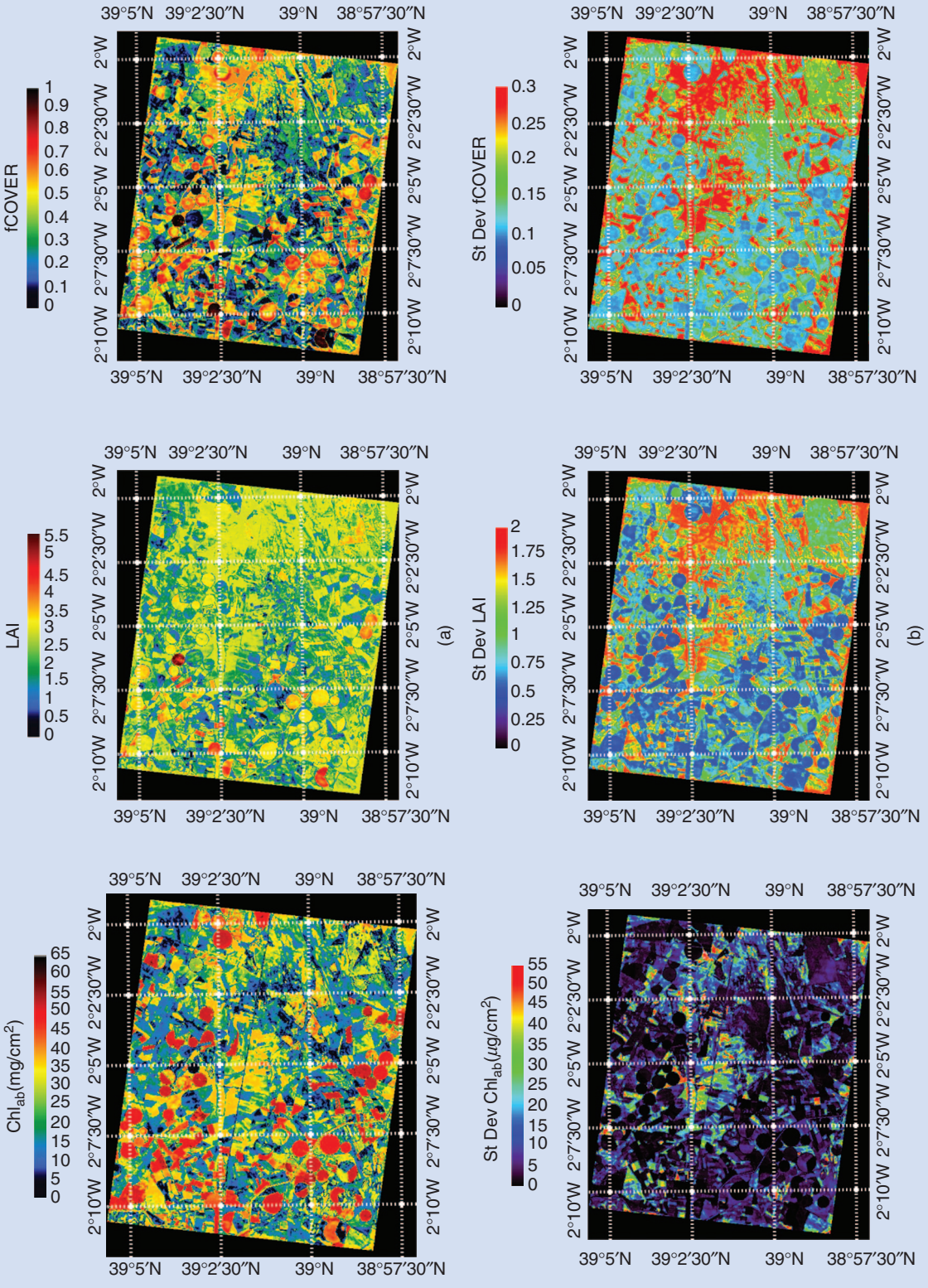

FIGURE 7. (a) The prediction maps and (b) the associated uncertainty intervals generated with GP and four bands of the CHRIS 12 July 2003 nadir image.

over the local Barrax site, as well images across the world, were processed; the role of an extended training set (TrEx, adding spectra of nonvegetated surfaces) was evaluated; and uncertainty values were analyzed. Using TrEx not only further improved performance but also allowed for a decrease in theoretical uncertainties, which underlines the importance of a broad and diverse training data set. More importantly, the GP models were successfully applied to simulated S2 images covering various sites, and 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, a 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 uncertainty ranges were observed within an image rather than between images.

As a final example, uncertainty estimates were exploited to assess the robustness of the retrievals at multiple spatial scales. In [26], the retrievals from hyperspectral airborne and spaceborne data over the Barrax area were compared. The GP developed a model that was excellently validated (R^2: 0.96) based on the spaceborne SPARC-2003 data set, and the SPARC-trained GP model was subsequently applied to airborne CASI flight lines (Barrax, 2009) to generate *Chl* maps. The accompanying uncertainty maps provided insight as to the robustness of the retrievals, and, in general, similar uncertainties were achieved by both sensors, which is encouraging in terms of 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, which was most likely 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 variation in land-cover types is being observed than at the spaceborne scale of CHRIS. Some examples of MEs and associated uncertainties are shown in Figure 8.

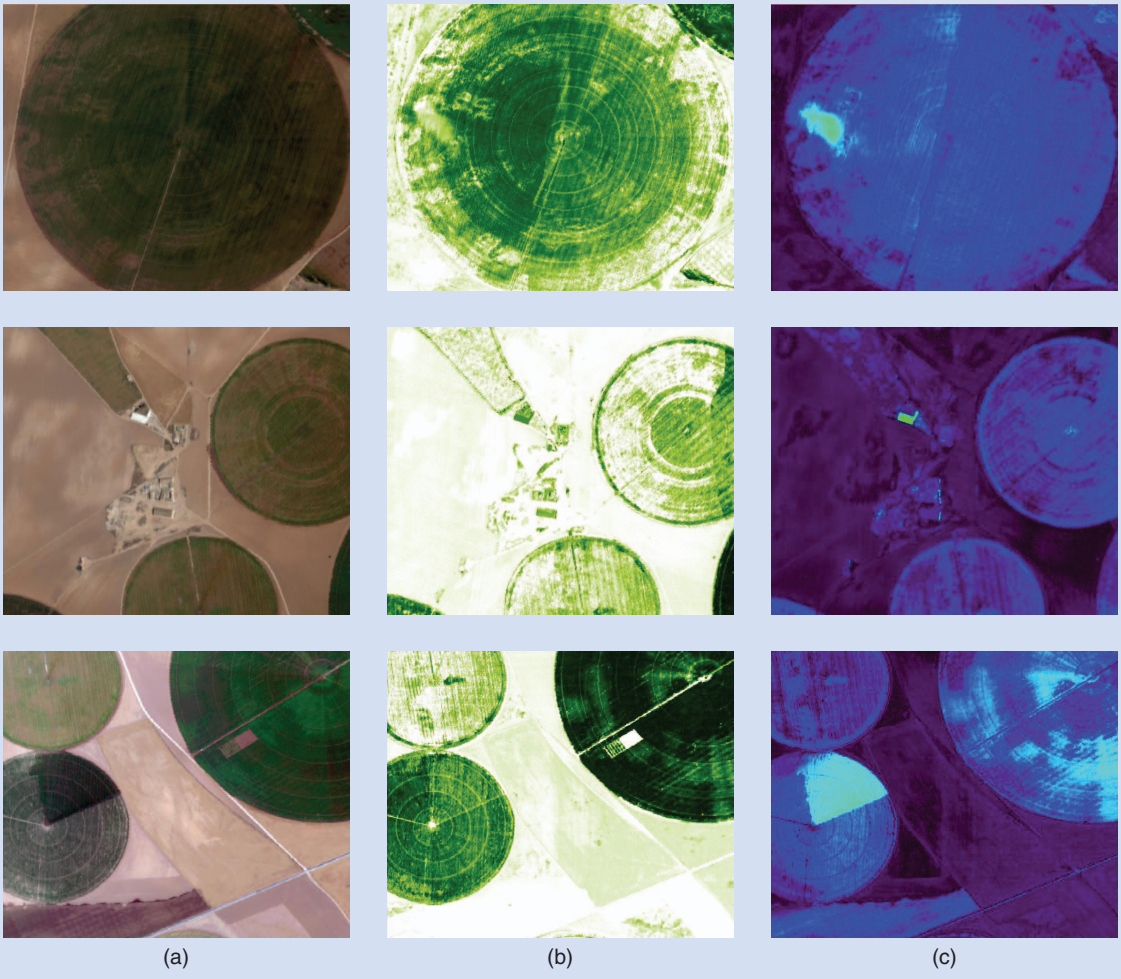

FIGURE 8. Three examples (top, middle, and bottom) of (a) CASI RGB snapshots, (b) *Chl* estimates, and (c) related uncertainty intervals.

FROM CORRELATION TO CAUSATION

Establishing causal relationships between random variables from empirical data is perhaps the most important challenge in science today. In this section, GP models are used for causal discovery, following the approach in [67] to discover causal relationships between the observed variables x and y. This methodology uses nonlinear regression from x→y (and vice versa, y→x) and assesses the independence of the forward, $r_f = \mathrm{y} - f(\mathrm{x})$, and backward residuals, $r_b = \mathrm{x} - g(\mathrm{y})$, with the input variable y (or x). The statistical significance of the independence test tells the right direction of causation. Essentially, the framework exploits nonlinear, nonparametric 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 discard the possibly strong assumption about noise distribution and to propose maximizing a dependency measure between residuals and regressors.

It is important to note that the estimation of causal relationships in this model suffers when the noise is not Gaussian and when linear models are used. Both scenarios pose serious identification problems that have led to an increased interest in nonlinear regression models that consider eventually non-Gaussian noise [69], [70]. The interest here is to assess the causality by discounting the elusive masking effects due to the assumption of Gaussian noise, as well as the possibly skewed distributions of the observation variable. This is why the standard GPR, VHGPR, and WGPR are used for comparison.

An approach in a relevant geoscience problem is exemplified. In the last few hundred years, human activities have precipitated an environmental crisis on Earth, which is commonly termed *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 Earth's atmosphere. Terrestrial ecosystems absorb approximately 120 Gt of carbon annually from the atmosphere, and approximately half is returned as plant respiration and the remaining 60 Gt yr^{-1} represents the net primary production (NPP). Out of this, approximately 50 Gt yr^{-1} is returned to the atmosphere as soil and litter respiration or via decomposition, while approximately 10 Gt yr^{-1} results in the net ecosystem production (NEP). The problem is in estimating the causal relationship between photosynthetic photon flux density (PPFD), which is a measure of light intensity, and the NEP, which results from the potential of ecosystems to sequestrate atmospheric carbon. Here, the total PPFD was measured as the number of photons falling on an area of 1 m^2/s, while the NEP was calculated by the photosynthetic uptake minus the release by respiration, which is known to be driven by either the total, diffuse, or direct PPFD. Discovering such relationships may be helpful in understanding the carbon fluxes and in establishing the sinks and sources of carbon across the globe. Three data sets are used, taken at a flux tower at site DE-Hai, involving PPFD (total), PPFD (diffuse), PPFD (direct) drivers and the NEP consequence variable [71]. The results for all three scenarios are shown in Table 6. These results generally confirm the good capability 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 GP models. As more flexible GP models are deployed, the sharpness in causal detection becomes more evident. Interestingly, heteroscedastic GP discounts the noise effects so that the dependency estimate becomes slightly more reliable.

TABLE 6. THE RESULTS OF THE PPFD AS A CAUSE OF THE NEP CASUAL 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

EMULATING RADIATIVE TRANSFER MODELS THROUGH GAUSSIAN PROCESSES

A slightly different approach for the use of GPs in remote sensing is to use them as fast approximations to complex physical models, which is 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, therefore, can be used to replace it, opening the door to more advanced biophysical-parameter-estimation methods (e.g., using data assimilation (DA) concepts] [73], [74], [76]).

FUNCTION APPROXIMATION, REGULARIZATION, AND EMULATION

A function is a mapping from an input parameter space to an output space. Consider that, for a particular application, a particular function is used, but the function can only be run a limited number of times (perhaps, for example, because the function is so complicated that it would take too long to run it repeatedly). For the purposes of this example, consider such function to be an RTM. One way to get around this limitation is to carry out an inference on the function itself, which would require the placement of a prior that encodes our understanding in the properties of the function (e.g., smoothness, continuity, or finite values)

and the use of 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 previously is a GP with an associated covariance function, as previously explained. Assuming the likelihood is also Gaussian and independent additive noise, what results is a reparameterization of the prior GP as the posterior, meaning that the output of our function for an arbitrary input $\mathbf{x}_*$ can now be predicted, conditional on the limited sampling of the original model's input and output pairings. The prediction will provide an estimate of the function value μ_{GP*}, and, more importantly, an estimate of the predictive uncertainty, σ^2_{GP*}. If the GP is able to correctly reproduce the function in which only a limited number of runs was available (which, in this context, is called the *simulator*), the GP can be used in its stead. This use of GPs is called *emulation*, and it is an exploitation of the versatility of GPs to effectively cope with varied mappings (or simulators).

Although emulators may appear to be trivial diversions, they have a number of important advantages. Firstly, if the simulator is computationally expensive, an emulator typically provides a very fast approximation to the simulator. Given a GP's ability to cope with fairly nonlinear problems, this method can be effective for a large number of complex physical models, such as RTMs that describe in some detail the scattering and absorption of photons by the atmosphere and vegetation. The emulator can thus be seen as a drop-in replacement for a complicated physical RTM. The fact that there is an associated uncertainty with emulator prediction is important; 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 access to fast physical models through machine learning opens new avenues that will be reviewed next.

FROM FORWARD AND BACKWARD MODELS TO STATISTICAL EMULATION

A particular problem often found in remote sensing is the inverse problem, in which a physical RTM is used to interpret observations of, e.g., surface directional reflectance or microwave backscatter, in terms of biophysical parameters, such as LAI or soil roughness. The computational complexity of the models at hand usually makes analytic inversions intractable; thus, the inversion method typically results in a least-squares problem in which the model's input parameters are varied until a minimum difference is found in the observations. However, remote sensing data are corrupted by uncertainties (e.g., additive noise and imaging artifacts) that degrade the data's information content, and observations are typically only available over small spectral or angular regions, giving a partial overview of the land surface, for example. Additionally, the processes that describe the interactions between photons and the scene are nonlinear. These effects conjure a situation in which many possible combinations of input parameters result in an adequate observation description and, therefore, a large amount of uncertainty in the retrieved parameters. To help circumvent the inverse problem, either more prior information or more evidence (i.e., observations) would need to be added. The flexibility of RTMs makes the latter strategy possible, as they can usually account for different sensor configurations (e.g., geometry and spectral sampling) while keeping a consistent description of the scene. New observations are typically hard to come by and will again be limited by uncertainty and partial observation of the whole system. Therefore, adding prior information is necessary to better constrain the inverse problem. Prior estimates include parameter distributions (derived, for example, from expert knowledge or historical data), expectations of smoothness in time and space, and physiological models of vegetation growth. Ultimately, the posterior calculation is a complicated problem that can typically be solved by MCMC methods, requiring many iterations (and therefore many executions of the RTM) or, under some assumptions, by a nonlinear cost-function-minimization problem. The latter is typically an iterative procedure, and gradient descent methods are required for efficiency. It is important to keep in mind that the goal here is to infer the land surface parameters conditioned on the remote sensing data and other prior knowledge, estimating the uncertainty of the parameters.

GAUSSIAN PROCESS MODELS AS EFFICIENT EMULATORS

The GP emulators can be used advantageously in complex inverse problems. The physical model can be emulated directly if MCMC methods are used, resulting in much faster parameter space exploration. In cost-function minimization, the emulator can be used instead of the full model, but GP may also be used to approximate the gradient of the emulated model,

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

From (10), it is seen 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-minimization approaches, the inverse of this matrix as the MAP point is the posterior covariance matrix, and thus a statement on the uncertainty of the retrieved parameters. A further benefit of numerically cheap approximations to the gradient is that local linearizations 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 nonlinear problem). Ultimately, having fast surrogate models of the most computationally demanding part of the inversion problem allows us to implement inversion strategies that

were practically impossible with these models and extend them to practical problem sizes.

A particular requirement in many RTMs is the prediction of spectral reflectance over the solar reflective domain (i.e., broadly from 400 to 2,500 nm) so instrument bandpass functions can be applied to the data. To emulate full spectra, the idea of the PCA of hyperspectral data can be extended, where there are large degrees of spectral redundancy. Let the output $\mathbf{y}$ be given a stacking of N_t spectra. Each of these spectra can be approximately reconstructed from

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

where only the first L principal components are considered, and σ_j is the jth score associated with the $\mathbf{w}_j$ principal component. In PCA, the principal components are orthogonal over the input set, so one 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 (11)].

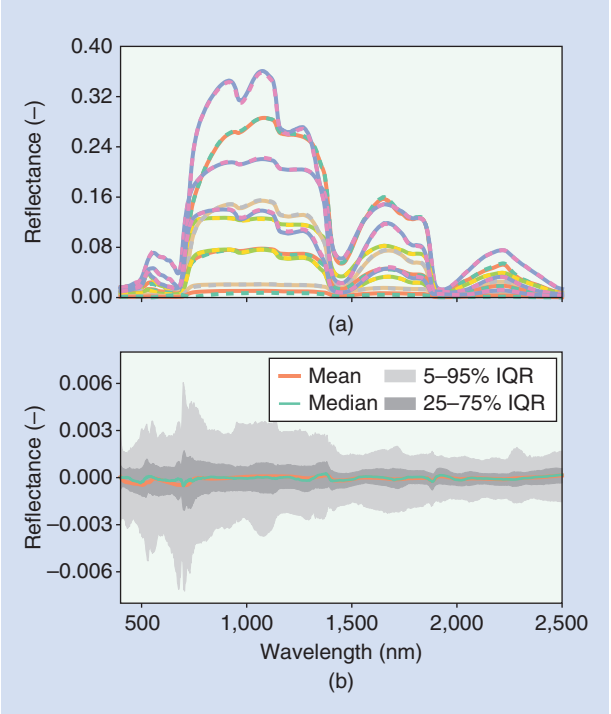

FIGURE 9. An example of RTM emulation with GPs. The PROSAIL soil-leaf-canopy RTM is emulated spectrally. (a) The complete model (full lines) and the emulated reflectance (dashed lines) for ten random input parameter sets. (b) The mean, median, 5–95%, and 25–75% interquartile ranges for the residuals 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 1,000 uniformly independent samples.

AN ILLUSTRATIVE EXAMPLE

As an example, consider a coupled, soil-leaf-canopy RTM over the solar reflective domain, PROSAIL [75]. A simple linear spectral mixture RTM for the soil (therefore assuming the soil properties are isotropic), the leaf optical properties spectra (PROSPECT) model, and the scattering by arbitrarily inclined leaves (SAIL) canopy RTM will be used. The aim is to map from a state made up of soil, leaf, canopy, and parameters such as LAI and chlorophyll content to top-of-canopy reflectance. This is an important example because the coupled model can be used within a DA system to infer the properties of the land surface (i.e., vegetation structure and biochemistry) from the atmospherically corrected directional surface reflectance. A validation of the emulation approach is shown in Figure 9, where the emulator has been trained with 250 input parameter-reflectance pairs, which were chosen using a Latin hypercube sampling design. Using the approach outlined in the previous section for multivariate output, L in (11) was chosen to be 11 so as to encompass 99% of the variance in the training set. It can be immediately seen that the emulator is virtually indistinguishable from the original model, with negligible bias in the validation, and a very small RMSE. Although PROSAIL is a fast RTM, this emulator is some 5,000-fold faster than the original in a contemporary PC. In evaluating the GP, the PROSAIL gradient is also calculated.

CONCLUSIONS AND FURTHER WORK

This article provides a comprehensive survey of GPs in the context of remote sensing data analysis, particularly for statistical biophysical parameter estimation. The GPs' main properties and their advantages over other estimation methods were summarized to find that GPs can essentially provide competitive predictive power, give error bars for estimations, allow design and optimization of sensible kernel functions, and analyze the encoded knowledge in a model via ARD kernels. The GP models also offer a solid Bayesian framework to formulate new algorithms that are well suited to signal characteristics. For example, it can be seen that, by incorporating proper priors, signal-dependent noise can be encompassed and parametric forms of warping the observations as an alternative to either ad-hoc filtering or linearization, respectively, can be inferred. A downside for GPs is the scalability issue, which is that, essentially, the optimization of GP models require computing determinants and invert matrices of size n × n, which runs cubically in computational time and quadratically in memory storage. In recent years, however, great advances have been made in machine learning, and it is now possible to train GPs with millions of points in almost linear time.

All of the developments were illustrated on local and global scales through a full set of illustrative examples in geosciences and remote sensing. In particular, addressed were important problems of ocean, land, and atmospheric sciences, from accurately estimating oceanic chlorophyll

content and pigments to vegetation properties (such as LAI, chlorophyll content, or fluorescence) from multi- and hyperspectral sensors, as well as estimating atmospheric parameters (such as temperature, moisture, and ozone) from infrared sounders.

This article has taken a step forward by introducing and illustrating two relevant uses of GP technology: 1) by studying the important issue of passing from regression to causation from empirical data, and 2) by considering the approximating physically based RTMs with GPs. Both approaches, yet in their infancy, are promising ways to develop flexible statistical models that discover and incorporate physical knowledge about the problem. More exciting developments are envisioned in the intersection of physics and machine intelligence.

ACKNOWLEDGMENTS

The authors wish to acknowledge the collaboration, comments, and fruitful discussions had with many researchers during the last decade on GP models for remote sensing and geoscience applications, including Miguel Lázaro-Gredilla (Vicarious), Robert Jenssen (University of Tromsø, Norway), Martin Jung (Max Planck Institute, Jena, Germany), and Salcho Salcedo-Saez (University of Alcalá, Madrid, Spain). This article has been partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO) under project TIN2012-38102-C03-01 and by the European Research Council (ERC) Consolidator Grant ERC-2014-CoG-647423. The authors also wish to acknowledge the MINECO for the FEDER-funded project GEOLEARN with TIN2015-64210-R.

AUTHOR INFORMATION

Gustau Camps-Valls (gcamps@uv.es) received a B.Sc. degree in physics in 1996 and in electronics engineering in 1998 and a Ph.D. degree in physics in 2002, all from the Universitat de València. He is currently an associate professor (hab. full professor) in the Department of Electronics Engineering. He is a research coordinator in the Image and Signal Processing group. He has been a visiting researcher at the Remote Sensing Laboratory, University Trento, Italy, in 2002; the Max Planck Institute for Biological Cybernetics, Tübingen, Germany, in 2009; and an invited professor at the École Polytechnique Fédérale de Lausanne, Switzerland, in 2013. He is the coauthor of 120 journal papers, more than 150 conference papers, and 20 international book chapters, and his interests are in the development of machine-learning algorithms for geoscience and remote sensing data analysis. He is a Senior Member of the IEEE. For further details, see http://www.uv.es/gcamps.

Jochem Verrelst (jverrelst@uv.es) received the M.Sc. degree in tropical land use and in geoinformation science, both in 2005, and the Ph.D. degree in remote sensing in 2010 from Wageningen University, The Netherlands. His dissertation focused on the spaceborne spectrodirectional estimation of forest properties. During 2010–2012, he was a Marie Curie Postdoctoral Fellow at the Laboratory for Earth Observation, Image Processing Laboratory, University of Valencia, Spain, where he is currently employed. He is involved in preparatory activities of the European Space Agency's Eighth Earth Explorer Fluorescence Explorer. His research interests include the retrieval of vegetation properties using airborne and satellite data, canopy radiative transfer modeling, and hyperspectral data analysis.

Jordi Muñoz-Marí (jordi@uv.es) received the B.Sc. degree in physics, the M.Sc. degree in electronics engineering, and the Ph.D. degree in electronics engineering from the Universitat de València, Spain, in 1993, 1996, and 2003, respectively. Currently, he is an associate professor with the Electronics Engineering Department, Universitat de València, where he teaches electronic circuits and programmable logical devices, digital electronic systems, and microprocessor electronic systems. He has been a visiting researcher at the Remote Sensing Laboratory, University of Trento, Italy, in 2003 and an invited professor at the Laboratory of Geographic Information Systems of the École Polytechnique Fédérale de Lausanne, Switzerland, in 2013. His research interests are tied to the development of machine-learning algorithms for signal and image processing. He is coauthor of 34 journal papers, 50 conference papers, and four international book chapters.

Valero Laparra (lapeva@uv.es) received a B.Sc. degree in telecommunications engineering in 2005, a B.Sc. degree in electronics engineering in 2007, a B.Sc. degree in mathematics in 2010, and a Ph.D. degree in computer science and mathematics in 2011. He is a postdoc in the Image and Signal Processing group at Universitat de València, and currently doing a research stay in the Laboratory for Computer Vision at New York University. For further details, see http://www.uv.es/lapeva.

Fernando Mateo-Jiménez (fmateo@uv.es) received a degree in telecommunication engineering from the Polytechnic University of Valencia, Spain, in 2005 and a Ph.D. degree in electronics engineering from the same university in 2012. He has carried out research stays in several research centers and universities such as Delft University of Technology, The Netherlands; Aalto University, Helsinki, Finland; and CERN, Geneva, Switzerland. At present, he works as a data scientist at the Intelligent Data Analysis Laboratory, University of València. He is author or coauthor of approximately 50 international peer-reviewed journal articles or book chapters. His research focuses on data mining and preprocessing, feature selection, machine-learning models both for regression and classification, clustering, and time-series forecasting.

José Gómez-Dans (j.gomez-dans@ucl.ac.uk) received the M.Eng. and Ph.D. degrees in electronic engineering from the University of Sheffield, United Kingdom. His Ph.D. thesis dealt with the use of polarimetric interferometry in synthetic aperture radar data to monitor

croplands. He has worked on cryopheric topographic mapping and in the combined use of Earth-observation (EO) data and dynamic crop models for regional crop monitoring on regional scales. In 2008, he joined University College London as a researcher working on the use of EO data for fire and disturbance monitoring. He is currently employed by the National Centre for Earth Observation, based at University College London, United Kingdom, where he is mostly responsible for disturbance monitoring and data assimilation studies for land vegetation and carbon monitoring. His research interests include assessment of the impact of fire in vegetation using EO, radiative transfer theory, advanced approaches to inverse problems that relate to land surface parameter retrieval, and the assimilation of EO data into dynamic vegetation models. He is a Student Member of the IEEE.

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," *Int. J. Appl. Earth Observation 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, Sept. 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, May 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," *Rem. Sens. Env.*, vol. 120, pp. 37–57, May 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 Astronaut.*, vol. 61, no. 1–6, pp. 115–120, June–Aug. 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, no. 15, pp. 83–101, Feb. 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 Astronaut.*, vol. 65, no. 9–10, pp. 1429–1436, Nov.–Dec. 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 *Proc. Int. Soc. Opt. Eng.*, vol. 8870, Sept. 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 J. Photogramm. Rem. Sens.*, Oct. 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, May 2012.

[11] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: 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 J. Photogramm. Rem. Sens.*, vol. 108, pp. 260–272, Oct. 2015.

[13] G. Camps-Valls, D. Tuia, L. Gómez-Chova, and J. Malo, " Remote sensing image processing," in *Synthesis Lectures on Image, Video, and Multimedia Processing*, Al Bovik, Ed. San Rafael, CA: Morgan & Claypool, 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, Dec. 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. Part 1: Principles of development and production," *Rem. Sens. Env*, vol. 137, pp. 299–309, Oct. 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," *Sci.*, vol. 329, no. 5993, p. 834, Aug. 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," *J. Geophys. Res.*, vol. 116, no. G3, pp. 1–16, Sept. 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, Apr. 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," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 14, pp. E1327–E1333, 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. Geosci. 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, Nov. 2006.

[23] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: 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, pp. 127–139, Mar. 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(2), pp. 1832–1843, Nov. 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 J. Select. Topics Appl. Earth Observ. Rem. Sens.*, vol. 6, no. 2, pp. 867–874, May 2013.

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

[28] M. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *J. Roy. Statist. Soc. Series B: Statist. Methodol.*, vol. 63, no. 3, pp. 425–450, 2001.

[29] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Neural Inform. Processing Syst. 8*, Cambridge, MA: MIT Press, 1995, pp. 598–604.

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

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

[32] A. O'Hagan and J. F. C. Kingman, "Curve fitting and optimal design for prediction," *J. Roy. Statist. Soc., Series B (Methodol.)*, vol. 40, no. 1, pp. 1–42, 1978.

[33] M. Reed and B. Simon, *Functional Analysis, (Methods of Modern Mathematical Physics)*, vol. 1, New York: Academic, 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. Rem. Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[35] P. Sampson and P. Guttorp, "Nonparametric estimation of nonstationary spatial covariance structure," *J. Amer. Statist. Assoc. Pub.*, 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 J. Select. Topics Appl. Earth Observ. Rem. Sens.*, vol. 1, no. 1, pp. 18–27, Mar. 2008.

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

[38] T. Khatib, A. Mohamed, and K. Sopian, "A review of solar energy modeling techniques," *Renewable Sustainable Energy Rev.*, vol. 16, no. 5, pp. 2864–2869, June 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 Trans. Geosci. Rem. Sens.*, vol. 50, no. 8, pp. 2934–2953, Jan. 2012.

[40] E. Lorenz, J. Hurka, D. Heinemann, and H. G. Beyer, "Irradiance forecasting for the power prediction of grid-connected photovoltaic systems," *IEEE J. Select. Topics Appl. Earth Observ. Rem. Sens.*, vol. 2, no. 1, pp. 2–10, Mar. 2009.

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

[42] M. E. Tipping, "The relevance vector machine," in *Neural Inform. Processing Syst. 12*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge, MA: MIT Press, 2000.

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

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

[45] M. Lázaro-Gredilla and M. K. Titsias, "Variational heteroscedastic gaussian process regression," in *Proc. 28th Int. Conf. Mach. Learning, 2011*. Bellevue, WA: ACM, pp. 841–848.

[46] E. Snelson, C. Rasmussen, and Z. Ghahramani, "Warped gaussian processes," in *Neural Inform. Processing Syst. 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2004.

[47] M. Lázaro-Gredilla, "Bayesian warped Gaussian processes," in *Neural Inform. Processing Syst. 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Cambridge, MA: MIT Press, 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," *J. Geophys. Res.*, 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, MA, NASA Tech. Memo. 2000-206892, vol. 11, pp. 3–8, 2000.

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

[51] A. G. Wilson, D. A. Knowles, and Z. Ghahramani, "Gaussian process regression networks," in *Proc. 29th Int. Conf. Mach. Learning*, J. Langford and J. Pineau, Eds. Edinburgh: Omnipress, June 2012.

[52] K. N. Liou, *An Introduction to Atmospheric Radiation*, 2nd ed. New York: Academic, 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, Mar. 1992.

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

[55] J. Quinonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1939–1959, Dec. 2005.

[56] E. Snelson and Z. Ghahramani, "Local and global sparse Gaussian process approximations," in *Proc. Artificial Intell. and Statist.*, 2007.

[57] M. Lázaro-Gredilla, J. Q. Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse spectrum Gaussian process regression," *J. Mach. Learn. Res.*, vol. 11, pp. 1865–1881, June 2010.

[58] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Neural Inform. Processing Syst. 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Cambridge, MA: MIT Press, 2007.

[59] A. G. Wilson, H. Nickisch, "Kernel interpolation for scalable structured Gaussian processes (KISS-GP)," in *Proc. 32nd Int. Conf. Mach. Learn. 2015*, Lille, France, 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, pp. 337–352, Apr. 2004.

[61] I. Filella and J. Penuelas, "The red edge position and shape as indicators of plant chlorophyll content, biomass and hydric status," *Int. J. Rem. Sens.*, 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, May 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," *J. Photochem. Photobiol. B: Biol.*, vol. 134, pp. 37–48, May 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 Biol.*, vol. 20, no. 4, pp. 1191–1210, 2014.

[65] GCOS. (2011, Dec.). Systematic observation requirements for satellite-based products for climate. [Online]. p. 138. 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," *J. Photogramm. Rem. Sens.*, vol. 86, pp. 157–167, Dec. 2013.

[67] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Neural Inform. Processing Syst. 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2008, pp. 689–696.

[68] J. Mooij, D. Janzing, J. Peters, B. Schölkopf, "Regression by dependence minimization and its application to causal inference in additive noise models," in *Proc. 26th Int. Conf. Mach. Learning*, A. Danyluk, L. Bottou, and M. Littman, Eds. Montréal, Canada, 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," *J. Mach. Learning Res.*, 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," *Mach. 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 Biol.*, vol. 16, no. 1, pp. 2737–2749, Aug. 2010.

[72] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and analysis of computer experiments," *Statist. Sci.*, vol. 4, no. 4, 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," *Rem. Sens. Env.*, vol. 112, no. 4, pp. 1347–1364, Apr. 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)," *Rem. Sens. Env.*, vol. 120, pp. 219–235, May 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, Sept. 2009.

[76] J. L. Gómez-Dans, P. E. Lewis, and M. Disney, "Efficient emulation of radiative transfer codes using Gaussian processes and application to land surface parameter inferences," *Remote Sens.*, vol. 8, no. 2, p. 119, Feb. 2016.

GRS