Land Cover Classification with Gaussian Processes using spatio-spectro-temporal features

Valentine Bellet, Mathieu Fauvel, Jordi Inglada

To cite this version:

Valentine Bellet, Mathieu Fauvel, Jordi Inglada. Land Cover Classification with Gaussian Processes using spatio-spectro-temporal features. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61, pp.4400221. 10.1109/TGRS.2023.3234527 . hal-03781332v2

HAL Id: hal-03781332
https://hal.science/hal-03781332v2

Submitted on 5 Jan 2023

Land Cover Classification with Gaussian Processes using spatio-spectro-temporal features

Valentine Bellet, *Student Member, IEEE,* Mathieu Fauvel, *Senior Member, IEEE,* and Jordi Inglada

***Abstract*—In this article, we propose an approach based on Gaussian Processes (GP) for large scale land cover pixel-based classification with Sentinel-2 satellite image time-series (SITS). We used a sparse approximation of the posterior combined with variational inference to learn the GP's parameters. We applied stochastic gradient descent and GPU computing to optimize our GP models on massive data sets. The proposed GP model can be trained with hundreds of thousands of samples, compared to few thousands for traditional GP methods. Moreover, we included the spatial information by adding the geographic coordinates into the GP's covariance function to efficiently exploit the spatio-spectro-temporal structure of the SITS. We ran experiments with Sentinel-2 SITS of the full year 2018 over an area of 200 000 km^2 (about 2 billion pixels) in the south of France, which is representative of an operational setting. Adding the spatial information significantly improved the results in terms of classification accuracy. With spatial information, GP models have an overall accuracy of 79.8. They are more than three points above Random Forest (the method used for current operational systems) and more than one point above a multi-layer perceptron. Compared to a Transformer-based model (which provides state of the art results in the literature, but are not applied in operational systems), GP models are only one point below.**

***Index Terms*—Satellite Image Time-Series (SITS), Sentinel-2, Land Cover Map, Pixel-Based, Classification, Large Scale, Sparse Variational Gaussian Processes, Earth Observation (EO), Remote Sensing.**

I. Introduction

THE increasing number of Earth observation satellites generates a huge amount of data with heterogeneous modalities (e.g. optical, radar, etc.) at various resolutions (e.g. sub-metric, decametric, etc). Among them, the Sentinel-2 constellation provides free and open data with a 5-day revisit time at high spectral and spatial resolutions (four spectral bands at 10m, six at 20m and three at 60m per pixel) [1]. This mission was designed to monitor Earth's surface changes. Models based on these optical satellite image time-series (SITS) can explain and predict the states and trends of our environment. They are essential to understand the challenges related to climate change [2]. Since their launch in 2015 and 2017, SITS from the twin Sentinel-2 satellites have already shown a clear benefit in biodiversity monitoring [3], [4], forest mapping [5], [6], water quality [7], [8], agricultural monitoring [9], [10] or disaster management [11], [12].

Every year, around one petabyte of Sentinel-2 SITS is generated [13]. These data, covering all continental surfaces with a short revisit cycle, bring the opportunity of large scale mapping (at national or even continental scales). To fully benefit from the information gathered by these massive geo-spatial data, automatic methods are needed for their analysis. Over the past 20 years, statistical methods [14] and then machine learning (ML) based methods [15] have shown great potential for various thematic applications. From those, the best known is certainly automatic land use/cover classification (LUCC), which consists in assigning a class among a set of predefined ones to each pixel from the area of interest. Three main challenges are associated to large scale LUCC:

1) The spatio-spectro-temporal structure of the SITS: each pixel has a local spatial correlation, as well as a class-dependent spectral and temporal correlation structure that needs to be taken into account for an accurate classification [16].
2) The non stationarity of the class-conditional probability distribution that implies a varying spectro-temporal signature over the spatial domain. This phenomenon is particularly critical for the large scale area classification problems where phenology and topographic conditions exacerbate this issue. Therefore, the learning algorithm has to be able to model spatially varying class-conditional probability distributions [17], [18].
3) The volume (defined as number of pixels times number of dates times number of spectral features) of the data to be processed, both at the learning and inference stages, which require fast and effective algorithms that scale well.

Kernel-based algorithms have shown to perform well for LUCC [19]. Support Vector Machines (SVM) were widely applied in land cover classification with multi and hyper-spectral images [15], [20]–[22]. However, the computational complexity of kernel-methods is cubic w.r.t. the number of samples used to train the model, and become quickly intractable as the number of samples increases. Therefore, kernel methods have rarely been used for large-scale mapping despite their learning capacity.

Another ML algorithm widely investigated is Random Forest (RF) [23]. It has shown to perform well for different case studies [24] and has been favorably applied to large scale classification problems, such as [25], [26]. Yet, RF do not allow to incorporate information about the structure of the SITS beyond the use of specific temporal features. Indeed, the temporal structure is not taken into account: switching the order of the data in time series leads to the same results.

This work is supported by the Natural Intelligence Toulouse Institute (ANITI) from Universite Federale Toulouse Midi-Pyrenees under grant agreement ANITI ANR-19-PI3A-0004 (this PhD is co-founded by CS-Group and by the Centre National d'Etudes Spatiales (CNES)).

V. Bellet, M. Fauvel and J. Inglada are with CESBIO, Universite de Toulouse, CNES/CNRS/INRAe/IRD/UPS, 31000 Toulouse, France (e-mail: valentine.bellet@univ-toulouse.fr, mathieu.fauvel@inrae.fr, jordi.inglada@cesbio.eu)

Furthermore, the learning step requires full simultaneous view on the training samples, which hinders parallel training implementations. In [25], the non-stationarity as well as the massive training data set were handled by doing a spatial stratification of the problem. The training data set was divided into strata: each stratum corresponds to an eco-climatic region as defined in [27]. For each stratum, an independent RF model was trained. However, no spatial constraints were imposed during the learning or the prediction steps and the models could behave differently at the boundaries between strata. Thus, the transition between two spatial strata can show artifacts due to the discontinuity in the predictions by models of adjacent strata.

In recent years, deep learning (DL) methods have quickly emerged in the remote sensing community due to the increased free distribution of Big Earth Observation Data, the development of computing resources (e.g. GPU, HPC, etc.) and the availability of open source deep learning frameworks (e.g. *Pytorch* [28] or *Tensorflow* [29]). A large variety of DL methods have been developed such as: multi-layer perceptrons (MLP) [30], convolutional neural networks (CNN) [31], recurrent neural networks (RNN) [32], auto-encoders (AE) [33] and generative adversarial networks (GAN) [34]. The main advantage of these methods is their ability to extract features (i.e. spatial, spectral and temporal patterns) instead of handcrafting them. By learning temporal patterns, long short-term memory (LSTM) (i.e. neural layers developed to solve the problem of vanishing gradients in RNN) have been a promising tool in SITS classification [35], [36]. To include spatial information, spatial-sequential RNN [37] and spectral-spatial RNN [38] have been developed. Besides, CNN models which are commonly used with images have also shown very good results [39], [40]. By adding coordinate information into feature maps, performance results with CNN have been improved in SITS classification [41], [42]. Therefore, methods combining both RNN and CNN networks have been developed [43], [44]. Temporal CNN, which combine features across time with convolutions, have also proved to be effective [45]. Recently, methods based on the attention principle have shown very interesting results [46], [47]. One major problem with all these DL methods is their "black-box" nature: their parameters are hardly interpretable.

Gaussian Processes (GP) are stochastic non-parametric approaches combining Bayesian and kernel methods for regression and classification problems [48]. They have been successfully applied in remote sensing for parameter estimation [49]–[52] or for classification [53]–[55]. Unlike DL models, GP can be interpretable through their parameters (e.g. temporal correlation for the length-scale parameter in a Radial Basis Function (RBF) covariance function [48], [55]). Furthermore, their Bayesian nature enables the estimation of posterior distributions rather than point-wise values which is useful to assess prediction uncertainties. Another interesting property of GP is the possibility to define suitable kernel functions, as with SVM, and to learn their parameters through gradient descent, unlike SVM [48, Chapter 5].

However, conventional GP are limited to few thousands of training inputs since their complexity scales cubically w.r.t. the number of training samples. In recent years, several solutions have been proposed to deal with large amounts of data [56]. For example, [57] proposed an approach based on the approximation of the posterior distribution that uses variational inference. Stochastic gradient descent and GPU computing can therefore be exploited to optimize GP models. Such recent methods drastically reduce the computing complexity and have been applied to large scale data in computer vision.

The contributions of this work are three-fold and follow the survey of GP for Earth Observation (EO) data analysis of G. Camps-Valls *et al.* [52]. First, we formally introduce a large scale GP model recently proposed in the computer vision community, and make connection with existing literature in EO data analysis with GP. Second, we propose two kernel functions allowing the structure of the SITS to be taken into account in the processing. Fundamentally, the spatial coordinates of a given pixel are included in the covariance function in addition to the spectro-temporal features to model the spatial dependency between pixels. The parameters of the covariance functions are optimized with a lower bound of the marginal likelihood. This point has not been investigated so far for SITS classification at large scale. Some works have been proposed to use contextual information with GP [58], [59] but spatial dependency was limited to a close neighborhood (e.g., 5×5 pixels) with a small training set size. Third, we report an intensive large scale classification benchmark with conventional methods and recent deep models. A novelty w.r.t. existing works concerns the analysis of spatial stratification and its effects on the continuity of the prediction across the different strata. A comparison with the non-stratified counterpart is also performed. As a by-product of the paper, we also release the data set and the source code[1].

The remainder of this paper is organized as follows. A formal review of conventional GP is presented in Section II. The proposed GP model used for the pixel-based large scale land cover classification is explained and described in Section III. The model parametrization choices are discussed in Section IV. The experimental setup is detailed in Section VI and the comparison with the state-of-the-art methods is provided in Section VII. Section VIII proposes an analysis of the specific characteristics of GP. Finally, Section IX concludes this paper and opens discussions on future works.

II. Gaussian Processes

The training data, defined as a pair of input and output data, is used to train ML algorithms. In remote sensing, input data are usually represented as a time-series (e.g. sequence of pixels organized in time order). Each pixel is itself a vector defined with the same number of features (e.g. usually spectral reflectances or indices). The output data, also called target, is the measure that wants to be predicted, in classification problem it is known as the class or the label.

In the following, the training set is denoted $\mathcal{S} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, where N is the number of pixels, $\mathbf{x}_i \in \mathbb{R}^{d+d'}$ is a pixel i represented by its corresponding d spectro-temporal measurements and its d' spatial measurements (e.g.

[1] https://gitlab.cesbio.omp.eu/belletv/land_cover_southfrance_gp

coordinates), $\mathbf{y}_i \in \mathbb{R}^P$ with $P \geq 1$ is the value to be predicted, or target, associated to pixel i. For instance, in a classification problem $\mathbf{y}_i$ corresponds to the membership degree of the pixel to each class.

A. Univariate Gaussian Processes

An univariate GP f is completely specified by its real-valued mean function m and its covariance function k: $f \sim \mathcal{GP}(m, k)$ [48]. In this paper, m and k are assumed to be modeled by parametric functions with hyper-parameters $\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}_k$, respectively, and k is constrained to be a positive semi-definite function [48, Chapter 4]. Noting $f(\mathbf{X})$ the random vector defined as $f(\mathbf{X}) = \big[f(\mathbf{x}_1), \dots, f(\mathbf{x}_N)\big]^\top$, $f(\mathbf{X})$ follows a multivariate Gaussian distribution: $f(\mathbf{X}) \sim \mathcal{N}_N(\boldsymbol{\mu}, \mathbf{K})$ with $\boldsymbol{\mu} = \big[m(\mathbf{x}_1), \dots, m(\mathbf{x}_N)\big]^\top$ and $\mathbf{K}$ such as $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\forall i, j \in \{1, \dots, N\}^2$.

Univariate GP are commonly used to regress a scalar target value ($y_i \in \mathbb{R}$) through a link function ψ that relates the univariate latent variable $f(\mathbf{x}_i)$ to the observed y_i. In the regression case, we denote $\mathbf{X} = \big[\mathbf{x}_1, \dots, \mathbf{x}_N\big]^\top$ and $\mathbf{y} = \big[y_1, \dots, y_N\big]^\top$. To model realistic situations, an usual approach is to consider a noisy version of the function value such as

$$y_i = \psi\big(f(\mathbf{x}_i)\big) = f(\mathbf{x}_i) + \epsilon_i \tag{1}$$

with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and σ the noise level. The likelihood is simply

$$p\big(y_i | f(\mathbf{x}_i)\big) = \mathcal{N}_1\big(y_i | f(\mathbf{x}_i), \sigma^2\big). \tag{2}$$

Assuming i.i.d. samples, the full likelihood is given by

$$p\big(\mathbf{y} | f(\mathbf{X})\big) = \prod_{i=1}^{N} p\big(y_i | f(\mathbf{x}_i)\big) = \mathcal{N}_N(\mathbf{y} | f(\mathbf{X}), \sigma^2 \mathbf{I}_N). \tag{3}$$

Given a new input $\mathbf{x}_*$ the prediction is done by taking the *maximum a posteriori* (MAP) of the predictive distribution obtained by marginalizing over the latent variables $f(\mathbf{x}_*)$. Every term follows a Gaussian distribution and therefore the posterior distribution is also Gaussian. Using standard Gaussian equalities, it can be written analytically [60, Chapter 2.3.2 and 2.3.3]:

$$p(y_* | \mathbf{y}, \mathbf{X}, \mathbf{x}_*) = \mathcal{N}_N(y_* | \mu_*, \sigma_*^2), \tag{4}$$

with

$$\mu_* = m(\mathbf{x}_*) + \mathbf{k}_*^\top (\mathbf{K} + \sigma^2 \mathbf{I}_N)^{-1} (\mathbf{y} - \boldsymbol{\mu}), \tag{5}$$

$$\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (\mathbf{K} + \sigma^2 \mathbf{I}_N)^{-1} \mathbf{k}_* + \sigma^2, \tag{6}$$

and with $\mathbf{k}_* = [k(\mathbf{x}_1, \mathbf{x}_*), \dots, k(\mathbf{x}_N, \mathbf{x}_*)]^\top$. For a Gaussian distribution, the MAP is given by the mean of the distribution, i.e., $\hat{y}_* = \mu_*$. Furthermore, the GP framework allows to estimate the uncertainty of the prediction through the variance of the posterior distribution σ_*^2.

The hyper-parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_m, \boldsymbol{\theta}_k, \sigma^2\}$ strongly influence the prediction since they appear in (5) and (6) ($\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}_k$ are respectively the parameters of the functions m and k). They are usually optimized by maximizing the log-marginal likelihood of the model on the training set $\mathcal{S}$ [48, Chapter 2]:

$$\begin{aligned}\log p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}) = &-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^{\mathrm{T}} \big(\mathbf{K} + \sigma^2 \mathbf{I}_N\big)^{-1} (\mathbf{y} - \boldsymbol{\mu}) \\ &- \frac{1}{2} \log\big(|\mathbf{K} + \sigma^2 \mathbf{I}_N|\big) - \frac{N}{2} \log(2\pi).\end{aligned} \tag{7}$$

The derivatives of Equation (7) are analytically tractable and the optimization of $\boldsymbol{\theta}$ can be done using constrained gradient descent [48, Chapter 5 and Appendix A.3].

In comparison to other non-linear prediction algorithms, such as SVM or kernel ridge regression, GP offer the possibility to automatically tune their hyper-parameters $\boldsymbol{\theta}$. GP also provide the variance of the point-wise estimation. These properties made GP for regression widely used by the remote sensing community in the last decade [61]–[64].

However, conventional GP scale poorly w.r.t. the number of training samples. The main bottleneck comes from the computational cost of the matrix inversion and the computation of the determinant in (7). These operations scale cubically with the number of training pixels and, moreover, the storage complexity is $\mathcal{O}(N^2)$. This is the main reason explaining why GP have only been used on data sets limited to a few thousand pixels [52].

We discuss in Section III some solutions that have been explored in the last decade to apply large data sets to GP. However, the extension of univariate GP to multivariate GP for the purpose of classification is presented first in the following section.

B. Multivariate Gaussian Processes

Likewise univariate GP, a P-multivariate GP $\mathbf{f}$ is specified by its vector-valued mean function $\mathfrak{m} \in \mathbb{R}^P$ and its positive matrix-valued covariance function $\mathcal{K} \in \mathbb{R}^{P \times P}$. We have $\mathbf{f} \sim \mathcal{GP}(\mathfrak{m}, \mathcal{K})$ with:

$$\begin{aligned}\mathfrak{m}(\mathbf{x}) &= [m_1(\mathbf{x}) \quad ... \quad m_P(\mathbf{x})]^\top, \\ \mathcal{K}(\mathbf{x}, \mathbf{x}) &= \begin{bmatrix} k_{11}(\mathbf{x}, \mathbf{x}) & ... & k_{1P}(\mathbf{x}, \mathbf{x}) \\ ... & k_{pp'}(\mathbf{x}, \mathbf{x}) & ... \\ k_{P1}(\mathbf{x}, \mathbf{x}) & ... & k_{PP}(\mathbf{x}, \mathbf{x}) \end{bmatrix},\end{aligned}$$

where $k_{pp'}(\mathbf{x}, \mathbf{x})$ is the covariance between two univariate GP: $f_p(\mathbf{x})$ and $f_{p'}(\mathbf{x})$ with $p, p' \in \{1, \dots, P\}$ and $\mathcal{K}$ of size $P \times P$. Similarly to univariate GP, all marginals follow a Gaussian distribution, noting

$$\mathbf{f}(\mathbf{X}) = [f_1(\mathbf{x}_1), \dots, f_P(\mathbf{x}_1), \dots, f_1(\mathbf{x}_N), \dots, f_P(\mathbf{x}_N)]^\top$$

the random vector of size NP, then $\mathbf{f}(\mathbf{X}) \sim \mathcal{N}_{NP}(\boldsymbol{\mu}_o, \mathbf{K}_o)$ with $\boldsymbol{\mu}_o = [\mathfrak{m}(\mathbf{x}_1), \dots, \mathfrak{m}(\mathbf{x}_N)]^\top$ and

$$\mathbf{K}_o = \begin{bmatrix} \mathcal{K}(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \mathcal{K}(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \mathcal{K}(\mathbf{x}_N, \mathbf{x}_1) & \cdots & \mathcal{K}(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}.$$

Multivariate GP are also known as multi-output or multi-task GP [65]. For instance, they are used when the learning task has several correlated outputs (e.g. several variables to regress given a single input, multi-class classification, etc.) [66].

The main challenge in multivariate GP is to define and optimize the cross-covariance function $k_{pp'}$ that:

1) lead to a valid covariance function $\mathcal{K}$,
2) exploit the multivariate structure of the problem to be inferred,
3) can be efficiently computed ($\mathbf{K}_o$ of size $NP \times NP$).
The most common approach is to consider separable kernels where one kernel acts on the input sample and another kernel models the interaction between the outputs [67]. The linear model of co-regionalization (LMC) exploits this formulation [68], [69]. It defines each marginal f_p as a linear combination of L independent univariate GP g_l:

$$\mathbf{f} = \mathbf{A}\mathbf{g} \tag{8}$$

with $\mathbf{A} \in \mathbb{R}^{P\times L}$ and $g_l \sim \mathcal{GP}(m_l, k_l)$. The processes $\{g_l\}_{l=1}^{L}$ are independent for $l \neq l'$. Many multivariate GP models from the literature are particular cases of the LMC, see for instance [67], [70]. In remote sensing, LMC was used to regress biophysical variables in [71] using MODIS time-series. Moreover, it was also used for land cover classification from Sentinel-2 time-series in [55].

Another common approach to remove the separable assumption is using convolution processes [72]–[74]. Convolution processes can capture more dependence between outputs than LMC (e.g. translation between outputs), but they lack a formulation that scales well with the number of training samples [75]. Therefore, the LMC is used in the following because efficient optimization procedures exist [75]–[77], as discussed in Section III-B.

C. LMC for Gaussian Process classification

In the case of classification with C classes, the target is such as $\mathbf{y}_i \in \{0,1\}^C$ with all its values set to zero except for the element $y_{ic} = 1$ for $\mathbf{x}_i$ of class c. In the classification case, we denote $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_N\right]^\top$ and $\mathbf{Y} = \left[\mathbf{y}_1, \ldots, \mathbf{y}_N\right]^\top$. A *softmax* function $\boldsymbol{\sigma}$ is used as link function to relate the multivariate latent variable $\mathbf{f}(\mathbf{x}_i) = [f_1(\mathbf{x}_i), \ldots, f_C(\mathbf{x}_i)]^\top$ and the observation $\mathbf{y}_i$:

$$\begin{aligned}\mathbf{y}_i =& \boldsymbol{\sigma}(\mathbf{f}(\mathbf{x}_i)) \\ =& \frac{1}{\sum_{c'=1}^{C} \exp(f_{c'}(\mathbf{x}_i))} \times \begin{bmatrix}\exp(f_1(\mathbf{x}_i)) \\ \vdots \\ \exp(f_C(\mathbf{x}_i))\end{bmatrix}.\end{aligned} \tag{9}$$

The associated likelihood for the sample i is written:

$$\begin{aligned}p(\mathbf{y}_i|\mathbf{f}(\mathbf{x}_i)) =& \prod_{c=1}^{C}\left[\frac{\exp(f_c(\mathbf{x}_i))}{\sum_{c'=1}^{C}\exp(f_{c'}(\mathbf{x}_i))}\right]^{y_{ic}} \\ =& \frac{\exp(\mathbf{y}_i^\top \mathbf{f}(\mathbf{x}_i))}{\sum_{c'=1}^{C}\exp(f_{c'}(\mathbf{x}_i))},\end{aligned} \tag{10}$$

or using the LMC

$$p(\mathbf{y}_i|\mathbf{g}(\mathbf{x}_i), \mathbf{A}) = \frac{\exp(\mathbf{y}_i^\top \mathbf{A}\mathbf{g}(\mathbf{x}_i))}{\sum_{c'=1}^{C}\exp(\mathbf{e}_{c'}^\top \mathbf{A}\mathbf{g}(\mathbf{x}_i))}, \tag{11}$$

with $\mathbf{e}_{c'}$ a C-dimensional vector made of zeros except at position c' for which the value is one and with $\mathbf{A} \in \mathbb{R}^{C\times L}$. Conventional GP for classification use a trivial LMC [48, Chapter 3]: the number of latent processes is equal to the number of classes ($L = C$), and all latent univariate GP share the same covariance operator[2].

Contrary to the univariate regression case, the likelihood (10) is not conjugate to the Gaussian distribution and thus analytic expressions of the marginal and predictive distributions are not available.

Sampling methods, such as Markov chain Monte Carlo (MCMC) [78], provide exact computation but at prohibitive computational costs that discard such approaches for large scale scenarios. Alternatively, two popular approximation methods overcoming the non-Gaussian likelihood were discussed for the pixel-wise land-cover classification of several satellite images [79]. These methods, namely the *Laplace approximation* [80] and the *Expectation Propagation* [81] were positively compared to SVM in terms of classification accuracy. However, as the authors of [79] concluded, the computational complexity is $\mathcal{O}(N^3C)$ and thus not applicable to large data sets.

To summarize this brief overview, GP for classification have interesting properties for remote sensing: they allow model selection with the optimization of the marginal likelihood and provide a full posterior predictive distribution rather than point estimates. Fig. 1 outlines the model's architecture used for multi-class classification. However, conventional GP for multi-class classification exhibit two bottlenecks: first, the prior leads to high computational load that scales cubically w.r.t. to the number of training samples; second, the likelihood does not lead to an analytical solution and other approximations are needed. In the following, advances that alleviate the computational cost of GP are presented.

[2] $\mathbf{A} = \mathbf{I}_C$ and $k_l = k,\ \forall l \in \{1, \ldots, C\}$.

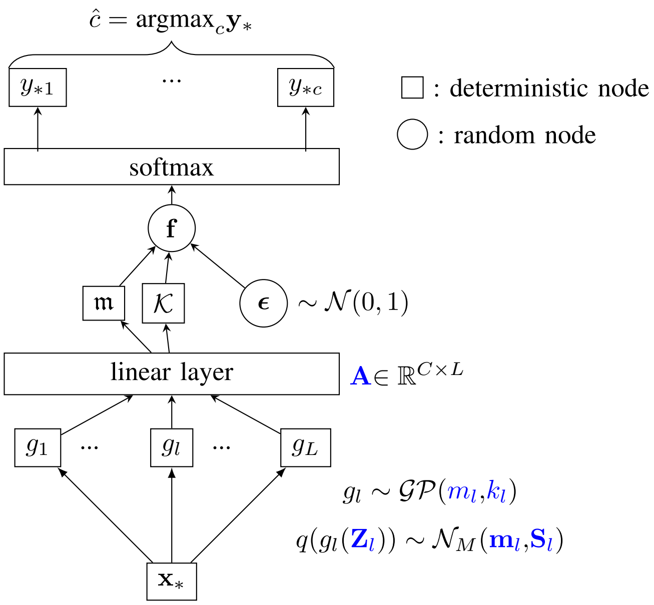


Fig. 1: Model's architecture proposed in this work. Using LMC, univariate GP g_l are combined to obtain multivariate GP $\mathbf{f}$. As conventional GP do not correctly perform for multi-class classification in large scale, some approximations are made. $\mathbf{f}_i$ correspond to the realizations obtained with MC sampling technique: details are provided in Section III. The trainable parameters are written in blue, a full description is available in Section IV.

III. Large Scale Multivariate Gaussian Process Classification

Approximations for large scale univariate GP can be mainly categorized into two approaches: model approximation and posterior approximation [75]. The former, which includes the sparse GP, was successfully used in remote sensing, as discussed in Section III-A, while the latter has barely been investigated in this context. Yet, it has shown superior results to model approximation in large scale classification in computer vision [57], [77]. We propose to use *Variational Inference*, one effective technique in posterior approximation, as described in Section III-B. We apply its extension to multi-class to SITS classification as detailed in Section III-C.

A. Model approximation

Approximation of a Gaussian process model consists in reducing the computational complexity when computing the prior $p\big(f(\mathbf{X})\big)$ or the joint prior $p\big(f(\mathbf{x}_*)|f(\mathbf{X})\big)$ [82]. Data structure can be taken into account to speed-up the inversion of $\mathbf{K}$, such as in [83, Chapter 5] and [84] where $\mathbf{K}$ is decomposed into a Kronecker product of smaller matrices. Using properties of the Kronecker product, all operations involving matrices can be done in $\mathcal{O}(N)$ time and space. However, this method does not scale with the number of features.

A more general and effective approach is to seek for a low rank approximation of $\mathbf{K}$ using a set of M inducing points $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^M$ with $M \ll N$ [85] and to assume that $f(\mathbf{X})$ and $f(\mathbf{x}_*)$ are conditionally independent given $f(\mathbf{Z})$. Such approximation reduces the complexity to $\mathcal{O}(NM^2)$ because only $\mathbf{K}_{MM}$ needs to be inverted, with $\mathbf{K}_{MM}$ the kernel matrix for $\mathbf{Z}$. To find these inducing points, different techniques were proposed: random projection [86], Nyström approximation [87] or Deterministic Training Conditional (DTC) approximation [88], etc.

In remote sensing, Bazi and Melgani [79] have used DTC for classification, which required less training time and provided similar accuracy. Still for classification, Morales-Alvarez *et al.* [89] have used random projection to construct the covariance matrix as well as variational posterior approximation.

An effective approach is to consider the optimization of the inducing points during the learning step, in complement to the mean and covariance function parameters, as proposed in [57], [90]. Such approach considers a variational approximation of the posterior (instead of model approximation) which gives superior results in large scale scenarios. Considering that, this work proposes to learn the inducing points by optimizing the posterior using variational inference, as discussed in the following part.

B. Posterior approximation by Variational Inference

Variational Inference (VI) aims to approximate the posterior distribution using a distribution q [38]. The core of the VI idea is to optimize parameters of q using a lower bound of the posterior (the *evidence lower bound - ELBO*). In the following, the inducing points are considered as latent variables that are optimized jointly with the prior parameters $\boldsymbol{\theta}$. Noting $\mathcal{E}$ the ELBO, the following result holds [38]:

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathbf{Z}) \geq \mathcal{E}(q),$$

where the model parameters are explicit. Hence optimizing $\mathcal{E}$ amounts to optimizing the log-marginal likelihood of the model.

In GP, the first formulation was proposed by Titsias [91] for the regression case and then was extended to classification problems by Hensman [57]. In remote sensing, VI was used to model heteroscedastic noise in GP regression [76] and for binary classification with model approximation [89].

Considering both the training and (non-observed) inducing points, the ELBO $\mathcal{E}$ is

$$\begin{aligned}\mathcal{E}(q) = \int & q\big(f(\mathbf{X}), f(\mathbf{Z})\big) \\ & \times \ln \left\{ \frac{p\big(\mathbf{y}, f(\mathbf{X}), f(\mathbf{Z})|\boldsymbol{\theta}\big)}{q\big(f(\mathbf{X}), f(\mathbf{Z})\big)} \right\} df(\mathbf{X}) df(\mathbf{Z})\end{aligned} \tag{12}$$

The variational distribution is defined as

$$q\big(f(\mathbf{X}), f(\mathbf{Z})\big) = p\big(f(\mathbf{X})|f(\mathbf{Z}), \boldsymbol{\theta}\big) q\big(f(\mathbf{Z})\big) \tag{13}$$

with $q\big(f(\mathbf{Z})\big) \sim \mathcal{N}_M(\mathbf{m}, \mathbf{S})$. We denote $\boldsymbol{\theta}^{\text{v}} = \{\mathbf{m}, \mathbf{S}\}$ the parameters of the variational distribution. Injecting (13) into (12) and simplifying leads to the bound proposed by [57]:

$$\begin{aligned}\mathcal{E}(q) = & \sum_{i=1}^{n} \mathbb{E}_{q(f(\mathbf{x}_i)|\boldsymbol{\theta}^{\text{v}}, \boldsymbol{\theta})} \Big[\log p\big(y_i | f(\mathbf{x}_i)\big) \Big] \\ & - \text{KL}\Big[q\big(f(\mathbf{Z})|\boldsymbol{\theta}^{\text{v}}, \boldsymbol{\theta}\big) \parallel p\big(f(\mathbf{Z})|\boldsymbol{\theta}\big) \Big],\end{aligned} \tag{14}$$

with

$$q\big(f(\mathbf{x}_i)|\boldsymbol{\theta}^{\text{v}},\boldsymbol{\theta}\big) \sim \mathcal{N}_1\Big(f(\mathbf{x}_i)|\ \mathbf{k}_{Mi}^{\top}\mathbf{K}_{MM}^{-1}\mathbf{m}, \\ k(\mathbf{x}_i,\mathbf{x}_i) - \mathbf{k}_{Mi}^{\top}\mathbf{K}_{MM}^{-1}\big(\mathbf{K}_{MM}-\mathbf{S}\big)\mathbf{K}_{MM}^{-1}\mathbf{k}_{Mi}\Big) \quad (15)$$

. The term KL is the Kullback-Leibler divergence between two distributions. Since the prior and the variational distribution are Gaussian, the Kullback-Leibler divergence can be computed and derived analytically [48, Chapter A.3.1]. The expectation term in (14) can be computed analytically for a regression problem [91]. The likelihoods cannot be calculated analytically as it was (10). However, it can be estimated using Gauss-Hermite quadrature (for binary problems) or by Monte Carlo (MC) sampling (for multi-class problems) [57]. The latter is discussed in the next section.

As explained in [92], the expectation term is factored over data points, it is thus possible to optimize (14) using stochastic optimization [93] without the $\mathcal{O}(N^3)$ computational and $\mathcal{O}(N^2)$ storage complexities. The resulting complexity is linear with the batch size and it is cubic with the number of inducing points. Using such strategy, Hensman *et al.* [57] optimized the whole model, i.e. $\{\boldsymbol{\theta},\boldsymbol{\theta}^{\text{v}},\mathbf{Z}\}$, on 700 000 points for a regression problem on a mono-CPU computer.

C. Variational Inference for multi-class GP classification

In this part, we describe how VI is applied to GP classification with LMC. First, as in [77], a more general LMC than in Section II-C is used: we assume that each univariate latent GP g_l has its own mean and covariance functions, with parameter $\boldsymbol{\theta}_l$. We also associate to each latent process a set of inducing points $\mathbf{Z}_l$ of size M[3] and we denote $\mathbf{g}(\mathbf{Z})$ the ML-dimensional random vector such as $\mathbf{g}(\mathbf{Z}) = \big[g_1(\mathbf{Z}_1),\dots,g_L(\mathbf{Z}_L)\big]^\top$. From the LMC definition in Section II-B, it follows that

$$p\big(\mathbf{g}(\mathbf{Z})|\boldsymbol{\Theta}\big) = \prod_{l=1}^{L} p\big(g_l(\mathbf{Z}_l)|\boldsymbol{\theta}_l\big)$$

with $p\big(g_l(\mathbf{Z}_l)|\boldsymbol{\theta}_l\big)$ Gaussian. Similarly, the variational distribution for $q\big(\mathbf{g}(\mathbf{Z})\big)$ is assumed to be such as

$$q\big(\mathbf{g}(\mathbf{Z})\big) = \prod_{l=1}^{L} q\big(g_l(\mathbf{Z}_l)\big)$$

with $q(g_l(\mathbf{Z}_l)) \sim \mathcal{N}_M(\mathbf{m}_l,\mathbf{S}_l)$. With these assumptions, the ELBO can be written as

$$\mathcal{E}(q) = \sum_{i=1}^{n}\mathbb{E}_{q(\mathbf{g}(\mathbf{x}_i)|\boldsymbol{\Theta}^{\text{v}},\boldsymbol{\Theta})}\Big[\log p\big(\mathbf{y}_i|\mathbf{g}(\mathbf{x}_i),\mathbf{A}\big)\Big] \\ - \sum_{l=1}^{L}\text{KL}\Big[q\big(g_l(\mathbf{Z}_l)|\boldsymbol{\theta}_l^{\text{v}},\boldsymbol{\theta}_l\big)\ \|\ p\big(g_l(\mathbf{Z}_l)|\boldsymbol{\theta}_l\big)\Big]. \quad (16)$$

with $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_L\}$, $\boldsymbol{\Theta}^{\text{v}} = \{\boldsymbol{\theta}_1^{\text{v}},\dots,\boldsymbol{\theta}_L^{\text{v}}\}$ and $q\left(\mathbf{g}(\mathbf{x}_i)|\boldsymbol{\Theta}^{\text{v}},\boldsymbol{\Theta}\right)$ being a L-dimensional Gaussian distribution with diagonal covariance matrix

$$q\left(\mathbf{g}(\mathbf{x}_i)|\boldsymbol{\Theta}^{\text{v}},\boldsymbol{\Theta}\right) \sim \mathcal{N}_L\big(\mathbf{g}(\mathbf{x}_i)|\mathfrak{m}^{\text{v}},\mathcal{K}^{\text{v}}\big). \quad (17)$$

Each marginal is given by (15), a consequence of the LMC: the latent processes become dependent on one another only during the computation of the likelihood. Specifically, the l^{th} element of the mean vector and of the diagonal of the covariance matrix are totally specified by the l^{th} latent process:

$$\mathfrak{m}_l^{\text{v}} = \mathbf{k}_{Mi}^{l^\top}\mathbf{K}_{MM}^{l^{-1}}\mathbf{m}_l, \quad (18)$$

$$\mathcal{K}_{ll}^{\text{v}} = k_l(\mathbf{x}_i,\mathbf{x}_i) - \mathbf{k}_{Mi}^{l^\top}\mathbf{K}_{MM}^{l^{-1}}\big(\mathbf{K}_{MM}^{l}-\mathbf{S}_l\big)\mathbf{K}_{MM}^{l^{-1}}\mathbf{k}_{Mi}^{l}. \quad (19)$$

As in the previous section, the KL terms can be computed and derived in closed-form. The expectation term needs to be approximated. MC sampling technique is used, similar to [57], [77]. It is combined with the so-called *reparametrization trick* from *variational auto-encoders* (VAE) to compute the derivative of the expectation during the stochastic gradient descent [94, section 2.4]. In practice, one realization is enough for the MC sampler during the training, as found in VAE [95], [57].

The prediction for a test sample uses the same variational approximation for the joint prior than in the marginal likelihood, and reduces to:

$$p(\mathbf{y}_*|\mathbf{Y},\mathbf{X},\mathbf{x}_*) = \mathbb{E}_{q(\mathbf{g}(\mathbf{x}_*)|\boldsymbol{\Theta}^{\text{v}},\boldsymbol{\Theta})}\Big[p\big(\mathbf{y}_*|\mathbf{g}(\mathbf{x}_*),\mathbf{A}\big)\Big] \quad (20)$$

with $q\left(\mathbf{g}(\mathbf{x}_*)|\boldsymbol{\Theta}^{\text{v}},\boldsymbol{\Theta}\right)$ given by (17). Again, the expectation is not analytically tractable: the approximation is obtained with MC sampling technique. The class is found by taking $\hat{c} = \arg\max_c \mathbf{y}_*$.

IV. Model description

In the previous section, a general large scale GP classification model based on variational approximation has been presented. In this section, we present the practical choices made for the classification of large scale SITS using this model: parametrization of the mean/covariance function, number of inducing points and initialization of the parameters. The trainable parameters are described in Fig. 1.

A. Mean function

For each latent function g_l, the mean function m_l is selected as a constant:

$$m_l(\mathbf{x}) = \mu_l, \quad (21)$$

the trainable parameter is therefore μ_l. It is initialized with the following value: $\mu_l = 0$.

B. Covariance function

In GP, the choice of the covariance function allows to introduce prior knowledge and to infer properties of GP posteriors [96], [48]. In remote sensing, the joint use of spatial, spectral and temporal information has shown to improve classification results [97]–[100]. A typical example is the use of composite kernels made of disjoint spatial and spectral parts for SVM hyper-spectral image classification [101]. As in SVM, the main idea is that GP can exploit the spatio-spectro-temporal structure of the data through the covariance function.

In this work, we define $k_l(\mathbf{x},\mathbf{x}')$ as a composition of a spatial covariance function $k_{l\phi}(\mathbf{x}_\phi,\mathbf{x}'_\phi)$ and a spectro-temporal

[3] For simplicity it is assumed that each latent process has the same number of inducing points.

covariance function $k_{l\lambda t}(\mathbf{x}_{\lambda t}, \mathbf{x}'_{\lambda t})$, with $\mathbf{x}_\phi$ and $\mathbf{x}_{\lambda t}$ being the spatial features and the spectro-temporal features, respectively. This configuration prevents two spatially distant pixels to be correlated even if they share a similar spectro-temporal profile. Indeed, distant pixels from different classes can have a similar vegetation phenology because of latitudinal and topographical effects on the biological cycle. Such modeling takes into account both the phenology and the spatial location in the studied area.

We propose to use the Radial Basis Function (RBF) kernel

$$k(\mathbf{x}, \mathbf{x}') = \alpha \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right)$$

for both k_ϕ and $k_{\lambda t}$. It has two parameters: $\alpha > 0$ and $\ell > 0$. This kernel uses isotropic distance between pixels in the spatial and spectro-temporal domain and the proximity between two pixels is controlled by the length-scale parameter ℓ: a small value tends to make all pixels uncorrelated ($k(\mathbf{x}, \mathbf{x}') \approx 0$) and a high value tends to increase the correlation between pixels ($k(\mathbf{x}, \mathbf{x}') \approx 1$).

Two different combinations of kernels have been investigated.

The first combination is the *sum of kernel*:

$$\begin{aligned} k_l^S(\mathbf{x}, \mathbf{x}') &= \alpha_{l\phi}^2 \times k_{l\phi}(\mathbf{x}_\phi, \mathbf{x}'_\phi) + \alpha_{l\lambda t}^2 \times k_{l\lambda t}(\mathbf{x}_{\lambda t}, \mathbf{x}'_{\lambda t}) \\ &= \alpha_{l\phi}^2 \exp\left(-\frac{\|\mathbf{x}_\phi - \mathbf{x}'_\phi\|_2^2}{2\ell_{l\phi}^2}\right) \\ &\quad + \alpha_{l\lambda t}^2 \exp\left(-\frac{\|\mathbf{x}_{\lambda t} - \mathbf{x}'_{\lambda t}\|_2^2}{2\ell_{l\lambda t}^2}\right) \end{aligned} \tag{22}$$

For each covariance function k_l^S, the trainable parameters are: $(\alpha_{l\phi}, \alpha_{l\lambda t}, \ell_{l\phi}, \ell_{l\lambda t})$. The scaling parameters $\alpha_{l\phi}$ and $\alpha_{l\lambda t}$ allow to give different weights to either spatial or spectro-temporal features. The second combination is the *product of kernel*:

$$\begin{aligned} k_l^P(\mathbf{x}, \mathbf{x}') &= k_{l\phi}(\mathbf{x}_\phi, \mathbf{x}'_\phi) \times k_{l\lambda t}(\mathbf{x}_{\lambda t}, \mathbf{x}'_{\lambda t}) \\ &= \exp\left(-\frac{\|\mathbf{x}_\phi - \mathbf{x}'_\phi\|_2^2}{2\ell_{l\phi}^2}\right) \\ &\quad \times \exp\left(-\frac{\|\mathbf{x}_{\lambda t} - \mathbf{x}'_{\lambda t}\|_2^2}{2\ell_{l\lambda t}^2}\right) \end{aligned} \tag{23}$$

For each covariance function k_l^P, the trainable parameters are: $(\ell_{l\phi}, \ell_{l\lambda t})$.

The trainable parameters are initialized with the following values:

- $\ell_{l\lambda t} = \sqrt{d}$, $\ell_{l\phi} = \sqrt{d'}$ with d and d' be respectively the square root of the features dimension, as it is usually done in kernel methods.
- $\alpha_{l\phi} = \alpha_{l\lambda t} = \ln 2$.

All parameters are reparameterized in log-scale to enforce positivity constraints during the learning step.

C. Inducing points (IP)

In this work, the number M of inducing points (IP) is common to each latent GP g_l. Different methods for the initialization of IP from the training set were investigated, such as random selection, clustering method (k-means) and defining a set of common or different IP per g_l. None of the investigated methods clearly outperforms the simplest one: random selection with the same set of M points for each g_l[4].

D. Model complexity

As described previously, for each latent function g_l, the same mean and kernel function were chosen as well as the same number of inducing points. Three different GP classification models were studied: λt-GP, $\phi\lambda t$-GPSC and $\phi\lambda t$-GPPC. λt-GP is a GP model using only the spectro-temporal covariance function $k_{\lambda t}(\mathbf{x}_{\lambda t}, \mathbf{x}'_{\lambda t})$. $\phi\lambda t$-GPSC and $\phi\lambda t$-GPPC are models with $k_l^S(\mathbf{x}, \mathbf{x}')$ and $k_l^P(\mathbf{x}, \mathbf{x}')$, whose covariance functions are defined in (22) and (23), respectively. Parameters for each model and their corresponding sizes are summarized in the Table I.

TABLE I: Description of the model parameters and their corresponding sizes. The last line corresponds to the total number of parameters for each model.

	λt-**GP**	$\phi\lambda t$-**GPSC**	$\phi\lambda t$-**GPSC**
k_l	1	4	2
m_l	1	1	1
$\mathbf{Z}_l$	$M \times d$	$M(d + d')$	$M(d + d')$
$\mathbf{m}_l$	M	M	M
$\mathbf{S}_l$	$\frac{M(M+1)}{2}$	$\frac{M(M+1)}{2}$	$\frac{M(M+1)}{2}$
$\mathbf{A}$	$L \times C$	$L \times C$	$L \times C$
Total	$L \times (1 + 1 +$ $d \times M +$ $M + \frac{M(M+1)}{2})$ $+ L \times C$	$L \times (4 + 1 +$ $(d + d') \times M +$ $M + \frac{M(M+1)}{2})$ $+ L \times C$	$L \times (2 + 1 +$ $(d + d') \times M +$ $M + \frac{M(M+1)}{2})$ $+ L \times C$

V. Data set

This section presents the different data sets used in the experiments and their corresponding pre-processing tasks. The *Southfrance* study area is located in the south of metropolitan France and it covers an area of approximately $200\,000\,\mathrm{km}^2$. It is composed of 27 Sentinel-2 tiles, as displayed in Fig. 2. The area provides a large variety of landscapes, ranging from coastal, through rural and urban, to mountainous areas for about two billion pixels.

A. SITS Sentinel-2

All available acquisitions of level 2A between January 2018 and December 2018 for the Sentinel-2 tiles were downloaded from the Theia Data Center[5].

[4] Some results are provided in the Supplementary Material, Section I.

[5] https://www.theia-land.fr/en/products/

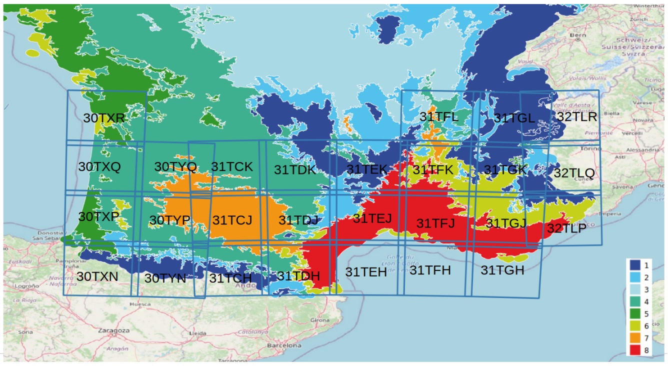


Fig. 2: Location of the 27 studied tiles where a blue square corresponds to one tile as provided by the Theia Data Center[5]. Each tile is displayed with its name in the Sentinel-2 nomenclature. Eco-climatic regions (region 1 to region 8) are displayed for the study area. (background map © OpenStreetMap contributors)

Surface reflectance time-series were produced using the MAJA processing chain, which corrects atmospheric, adjacency and slope effects, and provides cloud and shadow masks [102]. All spectral bands at a spatial resolution of 10 and 20m/pixel were used. Bands at 20m/pixel were spatially up-sampled to 10m/pixel using a bicubic interpolation, as implemented in the Orfeo ToolBox and its *SuperImpose* application [103]. In addition to the spectral channels, three spectral indices were also used: normalized difference vegetation index (NDVI), normalized difference water index (NDWI) and Brightness [104]. Furthermore, two geographic coordinates were also extracted for each pixel. These spatial features are in meters in the Lambert 93 projection. Thus for each pixel, 13 spectral features were extracted for each date in addition to two spatial features. To cope with the clouds/shadows and different temporal sampling among the tiles, the data have been linearly resampled onto a common set of virtual dates with an interval of 10 days, for a total of 37 dates [25].

Finally, a set of 483 features describes each pixel as $d + d'$ with:

- d : 37 interpolated dates × 13 spectral features,
- d' : 2 spatial features.

B. Reference data

The reference data used in this work is composed of 23 land cover classes ranging from artificial areas to vegetation and water bodies. It is the result of the fusion of different data sources:

1) CORINE Land Cover (CLC 2012): an inventory of land cover in 44 classes with a Minimum Mapping Unit of 25ha [105].
2) Urban Atlas (UA 2012): a geometrically accurate description of the various artificial cover types [106].
3) French National Geographic Institute (BD-Topo): a national topographical map [107].
4) Agricultural Land Parcel Information System, Registre Parcellaire Graphique (RPG 2018): a spatial register of agricultural parcels with the associated crop type as provided by farmer declarations [108].
5) The Randolph Glacier Inventory (RGI): a global inventory of glacier outlines [109].

Following the methodology described in [25], all the information from these different sources has been aggregated, both spatially and semantically, to create the reference data set. It is provided as a set of non-overlapping spatial polygons. The nomenclature of the 23 land cover classes can be found in Table II.

TABLE II: Land cover classes used for the experiments with their corresponding color code and their respective area.

Color	Code	Name	Area (km^2)
🟪	CUF	Continuous urban fabric	104
🟪	DUF	Discontinuous urban fabric	654
🟪	ICU	Industrial and commercial units	564
🟦	RSF	Road surfaces	62
🟨	RAP	Rapeseed	297
🟨	STC	Straw cereals	564
🟩	PRO	Protein crops	150
🟧	SOY	Soy	470
🟨	SUN	Sunflower	1 441
🟧	COR	Corn	1 030
🟦	RIC	Rice	77
🟫	TUB	Tubers / roots	49
🟫	GRA	Grasslands	1 167
🟪	ORC	Orchards and fruit growing	93
🟫	VIN	Vineyards	523
🟩	BLF	Broad-leaved forest	1 593
🟩	COF	Coniferous forest	4 934
🟩	NGL	Natural grasslands	3 386
🟩	WOM	Woody moorlands	1 713
🟥	NMS	Natural mineral surfaces	1 680
🟧	BDS	Beaches, dunes and sand plains	126
⬜	GPS	Glaciers and perpetual snows	164
🟦	WAT	Water bodies	14 567

C. Eco-climatic regions

As discussed in Section I, the complexity of the classification problem can be reduced by stratifying the spatial area into sub-regions. In this work, French metropolitan eco-climatic regions originaly proposed in [27] were used as strata. In each region, meteorological and topographical conditions are similar, thus the spectro-temporal variability of the pixel reflectance is reduced. All the eco-climatic regions are represented in the study area, but with varying proportions as shown in Fig. 3. Fig. 2 presents the eco-climatic regions over our *Southfrance* study area.

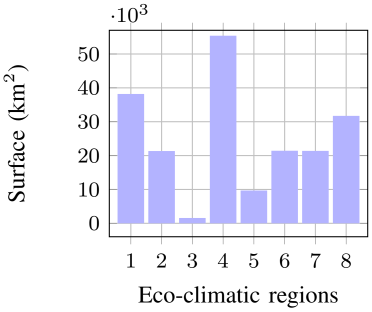


Fig. 3: Surface (in km^2) of each eco-climatic region in the *Southfrance* study area.

VI. Experimental setup

In this section, we explain how the data sets introduced in the previous section are used for the experiments. Moreover, we introduce three different competitive classification methods used in land cover classification for comparison.

A. Data generation

Two different data sets were produced using the `iota`[2] software [110]. A first data set called *classification* data set was used to train the models and to assess their accuracies. A second data set called *boundary* data set was used to evaluate the spatial continuity of the predictions in the boundary zones between two eco-climatic regions. A synthetic example of a boundary zone is represented in Fig. 4.

1) Classification data set: The *classification* data set was produced for each eco-climatic region. It is composed of three *spatially disjoint* data subsets: *training*, *validation* and *test*. The *training* subset was used to train the model while the *validation* subset was used to monitor the stochastic gradient descent and to detect over-fitting [93]. The *test* subset was used to estimate the performance of the model in terms of classification accuracy [111]. The term *spatially disjoint* indicates that pixels from one polygon fully belong to an unique data subset (either *training*, *validation* and *test*). 80 000, 20 000 and 100 000 polygons were extracted to build the training, validation and test polygons, respectively.

Next, pixels were randomly sampled from these polygons. Two sizes for the *training*-*validation* have been investigated for the learning step: (4 000, 1 000) and (16 000, 4 000) pixels per class, respectively called data set DS-A and data set DS-B. 10 000 pixels were extracted for the *test* set (except for the classes with fewer pixels, for which all were selected).

Two learning scenarios were considered: with and without stratification. For the former scenario (*stratification* configuration), a dedicated learning model was fit on each eco-climatic region, and global predictions were obtained by concatenating per-region model predictions over the full area. For the second scenario (*global* configuration), only one model was learned using pixels gathered from the eight *classification* data sets.

The performance of each model in terms of classification accuracy for the two scenarios was computed using classical classification metrics (overall accuracy (OA), F-score). To correctly estimate the classification metrics, 11 runs with different random pixel samplings were done. Table V in Appendix B provides the average number of pixels for each class and each region for the 11 *training*-*validation*-*test* pixels subsets.

2) Boundary data set: The spatial continuity of the model predictions at the border of two eco-climatic regions is assessed thanks to the *boundary* data set. A synthetic example of a *boundary* data set is represented in Fig. 4. It is composed of labeled and unlabeled pixels in a boundary zone around the boundary of two regions[6]. Several buffer sizes B have been investigated: $B \in \{100, 200, 500, 1000, 1500, 2000\}$ meters, the total width of the buffer being equal to $2 \times B$. All available labeled pixels were selected except those included in the *training* and *validation* data sets. From the available unlabeled pixels, approximately 1% were selected. Table VI in Appendix B summarizes the number of labeled and unlabeled pixels for each buffer size.

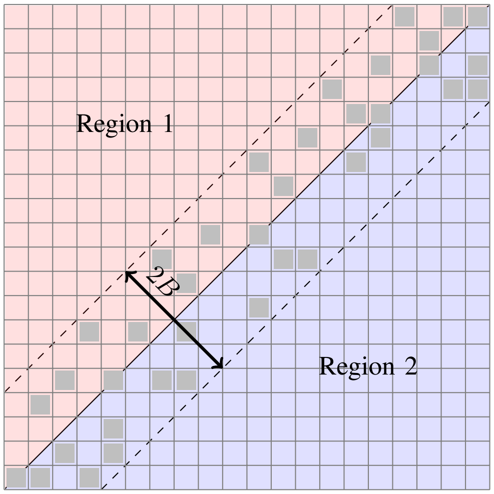


Fig. 4: Synthetic represention of a boundary zone: the full line represents the boundary between two eco-climatic regions and the area inside the dotted lines corresponds to the boundary zone. Gray pixels are selected to compose the *boundary* data set.

B. Pre-processing

Feature scaling was performed before the learning step. Mean and standard deviation were estimated for each feature on the training data set from the *classification* data set and then used to standardize the data on the different data sets (*training*, *validation*, *test* and *boundary*) [112]. The standardization was performed with the Scikit-Learn function *StandardScaler* [113].

C. Competitive methods

The GP model described in Section IV was implemented using the GPyTorch library. It is based on PyTorch and has been developed to exploit the usage of GPU hardware [114] for GP. The number of g_l latent functions was selected with $L = C$. The matrix $\mathbf{A}$ which is the linear relation between GP was initialized with random values drawn from a standard Gaussian distribution. The prediction was done by using 10 draws from the MC estimation. Studies have been made with a higher number of draws but results were similar.

Our GP model was compared with three different classification methods. The first two methods do not take into account the spectro-temporal structure of the data, e.g., modifying the order of the temporal acquisitions would not change the behavior of the algorithm. The last one takes the temporal structure into account to process the SITS.

a) Random Forest (RF): The Random Forest Classifier from the Scikit-Learn library [113] was used to train the RF model. Standard parameter settings were used: 100 trees with no maximum depth and the number of features considered for splitting at each leaf node was equal to the square root of the total number of features.

[6] Examples with real data are given in Supplementary Material, Section II.

b) Multi-layer Perceptron (MLP): The MLP model was built with four hidden layers. The number of neurons in the first layer was the number of features divided by two ($481/2 = 240$ or $483/2 = 241$) and in the last three layers: the number of classes multiplied by three ($23 \times 3 = 69$). The activation function used was the ReLU.

c) Lightweight Temporal Self-Attention (LTAE): In LTAE, temporal inputs were divided in channels distributed among several compact attention heads. Each head operated in parallel and extracted highly-specialized temporal features. These features were concatenated to create a single representation. A more detailed description and the parameters used from the LTAE model are given in [46]. The implementation was based on the Pytorch library.

For GP and neural networks, the Adam optimizer was used. Solver parameters are given in Table III. They were found by trial and error.

TABLE III: Parameter values for the Adam optimizer for *GP*, *MLP* and *LTAE*.

	GP	**MLP**	**LTAE**
Number of epochs E	100	300	100
Batch size β	1024	1000	1000
Learning rate η	1×10^{-3}	1×10^{-5}	1×10^{-5}

For each classification method previously defined, two different models were learned: λt-model and $\phi\lambda t$-model. λt-model was the model trained using only spectro-temporal features $\mathbf{x}_{\lambda t}$. $\phi\lambda t$-model was the one trained using spectro-temporal features $\mathbf{x}_{\lambda t}$ but also spatial features $\mathbf{x}_{\phi}$ as defined in Section V-A. The number of trainable parameters for each method in the *global* configuration classification is summarized in the Table IV.

TABLE IV: Number of trainable parameters for each model in the *global* configuration classification.

Model	**# of parameters**
λt-GP	584 200
$\phi\lambda t$-GPSC	586 569
$\phi\lambda t$-GPPC	586 523
λt-MLP	143 579
$\phi\lambda t$-MLP	144 612
λt-LTAE	239 521
$\phi\lambda t$-LTAE	240 005

VII. Results

First, this section describes the results obtained using the *classification* data set: the global performance accuracy of each method in the *Southfrance* area. Then, results concerning the spatial continuity of the predictions in the boundary zones using the *boundary* data set are studied. For each case, quantitative results but also qualitative ones are given.

A. *Performance results in the* Southfrance *area*

1) Quantitative results: Classification metrics were computed using the *test* data set from the *classification* data set in both configurations (*stratification* and *global*)[7]. They were computed over the 11 runs of each model trained either with the DS-A or the DS-B *training* data set. The studied models are: λt-GP, $\phi\lambda t$-GPSC, $\phi\lambda t$-GPPC, λt-RF, $\phi\lambda t$-RF, λt-MLP, $\phi\lambda t$-MLP, λt-LTAE and $\phi\lambda t$-LTAE.

The overall accuracy (OA) for each model trained with *training* data sets DS-A and DS-B are respectively presented using boxplots in Fig. 5a and Fig. 5b. The averaged F-score by class for each model trained with *training* data sets DS-A and DS-B on *global* and *stratification* configurations are presented using bar plots in Fig. 12, 13, 14, 15 in Appendix C. Raw results are reported in the supplementary materials.

The Wilcoxon rank-sum test [115] was used to assess the statistical significance of the observed differences over the MC runs, for each pair of classification methods. The null hypothesis was rejected at a significance level of alpha= 0.01. Results are reported in Fig. 6a. For the *stratification* configurations, all the results are significantly different. Similar results are found for the *global* configurations, except between λt-GP and $\phi\lambda t$-RF. For the larger training data set, results are reported in Fig. 6b. Some results are not significantly different in terms of classification accuracy.

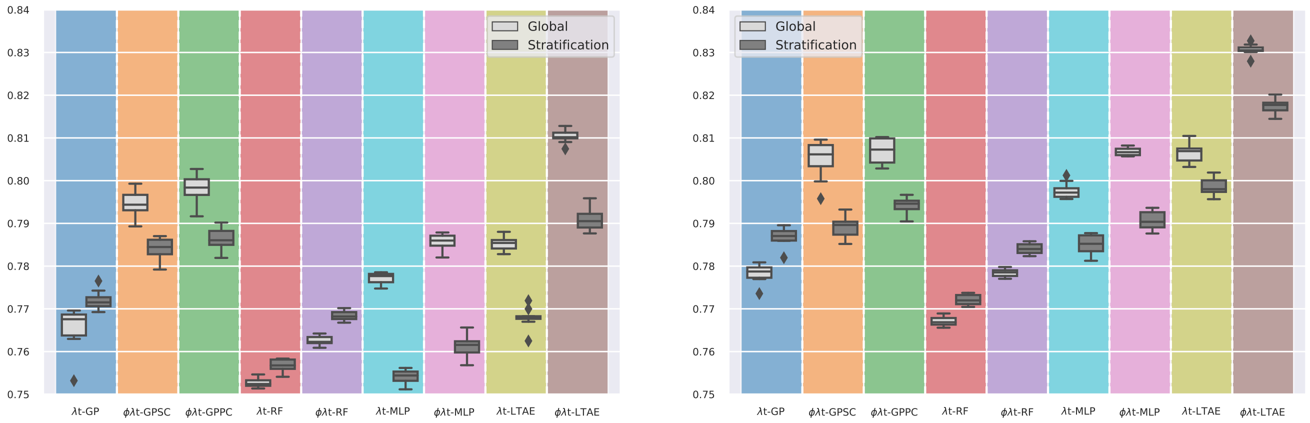
According to the Wilcoxon's tests, we can see that with the data set DS-A all models, for each configuration, benefit from the included spatial information. Indeed, the OA is increased by less than one point for RF and MLP models and between one and three points for GP and LTAE models. Furthermore, GP models take more advantage of the spatial information than the other methods: the OA is increased by three points compared to less than one point for RF and MLP and around two points for LTAE. Furthermore, GP models have the highest improvement, specifically in the *global* configuration. Only λt-GP and RF models have better results with the *stratification* configuration compared to the *global* one. Finally, by considering the best configuration, *global* with spatial information, GP models are in average three points above RF models, one point above MLP models and one point below LTAE models.

The averaged training and prediction times were computed for each region and each model over the 11 runs[8]. To process the RF models, 20 CPU with a total of RAM of 100 GB were available. For GP and DL models, 1 NVIDIA Tesla V100 GPU was used. In the *global* configuration, RF have the shortest training time followed by LTAE, MLP and finally GP. GP are more demanding, because of the MC sampling technique for the variational posterior. Moreover, $\phi\lambda t$-GPSC have higher training times compared to $\phi\lambda t$-GPPC, which can be explained by the presence of an indeterminate form for $\phi\lambda t$-GPSC.

2) Qualitative results: Land cover maps have been produced using the `iota`[2] processing chain [110] for both *stratification* and *global* configurations. Experiments were done for all the studied models previously described on two tiles: $T31TCJ$ and $T31TDJ$. For each region, predictions were done using the model trained with the *train* data set DS-A on the 27 tiles with the best OA over the 11 runs.

[7] A complete summary of classification metrics (OA, F-score, F-score per class, recall per class and precision per class) are provided in Supplementary Material, section IV.

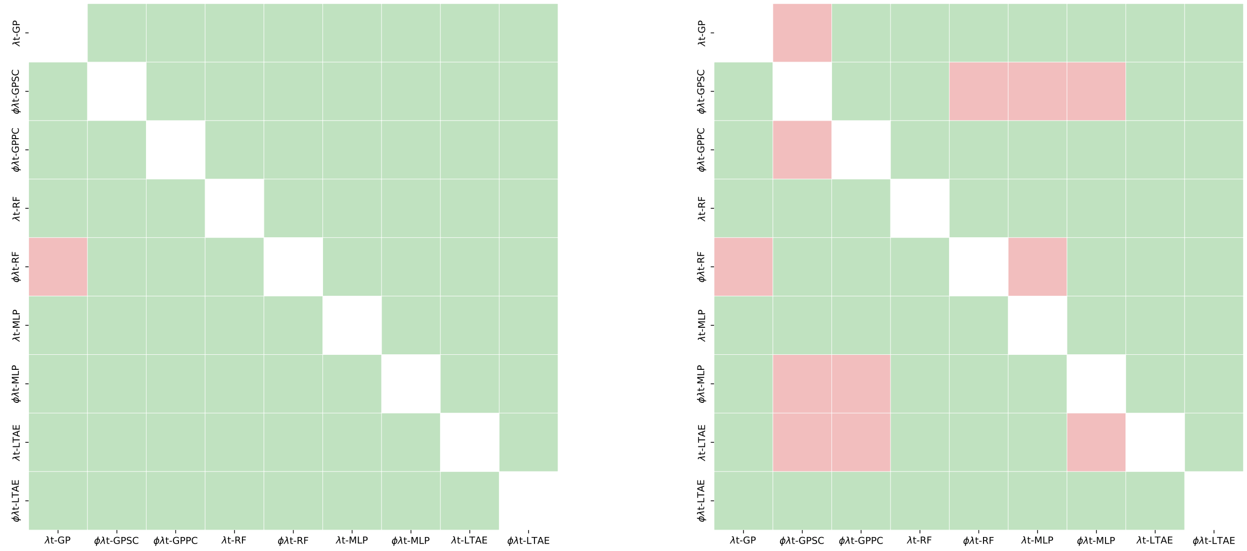
[8] Results are provided in the Supplementary Material, Section III.



(a) data set DS-A

(b) data set DS-B

Fig. 5: Boxplots of the overall accuracy for each model (λt-GP, $\phi\lambda t$-GPSC, $\phi\lambda t$-GPPC, λt-RF, $\phi\lambda t$-RF, λt-MLP, $\phi\lambda t$-MLP, λt-LTAE and $\phi\lambda t$-LTAE). Comparison between *global* and *stratification* configurations.



(a) data set DS-A

(b) data set DS-B

Fig. 6: Wilcoxon rank-sum tests results. Every pair of classification methods was tested and the null hypothesis was rejected at a significance level of alpha= 0.01. Red cells indicate that the observed differences in terms of F-score over the MC runs between the two classification methods are not significantly different. Green cells correspond to significant observed differences. The cells above the diagonal of the table contain Wilcoxon test results for the *stratification* configuration, while cells below the diagonal contain the results for the *global* configuration. Fig. 6a corresponds to results obtained with the data set DS-A and Fig. 6b corresponds to results obtained with the data set DS-B.

Fig. 7a and Fig. 7b represent the land cover map obtained with a GP model respectively trained without and with spatial information in the *global* configuration. The results obtained on this agricultural area around Toulouse show that the pixels are more homogeneous (with less salt and pepper classification noise [116]) when the spatial information is added[9]. Finally, all the land cover maps generated are available for download [10].

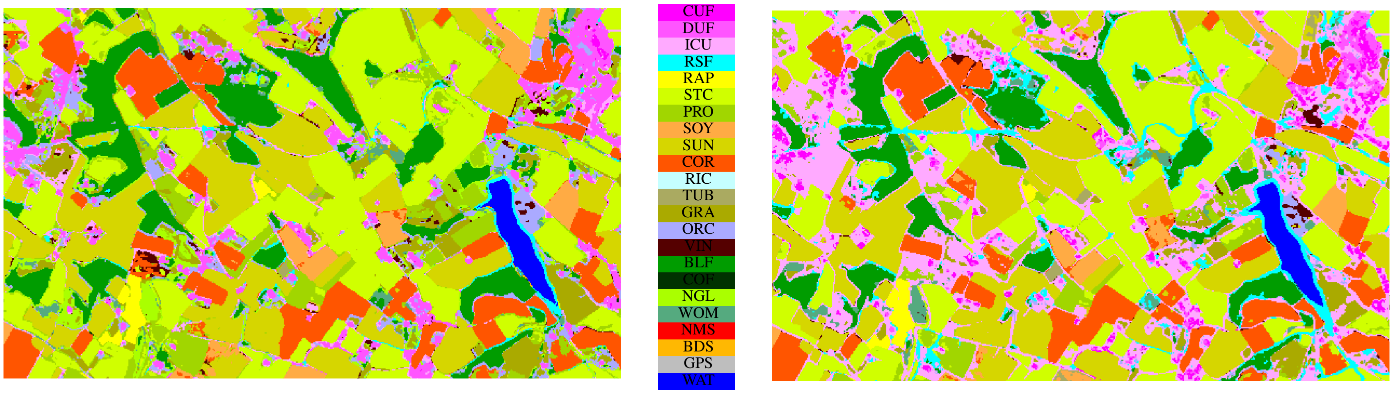
B. Continuity analysis in boundary zones

In the *stratification* configuration, two models surrounding a boundary zone were trained independently. The goal of this section is to evaluate the continuity of the predictions inside the boundary zone.

1) Quantitative results: All the pixels inside the boundary zone (i.e. *boundary* data set) were predicted by both models surrounding this zone. The number of agreements corresponds to the number of pixels predicted with the same label by both models. Thus, the percentage of agreement corresponds to the number of agreements divided by the total number of pixels predicted. This percentage was calculated for both unlabeled

[9] The land cover maps of this agricultural area around Toulouse for all the studied models in both configurations are provided in the Supplementary Material, Section V.

[10] DOI: 10.5281/zenodo.7077887



(a) model λt-GP

(b) model $\phi\lambda t$-GPPC

Fig. 7: Land cover maps obtained on an agricultural area around Toulouse (tile $T31TCJ$) (*global* configuration)

and labeled pixels which are correctly predicted. The size of the boundary has no influence as the results are similar for different boundary sizes $B \in \{100, 200, 500, 1000\}$, as shown in Table VII in Appendix D. In general, RF models have higher agreement than other models for both unlabeled and labeled pixels correctly predicted. However, the percentage of agreement with labeled pixels correctly predicted is around two points below the ones with the unlabeled pixels. Indeed, the continuity in predictions does not mean that the predictions are correct. The OA on labeled pixels for different boundary sizes for both *global* and *stratification* configurations are presented in Table VIII in Appendix D. For all methods, the OA in the *global* configuration is above the *stratification* one. The difference between both configurations is only two points for RF models and more than four points for DL methods. The *stratification* configuration is only beneficial for RF methods, as results found in Section VII-A1. For all models, the performances increase when the spatial information is added.

2) Qualitative results: As expected, there is no discontinuity in predictions between two eco-climatic regions for the *global* configuration with the λt-GP model as shown in Fig. 8b. However, for the *stratification* configuration discontinuities can be found. Specifically, we found discontinuities for zones where there is a high topography gradient and very scarce reference data, as shown in Fig. 8a. Similar results are obtained for all models[11]. Moreover, for the *stratification* configuration, adding the spatial information did not improve the prediction continuity[11] for every models.

VIII. Discussion

Beyond the quantitative and qualitative assessment of GP with respect to existing methodologies, we propose here an analysis of the specific attributes of GP. One advantage of GP with respect other ML or DL approaches is their ability to produce full posterior predictive distributions and not only point estimates. This aspect is discussed in Section VIII-A. As we claimed in Section I, another interesting feature of GP is the interpretability of the learned parameters. Section VIII-B focuses on the interpretation of some of these parameters and, in particular, their evolution during the training.

A. Posterior predictive distribution

As described in Section III, the posterior predictive distribution is not Gaussian and has to be estimated with MC sampling technique. For each sample, the class membership probabilities are computed by averaging the random draws. The class with the highest value is selected as the predicted class. In addition to the average value, standard deviation can also be computed as a measure of classifier uncertainty. Fig. 9 and Fig. 10 represent the approximate posterior predictive distributions obtained with 100 draws. The two largest class membership probabilities are represented for respectively a correctly predicted pixel and an incorrectly predicted one. In the case of a correctly predicted pixel, regardless of the draw, the model is very confident: the marginal distributions are tight, thus the variance is low. However, in the case of an incorrectly predicted pixel, we observe wide marginal distributions which can be interpreted as having higher uncertainty.

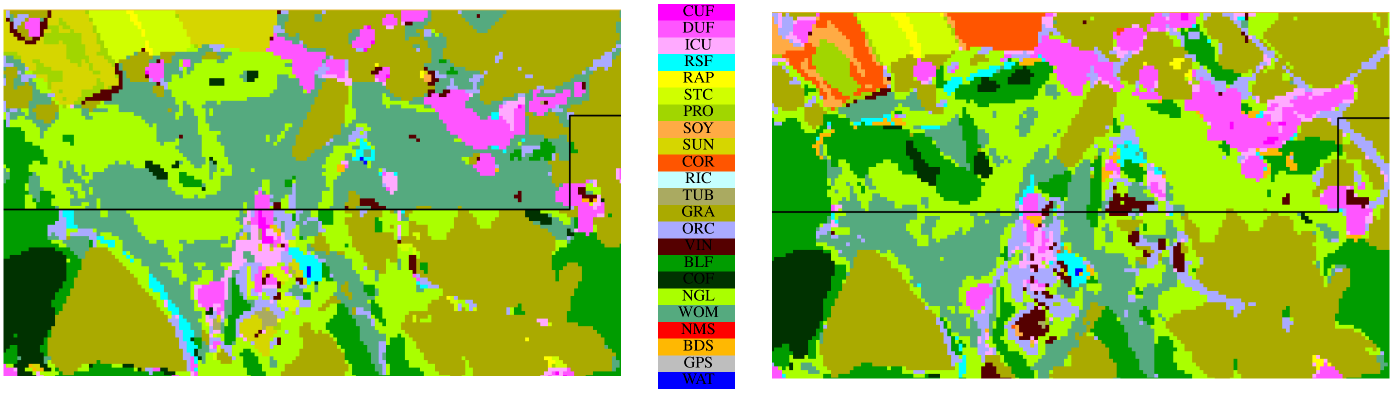
It is also possible to observe this trend by looking at the marginal distributions of the selected class membership for correctly or incorrectly predicted pixels. Indeed, Fig. 11 shows that, on average, the posterior predictive distribution of correctly predicted pixels has a higher mean but also a lower standard deviation than the posterior predictive distribution of incorrectly predicted pixels.

B. Interpretation of the learned parameters

As described in Section IV, different parameters are optimized during the training step. In the following, we focus our study on two different learned parameters: the inducing points $\mathbf{Z}_l$ and the matrix $\mathbf{A}$.

1) Spatial location of IP: Inducing points (IP) are used to approximate the posterior and their values are optimized to find a posterior as similar as possible to the true posterior on the training samples [75]. Relevant information can be obtained by looking at the IP after optimization. Visualizing the 481 spectro-temporal features is not possible and we restrict here to the two spatial features only, see Appendix E.

[11] Land cover maps in this specific boundary zone for all models are provided in Supplementary Material, Section VI.



(a) *stratification* configuration

(b) *global* configuration

Fig. 8: Land cover maps between two eco-climatic regions computed with the λt-GP model (tile $T31TDJ$). The black line represents the boundary between regions.

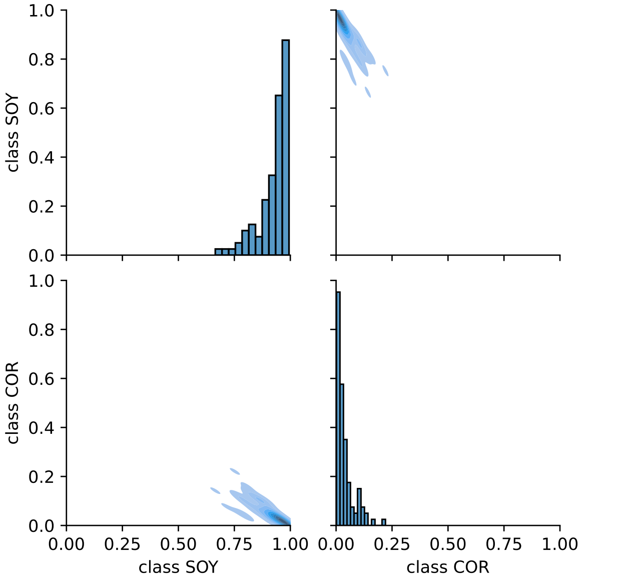


Fig. 9: Posterior predictive distributions, estimated with 100 draws, of the two largest class membership probabilities for a pixel correctly predicted (true class: SOY). Marginal distribution of each class is shown on the diagonal and joint distribution between two classes is shown on the off-diagonal.

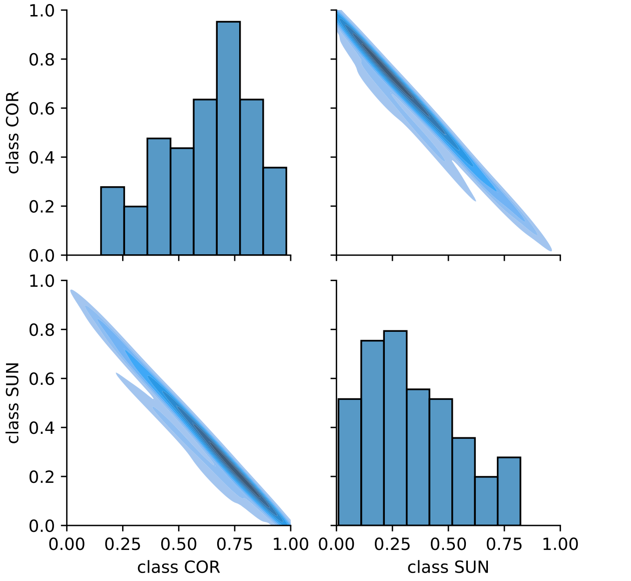


Fig. 10: Posterior predictive distributions, estimated with 100 draws, of the two largest class membership probabilities for a incorrectly predicted pixel (true class: SUN). Marginal distribution of each class is shown on the diagonal and joint distribution between two classes is shown on the off-diagonal.

The plotted ellipses represent the spatial area inside which the spatial correlation is greater than 0.9. The spatial distribution of the optimized IP can be qualified as regular: the points are more regularly spaced than in a random distribution. Also, the obtained spatio-length-scale ℓ_ϕ varies w.r.t. the latent GP, Appendix E represents their distribution. One possible interpretation is that the model achieves a multi-scale analysis in the spatial domain. Indeed, a latent GP with small spatio-length-scale perform a local analysis i.e. its spatial kernel rapidly tends to zero for closed pixels and thus limits its influence locally in the spatial domain. The latent GP with large spatio-length-scale performs a spatially wider analysis: the spatial kernel is always close to one, even for faraway pixels.

2) Weighting matrix of latent Gaussian Processes: As described in Section II-B, $\mathbf{A}$ is a mixing matrix: its coefficients a_{cl} are used to combine the L independent univariate latent GP g_l to estimate a final GP f_c such as $f_c = \sum_{i=1}^{L} a_{ci} g_i$. The a_{cl} can be interpreted as the contribution of a latent GP to the class-conditional posterior predictive distribution. Yet, we have found no specific pattern in $\mathbf{A}$ among the different results and we were not able to derive any specific interpretations: all GP contribute significantly. A possible extension would be to add sparsity constraints on $\mathbf{A}$ to improve the interpretability.

IX. Conclusions and perspectives

This work introduces an approach based on sparse variational Gaussian Processes (GP) for land cover pixel-based

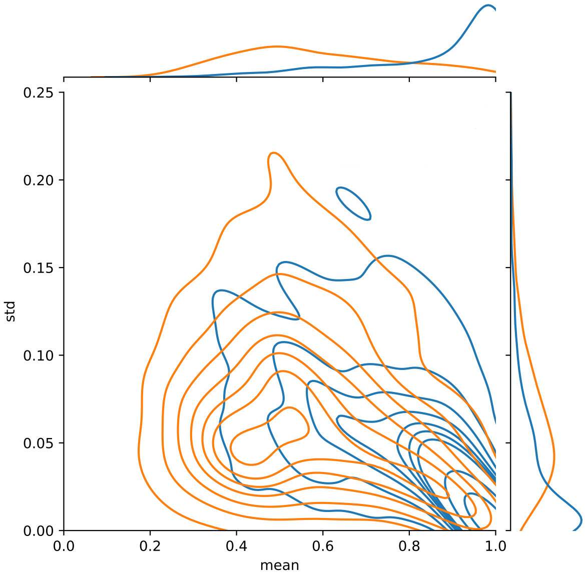


Fig. 11: Joint density of the standard deviation and the mean of the posterior predictive distribution for the selected class membership (obtained with 10 draws) and their respective marginal densities. — corresponds to 1000 correctly predicted pixels and — corresponds to 1000 incorrectly predicted pixels. The model $\phi\lambda t$-GPSC was trained on a *global* configuration.

classification at large scale. The discussed model combines sparse methods with variational inference and is able to scale to large data sets. The spatio-spectro-temporal structure of the SITS is taken into account through a dedicated covariance function. Experiments were conducted on Sentinel-2 SITS of the full year 2018 in an area of 200 000 km^2 in the south of France. In terms of accuracy, GP models outperformed conventional ML methods (i.e. RF) and DL methods (i.e. MLP). However, they are slightly worse than structured DL models (i.e. LTAE). Another finding is that spatial stratification is not necessary for advanced classifiers. Even worse, spatial discontinuities between adjacent regions are more severe for such classifiers w.r.t. RF.

Yet, spatial stratification in a large scale context can be of interest since the size of the training set is reduced and the different models can be trained in parallel. In such a case, a possible perspective would be to impose a smooth transition in terms of prediction between two spatial regions during the learning step. Following [117], we are considering to introduce an auxiliary GP linking pairs of adjacent regions at boundaries to constrain similar predictions in those areas.

Another perspective of this work is to implement feature extraction to take greater account of the spectro-temporal structure in the GP. The estimation of the inducing points involves a high number of parameters and is time-consuming: reducing the number of features could be beneficial for the convergence of the algorithm.

In the interest of reproducible research, the implementation of the models is made available in the following repository: https://gitlab.cesbio.omp.eu/belletv/land_cover_southfrance_gp.

Acknowledgment

The authors would like to thank Benjamin Tardy for his support and help during the generation of the different data sets and the production of land cover classification maps with the `iota`^{2} software. The authors would also like to thank CNES for the provision of its high performance computing (HPC) infrastructure to run the experiments presented in this paper and the associated help.

References

[1] F. Bertini, O. Brand, S. Carlier, U. Del Bello, M. Drusch, R. Duca, V. Fernandez, C. Ferrario, M. H. Ferreira, C. Isola, V. Kirschner, P. Laberinti, M. Lambert, G. Mandorlo, P. Marcos, P. Martimort, S. Moon, P. Oldeman, M. Palomba, and J. Pineiro, “Sentinel-2 esa's optical high-resolution mission for gmes operational services,” *ESA bulletin. Bulletin ASE. European Space Agency*, vol. SP-1322, 03 2012.

[2] C. Persello, J. D. Wegner, R. Hansch, D. Tuia, P. Ghamisi, M. Koeva, and G. Camps-Valls, “Deep Learning and Earth Observation to Support the Sustainable Development Goals: Current Approaches, Open Challenges, and Future Opportunities,” *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–30, 2022.

[3] C. Tarantino, M. Adamo, R. Lucas, and P. Blonda, “Change Detection in (Semi-) Natural Grassland Ecosystems for Biodiversity Monitoring Using Open Data,” in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 8981–8984, 2018.

[4] M. Fauvel, M. Lopes, T. Dubo, J. Rivers-Moore, P.-L. Frison, N. Gross, and A. Ouin, “Prediction of plant diversity in grasslands using Sentinel-1 and -2 satellite image time series,” *Remote Sensing of Environment*, vol. 237, p. 111536, 2020.

[5] N. Karasiak, D. Sheeren, M. Fauvel, J. Willm, J.-F. Dejoux, and C. Monteil, “Mapping tree species of forests in southwest France using Sentinel-2 image time series,” in *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, pp. 1–4, 2017.

[6] J. Dalimier, M. Claverie, B. Goffart, Q. Jungers, C. Lamarche, T. De Maet, and P. Defourny, “Characterizing the Congo Basin Forests by a Detailed Forest Typology Enriched with Forest Biophysical Variables,” in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 673–676, 2021.

[7] M. Pereira-Sandoval, A. Ruiz-Verdù, C. Tenjo, J. Delegido, P. Urrego, R. Pena, E. Vicente, J. Soria, J. Soria, and J. Moreno, “Calibration and Validation of Algorithms for the Estimation of Chlorophyll-A in Inland Waters with Sentinel-2,” in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 9276–9279, 2018.

[8] R. Lomelí-Huerta, H. Avila-George, J. P. Rivera-Caicedo, and M. De-la Torre, “Water Pollution Detection in Acapulco Coasts Using Merged Data from the Sentinel-2 and Sentinel-3 Satellites,” in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 1518–1521, 2021.

[9] S. Feng, J. Zhao, T. Liu, H. Zhang, Z. Zhang, and X. Guo, “Crop Type Identification and Mapping Using Machine Learning Algorithms and Sentinel-2 Time Series Data,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3295–3306, 2019.

[10] A. Moeini Rad, D. Ashourloo, H. Salehi Shahrabi, and H. Nematollahi, “Developing an Automatic Phenology-Based Algorithm for Rice Detection Using Sentinel-2 Time-Series Data,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 5, pp. 1471–1481, 2019.

[11] D. A. G. Dell'Aglio, M. Gargiulo, A. Iodice, D. Riccio, and G. Ruello, “Fire Risk Analysis by using Sentinel-2 Data: The Case Study of the Vesuvius in Campania, Italy,” in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6806–6809, 2020.

[12] J. Guo, Y. Luan, Z. Li, X. Liu, C. Li, and X. Chang, “Mozambique Flood (2019) Caused by Tropical Cyclone Idai Monitored From Sentinel-1 and Sentinel-2 Images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8761–8772, 2021.

[13] P. Soille, A. Burger, D. De Marchi, P. Kempeneers, D. Rodriguez, V. Syrris, and V. Vasilev, “A versatile data-intensive computing platform for information retrieval from big geospatial data,” *Future Generation Computer Systems*, vol. 81, pp. 30–40, 2018.

[14] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Newark, NJ: Wiley, 2005.

[15] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, “Advances in Hyperspectral Image Classification: Earth Monitoring with Statistical Learning Methods,” *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 45–54, 2014.

[16] P. J. Curran and P. M. Atkinson, “Geostatistics and remote sensing,” *Progress in Physical Geography: Earth and Environment*, vol. 22, no. 1, pp. 61–78, 1998.

[17] C. J. Paciorek and M. J. Schervish, “Spatial modelling using a new class of nonstationary covariance functions,” *Environmetrics*, vol. 17, no. 5, pp. 483–506, 2006.

[18] D. Higdon, J. Swall, and J. Kern, “Non-stationary spatial modeling,” 1998.

[19] M. Pal, “Kernel methods in remote sensing: a review,” *ISH Journal of Hydraulic Engineering*, vol. 15, no. sup1, pp. 194–215, 2009.

[20] G. Camps-Valls, L. Gomez-Chova, J. Calpe-Maravilla, J. Martin-Guerrero, E. Soria-Olivas, L. Alonso-Chorda, and J. Moreno, “Robust support vector method for hyperspectral data classification and knowledge discovery,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 7, pp. 1530–1542, 2004.

[21] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[22] Y. Bazi and F. Melgani, “Toward an Optimal SVM Classification System for Hyperspectral Remote Sensing Images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3374–3385, 2006.

[23] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] M. Belgiu and L. Drăguţ, “Random forest in remote sensing: A review of applications and future directions,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[25] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes, “Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series,” *Remote Sensing*, vol. 9, no. 1, 2017.

[26] C. Pelletier, S. Valero, J. Inglada, N. Champion, and G. Dedieu, “Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas,” *Remote Sensing of Environment*, vol. 187, pp. 156–168, 2016.

[27] D. Joly, T. Brossard, H. Cardot, J. Cavailhes, M. Hilal, and P. Wavresky, “Les types de climats en France, une construction spatiale,” *Cybergeo: European Journal of Geography*, June 2010.

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[29] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015. https://www.tensorflow.org/.

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning Representations by Back-propagating Errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[31] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation Applied to Handwritten Zip Code Recognition,” *Neural Computation*, vol. 1, pp. 541–551, 12 1989.

[32] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.

[33] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[35] M. Russwurm and M. Korner, “Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[36] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, “Land cover classification via multitemporal spatial data by deep recurrent neural networks,” *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1685–1689, 2017.

[37] X. Zhang, Y. Sun, K. Jiang, C. Li, L. Jiao, and H. Zhou, “Spatial sequential recurrent neural network for hyperspectral image classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 4141–4155, 2018.

[38] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[39] L. Zhong, L. Hu, and H. Zhou, “Deep learning based multi-temporal crop classification,” *Remote Sensing of Environment*, vol. 221, pp. 430–443, 2019.

[40] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, “Deep learning classification of land cover and crop types using remote sensing data,” *IEEE Geoscience and Remote Sensing Letters*, vol. 14, pp. 778–782, 2017.

[41] R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev, and J. Yosinski, “An intriguing failing of convolutional neural networks and the coordconv solution,” *Advances in neural information processing systems*, vol. 31, 2018.

[42] X. Yao, H. Yang, Y. Wu, P. Wu, B. Wang, X. Zhou, and S. Wang, “Land use classification of the deep convolutional neural network method reducing the loss of spatial features,” *Sensors*, vol. 19, no. 12, 2019.

[43] M. Rußwurm and M. Körner, “Multi-temporal land cover classification with sequential recurrent encoders,” *ISPRS International Journal of Geo-Information*, vol. 7, no. 4, 2018.

[44] R. M Rustowicz, R. Cheong, L. Wang, S. Ermon, M. Burke, and D. Lobell, “Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[45] C. Pelletier, G. Webb, and F. Petitjean, “Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series,” *Remote Sensing*, vol. 11, p. 523, Mar. 2019.

[46] V. Sainte Fare Garnot and L. Landrieu, “Lightweight temporal self-Attention for classifying satellite images time series,” in *Workshop on Advanced Analytics and Learning on Temporal Data*, AALTD, Sept. 2020.

[47] M. Russwurm, S. Lefèvre, and M. Körner, “BreizhCrops: A Satellite Time Series Dataset for Crop Type Identification,” in *Time Series Workshop of the 36th International Conference on Machine Learning (ICML)*, (Long Beach, United States), 2019.

[48] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[49] L. Pasolli, F. Melgani, and E. Blanzieri, “Gaussian Process Regression for Estimating Chlorophyll Concentration in Subsurface Waters From Remote Sensing Data,” *IEEE Geoscience and Remote Sensing Letters*, vol. 7, pp. 464–468, July 2010.

[50] Y. Bazi, N. Alajlan, and F. Melgani, “Improved estimation of water chlorophyll concentration with semisupervised gaussian process regression,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 7, pp. 2733–2743, 2012.

[51] M. Lazaro-Gredilla, M. K. Titsias, J. Verrelst, and G. Camps-Valls, “Retrieval of Biophysical Parameters With Heteroscedastic Gaussian Processes,” *IEEE Geoscience and Remote Sensing Letters*, vol. 11, pp. 838–842, Apr. 2014.

[52] G. Camps-Valls, J. Verrelst, J. Munoz-Mari, V. Laparra, F. Mateo-Jimenez, and J. Gomez-Dans, “A Survey on Gaussian Processes for Earth-Observation Data Analysis: A Comprehensive Investigation,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 58–78, 2016.

[53] M. Fauvel, C. Bouveyron, and S. Girard, “Parsimonious Gaussian Process Models for the Classification of Hyperspectral Remote Sensing

Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 12, pp. 2423–2427, 2015.

[54] P. Morales-Álvarez, A. Pérez-Suay, R. Molina, and G. Camps-Valls, "Remote Sensing Image Classification With Large-Scale Gaussian Processes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1103–1114, 2018.

[55] A. Constantin, M. Fauvel, and S. Girard, "Joint Supervised Classification and Reconstruction of Irregularly Sampled Satellite Image Times Series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 4403913, May 2021.

[56] H. Liu, Y.-S. Ong, X. Shen, and J. Cai, "When gaussian process meets big data: a review of scalable gps," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, pp. 4405–4423, Jan. 2020.

[57] J. Hensman, A. Matthews, and Z. Ghahramani, "Scalable Variational Gaussian Process Classification," in *In Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 351–360, 2015.

[58] H. Hassouna, F. Melgani, and Z. Mokhtari, "Spatial contextual gaussian process learning for remote-sensing image classification," *Remote Sensing Letters*, vol. 6, no. 7, pp. 519–528, 2015.

[59] S. Sun, P. Zhong, H. Xiao, and R. Wang, "Spatial contextual classification of remote sensing images using a gaussian process," *Remote Sensing Letters*, vol. 7, no. 2, pp. 131–140, 2016.

[60] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[61] S. S. Ghosh, S. Dey, N. Bhogapurapu, S. Homayouni, A. Bhattacharya, and H. McNairn, "Gaussian Process Regression Model for Crop Biophysical Parameter Retrieval from Multi-Polarized C-Band SAR Data," *Remote Sensing*, vol. 14, no. 4, 2022.

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

[63] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, "Retrieval of Vegetation Biophysical Parameters Using Gaussian Process Techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1832–1843, 2012.

[64] D. H. Svendsen, L. Martino, M. Campos-Taberner, F. J. García-Haro, and G. Camps-Valls, "Joint Gaussian Processes for Biophysical Parameter Retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1718–1727, 2018.

[65] E. V. Bonilla, K. Chai, and C. Williams, "Multi-task Gaussian Process Prediction," *NIPS Foundation*, vol. 20, 2007.

[66] L. Pipia, J. Muñoz-Marí, E. Amin, S. Belda, G. Camps-Valls, and J. Verrelst, "Fusing optical and sar time series for lai gap filling with multioutput gaussian processes," *Remote Sensing of Environment*, vol. 235, p. 111452, 2019.

[67] M. A. Álvarez, L. Rosasco, and N. D. Lawrence, "Kernels for Vector-Valued Functions: A Review," *Found. Trends Mach. Learn.*, vol. 4, p. 195–266, mar 2012.

[68] A. G. Journel and C. J. Huijbregts, "Mining geostatistics," *Academic Press*, 1976.

[69] P. Goovaerts and D. Goovaerts, *Geostatistics for Natural Resources Evaluation*. Applied geostatistics series, Oxford University Press, 1997.

[70] N. Durrande, D. Ginsbourger, and O. Roustant, "Additive Kernels for Gaussian Process Modeling," Jan. 2010.

[71] A. Mateo-Sanchis, J. Munoz-Mari, M. Campos-Taberner, J. Garcia-Haro, and G. Camps-Valls, "Gap Filling of Biophysical Parameter Time Series with Multi-Output Gaussian Processes," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, jul 2018.

[72] P. Boyle and M. Frean, "Dependent Gaussian Processes," *MIT Press*, vol. 17, 2004.

[73] D. Higdon, "Space and Space-Time Modeling using Process Convolutions," in *Quantitative Methods for Current Environmental Issues* (C. W. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi, eds.), pp. 37–56, London: Springer London, 2002.

[74] M. van der Wilk, C. E. Rasmussen, and J. Hensman, "Convolutional Gaussian Processes," in *NIPS* (I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds.), pp. 2849–2858, 2017.

[75] M. van der Wilk, V. Dutordoir, S. John, A. Artemev, V. Adam, and J. Hensman, "A Framework for Interdomain and Multioutput Gaussian Processes," 2020.

[76] P. Moreno-Muñoz, A. Artés, and M. Álvarez, "Heterogeneous multi-output gaussian process prediction," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[77] A. G. Wilson, Z. Hu, R. R. Salakhutdinov, and E. P. Xing, "Stochastic Variational Deep Kernel Learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[78] R. M. Neal, "Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification," *arXiv: Data Analysis, Statistics and Probability*, 1997.

[79] Y. Bazi and F. Melgani, "Gaussian Process Approach to Remote Sensing Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 1, pp. 186–197, 2010.

[80] C. Williams and D. Barber, "Bayesian classification with Gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998.

[81] T. P. Minka, "Expectation Propagation for Approximate Bayesian Inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, (San Francisco, CA, USA), pp. 362–369, Morgan Kaufmann Publishers Inc., 2001.

[82] J. Quiñonero-Candela and C. E. Rasmussen, "A Unifying View of Sparse Approximate Gaussian Process Regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1939–1959, Dec. 2005.

[83] Y. Saatci, *Scalable Inference for Structured Gaussian Process Models*. Ph.D. dissertation, University of Cambridge, 2011.

[84] A. G. Wilson, E. Gilboa, A. Nehorai, and J. P. Cunningham, "Fast Kernel Learning for Multidimensional Pattern Extrapolation," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, (Cambridge, MA, USA), pp. 3626–3634, MIT Press, 2014.

[85] M. Van der Wilk, *Sparse Gaussian process approximations and applications*. PhD thesis, University of Cambridge, 2019.

[86] M. Lázaro-Gredilla, J. Quiñnero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse Spectrum Gaussian Process Regression," *Journal of Machine Learning Research*, vol. 11, no. 63, pp. 1865–1881, 2010.

[87] C. Williams and M. Seeger, "Using the Nyström Method to Speed Up Kernel Machines," in *Advances in Neural Information Processing Systems 13*, pp. 682–688, MIT Press, 2001.

[88] M. Seeger, *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximation*. PhD thesis, University of Edinburgh, December 2003.

[89] P. Morales-Alvarez, A. Perez-Suay, R. Molina, and G. Camps-Valls, "Remote Sensing Image Classification with Large Scale Gaussian Processes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, pp. 1103–1114, Feb. 2018.

[90] E. Snelson and Z. Ghahramani, "Sparse gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing Systems 18*, pp. 1257–1264, MIT press, 2006.

[91] M. K. Titsias, "Variational Learning of Inducing Variables in Sparse Gaussian Processes," in *In Artificial Intelligence and Statistics 12*, pp. 567–574, 2009.

[92] J. Hensman, N. Fusi, and N. D. Lawrence, "Gaussian Processes for Big Data," in *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, (Arlington, Virginia, USA), p. 282–290, AUAI Press, 2013.

[93] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization Methods for Large-Scale Machine Learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[94] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[95] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *ICLR* (Y. Bengio and Y. LeCun, eds.), 2014.

[96] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[97] F. Pacifici, M. Chini, and W. J. Emery, "A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification," *Remote Sensing of Environment*, vol. 113, pp. 1276–1292, June 2009.

[98] D. Tuia, F. Pacifici, M. Kanevski, and W. Emery, "Classification of Very High Spatial Resolution Imagery Using Mathematical Morphology and Support Vector Machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, pp. 3866–3879, Nov. 2009.

[99] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Rojo-Alvarez, and M. Martinez-Ramon, "Kernel-Based Framework for Multitemporal and Multisource Remote Sensing Data Classification and Change

Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, pp. 1822–1835, June 2008.

[100] M. Fauvel, J. Chanussot, J. A. Benediktsson, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," in *2007 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4834–4837, 2007.

[101] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "A Spatial-Spectral Kernel-Based Approach for the Classification of Remote-Sensing Images," *Pattern Recogn.*, vol. 45, p. 381–392, jan 2012.

[102] L. Baetens, C. Desjardins, and O. Hagolle, "Validation of Copernicus Sentinel-2 Cloud Masks Obtained from MAJA, Sen2Cor, and FMask Processors Using Reference Cloud Masks Generated with a Supervised Active Learning Procedure," *Remote Sensing*, vol. 11, p. 433, Feb. 2019.

[103] O. D. Team, "Orfeo ToolBox 7.1," Mar. 2020. https://zenodo.org/record/3715021.

[104] V. Henrich, E. Götze, A. Jung, C. Sandow, D. Thürkow, and C. Gläßer, "Development of an online indices database: Motivation, concept and implementation," in *Proceedings of the 6th EARSeL Imaging Spectroscopy SIG Workshop Innovative Tool for Scientific and Commercial Environment Applications, Tel Aviv, Israel*, pp. 16–18, 2009.

[105] M. Bossard, J. Feranec, J. Otahel, and others, *CORINE land cover technical guide: Addendum 2000*, vol. 40. European Environment Agency Copenhagen, 2000.

[106] E. Montero, J. Van Wolvelaer, and A. Garzón, *The European Urban Atlas*. Springer Netherlands, 2014.

[107] E. Maugeais, F. Lecordix, X. Halbecq, and A. Braun, "Dérivation cartographique multi échelles de la BDTopo de l'IGN France: mise en oeuvre du processus de production de la Nouvelle Carte de Base," in *Proc 25th Int Cartogr Conf Paris*, pp. 3–8, 2011.

[108] P. Cantelaube and M. Carles, "Le registre parcellaire graphique: des données géographiques pour décrire la couverture du sol agricole," *Cahier des Techniques de l'INRA*, no. Méthodes et techniques GPS et SIG pour la conduite de dispositifs expérimentaux, pp. 58–64, 2014.

[109] W. T. Pfeffer, A. A. Arendt, A. Bliss, T. Bolch, J. G. Cogley, A. S. Gardner, J.-O. Hagen, R. Hock, G. Kaser, C. Kienholz, E. S. Miles, G. Moholdt, N. Mölg, F. Paul, V. Radić, P. Rastner, B. H. Raup, J. Rich, M. J. Sharp, and The Randolph Consortium, "The Randolph Glacier Inventory: a globally complete inventory of glaciers," *Journal of Glaciology*, vol. 60, no. 221, pp. 537–552, 2014.

[110] J. Inglada, A. Vincent, M. Arias, and B. Tardy, *iota2-a25386*, July 2016. https://doi.org/10.5281/zenodo.58150.

[111] C. M. Bishop, *Neural networks for pattern recognition*. Oxford : New York: Clarendon Press ; Oxford University Press, 1995.

[112] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall/CRC, 1 ed., July 2019.

[113] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[114] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, "Gpytorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration," in *Advances in Neural Information Processing Systems*, 2018.

[115] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[116] H. Hirayama, R. C. Sharma, M. Tomita, and K. Hara, "Evaluating multiple classifier system for the reduction of salt-and-pepper noise in the classification of very-high-resolution satellite images," *International Journal of Remote Sensing*, vol. 40, no. 7, pp. 2542–2557, 2019.

[117] C. Park and D. Apley, "Patchwork kriging for large-scale gaussian process regression," *J. Mach. Learn. Res.*, vol. 19, p. 269–311, jan 2018.

Valentine Bellet Valentine Bellet (Student Member, IEEE) received the master's degree in automatic control and electronics from INSA Toulouse, France, in 2019. She is currently pursuing the Ph.D. degree with the Centre d'Etudes Spatiales de la Biosphère (CESBIO) Laboratory, Toulouse, France. She is working on the subject of land cover pixel-based classification with satellite image time-series (SITS) at national scale.

Mathieu Fauvel Mathieu Fauvel (Senior Member, IEEE) received the Ph.D. degree in image and signal processing from the Grenoble Institute of Technology, Grenoble, France, in 2007. From 2008 to 2010, he was a Post-Doctoral Researcher with the MISTIS Team, National Institute for Research in Digital Science and Technology (INRIA). From 2011 to 2018, he was an Associate Professor with the DYNAFOR Lab (INRA), National Polytechnic Institute of Toulouse, Toulouse, France. Since 2018, he has been a Researcher at INRAe and CESBIO-lab. His research interests are remote sensing, machine learning, and image processing.

Jordi Inglada Jordi Inglada received the master's degree in telecommunications engineering from the Universitat Politècnica de Catalunya, Barcelona, Spain, and the École Nationale Supérieure des Télécommunications de Bretagne, Brest, France, in 1997, and the Ph.D. degree in signal processing and telecommunications from the Université de Rennes 1, Rennes, France, in 2000. He is currently with the Centre National d'Etudes Spatiales (French Space Agency), Toulouse, France, where he is involved in the field of remote sensing image processing at the Centre d'Etudes Spatiales de la Biosphère (CESBIO) Laboratory. He is involved in the development of image processing algorithms for the operational exploitation of Earth observation images, mainly in the field of multitemporal image analysis for land use and cover change.

Appendix A
Symbols and notations

Symbol	Meaning
$\mathbf{A}$	Mixing matrix, $\mathbf{A} \in \mathbb{R}^{C \times L}$
α_l, ℓ_l	Scaling, length-scale parameter of the covariance function k_l
C	Number of classes, $c \in \{1, ..., C\}$
d, d'	Number of spectro-temporal, spatial features
$f \sim \mathcal{GP}(m, k)$	Univariate GP with mean function m and covariance function k
$\mathbf{f} \sim \mathcal{GP}(\mathfrak{m}, \mathcal{K})$	P-multivariate GP such as $\mathbf{f} = \mathbf{A}\mathbf{g}$ with mean function $\mathfrak{m}$ and covariance function $\mathcal{K}$
$g_l \sim \mathcal{GP}(m_l, k_l)$	Univariate GP, the l^{th} latent GP with mean function m_l and covariance function k_l
$\mathbf{g}$	Vector of L independent univariate GP, $\mathbf{g} = [g_1, ..., g_L]$
$\mathbf{k}_*$	Covariance vector between the training inputs and the test inputs $\mathbf{k}_* = [k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_N, \mathbf{x}_*)]^\top$
$\mathbf{K}$	Covariance matrix such as $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\forall i, j \in \{1, \ldots, N\}^2$
$\mathbf{K}_o$	Covariance matrix such as $K_{o,ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$, $\forall i, j \in \{1, \ldots, N\}^2$
$\mathcal{K}^{\mathrm{v}}$	Covariance matrix of the L-dimensional distribution $q\left(\mathbf{g}(\mathbf{x}_i) \mid \boldsymbol{\theta}^{\mathrm{v}}, \boldsymbol{\theta}\right) \sim \mathcal{N}_L\big(\mathbf{g}(\mathbf{x}_i) \mid \mathfrak{m}^{\mathrm{v}}, \mathcal{K}^{\mathrm{v}}\big)$
$\mathcal{K}^{\mathrm{v}}_{ll}$	The diagonal l^{th} element of diagonal covariance matrix $\mathcal{K}^{\mathrm{v}}$
L	Number of latent processes, $l \in \{1, ..., L\}$
$\mathbf{m}$	Mean vector of the variational distribution $q(f(\mathbf{Z})) \sim \mathcal{N}_M(\mathbf{m}, \mathbf{S})$
$\mathfrak{m}^{\mathrm{v}}$	Mean matrix of the L-dimensional distribution $q\left(\mathbf{g}(\mathbf{x}_i) \mid \boldsymbol{\theta}^{\mathrm{v}}, \boldsymbol{\theta}\right) \sim \mathcal{N}_L\big(\mathbf{g}(\mathbf{x}_i) \mid \mathfrak{m}^{\mathrm{v}}, \mathcal{K}^{\mathrm{v}}\big)$
M	Number of inducing points
$\boldsymbol{\mu}$	Mean vector such as $\boldsymbol{\mu} = \big[m(\mathbf{x}_1), \ldots, m(\mathbf{x}_N)\big]^\top$
$\boldsymbol{\mu}_o$	Mean vector such as $\boldsymbol{\mu}_o = [\mathfrak{m}(\mathbf{x}_1), \ldots, \mathfrak{m}(\mathbf{x}_N)]^\top$
N	Number of training inputs
$\mathcal{N}_N(\boldsymbol{\mu}, \mathbf{K})$	Multivariate Gaussian distribution of N dimension with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$
P	Number of output GP, $p \in \{1, ..., P\}$
$q(f(\mathbf{Z}))$	Variational distribution $q(f(\mathbf{Z})) \sim \mathcal{N}_M(\mathbf{m}, \mathbf{S})$
$\mathbf{S}_l$	Covariance matrix of the distribution $g_l(\mathbf{Z}_l) \sim \mathcal{N}_M(\mathbf{m}_l, \mathbf{S}_l)$ q
$\boldsymbol{\Theta}$	Hyper-parameters of $\mathbf{g}$, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_L\}$
$\boldsymbol{\theta}_{lk}$	Parameters of the covariance function k_l
$\boldsymbol{\theta}_{lm}$	Parameters of the mean function m_l
$\boldsymbol{\theta}_l$	Hyper-parameters of the latent process g_l, $\boldsymbol{\theta}_l = \{\boldsymbol{\theta}_{lm}, \boldsymbol{\theta}_{lk}\}$
$\boldsymbol{\theta}^V$	Parameters of the variational distribution q, $\boldsymbol{\theta}^V = \{\mathbf{m}, \mathbf{S}\}$
$\boldsymbol{\Theta}^V$	Parameters of all the variational distributions $\boldsymbol{\Theta}^V = \{\boldsymbol{\theta}_l^V, ..., \boldsymbol{\theta}_L^V\}$
$\mathbf{x}_i, \mathbf{y}_i$	The i^{th} training input, target
$\mathbf{x}_{i\phi}$, $\mathbf{x}_{i\lambda t}$	Spatial, spectro-temporal features of the i^{th} pixel
$\mathbf{X}$	Set of training inputs $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$
$\mathbf{Y}$	Set of training targets $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n]$
$\mathbf{z}_i$	The i^{th} inducing point
$\mathbf{Z}_l$	Set of inducing points for the latent process g_l, $\mathbf{Z}_l = \{\mathbf{z}_{li}\}_{i=1}^M$

APPENDIX B
NUMBER OF PIXELS IN CLASSIFICATION AND BOUNDARY DATA SETS.

For both tables, the class code is provided in Table II.

TABLE V: Average number of pixels per class and regions for the *classification* data set. For a given class, the two first rows (data set DS-A and B) indicate the number of *training-validation* pixels per region and the third rows indicates the number of *test* pixels per region.

				Regions					Global
Class	1	2	3	4	5	6	7	8	
CUF	4 000 - 1 000 6 569 - 1 727 7 286	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 12 011 - 2 676 10 000	4 000 - 1 000 10 802 - 2 657 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	32 000 - 8 000 109 382 - 27 061 77 286
DUF	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	32 000 - 8 000 128 000 - 32 000 80 000
ICU	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	32 000 - 8 000 128 000 - 32 000 80 000
RSF	3 939 - 966 5 191 - 2 104 6 622	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 7 642 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 9 148 - 2 769 10 000	2 562 - 658 2 562 - 658 5 360	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	30 501 - 7 624 88 543 - 23 457 71 982
RAP	4 000 - 987 5 942 - 1 424 4551	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 941 7 261 - 2 125 10 000	32 000 - 7 928 109 204 - 27 549 74 551
STC	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	32 000 - 8 000 128 000 - 32 000 80 000
PRO	1 073 - 340 1 073 - 340 1 222	4 000 - 1 000 9 748 - 2 596 10 000	1 188 - 363 1 188 - 363 3 120	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 11 945 - 2 709 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 13 154 - 3 243 10 000	26 261 - 6 704 85 110 - 21 253 64 342
SOY	3 998 - 902 4 362 - 1 122 7 098	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 3 959 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 344 10 000	31 998 - 7 902 116 362 - 28 525 77 098
SUN	1 316 - 437 1 437 - 1 122 3 492	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 3 757 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 3 959 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 344 10 000	29 316 - 7 437 113 316 - 28 194 73 492
COR	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	32 000 - 8 000 128 000 - 32 000 80 000
RIC	0 - 0 0 - 0 0	0 - 0 0 - 0 0	0 - 0 0 - 0 0	0 - 0 0 - 0 0	0 - 0 0 - 0 0	4 000 - 1 000 16 000 - 4 000 10 000	0 - 0 0 - 0 0	4 000 - 1 000 16 000 - 4 000 10 000	8 000 - 2 000 32 000 - 8 000 20 000
TUB	1 604 - 411 1 604 - 411 1 816	3 836 - 912 3 928 - 1 078 5 185	2 757 - 676 2 757 - 676 5 864	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 988 8 688 - 2 563 10 000	4 000 - 1 000 11 518 - 3 296 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 3 985 10 000	28 199 - 6 988 76 497 - 20 011 62 865
GRA	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	32 000 - 8 000 128 000 - 32 000 80 000
ORC	844 - 173 844 - 173 657	4 000 - 1 000 15 967 - 3 930 10 000	1 175 - 343 1 175 - 343 3 026	4 000 - 1 000 16 000 - 4 000 10 000	3 236 - 800 3 236 - 965 3 590	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	25 256 - 6 317 85 223 - 21 412 57 273
VIN	672 - 207 672 - 207 574	4 000 - 987 5 399 - 1 545 5 115	4 000 - 1 000 6 255 - 1 649 9 200	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	28 672 - 7 194 92 327 - 23 402 64 889
BLF	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	32 000 - 8 000 128 000 - 32 000 80 000
COF	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	2 598 - 648 2 598 - 717 5 317	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 3 896 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	30 598 - 7 648 114 598 - 28 614 75 317
NGL	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	0 - 0 0 - 0 0	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	28 000 - 7 000 112 000 - 28 000 70 000
WOM	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	3 983 - 925 4 920 - 1 401 6 189	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	31 983 - 7 925 116 920 - 29 401 76 189
NMS	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	0 - 0 0 - 0 0	4 000 - 1 000 16 000 - 3 773 10 000	3 437 - 768 7 654 - 1 795 3 140	4 000 - 1 000 16 000 - 4 000 10 000	0 - 0 0 - 0 0	4 000 - 1 000 16 000 - 3 932 10 000	23 437 - 5 768 87 654 - 21 500 53 140
BDS	4 000 - 1 000 15 713 - 3 853 10 000	3 990 - 748 5 274 - 1 194 9 097	0 - 0 0 - 0 0	4 000 - 931 16 000 - 2 137 10 000	4 000 - 1 000 16 000 - 3 972 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 6 817 - 4 000 0	4 000 - 1 000 16 000 - 4 000 10 000	27 990 - 6 679 91 805 - 23 157 59 097
GPS	4 000 - 1 000 16 000 - 4 000 10 000	0 - 0 0 - 0 0	0 - 0 0 - 0 0	0 - 0 0 - 0 0	3 715 - 818 4 773 - 2 114 4 383	0 - 0 0 - 0 0	0 - 0 0 - 0 0	0 - 0 0 - 0 0	7 715 - 1 818 20 773 - 6 114 14 383
WAT	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 3 915 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 3 957 10 000	4 000 - 1 000 16 000 - 3 586 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	4 000 - 1 000 16 000 - 4 000 10 000	32 000 - 8 000 128 000 - 31 459 80 000

TABLE VI: Number of extracted pixels in the *boundary* data set for each buffer size. Labeled pixels for each class and also unlabeled pixels are represented.

Class	**Buffer size $2B$ (in meters)**					
	200	400	1 000	2 000	3 000	4 000
CUF	13 210	24 795	54 063	89 637	120 212	145 055
DUF	69 801	129 865	290 381	551 337	793 284	985 294
ICU	37 873	76 091	175 984	345 776	499 108	632 258
RSF	4 413	8 319	20 344	42 039	63 063	77 133
RAP	39 251	73 323	149 778	250 106	329 672	408 112
STC	62 048	119 209	250 463	440 889	583 845	710 629
PRO	13 729	27 975	68 267	124 310	158 918	196 369
SOY	54 631	107 367	243 272	404 260	536 731	667 757
SUN	140 271	262 218	574 634	987 998	1 315 013	1 597 642
COR	139 962	261 293	583 261	1 019 811	1 360 352	1 651 259
RIC	7 952	14 465	32 304	63 066	82 738	95 780
TUB	4 479	10 608	21 697	41 108	57 657	74 043
GRA	151 587	289 454	636 485	1 141 138	1 551 963	1 892 411
ORC	10 512	20 144	46 956	81 462	109 277	133 584
VIN	29 979	56 131	129 244	239 707	323 826	403 441
BLF	334 754	634 454	1 430 734	2 480 683	3 323 765	3 974 349
COF	623 400	1 175 363	2 615 784	4 755 157	6 669 116	8 524 143
NGL	458 962	881 752	1 977 349	3 410 308	4 606 858	5 621 974
WOM	236 179	443 113	944 710	1 542 605	2 040 969	2 511 469
NMS	81 900	155 856	324 391	483 110	618 084	785 524
BDS	8 480	16 246	47 651	69 107	91 400	112 524
GPS	7	7	608	2 887	5 311	5 390
WAT	262 745	507 158	1 170 362	2 177 128	3 098 221	3 910 482
Total	2 786 125	5 295 206	11 788 722	20 743 629	28 339 383	35 116 622
Unlabeled	466 238	887 200	1 966 564	3 427 563	4 639 251	5 710 571

APPENDIX C
QUANTITATIVE RESULTS: CLASSIFICATION DATA SET

Averaged F-score for each class computed over the 11 runs are represented by bar plots. Fig. 12 and Fig. 13 respectively correspond to models trained with the *training* data set DS-A or DS-B on *global* configuration. Fig. 14 and Fig. 15 respectively correspond to models trained with the *training* data set DS-A or DS-B on *stratification* configuration. The class code is provided in Table II.

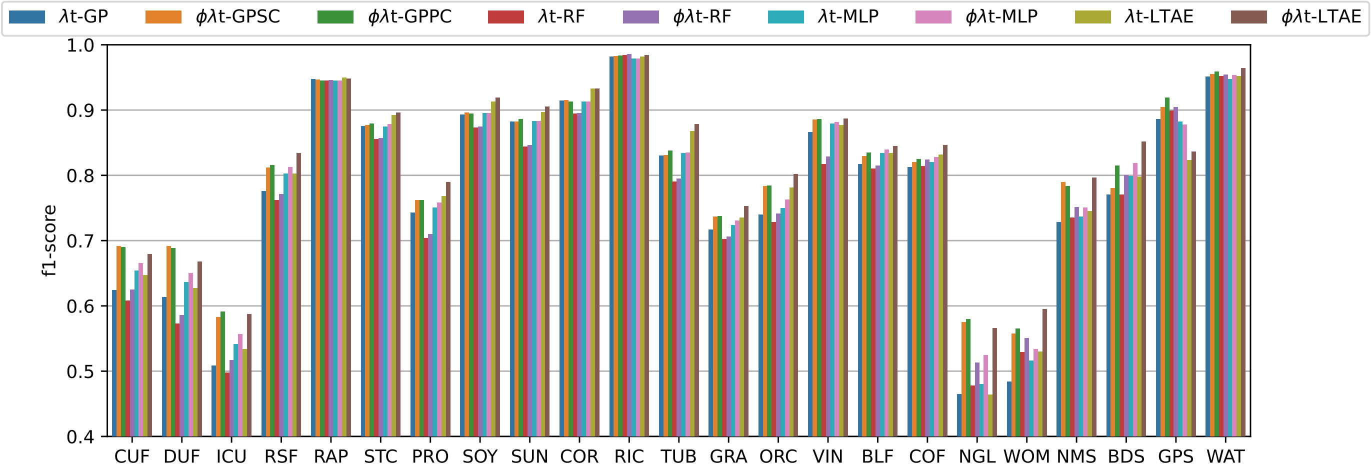


Fig. 12: F-score per class, *global* configuration (data set DS-A)

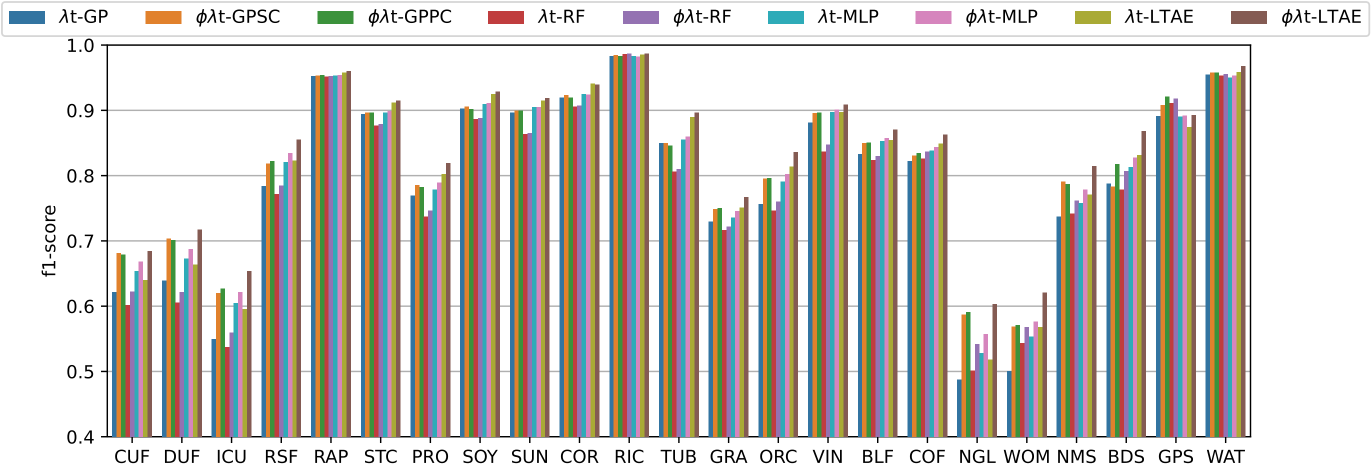


Fig. 13: F-score per class, *global* configuration (data set DS-B)

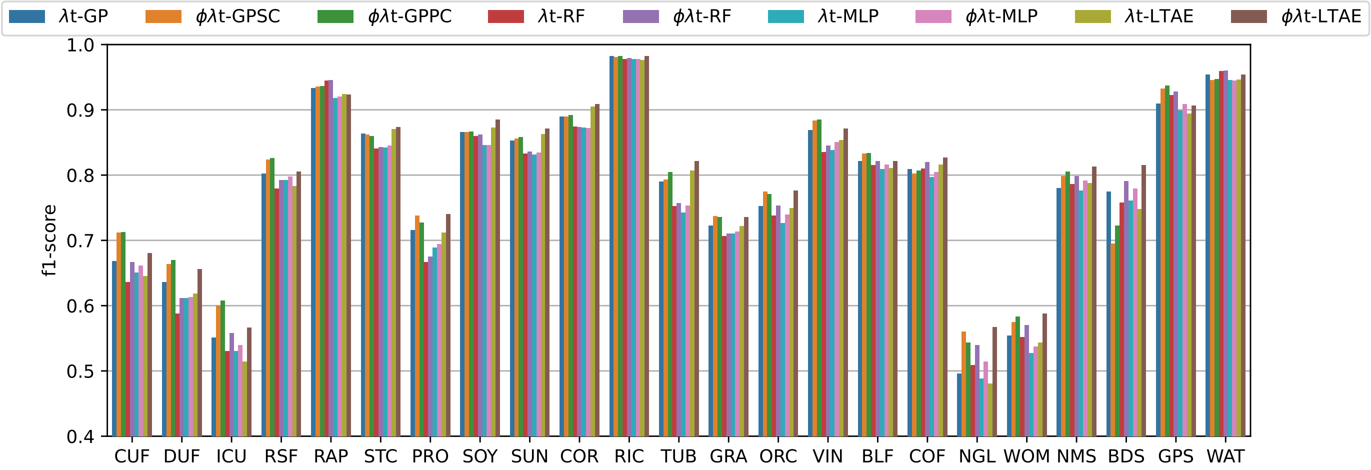


Fig. 14: F-score per class, *stratification* configuration (data set DS-A)

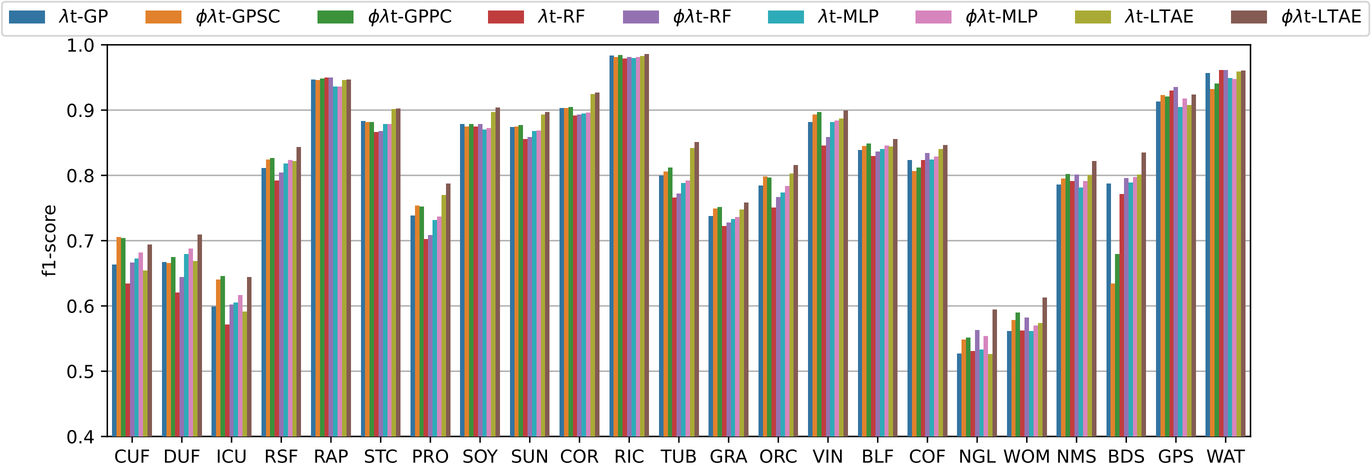


Fig. 15: F-score per class, *stratification* configuration (data set DS-B)

Appendix D
Quantitative results: boundary data set

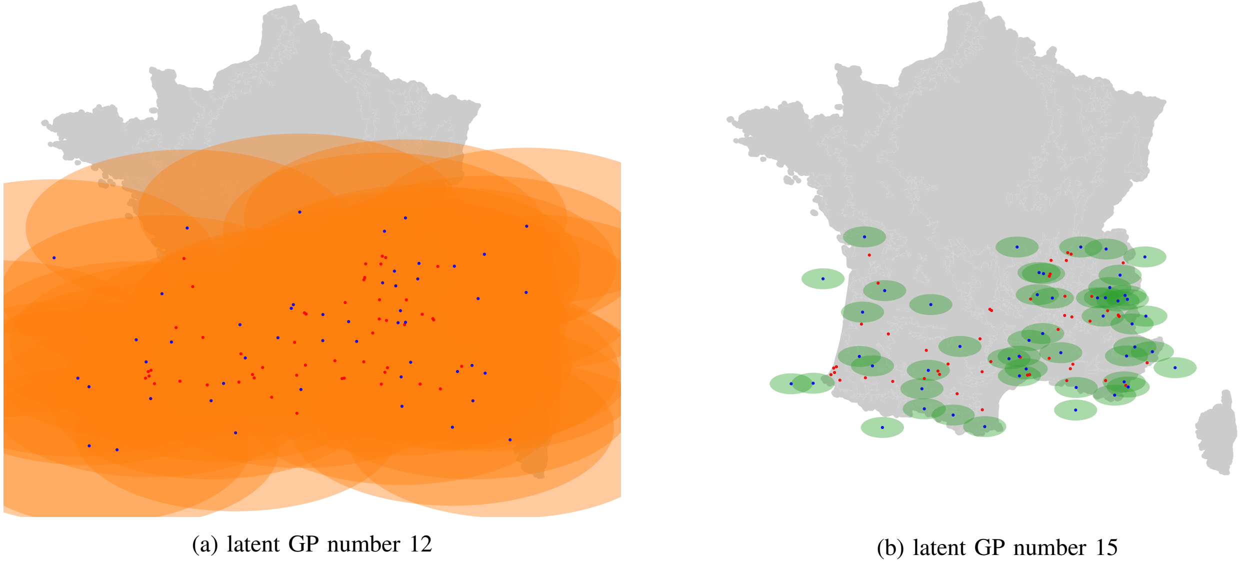
TABLE VII: Averaged percentage of agreement (between two adjacent models) for different sizes of boundary zones ($B \in \{100, 200, 500, 1000\}$) (mean % ± standard deviation computed with 11 runs). Comparison between unlabeled pixels and labeled pixels correctly predicted.

B	Pixels	λt-**GP**	$\phi\lambda t$-**GPSC**	$\phi\lambda t$-**GPPC**	λt-**RF**	$\phi\lambda t$-**RF**	λt-**MLP**	$\phi\lambda t$-**MLP**	λt-**LTAE**	$\phi\lambda t$-**LTAE**
100	unlabeled	66.3 ± 0.7	64.6 ± 1.0	66.2 ± 0.8	72.6 ± 0.5	72.1 ± 0.4	65.2 ± 0.6	64.4 ± 0.6	68.4 ± 0.6	66.0 ± 0.8
	labeled correctly predicted	66.6 ± 0.6	68.5 ± 0.6	69.8 ± 0.6	69.2 ± 0.4	70.5 ± 0.8	64.9 ± 0.4	65.6 ± 0.4	66.4 ± 0.4	68.2 ± 0.5
200	unlabeled	66.2 ± 0.7	64.7 ± 0.9	66.2 ± 0.8	72.6 ± 0.5	72.1 ± 0.3	65.1 ± 0.6	64.4 ± 0.6	68.3 ± 0.6	66.0 ± 0.9
	labeled correctly predicted	66.5 ± 0.6	68.3 ± 0.6	69.5 ± 0.6	69.2 ± 0.4	70.5 ± 0.4	64.9 ± 0.4	65.6 ± 0.4	66.3 ± 0.4	68.1 ± 0.5
500	unlabeled	66.0 ± 0.7	64.5 ± 0.9	66.1 ± 0.8	72.5 ± 0.5	71.8 ± 0.3	65.0 ± 0.5	64.2 ± 0.6	68.2 ± 0.6	65.9 ± 0.8
	labeled correctly predicted	66.6 ± 0.5	68.2 ± 0.5	69.4 ± 0.5	69.3 ± 0.4	70.5 ± 0.3	65.1 ± 0.4	65.8 ± 0.4	66.4 ± 0.4	68.2 ± 0.5
1000	unlabeled	65.8 ± 0.7	64.3 ± 0.9	65.8 ± 0.8	72.3 ± 0.5	71.8 ± 0.3	64.8 ± 0.6	64.0 ± 0.6	68.0 ± 0.6	65.7 ± 0.8
	labeled correctly predicted	66.9 ± 0.5	68.5 ± 0.5	69.7 ± 0.5	69.4 ± 0.4	70.8 ± 0.4	65.4 ± 0.3	66.2 ± 0.3	66.8 ± 0.4	68.6 ± 0.5

TABLE VIII: Averaged overall accuracy (OA) computed on labeled pixels for different sizes of boundary zones ($B \in \{100, 200, 500, 1000\}$) (mean % ± standard deviation computed with 11 runs). Comparison between global configuration and stratification configuration.

B	Pixels	λt-**GP**	$\phi\lambda t$-**GPSC**	$\phi\lambda t$-**GPPC**	λt-**RF**	$\phi\lambda t$-**RF**	λt-**MLP**	$\phi\lambda t$-**MLP**	λt-**LTAE**	$\phi\lambda t$-**LTAE**
100	global	77.1 ± 0.6	79.3 ± 0.7	79.9 ± 0.6	77.7 ± 0.1	78.7 ± 0.4	77.8 ± 0.2	78.8 ± 0.1	78.0 ± 0.4	80.6 ± 0.2
	stratification	74.6 ± 0.4	76.5 ± 0.4	77.3 ± 0.4	75.6 ± 0.2	76.8 ± 0.7	73.1 ± 0.3	74.0 ± 0.2	74.2 ± 0.3	76.2 ± 0.3
200	global	77.0 ± 0.6	79.2 ± 0.6	79.8 ± 0.6	77.6 ± 0.1	78.7 ± 0.1	77.8 ± 0.3	78.7 ± 0.1	78.0 ± 0.4	80.6 ± 0.2
	stratification	74.6 ± 0.4	76.5 ± 0.3	77.2 ± 0.3	75.6 ± 0.2	76.9 ± 0.2	73.2 ± 0.3	74.0 ± 0.2	74.1 ± 0.3	76.2 ± 0.3
500	global	77.3 ± 0.6	79.3 ± 0.7	79.9 ± 0.6	77.7 ± 0.1	78.7 ± 0.1	77.9 ± 0.2	78.9 ± 0.1	78.1 ± 0.3	80.6 ± 0.2
	stratification	74.8 ± 0.3	76.4 ± 0.4	77.2 ± 0.3	75.9 ± 0.2	77.0 ± 0.2	73.6 ± 0.2	74.4 ± 0.2	74.4 ± 0.3	76.4 ± 0.3
1000	global	77.5 ± 0.6	79.6 ± 0.7	80.1 ± 0.6	77.8 ± 0.1	79.0 ± 0.1	78.1 ± 0.2	79.1 ± 0.1	78.3 ± 0.3	80.9 ± 0.2
	stratification	75.4 ± 0.3	77.0 ± 0.4	77.7 ± 0.2	76.2 ± 0.3	77.5 ± 0.2	74.1 ± 0.3	75.0 ± 0.2	74.8 ± 0.3	76.8 ± 0.3

APPENDIX E
INTERPRETATION OF THE LEARNED PARAMETERS: INFLUENCE OF THE SPATIAL INFORMATION


(a) latent GP number 12

(b) latent GP number 15

Fig. 16: Spatial location of inducing points (IP) for 2 different latent GP: ● and ● represent spatio IP respectively before and after optimization. Orange and green ellipses correspond to the spatial area inside which the spatial correlation is greater than 0.9 respectively for the latent GP number 12 and 15. The model $\phi\lambda t$-GPPC was trained on a global configuration.

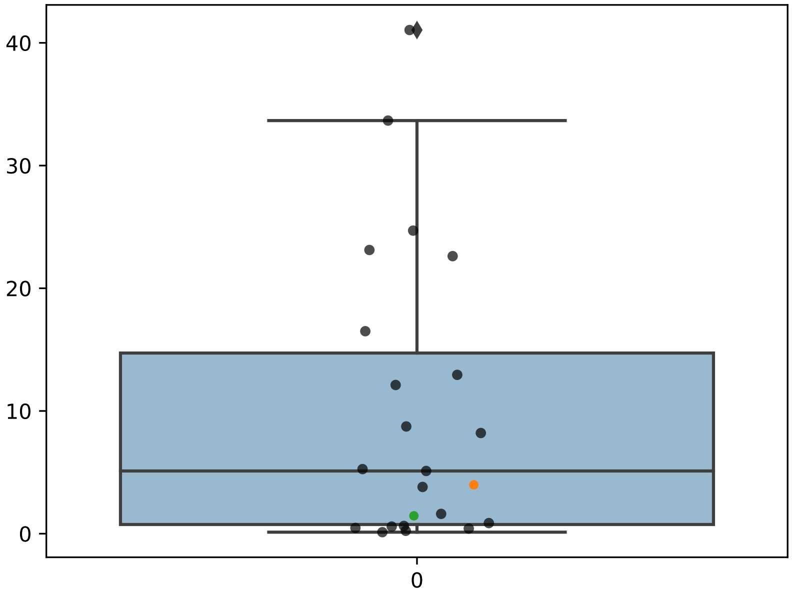


Fig. 17: Distribution of the spatio length-scale ℓ_ϕ for all the latent GP: ● and ● represent the spatio length-scale ℓ_ϕ respectively for the GP latent number 12 and 15.