

AGENCIA ESTATAL DE INVESTIGACIÓN - Convocatorias 2018  
Proyectos de I+D de GENERACIÓN DE CONOCIMIENTO y Proyectos I+D+i RETOS INVESTIGACIÓN

**AVISO IMPORTANTE - La memoria no podrá exceder de 20 páginas. Para rellenar correctamente esta memoria, lea detenidamente las instrucciones disponibles en la web de la convocatoria. Es obligatorio rellenarla en inglés si se solicita más de 100.000 €.**

**IMPORTANT – The research proposal cannot exceed 20 pages. Instructions to fill this document are available in the website. If the project cost exceeds 100.000 €, this document must be filled in English.**

**IP 1** (Nombre y apellidos): Ana Belén Ruescas Orient

**TÍTULO DEL PROYECTO (ACRÓNIMO):** MÉTODOS AVANZADOS DE CLASIFICACIÓN PARA TIPOS ÓPTICOS DE AGUAS CONTINENTALES (AQUACLASS)

**TITLE OF THE PROJECT (ACRONYM):** ADVANCED CLASSIFICATION METHODS FOR OPTICAL INLAND WATER TYPES (AQUACLASS)

## 1. PROPUESTA CIENTÍFICA

### 1.1. Introduction

Freshwater lakes, reservoirs and rivers are an essential resource for human and animal survival. They represent a resource for societies, and as a consequence they are constantly under anthropogenic pressure. Population growth coupled with change in land use, hydrologic regimes and climate are stressing these systems worldwide, threatening their function as sources for drinking water, socio-economic activities and ecological environments. Human activities impact the food web structure and biodiversity of lakes and reservoirs favoring the invasion of alien species and throughout intense fishing or aquaculture [1]. Nutrient inputs to lakes from crops lead to eutrophication and hypoxia phenomena.

Lake and river ecosystems need proper monitoring to gain better understanding of their dynamics and possible changes in their properties. Over the last decades, there has been an increase in the capacity and availability of remote sensing imagery from satellites to monitor lake systems worldwide, promoting usage and creating new demands and opportunities for reliable remotely sensed data sets. New and traditional techniques are used efficiently to follow the impact of protection policies (e.g. EU's Water Framework Directive, WFD). Remote sensing of ocean colour has an important role as a cost-effective tool for global and frequent observations of parameters like chlorophyll-a (Chla), suspended matter (SPM) or chromophoric dissolved organic matter (CDOM). Old and newly launched satellites, such as the the European Space Agency's (ESA) Sentinel-2, with the MultiSpectral Imager (MSI) on board, and especially the Sentinel-3 satellite, with the Ocean Land Colour Imager (OLCI) sensor, are devoted to this purpose. The constellation missions of Sentinel-2A/B and Sentinel-3A/B and follow ups provide unprecedented monitoring capabilities for lake water quality because of the favorable band settings, high signal-to-noise ratios, good spatial resolution, and overpass frequency.

The approaches used to derive bio-geochemical parameters from remote sensing data are based on inversion algorithms, in many cases specifically designed for a delimited area or region, and are usually based on field data. In fact, many empirical or semi-analytical inversion algorithms depend too much on the *in situ* data used for their development, and the scalability to larger areas or different water conditions is difficult. Since the variability of water properties, especially on coastal areas, is so high, many local or regional inversion algorithms are needed to cover the temporal and spatial variability of the many different water property conditions. This task could be facilitated if the water type would be known in advanced, i.e. by using an optical classification, assuming that the results obtained for one optical water type (OWT) are applicable to any water body associated with that type; that is, "grouping waters with similar optical traits and develop adapted algorithms for each water class" [2, 3].

## 1.2. Previous studies and state of the art

Classification schemes are more common to terrestrial imagery, but are gaining traction in aquatic applications, and share basic similarities [4–6]. Both in land and water, the classification systems are based on features (i.e., spectral channels) in a spectral signal related to underlying types with ecological meaning. The features are usually derived from the spectral reflectance shape and magnitude, and are ultimately limited by the spectral resolution of the sensors when utilized for image classification. The notion of optical type has its origins in the work of Jerlov [7, 8], where water types were defined by the diffuse attenuation coefficient of downwelling light. These *Jerlov* water types are still used in marine applications [9], and were used in a recent modeling study to generate inherent optical properties (IOPs) for each type [10], directly utilizing type-specific parameters.

In practice, optical water types (OWTs) are derived from averaging grouped  $R_{rs}(\lambda)$  spectra that share characteristics, where each individual spectrum is an instance along an optical continuum bound by the outer ranges of the environmental and optical conditions of all water systems. The goal of the OWT is a meaningful partitioning of the full multi-dimensional  $R_{rs}(\lambda)$  space into a set of optical water types. One of the more popular type of OWT implementations is that of Moore et al. (2001) [11], based on the application of the fuzzy c-mean (FCM) algorithm [12] to a reduced  $R_{rs}(\lambda)$  dataset. These data are transformed to sub-surface values following [13]. It should be noted that the clustering and ensuing membership functions use the below-water quantity, but we will retain referencing any spectra as  $R_{rs}(\lambda)$  for simplicity.

The FCM algorithm partitions the input data into a specified number of clusters. The function operates by minimizing the distance between the data points and the prototype cluster centres (means), which are iteratively adjusted until some optimization criteria are met. Since the number of clusters is not known beforehand, FCM is applied to the data set over a range clusters set from 2 to 20. Cluster validity functions are used to assess the effectiveness of the cluster performance for each outcome. These functions measure various aspects of the entire cluster partitioning, and were used to guide the ultimate choice for the number of optimal clusters [3, 6]. The clusters define the OWTs through their means and covariance matrices. The OWTs are created with the full hyperspectral data allowing for the construction of a membership function - the main component of the classification tool that produces the image classification - to operate on any band configuration within the range of hyperspectral data (400 – 800 nm), and thus on any satellite sensor.

There are two different forms of  $R_{rs}(\lambda)$  used in classification schemes for depicting OWTs: area-normalized  $R_{rs}(\lambda)$  e.g., [14, 15] and un-modified or non-normalized  $R_{rs}(\lambda)$  e.g. [3]. The rationale behind normalizing is to remove the influence of magnitude on clustering and stressing the spectral shape. Vantrepotte et al. 2011 [14] showed that in the case of coastal turbid waters, those are susceptible to magnitude shifts based on concentration of particles of the same type, which are sorted into the same cluster when normalized. Absorption characteristics have more impact on clustering.

The separation of the waters in optical groups or classes have been a matter of previous studies with developed approaches based on sets of different bio-optical parameters [16–18] or using remote sensing reflectance spectra  $R_{rs}(\lambda)$  together with *in situ* data [11, 19, 2]. Since the use of *in situ* data is not sufficient for relating those with in-water optical properties, usually due to a low number of samples, satellite observations have been used as basis for the characterization of water types [15].

Regardless of the method, optical water types (OWT) provide information on the spatial distribution of optical states across image scenes when applied to satellite data. These mapped products function as weighting factors for optimizing bio-optical algorithms and product uncertainties for image scenes [3, 13, 19, 20]. In these cases, they are intermediary products that are not needed themselves for analysis, and are invisible to users. However, OWTs are depictions of optical states, providing information on underlying water conditions that in, and of themselves, have intrinsic ecological value. They have been used directly for interpretive analysis for ecological diversity [15] and ecological patterns [21] that may not be obvious from other

bio-optical products such as chlorophyll-a concentration, which, at the same time, may be hard to retrieve in complex lake waters because of the complex atmospheric and in-water optical properties. In some cases, OWTs have been linked to distinct optical phenomena that relate to specific phytoplankton types [22]. These studies collectively illustrate the varying roles and uses for water types, whether freshwater or marine, when applied to remote sensing data.

### 1.3. Optical water types classification applications for environmental monitoring

In this project, we are mainly interested in the development of techniques to improve the OWT classification. However, a way of testing the accuracy of the results is precisely to check how useful they could be in different applications.

OWT results are used for different environmental studies, usually as intermediate products that help to obtain other parameters:

- Ocean colour algorithm development and blending at pixel scale to obtain highly accurate water quality concentration values [3, 23, 6].
- Phytoplankton functional type identification [24].
- Algae bloom identification: cyanobacteria [25] and coccolitophore [22].
- Water dynamics and change monitoring [15, 26].

Data collection and model development will take most of the activities in the first year. Two main applications have been identified and will be performed in the second year of the project: (1) compare the performance of chlorophyll-retrieval algorithms in each optical class and blend them at per-pixel scale on several well-known lakes; (2) to check the capabilities of the new OWT to detect several types of algae blooms, like cyanobacteria in the Albufera Lagoon or the Baltic Sea.

### 1.4. Goals and objectives

The main goal of the AQUACLASS project is to generate optical classification schemes for inland waters that can be useful for ocean colour extractions and its applications. Alternatives to the semi-supervised fuzzy-c means, and novel machine learning classification approaches, will be used for this purpose. Machine learning approaches will be considered, having in mind that algorithms for EO applications need to be guided both by data and by prior physical knowledge. A training data set will be used to define the optical classes. This data set will consist of satellite and *in situ* data, which should be representative for the optical variability found in nature in inland waters, yet maximizing the geographical coverage and seasonal variance.

The methodologies proposed here will follow two different pathways: (semi)supervised classification through different flavours of kernel machines, random forests and neural nets; and as alternative we will also explore the field of nonlinear dimensionality reduction followed by linear (or simple nearest neighbour) classification. The first approach is a direct one and promises improved accuracy, while the second approach will yield a set of useful nonlinear spatial-spectral features driving the different water types.

Regarding the challenges identified in the R+D+I Spanish Strategy Plan, this project will deliver novel tools and methods addressed to the challenge (5): actions on "Climate Change and efficient use of resources". In addition, these challenges perfectly fit within the European Commission framework programmes for research and innovation Horizon 2020 (CyanoAlert, GLaSS) and the following Horizon Europe, which is aimed at securing Europe's global competitiveness. The benefits of these EU programs allow the creation of new business opportunities and, transversely, to all sectors of the economy both nationally and regionally. Our proposal will help achieve these objectives through the development and improvement of machine learning models for (pre)classifying water types with satellite images. From a socio-economical perspective, the AQUACLASS project will contribute to the pre-operational capacities of the EU Copernicus programme in the context of ocean colour by providing advanced methodologies suitable for reliable classification, analysis, and retrieval of water products. This would have a direct effect on the accuracy of water quality products (WQ), making EO data more reliable and usable, i.e. for reporting to the EU Water Framework Directive (WFD). Less uncertainty in WQ means

better analysis and improved knowledge about lakes and reservoirs of Europe, both spatially and temporally. It helps detect water emergencies faster and shortens the reaction time for mitigation in case of e.g. water pollution or anthropic eutrophication, harmful algal blooms, etc. Economically, this would mean a reduction in the regional and national budgets spent on environmental emergencies, and socially, faster reaction leads to less, or at least more controlled, damage to those using water for recreational, watering or drinking purposes. The project's links with other international initiatives such as the H2020 CyanoAlert project, Globolakes, the future IOCCG OWT-WG, is an additional advantage for starting new international project proposals in the future, through the Horizon Europe or other calls from ESA or EUMETSAT.

The main goal translates into the following specific objectives:

1. **Improve the characterization of uncertainties per optical water type.** To use a large database of remote sensing reflectance, chlorophyll concentrations, inherent optical properties and diffuse attenuation coefficients (i.e. GLaSS, CyanoAlert, Limnades databases) to compute the uncertainties estimates.
2. **Improve OWT classification results using machine learning approaches.** Testing different approaches and compare results with the fuzzy-c means established approach will determine the performance of the proposed methods. Metrics like the bias and the root-mean-square difference (RMSD) at per-pixel level will be used.
3. **Check the applicability of the OWT results using practical use cases.** Water body classification is, in fact, and application in itself, as a requirement to derive information for the WFD in some European countries (e.g. Sweden). Other study cases, like algorithm blending to obtain improved retrievals; or cyanobacteria bloom detection and alert in high eutrophic lakes or regions, will be tested. For the former use case, we will compare the performance of various chlorophyll-retrieval algorithms in each optical class. One possible application of the knowledge acquired would consist in blending algorithms at each pixel to obtain a highly accurate chlorophyll concentration value [6]. For the latter, we will work together with the CyanoAlert team, and investigate if the new classification schemes could provide useful pre-knowledge for the application of algorithms like the maximum peak height (MPH) for cyanobacteria prompt alert.

## 1.5. Proposed methodology.

### ***Work Package 1 (WP1). Collection of the spectral data set***

Optical classification methods use spectral radiance/reflectance measurements (usually  $R_{rs}$ ) to define the characteristics of the different typologies of the water. The *in situ* information is related with in-water optical properties, and a complete overview of all possible relationships is needed, but complicated to obtain exclusively through *in situ* measurements, due to the scarcity of samples, in relation with the number on inland waters and its dynamics. The spatial and temporal dynamics of the inherent optical properties (IOPs) associated to each optical water class are the main features analyzed for the classification, since the light attenuation of the optically active components of the water regulate its colour. Of course, these IOPs are related to physical and bio-geochemical processes. In the proposed project we will use a combination of *in situ* knowledge and satellite information for the definition of the water typologies. Satellite observations add valuable (and numerous) samples used to extent spatially the data set, and can help to define the characteristics of the water types, i.e. through the manually labeling of pixels by an expert, associating the spectral to representative water classes. The breakdown of this workpackage presents the *modus operandi*.

- **WP1.1 Check availability and usability of external spectral data sets.** There are several *in situ* datasets available that have been used or could be used as basis for the classification development. To start with, the GLaSS project developed an OWT implementation based on that of Moore et al. (2014) [3], with the addition of a larger variety of lakes. The objective of the GLaSS OWT implementation was to develop a non lake-specific classification tool for all lakes and conditions using the fuzzy-c mean clustering approach. To achieve this, GLaSS assembled a data set of *in situ* hyperspectral  $R_{rs}(\lambda)$  with co-measured Chl-a and



TSM concentrations and absorption of CDOM at 443 nm from multiple sources covering a wide dynamic range in optical and environmental conditions. This data set includes the "lake only" dataset portion (N=320) from [3], which consists of measurements from the northeast U.S., the Great Salt Lake [27] and across Spain [28]. These data were combined with the GLaSS data, which consists of  $R_{rs}(\lambda)$  with co-measured Chl-a, TSM and aCDOM from different countries. A large range of water quality concentrations are covered, including very high concentrations (Chl-a  $> 900 \text{ (mgm}^{-3}\text{)}$ , TSM  $> 200 \text{ (mg}^{-3}\text{)}$ , CDOM  $> 30 \text{ (443 m}^{-1}\text{)}$ ), representing a large variety of optical conditions. A big portion of the GLaSS dataset was later included in the LIMNADES dataset (<https://www.limnades.org/datasets.psp>). LIMNADES is a centralised database for in situ bio-optical measurements and satellite match-up data from lakes and other inland waters worldwide. The database is held in trust and maintained by the UK GloboLakes project ([www.globolakes.ac.uk](http://www.globolakes.ac.uk)). Some other projects that store lake's *in situ* data are the CyanoAlert (H2020) and the ESAQS (Prometeo, GVA). The *in situ* datasets will be studied in order to analyse the relation between the IOPs, concentrations and radiometry for a set of selected lakes or reservoirs. The use of auxiliary available variables, e.g. climate or topographic data, and previous knowledge of the study areas will be also considered for the feature selection and preparation of the classification models.

- **WP1.2 Collect satellite images.** The Sentinel-2 Multispectral imager (MSI) sensor will be used. Level-1C data (Top-Of-Atmosphere reflectance) will be downloaded from the Copernicus Open Access Hub (<http://isp.uv.es/soft.htm>) and will be processed to Level-2 geophysical products using the Case 2 Regional Coast Colour, C2RCC [29] algorithm available in the SentiNel Application Platform toolboxes (<http://step.esa.int/main/toolboxes/snap/>). The developments will be focused on a set of well-studied lakes linked to H2020 projects where the proposed PI was/is part of the research team or has collaborated as external expert: CyanoAlert (<http://cyanoalert.com/>), Global Lakes Sentinel Services (<http://glass-project.eu/>) and ESAQS. The MSI sensor has appropriate spatial and temporal resolution for the study of a large number of global lakes and other types of inland water. The Sentinel-2 mission is based on a constellation of two satellites (Sentinel-2A and Sentinel-2B), both orbiting Earth at an altitude of 786 km but 180° apart. This configuration optimizes coverage and global revisit times. As a constellation, the same spot over the equator is revisited every five days, and this is even faster at higher latitudes. The MSI covers 13 spectral bands (443 nm–2190 nm) with a swath width of 290 km and spatial resolutions of 10 m (4 visible and near-infrared bands), 20 m (6 red-edge/shortwave-infrared bands) and 60 m (3 atmospheric correction bands). One example of a red-green-blue composition of an MSI image over the Karavasta Lagoon is shown in Figure 1. MSI collection will be made at the beginning of the project, and it will be used for three tasks: 1) extraction and labeling by an expert of  $R_{rs}(\lambda)$  of several pixels from selected lakes, with uniform distribution through the available time series (2015-2019), to increase the number of observations part of the training data set (see WP1.3); 2) use the images over the selected lakes to apply the OWT classification and analyze/validate the results (see WP2.1, WP2.2); 3) use the images for the case studies indicated in WP3.1 and WP3.2.
- **WP1.3 Collection of new data by labeling pixels.** The sampling protocol to extract reflectance information from MSI pixels will be carefully planned once the decision on the selected lakes has been done. The sampling constitutes a previous important step since we expect to have a high representation of the different water types that can be found in nature. The insufficient representation of those types in the *in situ* datasets will determine which data are missing and we will try to complete it with EO labeled samples. The labeling of the pixels of interest can be done in two ways: 1) manually labeling the valid pixels by an expert, previously screen out for avoiding cloud contaminated or other problems that could lead to false reflectance data. Brockmann Consult GmbH developed a few years ago the PixBox interface method for this purpose [30]; b) label pixels in a semi-automatic way with a user-driven methodology proposed in the PhD thesis of Dr. Gomez-Chova (2007) for MERIS, where the labeling of clusters found in the image is done by the expert too.



Figure 1: The image of the Karavasta Lagoon in Albania is a subset from the first acquisition by Sentinel-2B on 15 March 2017, RGB enhanced composition. Credits: ESA.

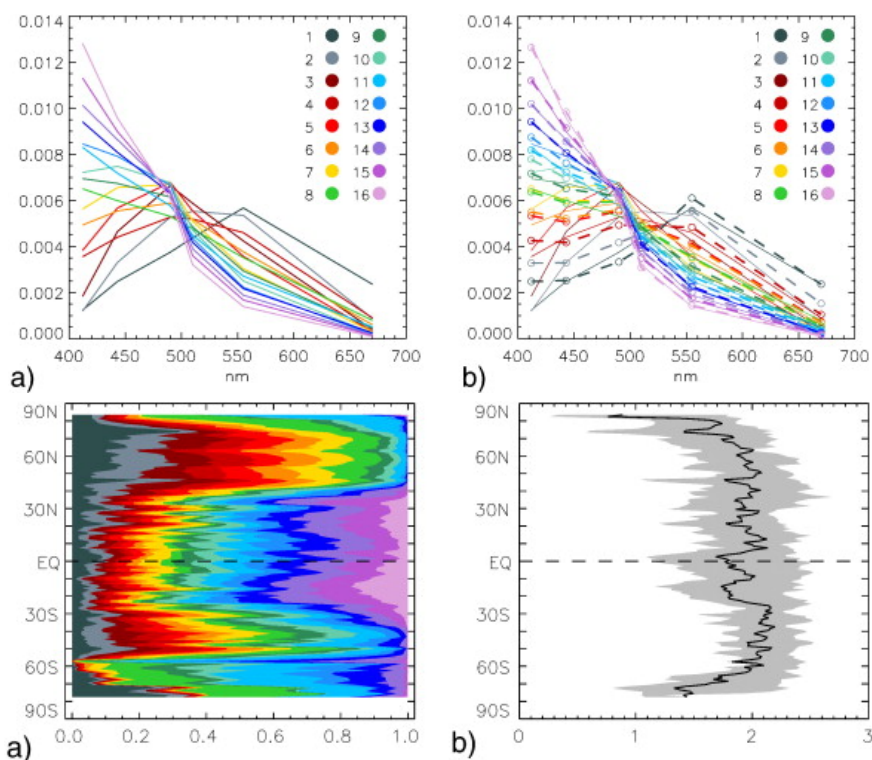


Figure 2: Top row: Example of derived global marine OWT a) with average normalized spectra, and b) with Case 1 model (top). Bottom row: The latitudinal dependence of a) the average class, and b) the Shannon index (from Mélin and Vantrepotte, 2015).

### Work Package 2 (WP2). Improve classification of optical water types

The fuzzy logic classification scheme has been used profusely in remote sensing ocean colour studies. The method, developed and improved by Tim Moore and colleagues (2001,2014,2017), "allows for pixels to be assigned partial or grader class memberships to different water types with which they share spectral characteristics" using a fuzzy membership function that shows the likelihood of one pixel to belong to a class or another. A number of plausible classes is defined, and the class membership is use to weight the output of class-specific algorithms. A definition of the water classes is then expected before-hand and for this the spectral reflectance characteristics of the distinct water set must be defined. This definition is done by cluster analysis, and the statistical properties for each cluster found with the training set becomes the basis for the membership to each class. This approach is valid for any other type of semi-supervised

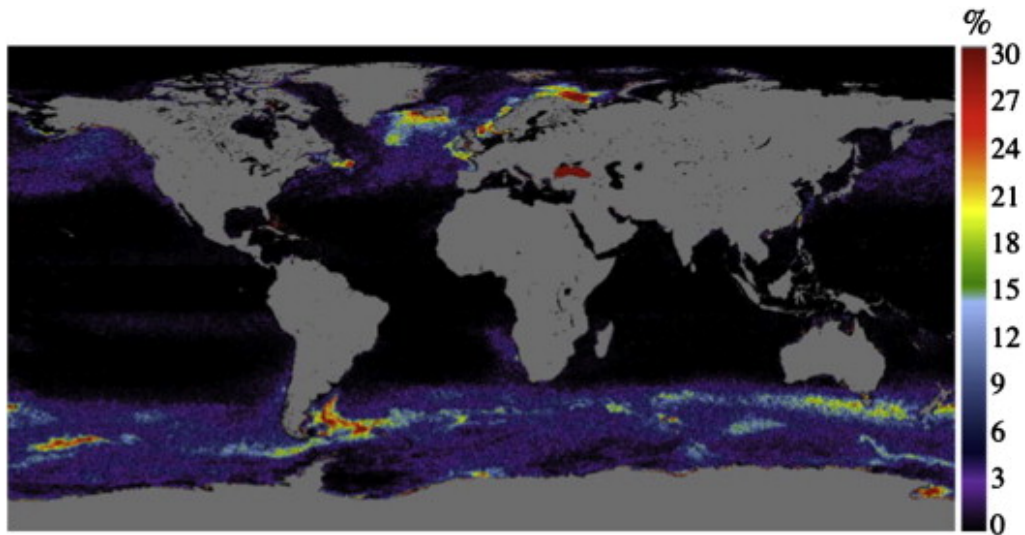


Figure 3: Example of the global frequency for coccolithophore blooms derived from 8-days averages of the dominant OWT maps (from Moore et al., 2012).

classification method, like the ones proposed in this project. The first step is, therefore, understanding the training data and how the different available variables are related to be able to define water typologies.

- **WP2.1 Unsupervised feature learning for classification and understanding.**

Our first step to improve and deal with OWT classification problems is to extract and learn key features from data sources. To this end we will rely on statistical feature extraction methods. There are two main reasons supporting this decision. First, feature selection algorithms are conceived to extract key information from the data, which allows to employ simpler classification algorithms instead of complicated, more costly ones. For instance, it is possible to obtain equal or better results than powerful classification methods by simply using a combination of a Principal Component Analysis (PCA) followed by a simple, eventually linear, classifier. The second advantage of applying feature selection algorithms is that, for medium/high dimensional input samples, the selected features can be represented and visualized, and experts and scientists can extract important information from them.

- **Nonlinear kernel-based multivariate data analysis.** Statistical multivariate analysis (MVA) is commonly used for feature extraction in order to understand the relationships between variables and their relevance to determine a dimensional reduction. The feature extraction process is performed by projections that synthesize the information of the original system into a subset of independent and uncorrelated variables (or features) with less dimensions. The original samples are projected onto the most relevant directions of the manifold by means of mathematical transformations depending on the characteristics of the original samples. The transformations are defined by a mathematical function:  $f : \mathbb{R}^d \rightarrow \mathbb{R}^{d_f}$ , where  $d$  is the number of dimensions of the examples in the original space, and  $d_f$  is the dimension of the transformed space, typically  $d_f \leq d$ . The feature extraction methods can be divided in different groups depending on the type of the transformation as well as the dependence on the available labeled data. Among all MVA feature extraction methods, (PCA) [31] and partial least squares (PLS) [32] are the most common methods, which assume that there exists a *linear* relation between the original features.

Linear MVA methods are differentiated between unsupervised and supervised methods. While PCA [31] does not require or ignores the output labels (target variables), other methods like PLS [33], orthonormalized PLS (OPLS), and Canonical Correlation Analysis (CCA) [34], may exploit such information to define the transform and act as supervised methods. The main assumption of the linear feature extraction methods is that there exists a *linear* relation between the input and/or output features. However, this is not true in general in real remote sensing problems. The multi-scattering in the acquisition process, the hetero-



geneity at subpixel level, as well as the impact of atmospheric and geometric distortions are some of the factors related to nonlinearities. Therefore, using non-linear methods has become a hot topic in the last years in remote sensing. A natural way of transforming linear to nonlinear MVA methods is by means of kernel methods (see [35] for a comprehensive review). To obtain the kernel version we only have to replace the original centered data matrix by the centered data in a feature space where the samples has been mapped to using a nonlinear function, and then apply the *kernel trick* to replace all appearances of dot products with a reproducing kernel function (see [36] for details). Once the nonlinear feature extraction method has been applied, linear classification or regression methods can be used. Thanks to the nonlinear feature extraction method, and despite the original data being nonlinear, linear methods applied to the projected data achieve high efficiency because of mapping the original data into a Hilbert, high dimensional space.

Among the feature extraction methods that we will apply to the OWT classification challenge, it is worth to mention PCA [31], PLS and OPLS [32], CCA [34], and their corresponding nonlinear kernel counterparts, kPCA [37], kPLS and kOPLS, and kCCA. A comprehensive review of all these methods is in [38].

- **Sparse coding and convolutional neural networks.** As an alternative pathway to kernel methods, we will consider deep convolutional neural networks (DCNN) [39], and in particular we will rely on sparse coding as a guiding criterion for feature extraction.

In remote sensing, autoencoders have been widely used [40–43]. However, a number of critical free parameters are to be tuned; regularization is an important issue, which is mainly addressed by limiting the network's structure heuristically; and only shallow structures are considered mainly due to the limitations on computational resources and efficiency of the training algorithms. On top of this, very often, autoencoders employ only the spectral information, and in the best of the cases, spatial information is naively included through stacking hand-crafted spatial features.

Despite the wide adoption of *supervised* deep convolutional networks in our field, there is still little evidence of performance of *unsupervised* deep architectures in remote sensing image classification: [44] introduces a deep learning algorithm for classification of low-dimensional VHR images; [45] explores the robustness of deep networks to noisy class labels for aerial image classification; and [46] introduces hybrid Deep Neural Networks to enable the extraction of variable-scale features to detect vehicles in satellite images; [47] proposes a hybrid framework based on Stacked Auto-Encoders for classification of hyperspectral data. Although deep learning methods can cope with the difficulties of non-linear spatial-spectral image analysis, the issues of sparsity in the feature representation and efficiency of training algorithms are not obvious in state-of-the-art frameworks.

In this project we aim the combination of *greedy layer-wise unsupervised pre-training* [48–51] coupled with the highly efficient Enforcing Lifetime and Population Sparsity (EPLS) algorithm [52] for *unsupervised* learning of *sparse* features and show the applicability and potential of the method to extract hierarchical (i.e. deep) sparse feature representations of remote sensing images. The EPLS seeks a sparse representation of the input data (remote sensing images) and allows to train systems with large numbers of input channels efficiently (and numerous filters/parameters), *without requiring any meta-parameter tuning*. Thus, deep convolutional networks are trained *efficiently* in an unsupervised greedy layer-wise fashion using the EPLS algorithm to learn the network filters. The learned hierarchical representations of the input remote sensing images are used for image/pixel classification, where lower layers extract low-level features and higher layers exhibit more abstract and complex representations. To our knowledge, this will be the first work dealing with *sparse unsupervised deep convolutional networks* in water applications.

## • WP2.2 Supervised classification schemes.

While WP2.1 is devoted to extract key features from data sources, the second is tied to the OWT classification acting over these extracted features. In what follows, we describe some of the classification algorithms we will apply.



- **Support Vector Classifiers.** The support vector machine for classification, SVC [53], is a well established, state of the art algorithm for nonlinear binary and multiclass classification. It works by mapping the input (training) samples into a high dimensional Hilbert space using a (in principle unknown) nonlinear function where a linear classification is performed by defining an hyperplane that separates the different classes with maximum margin. The most relevant characteristics of this type of classifiers are that i) they obtain an sparse solution, where only the most relevant training samples are used, which are called *support vectors*, ii) they are able to deal with noisy input samples, iii) they generalize well, that is, they are able to make good predictions on new, unseen samples, and iv) they are relatively fast to train, and once a model is obtained, very fast for making predictions on new input samples.

Besides the original SVC algorithm, we have worked on improvements for this algorithm that are relevant for the current project, such as:

- \* *Semi-supervised support vector machine.* The standard SVM is a supervised algorithm which needs labeled training samples to build a model<sup>1</sup>. SS-SVM [54–56] is a modification of the SVM that allows to include unlabeled samples into the SVM formulation in order to obtain a model that uses both labeled and unlabeled samples, improving classification results. The information of unlabeled samples is included by building a graph Laplacian using all samples (labeled and unlabeled), and then a spreading function is applied iteratively until convergence. It incorporates contextual information through a full family of composite kernels. It has been successfully applied to hyperspectral image classification [57].
- \* *Bag-SVM.* Like the SS-SVM, the purpose of the Bag-SVM algorithm [58, 59] is to exploit the information provided by unlabeled samples. In this case, the method exploits the wealth of unlabeled samples by regularizing the training kernel representation locally by means of cluster kernels. The method learns a suitable kernel directly from the data, and thus avoids assuming *a priori* signal relations obtained when using a predefined kernel structure. The method scales almost linearly with the number of unlabeled samples and provides out-of-sample predictions.

In the context of the current project, the use of this kind of semi-supervised models is very promising and relevant, because in OWT classification we often have to deal with datasets having few labeled samples, but with many unlabeled samples that, despite being of unknown class, contain important information that can be exploited to make better classification.

- \* *One-class support vector machine.* OC-SVM [60], also formulated as one-class support domain description (OC-SVDD) [61], is a type of algorithm belonging to the family of one-class classifiers. The goal of these classifiers is to identify just one, particular class, rejecting the others. In these type of problems, one has enough information about the properties of a type of object (a particular class), but has little or none information on others objects or classes. We have successfully used OC-SVM for one-class remote sensing classification problems, and also to build multi-class classifiers by employing ensembles of one-class classifiers [62]. Moreover, there is also an extension of the OC-SVM, the semi-supervised OC-SVM, that, as in the SS-SVM, allows to include unlabeled samples in the training process, improving the classification results [63].
- \* *Virtual SVM.* The V-SVM is an extension of the SVM that adds new, virtual labeled samples to the original training set [64]. These new samples are built from the original training samples by applying spatial operators such as rotations, inversions, mirroring, etc. The advantage of this simple, yet powerful strategy, is two fold: first, it includes a variety of new examples in the training set, making the model more robust to changes in scenes like rotations, shifts, etc. Second, it increases the size of the training set, alleviating the problem of having datasets with a low number of labeled samples.

- **Neural networks.** NN [65, 66] are classical methods for classification. In the last years,

<sup>1</sup> In classification problems, labeled samples mean that we known the class of these samples.

they have received further attention and obtained impressive results in classification problems thanks to the use of Deep Convolutional Neural Networks (DCNN) [39], a kind of NNs with many layers, containing hundreds or even thousands of neurons, which implement convolutional filters, making them specially suited for classification scenarios involving images. In essence, the different layers of these DCNNs extract key spatio-spectral features from images, which in turn greatly facilitates the classification of the scenes and the objects present in images. Although mathematically theorized years ago, it was very hard, almost impossible to train these DCNNs until the recent appearance of new hardware facilities, such as graphical processing units (GPUs), having thousands of processing units that allow training these networks, usually having millions of parameters to adjust. There are also new powerful software and library tools, such as Theano [67] or TensorFlow [68], that make these resources easy to use for scientists.

- **Random Forests.** RFs [69] are also classical algorithms based on the ensemble of hundreds or thousands of classification (or regression) trees. The main idea of RFs is that the results obtained by averaging simple classifiers (trees in this case) can obtain better results than the ones obtained by single, more sophisticated and powerful classifiers. Important and relevant features of RFs are that they are fast to train and in making predictions, that they are easily parallelizable, making them specially suitable for new hardware composed of multi-core systems, and that they provide with a feature ranking (or variable importance) which indicates the most important variables. They are known for obtaining quite good results in remote sensing classification problems [70–72]. For the projected project, RFs can perform better than the classification algorithms used so far, and can also be very useful when dealing with large datasets such as LIMNADES and others.
- **WP2.3 Developed and release classification software.** The classification approaches tested will be coded in Python and updated regularly using the GitHub facilities. Several methods are proposed in WP2.2, and many of them are already available to the community. Jupyter notebooks will be prepared for other scientists to test the approaches with their own data. A link to the HypeLabelMe <https://hyperlabelme.uv.es/> benchmark system will also be done. A plug-in for SNAP, the Sentinel Application Platform (<http://step.esa.int/main/toolboxes/snap/>) will be provided as well. SNAP is an open source common architecture for ESA Toolboxes ideal for the exploitation of Earth Observation data. The IPL team regularly releases software and databases to the scientific community at: <http://isp.uv.es/soft.htm>. This will not be an exception and the possible improvements of the classification approaches for OWT will be uploaded and properly documented as soon as the project is finalised.

### ***Work Package 3 (WP3). Applications for ocean colour and environmental studies.***

The combination of the OWT results and the analysis of similarities between lakes or regions, together with factors related to light availability or aquatic life communities could be of support for understanding the variability of the biogeographic provinces [73]. Lacustre biodiversity can also be linked to the results of optical diversity, due to the variations in optical types at any location through a time period (Figure 2). Since the datasets to be used in this project will be limited by the number of data sets of the selected lakes and the short operational period of the Sentinel-2 data (2015-present), we will focus the applications on local or regional case studies. Two possible cases are a) the use of the OWT for the selection of the right OC algorithm; b) the use of OWT result for identifying algae blooms of different species.

- **WP3.1 OWT schemes for improving ocean colour retrievals.** One application of the OWT results is the selection of algorithms for specific regions; or if it be used to define the applicability of the different algorithms or to reduce their uncertainty. Blending of algorithms that work properly for one or another OWT, is also one recurrent use [19, 2].
- **WP3.2 OWT for algae bloom types discrimination.** Links between optical diversity and biodiversity are related to the light variation in quantity and spectrum. For instance, coccolithophores produce small calcium plates or coccoliths to cover the living cell. A coccolithophore bloom is visible from space since they have a very clear signature in the visible

range, generating a very bright patch of turquoise color water. This phytoplankton species live under very variable conditions, but they can be mainly found at mid to high latitudes, with a specific light dynamic and salty waters. OWT methods have been used to detect coccolithophore bloom in the ocean with success, but still with some room for improvement (Figure 3). Similar approaches can be developed to detect cyanobacteria blooms in inland water, taking into account the spectral reflectance characterization of the cyanobacteria. Cyanobacteria are photosynthetic bacteria that share some properties with algae and are found naturally in lakes, streams, ponds, and other surface waters. Since these blooms can be harmful, early detection is compulsory to keep the good quality of the water. This is precisely the main objective of the CyanoAlert project, where AQUACLASS results can be very useful to help with this task.

#### **Work Package 4 (WP4). Project management and technology transfer.**

The AQUACLASS project will involve some managerial as well as technological transfer tasks. The general managerial activities of AQUACLASS 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; and (3) ensure timeliness of 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 AQUACLASS project, e.g. see <http://isp.uv.es/projects.htm>. Frequent follow-up meetings in the group will be minuted on a monthly basis. In summary, the WP4 will imply three main activities:

- **WP4.1 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, physics.geo-ph), and we plan to continue with this open access philosophy.
- **WP4.2 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/>.
- **WP4.3 Attendance of conferences and workshops.** We will disseminate the main achievements of the project through conferences and workshops i.e. the Living Planet Symposium, the International Ocean Colour Science meeting, etc. getting together with key scientists in the fields of ocean remote sensing and machine learning, and with interested users. Three members of the working team are included in a recently created international OWT working group, with scientist from several countries in Europe and USA, that are proposing the integration of the group into the International Ocean Colour Coordinating Group (IOCCG, <http://ioccg.org/>).

#### **1.5. Chronogram**

The AQUACLASS project will develop new methodologies for ocean colour and remote sensing data processing. The overall Workpackage Breakdown structure of the project is shown in Table 1. We have identified three types of workpackages: *theoretical* (WP1-WP2), *applied* (WP3) and *managerial* (WP4).






The AQUACLASS *research team* is formed by two professors:

- **Asst Prof. Ana B. Ruescas Orient** (Department of Geography & Image and Signal Processing (ISP), Univ. València, Spain). *Principal investigator. Expertise in EO processing and interpretation: retrieval and validation of OC biophysical parameters, optical and thermal image processing.*
- **Prof. Dr. Jordi Muñoz-Marí** (Image and Signal Processing (ISP), Univ. València, Spain). *Expertise in retrieval of biophysical parameters, support vector machines for regression and anomaly detection, optimization and large-infrastructure management.*

The AQUACLASS *working team* is formed by scientists from ISP and external collaborators in

Table 1: *Workpackage breakdown structure and Gantt diagram of the AQUACLASS project. Project leader in each task is **boldfaced**.*

Work package	2019	2020	Personnel <sup>†</sup>
WP1. Collection of datasets			
1.1. Check external datasets			<b>AR</b> , C1
1.2. Collect satellite images			<b>AR</b> , C1
1.3. Collection of new data by labeling pixels			<b>AR</b> , PP, C1
WP2. Improve classification of OWT			
2.1. Unsupervised feature learning			AR, <b>JM</b> , TM, GC, C1
2.1. Supervised classification			<b>JM</b> , AR, TM, LG, C1
2.2. Classification software			<b>JM</b> , AR, GM, C1
WP3. Applications for ocean colour and environmental studies			
3.1. OWT schemes for improving OC retrievals			TM, <b>AR</b> , GC, C1
3.2. OWT for algae bloom types discrimination			PP, <b>AR</b> , C1
WP4. Management and transfer			
4.1. Reports			<b>AR</b> , JM
4.2. Toolboxes and databases			<b>JM</b> , GM, C1
4.3. Conferences/workshops			All

<sup>†</sup> AR: AB. Ruescas; JM: J. Muñoz; GC: G. Camps-Valls; PP: P. Philipson; LG: L. Gómez-Chova; TM: T. Moore; GM: G. Mateo-García; Contract C1 : Publications, : Databases, : Sentinels EO data, : Software, : Conferences or workshops.

different aspects, both theoretical and applied:

- **Prof. Gustau Camps-Valls** (ISP, Univ. València, Spain). *Expertise in machine learning and signal processing for EO data analysis.*
- **Prof. Luis Gómez-Chova** (ISP, Univ. València, Spain). *Expertise in machine learning and signal processing for EO data analysis, with focus on classification approaches.*
- **PhD Candidate Gonzalo Mateo-García** (ISP, Univ. València, Spain). *Programmer and machine learning expert, focused on deep learning and segmentation for EO imagery.*

The two external collaborators working on different aspects, both theoretical and applied, relevant for WP2 and WP3 are:

- **Dr. Petra Philipson** (Brockmann Geomatics Sweden AB). *Remote sensing consultant with expertise in ocean colour, processing and analysis of in situ and EO data.*
- **Dr. Timothy Moore** (Univ. New Hampshire, USA). *Expertise in ocean colour, processing and analysis of in situ and EO data, with focus on classification and retrieval of OC biophysical parameters.*

## 1.6 Materials, infrastructures and equipment currently at our disposal

The members of the AQUACLASS team are part of the Image and Signal Processing (ISP) group, a multidisciplinary team of geo-physicists, mathematicians and engineers with common interests in remote sensing image processing and statistical learning. The ISP is part of the interdisciplinary research unit, Image Processing Laboratory (IPL), located in the 'Parc Científic' of the University of Valencia. The IPL is divided in four groups (about 40 researchers in total) covering various aspects of image processing for remote sensing: ISP for signal and image remote sensing data processing, UCG for global change studies, GACE for space technology, and LEO, the Laboratory for Earth Observation, for statistical retrieval of biophysical parameters.

The IPL is equipped with excellent facilities, access to image acquisition equipment, antennas for reception of satellite images as well as access to satellite images through several ESA, EUMETSAT, Google, and NASA projects. We count also with computer servers and have access to external computer grids, such as *Tirant* and *MareNostrum* facilities.



## 1.7 Justification for contracting people

The ISP (<http://isp.uv.es>) is an relatively new emerging group: 6 permanent, 1 Ramón y Cajal and 12 temporal members; and a highly productive research team in scientific terms ( $h$ -index=63 and 15 JCR journal papers/year of the research team). ISP is obtaining excellent research results and recognition in the last years through National projects, ESA and EUMETSAT projects and one ERC Consolidator Grant (SEDAL 2015-2019) to Prof. Camps-Valls <http://isp.uv.es/projects.htm>. We are internationally recognized and networked, we are involved in the organization of international workshops, special sessions and IEEE conferences, we are invited to keynote talks, etc. Key collaborations are currently on-going with the Barcelona Expert Center (Dr. A. Turiel and Dr. M. Vall-Ilossera), Dept. Biogeochemistry at MPI-Jena (Prof. M. Reichstein), German Aerospace Center (Dr. T. Jagdhuber) and the Massachusetts Institute of Technology (Prof. Dara Entekhabi), ESA (ISP is involved in the Phi-lab) and EUMETSAT. The group's multidisciplinary approach allows to tackle geoscience and remote sensing problems from a broad perspective, ranging from image classification, anomaly detection, and time series analysis, to bio-geophysical parameter retrieval and model inversion. This multidisciplinary background places the group in a privileged position for both developing innovative approaches and for training new PhD/master students.

We think that the ISP can be consolidated as a key group at European level in a few years from now. At this moment, we only count with 2 full-time PhD students (Gonzalo Mateo and José A. Padrón), while the contracts of the other 12 students and postdocs depend on projects, some of them ending as soon as next year.

The research team in this proposal is formed by 1 assistant professor (Prof. Ayudante Doctor) and 1 permanent associate professors (Prof. Titular). However, we will count with the support and expertise of other members of the ISP and external collaborators, especially for better scientific understanding and analysis of the results. It would be beneficial for the good performance of the project to count with a contracted technician that could dedicated 100% of her/his time for the duration of the project (2 years). The contracted technician (C1) is required because: (1) some tasks are very time consuming, especially in WP1, and should be supported by someone with experience with databases, and with computer science skills (IT); (2) C1 will also help in the scientific development. The tasks to be developed by the contractor will include support in: collection and analysis of data sets (WP1); efficient implementation of methods and algorithms of WP2; support with applications of WP3.

## 2. IMPACTO ESPERADO DE LOS RESULTADOS

The objectives of this project perfectly fit within the R+D+I Spanish National Plan and the EU's old and new framework programme for research and innovation Horizon 2020 and Horizon Europe. Space research was identified in Horizon 2020 as one of Europe's *key industrial technologies*, providing the tools to address key global challenges, and in particular to stimulate the wider use of Copernicus Sentinels data. Benefits of these programmes for the EU enable creating new business opportunities and cut across all sectors of the economy, such as transport, telecommunications, environment and security.

Our research proposal will help achieving such goals by developing and improving efficient machine learning models for remote sensing. A significant impact of the proposed research and the project activities is expected through publications in top ranked international journals and conferences, as well as the expected technological transfer to European and international institutions, such as ESA, EUMETSAT and IOCCG, as we have done in previous projects and contracts. For instance, the AQUACLASS project will contribute towards the pre-operational capacities of the EU Copernicus programme in the ocean colour context by providing new methodologies suitable for reliable classification, analysis, and retrieval of water products.

About the **impact and benefits of the project**, our scientific and technical contribution deals with developing and improving efficient machine learning models that will be applied to the classification of water types, improving the accuracy of the retrieval of biophysical parameters. The expected scientific and technical advances and derived benefits from the project will be

very useful for the image processing and machine learning communities, and the EU space research.

Project results will be publicized following the project's **dissemination plan**. The high publication record of the research team, the organization of workshops, conferences and special sessions, as well as the involvement in Networks of Excellence, demonstrates the quality of the team both at national and international levels. According to our publication records, we estimate publishing about 15 JCR international journal and a high number of conference papers per year (more than 40 in 2017) in top ranked proceedings and conferences. From this total number we estimate that at least 1.5 international papers and 2 to 4 conference proceedings per year will be related to the project. Since the working team is integrated in an international OWT working group, we will be involved in the organization and attendance of an international user workshop in 2019, more action to come in the near future.

As for the **transfer of results**, the project will provide advanced methods and toolboxes that can be used in a variety of remote sensing applications. The provision of these methods could have a direct benefit to the processing chains and ocean colour products derived from the sensors on board of Sentinel 2 and Sentinel 3 satellites. This technology will be easily transferred to the industry by improving and developing new software tools and dedicated modules for the EU Copernicus programme. We have a broad experience in developing modules for the ESA BEAM-Visat and SNAP Toolboxes; for the Instituto Cartográfico Valenciano (ICV); the Instituto Geográfico Nacional; and EUMETSAT. We have also implemented an operational cloud service on a dynamic parallel processing infrastructure, based on Hadoop and OpenNebula, exploiting the capabilities of grid computing (within the Sentinels Synergy Framework (SenSyF) FP7 project and the ERC-SEDAL grant). In addition, the team regularly releases software and databases to the scientific community at <http://isp.uv.es/soft.htm>.

### 3. CAPACIDAD FORMATIVA DEL EQUIPO SOLICITANTE

#### 3.1. Training capabilities and plan

The ISP research team is a group of engineers, from the computer science and electrical engineering fields, (geo)physicists and mathematicians. The group pursues theoretical and applied research focused on the design of statistical learning methods for signal processing and exploitation of Earth observation data. This multidisciplinary background places the group in a privileged position for both developing innovative approaches and for training new PhD/master students in a field so dynamic and diverse as machine learning.

#### Teaching and training experience of students

- *Courses*. The members of the AQUACLASS team are deeply involved in the PhD program of 'Electrical Engineering' and the 'Master in Remote Sensing' at the Universitat de València with several subjects related to image processing, vision, data analysis, and space electronics. These activities have been awarded a 'Top Quality Mention'. ISP members are also involved in other Masters and Doctoral programs in other universities: advanced time series analysis in UC3M (Madrid), kernel methods for computer vision at the CVC in UAB (Barcelona), and advanced learning at Univ. Lausanne (Switzerland). Educational activities have also been developed for ESA and EUMETSAT, with specific advanced ocean colour or land remote sensing courses; as well as being part of the ESA-Academy team.
- *Master Thesis projects*. In addition to the FPI fellowships, shorter duration contracts associated to projects have enabled us to train graduates in various topics related to the different projects. We have directed approximately 20 Master theses in the last 5 years.
- *Fostering students excellence*. The IPL has a strict policy and criteria for admission of students. Several postdoctoral fellowships have been awarded, including the highly competitive 'Ramón y Cajal' and 'Juan de la Cierva', as well as similar grants from the local Government.
- The Statistical Learning for Earth Observation Data Analysis project (SEDAL, <http://isp.uv.es/sedal.html>), an ERC Consolidator grant, has allowed the ISP group to open several PhD and post-doctoral positions. Some of our PhD students are listed in section 3.2.

## Environment

- *Invited talks and reading groups.* The ISP group is integrated in the institute Image Processing Laboratory (IPL), that consists of about 40 researchers in the areas of image processing and remote sensing, which has its own space and resources at the Parc Científic (UVEG). In this environment, we organize regular invited talks where doctoral students and experts interact. We held weekly meetings for discussion of the research topics. See program of talks in <http://isp.uv.es/talks.htm>. During the last 5 years we have received visits of renown researchers to our lab for invited talks and short stays: Dr. Jenssen (Tromsø Univ., Norway), Dr. Rakotomamonjy (INSA-Rouen, France), Prof. Pappas (EIC of the IEEE Trans. Image Proc.), Dr. Simoncelli (NYU), Dr. Wang, Dr. Tuia (Univ. Wageningen) and Dr. Volpi (ESDC, Switzerland), Dr. Ratle (Nuance Inc, Belgium), Dr. Calbet (EUMETSAT, Germany), Dr. Canu (INSA, Rouen, France), Dr. Van de Weijer and Dr. Vanrell (CVC, Barcelona), Dr. Molina (UGR, Granada), Dr. Martínez-Ramón (Univ. New Mexico, USA), and Dr. Santos (Univ. Bristol, UK), Dr. Rojo-Álvarez (URJC, Madrid), Drs. Zhou Wang (Waterloo Univ., CA), Dr. Johannes Balle (NYU, USA), Dino Sejdinovic (Oxford), Jakob Runge (DLR), Prof. Markus Reichstein (MPI Biogeochemistry, Germany), and Prof. Tulay Adali (University of Maryland, USA). This international environment is ideal for learning and scientific motivation of students.
- *Collaborators, stays and visits.* We have a vast network of international contacts and collaborators (see <http://isp.uv.es/collaborators.htm>), which certainly can make a key experience for training students.

### 3.2. On-going PhD thesis

1. Dr. Ana B. Ruescas, Jesús Delegido and José Moreno are advisors of the PhD student Marcela Pereira-Sandoval with the thesis entitled "*Obtención de parámetros biofísicos de la vegetación y de la calidad de los cuerpos de agua continentales a partir de Sentinel-2*". Foreseen defence in 2021.
2. Dr. Jordi Muñoz-Marí is the supervisor of PhD student Anna Mateo García with the thesis entitled "*Machine learning algorithms for remote sensing data processing*". Foreseen defence in 2020.
3. Dr. Gómez-Chova is the advisor of the PhD student Gonzalo Mateo García with the thesis entitled "*Cloud Detection in the Cloud*". The thesis will be developed in the context of the Google Earth Engine Award 2015 founded by Google. Foreseen defence in 2019. PhD student Dan López Puigdollers is also supervised by Dr. Gómez-Chova with the thesis entitled "*Change detection machine learning algorithms for cloud masking of remote sensing image time series*". Foreseen defence in 2021.
4. Dr. Camps-Valls is currently advising (or co-advising) 8 PhD theses in topics marginally related to the project, mainly along the lines of classification and anomaly detection (J. A. Padron), regression (Daniel Svendsen, Anna Mateo, and Sara Bjørk at Tromsø University), dependence estimation and manifold learning (Emmanuel Johnson), dimensionality reduction (Diego Bueso and Guido Kraemers at the Max Planck Institute in Jena, Germany) and causal inference (Emiliano Díaz).

### 3.3. Trained doctors and current positions

The training capacity of the research group is confirmed by the direction of the following PhD theses on topics related to the project between 2010-present, and the scientific/professional development of the graduated doctors. All investigators trained in our team are now working in the scientific-technical public institutes or in R&D companies:

1. Valero Laparra defended his PhD co-directed by Dr. Malo and Dr. Camps-Valls in 2011. His thesis work focused on the design of nonlinear transforms adapted to the statistics of images and its application in neuroscience and image processing. Title: "*Learning efficient image representations: Connections between statistics and neuroscience*." He is currently contracted as researcher in the ERC-SEDAL project.

2. Julia Amorós López, co-directed by Dr. Gómez-Chova and Dr. Calpe-Maravilla, defended her PhD thesis in 2012, focused on image fusion, developing downscaling methods to tackle multitemporal image fusion and change detection problems with improved spatio-spectral-temporal resolutions. Title: “*Multi-resolution spatial unmixing for MERIS and Landsat TM image fusion*”. She is now tenure track assistant professor (contratada doctor) at the Electronics Eng. Dept. at UVEG.
3. Emma Izquierdo Verdiguier, co-directed by Dr. Gómez-Chova and Dr. Camps-Valls, defended her PhD thesis in 2014, focused on nonlinear feature selection/extraction based on graphs and kernels, and the encoding of invariances in image recognition systems. Title: “*Kernel feature extraction methods for remote sensing data analysis*”. She is currently assistant researcher at the Universität für Bodenkultur Wien, Austria.
4. Manuel Campos-Taverner, co-directed by Dr. Camps-Valls defended his PhD thesis in 2017. The thesis entitled “*Advanced methods and processing chain for bio-physical parameter estimation of rice crops*”. The thesis was developed in the context of the FP7 ERMES project, <http://www.ermes-fp7space.eu/>.

#### 4. IMPLICACIONES ÉTICAS Y/O DE BIOSEGURIDAD

This project does not imply any sensitive ethical or biosecurity issues.

### Bibliografía

- [1] Kelly R. Stewart, Rebecca L. Lewison, Daniel C. Dunn, Rhema H. Bjorkland, Shaleyla Kelez, Patrick N. Halpin, and Larry B. Crowder. Characterizing fishing effort and spatial extent of coastal fisheries. *PLOS ONE*, 5(12):1–8, 12 2011.
- [2] V. Vantrepotte, H. Loisel, D. Dessailly, and X. Mériaux. Optical classification of contrasted coastal waters. *Remote Sensing of Environment*, 123:306 – 323, 2012.
- [3] Timothy S. Moore, Mark D. Dowell, Shane Bradt, and Antonio Ruiz Verdu. An optical water type framework for selecting and blending retrievals from bio-optical algorithms in lakes and coastal waters. *Remote Sensing of Environment*, 143:97 – 111, 2014.
- [4] Annelies Hommersom, Marcel R. Wernand, Steef Peters, Marieke A. Eleveld, Hendrik Jan van der Woerd, and Jacob de Boer. Spectra of a shallow sea-unmixing for class identification and monitoring of coastal waters. *OCEAN DYNAMICS*, 61(4):463–480, APR 2011.
- [5] M. R. Wernand, A. Hommersom, and H. J. van der Woerd. MERIS-based ocean colour classification with the discrete Forel-Ule scale. *OCEAN SCIENCE*, 9(3):477–487, 2013.
- [6] Thomas Jackson, Shubha Sathyendranath, and Frédéric Mélin. An improved optical classification scheme for the ocean colour essential climate variable and its applications. *Remote Sensing of Environment*, 203:152 – 161, 2017. Earth Observation of Essential Climate Variables.
- [7] N.G. Jerlov. *Marine Optics*. Elsevier, 1976.
- [8] N.G. Jerlov. Optical studies of ocean waters. *Rep. Swed. Deep-Sea Exp.*, 3, 1951.
- [9] E. Aas, N.K. Hojerslev, J. Hokedal, and K. Sorensen. Optical water types of the nordic seas and adjacent areas. *Oceanologia*, 55(2):471–482, 2013.
- [10] M.G. Solonenko and C.D. Mobley. Inherent optical properties of jerlov water types. *Applied Optics*, 54(17):5392–5401, 2015.
- [11] T. S. Moore, J. W. Campbell, and Hui Feng. A fuzzy logic classification scheme for selecting and blending satellite ocean color algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 39(8):1764–1776, Aug 2001.
- [12] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer, New York: Plenum, 1981.
- [13] Chengfeng Le, Yunmei Li, Yong Zha, Deyong Sun, Changchun Huang, and Hong Zhang. Remote estimation of chlorophyll a in optically complex waters based on optical classification. *Remote Sensing of Environment*, 115(2):725 – 737, 2011.
- [14] V. Vantrepotte, H. Loisel, F. Mélin, D. Desailly, and L. Duforêt-Gaurier. Global particulate matter pool temporal variability over the seawifs period (1997–2007). *Geophysical Research Letters*, 38(2):n/a–n/a, 2011. L02605.



- [15] F. Melin and V. Vantrepotte. How optically diverse is the coastal ocean? *Remote Sensing of Environment*, 160:235 – 251, 2015.
- [16] Karen S. Baker and Raymond C. Smith. Bio-optical classification and model of natural waters. 21. *Limnology and Oceanography*, 27(3):500–509.
- [17] Louis Prieur and Shubha Sathyendranath. An optical classification of coastal and oceanic waters based on the specific spectral absorption curves of phytoplankton pigments, dissolved organic matter, and other particulate materials1. *Limnology and Oceanography*, 26(4):671–689.
- [18] Anu Reinart, Antti Herlevi, Helgi Arst, and Liis Sipelgas. Preliminary optical classification of lakes and coastal waters in estonia and south finland. *Journal of Sea Research*, 49(4):357–366, 2003.
- [19] Timothy S. Moore, Janet W. Campbell, and Mark D. Dowell. A class-based approach to characterizing and mapping the uncertainty of the modis ocean chlorophyll product. *Remote Sensing of Environment*, 113(11):2424 – 2430, 2009.
- [20] Robert J. W. Brewin, Stefano Ciavatta, Shubha Sathyendranath, Thomas Jackson, Gavin Tilstone, Kieran Curran, Ruth L. Airs, Denise Cummings, Vanda Brotas, Emanuele Organelli, Giorgio Dall’Olmo, and Dionysios E. Raitsos. Uncertainty in ocean-color estimates of chlorophyll for phytoplankton groups. *Frontiers in Marine Science*, 4:104, 2017.
- [21] John T. Trochta, Colleen B. Mouw, and Timothy S. Moore. Remote sensing of physical cycles in lake superior using a spatio-temporal analysis of optical water typologies. *Remote Sensing of Environment*, 171:149 – 161, 2015.
- [22] T.S. Moore, M.D. Dowell, and B.A. Franz. Detection of coccolithophore blooms in ocean color satellite imagery: a generalized approach for use with multiple sensors. *Remote Sensing of Environment*, 117:249–263, 2012.
- [23] M. Hieronymi, D. Mueller, and R. Doerffer. The OLCI Neural Network Swarm (ONNS): A bio-geo-optical algorithm for open ocean and coastal waters. *Frontiers in Marine Science*, 4:140, 2017.
- [24] Hongyan Xi, Martin Hieronymi, Hajo Krasemann, and Rüdiger Röttgers. Phytoplankton group identification using simulated and in situ hyperspectral remote sensing reflectance. *Frontiers in Marine Science*, 4:272, 2017.
- [25] Mark William Matthews and Daniel Odermatt. Improved algorithm for routine monitoring of cyanobacteria and eutrophication in inland and near-coastal waters. *Remote Sensing of Environment*, 156:374 – 382, 2015.
- [26] Marieke A. Eleveld, Ana B. Ruescas, Annelies Hommersom, Timothy S. Moore, Steef W. M. Peters, and Carsten Brockmann. An optical classification tool for global lake waters. *Remote Sensing*, 9(5), 2017.
- [27] S. R. Bradt. *Development of bio-optical algorithms to estimate chlorophyll in the Great Salt Lake and New England lakes using in situ hyperspectral measurements*. PhD thesis, The University of New Hampshire., 2012.
- [28] Antonio Ruiz-Verdú, Stefan G.H. Simis, Caridad de Hoyos, Herman J. Gons, and Ramon Peña-Martinez. An evaluation of algorithms for the remote sensing of cyanobacterial biomass. *Remote Sensing of Environment*, 112(11):3996 – 4008, 2008. Applications of Remote Sensing to Monitoring Freshwater and Estuarine Systems.
- [29] C. Brockmann, R. Doerffer, M. Peters, K. Stelzer, S. Embacher, and A. Ruescas. Evolution of the c2rcc neural network for sentinel 2 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters. Living Planet Symposium, 2016.
- [30] C. Brockmann, M. Paperin, O. Danne, and A. Ruescas. Multisensor cloud screening and validation: idepix an pixbox. Living Planet Symposium, 2013.
- [31] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.
- [32] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*, volume 3940 of *LNCS*, pages 34–51. Springer, 2006.
- [33] H. Wold. Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, pages 391–420, 1966.
- [34] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–337,

1936.

- [35] E. Izquierdo-Verdiguier, L. Gómez-Chova, and G. Camps-Valls. Kernels for remote sensing image classification. *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–23, June 2015.
- [36] Bernhard Schoelkopf and Alexander Smola. *Learning with kernels*. MIT Press, 2002.
- [37] Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *Proceedings of the 11th International Conference on Neural Information Processing Systems*, NIPS'98, pages 536–542, Cambridge, MA, USA, 1998. MIT Press.
- [38] Emma Izquierdo-Verdiguier, V Laparra, J Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Advanced feature extraction for earth observation data processing. In *Reference Module in Earth Systems and Environmental Sciences*. 12 2017.
- [39] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521, 2015.
- [40] A. Baraldi and F. Parmiggiani. A neural network for unsupervised categorization of multi-valued input patterns: an application to satellite image clustering. *IEEE Transactions on Geoscience and Remote Sensing*, 33(2):305–316, Mar 1995.
- [41] S. Ghosh, L. Bruzzone, S. Patra, F. Bovolo, and A. Ghosh. A context-sensitive technique for unsupervised change detection based on hopfield-type neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):778–789, March 2007.
- [42] F. Del Frate, G. Licciardi, and R. Duca. Autoassociative neural networks for features reduction of hyperspectral data. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS '09. First Workshop on*, pages 1–4, Aug 2009.
- [43] G. Licciardi, F. Del Frate, and R. Duca. Feature reduction of hyperspectral data using autoassociative neural networks algorithms. In *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*, volume 1, pages I–176–I–179, July 2009.
- [44] C. Vaduva, I. Gavat, and M. Datcu. Deep learning in very high resolution remote sensing image information mining communication concept. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2506–2510, Aug 2012.
- [45] Volodymyr Mnih and Geoffrey Hinton. Learning to label aerial images from noisy data, 2012.
- [46] X. Chen, S: Xiang, C.-L. Liu, and C.-H. Pan. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.*, 11(10):1797–1801, 2014.
- [47] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu. Deep learning-based classification of hypersepctral data. *IEEE J. Sel. Topics Appl. Earth Observ.*, 7(6):2094–2107, 2014.
- [48] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006.
- [49] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, pages 153–160, 2006.
- [50] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 609–616, New York, NY, USA, 2009. ACM.
- [51] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I*, ICANN'11, pages 52–59, Berlin, Heidelberg, 2011. Springer-Verlag.
- [52] A. Romero, P. Radeva, and C. Gatta. Meta-parameter free unsupervised sparse feature learning. *Accepted to IEEE Transaction on Pattern Analysis and Machine Intelligence*, :-, 2014.
- [53] Corinna Cortes and Vladimir Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.
- [54] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, December 2006.

- [55] G. Camps-Valls, T.V. Bandos Marsheva, and D. Zhou. Semi-supervised graph-based hyperspectral image classification. *IEEE Trans. Geosci. Rem. Sens.*, 45(10):3044–3054, 2007.
- [56] L. Gómez-Chova, G. Camps-Valls, J. Muñoz-Marí, and J. Calpe. Semi-supervised image classification with laplacian support vector machines. *IEEE Geosci. Remote Sens. Lett.*, 5(4):336–340, 2008.
- [57] G. Camps-Valls, N. Shervashidze, and K. M. Borgwardt. Spatio-spectral remote sensing image classification with graph kernels. *IEEE Geoscience and Remote Sensing Letters*, 7(4):741–745, Oct 2010.
- [58] D. Tuia and G. Camps-Valls. Semi-supervised remote sensing image classification with cluster kernels. *IEEE Geosc. Rem. Sens. Lett.*, 6(1):224–228, 2009.
- [59] 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.
- [60] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 582–588, Cambridge, MA, USA, 1999. MIT Press.
- [61] David Tax and Robert Duin. Support vector domain description. 20:1191–1199, 11 1999.
- [62] J. Muñoz-Marí, L. Bruzzone, and G. Camps-Valls. A support vector domain description approach to supervised classification of remote sensing images. *IEEE Trans. Geosc. Rem. Sens.*, 45(8):2683–2692, Aug 2007.
- [63] J. Muñoz-Marí, F. Bovolo, L. Gómez-Chova, L. Bruzzone, and G. Camps-Valls. Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE Trans. Geosc. Rem. Sens.*, 48(8):3188–3197, 2010.
- [64] E. Izquierdo-Verdiguier, V. Laparra, L. Gómez-Chova, and G. Camps-Valls. Encoding invariances in remote sensing image classification with svm. *IEEE Geoscience and Remote Sensing Letters*, 10(5):981–985, Sept 2013.
- [65] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA, 1995.
- [66] S. Haykin and S.S. Haykin. *Neural Networks and Learning Machines*. Number v. 10 in Neural networks and learning machines. Prentice Hall, 2009.
- [67] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.
- [68] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke and Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [69] Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [70] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson. Random forest classification of multisource remote sensing and geographic data. In *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 2, pages 1049–1052 vol.2, Sept 2004.
- [71] M. Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005.
- [72] Mariana Belgiu and Lucian Drăguț. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24 – 31, 2016.
- [73] A. R. (Alan R.) Longhurst, EBSCOhost, and Elsevier. *Ecological geography of the sea*. Amsterdam ; Boston, MA : Academic Press, 2nd ed edition, 2007.