Convocatorias 2015
Proyectos EXCELENCIA y Proyectos RETOS
Dirección General de Investigación Científica y Técnica
Subdirección General de Proyectos de Investigación

SUMMARY OF THE PROPOSAL

RESEARCHERS: Jordi Muñoz Marí, Gustau Camps-Valls

TITLE OF THE PROJECT: ADVANCES IN MACHINE LEARNING FOR LARGE SCALE REMOTE SENSING DATA PROCESSING

ACRONYM: GEOLEARN

SUMMARY:

GEOLEARN is an interdisciplinary project that aims to develop novel machine learning algorithms to analyze Earth Observation (EO) data. In the last decade, machine learning models have helped to monitor land and atmosphere through the analysis and estimation of climate variables and biophysical parameters. Current approaches, however, cannot deal with the particular characteristics of remote sensing data efficiently. In the upcoming years, this problem will largely increase: several satellite missions, such as the operational EU Copernicus Sentinels and the future Meteosat Third Generation (MTG), will be launched and we will face the urgent need to process and understand a huge amount of complex, heterogeneous and structured data streams in order to monitor our Planet.

GEOLEARN aims to develop the next generation of machine learning algorithms for EO data analysis, and will be addressed in four workpackages (WP). In the first WP, we address the problem of adapting machine learning algorithms to remote sensing (RS) data, dealing with heterogeneous data types at different (temporal, spatial and spectral) scales and resolutions. Multivariate outputs are also required in many EO data processing problems, in order to constrain algorithms to sensible predictions. Furthermore, the uncertainty in the predictions will be improved in order to evaluate their reliability and propagation. Finally, physical knowledge will be included in the mathematical models by means of radiative transfer models (RTMs) inversion (also known in the field as emulation).

The second GEOLEARN WP deals with developing computationally efficient algorithms able to manage large amounts of data quickly and accurately. The tasks in this WP will be tackled from three complementary perspectives: (1) from a theoretical approach, by reducing the inherent complexity of the models based on Gaussian processes, kernel methods and deep architectures; (2) exploiting new hardware and software resources for parallelization; and (3) following divide-and-conquer strategies adapted to the particularities of remote sensing data.

In the third WP we will extract knowledge from the developed algorithms through global sensitivity analysis, and will propose novel algorithms for causal inference on relevant climate science applications.

Finally, these developments will be guided in the fourth WP by the challenging problems of emulating radiative transfer models, estimating biophysical parameters (e.g. vegetation cover or chlorophyll content) and atmospheric variables (e.g. temperature and ozone profiles) at both local and global planetary scales, and the estimation of global time-resolved carbon and heat fluxes, which will allow for a rapid development of policy responses on climate change.

The outcomes from the GEOLEARN project will provide, among others, remote sensing applications and dedicated modules to benefit the processing chains and products derived from the sensors on board Sentinel 2 and 3 satellites of the EU Copernicus program and EUMETSAT Meteosat satellites.

KEYWORDS: Machine learning, remote sensing, geosciences, climate science, image processing, time series analysis, kernel methods, Gaussian processes, deep learning.

SCIENTIFIC DOCUMENT

C.1. SCIENTIFIC PROPOSAL

C.1.1. Previous works and state of the art

Human activities, in particular those involving the combustion of fossil fuels and the conversion of land for forestry and agriculture, have ever increased since the Industrial Revolution. These activities have had a definite impact on the Earth's climate system. Undoubtedly, the Earth is experiencing important climate changes [1], and the attribution to natural or anthropogenic causes is matter of current and intense research [2]. Nowadays, *monitoring* and *understanding* the Earth's climate system is one of the main challenges in Science, but also it is crucial for adopting appropriate policies by decision-makers. Measuring key parameters of climate evolution by Earth Observation (EO) satellite missions and *in situ* measurements, along with the exploitation of quantitative statistical methods, are essential in current climate science. This unique combination of data and techniques allows us monitoring our Planet at continental and global scales. The field has obvious societal, environmental and economical implications, given the rapidly growing demand of bio-fuels and food.

Earth Observation aims to monitoring Earth's changes and evolution using the information provided by satellites, airborne sensors and ground measures of physical parameters. EO satellites, endowed with a high temporal resolution, enable the retrieval and hence monitoring of climate and bio-geo-physical variables [3]. With the forthcoming super-spectral Copernicus Sentinel-2 (S2) [4] and Sentinel-3 missions [5], the upcoming Meteosat Third Generation Infrared Sounder (MTG-IRS), as well as the planned german EnMAP [6], and European Space Agency (ESA) Earth Explorers candidate mission FLEX [7], an unprecedented data stream for land, ocean and atmosphere monitoring will soon become available to a diverse user community. The problem of managing and processing massive data volumes requires enhanced processing techniques on *accuracy*, *robustness* and *computational cost*. In addition, the statistical models should be also *self-explanatory*, in the sense that they should capture plausible physical relations and *explain causal* links between the climate variables and observations.

Statistical machine learning (ML) has proven successful in many disciplines of Science and Engineering [8]. In the last decade, statistical inference has widely contributed to the estimation of particular essential climate variables (ECVs) and related bio-geo-physical parameters, such as temperature, ozone, or chlorophyll content [9]. For example, current operational leaf-area-index (LAI) global maps are typically produced with neural networks, Gross primary production (GPP) is estimated using ensembles of random forests, neural networks and process-based models [10, 11], biomass has been estimated with stepwise multiple regression [12], and support vector regression showed high efficiency in modelling LAI and evapotranspiration [13]. It is worth noting that recently much attention has been payed to Gaussian Process Regression (GPR) [14], as they provided very good results in chlorophyll content estimation [15, 16], GPP [17] and atmospheric variables [18].

Despite all these advances in the statistical treatment of EO data, current ML algorithms do not cope efficiently with some data characteristics. We identify three main aspects that require urgent attention in the current and upcoming scenario of Earth monitoring with ML techniques, and that we will approach in this project: namely, (i) how to improve model's accuracy that respects both data structures and physical facts, (ii) how to scale to huge EO data streams, and, more importantly, (iii) how to extract knowledge from EO machine learning models to gain in problem understanding.

Adapting machine learning to EO data characteristics.

Very often, regression algorithms are applied blindly to *remote sensing* (RS) data with few, or none, adaptation to respect data characteristics [9]. From a pure machine learning point of view, EO data is essentially structured, multisource, and multimodal [19]. However, few approaches have considered fusion of multisensor data for climate variable prediction, and only recently we have imposed spatial or temporal structures in the retrieval models [17, 20],

and work with exogenous time series with proper regression models [21, 22]. From a signal processing standpoint, the acquired time series of, e.g. carbon, heat and water fluxes, exhibit heteroscedastic relations, strong correlations between observations, and the bio-geo-chemical processes occur at different temporal and spatial scales. When models do not match the time/space structures, not only *prediction but also uncertainty* estimation are compromised, with strong implications on further studies that rely on previous models. Strikingly enough, the international Global Climate Observation System (GCOS) [23] panel recommends less than 20% of prediction uncertainty in the models, but this is still difficult to achieve for the estimation of many important ECVs. Poor uncertainty estimates directly reflect an ubiquitous problem: statistical models do not incorporate *physical knowledge* and *a priori* information. This hampers current machine learning models being widely accepted by the EO community as the preferred model to generate products. Actually, they are typically used just as *first guess* estimators for data assimilation with physical models. We posit that machine learning methods should be constrained by physical models to provide sensible and consistent predictions.

Scaling machine learning regression models to huge data streams.

Dealing with this unprecedented amount of EO data requires efficient implementations of the algorithms ready to handle this immense and heterogeneous data volume. ESA Sentinels[1] [15, 24] will deliver improved spectral and temporal resolutions, while the MTG-IRS infrared sounder[2] [25] will acquire each pixel (field of view) in about 1800 spectral dimensions. Yet, also the output variable space is increasing: for instance, ECVs and time series of carbon, water and heat fluxes in the FLUXNET activities[3] [26] need to be predicted simultaneously to attain consistent models, and atmospheric state vectors define hundreds of correlated and structured output variables, such as the temperature or moisture values across the atmospheric column. This upcoming EO data streams requires new automatic tools and algorithms able to adapt and exploit the relevant information within the data. Machine learning algorithms could be of paramount importance in solving these new challenges. Currently, the used state-of-the-art kernel methods and Gaussian Process Regression (GPR) models do not scale well to more than a few thousand points, and need intensive training in cluster facilities. This hampers adoption of ML by regular users, and keep this technology obscure to organizations such as ESA and EUMETSAT.

Unveiling knowledge in machine learning EO models.

Statistical learning models should be transparent and provide information about the learned relations, as process-based and physical models do. Unfortunately, machine learning has traditionally focused on *fitting rather than understanding*. On the one hand, very few works have studied the relative relevance of advanced statistical retrieval methods [27, 26], but it is still unclear if the identified relations are too naïve or even biased, given the limited datasets considered to train the models. On the other hand, a quite limited number of works have explained causal relations between vegetation variables via graphical models: [28] used Bayesian modelling to assess the impact of climate change on biofuel production, while [29] used constrained structure learning to derive hypotheses of causal relationships between prominent modes of temporal atmospheric variability, and very recently [30] used the causal counter-factual theory for the attribution of weather and climate-related events. Again, the methods rely on limited amount of data, use standard structure learning algorithms driven by uni-variate dependence estimates, and seldom identify new drivers or confounding factors.

Monitoring land, vegetation and atmosphere with advanced machine learning.

The aforementioned deficiencies are common to many geoscience and EO problems. In this project, we will focus on three selected meta-studies for global monitoring with statistical inference, in which we are experts: (1) *Estimation of vegetation parameters* in the context of the upcoming Sentinels missions; (2) *Estimation of atmospheric temperature and moisture profiles* in the context of the upcoming MTG-IRS super-spectral infrared sounder; and (3) *Estimation of*

[1] https://sentinel.esa.int
[2] http://esamultimedia.esa.int/docs/MinisterialCouncil/MC-MTG_1811.pdf
[3] http://fluxnet.ornl.gov/

global time-resolved carbon and heat fluxes in the context of FLUXNET activities. These applications of global monitoring are ideal testbeds for the proposed methodological developments. They are challenging, large scale, structured input-output domains in Earth climate science that may lead to important and ground-breaking achievements.

C.1.2. Hypothesis and general objective

The GEOLEARN project is aimed to develop a new generation of machine learning algorithms for Earth Observation global monitoring. We advocate that machine learning algorithms for EO applications need to be guided both by data and by prior physical knowledge. This combination is the way to restrict the family of possible solutions and thus to obtain non-parametric flexible models that respect the physical rules governing the Earth climate system. Current algorithms need to be adapted to the particular specifics of the EO data streams, i.e. need to deal with multivariate outputs, spatial and temporal complex structures, and massive datasets. We are equally concerned about the 'black-box' criticism to statistical learning algorithms, for which we aim to design self-explanatory models and take a leap towards the relevant concept of causal inference from empirical EO data. The guiding hypothesis of GEOLEARN is illustrated in Fig. 1.

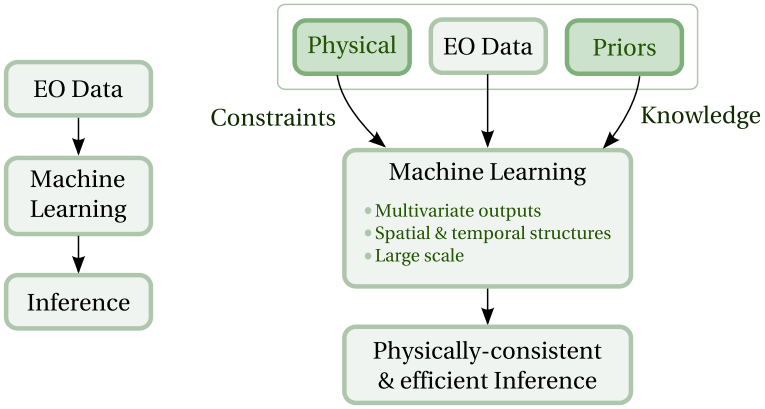

Figure 1: *Traditional algorithms are fed with remote sensing data to perform inferences, but no attention is payed to the underlying physical processes (left). In the proposed approach (right), machine learning models include physically meaningful constraints and prior knowledge about the time and spatial scales of the processes, and scale to large datasets. The approach will deliver a framework in which models will leverage efficient, and physically-consistent inferences.*

The advances in the GEOLEARN project will represent an improvement in future machine learning models for regression and causation, and could answer relevant questions in today's Climate Science:

- *Impact of climate and remote sensing variables.* What is the impact of warmer temperatures to soil mineralization? What is the relative relevance of nitrogen to the terrestrial carbon uptake? How the change in land use impacts uncertainty estimation of GPP [10, 11]?
- *Impact of alternative variables.* Are bio/geo-diversity variables directly mediated by gross/net production? And viceversa: Can biodiversity indicators constitute good covariates for GPP estimation?
- *Sun-induced fluorescence (SIF) as potential ECV.* Assessing whether the observed SIF-GPP relations at the leaf-level also hold at the synoptic/monthly scale, and what vegetation and meteorological variables drive the SIF and GPP signals. This is still an elusive problem, mainly due to the difficulty in retrieval and validation [31–33].

C.1.3. Specific goals

The main goal of the GEOLEARN project is to develop new machine learning models for the efficient treatment of biophysical land parameters and related covariates at continental and global scales. This main scientific goal translates into the following specific objectives:

1. **Improve prediction models by adaptation to Earth Observation data characteristics.** Current practice apply off-the-shelf machine learning algorithms directly for biophysical parameter estimation problem as a regression problem. Models do not respect relevant EO

data characteristics, such as non-Gaussianity, presence of heteroscedastic and nonstationary processes, and non-i.i.d. (spatial and temporal) relations. Models must be improved in terms of accuracy, reduced uncertainty of the estimates, and consistency of multiple output predictions. In addition, ML models need to encode prior (physical) knowledge. Emulation of physical radiative transfer models with nonparametric algorithms will be approached here. We will also design regularizers that enforce structure in phenological cycles and include rules from physical vegetation models. Advanced structured, multioutput Gaussian processes and deep nets will be developed here as well (see Fig. 2 for a real example).

2. **Scaling Machine Learning for EO data processing.** Efficiency of algorithms is also tied to computational burden given the large heterogeneous data streams. We will tackle this EO 'big data' problem from three complementary perspectives: (1) from a theoretical approach, by reducing the inherent complexity of the models, (2) from a computational perspective, using new hardware and software resources for parallelization, and (3) adapting divide-and-conquer strategies to EO data specifics (see Fig. 3 for a real example).
3. **Extract knowledge from Earth observation data and models.** Explaining the potentially complex interactions between the involved covariates for ECV estimation is essential to *understand* the climate mechanisms. So far, statistical models are treated as pure black box models. There is an urgent need both to unveil the knowledge encoded in non-parametric retrieval models, and to advance in the evaluation of potentially useful alternative covariates. We will investigate feature selection and ranking, in the form of sensitivity analysis of the predictive mean/variance of GPs and deep nets, as well as regression-based causal schemes applied to large heterogeneous EO data streams (see Fig. 4 for a real example).

C.1.4 Proposed methodology.

Work Package 1 (WP1). Improvement of algorithms.

This task will develop new algorithm regression models (mainly based on GPs and deep architectures) to cope with the shortcomings identified before. All the tasks will be consistent with the particular needs of the problems described in WP4. We detail the ideas and foreseen developments in what follows.

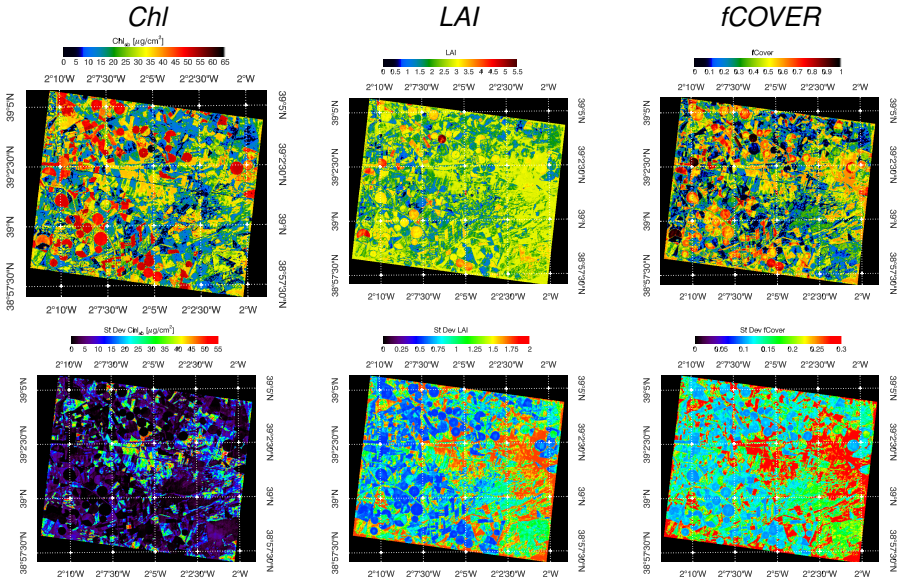

Figure 2: *Predictive mean maps (top) and associated predictive variance (bottom) for three important biophysical parameters (chlorophyll content, leaf area index and fractional coverage) describing vegetation status generated with Gaussian Processes using a hyperspectral CHRIS image [17, 27]. Current uncertainty maps fairly meet the Global Climate Observing System (GCOS) prescriptions of an uncertainty maximum 20% [23], and can be used as a quality mask. However, current GP models are very limited as they do not exploit neither the spatial or temporal information, models are generated independently for each observed variable, and are so far guided by data alone, without any inclusion of prior physical knowledge.*

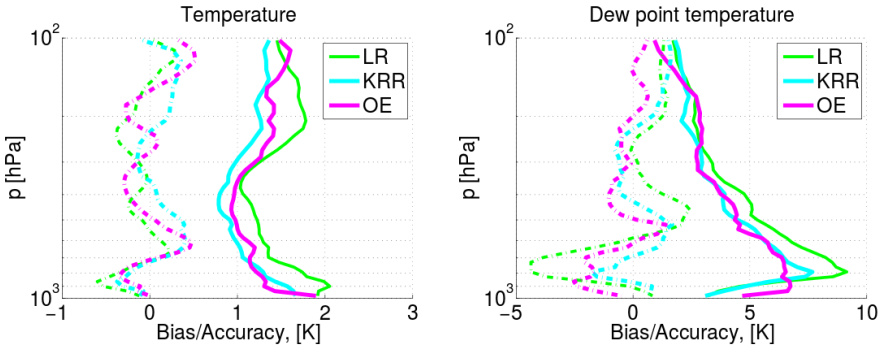

Figure 3: *Prediction error profiles across the atmospheric column using linear regression (LR), Kernel Ridge Regression (KRR), and a physically-based optimal estimation (OE) using IASI infrared sounder data for temperature (left) and moisture (right) [25]. Current machine learning models are highly competitive in RMSE terms (solid) and bias (dashed) versus OE and extremely efficient nowadays at the prediction test: OE takes 8.19 sec per pixel, while KRR only 0.043 sec, a gain of* $\times 190$*. However, current statistical models do not scale well for training, and do not take physical constraints into account.*

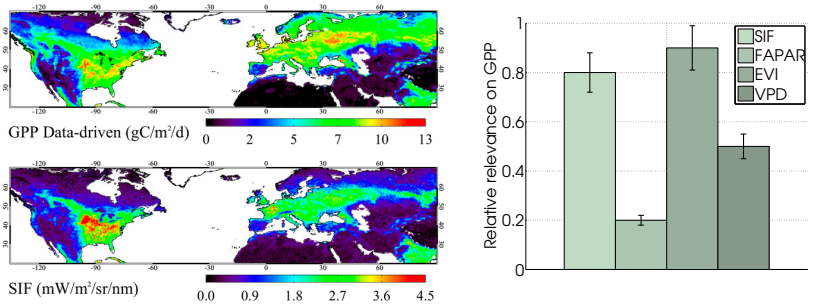

Figure 4: *Spatial patterns of maximum monthly GPP using random forests and maximum monthly SIF with linear regression [33] (left). Improvements in advanced regression methods are expected in GEOLEARN when exploiting structure and non-stationarity in the models. GEOLEARN will also develop feature rankings of drivers that impact essential climate variables (right).*

- **WP1.1 Structured multiscale and multiresolution domains.** Earth observation data are mainly structured data, i.e. it exhibits clear and strong spatial, temporal and spectral correlations at different scales and hierarchies that need to be exploited. Very often such obvious observation is not respected, and data (images of different spatial-spectral resolutions, times series of fluxes, biophysical/atmospheric parameters) are wrongly treated as i.i.d. objects, and are processed with fixed *ad hoc* receptive fields and time embeddings. In recent years, we have proposed a series of algorithms, based on kernel methods [34], Gaussian Processes [14] and deep learning [35] to account for spatial and temporal structures [36–38], multiple temporal resolutions [20], multiple tasks [39] and multimodal fusion [40], in RS data. We plan to extend some of these models to other relevant settings in RS, by developing (1) *tensorial* convolutional nets in space-spectrum domains that better match the properties of the EO data in both optical and infrared microwave sensors [9]; (2) *spatio-temporal* finite impulse response (FIR) deep nets [41], as a novel alternative to the current use of convolutional nets plus Long Short-Term Memory (LSTM) units; and (3) *structured deep kernel regression* without the need of pre-images extending [42].
- **WP1.2 Multivariate outputs.** Bio-geo-physical parameter estimation are fundamentally multi-output regression problems in which the output covariates are typically correlated and show dependencies. For example, plant density parameters like LAI and fCover are correlated, but very often they are predicted using different regression models that lead to inconsistencies. Another paradigmatic case is the extremely high output correlation in atmospheric state vectors: typical sampling of the atmospheric column yields between 100 and 200 outputs that can be compressed into 6 to 10 principal components. Developing individual models or linear compression of the output space do not constrain predictions to sensible (and biophysically consistent) levels, nor account for nonlinear correlations. Actually, constraining

the predictions to sensible and consistent levels might have relevant implications in posterior climate prediction models [43]. In machine learning, the field is known as *structured-output learning*, and is tightly related to *multitask learning*, in which we have developed novel implementations for data classification [39, 44]. Extension to the regression setting is far from being a trivial task, as the number of constraints increases cubically with the number of samples and outputs. In this sub-task, we will (1) extend Gaussian process regression networks (GPRN) [45] in order to account for explicit physically-inspired output constraints; and (2) propose orthonormalized kernel feature extraction methods [46] as efficient alternatives to both GPRN and PCA projections of the output space. The algorithms will be exploited in WP4 to generate global scale products of carbon, energy and/or heat fluxes simultaneously, to estimate atmospheric profiles, and for the inversion of physical models.

- **WP1.3 Uncertainty estimation and propagation.**[4] An important cornerstone in geoscience model analysis is related to the analysis of uncertainty of the estimates [47, 48]. Two different approaches will be followed here: On the one hand, we will take advantage of developments in previous tasks to derive tighter predictive variances for GPs than we did before [15, 49]. In these works, the associated uncertainty provided information about the success of transporting a locally-trained model to other sites and conditions [50]. However, the trustworthiness of GP posteriors has been questioned [51] given that GP covariances leverage global models not necessarily capturing the data manifold local structure. We will approach this important problem in two ways: (1) by exploring our coarse-to-fine unsupervised covariances [52] in the context of GPs; and (2) exploring combination of experts via standard averaging [53] or by sophisticated geometric distance weighting to approximate the prior far from the sampled space [54]. In both cases, we will adapt algorithms to emulators of vegetation and atmosphere (see WP4.1 and WP4.2). On the other hand, a direct measure of the uncertainty in the prediction is the Jacobian of the transformation, which characterizes the variation of the prediction given the input data. A high determinant of the Jacobian means that a small change in the input will affect the prediction drastically. We have explored these issues to derive average feature rankings and sensitivity maps in GPs (see more in WP3.2). We will extend the analysis to deep architectures that incorporate unsupervised sparse dictionary learning [40, 55], multi-task [56] and time-dependent [41] regularizers in the context of carbon and heat fluxes upscaling [26] (see WP4.3). Results obtained in this task can be very useful to meet GCOS recommendations on EO products [23], that have clear implications on vegetation, ecosystem, and crop yield models, and to constrain subsequent climate models that need tight uncertainty estimates [43].
- **WP1.4 Including physical knowledge via emulation of physical radiative transfer models.** Traditional approaches to remote sensing data problems from machine learning are too naïve: out-of-the-shelf algorithms are fed with data and do inferences in the form of predictions and ideally error bars. In our proposed approach, we aim to include physically-meaningful constraints and prior knowledge about the time and spatial scales of feature relations. The approach will deliver a framework in which models will leverage physically-consistent inferences. To do this, we will rely on building *machine learning* models that *emulate* their physical models counterparts. These ML models are known in the RS field as *emulators*. Emulators are essentially function approximation algorithms trained to mimic physical models (commonly referred in the literature as radiative transfer models, RTMs). Emulators are currently capturing much attention because they act as extremely fast surrogates of expensive (in computational terms) RTMs: rather than using memory and CPU expensive physical models or *ad hoc* look-up tables (LUTs), a flexible ML method can replace RTMs efficiently. But, more importantly than just the computational convenience, we advocate that emulators provide us with readily useful non-parametric models that incorporate physical knowledge. We will develop a full toolbox of emulators for common RTMs: (1) for vegetation applications, such as the case of the widely used coupled soil-leaf-canopy model over the solar reflective domain, PROSAIL [57]; and (2) for atmospheric applications we will invert the

[4] Workpackages marked in gray as this one were proposed in the original document but discarded later after the concesion of the project. Most likely we won't have enough human resources to complete them under this project, but we are still interested and we will try to carry out them.

standard Optimal Spectral Sampling (OSS) [58], which is a well-suited method for both RS applications and assimilation of satellite observations in numerical weather prediction models. In both cases, *efficiency* and *accuracy* in the calculation of radiances and Jacobians will be studied. We will consider standard multi-output regression models and the proposed advances, and will resort to closed-form kernel-based solutions for the Jacobians [17].

Work Package 2 (WP2). Efficient implementations.

Earth observation data come in huge quantities, and multimodalities of diverse spatial, spectral and temporal resolutions. The algorithms in GEOLEARN will deal with large quantities of samples that must be processed in operational times. We will tackle this particular 'big data' problem from three complementary directions: (1) redesign the algorithms to make them capable to handle big EO data, (2) using new hardware resources, like multicore CPUs and GPUs, for implementing the algorithms to process big data in operational times, and (3) following divide-and-conquer strategies adapted to the particularities of remote sensing data.

- **WP2.1 Engineering kernel and deep architectures for EO big data.** Despite the good accuracies obtained by kernel methods in EO data processing [38], they have high computational cost in terms of time and memory requirements. The naïve implementation of kernel methods, such as GP regression, requires the inversion of a kernel matrix, which has complexity $\mathcal{O}(n^3)$ in computation and $\mathcal{O}(n^2)$ in storage, where n denotes the number of samples to build the model. This makes their use unfeasible when dealing with a relatively high number of samples (e.g., $n > 10,000$). Most strategies deal with this issue by using approximate kernel functions instead of the exact ones, and are typically based on the Incomplete Cholesky Factorization (ICF) or the Nyström method [59]. Specifically to kernel methods, most approaches are based on approximating the kernel function using a series of *inducing variables* [60], which allows to express the original n input variables as a weighted combination of m inducing variables, where $m \ll n$, reducing notably the computational and storage costs. Other methods to reduce the kernel matrix rank are based on approximating the kernel function using m random basis functions [61]. Extending the same idea, Fastfood kernels [62] replace the kernel function using a combination of random and Walsh-Haddamard matrices, which reduces the computational cost further to $\mathcal{O}(n \log d)$. Although all these approaches work well for any dataset in general, in this project we will exploit the statistical spatial-spectral spectrum of remote sensing data [9]. In particular, we will optimize the set of inducing variables and random features for the particular characteristics of remote sensing data (see WP4). For instance, emulation of RTMs spectra (WP4.1, WP4.2) requires paying more attention to critical (absorption) spectral regions, where model should be more accurate and yield lower predictive variances. We will allocate more inducing variables in those critical regions following physical [17], optimal coding [63], and interpolation [64] approaches. Equivalently, atmospheric profiles (WP4.2) change smoothly, suggesting that using more Fourier basis functions in low frequencies would describe better the data.
- **WP2.2 Adapting the algorithms for multicore CPUs and GPUs.** One of the reasons Deep Neural Networks (DNN) have become so popular in the last years is because of the increase in computational power available nowadays. In particular, the use of multi-core processors and co-processors, together with Graphic Processing Units (GPUs), has reduced the training process to a reasonable amount of time. The aim of this project is to adapt our current and proposed algorithms to make full profit of this multi-core hardware architectures. To this end, we will use of *the facto* standard libraries, such as NVIDIA's CUDA[5], as well as ready-to-use high-level implementations present in Pyhton, such as Theano[6] or MATLAB(tm)[7].
- **WP2.3 Divide-and-conquer strategies for remote sensing data.** Divide-and-conquer strategies can be defined either at an *algorithm level* (related to the approximation techniques described above), or at a *data/high level*. We will consider the second strategy in GEOLEARN. In particular, a direct approach consists in dividing the input dataset in parts and analyze (i.e., develop a model for) each part locally. Despite of being a straightforward methodology,

[5] https://developer.nvidia.com/cuda-zone
[6] http://deeplearning.net/software/theano/
[7] http://es.mathworks.com/products/parallel-computing/

it readily has important advantages, and actually is used for instance in *local* GPs [65] with good results. Among the main advantages of this approach, we should mention (i) local models are small in size, thus faster to train and use in prediction; (ii) local models are accurate in their defined regions, usually better than global models that try to cover all regions; (iii) it is straightforward to obtain a parallelized implementation; and (iv) local models are often simpler and easier to interpret. On the other hand, they have some issues like presenting discontinuities in the boundaries of defined regions, and show low accuracies far from the regions where they have been trained in. In this project, we will design optimal ways to partition remote sensing data to distribute its analysis among several machine learning models. For instance, spatial and time structures in EO data are of crucial importance, and a proper definition of local data batches may make a huge difference.

We want to stress the fact that the three approaches are not exclusive, but rather complementary, and they can and will be used in this project together to further improve processing times and the ability to process more quantities of data.

Work Package 3 (WP3). Extracting information from models.

- **WP3.1 Feature ranking and global sensitivity analysis.** Sensitivity analysis evaluates the relative importance of each input variable and can be used to identify the most influential in determining the variability of model outputs. In *local sensitivity analysis*, also known as 'One-factor-at-a-time' (OAT), one changes one input variable at a time whilst holding all other at their central values. OAT methods do not cover the whole input variable space, so they are inadequate for analyzing complex models, which may have many variables and may be high-dimensional and/or non-linear. On the contrary, *global sensitivity analysis* explores the full input variable space. The contribution of each input variable to the variation in outputs is averaged over the variation of all input variables, i.e. all input variables are changed together. Global sensitivity analysis (GSA) techniques, which quantify the relative importance of each input variable to model outputs, can help setting safe default values for those less influential input variables. GSA can greatly simplify model calibration through enabling the most influential variables to be targeted for data acquisition and refinement. Essentially, GSA: (i) is a useful tool to gain insight into radiative transfer fluxes and model performances; (ii) enables to configure simplified models for retrieval of specific outputs (e.g. sun-induced fluorescence signal); and (iii) constitute a useful tool to identify RTM key and non-influential variables. Depending on the RTM, not only insight in driving variables along spectral domain, but also of fluxes. In this task, we will (1) develop GSA techniques for the predictive mean and variance in GPs for several covariance functions, and study explicit closed-form solutions for the sensitivity of the predictive mean and variance in the GP framework [26]; (2) introduce kernel versions of standard methods for GSA mainly based on estimating Euclidean distances in input spaces [66, 67]; and (3) analyze the sensitivity scores for physical RTM emulators and its optimization [17].
- **WP3.2 Inspecting deep features.** One of the key points in the next years regarding DNNs will be to interpret the model in order to extract information about the studied problem. We will extract information from the learned model by analyzing the learned transformation at each layer. This transformation will tell us how the different input variables are combined and therefore we will obtain information on which input variables should be related and what is the amount of relation between them. This is straightforward in the first layer. For the next layers, we will employ a similar strategy than the uncertainty propagation analysis employed in WP1, i.e. study how variations (i.e., the Jacobian) affect the outputs and the inputs. This will give us numerical values of the relative importance of the information transmitted by the deep filters.
- **WP3.3 Causality.** Establishing causal relations between random variables from empirical data is perhaps the most important challenge in today's Science. In this task, we will explore several pathways to establish causal relations in important geoscience problems. We will work on inferring cause-effect links between random variables from empirical data, given that an interventional framework is obviously not possible in climate science. We will work in non-deterministic, empirical data-driven approaches:

1. *Regression.* We will follow the framework in [68], which exploits nonlinear, non-parametric regression to assess the plausibility of the causal link between two random variables in both forward (predicting y from x) and barckward (x from y) directions: statistically significant residuals in just one direction indicate the true data-generating mechanism. We have recently shown that heteroscedastic and warped GPs can better identify such causal relations, as they 'discount' the signal-dependent noise effects [17, 69]. We will deploy the algorithms developed in previous tasks to extend the framework to multi-dimensional problems (with possibly dependent –but also confounder– co-variates). Accounting for the estimated (eventually tighter) predictive variances might improve identifiability and trustworthiness (application in WP4.2).
2. *Time series.* Canonical causal inference in Hume terms reduces to identify the arrow of time in a set of exogenous time series. Many methods have been proposed for this, being the Granger causal analysis the most well-known approach. The approach typically exploits (linear) auto-regressive models, such as VAR or ARMA. Linearity and Gaussianity are however strong assumptions. In this subtask, we will rely on our kernel-based framework for signal processing [70], and in particular on kernel ARMA modeling in either implicit [71] or explicit reproducing kernel Hilbert spaces [21], to account for richer dynamic structures to identify the direction of the time series (application in WP4.3).
3. *Asymmetries.* Causal inference can be cast as a problem about finding asymmetries in the density function of the effect given the uniform density of the cause variable. Actually, establishing such asymmetries extends to the important issue of independence between the cause and the mechanics that generated the data. Both problems boil down to the challenging (and still unsolved) problem of estimating (eventually conditional) multidimensional densities from a finite number of observations. In this subtask, we plan to approach this with (conditional) density estimation using our multivariate Gaussianization method [72], which allows invertible transformations and explicit calculation of the Jacobian, that may lead to improved identifiability. Applications in WP4.2 and WP4.3 will show the validity of the approach.

Work Package 4 (WP4). Applications for remote sensing and geosciences.

The last technical WP is devoted to the adaptation and application of the previous algorithms in three particularly challenging research domains in Climate Science, in which we have solid expertise: vegetation monitoring in the context of the upcoming Sentinels missions[8] [15, 24, 50], atmospheric variable prediction (temperature and moisture atmospheric profiles and emissivities) in the context of the upcoming MTG-IRS sensor[9] [25, 73], and the upscaling of carbon and heat fluxes from eddy covariance measures and remote sensing data in the context of the FLUXNET activities[10] [26, 74]. We are deeply involved in all these applications through specific projects (ESA, EUMETSAT, consortia), but neither the proposed approaches nor the GEOLEARN goals are addressed therein. In fact, these previous works allowed us identifying urgent needs and plausible improvements on the machine learning techniques adapted to the particularities of EO data. In all three application domains, the key steps of data collection are ensured by our external collaborators and our own databases. We show in Table 1 the methods used in each one of the meta-application domains. In what follows we specify the subtasks of WP4.

Table 1: *Methods and developments that will be used in each metacase study.*

	WP1				WP2	WP3		
	Structure	Multioutput	Uncertainty	Emulate	Efficient	Rank	Deep	Causal
WP4.1. Vegetation	×	√	√	√	√	√	×	√
WP4.2. Atmosphere	√	√	√	√	√	√	√	√
WP4.3. Carbon/heat	√	√	√	×	√	√	×	√

[8] https://sentinel.esa.int
[9] http://esamultimedia.esa.int/docs/MinisterialCouncil/MC-MTG_1811.pdf
[10] http://fluxnet.ornl.gov/

- **WP4.1. Sentinel 2/3 data processing through RTM emulation**. Plant and atmospheric RTMs are currently used in End-to-End simulators that function as a virtual laboratory in the development of new optical sensors, for instance in preparation of the upcoming Sentinels missions. Over the last three decades, a large number of RTMs have been developed with different degrees of complexity, and gradual improvements and increase in complexity have diversified RTMs from simple turbid medium RTMs towards advanced Monte Carlo RTMs that allow for explicit of 3-D representations of complex atmospheric models or canopy architectures. This evolution resulted in an increase in the computational requirements to run the model, and therefore in our ability to invert the model [17, 75]. In this context, we will develop accurate and efficient RTM emulators with machine learning algorithms in preparation of the Sentinel 2/3 data. This calls for the generation of consistent and diverse databases from accurate (but expensive) RTMs. We will focus on the inversion of the PROSAIL RTM, which is the combination of the PROSPECT leaf optical properties model and the SAIL canopy bidirectional reflectance model. Essentially, PROSAIL links the spectral variation of canopy reflectance, which is mainly related to leaf biochemical contents, with its directional variation, which is primarily related to canopy architecture and soil/vegetation contrast. This link is key to simultaneous estimation of canopy biophysical/structural variables for applications in agriculture, plant physiology, and ecology at different scales. PROSAIL has become one of the most popular radiative transfer tools due to its ease of use, robustness, and consistent validation by lab/field/space experiments over the years. In our previous work [18] we already used PROSAIL to generate 1,000,000 pairs of Sentinel-2 spectral (13 channels) and 7 associated parameters. We used random kitchen sinks [61] to do the inversion with different basic structures, but the algorithm suffered in this challenging multi-output regression problem, given the diversity in outputs uncertainty and uneven observation dependencies. We here plan (1) to generate richer RTM data pairs through the use of ancillary phenological models of crop evolution; (2) to exploit the advances in WP1 on structured models and sensible uncertainty estimation; and (3) to improve computational efficiency of algorithms following smart physically-inspired *inducing features* allocation (see WP2.1).
- **WP4.2. Estimation of atmospheric profiles with super-spectral infrared sounders.** We will focus on some of the most important state vectors in climate science: temperature and water vapor are critical atmospheric parameters for weather forecast and atmospheric chemistry studies [76]. Observations from space-borne high spectral resolution infrared sounding instruments can be used to calculate the profiles of such atmospheric parameters with unprecedented accuracy and vertical resolution [77]. EUMETSAT will provide remote sensing and meteorological data for statistical retrieval of temperature and humidity from super-spectral infrared sounders, in which we have large experience. We will use data coming from the Infrared Atmospheric Sounding Interferometer (IASI) that provides radiances in 8461 spectral channels [78]. Its spatial resolution is 25 km at nadir with an Instantaneous Field of View (IFOV) size of 12 km at an altitude of 819 km. The data used in this task will be both real data (IASI and resampled to MTG-IRS resolutions, plus ECMWF re-assimilation profiles) and simulated data, using appropriate RTMs, such as OSS [58]. The huge datasets typically require computationally efficient processing techniques. The sub-tasks here will involve extensive dedication to data collection and harmonization. Then, two main applied objectives will be tackled here: (1) to deliver more accurate predictions of relevant atmospheric state vectors with advanced statistical models; and (2) to develop efficient emulators for the standard OSS RTM. To accomplish these goals, we will rely on algorithms in WP1 and their efficient implementations described in WP2.
- **WP4.3. Estimation of global time-resolved carbon and heat fluxes.** Estimations on the biosphere-atmosphere fluxes at continental and global scales are currently essential for a rapid development of policy responses on climate change. In the last decade, global spatial-temporal fields of FLUXNET derived carbon and energy fluxes are increasingly used for analyzing variations of the global carbon and energy cycles, and to evaluate global land surface models. Model/process-based and data-driven algorithms are the two main approaches to upscale data acquired from flux towers [10, 11]. In the last few years, nevertheless, data-driven statistical learning algorithms have attained outstanding results in the estimation of cli-

mate variables and related bio-geo-physical parameters at local and global scales [9]. These algorithms avoid complicated assumptions and provide flexible nonparametric models that fit the observations using massive heterogeneous data. In a recent work [26] we estimated global flux products derived from upscaling FLUXNET eddy covariance observations using GP models, and also assessed the relative relevance of the remote sensing and meteorological variables. For this we used the global long time series of MODIS satellite data and La Thuile FLUXNET synthesis data set, which is composed of half-hourly FLUXNET eddy covariance measurements processed using standardized procedures of gap-filling and quality control [79, 80]. The fluxes were subsequently aggregated into 8-daily means to conform to the temporal resolution of MODIS products. This is actually an important limitation when trying to understand the processes and inter-relations, as the seasonal variability may be considered as a strong co-founder. In this project we plan to extend this database to situations of increased sampling, thus going for hourly upscaling datasets. Additionally, analysis of relevant variables and causality will also be explored using the methodologies proposed on WP3. Dealing with such huge amount of data, uneven resolutions and scales, and noise sources, constitutes an extremely challenging problem for current prediction algorithms and causal inference approaches. Of course, this is the most risky task in the GEOLEARN project, but at the same time an ideal testbed for our methodological proposals, and if successful, it would make a definite leap towards understanding essential climate variables through automatic reasoning.

Work Package 5 (WP5). Project management and technology transfer.

The GEOLEARN project will involve some managerial as well as technological transfer tasks. The general managerial activities of GEOLEARN will involve: (1) coordination of members and manage all activities in the group, as data collection and algorithm design are mutually dependent; (2) control the overall project schedule as some tasks are also timely; and (3) ensure timeliness of all deliverables (e.g. software packages should be available for further applications) and planned reports. For dissemination, we will design and implement a website/wiki for the GEOLEARN project, as we did before in previous projects, e.g. see `http://isp.uv.es/projects.htm`. Frequent follow-up meetings in the group will be minuted in a three-monthly basis. In summary, the WP5 will imply three main activities:

- **WP5.1 Reports and documentation.** We will generate a bi-annual progress report that puts together the scientific/technical achievements, and summarizes ongoing activities, identified risks and contingency plans/ideas, as well as the dissemination plan. We typically publish pre-print versions of relevant papers in ArXiv (areas: stat.ML, cs.CV, physics.geo-ph), and we plan to continue with this philosophy in green open access.
- **WP5.2 Open software, toolboxes, and harmonized databases.** We will release a number of open source software packages and standardized databases for the sake of reproducibility of the attained results in `http://isp.uv.es/soft.htm`. Code will be also eventually released at `https://github.com/` and `http://mloss.org/`.
- **WP5.3 Special sessions and a workshop.** We plan to organize special sessions in flagship conferences both on remote sensing (e.g. IEEE IGARSS, AGU) and machine learning (NIPS or ICML). In addition, a dedicated small workshop will disseminate the main achievements of the project, trying to get together key scientists in the fields (remote sensing and machine learning), users (cartographic institutes, members of international organizations), and interested stakeholders.

References

[1] IPCC. Intergovernmental Panel on Climate Change. Fourth Assessment Report: Climate Change 2007: The AR4 Synthesis Report, 2007.

[2] F.E.L. Otto. Climate change: Attribution of extreme weather. *Nature Geoscience*, 2015.

[3] M.E. Schaepman, S.L. Ustin, A.J. Plaza, T.H. Painter, J. Verrelst, and S. Liang. Earth system science related imaging spectroscopy-An assessment. *Rem. Sens. Env.*, 113(1):S123–S137, 2009.

[4] 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.*, 120:25–36, 2012.

[5] 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. Envir.*, 120:37–57, 2012.

[6] T. Stuffler, C. Kaufmann, S. Hofer, K.P. Farster, G. Schreier, A. Mueller, A. Eckardt, H. Bach, B. Penné, U. Benz, and R. Haydn. The EnMAP hyperspectral imager-An advanced optical payload for future applications in Earth observation programmes. *Acta Astronautica*, 61(1-6):115–120, 2007.

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

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

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

[10] C. Beer et al. Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. *Science*, 329(834), 2010.

[11] M. Jung et al. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research: Biogeosciences*, 116(G3), 2011.

[12] L. R. Sarker and J. E. Nichol. Improved forest biomass estimates using ALOS AVNIR-2 texture indices. *Rem. Sens. Env.*, 115(4):968–977, 2011.

[13] D. Tuia, J. Verrelst, L. Alonso, F. Pérez-Cruz, and G. Camps-Valls. Multioutput support vector regression for remote sensing biophysical parameter estimation. *IEEE Geosc. Rem. Sens. Lett.*, 8(4):804–808, 2011.

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

[15] J. Verrelst, J. Muñoz, L. Alonso, J. Delegido, J.P. Rivera, G. Camps-Valls, and J. Moreno. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Rem. Sens. Env.*, 118:127–139, 2012.

[16] R. Furfaro, R. D. Morris, A. Kottas, M. Taddy, and B. D. Ganapol. A Gaussian Process Approach to Quantifying the Uncertainty of Vegetation Parameters from Remote Sensing Observations. *AGU Fall Meeting Abstracts*, pages A261+, Dec 2006.

[17] G. Camps-Valls, J. Verrelst, J. Muñoz-Marí, V. Laparra, F. Mateo-Jimenez, and J. Gomez-Dans. A survey on gaussian processes for earth observation data analysis. *IEEE Geoscience and Remote Sensing Magazine*, 2015. Accepted.

[18] V. Laparra, D. Marcos, D. Tuia, and G. Camps-Valls. Large-scale random features for kernel regression. In *IGARSS 2015*, 2015.

[19] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls. Multimodal analysis of remote sensing images: A review and future directions. *Proc. IEEE*, 2014. Invited paper, to appear.

[20] S. Salcedo-Sanz, C. Casanova-Mateo, J. Muñoz-Marí, and G. Camps-Valls. Prediction of daily global solar irradiation using temporal gaussian processes. *Geoscience and Remote Sensing Letters, IEEE*, pages 1–5, 2014.

[21] D. Tuia, J. Muñoz Marí, J.L. Rojo-Álvarez, M. Martínez-Ramón, and G. Camps-Valls. Explicit recursive and adaptive filtering in reproducing kernel hilbert spaces. *Neural Networks and Learning Systems, IEEE Transactions on*, 25(7):1413–1419, July 2014.

[22] J.L. Rojo-Álvarez, M. Martínez-Ramon, J. Muñoz Marí, and G Camps-Valls. *Digital Signal Processing with Kernel Methods*. Wiley & Sons, 2015.

[23] GCOS. Systematic observation requirements for satellite-based products for climate, 2011, 2011.

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

[25] G. Camps-Valls, J. Muñoz-Marí and, L. Gómez-Chova, L. Guanter, and X. Calbet. Nonlinear statistical retrieval of atmospheric profiles from MetOp-IASI and MTG-IRS infrared sounding data. *IEEE Trans. Geosc. Rem. Sens.*, 50(5):1759–1769, 2012.

[26] G. Camps-Valls, M. Jung, K. Ichii, D. Papale, G. Tramontana, P. Bodesheim, C. Schwalm, J. Zscheischler, M. Mahecha, and M. Reichstein. Ranking drivers of global carbon and energy fluxes over land. In *IEEE IGARSS*, 2015.

[27] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno. Retrieval of vegetation parameters using Gaussian processes techniques. *IEEE Trans. Geosc. Rem. Sens.*, 49:1832–1843, 2012.

[28] C Peter, W de Lange, JK Musango, K. April, and A Potgieter. Applying Bayesian modelling to assess climate change effects on biofuel production. *Clim Res*, 40:249–260, 2009.

[29] I. Ebert-Uphoff and Y. Deng. Causal discovery for climate research using graphical models. *J. Climate*, 25:5648–5665, 2012.

[30] A. Hannart, J. Pearl, F.E.L. Otto, P. Naveau, and M. Ghil. Causal counterfactual theory for the attribution of weather and climate-related events. *Bulletin of the American Meteorological Society*, 2015.

[31] M. Meroni, M. Rossini, L. Guanter, L. Alonso, U. Rascher, R. Colombo, and J. Moreno. Remote sensing of solar-induced chlorophyll fluorescence: Review of methods and applications. *Rem. Sens. Envir.*, 113(10):2037–2051, 2009.

[32] Frankenberg et al. New global observations of the terrestrial carbon cycle from gosat: Patterns of plant fluorescence with gross primary productivity. *Geophysical Research Letters*, 38(17), 2011.

[33] L. Guanter et al. Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence. *Proc. National Academy of Sciences, PNAS*, 2014.

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

[35] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521, 2015.

[36] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla. Composite kernels for hyperspectral image classification. *IEEE Geosc. Rem. Sens. Lett.*, 3(1):93–97, 2006.

[37] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Luis Rojo-Álvarez, and M. Martínez-Ramón. Kernel-based framework for multi-temporal and multi-source remote sensing data classification and change detection. *IEEE Trans. Geosc. Rem. Sens.*, 46(6):1822–1835, 2008.

[38] G. Camps-Valls and L. Bruzzone, editors. *Kernel methods for Remote Sensing Data Analysis*. Wiley & Sons, UK, Dec 2009.

[39] J. Leiva, L. Gómez-Chova, and G. Camps-Valls. Multitask remote sensing data classification. *IEEE Trans. Geosc. Rem. Sens.*, 50, Oct 2012.

[40] M. Camps-Taberner, A. Romero, C. Gatta, and G. Camps-Valls. Shared feature representations of lidar and optical images: trading sparsity for semantic discrimination. In *IEEE Workshop on Hyperspectral Image and Signal Processing, Whispers 2014*, Milano, Italy, 2015. IEEE Press. Winner of the 2015 IEEE GRSS Data Fusion Contest.

[41] G. Camps-Valls, Marcelino Martínez-Sober, Emilio Soria-Olivas, Rafael Magdalena-Benedito, Javier Calpe-Maravilla, and Juan Guerrero-Martínez. Foetal ECG recovery using dynamic neural networks. *Artificial Intelligence in Medicine*, 31(3):197 – 209, 2004.

[42] Hachem Kadri, Mohammad Ghavamzadeh, and Philippe Preux. A generalized kernel approach to structured output learning. In *ICML*, volume 28 of *JMLR Proceedings*, pages 471–479. JMLR.org, 2013.

[43] R. H. Moss, J.A. Edmonds, K. A. Hibbard, M.R. Manning, S.K. Rose, D.P. van Vuuren, T.R. Carter, S. Emori, M. Kainuma, T. Kram, G.A. Meehl, J.F.B. Mitchell, N. Nakicenovic, K. Riahi, S.J. Smith, R.J. Stouffer, A.M. Thomson, J.P. Weyant, and T.J. Wilbanks. The next generation of scenarios for climate change research and assessment. *Nature*, 463:747–756, 2010.

[44] D. Tuia, J. Muñoz-Marí, Mikhail F. Kanevski, and G. Camps-Valls. Structured output

SVM for remote sensing image classification. *Signal Processing Systems*, 65(3):301–310, 2011.

[45] A. G. Wilson, D. A. Knowles, and Z. Ghahramani. Gaussian process regression networks. In J. Langford and J. Pineau, editors, *ICML*, Edinburgh, June 2012. Omnipress.

[46] J. Arenas-García and K.B. Petersen. *Kernel Methods for Remote Sensing Data Analysis*, chapter Kernel Multivariate Analysis in Remote Sensing Feature Extraction, pages 329–352. J. Wiley & Sons Inc., UK, 2009.

[47] Anthony O'Hagan. Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux. *Environmental Modelling & Software*, 36(0):35 – 48, 2012. Thematic issue on Expert Opinion in Environmental Modelling and Management.

[48] J. Caers, editor. *Modeling Uncertainty in the Earth Sciences*. Wiley-Blackwell, UK, June 2011.

[49] J. Verrelst, L. Alonso, J. P. Rivera, J. Moreno, and G. Camps-Valls. Gaussian Process Retrieval of Chlorophyll Content from Imaging Spectroscopy Data. *IEEE JSTARS*, 6(2):867–874, Apr 2013.

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

[51] Grégoire Montavon, Mikio L. Braun, Tammo Krueger, and Klaus-Robert Müller. Analyzing local structure in kernel-based learning: Explanation, complexity and reliability assessment. *Signal Processing Magazine, IEEE*, 30(4):62–74, 2013.

[52] E. Izquierdo-Verdiguier, R. Jenssen, Luis Gómez-Chova, and G. Camps-Valls. Spectral clustering with the probabilistic cluster kernel. *Neurocomputing*, 149, Part C(0):1299 – 1304, 2015.

[53] Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Divide and conquer kernel ridge regression. In *COLT*, pages 592–617, 2013.

[54] M.P. Deisenroth and J.W. Ng. Distributed gaussian processes. In *ICML*, pages 1481–1490, 2015.

[55] A. Romero, C. Gatta, and G. Camps-Valls. Unsupervised deep feature extraction for remote sensing data analysis. *IEEE Trans. Geosc. Rem. Sens.*, 2015.

[56] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167, New York, NY, USA, 2008. ACM.

[57] S. Jacquemoud, W. Verhoef, F. Baret, C. Bacour, P.J. Zarco-Tejada, G.P. Asner, C. François, and S.L. Ustin. PROSPECT + SAIL models: A review of use for vegetation characterization. *Rem. Sens. Env.*, 113(SUPPL. 1):S56–S66, 2009.

[58] J. Moncet, G. Uymin, A. E. Lipton, and H. E. Snell. Infrared radiance modeling by optimal spectral sampling. *J. Atmos. Sci.*, 65:3917–3934, 2008.

[59] P. Drineas and M.W. Mahoney. On the Nÿstrom method for approximating a gram matrix for improved kernel-based learning. *JMLR*, 6:2153–2175, 2005.

[60] J. Quinonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *JMLR*, 6:1939–1959, 2005.

[61] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1313–1320. Curran Associates, Inc., 2009.

[62] Q. Le, T. Sarlos, and A. Smola. Fastfood - approximating kernel expansions in loglinear time. In *ICML*, 2013.

[63] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Press, Boston, 1992.

[64] A.G. Wilson and H. Nickisch. Kernel interpolation for scalable structured gaussian processes (KISS-GP). In *ICML*, 2015.

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

[66] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and

S. Tarantola. *Global Sensitivity Analysis. The Primer*. John Wiley & Sons, Ltd, 2008.

[67] I.M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1–3):271 – 280, 2001. The Second IMACS Seminar on Monte Carlo Methods.

[68] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NIPS*, pages 689–696, 2008.

[69] L. Gómez-Chova and G. Camps-Valls. Learning with the kernel signal-to-noise ratio. In *IEEE MLSP'12*, Santander, Spain, September 2012.

[70] J.L. Rojo-Álvarez, M. Martínez-Ramón, J. Muñoz Marí, and G. Camps-Valls. A unified SVM framework for signal estimation. *Digit. Signal Process.*, 26:1–20, March 2014.

[71] M. Martínez-Ramón, José Luis Rojo-Álvarez, G. Camps-Valls, J. Muñoz-Marí, Angel Navia-Vazquez, Emilio Soria-Olivas, and Aníbal Ramón Figueiras-Vidal. Support vector machines for nonlinear kernel ARMA system identification. *IEEE Trans. Neur. Netw.*, 17(6):1617–1622, November 2006.

[72] V. Laparra, G. Camps, and J. Malo. Iterative gaussianization: from ICA to random rotations. *IEEE Trans. Neur. Nets.*, 22(4):537–549, 2011.

[73] X. Calbet and P. Schlüssel. Analytical estimation of the optimal parameters for the EOF retrievals of the IASI Level 2 product processing facility and its application using AIRS and ECMWF data. *Atmos. Chem. Phys.*, 6:831–846, 2005.

[74] G. Tramontana, K. Ichii, G. Camps-Valls, E. Tomelleri, and D. Papale. Uncertainty analysis of gross primary production predictions using random forests, remote sensing and eddy covariance data. *Rem. Sens. Envir.*, 2015. Accepted.

[75] J.P. Rivera, J. Verrelst, J. Gómez-Dans, J. Muñoz Marí, J. Moreno, and G. Camps-Valls. An emulator toolbox to approximate radiative transfer models with statistical learning. *Remote Sensing*, 2015. Accepted.

[76] K. N. Liou. *An Introduction to Atmospheric Radiation*. Academic Press, Hampton, USA, second edition, 2002.

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

[78] G. Chalon, F. Cayla, and D. Diebel. IASI: an advanced sounder for operational meteorology. In *Proc. 52nd Congress of IAF*, Toulouse, France, 1-5 October 2001, 2001.

[79] A.M. Moffat et al. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, 147(3-4):209–232, 2007.

[80] D. Papale, M. Reichstein, M. Aubinet, E. Canfora, C. Bernhofer, W. Kutsch, B. Longdoz, S. Rambal, R. Valentini, T. Vesala, and D. Yakir. Towards a standardized processing of net ecosystem exchange measured with eddy covariance technique: algorithms and uncertainty estimation. *Biogeosciences*, 3(4):571–583, 2006.