Chapter 2
Machine Learning Methods for Spatial and Temporal Parameter Estimation

Álvaro Moreno-Martínez, María Piles, Jordi Muñoz-Marí, Manuel Campos-Taberner, Jose E. Adsuara, Anna Mateo, Adrián Perez-Suay, Francisco Javier García-Haro and Gustau Camps-Valls

Abstract Monitoring vegetation with satellite remote sensing is of paramount relevance to understand the status and health of our planet. Accurate and constant monitoring of the biosphere has large societal, economical, and environmental implications, given the increasing demand of biofuels and food by the world population. The current democratization of machine learning, big data, and high processing capabilities allow us to take such endeavor in a decisive manner. This chapter proposes three novel machine learning approaches to exploit spatial, temporal, multi-sensor, and large-scale data characteristics. We show (1) the application of multi-output Gaussian processes for gap-filling time series of soil moisture retrievals from three spaceborne sensors; (2) a new kernel distribution regression model that exploits multiple observations and higher order relations to estimate county-level crop yield from time series of vegetation optical depth; and finally (3) we show the combination of radiative transfer models with random forests to estimate leaf area index, fraction of absorbed photosynthetically active radiation, fraction vegetation cover, and canopy water content at global scale from long-term time series of multispectral data exploiting the Google Earth Engine cloud processing capabilities. The approaches demonstrate that machine learning algorithms can ingest and process multi-sensor data and provide accurate estimates of key parameters for vegetation monitoring.

Á. Moreno-Martínez, M. Piles, J. Muñoz-Marí, M. Campos-Taberner, J. E. Adsuara, G. Camps-Valls—Authors contributed equally.

Á. Moreno-Martínez (✉) · M. Piles · J. Muñoz-Marí · J. E. Adsuara · A. Mateo · A. Perez-Suay · G. Camps-Valls
Image Processing Laboratory (IPL), Universitat de València, Valencia, Spain
e-mail: alvaro.moreno@uv.es

G. Camps-Valls
e-mail: gustau.camps@uv.es

M. Campos-Taberner · F. Javier García-Haro
Department of Earth Physics and Thermodynamics, Universitat de València, Valencia, Spain
e-mail: manuel.campos@uv.es

S. Prasad and J. Chanussot (eds.), *Hyperspectral Image Analysis*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-38617-7_2

2.1 Introduction

2.1.1 Remote Sensing as a Diagnostic Tool

The Earth is a complex, dynamic, and networked system, and this system is under pressure and in continuous change. Population is increasingly demanding more food and biofuels, at a faster pace, worldwide. Consequently, monitoring the planet in a spatially explicit and timely resolved manner is an urgent need to address important societal, environmental, and economical questions. This is exactly the main goal of Earth Observation (EO) from space, and current satellite sensors operating in different bands of the electromagnetic spectrum help in this challenge as accurate diagnostic tools.

The analysis of the acquired sensor data can be done either at local or global scales by looking at biogeochemical cycles, atmospheric situations, and vegetation dynamics [1–5]. All these complex interactions are studied through the definition of bio-geophysical parameters, either representing different properties for land (e.g., surface temperature, soil moisture, crop yield, defoliation, biomass, leaf area coverage), water (e.g., yellow substance, ocean color, suspended matter, or chlorophyll concentration), or the atmosphere (e.g., temperature, moisture, or trace gases). Every single application considers the specific knowledge about the physical, chemical, and biological processes involved, such as energy balance, evapotranspiration, or photosynthesis.

However, remotely sensed observations only sample the energy reflected or emitted by the surface and thus, an intermediate modeling step is necessary to transform the measurements into estimations of the biophysical parameters [6]. From a pure statistics standpoint, this is considered to be as an *inverse modeling* problem, because we have access to observations generated by the system and we are interested in the unknown parameters that generated those. A series of international study projections, such as the International Geosphere-Biosphere Programme (IGBP), the World Climate Research Programme (WCRP), and the National Aeronautics and Space Administration (NASA) Earth Observing System (EOS), established remote sensing model inversion as one of the most important problems to be solved with EO imagery in the near future.

2.1.2 Data and Model Challenges

Current EO, however, faces two very important challenges that we hereby define as the *data problem* and the *model problem*:

- *The data problem*: The data involved in EO applications is *big, diverse, and unstructured*. We often deal with remote sensing data acquired by many satellite sensors working with different and ever-increasing spatial, temporal, and ver-

tical resolutions. Not to mention that data may also come from high-resolution simulations and re-analysis. At the same time, data is heterogeneous and covers space and time with uneven resolutions, different footprints, signal and noise levels, and feature characteristics. EO applications on land monitoring have mainly considered optical sensors, like the **NASA A-Train** (http://atrain.nasa.gov/) satellite constellations including MODIS and Landsat, and recently the European Space Agency (ESA) Sentinels 2–3 sensors. More recently, sensors operating in the microwave range of the spectrum were introduced. Unlike optically based technologies, microwaves are not affected by atmospheric conditions, and a total coverage of the Earth's surface is obtained every 2–3 days. Microwave radiometry is optimal for sensing the water content in soils and vegetation, but the passive measurement is presently limited in spatial resolution by the size of the instrument antenna aperture to ∼25 km (e.g., ESA's SMOS, NASA's SMAP). Active microwave remote sensing can overcome this limitation but often it is accompanied by constraints on spatial coverage and temporal data refresh rate and require complex scattering models for inversion of geophysical parameters (e.g., ESA's Sentinel 1). Optical sensing technology, in turn, is at a maturity level today that allows providing very fine spatial resolution on a weekly basis (e.g., ESA's Sentinel 2). Undoubtedly, the combination of satellite-based microwave and optical sensory data offers an unprecedented opportunity to obtain a unique view of the Earth system processes.

- *The model problem*: Dealing with such data characteristics and big data influx requires (semi)automatic processing techniques that should be accurate, robust, reliable, and fast. Over the last few decades, a wide diversity of bio-geophysical retrieval methods have been developed, but only a few of them made it into operational processing chains. Lately, machine learning has attained outstanding results in the estimation of climate variables and related bio-geophysical parameters at local and global scales [1]: leaf area index (LAI) [7] and Gross Primary Production (GPP) [8–11] are currently derived with neural networks, kernel methods, and random forests, while multiple regression is used for retrieving biomass [12], support vector methods were also proposed to derive vegetation parameters [13, 14], and kernel methods and Gaussian processes (GPs) [15] have been paid wide attention in the last years in deriving vegetation properties [16]. However, it is important to observe here that, very often, these methods are applied blindly, without being adapted to the data specificities. On the one hand, data exhibits clear spatial and temporal structures that could be useful to design new kernel functions in GPs [17] or rely on convolutional networks [18]. On the other hand, data from different sensors should be synergistically combined in the model, but this is often done via *ad hoc* data re-sampling or statistics summarization, as a convenient way to data preparation for the algorithm. These practices are far from being optimal, and a lot is yet to be done in the algorithm development arena to improve algorithms that respect data characteristics, learn structures from data, fuse heterogeneous multisensor and multi-resolution data naturally, and scale well to big data volumes.

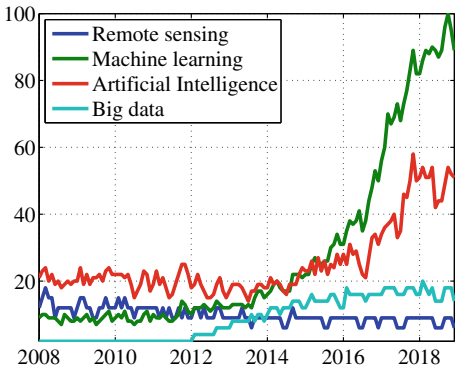

Fig. 2.1 Normalized worldwide interest (i.e., popularity) of terms "remote sensing", "machine learning", "artificial intelligence", and "big data" in the last decade, as measured in Google trends©

Tackling the two sides of the EO challenge is nowadays possible. The current popularization of machine learning, big data, and high processing capabilities allows us to take such an endeavor in a decisive manner, cf. Fig. 2.1.

Nowadays, both data and algorithms are mostly freely available, while large-scale data processing platforms, clusters, and infrastructures are accessible to everyone:

- Machine learning code is now ready to (re)use in different forms: from excellent packages and frameworks like scikit-learn or TensorFlow, to open accessible repositories and developer's platforms like GitHub.
- Earth observation data is also currently accessible through the main space agencies hubs: for example, ESA provides Sentinels data through the ESA open access hub, and NASA grants access via its NASA open data portal.

This unprecedented situation has sowed the seed for the development of applications and the creation of EO-centered companies. Google allows not only accessing but also processing data through the Google Earth Engine, which will be subject of study in this chapter (cf. Sect. 2.4), Descartes Labs offers an EO data processing facility in the cloud, and an increasing number of SMEs has grown around and created what is called the "EO exploitation ecosystem". Altogether, they have allowed tackling problems that were unthinkable just a decade ago.

Earth observation through remote sensing offers great opportunities to monitor our planet by the estimation of key parameters of the land, ocean, and atmosphere. The combined action of machine learning, big data, and high-performance computing platforms, like the Google Earth Engine (GEE), is currently paving the way toward this goal.

2.1.3 Goals and Outline

In this chapter, we will focus on modern machine learning methods for deriving land parameters (e.g., about the vegetation status and crop production) from remote sensing data: we will introduce three recent ML developments that can deal with multi-sensor and multi-resolution data, that exploit nonlinear feature relations and higher order moments of data (observational) distributions, and that can be implemented in the Google cloud platform to derive global maps of parameters of interest. We will mainly focus on new kernel methods, Gaussian processes, and random forests, which fulfill the needs of the field: mathematical tractability and big data scalability, respectively.

We will treat three main problems with different particular data characteristics:

- *Non-uniform temporal sampling and sensor fusion*: First, we will focus on problems of interpolating remote sensing parameters when several variables are available and heavy non-uniform sampling is present. This is a common problem when trying to fuse information from different sensors or in optical remote sensing due to the presence of clouds. Microwave remote sensing is not affected by clouds, but measurements can also be limited in some regions due to combined effects of Radio Frequency Interferences (RFIs), presence of snow, dense vegetation canopies, and high topography [19]; since these effects are sensor- and frequency-dependent, the optimal blend of available microwave-based soil moisture products holds great promise, particularly for observational climate data records [20]. In Sect. 2.2, we will show the exploitation of multi-output Gaussian processes to fill in the temporal gaps in satellite-based estimates of soil moisture from SMOS (L-band passive), AMSR2 (C-band passive), and ASCAT (C-band active) [21, 22]. The method will allow to treat non-uniform sampling and "transfer information across sensors" when samples are missing.
- *Non-uniform spatial sampling*: In remote sensing and geospatial applications, we often encounter problems where one aims to spatialize a variable of interest from a sparse set of measurements, while having access to a finer grid of observations. This is the case of non-uniform spatial sampling. This mismatch in quantity and location is typically resolved by summarizing (e.g., averaging) the observations and co-locating them with the measure. This procedure is ad hoc and suboptimal. In Sect. 2.3, we introduce a new kernel distribution regression model that exploits multiple observations to estimate county-level yield of major crops (wheat, corn, and soybean) from SMAP-based vegetation optical depth (VOD) time series [23]. The method exploits all the available observations and their feature relations.
- *Uniform spatial–temporal data spatialization*: Finally, we deal in Sect. 2.4 with the exploitation of big data in the cloud by spatializing vegetation parameters of interest when long time series of data are available. We will show the combination of radiative transfer models (RTMs) with random forests to estimate various vegetation parameters, namely, LAI, Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Fraction Vegetation Cover (FVC), and Canopy water content (CWC), globally from long-term time series of MODIS data exploiting the GEE.

The platform will allow us to generate products of almost any variable of interest modeled in an RTM [24, 25].

We conclude in Sect. 2.5 with some remarks and an outline of future work. The approaches demonstrate that machine learning algorithms can ingest and process multi-sensor data and provide accurate estimates of key parameters for vegetation monitoring.

2.2 Gap Filling and Multi-sensor Fusion

Measurements of soil moisture (SM) are needed for a better global understanding of the land surface-climate feedbacks at both local and global scales. Satellite sensors operating in the low-frequency microwave spectrum (from 1 to 10 GHz) have proven to be suitable for soil moisture retrievals. These sensors now cover nearly 4 decades, thus allowing for global multi-mission climate data records. The ESA Climate Change Initiative (CCI) soil moisture product combines various single-sensor active and passive microwave soil moisture products into three harmonized products: an only-active, an only-passive, and a combined active–passive microwave product [26]. In its current version, the presence of data gaps in time and space has been acknowledged as a major shortcoming which makes it difficult for users to integrate the data in their applications [20]. From a scientific perspective, the presence of "intermittent" data gaps in satellite-based soil moisture estimates impacts the analysis of spatiotemporal dynamics and trends, which may be limited to certain regions [27]. Also, the presence of missing data in time series prevents a robust computation of temporal autocorrelation and *e*-folding times, as a measure of soil moisture persistence [22]. In this regard, recent studies on the use of Gaussian process regression techniques to mitigate the effect of missing information in Earth observation data are very promising (e.g., [17, 21]).

The presence of gaps in EO data limits their applicability in a number of applications. In contrast with the standard temporal interpolation techniques, the LMC multi-output GP-based gap-filling regression allows taking into account information from other collocated sensors measuring the exact same variable. The method learns the relationships among the different sensors and builds a cross-domain kernel function able to transfer information across the time series and do predictions and associated confidence intervals on regions where no data are available.

In this section, a subset of 6 years of SMOS L-band passive, ASCAT C-band active, and AMSR2 C-band passive soil moisture measurements, starting in June 2010, have been used. SMOS and ASCAT estimates are available for the whole period, whereas AMSR2 estimates start on May 18, 2012 (its launch date). Each

product presents different observational gaps due to the presence of RFI at their operating frequency or a too high uncertainty in their inversion algorithm (e.g., due to the presence of snow masking observations, dense vegetation, or high topography). The problem we face here is that we need a gap-filling methodology able to handle several outputs together and force a "sharp" reconstruction of the time series so that fast dry-down and wetting-up dynamics are preserved (avoid smoothing). We show how we can efficiently deal with our problem by employing a multi-output Gaussian Process model based on the Linear Model of Corregionalization (LMC) [28]. This model implicitly exploits the relationships among the three microwave sensors and predicts an output for each of them. The reconstructed time series are provided with an estimate of its uncertainty and are shown to preserve the statistics from comparison to in situ data over a selection of catchments from the International Soil Moisture Network.

2.2.1 Proposed Approach

The presence of temporal data gaps in satellite-based estimates of soil moisture limits their applicability in a number of applications that need continuous estimates. Standard techniques for gap-filling temporal series such as linear or cubic interpolation, or auto-regressive functions fail to reconstruct sharp transitions or long data gaps and do not take into account information from other collocated sensors measuring exactly the same biophysical variable. Given that we have three different soil moisture products presenting no data in different time and space locations, we employ here an LMC multi-output GP regression (LMC-GP) to maximize the spatiotemporal coverage of the datasets. We illustrate the procedure at three in situ soil moisture networks where the SMOS satellite presents good, average, and poor temporal coverage, see Fig. 2.2. We will show how LMC-GP exploits the relationships among SMOS, ASCAT, and AMSR2 soil moisture time series to do inferences on regions where no data (gaps) are available, and provides a reconstructed prediction with and associated uncertainty for each dataset. Statistical scores from comparison with in situ data at the selected sites of the original and reconstructed time series will be shown.

2.2.2 LMC-GP

First, we will start introducing the formulation of standard GP models. Then, we will extend it to the LMC-GP model.

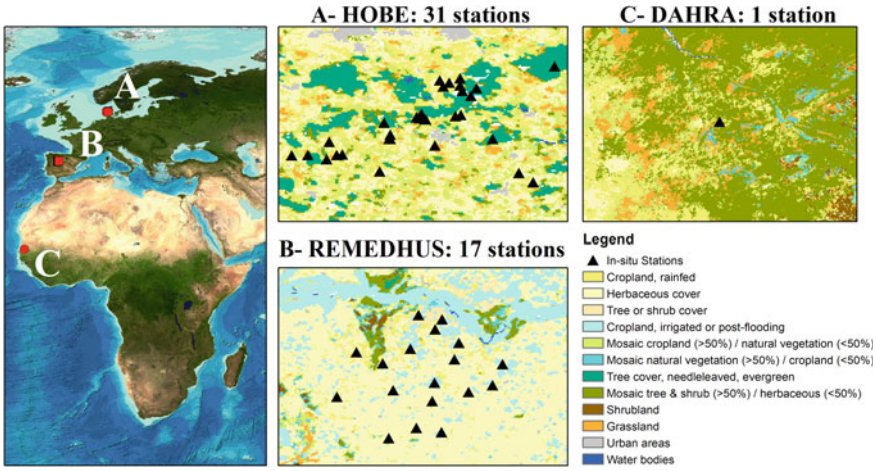

Fig. 2.2 Location and land use map of the three International Soil Moisture Network (ISMN) validation sites used in the study: **a** HOBE in Denmark (31 stations), **b** REMEDHUS in Spain (17 stations), and **C** DAHRA in Senegal (1 station)

2.2.2.1 Gaussian Processes

GPs [15] are state-of-the-art statistical methods for regression and function approximation, and have been used with great success in biophysical variable retrieval by following statistical and hybrid approaches [29]. We start assuming that we are given a set of n pairs of measurements, $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i$ is the feature/measurement space and y_i is the biophysical parameter from field data or other sources, perturbed by an additive independent noise e_i. We consider the following model:

$$y_i = f(\mathbf{x}_i) + e_i, \quad e_i \sim \mathcal{N}(0, \sigma_n^2), \tag{2.1}$$

where $f(\mathbf{x})$ is an unknown latent function, $\mathbf{x} \in \mathbb{R}^d$, and σ_n^2 represents the noise variance. Defining $\mathbf{y} = [y_1, \ldots, y_n]^\intercal$ and $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\intercal$, the conditional distribution of $\mathbf{y}$ given $\mathbf{f}$ becomes $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma_n^2 \mathbf{I}_n)$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. It is assumed that $\mathbf{f}$ follows a n-dimensional Gaussian distribution $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$. The covariance matrix $\mathbf{K}$ of this distribution is determined by a squared exponential (SE) kernel function with entries $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2))$, encoding the similarity between input points. In order to make a new prediction y_* given an input $\mathbf{x}_*$, we obtain the joint distribution over the training and test points,

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{C_n} & \mathbf{k}_*^\intercal \\ \mathbf{k}_* & c_* \end{bmatrix}\right),$$

where $\boldsymbol{C_n} = \mathbf{K} + \sigma_n^2 \mathbf{I}_n, \mathbf{k}_* = [k(\mathbf{x}_*, \mathbf{x}_1), \ldots, k(\mathbf{x}_*, \mathbf{x}_n)]^\intercal$ is an $n \times 1$ vector and $c_* = k(\mathbf{x}_*, \mathbf{x}_*) + \sigma_n^2$. Using the standard Bayesian framework, we obtain the distribution

over y_* conditioned on the training data, which is a normal distribution with predictive mean and variance given by

$$\begin{aligned}\mu_{\mathrm{GP}}(\mathbf{x}_*) &= \mathbf{k}_*^\intercal(\mathbf{K}+\sigma_n^2\mathbf{I}_n)^{-1}\mathbf{y},\\ \sigma^2_{\mathrm{GP}}(\mathbf{x}_*) &= c_* - \mathbf{k}_*^\intercal(\mathbf{K}+\sigma_n^2\mathbf{I}_n)^{-1}\mathbf{k}_*.\end{aligned} \tag{2.2}$$

One of the most interesting things about GPs is that they yield not only predictions $\mu_{\mathrm{GP}*}$ for test data, but also the uncertainty of the mean prediction, $\sigma_{\mathrm{GP}*}$. Model hyperparameters $\boldsymbol{\theta} = (\sigma, \sigma_n)$ determine, respectively, the width of the SE kernel function and the noise on the observations, and they are usually obtained by maximizing the log-marginal likelihood.

2.2.2.2 Linear Model of Corregionalization for GPs

LMC-GPs [28] extend standard GPs so it is possible to both handle several outputs at the same time (i.e., it is a multi-output model) and to deal with missing data in the considered outputs. This model is well known in the field of geostatistics as *co-kriging* [30].

In the LMC-GP model, we have a vector function, $\mathbf{f} : \mathcal{X} \to \mathbb{R}^D$, where D is the number of outputs. Given a reproducing kernel, defined as a positive definite symmetric function $\mathbf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{n\times n}$, where n is the number of samples of each output, we can express $\mathbf{f}(\mathbf{x})$ as

$$\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{n} \mathbf{K}(\mathbf{x}_i, \mathbf{x})\mathbf{c}_i, \tag{2.3}$$

for some coefficients $\mathbf{c}_i \in \mathbb{R}^n$. The coefficients $\mathbf{c}_i$ can be obtained by solving the linear system, obtaining

$$\bar{\mathbf{c}} = (\mathcal{K}(\mathbf{X}, \mathbf{X}) + \lambda n\mathbf{I})^{-1}\bar{\mathbf{y}}, \tag{2.4}$$

where $\bar{\mathbf{c}}, \bar{\mathbf{y}}$ are nD vectors obtained by concatenating the coefficients and outputs, respectively, and $\mathcal{K}(\mathbf{X}, \mathbf{X})$ is an $nD \times nD$ matrix with entries $(\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j))_{d,d'}$ for $i, j = 1, \ldots, n$ and $d, d' = 1, \ldots, D$. The blocks of this matrix are $(\mathbf{K}(\mathbf{X}_i, \mathbf{X}_j))_{i,j}$ $n \times n$ matrices. Predictions are given by

$$\mathbf{f}(\mathbf{x}_*) = \mathcal{K}_{\mathbf{x}_*}^\top \bar{\mathbf{c}}, \tag{2.5}$$

with $\mathcal{K}_{x_*} \in \mathbb{R}^{D\times nD}$ composed of blocks $(\mathbf{K}(\mathbf{x}_*, \mathbf{x}_j))_{d,d'}$. When the training kernel matrix $\mathcal{K}(\mathbf{X}, \mathbf{X})$ is block diagonal, that is, $(\mathbf{K}(\mathbf{X}_i, \mathbf{X}_j))_{i,j} = \mathbf{0}$ for all $i \neq j$, then each output is considered to be independent of the others, and we thus have individual GP models. The non-diagonal matrices establish the relationships between the outputs.

In the LMC-GP model, each output is expressed as a linear combination of independent latent functions,

$$f_d(\mathbf{x}) = \sum_{q=1}^{Q} a_{d,q} u_q(\mathbf{x}), \tag{2.6}$$

where $a_{d,q}$ are scalar coefficients, and $u_q(\mathbf{x})$ are latent functions with zero mean and covariance $k_q(\mathbf{x}, \mathbf{x}')$. It can be shown [28] that the full covariance matrix of this model can be expressed as

$$\mathcal{K}(\mathbf{X}, \mathbf{X}) = \sum_{q=1}^{Q} \mathbf{B}_q \otimes k_q(\mathbf{X}, \mathbf{X}), \tag{2.7}$$

where $\otimes$ is the Kronecker product. Here, each $\mathbf{B}_q \in R^{D \times D}$ is a positive definite matrix known as a *co-regionalization matrix*, and it encodes the relationships between the outputs.

2.2.3 Data and Setup

The temporal period of study is 6 years, starting in June 2010. Three global satellite soil moisture products have been extracted for the study period: SMOS BEC L3 (1.4 GHz, L3 SM v3.0), Metop A/B ASCAT (5.3 GHz, Eumetsat H-SAF), and GCOM W1 AMSR2 L3 (6.9 GHz, LPRM v05 retrieval algorithm, NASA). ASCAT and AMSR2 products have been resampled from their 0.25° grid to the SMOS EASE2 25-km grid using bilinear interpolation. These products have been widely validated under different biomes and climate conditions by comparison with ground-based observations (e.g., [26, 31, 32]) and outputs of land surface models (e.g., [33–35]).

We show the robustness of the multi-sensor gap-filling approach at three in situ soil moisture networks: REMEDHUS in Spain (17 stations [36]), HOBE in Denmark (31 stations [37]), and DAHRA in Senegal (1 station [38]). In terms of temporal coverage, they are representative of best-case (REMEDHUS), average-case (HOBE), and wort-case (DAHRA) scenarios, with SMOS providing a coverage during the study period of 96, 65, and 45%, respectively. The locations and land use maps of the in situ networks used for this study are presented in Fig. 2.2.

2.2.4 Results

Let us start with an illustrative example of method's performance. Figure 2.3 shows with a real example how the LMC-GP transfers information across SMOS, ASCAT, and AMSR2 satellite time series for the predictions when no data is available and provides associated confidence intervals.

Fig. 2.3 Time series of original (orange dots) and reconstructed (blue lines) SMOS, ASCAT, and AMSR2 time series using the LMC-GP gap-filling technique. The uncertainty on the predictions is shown in shaded gray. The orange square points out a specific rainfall event that was captured only by SMOS and is accounted for in the reconstruction of ASCAT and AMSR2 time series. The green square exemplifies how the method reconstructs long data gaps in AMSR2 based on no-rain information from the other two sensors, assigning a higher uncertainty when no original data is available

A more thorough experimental analysis follows. Results of the application of the proposed LMC-GP over REMEDHUS, HOBE, and DAHRA networks are shown in Fig. 2.4, together with the original satellite time series and the in situ data as a benchmark. It can be seen that the reconstructed soil moisture time series follow closely the original time series, capturing the wetting-up and drying-down events and filling the missing information (e.g., see in HOBE the dry-down in February 2014 which was captured only by AMSR2 during consecutive days and is reproduced by the three reconstructed time series). In DAHRA, the limited temporal coverage of AMSR2 in the dry seasons is completed in the reconstructed time series using information from the other two sensors. It is worth to remark that for AMSR2 the reconstructed time series back-propagate to dates where the satellite was not yet launched (shown here for illustration purposes), yet they look very consistent with the real satellite data. Also importantly, we fixed the kernel lengthscale parameter in LMC-GP model to force a sharp reconstruction, to prevent the predictions being smoothed with respect to the original time series.

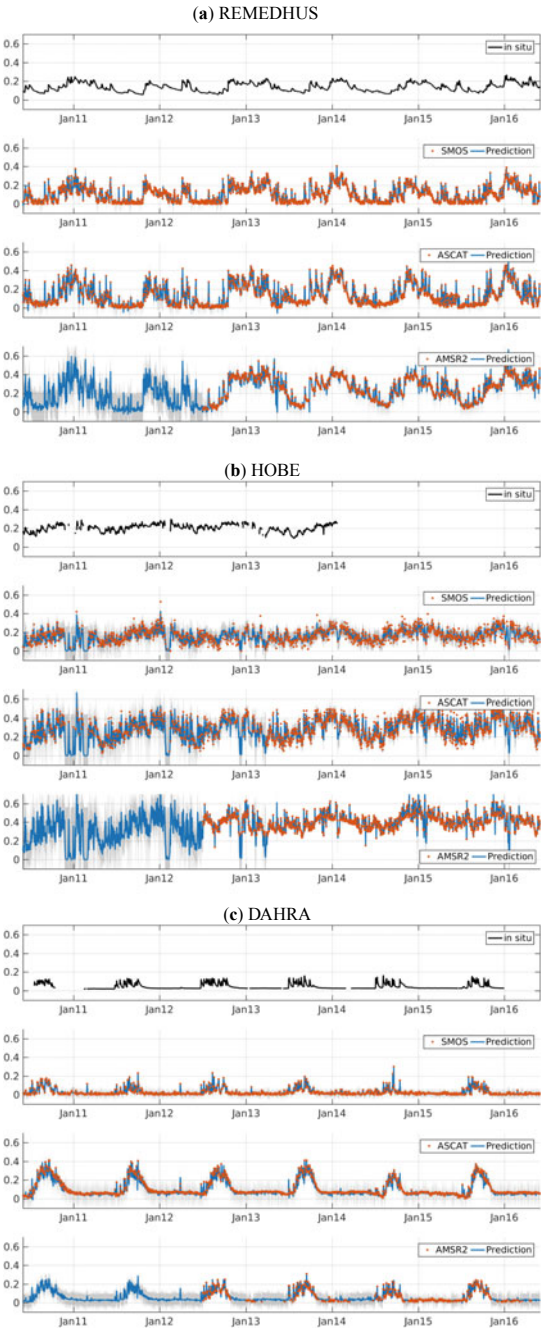

Fig. 2.4 Time series of in situ (black lines) and satellite-based soil moisture estimates from SMOS, ASCAT, and AMSR2 (orange dots denote the original time series and blue lines the predicted using the LMC-GP gap-filling technique) over **a** REMEDHUS, **b** HOBE, and **c** DAHRA networks

A statistical analysis of the original and reconstructed satellite time series has been undertaken following the recommended performance metrics in [39]. Table 2.1 shows that Pearson's correlation coefficient R, bias (as estimated by the mean error, ME) and root-mean-squared error (RMSE) with respect to in situ data in the three networks are not affected to a high degree by the reconstruction, and they remain within reasonable bounds. For SMOS, the reconstructed time series preserve the statistical scores of original time series in REMEDHUS and DAHRA and improve the R in HOBE from 0.62 to 0.68 (note the other sensors in HOBE have higher correlations of 0.66 and 0.73). The increase in coverage is notable, with an improvement of 37% for HOBE and of 54% for DAHRA. SMOS has the largest coverage over REMEDHUS, and the improvement of coverage is therefore limited (of 8%). For ASCAT, the statistical scores are preserved in the reconstructed time series, and the increase in coverage is also remarkable: 23% for REMEHDUS, 31% for HOBE, and 36% for DAHRA. For AMSR2, the validation is limited to four annual cycles (from its launch date in May 18, 2012, onward). Over REMEDHUS, AMSR2 presents a wet bias with respect to the in situ data that is reduced in the reconstructed time series; its correlation is reduced from 0.86 to 0.81, probably due to the lower correlations of the other two sensors, and the increase in coverage is of 27%. Similar results are obtained for reconstructed AMSR2 over HOBE, but with a lower number of collocated observations due to the lack of in situ data in early January 2014. Over DAHRA, correlation is improved from 0.73 to 0.79, with a 66% improvement of coverage. These results provide confidence in the proposed technique and show how it exploits the complementary spatiotemporal coverage of the three microwave sensors.

2.3 Distribution Regression for Multiscale Estimation

Non-uniform spatial sampling is a common problem in geostatistics and spatialization problems. When the variable of interest is available at the same resolution that the remote sensing observations, standard algorithms such as random forests, Gaussian processes, or neural networks are available to establish the relationship between the two. Nevertheless, we often deal with situations where the target variable is only available at the group level, collectively associated to a number of remotely sensed observations. This kind of problem is known in statistics and machine learning as *multiple instance learning* (MIL) or *distribution regression* (DR). Chapter 6 introduces the MIL framework and methodology, and reviews different approaches to address the particular issue of imprecision in hyperspectral images analysis. We here present a nonlinear method based on kernels for distribution regression that solves the previous problems without making any assumption on the statistics of the grouped data. The presented formulation considers distribution embeddings in reproducing kernel Hilbert spaces and performs standard least squares regression with the empirical means therein. A flexible version to deal with multisource data of different dimensionality and sample sizes is also introduced. The potential of the

Table 2.1 Mean error (ME) (m^3m^{-3}), unbiased RMSE (ubRMSE) (m^3m^{-3}) and Pearson correlation (R) for the original and reconstructed satellite time series against in situ measurements from REMEDHUS, HOBE, and DAHRA networks. Variable "days" reports the number of collocated satellite and in situ data available to compute the statistics

	REMEDHUS				**HOBE**				**DAHRA**			
	ME	ubRMSE	R	Days	ME	ubRMSE	R	Days	ME	ubRMSE	R	Days
SMOS	−0.032	0.003	0.81	2004	−0.042	0.002	0.62	741	−0.0143	0.001	0.79	841
SMOS rec	−0.033	0.003	0.81	2192	−0.048	0.002	0.68	1185	−0.014	0.001	0.78	1834
ASCAT	0.002	0.004	0.79	1673	0.086	0.003	0.73	816	0.071	0.002	0.70	1171
ASCAT rec	−0.001	0.004	0.78	2192	0.071	0.003	0.70	1185	0.064	0.002	0.70	1834
AMSR2	0.118	0.005	0.86	1047	0.199	0.002	0.66	486	0.026	0.002	0.73	417
AMSR2 rec	0.084	0.005	0.81	1428	0.153	0.003	0.62	513	0.019	0.001	0.79	1242

presented approach is illustrated by using SMAP VOD time series for the estimation of crop production in the US Corn Belt.

2.3.1 Kernel Distribution Regression

In distribution regression problems, we are given several sets of observations each of them with a single output target variable to be estimated. The training dataset $\mathcal{D}$ is formed by a collection of B bags (or sets) $\mathcal{D} = \{(\mathbf{X}_b \in \mathbb{R}^{n_b \times d}, y_b \in \mathbb{R}) | b = 1, \ldots, B\}$. A training set from a particular bag b is formed by n_b examples, here denoted as $\mathbf{X}_b = [\mathbf{x}_1, \ldots, \mathbf{x}_{n_b}]^\intercal \in \mathbb{R}^{n_b \times d}$, where $\mathbf{x}_i \in \mathbb{R}^{d \times 1}$. Let us denote all the available data collectively grouped in matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where $n = \sum_{b=1}^{B} n_b$, and $\mathbf{y} = [y_1, \ldots, y_B]^\intercal \in \mathbb{R}^{B \times 1}$. In this setting, the direct application of regression algorithms is not possible because not just a single input point $\mathbf{x}_b$ but a set of points $\mathbf{X}_b$ is available for each target output, and latter for prediction we may have test points or sets from each bag, $\mathbf{x}_b^* \in \mathbb{R}^{d \times 1}$ or $\mathbf{X}_b^* \in \mathbb{R}^{m_b \times d}$, which we denote with a star superscript. The problem boils down to finding a function f that learns the mapping from $\mathbf{x}$ to y exploiting the many-to-one dataset. To solve the problem, two main approaches are typically followed: (1) *output expansion*, that is, replicating the label y_b for all points in bag b; or (2) *input summary* most notably with the empirical average $\bar{\mathbf{x}}_b = \frac{1}{n_b} \sum_i \mathbf{x}_i$, or a set of centroids $\mathbf{c}_b$, $b = 1, \ldots, B$. What makes DR distinctive is that it instead exploits the rich structure in $\mathcal{D}$ by performing regression with the group distributions directly. Statistically, this consists of considering all higher order statistical relationships between the groups, not just the first- or second-order moments. The method we are going to introduce here works by embedding the bag distribution in a Hilbert space and performing linear regression therein. We essentially need the definition of a mean embedding, its induced kernel function, and how the regression is done with it.

> Distribution regression problems rely very often on using non-uniformly spatial sampled datasets, where the variables of interest are associated with sets of observations instead of single observations. While some approaches summarize the sets of observations using some kind of aggregation, such as the mean of the standard deviation, kernel distribution regression uses all higher moments by computing mean map embeddings in high-dimensional Hilbert spaces, and hence improved ability for function approximation.

2.3.1.1 Mean Map Embeddings

We frame the problem in the theory of mean map embeddings of distributions [40–42]. The kernel mean map from the set of all probability distributions $\mathcal{B}_{\mathcal{X}}$ into $\mathcal{H}$ is defined as

$$\boldsymbol{\mu} : \mathcal{B}_{\mathcal{X}} \to \mathcal{H}, \quad \mathbb{P} \to \int_{\mathcal{X}} k(\cdot, \mathbf{x}) \mathrm{d}\mathbb{P}(\mathbf{x}) \in \mathcal{H}.$$

Assuming that $k(\cdot, \mathbf{x})$ is bounded for any $\mathbf{x} \in \mathcal{X}$, it can be shown that for any $\mathbb{P}$, letting $\boldsymbol{\mu}_{\mathbb{P}} = \boldsymbol{\mu}(\mathbb{P})$, the $\mathbb{E}_P[f] = \langle \boldsymbol{\mu}_{\mathbb{P}}, f \rangle_{\mathcal{H}}$, for all $f \in \mathcal{H}$. Here $\boldsymbol{\mu}$ represents the expectation function on $\mathcal{H}$. Every probability measure has a unique embedding and the $\boldsymbol{\mu}$ fully determines the corresponding probability measure [41]. Here, we show how to estimate the mean map embeddings from empirical samples. For one particular bag, $\mathbf{X}_b$, drawn i.i.d. from a particular $\mathbb{P}_b$, the empirical mean estimator of $\boldsymbol{\mu}_b$ can be computed as

$$\widehat{\boldsymbol{\mu}}_b = \boldsymbol{\mu}_{\mathbb{P}_b} = \int k(\cdot, \mathbf{x}) \hat{\mathbb{P}}(\mathrm{d}\mathbf{x}) \approx \frac{1}{n_b} \sum_{i=1}^{n_b} k(\cdot, \mathbf{x}_i). \tag{2.8}$$

This is an empirical mean map estimator whose dot product can be computed via kernels:

$$\langle \widehat{\boldsymbol{\mu}}_{\mathbb{P}_b}, \widehat{\boldsymbol{\mu}}_{\mathbb{P}_{b'}} \rangle_{\mathcal{H}} = \frac{1}{n_b n_{b'}} \sum_{i=1}^{n_b} \sum_{j=1}^{n_{b'}} k(\mathbf{x}_i^b, \mathbf{x}_j^{b'}), \tag{2.9}$$

which is the base of a useful kernel algorithm for hypothesis testing named maximum mean discrepancy (MMD) [41, 42] and estimates the distance between two sample means in a reproducing kernel Hilbert space $\mathcal{H}$ where data are embedded

$$\mathrm{MMD}(\mathbb{P}_b, \mathbb{P}_{b'}) := \| \boldsymbol{\mu}_{\mathbb{P}_b} - \boldsymbol{\mu}_{\mathbb{P}_{b'}} \|_{\mathcal{H}}^2.$$

This can be computed using kernel functions in Eq. (2.9). Figure 2.5 shows how MMD and mean map embeddings can detect differences between distributions in higher order moments.

2.3.1.2 Distribution Regression with Kernels

The distribution regression task is carried out by standard least squares regression using the mean embedded data in Hilbert spaces. The solution leads to the kernel ridge regression (KRR) algorithm [43] working with mean map embeddings. We need to minimize a loss function composed of two terms: the least square errors of the approximation of the mean embedding, and a regularization term that acts over the class of functions to be learned in Hilbert space $f \in \mathcal{H}$:

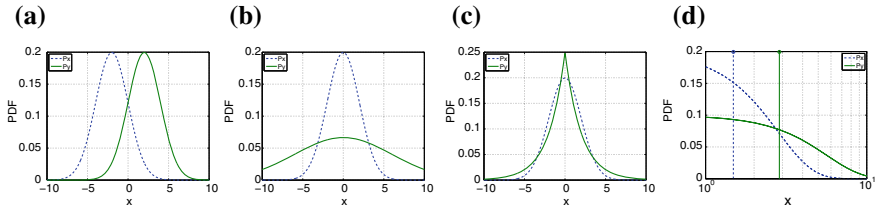

Fig. 2.5 The two-sample problem consists of detecting whether two distributions $\mathbb{P}_x$ and $\mathbb{P}_y$ are different or not. When they have different means (**a**), a simple t-test can differentiate them. When they have the same first moments (mean in **b**, mean and variance in **c**) but different higher order moments, mapping the data to higher dimensional spaces allows to distinguish them (**d**). Kernel mean embeddings are able to do so without having to map the data explicitly

$$f^* = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \| y_i - f(\boldsymbol{\mu}_i) \|^2 + \lambda \| f \|_{\mathcal{H}}^2 \right\},$$

where $\lambda > 0$ is the regularization term. The ridge regression has an analytical solution for a test set given a set of training examples:

$$\hat{f}_{\boldsymbol{\mu}_t} = \mathbf{k}(\mathbf{K} + n\lambda \mathbf{I})^{-1} \mathbf{y}, \tag{2.10}$$

where $\boldsymbol{\mu}_t$ is the mean embedding of the test set $\mathbf{X}_t$, $\mathbf{k} = [k(\boldsymbol{\mu}_1, \boldsymbol{\mu}_t), \ldots, k(\boldsymbol{\mu}_n, \boldsymbol{\mu}_t)]^\intercal \in \mathbb{R}^{n \times 1}$, $\mathbf{K} = [k(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)] \in \mathbb{R}^{n \times n}$ and $\mathbf{y} = [y_1, \ldots, y_n]^\intercal$ represents all outputs. Now, for a set of B bags each one containing n_b samples, and exploiting (2.9), one can readily compute the kernel entries of $\mathbf{K}$ as follows:

$$[\mathbf{K}]_{b,b'} = \boldsymbol{\mu}_b^\intercal \boldsymbol{\mu}_{b'} = \frac{1}{n_b n_{b'}} \mathbf{1}_{n_b}^\intercal \mathbf{K}_{bb'} \mathbf{1}_{n_{b'}},$$

where the matrix $\mathbf{K}_{bb'} \in \mathbb{R}^{n_b \times n_{b'}}$. Therefore, we have an analytic solution of the problem in (2.10):

$$\hat{y}_b^* = \frac{1}{m_b n} \mathbf{1}_{m_b}^\intercal \mathbf{K}_{bb'} \mathbf{1}_{n_{b'}} \boldsymbol{\alpha}, \tag{2.11}$$

where $\mathbf{K}_{bb'} \in \mathbb{R}^{m_b \times n_{b'}}$ which is computed given a valid Mercer kernel function k.

Kernel methods also allow to combine multisource (also known as multimodal) information in each bag, as was previously done with standard paired settings in either remote sensing or signal processing applications [42, 44, 45]. This is the case when bags have different numbers of both features and sizes, e.g., we aim to combine different spatial, spectral, or temporal resolutions. Notationally, now we have access to different matrices $\mathbf{X}_f^b \in \mathbb{R}^{n_b^f \times f}$, $f = 1, \ldots, F$. The multimodal kernel distribution method summarizes each dataset into a mean and then exploits the direct sum of Hilbert spaces in the mean embedding space. Therefore, we define F Hilbert spaces $\mathcal{H}_f$, $f = 1, \ldots, F$, and the direct sum of all of them, $\mathcal{H} = \oplus_{f=1}^{F} \mathcal{H}_f$. We

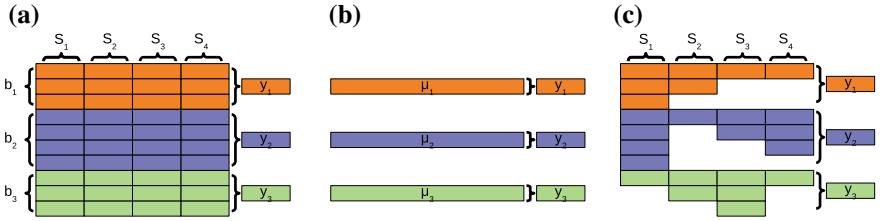

Fig. 2.6 Distribution regression approaches presented in this chapter. The DR problem is illustrated in **a** for $B = 3$ bags different numbers of samples per bag ($n_1 = 3$, $n_2 = 4$, $n_3 = 3$), three corresponding target labels, y_b, $b = 1, 2, 3$, and columns represent different features (sources, sensors) S_i, $i = 1, 2, 3, 4$. The standard approach **b** summarizes the distributions $\mathbb{P}_b$ with the mean vectors μ_b and then applies standard regression methods. Alternatively, this can be done in Hilbert spaces too with the advantage of considering all moments of the distributions. In **c**, we show the case of multisource distribution regression (MDR) in which some features are missing for particular bags and samples, which is often the case when different sensors are combined

summarize the bag feature vectors with a set of mean map embeddings of samples in bag b, which we denoted as $\boldsymbol{\mu}_b^f$. The collection of all mean embeddings in $\mathcal{H}$ is defined as $\boldsymbol{\mu}_b = [\boldsymbol{\mu}_b^1, \ldots, \boldsymbol{\mu}_b^F] \in \mathcal{H}$, and then we define the mean map embedding as $\mathbf{M} = [\boldsymbol{\mu}_1 | \cdots | \boldsymbol{\mu}_B]^\intercal \in \mathbb{R}^{B \times H}$. The multimodal kernel matrix is computed as follows:

$$[\widetilde{\mathbf{K}}]_{b,b'} = \boldsymbol{\mu}_b^\intercal \boldsymbol{\mu}_{b'} = \sum_{f=1}^{F} \frac{1}{n_b^f n_{b'}^f} \mathbf{1}_{n_b^f}^\intercal \mathbf{K}_{bb'} \mathbf{1}_{n_{b'}^f}, \tag{2.12}$$

Fig. 2.6 graphically illustrates the DR approaches used in this chapter. The algorithm reduces to the application of a standard kernel ridge regression with the kernel function Eq. (2.11) for the standard case or Eq. (2.12) for the multisource case. We provide source code of our methods in http://isp.uv.es/code/dr.html.

2.3.2 Data and Setup

We show results for crop yield estimation, which is a particular problem of distribution regression in the context of remote sensing. We show results for our KDR (kernel distribution regression) and several baseline standard approaches like least squares regularized linear regression model (RLR) and its nonlinear (kernel) counterpart, the kernel ridge regression (KRR) method, both working on the empirical means of each bag as input feature vectors. We use as evaluation criteria the standard mean error (ME) to account for bias, the root-mean-square-error (RMSE) to assess accuracy, and the coefficient of determination or explained variance (R^2) to account for the goodness-of-fit.

Specifically, for the crop yield estimation, satellite-based retrievals of vegetation optical depth (VOD) from SMAP [46] is related to crop production data from the

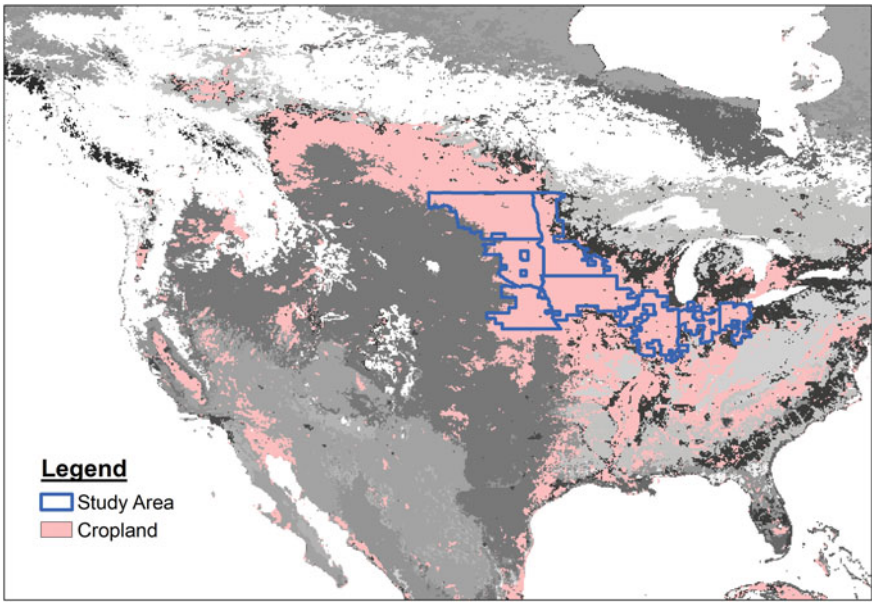

Fig. 2.7 Area of study. It includes both the eight states and the cropland mask following the MODIS IGBP land cover classification

2015 US agricultural survey (total yield and yield per crop type), and then the proposed methods are evaluated. VOD is a measure of the attenuation of soil microwave emissions when they pass through the vegetation canopy, being sensitive not only to the amount of living biomass, but also to the amount of water stress experienced by the vegetation [47]. SMAP VOD has been shown to carry information about crop growth and yield in a variety of agro-ecosystems [48, 49].

We focus on eight states within the so-called Corn Belt of the Midwestern United States: Illinois, Indiana, Iowa, Minnesota, Nebraska, North Dakota, Ohio, and South Dakota (Fig. 2.7). Also, the United States Department of Agriculture, in particular, the National Agricultural Statistics Service (USDA-NASS), publish reports and survey of agricultural information every year at the country, state and county levels. There is a total of 385 counties with yield and satellite data for prediction of total yield. We also predict per crop type. In particular, the three main crops in the region, i.e., corn, soybean, and wheat, are predicted. All the 363, 361, and 204 counties reporting corn, soybean, and wheat yields, independent of their relative importance at the county level, are included in the corresponding crop-specific experiments.

2.3.3 Results

The methodology for evaluating the algorithms is as follows. A 66% of the counties (bags) are used to train/validate and the remaining 33% are used for testing. With the first ones, we perform a fivefold cross-validation also at a bag level, i.e., we split the data into five subsets, one reserved for validation and the rest used for training the regression model. After this, we only apply the best model found to the test data. Finally, all this process is repeated ten times, and the average over all test results is computed. Only test errors are reported.

Table 2.2 shows the crop yield predictions for all the approaches. Notably, these results outperform those obtained in previous literature for corn–soy croplands ([48] and references therein), even with the simplest models like RLR and KRR. Results of the best regression model between VOD and official corn yields at county level are illustrated in Fig. 2.8. Except in few counties, corn predictions are reasonably good, with relative errors below 3%. The proposed DR approaches will be particularly useful for regional crop forecasting in areas covering different agro-climatic conditions and fragmented agricultural landscapes (e.g., Europe), where scale effects need to be properly addressed for adequate analysis and predictions [50].

Table 2.2 Results for prediction of total yield and crop yield prediction using VOD (Kg m^{-2})

Total crop yield	ME×1000	RMSE ×100	R^2
RLR	1.19 ± 7.36	9.67 ± 0.74	0.80 ± 0.02
KRR	2.22 ± 10.77	9.34 ± 0.73	0.81 ± 0.02
KDR	2.27 ± 10.95	9.35 ± 0.71	0.81 ± 0.02
Corn yield	ME×1000	RMSE ×100	R^2
RLR	-1.20 ± 5.89	7.54 ± 0.50	0.85 ± 0.02
KRR	1.68 ± 8.52	6.54 ± 0.72	0.88 ± 0.02
KDR	1.59 ± 7.88	6.47 ± 0.74	0.89 ± 0.02
Soybean yield	ME×1000	RMSE ×100	R^2
RLR	-1.99 ± 1.85	2.45 ± 0.13	0.85 ± 0.03
KRR	-0.70 ± 2.92	2.47 ± 0.21	0.85 ± 0.04
KDR	-0.64 ± 2.43	2.40 ± 0.21	0.86 ± 0.03
Wheat yield	ME×1000	RMSE ×100	R^2
RLR	2.72 ± 6.65	5.46 ± 0.48	0.64 ± 0.08
KRR	2.42 ± 8.47	5.07 ± 0.38	0.69 ± 0.05
KDR	2.91 ± 7.31	5.10 ± 0.40	0.69 ± 0.05

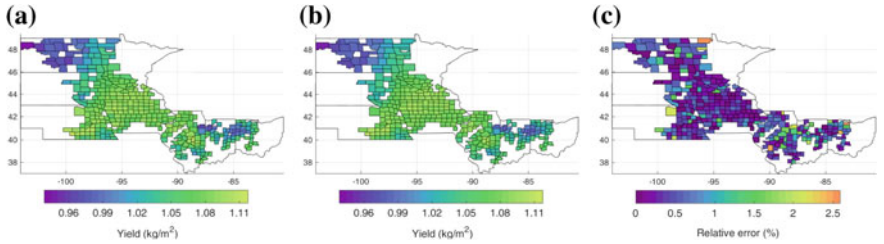

Fig. 2.8 **a** Map of official corn yield for year 2015 from USDA-NASS survey given in (Kg/m^2). **b** KDR predicted corn yield and **c** KDR relative error prediction per county (%)

2.4 Global Parameter Estimation in the Cloud

From an operational point of view, the implementation of biophysical parameter retrieval chains on ongoing basis demands high storage capability and efficient computational power, mainly when dealing with long time series of remote sensing data at global scales. There exist a wide variety of free available remote sensing data which could be potentially ingested in these processing chains. Among them, one can find remote sensing data disseminated by NASA (e.g., MODIS), the United States Geological Survey (USGS) (e.g., Landsat), and ESA (e.g., data from the Sentinel constellation). To deal with this huge amount of data, Google developed the Google Earth Engine [51], a cloud computing platform specifically designed for geospatial analysis at the petabyte scale. Due to its unique features, GEE is the state of the art in remote sensing big data processing. The GEE data catalog is composed by widely used geospatial datasets. The catalog is continuously updated and data are ingested from different government-supported archives such as the Land Process Distributed Active Archive Center (LP DAAC), the USGS, and the ESA Copernicus Open Access Hub. The GEE data repository embrace a wide variety of remote sensing datasets including meteorological records, atmospheric estimates, vegetation, and land properties and also surface reflectance data. Data processing is performed in parallel on Google's computational infrastructure, dramatically improving processing efficiency and speed. These features, among others, make GEE an extremely valuable tool for multitemporal and global studies which include vegetation, temperature, carbon exchange, and hydrological processes [24, 52, 53].

Here, we present an example of biophysical parameter estimation in the GEE cloud computing platform. The developed processing chain includes the joint estimation of LAI, FAPAR, FVC, and CWC parameters at global scale from long-term time series (15 years) of MODIS data exploiting the GEE cloud processing capabilities. The retrieval approach is based on a hybrid method, which combines the physically based PROSAIL radiative transfer model with random forests (RFs) regression. The implementation on GEE platform allowed us to use global and climate data records (CDR) of both MODIS surface reflectance and LAI/FAPAR datasets which provided

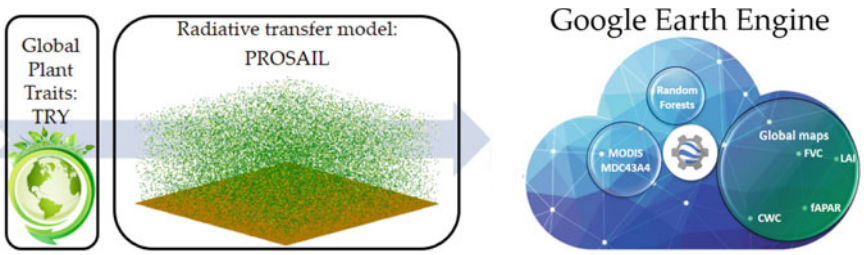

Fig. 2.9 Schema of the developed biophysical retrieval chain in the cloud

us with global biophysical variable maps at unprecedented timeliness. Figure 2.9 shows an schema summarizing the developed retrieval chain.

Cloud-based geospatial computing platforms such as Google Earth Engine offer opportunities to create a broad range of applications with precision and accuracy over unprecedented large areas with medium and high spatial resolutions. In this section, we illustrate the advantages of using algorithms implemented in a cloud computing infrastructure dealing with a common problem in remote sensing science, the retrieval of land biophysical parameters.

2.4.1 Data and Setup

As shown in Fig. 2.9, to model the spectral response of the vegetation we chose the PROSAIL radiative transfer model. This model results from the PROSPECT leaf optical reflectance model [54] coupled with the SAIL canopy model [55]. PROSAIL has been widely used in many remote sensing studies [56] and successfully applied for local and global parameter estimation [24, 57–59]. PROSAIL assumes the canopy as a turbid medium and simulates vegetation reflectance along the optical spectrum (from 400 to 2500 nm) depending on the leaf biochemistry, structure of the canopy, as well as the background soil reflectance and the sun–satellite geometry. At leaf level, the parametrization was based on the distributions derived from a massive global leaf trait measurements (TRY) [60] in order to account for a realistic representation of global leaf trait variability to optimize PROSAIL at global scale, whereas distributions of the canopy variables were similar to those adopted in other global studies [59]. The TRY database embrace 6.9 million trait records for 148,000 plant taxa at unprecedented spatial and climatological coverage [60]. Although the database is recent, due to the TRY unique properties, these data have been widely used and hundreds of top publications (TRY database) have been presented covering topics ranging from ecology and plant geography to vegetation modeling and

Table 2.3 General information about leaf traits measurements used in this work

Trait	No. samples	No. of species
C_{ab}	19,222	941
C_{dm}	69,783	11,908
C_w	32,020	4802

Table 2.4 Spectral specifications of the MODIS MCD43A4 product

MODIS band	Wavelength (nm)
Band 1 (red)	620–670
Band 2 (NIR)	841–876
Band 3 (blue)	459–479
Band 4 (green)	545–565
Band 5 (SWIR-1)	1230–1250
Band 6 (SWIR-2)	1628–1652
Band 7 (MWIR)	2105–2155

remote sensing [25, 61]. In this section, instead of using the usual lookup tables available in the literature, we use the TRY to parametrize PROSAIL. Our objective is to exploit the TRY database to infer more realistic distributions and correlations among some key leaf traits such as leaf chlorophyll (C_{ab}), leaf dry matter (C_{dm}), and water (C_w) contents. Table 2.3 shows some basic information about the considered traits extracted from the TRY.

The reflectance simulations obtained with PROSAIL were set up to mimic the MCD43A4 product bands which are available in GEE. The MCD43A4 MODIS product is generated combining data from Terra and Aqua spacecrafts, being disseminated as a level-3 gridded dataset. This product provides a bidirectional reflectance distribution function (BRDF) from a nadir view in the seven land MODIS bands (see Table 2.4 for more details), thus offering global surface reflectance data at 500 m spatial resolution with 8-day temporal frequency.

PROSAIL's forward mode provides a reflectance spectrum given a set of input parameters (leaf chemical components/traits, structural parameters of the vegetation canopy, etc.). After running PROSAIL in forward mode, its inversion was undertaken using RFs. This inversion allows, in turn, to retrieve the selected biophysical parameters (LAI, FAPAR, FVC, and CWC). RFs have been applied both for classification and regression in multitude of remote sensing studies [62] including forest ecology [63, 64], land cover classification [65], and feature selection [66]. We chose RFs to invert the PROSAIL model mainly because they can cope with high-dimensional problems due to their optimal pruning strategy and efficiency. RF is an ensemble method that builds up a stack of decision trees. This approach has been proven to be very beneficial to alleviate over-fitting problems in single decision tree models. On the ensemble, every tree is trained with different subsets of features and examples

(selected randomly) yielding an individual prediction. The combined prediction (usually the mean value) of the considered trees composing the RFs is the final prediction of the model [67]. The computed simulations obtained with PROSAIL were split into two groups: (1) a training dataset to optimize the models, and (2) an independent test set which was only used to assess the models (RFs). After our models were trained and validated, we predicted the chosen biophysical variables using real MODIS spectral information (land bands, see Table 2.4). In addition, RFs, once trained, are easily parallelized to cope with large-scale problems routinely encountered in global remote sensing applications. This is specifically the case of the problem described here, where we exploit large datasets and run predictions covering many years within the Google Earth Engine platform. A toy example of the code is available at https://code.earthengine.google.com/e3a2d589395e4118d97bae3e85d09106.

2.4.2 Results

The PROSAIL simulations were uploaded to GEE and randomly split into train (2/3 of the simulations) and test (the remaining 1/3 of the samples never used in the RFs training) datasets. The RFs theoretical performance evaluated in the GEE platform (assessed over the test dataset) revealed high correlations (R^2 = 0.84, 0.89, 0.88, and 0.80 for LAI, FAPAR, FVC, and CWC, respectively), low errors (RMSE = 0.91 m^2/m^2, 0.08, 0.06, and 0.27 kg/m^2 for LAI, FAPAR, FVC, and CWC, respectively), and practically no biases in all cases. Subsequently, the RFs retrieval model was executed over the computing cloud to obtain 15 years of global biophysical parameters from the MCD43A4 product available on GEE. Figure 2.10 shows the global mean values of LAI, FAPAR, FVC, and CWC derived from 2010 to 2015. The computation of the mean biophysical maps implied processing 230 (46 yearly images × 5 years) FAPAR images at 500 m spatial resolution (∼440 million cells), and compute their annual mean, which took around 6 h.

Validation of the estimates was achieved by means of intercomparison over a network of sites named BELMANIP-2.1 (Benchmark Land Multisite Analysis and Intercomparison of Products) especially selected for representing the global variability of Earth vegetation. Over this network, we compared the LAI and FAPAR estimates against the official LAI/FAPAR MODIS product (MCD15A3H) on GEE. We selected the MODIS pixels for every BELMANIP-2.1 location, and then we computed the mean value of the MODIS valid pixels within a 1 km surrounding area. In addition, since the MCD15A3H and MCD43A4 differ in temporal frequency, only the coincident dates between them were selected for comparison. For validation, we selected only high-quality MODIS pixels which resulted in ∼60000 valid pixels from 2002–2017 accounting for vegetation biomes: evergreen broadleaf forests (EBF), broadleaf deciduous forest (BDF), needle leaf forest (NLF), cultivated (C), shrublands (SH), herbaceous (H), and bare areas (BA). For FAPAR, very good agreement (R^2 ranging from 0.89 to 0.92) and low errors (RMSE ranging from 0.06 to 0.08) were found between retrievals and the MODIS FAPAR product over bare areas, shrub-

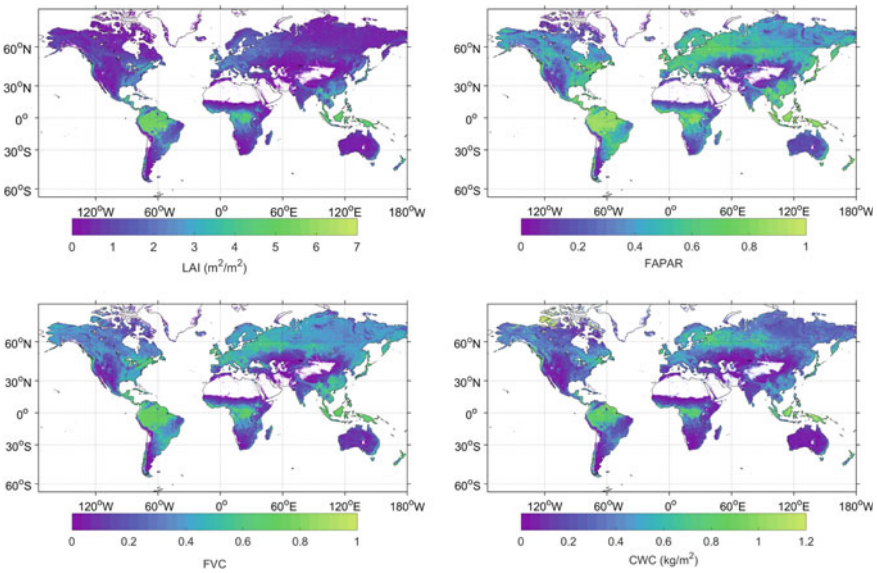

Fig. 2.10 LAI, FAPAR, FVC, CWC global maps corresponding to the mean values estimated by the proposed retrieval chain for the period 2010–2015

lands, herbaceous, cultivated, and broadleaf deciduous forest biomes. For needle-leaf and evergreen broadleaf forests, lower correlations (R^2 = 0.57 and 0.41) and higher errors (RMSE = 0.18 and 0.09) were obtained. It is worth mentioning that over bare areas, the MODIS FAPAR presents an unrealistic minimum value (∼0.05) through the entire period. In the case of LAI, goodness-of-fit ranging from 0.70 to 0.86 and low errors (RMSE ranging from 0.23 to 0.57 m^2/m^2) were found between estimates in all biomes except for evergreen broadleaf forest, where R^2 = 0.42 and RMSE = 1.13 m^2/m^2 are reported.

Figure 2.11 shows the LAI and FAPAR difference maps calculated using the mean outcomes (2010–2015) of our processing chain and the mean reference MODIS LAI/FAPAR product for the same period. The mean difference LAI map shows that the discrepancies among both products range within the ± 0.5 m^2/m^2 range, indicating that both products are consistent. However, our high LAI values present a significant underestimation over heavily vegetated areas (dense canopies) that reaches values up to 1.4 m^2/m^2. When comparing both FAPAR products, a constant negative bias of ≈ 0.05 m^2/m^2 our estimates is observed. These differences could be related with a documented systematic overestimation of operational MODIS FAPAR [68], meaning that our approach is partly correcting some of the flaws of the official MODIS product.

State-of-the-art cloud computing platforms like GEE provides routinely time series of global land surface variables related with vegetation status and an unprecedented computational power. Despite the variety of regression and classification methods implemented in GEE, the user could be limited by the number of state-of-

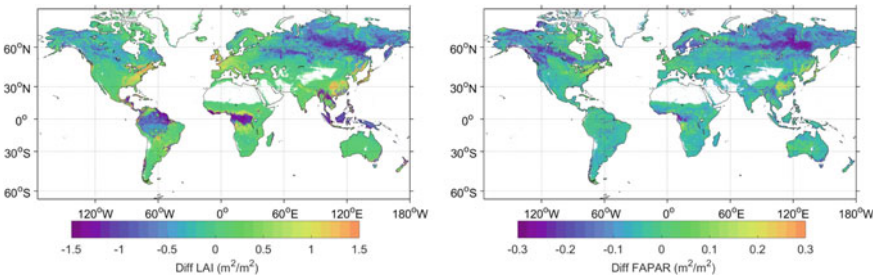

Fig. 2.11 LAI and FAPAR global maps corresponding to the difference of mean values between derived estimates by the proposed retrieval chain and the GEE MODIS reference product for the period 2010–2015

the-art algorithms which are currently implemented. However, GEE is being updated at a fast pace due to an increasing number of users developing new approaches and methods that may be potentially implemented in GEE for a wide range of geoscience applications. Here, we have illustrated an application that takes advantage of GEE capabilities to retrieve standard biophysical variables at a global scale. The validation of our estimates revealed, in general, good spatial consistency. However, differences in mean LAI values over dense forests are still noticeable and could be attributed mostly to differences in retrieval approaches. Other possible source for discrepancies shown could be associated to (i) product definition, such as those related with considering or not vegetation clumping [69], (ii) embedded algorithm assumptions (RTM, optical properties, canopy architecture), and (iii) satellite input data and processing. In relation with the FAPAR, as mentioned above, an overall negative bias is found for all biomes, which is not an issue since different studies have pointed out a systematic overestimation of MODIS retrievals in both C5 and C6 at low FAPAR values. Finally, it is worth mentioning that neither the FVC nor the CWC products are available on GEE. Moreover, there is no global and reliable CWC product with which compare the CWC estimates derived by the proposed retrieval chain. Regarding FVC, there are only a few global products that differ in retrieval approaches and spatiotemporal features.

2.5 Conclusions

This chapter focused on the problem of parameter estimation from remotely sensed optical sensor data. We identified two main challenges related to the data and the used models. To satisfy the urgent needs of fast and accurate data processing and product generation, we relied on three main building blocks: advanced machine learning, big and heterogeneous EO data, and large-scale processing platforms. In this scenario, machine learning has to be redesigned to accommodate data characteris-

tics (spatiotemporal and multi-sensor structures, higher order, and nonlinear feature relations), to be accurate and flexible, and to scale well to millions of observations.

To deal with these challenges, we introduced three machine learning approaches to exploit different spatial, multi-sensor, temporal, and large-scale data characteristics. In particular, we exploited multi-output Gaussian processes for gap-filling time series, kernel distribution regression models that exploits multiple observations and avoid working with arbitrary summarizing statistics, and random forests trained on RTM simulations and implemented in the GEE computation cloud. The approaches allow us to estimate key land parameters from optical and microwave EO data synergistically: SM, LAI, FAPAR, FVC, CWC, and crop yield.

Synergistic benefits of machine learning, big data, and scalable cloud computing are here to stay, and we envision many exciting developments in the near future. EO data allows to monitor continuously in space and time the Earth and can be used to "spatialize" almost any arbitrary quantity measured on the ground or simulated with appropriate transfer codes. Plant, vegetation, and land parameters will readily benefit from ML-based approaches in the cloud to make reliable and accurate products accessible to everyone.

References

1. Camps-Valls G, Tuia D, Gómez-Chova L, Jiménez S, Malo J (eds) (2011) Remote Sens Image Process. Morgan & Claypool Publishers, LaPorte, CO, USA
2. Liang S (2004) Quantitative Remote Sensing of Land Surfaces. Wiley, New York
3. Liang S (2008) Advances in land remote sensing: system, modeling. inversion and applications. Springer, Germany
4. Lillesand TM, Kiefer RW, Chipman J (2008) Remote sensing and image interpretation. Wiley, New York
5. Rodgers CD (2000) Inverse methods for atmospheric sounding: theory and practice. World Scientific Publishing Co., Ltd
6. Baret F, Buis S (2008) Estimating canopy characteristics from remote sensing observations: review of methods and associated problems. In: Advances in land remote sensing: system, modeling, inversion and applications. Springer, Germany
7. Baret F, Weiss M, Lacaze R, Camacho F, Makhmara H, Pacholcyzk P, Smets B (2013) GEOV1: LAI and FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. part1: principles of development and production. Remote Sens Environ 137(0):299–309
8. Beer C, Reichstein M, Tomelleri E, Ciais P, Jung M, Carvalhais N, Rödenbeck C, Arain MA, Baldocchi D, Bonan GB, Bondeau A, Cescatti A, Lasslop G, Lindroth A, Lomas M, Luyssaert S, Margolis H, Oleson KW, Roupsard O, Veenendaal E, Viovy N, Williams C, Woodward FI, Papale D (2010) Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. Science 329(5993):834–838
9. Jung M, Reichstein M, Margolis HA, Cescatti A, Richardson AD, Arain MA, Arneth A, Bernhofer C, Bonal D, Chen J, Gianelle D, Gobron N, Kiely G, Kutsch, W, Lasslop G, Law BE, Lindroth A, Merbold L, Montagnani L, Moors EJ, Papale D, Sottocornola M, Vaccari F, Williams C (2011) 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: Biogeosci 116(G3)

10. Jung M, Reichstein M, Schwalm CR, Huntingford C, Sitch S, Ahlström A, Arneth A, Camps-Valls G, Ciais P, Friedlingstein P, Gans F, Ichii K, Jain AK, Kato E, Papale D, Poulter B, Raduly B, Rödenbeck C, Tramontana G, Viovy N, Wang YP, Weber U, Zaehle S, Zeng N (2017) Compensatory water effects link yearly global land CO_2 sink changes to temperature. Nature 541(7638):516–520
11. Tramontana G, Jung M, Camps-Valls G, Ichii K, Raduly B, Reichstein M, Schwalm CR, Arain MA, Cescatti A, Kiely G, Merbold L, Serrano-Ortiz P, Sickert S, Wolf S, Papale D (2016) Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. Biogeosci Discuss 2016:1–33. https://doi.org/10.5194/bg-2015-661
12. Sarker LR, Nichol JE (2011) Improved forest biomass estimates using ALOS AVNIR-2 texture indices. Remote Sens Env 115(4):968–977
13. Durbha SS, King RL, Younan NH (2007) Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. Rem Sens Env 107(1–2):348–361
14. Yang F, White MA, Michaelis AR, Ichii K, Hashimoto H, Votava P, Zhu AX, Nemani RR (2006) Prediction of continental-scale evapotranspiration by combining MODIS and ameriflux data through support vector machine. IEEE Trans Geosci Remote Sens 44(11):3452–3461
15. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press, Cambridge, MA
16. Verrelst J, Muñoz J, Alonso L, Delegido J, Rivera JP, Moreno J, Camps-Valls G (2012) Machine learning regression algorithms for biophysical parameter retrieval: opportunities for Sentinel-2 and -3. Remote Sens Environ 118:127–139
17. Camps-Valls G, Verrelst J, Muñoz-Marí J, Laparra V, Mateo-Jimenez F, Gómez-Dans J (2016) A survey on gaussian processes for earth-observation data analysis: a comprehensive investigation. IEEE Geosci Remote Sens Mag 4(2):58–78
18. Reichstein M, Camps-Valls G, Stevens B, Denzler J, Carvalhais N, Jung M (2019) Prabhat: deep learning and process understanding for data-driven Earth system science. Nature
19. Ulaby FT, Long D, Blackwell W, Elachi C, Fung A, Ruf C, Sarabandi K, van Zyl J, Zebker H (2014) Microwave radar and radiometric remote sensing. University of Michigan Press
20. Dorigo W, Wagner W, Albergel C, Albrecht F, Balsamo G, Brocca L, Chung D, Ertl M, Forkel M, Gruber A, Haas E, Hamer PD, Hirschi M, Ikonen J, de Jeu R, Kidd R, Lahoz W, Liu YY, Miralles D, Mistelbauer T, Nicolai-Shaw N, Parinussa R, Pratola C, Reimer C, van der Schalie R, Seneviratne SI, Smolander T, Lecomte P (2017) ESA CCI soil moisture for improved earth system understanding: state-of-the art and future directions. Remote Sens Environ 203:185–215. Earth Observation of Essential Climate Variables
21. Mateo-Sanchis A, Muñoz-Marí J, Campos-Taberner M, García-Haro J, Camps-Valls G (2018) Gap filling of biophysical parameter time series with multi-output gaussian processes. In: IGARSS 2018—2018 IEEE international geoscience and remote sensing symposium, pp 4039–4042
22. Piles M, van der Schalie R, Gruber A, Muñoz-Marí J, Camps-Valls G, Mateo-Sanchis A, Dorigo W, de Jeu R (2018) Global estimation of soil moisture persistence with L and C-band microwave sensors. In: IGARSS 2018—2018 IEEE international geoscience and remote sensing symposium, pp 8259–8262
23. Adsuara JE, Pérez-Suay A, Muñoz-Marí J, Mateo-Sanchis A, Piles M, Camps-Valls G (2019) Nonlinear distribution regression for remote sensing applications. IEEE Trans Geosci Remote Sens (2019) (Submitted)
24. Campos-Taberner M, Moreno-Martínez A, García-Haro FJ, Camps-Valls G, Robinson NP, Kattge J, Running SW (2018) Global estimation of biophysical variables from google earth engine platform. Remote Sens 10:1167
25. Moreno A, Camps G, Kattge J, Robinson N, Reichstein M, van Bodegom P, Kramer K, Cornelissen J, Reich P, Bahn M et al (2018) A methodology to derive global maps of leaf traits using remote sensing and climate data. Remote Sens Environ 218:69–88
26. Dorigo WA, Gruber A, Jeu RAMD, Wagner W, Stacke T, Loew A, Albergel C, Brocca L, Chung D, Parinussa RM, Kidd R (2015) Evaluation of the ESA CCI soil moisture product using ground-based observations. Remote Sens Environ 162:380–395

27. Piles M, Ballabrera-Poy J, Muñoz-Sabater J (2019) Dominant features of global surface soil moisture variability observed by the SMOS satellite. Remote Sens 11(1):95
28. Alvarez MA, Rosasco L, Lawrence ND (2011) Kernels for vector-valued functions: a review. arXiv:1106.6251 [cs, math, stat]. ArXiv: 1106.6251
29. Verrelst J, Alonso L, Camps-Valls G, Delegido J, Moreno J (2012) Retrieval of vegetation biophysical parameters using gaussian process techniques. IEEE Trans Geosci Remote Sens 50(5/P2):1832–1843
30. Journel A, Huijbregts C (1978) Mining geostatistics. Academic Press
31. Albergel C, de Rosnay P, Gruhier C, Muñoz-Sabater J, Hasenauer S, Isaksen L, Kerr Y, Wagner W (2012) Evaluation of remotely sensed and modelled soil moisture products using global ground-based in situ observations. Remote Sens Environ 118:215–226
32. González-Zamora Á, Sánchez N, Martínez-Fernández J, Gumuzzio Á, Piles M, Olmedo E Long-term SMOS soil moisture products: a comprehensive evaluation across scales and methods in the duero basin (spain)
33. Al-Yaari A, Wigneron JP, Ducharne A, Kerr YH, Wagner W, Lannoy GD, Reichle R, Bitar AA, Dorigo W, Richaume P, Mialon A (2014) Global-scale comparison of passive (SMOS) and active (ASCAT) satellite based microwave soil moisture retrievals with soil moisture simulations (MERRA-land). Remote Sens Environ 152:614–626
34. Albergel C, Dorigo W, Balsamo G, noz Sabater JM, de Rosnay P, Isaksen L, Brocca L, de Jeu R, Wagner W (2013) Monitoring multi-decadal satellite earth observation of soil moisture products through land surface reanalyses. Remote Sens Environ 138:77–89
35. Polcher J, Piles M, Gelati E, Barella-Ortiz A, Tello M (2016) Comparing surface-soil moisture from the SMOS mission and the ORCHIDEE land-surface model over the iberian peninsula. Remote Sens Environ 174:69–81
36. Sanchez N, Martinez-Fernandez J, Scaini A, Perez-Gutierrez C (2012) Validation of the SMOS L2 soil moisture data in the REMEDHUS network (Spain). IEEE Trans Geosci Remote Sens 50(5):1602–1611
37. Bircher S, Skou N, Jensen KH, Walker JP, Rasmussen L (2012) A soil moisture and temperature network for SMOS validation in western denmark. Hydrol Earth Syst Sci 16(5):1445–1463
38. Torbern T, Rasmus F, Idrissa G, Olander RM, Silvia H, Cheikh M, Monica G, Stéphanie H, Inge S, Bo HR, Marc-Etienne R, Niklas O, Jørgen LO, Andrea E, Mathias M, Jonas A (2014) Ecosystem properties of semiarid savanna grassland in west africa and its relationship with environmental variability. Global Change Biol 21(1):250–264
39. Entekhabi D, Reichle RH, Koster RD, Crow WT (2010) Performance metrics for soil moisture retrievals and applications requirements. J Hydrometeorol 11:832–840
40. Harchaoui Z, Bach F, Cappe O, Moulines E (2013) Kernel-based methods for hypothesis testing: a unified view. IEEE Signal Proc Mag 30(4):87–97
41. Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B (2016) Kernel mean embedding of distributions: a review and beyond. now foundations and trends
42. Rojo-Álvarez JL, Martínez-Ramón M, Muñoz-Marí J, Camps-Valls G (2017) Digital signal processing with Kernel methods. Wiley, UK
43. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge, MA, USA
44. Camps-Valls G, Bruzzone L (2009) Kernel methods for remote sensing data analysis. Wiley
45. Camps-Valls G, Gómez-Chova L, Muñoz-Marí J, Vila-Francés J, Calpe-Maravilla J (2006) Composite kernels for hyperspectral image classification. IEEE Geosci Remote Sens Lett 3(1):93–97
46. Konings AG, Piles M, Das N, Entekhabi D (2017) L-band vegetation optical depth and effective scattering albedo estimation from SMAP. Remote Sens Environ 198:460–470
47. Jackson RD, Huete AR (1991) Interpreting vegetation indices. Prev Vet Med 11(3–4):185–200
48. Chaparro D, Piles M, Vall-llossera M, Camps A, Konings AG, Entekhabi D (2018) L-band vegetation optical depth seasonal metrics for crop yield assessment. Remote Sens Environ 212:249–259

49. Piles M, Camps-Valls G, Chaparro D, Entekhabi D, Konings AG, Jagdhuber T (2017) Remote sensing of vegetation dynamics in agro-ecosystems using smap vegetation optical depth and optical vegetation indices. In: IGARSS17, pp 4346–4349
50. López-Lozano R, Duveiller G, Seguini L, Meroni M, García-Condado S, Hooker J, Leo O, Baruth B (2015) Towards regional grain yield forecasting with 1km-resolution EO biophysical products: strengths and limitations at pan-european level. Agric For Meteorol 206:12–32
51. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ 202:18–27
52. He M, Kimball JS, Maneta MP, Maxwell BD, Moreno A, Beguería S, Wu X (2018) Regional crop gross primary productivity and yield estimation using fused landsat-MODIS data. Remote Sens 10:372
53. Robinson NP, Allred B, Jones MO, Moreno A, Kimball JS, Naugle D, Erickson TA, Richardson AD (2017) A dynamic landsat derived normalized difference vegetation index (NDVI) product for the conterminous united states. Remote Sens 9:823
54. Jacquemoud S, Baret F (1990) PROSPECT: a model of leaf optical properties spectra. Remote Sens Environ 43:75–91
55. Verhoef W (1984) Light scattering by leaf layers with application to canopy reflectance modeling: the SAIL model. Remote Sens Environ 16:125–141
56. Berger K, Atzberger C, Danner M, D'Urso G, Mauser W, Vuolo F, Hank T (2018) Evaluation of the PROSAIL model capabilities for future hyperspectral model environments: a review study. Remote Sens 10:85
57. Campos-Taberner M, García-Haro FJ, Camps-Valls G, Grau-Muedra G, Nutini F, Busetto L, Katsantonis D, Stavrakoudis D, Minakou C, Gatti L, Barbieri M, Holecz F, Stroppiana D, Boschetti M (2017) Exploitation of SAR and optical sentinel data to detect rice crop and estimate seasonal dynamics of leaf area index. Remote Sens 9:248
58. Campos-Taberner M, García-Haro FJ, Camps-Valls G, Grau-Muedra G, Nutini F, Crema A, Boschetti M (2016) Multitemporal and multiresolution leaf area index retrieval for operational local rice crop monitoring. Remote Sens Environ 187:102–118
59. García-Haro FJ, Campos-Taberner M, noz Marí JM, Laparra V, Camacho F, Sánchez-Zapero J, Camps-Valls G (2018) Derivation of global vegetation biophysical parameters from EUMETSAT polar system. ISPRS J Photogramm Remote Sens 139:57–75
60. Kattge J, Díaz S, Lavorel S, Prentice I, Leadley P, Bönisch G et al (2011) TRY-a global database of plant traits. Glob Change Biol 17:2905–2935
61. Madani N, Kimball J, Ballantyne A, Affleck D, van Bodegom P, Reich P, Kattge J, Sala A et al (2018) Future global productivity will be affected by plant trait response to climate. Sci Rep 8(2870)
62. Belgiu M, Lucian D (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31
63. De'ath G, Fabricius K (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192
64. Evans J, Cushman S (2009) Gradient modeling of conifer species using random forests. Landsc Ecol 24:673–683
65. Cutler D, Edwards J, Thomas C, Beard K, Cutler A, Hess K, Gibson J, Lawler J (2007) Random forests for classification in ecology. Ecology 88:2783–2792
66. Chen YW, Lin CJ (2006) Combining SVMs with various feature selection strategies. In: Guyon I, Nikravesh M, Gunn S, Zadeh LA (eds) Feature extraction. Studies in fuzziness and soft computing, vol 207. Springer, Berling, Heidelberg

67. Breiman L, Friedman J (1985) Estimating optimal transformations for multiple regression and correlation. J Am Stat Assoc 391:1580–1598
68. Yan K, Park T, Yan G, Liu Z, Yang B, Chen C, Nemani R, Knyazikhin Y, Myneni R (2016) Evaluation of MODIS LAI/FPAR product collection 6. part 2: validation and intercomparison. Remote Sens 8(460)
69. Campos-Taberner M, j García-Haro F, Busetto L, Ranghetti L, Martínez B, Gilabert MA, Camps-Valls G, Camacho F, Boschetti M (2018) A critical comparison of remote sensing leaf area index estimates over rice-cultivated areas: from Sentinel-2 and Landsat-7/8 to MODIS, GEOV1 and EUMETSAT polar system. Remote Sens 10:763