Remote Sensing of Environment 253 (2021) 112209

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

# Automatic water detection from multidimensional hierarchical clustering for Sentinel-2 images and a comparison with Level 2A processors

Maurício C.R. Cordeiro [a,b,*], Jean-Michel Martinez [b], Santiago Peña-Luque [c]

[a] Agência Nacional de Águas (ANA), Setor Policial Sul, Área 5, Quadra 3, Brasília (DF) 70610-200, Brazil
[b] Géosciences Environnement Toulouse (GET), Unité Mixte de Recherche 5563, IRD/CNRS/Université, Toulouse 31400, France
[c] Centre National d'Etudes Spatiales (CNES), Toulouse 31401, France

ARTICLE INFO

*Keywords:*
Water detection
Water mask
Sentinel-2
Unsupervised clustering
Machine learning
naïve bayes classifier

ABSTRACT

Continuous monitoring of water surfaces is essential for water resource management. This study presents a nonparametric unsupervised automatic algorithm for the identification of inland water pixels from multispectral satellite data using multidimensional clustering and a high-performance subsampling approach for large scenes. Clustering analysis is a technique that is used to identify similar samples in a multidimensional data space. The spectral information and derived indices were used to characterize each scene pixel individually. A machine learning approach with random subsampling and generalization through a Naïve Bayes classifier was also proposed to make the application of complex algorithms to large scenes feasible. Accuracy was evaluated using an independent dataset that provides water bodies in 15 Sentinel-2 images over France acquired in different seasons and that covers a large range of water bodies and water colour types. The validation dataset covers a water surface of more than 1200 km$^2$ (approximately 12 million pixels) including over 80,000 water bodies outlined using a semiautomatic active learning method, which were manually revised. The classification results were compared to the water pixel classification using three of the major Level 2A processors (MAJA, Sen2Cor and FMask) and two of the most common thresholding techniques: Otsu and Canny-edge. An input mask was used to remove coastal waters, clouds, shadows and snow pixels. Water pixels were identified automatically from the clustering process without the need for ancillary or pretrained data. Combinations using up to three water indices (Modified Normalized Difference Water Index-MNDWI, Normalized Difference Water Index-NDWI and Multi-band Water Index-MBWI) and two reflectance bands (B8 and B12) were tested in the algorithm, and the best combination was NDWI-B12. Of all the methods, our method achieved the highest mean kappa score, 0.874, across all tested scenes, with a per-scene kappa ranging from 0.608 to 0.980, and the lowest mean standard deviation of 0.091. Standard Otsu's thresholding had the worst performance due to the lack of a bimodal histogram, and the Canny-edge variation achieved an overall kappa of 0.718 when used with the MNDWI. For water masks provided by generic processors, FMask outperformed MAJA and Sen2Cor and obtained an overall kappa of 0.764. In-depth analysis shows a quick drop in performance for all of the methods in identifying water bodies with a surface area below 0.5 ha, but the proposed approach outperformed the second best method by 34% in this size class.

## 1. Introduction

Continuous monitoring of water bodies is essential for water resources management, but for many countries, especially those with continental extensions, the lack of field monitoring is an issue (Barbosa, 2019). In this regard, earth observations from spaceborne sensors have shown the potential to complement field surveys and provide scientific information in various water-related domains (Bukata, 2013).

Recently, there has been increasing capacity of remote sensing earth observation satellites, such as in the Sentinel missions, which offer high spatial, spectral and temporal resolutions at a global scale without cost. This capacity enables scientists to explore different possibilities, including surface water coverage (Buma et al., 2018; Feyisa et al., 2014; Markert et al., 2018; Pekel et al., 2016; Souza et al., 2019), flood and inundation mapping (Kordelas et al., 2019, 2018; Martinis et al., 2011; Wieland and Martinis, 2019), water quality monitoring (Ansper and

* Corresponding author at: Agência Nacional de Águas (ANA), Setor Policial Sul, Área 5, Quadra 3, Brasília (DF) 70610-200, Brazil.
*E-mail addresses:* mauricio@ana.gov.br (M.C.R. Cordeiro), martinez@ird.fr (J.-M. Martinez), santiago.penaluque@cnes.fr (S. Peña-Luque).

https://doi.org/10.1016/j.rse.2020.112209
Received 19 June 2020; Received in revised form 19 November 2020; Accepted 22 November 2020
Available online 8 December 2020

Alikas, 2018; Delegido et al., 2014; Frampton et al., 2013; Lins et al., 2017; Toming et al., 2016; Yadav et al., 2019), suspended sediment assessment (Condé et al., 2019; Martinez et al., 2009; Yepez et al., 2018), among others.

In this context, the correct identification of water pixels within an image is the first process required for many subsequent applications. The proposed solutions in the literature include a number of approaches that are suitable for different objectives, scales (local, regional or global) and sensor characteristics (Feng et al., 2016; Kordelas et al., 2018; Pekel et al., 2016; Souza et al., 2019).

Supervised (Hollstein et al., 2016) and unsupervised classification (Yousefi et al., 2018), hue saturation and value (HSV) transformation (Dinh Ngoc et al., 2019; Pekel et al., 2016, 2014), spectral mixture analysis (SMA) (Feng et al., 2015; Souza et al., 2019), water index thresholding (Donchyts et al., 2016; Du et al., 2016; Kordelas et al., 2018; Zhang et al., 2018), object segmentation (Kaplan and Avdan, 2017) and, more recently, deep learning (Wieland and Martinis, 2019) are examples of techniques that have been used to separate water bodies from land cover in radar-based and optical images.

Although radar-based water mapping has the advantage that it is not affected by cloud or weather conditions and detects water through a thin canopy, optical images, when available, are more straightforward for detecting water (Shen et al., 2019) and have the advantage of making it possible to use the spectral bands to perform other analyses beyond water mapping using an increased number of parameters (Markert et al., 2018).

Several optical water masking approaches rely on a stack of images acquired on different dates to avoid clouds, shadows or undesirable complex water conditions (Pekel et al., 2016), or they use ancillary data, such as Digital Elevation Models (DEM) or pre-existing cartography (Donchyts et al., 2016; Markert et al., 2018). Although these approaches can achieve high accuracy scores, some of them are complex, time consuming or require human intervention to prepare and process data (Zhang et al., 2018).

Because this work is focused on providing a water pixel baseline for the development of subsequent products within single scenes, especially for the production of water quality products, this article focuses on automated methods for processing optical images without the use of any auxiliary data or time-series mosaicking.

For automatic water detection, the most commonly used methodologies include thresholding in one or more spectral bands or indices and pretrained machine-learning algorithms (Wieland and Martinis, 2019). The literature on supervised machine learning models for water mapping includes commonly used methods, such as support vector machines (Nandi et al., 2017), decision trees (Acharya et al., 2016), random forests (Feng et al., 2015; Ko et al., 2015), multilayer perceptron (Jiang et al., 2018; Mishra and Prasad, 2015) and convolutional neural networks (Pu et al., 2019; Wang et al., 2020).

The use of supervised machine learning models is preferable over thresholding for accurate results, but due to having different water and atmospheric conditions at different sites and their dependency on the training data, supervised machine learning might not be the best option when multiple sites at a global scale are addressed (Bangira et al., 2019).

According to Bangira et al. (2019), varying concentrations of suspended sediments (turbidity), photosynthetic pigments in algae (e.g., chlorophylls, carotenoids), dissolved organic matter and aquatic plants make the implementation of supervised optical remote sensing-based water extraction methods difficult because training data must be universally applicable and frequently updated, especially in the case of water bodies that present highly variable reflectance over space and time.

With regard to unsupervised water mapping from optical images, to date, the majority of developed methods rely on water indices, such as the Normalized Difference Water Index (NDWI) (McFeeters, 1996), Modified Normalized Difference Water Index (MNDWI) (Xu, 2006), Multiband Water Index (MBWI) (Wang et al., 2018) and Automated Water Extraction Index (AWEI) (Feyisa et al., 2014). For automated and operational purposes, these indices are usually employed with an automated thresholding procedure, such as Otsu's thresholding (Otsu, 1979), which minimizes within-class variance in bimodal histograms. This procedure can be combined with other bands or ancillary data in complex workflows to overcome some of the indices' limitations (Dinh Ngoc et al., 2019; Donchyts et al., 2016; Kordelas et al., 2018; Markert et al., 2018; Yang et al., 2018).

The threshold values can be obtained locally or globally, however a unique global threshold is ineffective because it can vary with satellite altitude, illumination, angle, atmospheric conditions (Ji et al., 2009); and water constituents (sediments, organic dissolved matter and chlorophyll). Working over one river catchment, but within boxes of 20 km × 20 km, Donchyts et al. (2016) found that the MNDWI optimal threshold varied from −0.25 to 0.4, which highlights the need to adjust the threshold dynamically over space and time.

In this context, little attention has been paid to large areas, such as an entire Sentinel 2 scene, in which all the complex variabilities of water can lead to erroneous assessments due to a diversity of spectra occurring at the same time. One exception that is worth mentioning is Global Surface Water (Pekel et al., 2016), which was developed to assess water surfaces worldwide and their long-term changes. The equations for its inference system were obtained by manually drawing hulls in a multidimensional feature-space and interactively adjusting them to pixel samples. Ambiguous classification and spectral overlaps were solved by evidential reasoning, followed by a final visual inspection.

Bangira et al. (2019) compared the thresholding method with machine learning classifiers and concluded that the use of only one feature (i.e., one water index) produced relatively poor and unstable results compared to using a basic threshold combination and that more work is needed to investigate the efficacy of other combinations. To overcome some of the drawbacks of using water indices, their combination with other bands in a fully automated method has been an object of study (Dinh Ngoc et al., 2019; Kordelas et al., 2018, 2019), but in a rule-based system, their use in optically complex water remains a challenge (Bangira et al., 2019).

In addition, it appears that certain multidimensional unsupervised methods, such as clustering, have not been as popular as other methods for automatic water mapping based on the number of publications to date. Thus, clustering could be used as a solution for combining multiple features (reflectance bands and water indices) in a single automated procedure.

For this reason, the purpose of this study is to provide a robust method for water extraction from single scenes using optical high-resolution sensors that can be applied at a large scale. Different from other thresholding approaches that use only one dimension, usually a water index, this article aims to analyze the applicability of combining different water indices and spectral bands in unsupervised multidimensional hierarchical clustering, optimized for large-scale processing.

The obtained results are compared to water index thresholding, the most common method to separate water, and other major identification algorithms, such as Sen2Cor (Mueller-Wilm et al., 2019), FMask (Zhu et al., 2015; Zhu and Woodcock, 2012) and MAJA (Hagolle et al., 2010).

## 2. Materials and methods

### 2.1. Reference maps and study area

To test and validate the new water extraction methodology, it is necessary to have a good reference dataset. However, obtaining a good reference dataset can be a challenge for large areas because in-situ reference databases are scarcely available at a broad scale and remote sensing databases require costly, time-intensive investment (Baetens et al., 2019). Although previous studies have been performed to provide global water datasets, such as Global Surface Water data, provided by the European Commission (Pekel et al., 2016), or Global Water Bodies –

GLOWABO data (Verpoorter et al., 2014), our goal is to provide a per-scene pixel classification that is affected by instant climate and water conditions.

To overcome the difficulty of reference mask generation with a limited amount of manual work, Baetens et al. (2019) developed an active machine learning algorithm called Active Learning Cloud Detection (ALCD), which can be used for any classification purpose.

The approach of ALCD consists of an iterative process in which the operator manually selects the reference points, trains the machine learning model based on random forest and classifies the scene. After the classification, the operator visually determines the possible imperfections and labels new pixels where the classification went wrong or was uncertain and then repeats the cycle until a desirable result is achieved. The operator performs a manual correction of persistent errors if necessary.

For the continental water domain, the ALCD algorithm was used to produce the CNES ALCD Open Water Masks dataset (Santiago, 2019), which is available online. This open dataset comprises open water masks for 16 Sentinel-2 scenes over 9 different regions in France that were selected to minimize the presence of clouds while including areas with diverse land coverage (Fig. 1). Some areas have reference masks in different seasons, such as the end of summer and the end of winter periods. The masks were generated at a 10 m resolution, with each tile covering 110 km × 110 km. Among the available scenes, tile T31TGK, from the Alpes region during winter, was discarded because of the unavailability of ancillary data related to this scene that are required to complete the study.

The selected dataset has an inland area coverage of 96.315 km$^2$ with approximately 1.239 km$^2$ of surface water and water quality characteristics that would match most European countries. The complete list of scenes used in this study, with basic characteristics and water coverages and percentages, is described in Table 1.

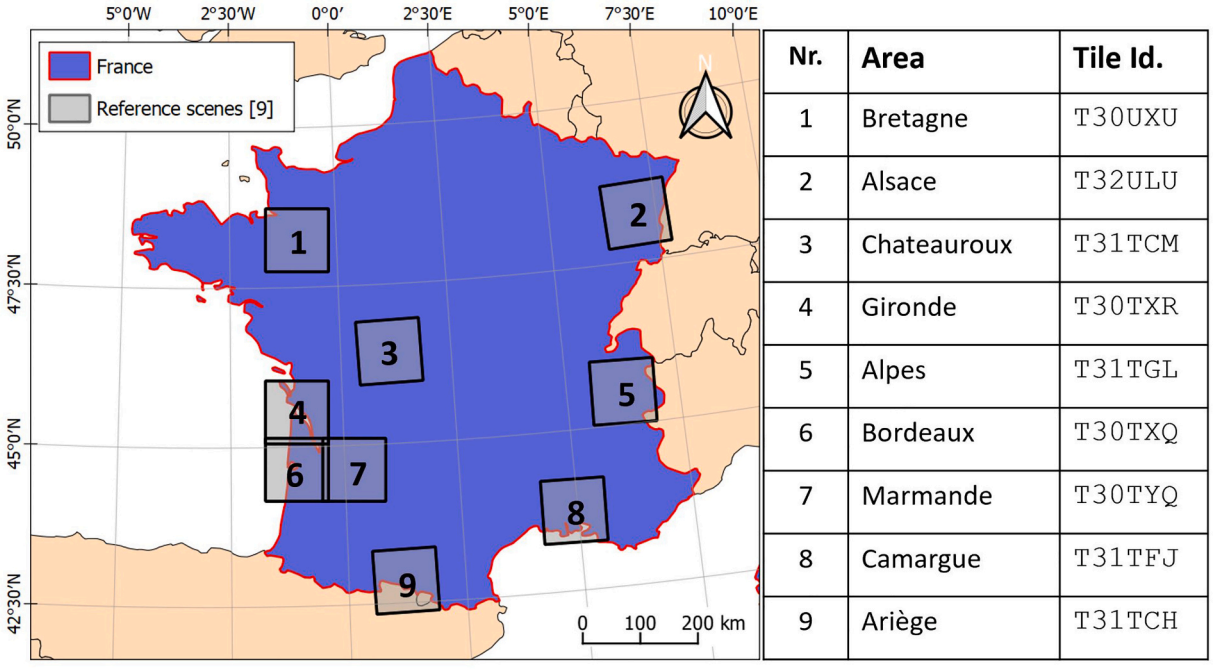
### 2.2. Data collection and preprocessing

The Copernicus Sentinel 2 mission consists of two polar-orbiting satellites (Sentinel 2-A and Sentinel 2-B) that were developed to monitor variability in land surface conditions with a revisit time from 2

**Table 1**
Description of the scenes used for reference with the corresponding continental water surface in km$^2$ and as a percentage.

| Region | Tile | Dates | Continental Water Area km$^2$ (%)[1] | Scene content |
|---|---|---|---|---|
| Alpes | 31TGL | 28 August 2018 | 125.04 (1.03) | Mountains |
| Alsace | 32ULU | 12 September 2018<br>21 March 2019 | 83.55 (0.69)<br>100.12 (0.83) | Lowlands and gentle slopes |
| Ariège | 30TCH | 23 October 2018<br>22 March 2019 | 39.94 (0.34)<br>31.42 (0.26) | Mountains, small water bodies |
| Bordeaux | 30TXQ | 11 September 2018<br>23 February 2019 | 184.40 (2.47)<br>111.36 (2.51) | Large water bodies, very turbid and clear water |
| Bretagne | 30UXU | 08 July 2018<br>23 February 2019 | 50.81 (0.42)<br>54.93 (0.46) | Lowlands, Wetlands, small water bodies |
| Camargue | 31TFJ | 27 September 2018<br>31 March 2019 | 433.83 (4.18)<br>489.10 (4.71) | Large water bodies, floodplain/delta |
| Chateauroux | 31TCM | 19 August 2018<br>25 February 2019 | 133.49 (1.10)<br>121.42 (1.00) | Lowlands |
| Gironde | 30TXR | 23 February 2019 | 92.63 (1.37) | Large water bodies, very turbid and clear water |
| Marmande | 30TYQ | 22 February 2019 | 95.33 (0.79) | Lowlands, small water bodies |

[1] The continental water area and percentage were calculated while accounting for only the valid pixels within each scene.




| Nr. | Area | Tile Id. |
|---|---|---|
| 1 | Bretagne | T30UXU |
| 2 | Alsace | T32ULU |
| 3 | Chateauroux | T31TCM |
| 4 | Gironde | T30TXR |
| 5 | Alpes | T31TGL |
| 6 | Bordeaux | T30TXQ |
| 7 | Marmande | T30TYQ |
| 8 | Camargue | T31TFJ |
| 9 | Ariège | T31TCH |

**Fig. 1.** Spatial distribution of the reference scenes over France used for validation in this work.

to 5 days according to latitude. Both satellites are equipped with an optical Multispectral Instrument (MSI) that can monitor 13 spectral bands with wavelengths ranging from 440 to 2200 nm at high resolution, from 10 to 60 m depending on the spectral band (Suhet, 2015).

Different processors have been developed for Sentinel 2 to correct images for atmospheric conditions and to provide the surface reflectance (Level 2A). One important subproduct from these processes is the pixel classification, which provides Clouds, Shadows, Land, and Water masks among other ancillary data. Sentinel 2 - Level 2A products distributed by ESA are generated by the Sen2Cor processor (Mueller-Wilm et al., 2019), while the French Land Data Center (named Theia) uses the MAJA processor (Gascoin et al., 2019; Hagolle et al., 2010). Another important provider, the United States Geological Survey (USGS), generates the final pixel classification using the Function of Mask (FMask) algorithm (Qiu et al., 2019; Zhu et al., 2015).

To compare the performance of the proposed algorithm with the water mask provided by the FMask, MAJA and Sen2Cor processors, all 15 scenes selected within the reference dataset were downloaded from two different sources. The level-2A MAJA corrected images were downloaded from CNES's Theia Land Data Center portal (https://theia.cnes.fr/). For the FMask mask, we downloaded level-1C images from ESA's Scientific Data Hub, and for the Sen2Cor masks, we downloaded Level-2A products from the same source. Then, version 4.0 of the FMask processor, which is available online (https://github.com/GERSL/Fmask), was used to produce the FMask categorical map.

Considering the importance of the atmospheric correction to better characterize the spectral signature of the desired targets (water and nonwater) and obtain a uniform and consistent dataset (Dinh Ngoc et al., 2019; Jiang et al., 2018; Wang et al., 2018), all of the images used to run the proposed algorithm were Level 2-*A*. *MAJA* was selected, considering the quality of its atmospheric correction and cloud mask in comparison to other processors (Baetens et al., 2019).

To avoid undesired pixels that are not objects of this study and that could bias the results from the algorithm, one external reference mask was produced for each scene. Five categories of pixels were masked using different specialized sources, as shown in Fig. 2. The undesired categories are snow, cloud, cloud shadow, geographic shadow and ocean/coastal waters, and the respective sources for each category are listed below:

- Snow: snow cover was downloaded from the Theia Snow collection (Gascoin et al., 2019), which is an operational snow detector with a 20 m resolution, developed by the CESBIO laboratory;
- Ocean/coastal waters: CNES ALCD Open Water Masks (Santiago, 2019) has a version called "inland masks" that delineates the oceans/coastal waters. This dataset was produced considering the coastal lines available at the Global Self-consistent, Hierarchical, High-resolution Geography Database - GSHHG (https://www.soest.hawaii.edu/pwessel/gshhg/) (Wessel and Smith, 1996) at level H and using an erosion filter of 400 m toward the continent. The shoreline was produced at a working scale of 1:100.000 according to the World Vector Shoreline database;
- Clouds and shadows: For the treatment of clouds, cloud-shadows and geographic shadows, the geophysical – MG2 mask from MAJA was selected because its multitemporal approach has been proven to be superior to other cloud detection processors, as shown in previous studies (Baetens et al., 2019). Although the MG2 layer is provided at 10 m and 20 m resolutions, these masks were computed at 240 m for the Sentinel-2 imagery.

### 2.3. Computational framework and infrastructure

Considering the amount of data and the many experiments required to complete this study, as described in section 3, all the processing was performed on a high-performance cluster within the CNES's computational infrastructure.

The main algorithm was developed in Python 3.7. The clustering, qualification indices and machine learning algorithms were implemented using the SciKit Learning library (Pedregosa et al., 2011), and the Otsu's thresholding algorithm (Otsu, 1979) was from the SciKit Image library (van der Walt et al., 2014). For the geospatial data manipulation, the GDAL library (GDAL/OGR contributors, 2020) was employed.

The comparison task was performed using the validation module of SurfWater software developed by CNES's team, which computes the confusion matrix with more relevant parameters, including Kappa, Precision, Recall, F1 Score, among others.

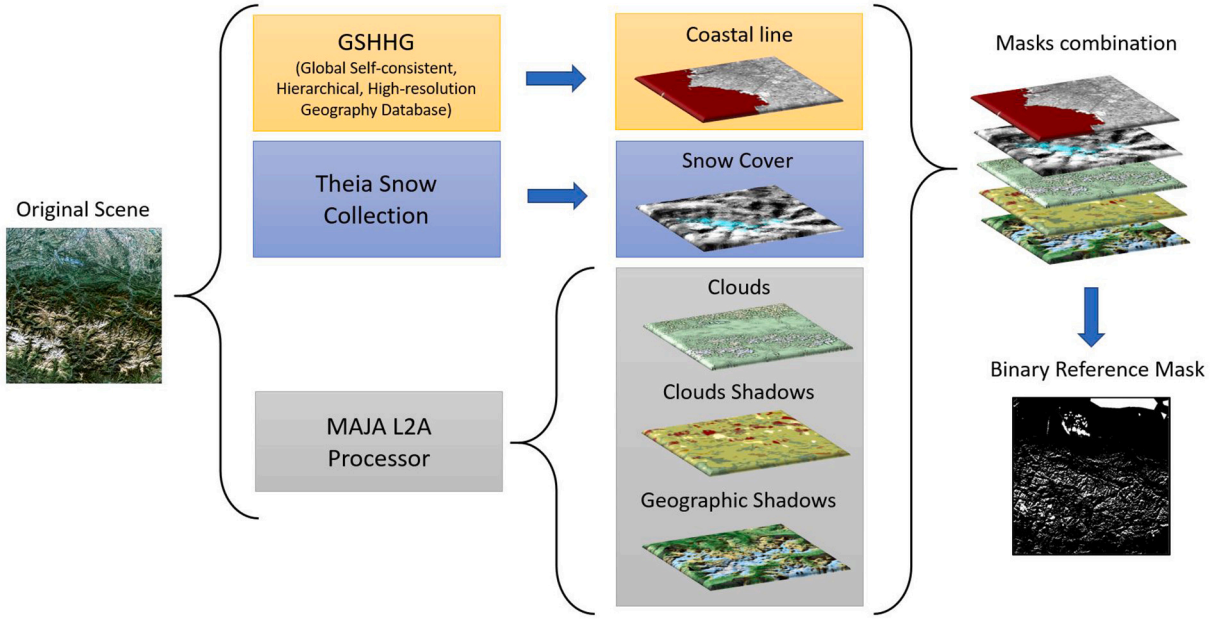



**Fig. 2.** Schematic view of the sources and categories of the pixels selected to produce the reference mask for each scene.

## 3. Methodology

### 3.1. Concept and rationale

Despite their simplicity, simple water extraction methods, such as two-band water indices and single-band thresholding, are not able to accurately distinguish water in complex environments that include build-up areas, dark surfaces or mountain shadows (Feyisa et al., 2014; Wieland and Martinis, 2019). In addition, when addressing large areas (i.e., 110 km × 110 km when considering a full Sentinel-2 scene), it is common to have different environments at the same time. The combination of water indices results with other reflectance bands has been tested in some rule-based water extraction approaches (Dinh Ngoc et al., 2019; Kordelas et al., 2018), but the optimal combination of different thresholding results remains uncertain and often requires tuning to local conditions (Bangira et al., 2019).

In addition, one of the major criticisms of the Otsu thresholding method is the instability that occurs when the histogram of the pixel values is not bimodal (Donchyts et al., 2016). That circumstance can occur in different situations, depending on the relations of the land and water pixels in the selected scene or study area, for example, in areas with very few water pixels compared to land pixels.

The method proposed in this study combines different water indices and reflectance bands into a single unsupervised multidimensional clustering approach for large areas; we compare its results with those from the most commonly used threshold methods and ready-to-use processors. The proposed method is applied on individual Sentinel-2 scenes and is intended to be used as an input water mask for different water resources applications, such as inland water quality monitoring, water temperature assessment, surface water storage or geomorphological analysis of river streams.

### 3.2. Feature selection

An important step in machine learning is called feature engineering, in which features are the input information provided to the algorithm. The primary features considered are the reflectance bands.

Pure water is known to have very strong light absorption for wavelengths above 800 nm, covering the near-infrared (NIR) and short-wave infrared (SWIR) Sentinel-2 bands, compared to the visible bands. This behavior serves as the rationale for the most commonly used water indices (NDWI and MNDWI) (McFeeters, 1996; Xu, 2006). In addition, Bangira et al. (2019) assessed the utilization of each of the Sentinel 2 spectral bands, among other indices, for water discrimination through Otsu-based thresholding, and the single bands that achieved the highest accuracies were B8 (NIR) and B11 (SWIR).

Because Sentinel-2 has two bands in the short-wave infrared spectra and because the B11 band is already used in the MNDWI and the MBWI indices, B12 was selected considering its lowest response over water surfaces. In addition to the selected B8 and B12 Sentinel-2 bands, the most commonly used indices for water segmentation, NDWI and MNDWI, were generated as input features for each scene. Additionally, the recently proposed MBWI (Wang et al., 2018), which was developed to maximize the spectral difference between water and nonwater surfaces in Landsat 8 scenes, especially for confused backgrounds (i.e., mountainous shadows and dark built-up areas), was adapted to work with Sentinel 2 bands and was also included in the current analysis (Table 2).

Lower resolution bands (B11 and B12) were upsampled to 10 m using the simple average algorithm. The two reflectance bands, B8 and B12, and three indices were then combined in different clustering experiments with up to five dimensions when all of the features were considered together. Considering that the order of the features is not important for the clustering method, the number of necessary experiments can be assessed by the combination formula (Eq. (1)):

**Table 2**
Features and calculation methods considered in the clustering experiment.

| Feature | Resolution | Notation | Equation |
|---|---|---|---|
| Visible green (537-582 nm) | 10 m | B3 | – |
| Visible red (646-685 nm) | 10 m | B4 | – |
| Near-infrared (767-908 nm) | 10 m | B8 | – |
| Short-wave infrared (1539–1681) | 20 m | B11 | Average up-sampling |
| Short-wave infrared (2072–2312) | 20 m | B12 | Average up-sampling |
| Modified normalized difference water index | 10 m | MNDWI | $MNDWI = \frac{B3 - B11}{B3 + B11}$ |
| Normalized difference water index | 10 m | NDWI | $NDWI = \frac{B3 - B8}{B3 + B8}$ |
| Multi-band water index | 10 m | MBWI | $MBWI = 3 * B3 - B4 - B8 - B11 - B12$ |

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad (1)$$

where n stands for the number of features and k is the number of dimensions. To test the applicability of B11 instead of B12, we ran extra tests by changing the input band to B11, but the results were systematically less robust than the same experiment with B12.

Additionally, as part of the experiment, for comparison purposes, Otsu-based thresholding and its modified version using the Canny-edge filter were also adopted for each index, achieving a total of 29 experiments for each scene. Table 3 summarizes the water mapping experiments and features used in this study.

### 3.3. Clustering method

One of the most commonly used clustering algorithms in remote sensing applications is k-means because of its high performance, low complexity and the fact that it is implemented in most image packages. The k-means method partitions data into K given clusters, in which each observation belongs to the cluster with the nearest centroid (the mean value in each dimension). This process partitions the data space into Voronoi polygons and results in clusters of similar size.

This characteristic has strong implications when attempting to separate water from land pixels, because depending on the scene, the bands included in the analysis and the number of water pixels compared to land pixels can imply very different cluster sizes.

Agglomerative clustering, which was selected for this study, can address the difference in cluster sizes because it does not have any constraint with regard to the sizes of the resulting clusters. Agglomerative clustering is a subtype of hierarchical clustering that follows a bottom-up approach, in which each observation starts in its own cluster and is then merged iteratively until the desired number of clusters (K) is reached (Nielsen, 2016).

The key parameters for the algorithm to decide whether to merge two clusters are the metric (Eq. (2)) and the linkage (Eq. (3)). The metric specifies the measure of the distance between pairs of observations, and we used the simple Euclidean distance - ED, described in (Eq. (2)):

**Table 3**
Number of water mapping experiments performed in this study for each method.

| Method | Dimensions | Number of experiments |
|---|---|---|
| Otsu Thresholding | 1 | 3 |
| Canny-edge Thresholding | 1 | 3 |
| Clustering | 2 | 10 |
| Clustering | 3 | 10 |
| Clustering | 4 | 5 |
| Clustering | 5 | 1 |

$$ED_{ab} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}, \quad (2)$$

where $ED_{ab}$ is the distance between points $a$ and $b$ in an $n$-dimensional space and $i$ is the considered dimension.

In addition to the metric, the linkage criteria determine how to compute the distance between clusters as a function of the pairwise distances between observations. The Average Linkage - AL, used in this study, considers the mean distance considering all the points in each cluster and can be described as in (Eq. (3)):

$$AL = \frac{1}{|A||B|} \sum_{a \epsilon A} \sum_{b \epsilon B} ED_{ab}, \quad (3)$$

where a and b are the coordinates of points in a n-dimensional space and |A| and |B| are the total number of observations in each cluster.

During each iteration, the algorithm merges the two clusters, among all of the clusters, that are closest to each other considering the criteria described above. The iteration continues until the specified number K is reached.

### 3.4. Selecting the ideal number of clusters

Agglomerative clustering continues to merge clusters until a targeted number of clusters is achieved. Because the final objective is to develop a nonparametric fully automated method, one important step is to identify the best value of K. One way to evaluate the performance of a clustering process when the true classes are not known is to employ a coefficient that measures whether the clusters obtained are dense and well separated.

The coefficient used in this study is the Variance Ratio Criterion, which is also called the Calinsk-Harabasz Index (Caliński and JA, 1974). The Variance Ratio Criterion considers intracluster (Eq. (4)) and intercluster (Eq. (5)) variances, represented by $W_k$ and V, respectively, which are defined as $S_{ch}$ in (Eq. (6)):

$$W_k = \frac{1}{|k|} \sum_{x \epsilon k} ED_{x\mu_k}, \quad (4)$$

$$V = \sum_{k=1}^{K} |k| ED_{\mu_k \mu}, \quad (5)$$

$$S_{CH} = \frac{(N-K)V}{(K-1)\sum_{k=1}^{K} W_k}, \quad (6)$$

where x is a point in cluster k, $|k|$ is the number of samples in cluster k, $\mu_k$ is the centroid of cluster k, $\mu$ is the centroid of all clusters, N is the total number of points, and K is the total number of clusters.

To achieve the best value of K, the Calinsk-Harabasz Index is evaluated for multiple experiments with K ranging from 2 to 10. The superior limit of 10 was defined empirically in the context of this study. More information on the selection of the maximum number of clusters as well as the number of sampled pixels is proved in section 4.4.

### 3.5. Identifying the water cluster

Once the best number of clusters is known, the next step is to identify which cluster, among all possibilities, includes the water points. To achieve this goal, the MBWI is calculated for the centroids of each cluster. The cluster that presents the highest MBWI value is selected as the water cluster, and the others are labeled as nonwater clusters. The MBWI was selected due to its robustness over confused backgrounds, which could result in false positives with the other indices (Wang et al., 2018).

### 3.6. Pixel sampling and generalization

One disadvantage of agglomerative clustering compared to k-means is that it has a space complexity (the amount of memory needed to compute) of $O(n^2)$ and a time complexity of $O(n^3)$ (Firdaus and Uddin, 2015), which make it inefficient for even medium-sized datasets.

In addition, the clustering procedure is performed many times for each different K value until the method can obtain the best number of clusters. Considering that each Sentinel-2 scene includes approximately 120 million pixels (10,980 by 10,980 pixels) and that our objective is to provide an operational method for monitoring large areas, it would not be feasible to apply the process to the whole scene at once.

The proposed method, therefore, randomly selects a subset of pixels in the scene to apply the clustering until the best K is found and the water cluster is identified. Once the best clustering solution is achieved and the subset pixels are labeled for water and nonwater, we generalize the solution to all the pixels in the scene through a supervised machine learning classifier. After several tests, including Naïve Bayes, support vector machines and multilayer perceptron models, Naïve Bayes was selected because it offered the best results within a short training time. The schematic flow-chart of the processing chain is shown in Fig. 3.

Even though most of the processing is performed using a high-performance cluster, as mentioned in section 2.3, a whole Sentinel-2 scene can be classified in 3 to 6 min using an intel i7 computer with 32 Gb of memory running the Windows 10 Pro operating system.

### 3.7. Image thresholding

The most straightforward method for water body mapping is to apply a threshold to an index and select the pixels whose values are greater than a selected threshold value. A dynamically selected threshold is used as a standard because the values can vary both temporally and spatially among different regions, depending on the different image and water characteristics. Among the variety of methods for automating the threshold selection proposed in the literature (Al-Bayati and El-Zaart, 2013; Yang et al., 2017), the Otsu algorithm, developed initially to separate background from foreground in computer vision applications, is one of the most commonly used approaches. This method aims to maximize interclass variance and minimize intraclass variance based on the histogram of a variable. Considering $\boldsymbol{\sigma}$ to be the interclass variance between classes separated by a threshold t (Eqs. (7)–(9)), the algorithm exhaustively searches for the value of t that maximizes $\boldsymbol{\sigma}$, as follows:

$$\sigma^2(t) = P_0(\mu_0 - \mu_T)^2 + P_1(\mu_1 - \mu_T)^2, \quad (7)$$

$$P = \frac{\sum Pixels\ in\ class}{Total\ Pixels}, \quad (8)$$

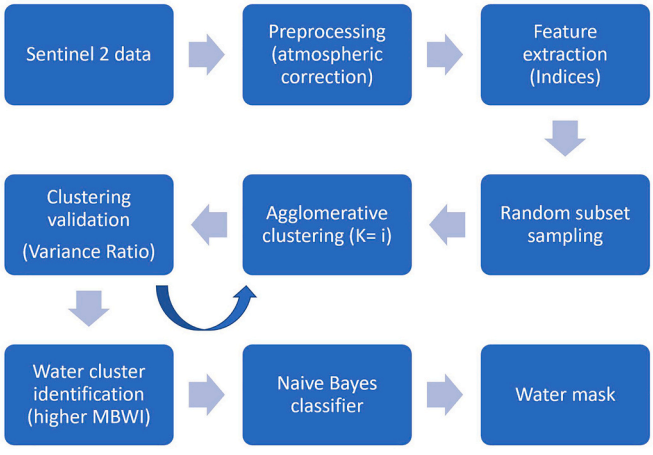



**Fig. 3.** Schematic flow diagram of the proposed method.

$$Threshold = \underset{min(t) \le t \le max(t)}{ArgMax} \left[\sigma^2(t)\right], \quad (9)$$

where P is the probability of a pixel belonging to one of the classes, $\mu_0$ and $\mu_1$ are the mean values of the variable for both classes, $\mu_T$ is the mean value of the whole histogram, and the threshold value is the value of t that maximizes **σ**.

To analyze the proposed multidimensional clustering method, we compare its results with those obtained by the plain Otsu's method applied to the three different indices, MNDWI, NDWI and MBWI. It has been shown, however, that this method yields unstable results if the amounts of water and land pixels are not equivalent and do not form a bimodal histogram (Donchyts et al., 2016).

To overcome these limitations, another set of experiments is performed using the improved Canny-edge method introduced by Donchyts et al. (2016), which first identifies the edges and then applies the Otsu algorithm on the pixels around those edges.

Both methods are implemented in Python, while taking advantage of the algorithm implementations available in the Scikit Image package (van der Walt et al., 2014). The implementation uses a mask for invalid pixels (clouds, no-data, shadows and snow). An iterative process is introduced to select the strongest edges. To provide similar conditions to the clustering technique, the final threshold is limited by preconfigured min/max values, differing, in this regard, from the original implementation.

## 4. Results

### 4.1. Multidimensional clustering

The kappa coefficients obtained from the best-performing combinations of indices/bands are presented in Table 4, as well as the average and standard deviation for all of the scenes. Within this subset, all the combinations achieved a mean kappa (calculated as the average kappa among the scenes) higher than 0.8 for all of the scenes.

There is variation of accuracy for clustering combinations with two, three or four dimensions, and therefore, no clear benefit is obtained by increasing the number of channels. In addition, all of the top combinations, including NDWI, MNDWI or both, corroborate with a previous study (Bangira et al., 2019) regarding the capacity of these indices to correctly separate water and land pixels.

The most successful combination comprises NDWI and the B12 (SWIR) reflectance band; it achieved a mean kappa of 0.874, with a high score of 0.980 in the Bordeaux summer scene and good results in most areas. Even with a considerably high mean kappa and low standard deviation (std), the best combinations failed to produce reliable results in Ariège, especially for the summer scene. In addition, all of the combinations worked comparably well in most scenes, and the major differences appeared on more challenging scenes, such as Ariège or Alpes, which include snow, a very mountainous area with topographic shadows and the lowest percentage of water surface. In such complex cases, the used of a higher-dimension feature space can increase performance, as is the case of the combinations (MNDWI, NDWI, B12, MBWI) and (NDWI, MBWI, B8, B12), which scored the best in Ariège winter (kappa = 0.780) and Ariège summer (kappa = 0.740), respectively. We also noted that the inclusion of the MBWI index improved the results, especially in the presence of very low reflectance waters.

The accuracies among the scenes varied substantially. Fig. 4 presents the mean kappa coefficient and the standard deviation for each area calculated over the best combinations. The water pixel percentage for each scene is shown in the right Y-axis (red line); it was calculated considering all of the unmasked pixels. There appears to be an inverse relationship between accuracy and water percentage. Clearly, the best accuracies were obtained from scenes with a water percentage above 1.5%. For the most challenging scenes (Ariège, Alsace and Alpes), only a few combinations were able to produce accurate results, which can be seen from the low mean and high standard deviation.

Examples of the clustering results, with the combination of NDWI and B12 channels, are shown in Fig. 5, as well as a comparison with the labels from the reference masks. The X-axis represents the reflectance, in $sr^{-1}$, of each pixel in the SWIR band (B12), and the Y-axis represents its NDWI value. The left column graphs show the pixels of the reference masks, where blue represents water pixels and green nonwater pixels. As expected, the water pixels are grouped in the higher NDWI and lower SWIR values due to the higher absorption of water in this wavelength.

The right column graphs show the results of the clustering algorithm for the same scenes. The pixel colors represent different clusters found throughout the clustering process. The best solution is achieved by testing a different fixed number of clusters (K), from two up to a predefined maximum, and selecting the value of K that maximizes the Variance Ratio index. It can be observed that the quantity of clusters varies depending on the overall characteristics of the scene, such as different coverage types, the presence of build-up areas and other features. Once the clusters are well-defined, the algorithm automatically identifies which cluster is the water cluster based on the higher MBWI value of the cluster centroids. The influence of the maximum number of clusters on the overall model accuracy is assessed in *Section 5.1*.

From the examples shown in Fig. 5, it can be observed that the NDWI (Y-axis) discriminates between land and water well, but the threshold value (lower boundary of the water pixels) can vary from almost 0, as in Alsace, to as low as −0.2, in Bordeaux. This finding agrees with those of other studies that found significant thresholding differences among

**Table 4**
Mean and standard deviation values of the kappa coefficients for the six best-performing clustering experiments. Best values per scene are highlighted in Bold.

| Scene | NDWI B12 | MNDWI NDWI B12 | MNDWI NDWI B12 MBWI | MNDWI NDWI MBWI | NDWI MBWI B8 B12 | MNDWI NDWI B8 |
|---|---|---|---|---|---|---|
| Bordeaux Summer | 0.980 | **0.984** | **0.984** | 0.953 | 0.894 | 0.921 |
| Camargue Winter | **0.970** | 0.966 | 0.960 | 0.946 | 0.969 | 0.893 |
| Camargue Summer | **0.977** | 0.969 | 0.912 | 0.912 | 0.976 | 0.924 |
| Bordeaux Winter | 0.925 | 0.968 | 0.879 | 0.957 | **0.969** | 0.935 |
| Chateauroux Summer | 0.869 | 0.832 | 0.936 | 0.821 | **0.942** | 0.854 |
| Alsace Summer | 0.857 | 0.924 | **0.938** | 0.886 | 0.757 | 0.878 |
| Alpes Summer | **0.947** | 0.937 | 0.942 | 0.912 | 0.560 | 0.943 |
| Alsace Winter | **0.917** | 0.906 | 0.777 | 0.768 | 0.905 | 0.868 |
| Chateauroux Winter | 0.887 | 0.843 | **0.901** | 0.876 | 0.753 | 0.733 |
| Bretagne summer | **0.884** | 0.756 | 0.859 | 0.789 | 0.874 | 0.799 |
| Marmande | 0.866 | 0.915 | **0.919** | 0.815 | 0.733 | 0.662 |
| Gironde | 0.811 | 0.846 | 0.798 | 0.736 | 0.778 | **0.848** |
| Bretagne winter | **0.856** | 0.813 | 0.794 | 0.685 | 0.847 | 0.792 |
| Ariège winter | 0.749 | 0.721 | **0.780** | 0.779 | 0.623 | 0.613 |
| Ariège summer | 0.608 | 0.545 | 0.405 | 0.512 | **0.740** | 0.508 |
| Mean | **0.874** | 0.862 | 0.852 | 0.823 | 0.821 | 0.812 |
| Std | **0.091** | 0.111 | 0.132 | 0.112 | 0.120 | 0.122 |

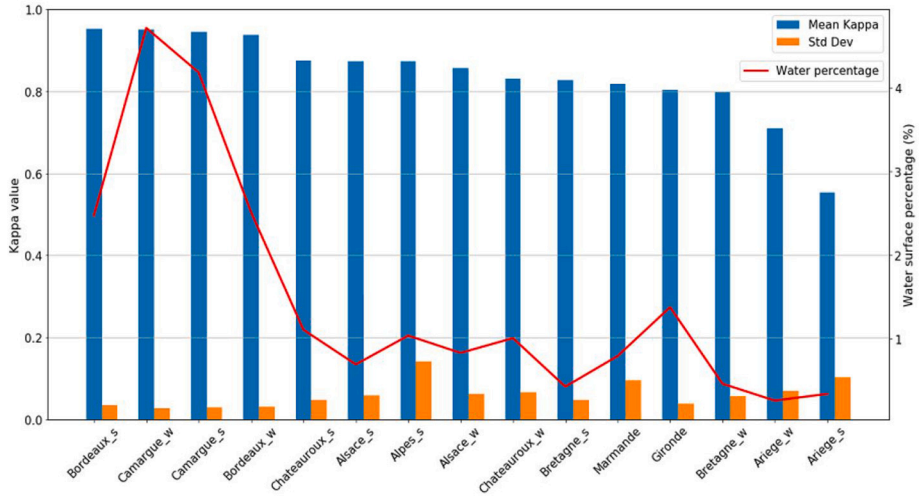



**Fig. 4.** Mean kappa score and standard deviation (left Y-axis) for each scene and comparison with the mean Water Area / Water Perimeter ratio (right Y-axis). In the scenes' labels, "_s" stands for summer, and "_w" stands for winter.

regions (Donchyts et al., 2016), which demonstrates the need for a dynamic method.

The fine separation ability provided by the NDWI, however, is not sufficient to correctly extract water bodies in some scenes. In scenes such as Bordeaux summer and Bretagne winter, for example, a single NDWI threshold would include many nonwater pixels (commission error) or miss many water bodies (omission error). The high variability of the NDWI in the Bordeaux region is due to the presence of water with a high suspended sediment matter (SPM) concentration (i.e., turbid waters), as is the case for the Garonne River. The higher the SPM concentration, the lower the NDWI index due to the increased reflectance response on the near-infrared band (B8).

Another issue is the presence of land pixels with a high NDWI response, which can occur in high albedo or shadow areas (Bangira et al., 2019; Feyisa et al., 2014; Wang et al., 2018). In such cases, the addition of complementing information, such as the SWIR band (B12), improves the results, as can be visually observed from the clustering results (Fig. 5).

### 4.2. Otsu and canny-edge Thresholding

To compare our results with those of previously published methods, we applied several water masking algorithms to the same scenes that we analyzed with the clustering technique. A simple thresholding experiment was conducted using the Otsu's algorithm in each of the three selected indices (MNDWI, NDWI and MBWI) with all of the pixels (no subsampling is necessary) while considering the reference mask to treat the clouds, shadows, snow and coastal waters. The results obtained from Otsu's thresholding algorithm, presented in Table 5, were rather disappointing because the model failed to produce reliable results in most scenes. These results could be explained, as previous studies have already noted (Donchyts et al., 2016), by the fact that in Otsu's algorithm, it is necessary to have a bimodal histogram for the independent variable to be able to produce reliable results.

To overcome this Otsu thresholding issue with a very low percentage of water pixels (not a bimodal histogram), we tested the same scenes using the method proposed by Donchyts et al. (2016), in which a bimodal histogram is obtained by selecting pixels around the contour between water and land through the use of the Canny-edge filter. With this approach, the results were greatly improved, but they are still not reliable for the most challenging scenes (Ariège and Alsace).

The Canny-edge algorithm using the MNDWI performed better in comparison with other thresholding experiments, with a mean kappa of 0.718 when considering all scenes. Plain Otsu's algorithm, without adjustments for the bimodal histogram, is not a feasible option when processing large scenes. A similar problem can be observed with the MBWI index because it fails to provide a good separation between water and land in all of the scenes, regardless of whether the Canny-edge algorithm or Otsu's algorithm is used.

To further compare the different methods and to reduce the errors caused by the small number of water pixels, we tested a modification of the Otsu's algorithm, which is named here Modified Otsu (M-Otsu). M-Otsu takes advantage of using the same sampling method implemented for clustering, with minor adjustments. During the sampling phase, instead of selecting a random sample of pixels, the algorithm is modified to select equivalent amounts of pixels with high and low MNDWIs, defined as MNDWI $<0$ (low) and MNDWI $>0$ (high), and the threshold value is defined using this new set of pixels. The results were much improved, and the accuracy surpassed that of Canny-edge in all of the tested scenes. Table 6 presents the new results, marked as M-Otsu, and compares them with the results obtained by the best clustering and Canny-edge approaches.

With the improved method, Modified Otsu thresholding achieved better results than those for the simple thresholding, but the final mean accuracy of M-Otsu was lower than the clustering, with mean kappas of 0.807 and 0.874, respectively. When comparing the overall behavior of the methods (Fig. 6), the results obtained from the best clustering combinations presented better accuracies than those from thresholding, with a lower standard deviation. In addition, one can note that recall (also known as sensibility) is significantly higher than precision in all of the thresholding methods, especially for Canny-edge. This behavior can be explained because certain types of objects, such as roads, shadows, snow and some built-up areas, can have similar values to water bodies after index calculation due to similarities in their reflectance patterns (Feyisa et al., 2014; Wang et al., 2018; Yang et al., 2018). This characteristic leads to an overestimation of water surfaces with single indices, regardless of the selected threshold. This phenomenon can be more pronounced depending on the percentage of these confusion elements in the scene, and a single index thresholding might not be sufficient to discriminate between water and nonwater surfaces (Bangira et al., 2019; Feyisa et al., 2014).

For some scenes, especially those with larger water surfaces, the results are comparable, with the advantage that thresholding is faster and more straightforward. For more difficult scenes (i.e., a small number of water pixels, small water bodies and increased presence of confusing targets), increasing the number of channels in the clustering approach gives more stable and accurate results, such as the cases in which snow and topographic shadows are presented (mountainous scenes).

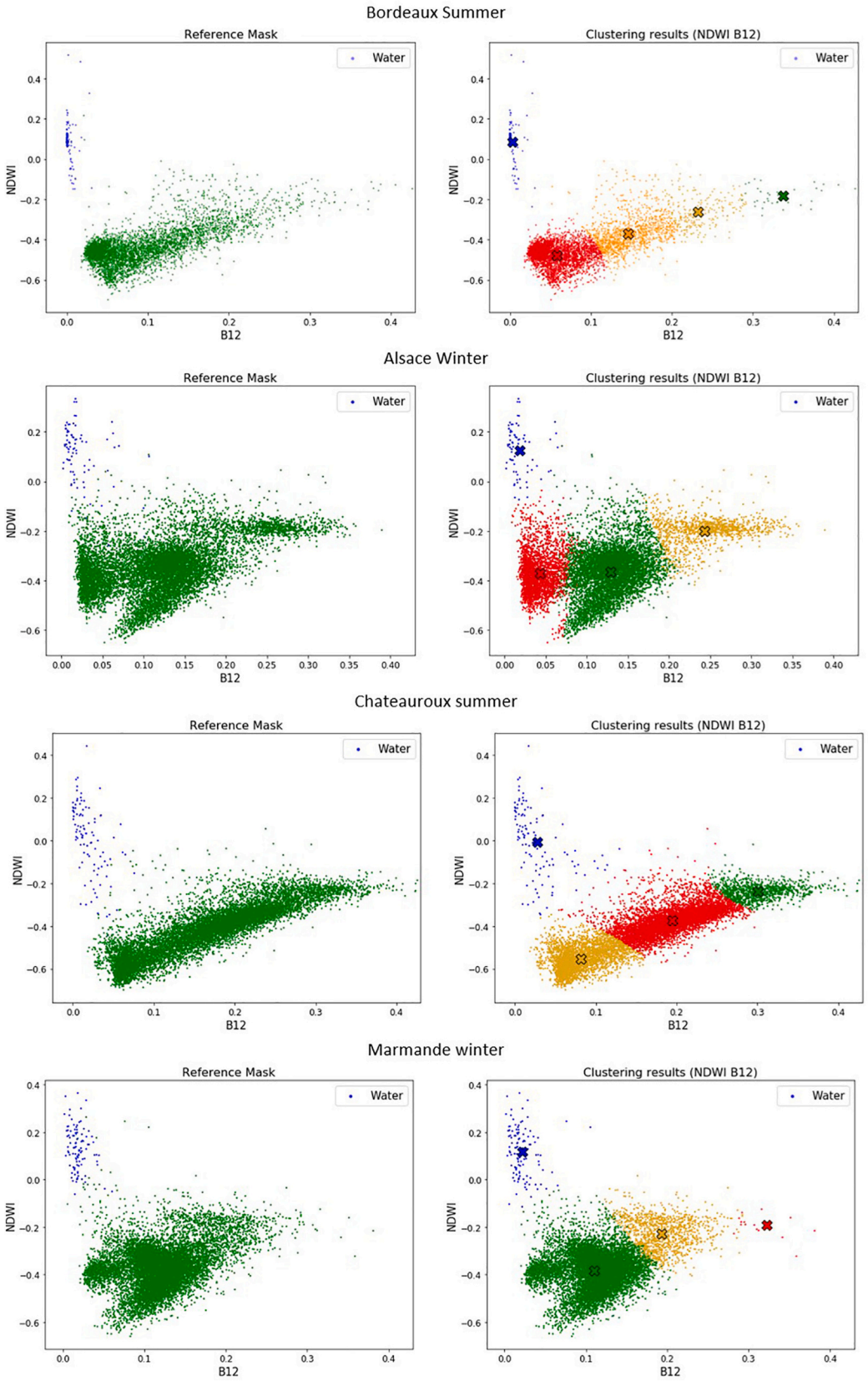



**Fig. 5.** Scatter plots of the NDWI (Y-axis) and B12 reflectance (X-axis) in $sr^{-1}$ for four different scenes. The left column graphs represent the reference masks, in which the water pixels are plotted in blue and the nonwater pixels are in green. In the right column graphs, each colour represents a different cluster, and the colored crosses show the cluster centroid. The cluster identified as the water group is plotted in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 5**
Results (kappa) from Otsu's thresholding and modified Canny-edge thresholding on the MNDWI, NDWI and MBWI indices. Best values per scene are highlited in Bold.

| Scene | Canny MNDWI | Canny NDWI | Otsu MNDWI | Canny MBWI | Otsu NDWI | Otsu MBWI |
|---|---|---|---|---|---|---|
| Bordeaux summer | **0.968** | 0.675 | 0.089 | 0.024 | 0.338 | 0.046 |
| Camargue winter | **0.938** | 0.883 | 0.904 | 0.179 | 0.467 | 0.121 |
| Camargue summer | **0.909** | 0.794 | 0.878 | 0.104 | 0.481 | 0.098 |
| Bordeaux winter | **0.962** | 0.834 | 0.091 | 0.029 | 0.234 | 0.041 |
| Chateauroux summer | **0.829** | 0.535 | 0.045 | 0.504 | 0.017 | 0.023 |
| Alsace summer | **0.808** | 0.364 | 0.032 | 0.035 | 0.023 | 0.011 |
| Alpes summer | **0.503** | 0.163 | 0.061 | 0.043 | 0.085 | 0.030 |
| Alsace winter | 0.515 | **0.661** | 0.043 | 0.054 | 0.026 | 0.024 |
| Chateauroux winter | **0.827** | 0.776 | 0.063 | 0.282 | 0.018 | 0.036 |
| Bretagne summer | **0.724** | 0.291 | 0.016 | 0.491 | 0.017 | 0.008 |
| Marmande | **0.853** | 0.714 | 0.038 | 0.128 | 0.020 | 0.025 |
| Gironde | **0.852** | 0.672 | 0.142 | 0.118 | 0.054 | 0.044 |
| Bretagne winter | **0.702** | 0.362 | 0.014 | 0.131 | 0.009 | 0.019 |
| Ariège winter | 0.136 | **0.177** | 0.023 | 0.009 | 0.004 | 0.009 |
| Ariège summer | **0.245** | 0.113 | 0.016 | 0.007 | 0.018 | 0.008 |
| Mean | **0.718** | 0.534 | 0.164 | 0.143 | 0.121 | 0.036 |
| Std | 0.248 | 0.256 | 0.287 | 0.157 | 0.166 | **0.031** |

### 4.3. Results for standard processors

As already mentioned, one of the objectives of this study is to provide a robust, nonparametric and fully automated process for identifying water pixels in large scenes. In this context, large-scale water masks for Sentinel 2 can be produced using ready-to-use multipurpose classification processors, such as Sen2Cor, MAJA and FMask. Sen2Cor and MAJA have been used as operational ground segments for Sentinel 2 images by ESA and CNES, respectively. In this case, the classified products were downloaded directly from the sites mentioned in section 2.2. FMask classification maps were produced running version 4.0, which is available online.

To compare the results from these processors to the reference masks, the multiclass classification outputs were transformed into a binary classification, with the water pixels being assigned a value of 1 and the nonwater pixels a value of 0. Because these processors are meant to correctly classify clouds, shadows and snow, the only external no-data mask considered was the coastal line, which was obtained from CSSHG with a 400 m dilation toward the continent.

FMask and Sen2Cor compute the classification at a 20-m resolution, while MAJA, although it outputs the EDG layer at a 10-m resolution, produces the water mask at a 240-m resolution. To compare the results, the reference data are resampled to the resolution of the input mask at the moment of the comparison, and no-data pixels are excluded from the statistics. This procedure is performed automatically by the validation module of the surf-water system developed by CNES.

The results presented in Table 7 show a clear advantage for FMask and Sen2Cor over the MAJA processor, with mean kappas of 0.764, 0.726 and 0.355, respectively. These results could be explained by the lower spatial resolution used by MAJA to produce the water masks and the relatively high proportion of small water body areas found in most France scenes. This comparison shows that the clustering technique outperforms the other water masks on average and, more specifically, on the scenes that have the lowest accuracy levels. Among the four scenes with the lowest kappa scores, Ariège summer, Ariège winter, Bretagne summer and Chateauroux summer only Ariège winter processed by Sen2Cor obtained a kappa value higher than 0.6 (kappa = 0.673). All of these scenes share the characteristics of having a small number of water pixels and small water body areas. The same scenes obtained 0.608, 0.780, 0.884 and 0.936, respectively, with the clustering analysis. The

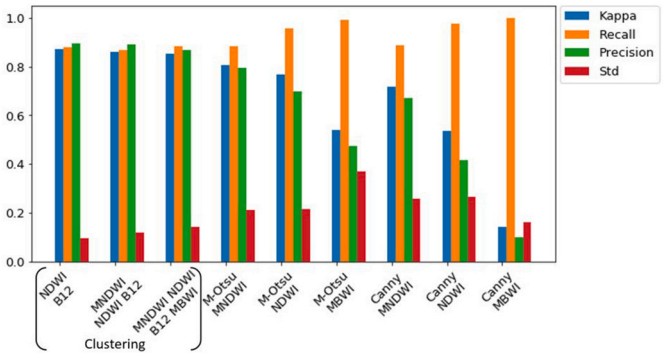



**Fig. 6.** Values of kappa, recall, precision and standard deviation for the clustering method (identified in red squares) considering different input channels, and for other methods tested, calculated for all scenes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 6**
Results (kappa) from Modified Otsu's thresholding and a comparison with Canny-edge thresholding on the MNDWI (Table 5) and the three best clustering combinations (Table 4). Best values per scene are highlighted in Bold.

| Scene | M-Otsu MNDWI | M-Otsu NDWI | M-Otsu MBWI | Canny MNDWI | NDWI B12 | MNDWI NDWI B12 | MNDWI NDWI B12 MBWI |
|---|---|---|---|---|---|---|---|
| Bordeaux summer | **0.984** | 0.967 | 0.982 | 0.968 | 0.980 | **0.984** | **0.984** |
| Camargue winter | **0.971** | 0.958 | 0.897 | 0.938 | 0.970 | 0.966 | 0.960 |
| Camargue summer | 0.966 | 0.947 | 0.888 | 0.909 | **0.977** | 0.969 | 0.912 |
| Bordeaux winter | **0.988** | 0.980 | 0.977 | 0.962 | 0.925 | 0.968 | 0.879 |
| Chateauroux summer | 0.825 | 0.789 | 0.693 | 0.829 | 0.869 | 0.832 | **0.936** |
| Alsace summer | 0.870 | 0.713 | 0.188 | 0.808 | 0.857 | 0.924 | **0.938** |
| Alpes summer | 0.589 | 0.196 | 0.155 | 0.503 | **0.947** | 0.937 | 0.942 |
| Alsace winter | 0.642 | 0.815 | 0.116 | 0.515 | **0.917** | 0.906 | 0.777 |
| Chateauroux winter | 0.821 | 0.838 | 0.508 | 0.827 | 0.887 | 0.843 | **0.901** |
| Bretagne summer | 0.841 | 0.746 | 0.843 | 0.724 | **0.884** | 0.756 | 0.859 |
| Marmande | 0.883 | 0.798 | 0.251 | 0.853 | 0.866 | 0.915 | **0.919** |
| Gironde | **0.963** | 0.905 | 0.847 | 0.852 | 0.811 | 0.846 | 0.798 |
| Bretagne winter | **0.909** | 0.750 | 0.700 | 0.702 | 0.856 | 0.813 | 0.794 |
| Ariège winter | 0.208 | 0.750 | 0.039 | 0.136 | 0.749 | 0.721 | **0.780** |
| Ariège summer | **0.652** | 0.388 | 0.043 | 0.245 | 0.608 | 0.545 | 0.405 |
| Mean | 0.807 | 0.769 | 0.542 | 0.718 | **0.874** | 0.862 | 0.852 |
| Std | 0.203 | 0.209 | 0.356 | 0.248 | **0.094** | 0.111 | 0.132 |

**Table 7**
Validation results (kappa) for the MAJA, FMask and Sen2Cor water masks compared with the results obtained from the three best clustering combinations (Table 4). Best values per scene are highlighted in Bold.

| SCENE | FMask | Sen2Cor | MAJA | NDWI B12 | MNDWI NDWI B12 | MNDWI NDWI B12 MBWI |
|---|---|---|---|---|---|---|
| Bordeaux summer | 0.982 | 0.951 | 0.773 | 0.980 | **0.984** | **0.984** |
| Camargue winter | 0.968 | 0.919 | 0.661 | **0.970** | 0.966 | 0.960 |
| Camargue summer | **0.977** | 0.960 | 0.738 | **0.977** | 0.969 | 0.912 |
| Bordeaux winter | **0.977** | 0.965 | 0.797 | 0.925 | 0.968 | 0.879 |
| Chateauroux summer | 0.573 | 0.329 | 0.069 | 0.869 | 0.832 | **0.936** |
| Alsace summer | 0.908 | 0.815 | 0.123 | 0.857 | 0.924 | **0.938** |
| Alpes summer | 0.672 | 0.886 | 0.704 | **0.947** | 0.937 | 0.942 |
| Alsace winter | 0.903 | 0.752 | 0.280 | **0.917** | 0.906 | 0.777 |
| Chateauroux winter | 0.837 | 0.705 | 0.025 | 0.887 | 0.843 | **0.901** |
| Bretagne summer | 0.512 | 0.352 | 0.124 | **0.884** | 0.756 | 0.859 |
| Marmande | 0.772 | 0.678 | 0.014 | 0.866 | 0.915 | **0.919** |
| Gironde | **0.950** | 0.892 | 0.738 | 0.811 | 0.846 | 0.798 |
| Bretagne winter | 0.751 | 0.657 | 0.042 | **0.856** | 0.813 | 0.794 |
| Ariège winter | 0.443 | 0.673 | 0.149 | 0.749 | 0.721 | **0.780** |
| Ariège summer | 0.239 | 0.352 | 0.088 | **0.608** | 0.545 | 0.405 |
| Mean | 0.764 | 0.726 | 0.355 | **0.874** | 0.862 | 0.852 |
| Std | 0.223 | 0.217 | 0.317 | **0.094** | 0.115 | 0.137 |

largest differences were observed in Ariège summer and Chateauroux summer, which had 72.2% and 63.3% higher accuracies, respectively, with clustering compared to the best processor in each scene.

## 5. Discussion

### 5.1. Overall comparison

A visual comparison among the tested methods in the Camargue summer scene, near Avignon city, is shown in Fig. 7. This area comprises rivers that vary from 50 m to 450 m wide and have dense cities on their margins. We note that the Canny-edge MNDWI (Fig. 7-c) appears flooded compared to the reference mask, with many false-positives, especially in cities and other built-up areas. On the contrary, the three processors (Fig. 7-d, e, f) miss smaller water bodies and narrower river stretches. In this regard, the clustering technique represents a good compromise between precision and recall.

The results show that the overall performance of all of the tested methods is greatly influenced by the scene characteristics, such as the quantity and size of the water bodies and the presence of snow, mountains, shadows and other features. Despite being one of the most commonly used methods, Otsu thresholding achieved the poorest results; it only provided satisfactory results in the Camargue area, which has the highest fraction of water pixels (approximately 4–5%). Consequently, the Otsu thresholding method should be seen as a viable option for smaller study areas, with a well-balanced amount of water and land pixels, but it should not be employed as a standard when large scenes are considered as a whole.

Utilization of the Canny-edge filter to identify the edges before the use of Otsu thresholding can considerably improve results because it surpasses the limitation of nonbimodal histograms, but there are still areas where it fails to identify a threshold, such as the Ariège summer and winter, Alpes summer and Alsace winter scenes, in which the kappa values were 0.245, 0.136, 0.503 and 0.515, respectively. The proposed M-Otsu method achieved kappa values of 0.652, 0.208, 0.589 and 0.642 in the same set of scenes, which shows its clear advantage over Canny-edge in these complex environments. In addition, the mean overall kappa was also higher (0.807) than that for Canny-edge (0.718), with a lower standard deviation (0.203 vs. 0.248).

For the three thresholding methods (M-Otsu, Canny-edge and Otsu), the MNDWI has been shown to produce better results, with greater discrimination of the classes. These results are comparable with those from Xu (2006) and Zhai et al. (2015), who found that the MNDWI performed substantially better than the other indices in extracting water bodies under different conditions, even considering that those studies used bands from Landsat satellites instead of the Sentinel 2 results presented here.

Among the processors, MAJA is the worst performer for water detection, most likely because of the lower spatial resolution of the water mask (i.e., 240 m) and the important areas of small water bodies in the selected scenes. These results contrast with those observed for cloud identification, where MAJA outperformed the other two processors (Baetens et al., 2019). The best overall processor performance was obtained for FMask. It delivered better kappa accuracies than Sen2Cor in twelve scenes, and its performance was only worse for the Alpes and Ariège scenes. To the best of our knowledge, no other study has been conducted comparing the performance of these processors with regard to water pixel extraction.

The results obtained using the proposed automated clustering were overall better than those obtained from the Level-2 processors or from the three thresholding approaches. The best configuration was obtained by combining the NDWI and the B12 band into a single bidimensional clustering. This combination achieved a mean kappa score of 0.874, which surpassed all the processors and thresholding procedures.

Comparing the results of the present study to those of previous studies, Zhai et al. (2015) evaluated their results from water body extraction using different indices based on Landsat-8 and Sentinel-2 imageries at two different sites, a city and village. In the city, the best kappa (0.68) was obtained using the AWEI, and for the village, the best kappa (0.89) was from the MNDWI. In both cases, the thresholds were selected manually. Acharya et al. (2018) evaluated the performances of the indices for water extraction in a 37,127.3 $km^2$ area in eastern Nepal using the Landsat 8 sensor. The area's elevation ranges from 60 m to 8848 m and contains flat lands and mid-hilly regions as well as rugged mountains with snow and glacial lakes. The optimal threshold was searched to provide a higher accuracy. The best kappa score was 0.596, which was obtained from the NDWI. Bangira et al. (2019) also compared the performance of the thresholding method using Sentinel-2 images across eight sites located in South African with a Mediterranean climate and areas ranging from 0.4 to 4.88 $km^2$. Different from the previously mentioned studies, the optimal threshold was obtained automatically using the standard Otsu algorithm. After analyzing more than two hundred different bands in combination, the best solution was achieved with the NDWI band, with kappa scores ranging from 0.78 to 0.92, and a mean kappa of 0.82 considering the eight sites. In the wetlands of Doñana, in Southwest Spain, Kordelas et al. (2018) proposed an automatic thresholding method for inundation mapping, using the MNDWI and SWIR bands, and compared it to a supervised approach. The unsupervised thresholding method resulted in a kappa score of 0.882 for the whole area, with 0.90 for recall and 0.88 for precision.

### 5.2. Influence of the quantity of sampling points and maximum clusters

Considering that the algorithm works initially with a subsample of the entire scene for unsupervised clustering and then generalizes results using a supervised machine learning technique (in this study, Naïve Bayes), the results can be influenced by the number of random points selected initially. In addition, another important parameter for the algorithm is the maximum number of clusters to limit the unsupervised

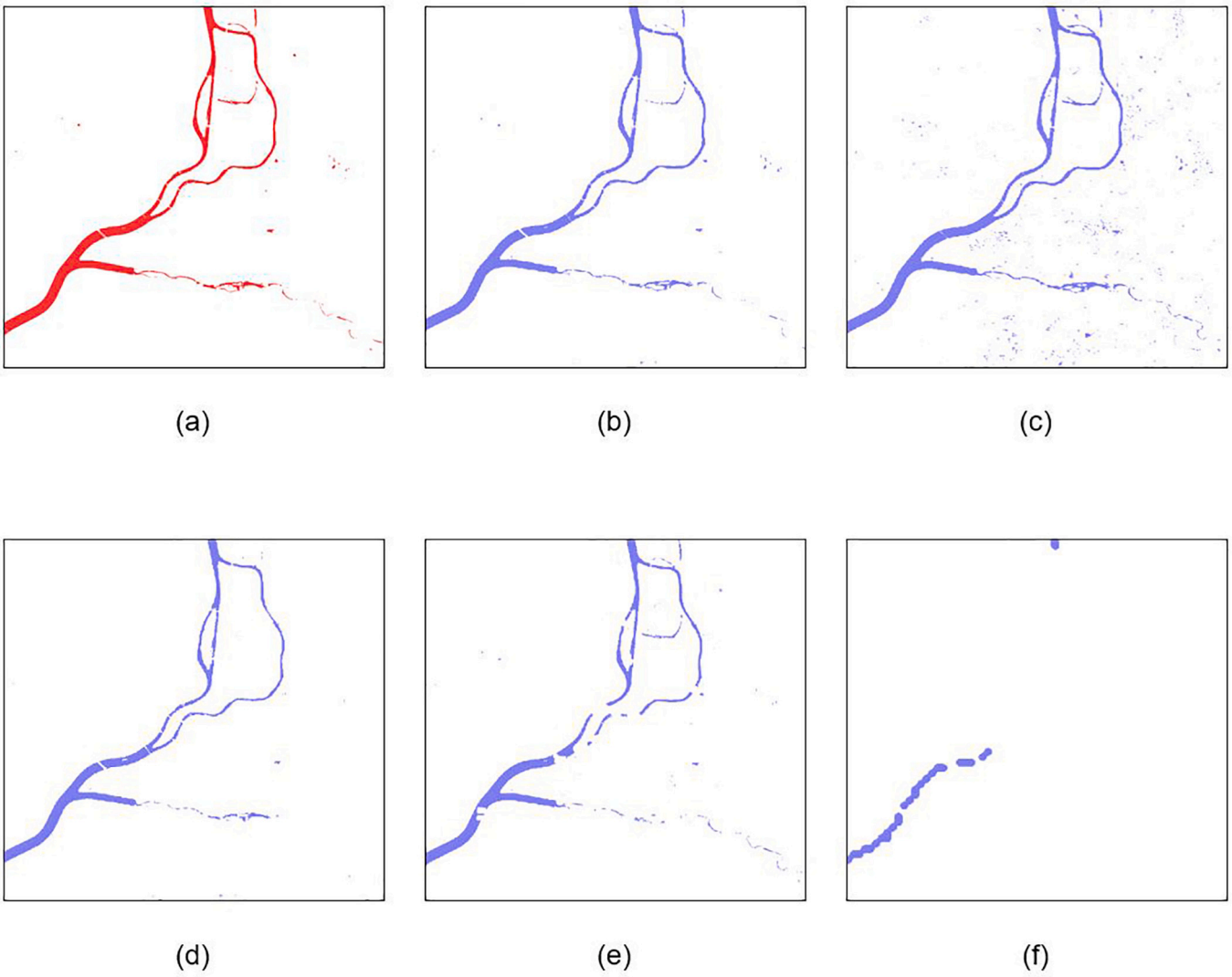



**Fig. 7.** Water segmentation comparison in a Camargue summer scene using the following methods: (a) reference mask, (b) clustering (NDWI x B12) result; (c) Canny-edge MNDWI thresholding; (d) Sen2Cor mask; (e) FMask mask; (f) MAJA mask.

step.

To assess the effects of the number of sampling pixels and the maximum limit on the number of clusters in the overall results, an additional experiment was conducted with the six best clustering combinations.

The mean kappa over all scenes and over the six best clustering combinations was evaluated for the number of sampling points varying from $1 \times 10^3$ to $25 \times 10^3$, with steps of $2.5 \times 10^3$, $5 \times 10^3$, and $10 \times 10^3$. The maximum number of allowed clusters was also fixed at K = 5 and K = 15 for testing purposes. Additionally, the time for processing a single

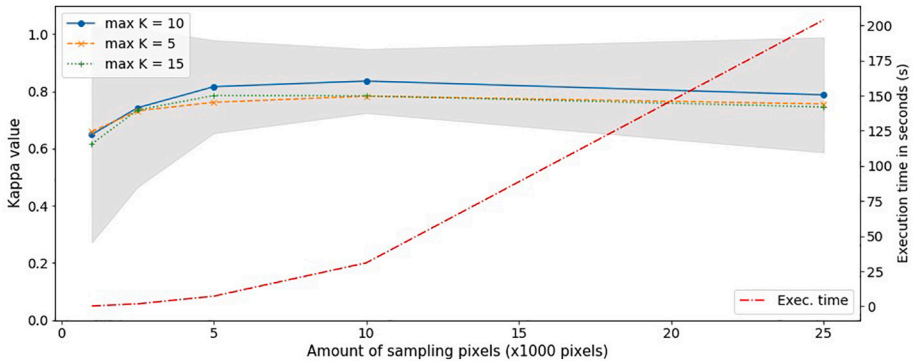



**Fig. 8.** Variation of the mean kappa according to number of sampling pixels and the maximum allowed number of clusters (K). The shaded area represents the standard deviation of the accuracy among the scenes when considering the max K = 10. The red line indicates the time for processing one single scene. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

scene was also measured according to the number of sampling points. The results were obtained running the algorithm on a node of CNES's cluster with 16 cores and 60 Gb of RAM memory. The results are presented in Fig. 8.

The figure shows that increasing the number of sampling pixels does not translate into a better overall accuracy and that the computation cost increases exponentially, as expected from the time complexity of agglomerative clustering, which is $O(n^2)$ (Firdaus and Uddin, 2015). On the other hand, a very small number of pixels is also prejudicial for the model to work because the sampling might not include all of the different pixel signatures presented in the scene. A maximum number of sampling pixels between $5 \times 10^3$ and $10 \times 10^3$ appears to provide a better trade-off between accuracy and computational cost.

With regard to the maximum number of clusters K, we note that there is an optimal number of clusters for the classification accuracy. Not enough clusters mean that the classes can be defined too roughly because each cluster can aggregate pixels with very different spectral behaviors. Too many clusters can lead to the creation of different groups of water pixels, resulting in significant errors because the algorithm looks for one unique water pixel cluster.

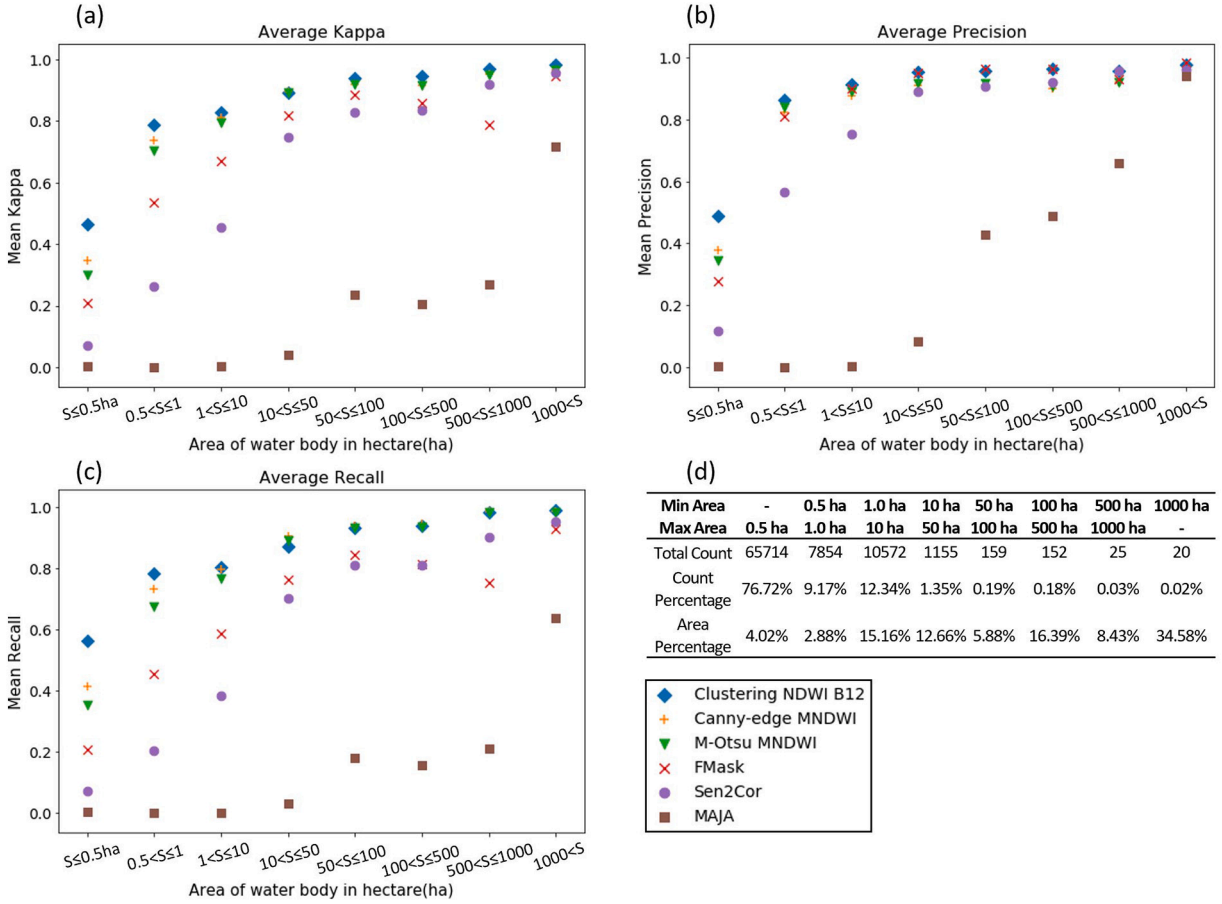
### 5.3. Influence of water body sizes on the clustering performance

To understand how the different methods are compared to each other according to the sizes of the water bodies in the scenes, we stratified the results by eight different classes, 0.0, 0.5, 1, 10, 50, 100, 500 and 1000 ha of water surface.

The water surface was estimated through a polygonization procedure of the reference dataset, and each water body was accounted for separately to make an accuracy assessment. The water bodies from all of the scenes were grouped together in the analysis to obtain an overall mean for the size range across all scenes.

Another important remark is that no further adjustments were made to differentiate rivers from lakes or reservoirs. In such a scenario, it is possible to have a large water surface in a thin water body, which would affect the accuracy of the algorithms by the introduction of mixed pixels, for example.

The results from the stratified analysis are shown in Fig. 9. The mean kappa score is presented in panel (a). Panels (b) and (c) present the average precision and recall indices for the methods, respectively. It can be seen that there is a clear advantage for the clustering approach in every range. The Canny-edge method has the closest performance, followed by M-Otsu, FMask, Sen2Cor and MAJA, in that order. Considering water bodies of sizes between 1 and 10 ha, only clustering (0.826) and Canny-edge (0.809) were able to provide a kappa score above 0.8. In the size ranges from 0.5 to 1 ha and below 0.5 ha, clustering outperformed the second best method, Canny-edge, by 7% and 34%, respectively. All of the parameters that describe accuracy presented lower values for small water bodies, especially for a water surface area beneath 0.5 ha, which represents a 50-pixel area in a Sentinel-2 10-m resolution image. For that class, the clustering technique presented much better results than other techniques, with a mean kappa of 0.47, a mean recall of 0.56 and a mean precision of 0.5. These results are much better than the second-best technique, canny edge based on MNDWI images, with a mean kappa of 0.35, mean recall of 0.42 and mean precision of 0.37. Very small water bodies are especially challenging because all of the




| Min Area / Max Area | - / 0.5 ha | 0.5 ha / 1.0 ha | 1.0 ha / 10 ha | 10 ha / 50 ha | 50 ha / 100 ha | 100 ha / 500 ha | 500 ha / 1000 ha | 1000 ha / - |
|---|---|---|---|---|---|---|---|---|
| Total Count | 65714 | 7854 | 10572 | 1155 | 159 | 152 | 25 | 20 |
| Count Percentage | 76.72% | 9.17% | 12.34% | 1.35% | 0.19% | 0.18% | 0.03% | 0.02% |
| Area Percentage | 4.02% | 2.88% | 15.16% | 12.66% | 5.88% | 16.39% | 8.43% | 34.58% |

**Fig. 9.** Stratified results according to the area of the water body. Average kappa (a), precision (b) and recall (c) for clustering (NDWI B12 combination), Canny-edge (MNDWI), FMask, Sen2Cor and MAJA. Number of water bodies stratified by size and their respective percentages related to the total count and total area, in panel (d).

pixels are located at or near a shore. Consequently, water reflectance is affected more severely, proportionally, by adjacency effects compared with larger water bodies. Other artifact effects can also lead to errors in the mapping accuracy for such small water bodies: landscape shadows projected onto the water surface (terrain, vegetation), lake bottom visibility or reduced water/land edge visibility. It appears that clustering techniques make it possible to soften the impact of these artifacts because this technique provides much better results for the smallest water bodies than any other methods considered in this work. For the 0.5–1 ha water body size class, the clustering technique leads to an accurate mapping assessment (kappa = 0.79, precision = 0.86, recall = 0.78), followed by Canny edge based on the MNDWI and the M-Otsu technique. Considering water bodies of sizes between 1 and 10 ha, only clustering and Canny-edge were able to provide a kappa score above 0.8. For recall (Fig. 9-c), clustering and Canny-edge performed very similar above 0.5 ha, while F-mask, Sen2cor and MAJA missed many water bodies below 50 ha. *MAJA* was the worst performer for identifying water bodies in all of the aspects analyzed due to its lower resolution.

### 5.4. Performance on challenging scenes

To understand the best performance achieved by multidimensional clustering over single band thresholding, Fig. 10 shows scatter plots with a comparison of the pixel separation of water and land pixels for these methods and the reference mask in the Alsace winter scene. The red horizontal lines represent the best thresholding value for the index on the Y-axis, according to the M-Otsu method (the threshold for MNDWI is −0.033, and the threshold for NDWI is −0.044).

It can be observed that regardless of whether the index is used for thresholding, there are nonwater pixels above the threshold (commission errors) and water pixels below the threshold (omission errors), which a single band cannot fully resolve. In such complex scenes, the use of clustering with the addition of more bands can improve the results.

The Ariège winter scene performed poorly with most algorithms and clustering combinations. A careful inspection shows that there is a large difference between the water signature of the water bodies in the mountainous region and the water bodies in the lowland area. The water reflectance in the green wavelength is of the order of 0.01 $sr^{-1}$ in the mountainous lakes, which is mostly due to the very clear water with a low concentration of SPM in this region, whereas it reaches more than 0.06 $sr^{-1}$ in flat areas as a result of higher SPM concentrations. Because the indices used for water detection are mostly a relationship between visible and near or middle infrared wavelengths, this large difference in the reflectance at the green band can result in two or more different water classes for the NDWI or MNDWI. Using the clustering method, the best accuracy for this scene was obtained using four dimensions including combination of the three indices, NDWI, MNDWI and MBWI, and the B12 band, and it achieved a kappa of 0.780. Although the NDWI

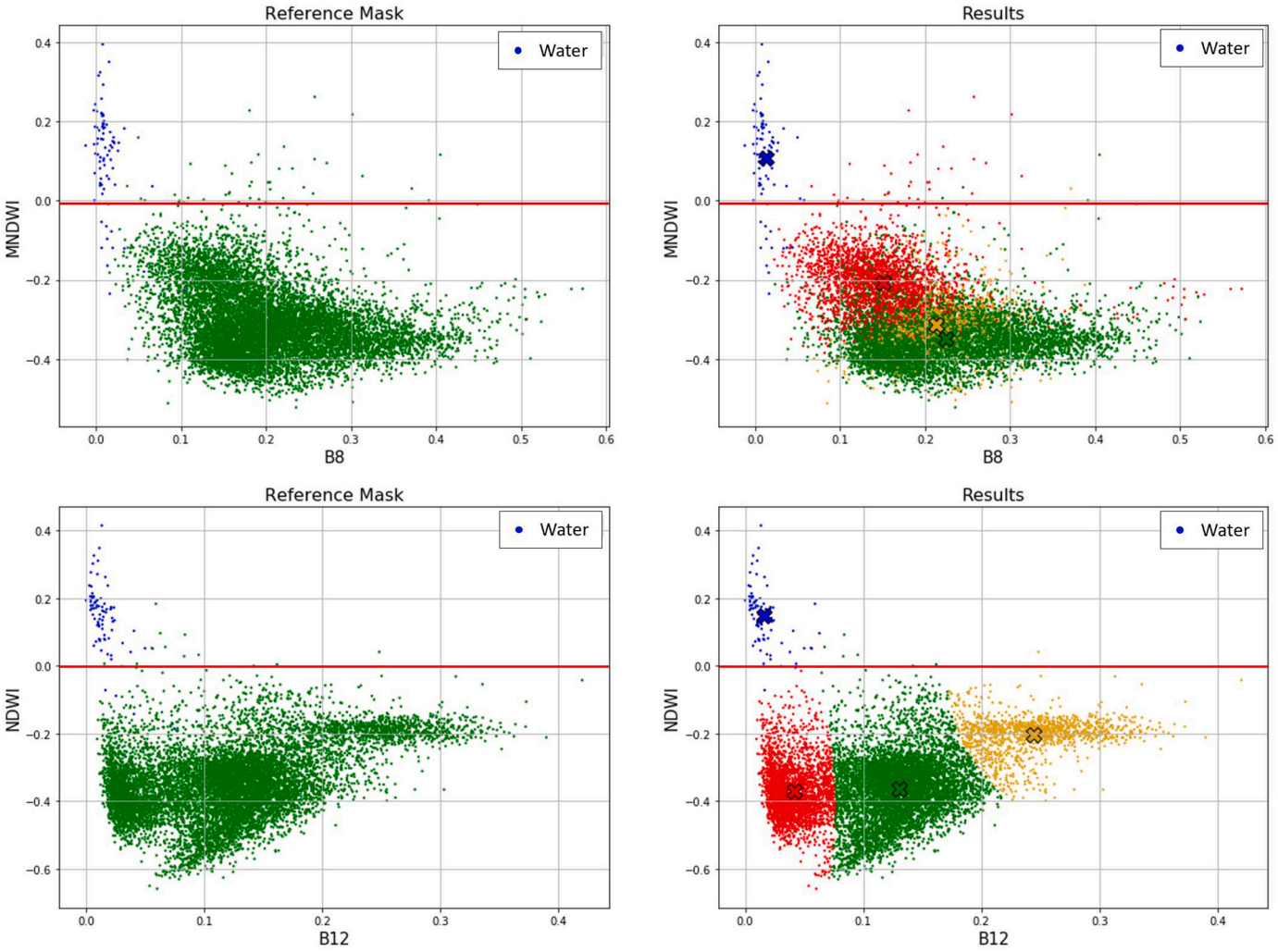



**Fig. 10.** Water segmentation comparison in the Alsace winter scene. The red horizontal lines represent the thresholding solution obtained by the M-Otsu method, and the pixel colors on the right panes are the clustering solution. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

B12 clustering combination can provide good results for most areas, this finding shows that no single combination fits the complexities of every scene and that sometimes it is necessary to use different combinations and even increase the number of clustering dimensions. Moreover, the clustering approach performs better in these mountainous areas and gives more flexibility by allowing the possibility of multiple combinations. A next step would be to develop an adaptative selection technique that can decide the ideal number of clustering dimensions based on some a priori parameters retrieved from the scene itself or from external knowledge. Information on the presence of snow, on solar zenith angle for shadow prediction or on the expected water coverage would allow to automatically assess the best clustering dimension and sampling size.

Although a direct comparison to previous studies is not straightforward, considering the differences in the study areas or satellite data, most unsupervised approaches were applied and validated over smaller regions or depend on subjectively assigned thresholds and ancillary data. In contrast, our method shows robustness (with a kappa standard deviation of 0.094 across all scenes) while considering the diversity of land coverage and environmental conditions (winter and summer scenes) in the present study. Consequently, our method can be easily applied to other regions without adaptation. The clustering technique coupled with a machine learning approach with random subsampling and generalization through a Naïve Bayes classifier presented an efficient running time (approximately 4 min for a Sentinel-2 tile). With regard to its disadvantages, the clustering analysis does not account for any special relationships among water pixels, such as spatial connectivity. Further testing against confusing targets, such as snow, clouds or landscape shadows, could allow us to fully assess the sensitivity of the method to extremely adverse conditions, such as lakes located in mountainous ranges covered by snow or ice or overcast by shadows.

## 6. Conclusions

Continuous monitoring of water surfaces is essential in many applications for water resource management, and the use of satellite images has been increasing with a higher availability of finer spatial and temporal resolution data.

This study proposed and tested a new method for the automation of water pixel classification of Sentinel-2 images that combines reflectance and water indices in a multidimensional clustering approach. Additionally, a machine learning method was proposed to subsample the whole image into a smaller classification set to make the process feasible for an entire Sentinel-2 scene. The method was tested on a set of 15 scenes from France, for which the reference masks are available for download (Santiago, 2019). The reference dataset had diverse coverage, including turbid water and some snow, which matched most European countries but did not consider extreme weather or complex canopy conditions, as in Sahelian areas, tundra plains or tropical forests. The validation process was performed at a 10-m spatial resolution, covering an inland area of more than 96,315 $km^2$, with an approximately 1239 $km^2$ water surface and more than 80,000 water bodies of different sizes and with a high percentage (76%) of water bodies smaller than 0.5 ha. Undesired pixels were removed according to established methodologies, as explained in section 2.2.

Comparing our results to the water classification performance of the three major Level-2 processors, namely, MAJA, Sen2Cor and FMask, the most common thresholding approaches showed that the proposed clustering method achieved higher accuracy results with a lower standard deviation, which indicate the better reliability of the proposed method across different scenes. Stratified analysis showed that the clustering method achieved the best mapping accuracy (i.e., kappa) for all water body size ranges from 0.5 ha to 100 ha. However, there was a quick drop in performance for water bodies that were smaller than 0.5 ha for all of the methods, and there were good results (kappa greater than 0.8) for water bodies above 1 ha with the clustering and thresholding methods. In addition, the newly proposed Modified Otsu thresholding method produced good results with very little complexity and should be further evaluated to test its applicability in a wider range of situations. All of the tested methods, however, presented a lower accuracy rate in complex scenes with mountainous terrains, where the correct water identification remains a challenge, especially without the use of ancillary data.

This study demonstrates that adding more spectral information, as new dimensions, increases the water detection accuracy, especially for smaller water bodies and more challenging environments. From a total of 32 different combinations of spectral indices and bands over 15 scenes (a total of 480 experiments), the combination of the NDWI and the B12 band produced the best overall accuracy across all of the considered scenes, and a four-dimensional combination including NDWI, MNDWI, MBWI, and B12 improved the most difficult scenes.

In summary, the proposed clustering method shows promise in automating water pixel extraction in large scenes from optical images without need for ancillary or pretrained data. Moreover, the subsampling approach leads to a gain in performance and makes it feasible to apply complex algorithms to large amounts of data without a penalty in accuracy. Currently, the use of good input masks for clouds, shadows, snow and coastal water is still critical for current methods to avoid misclassification, and future studies should be conducted to evaluate the methods on images while considering these elements.

The code necessary to run the algorithm is available for download at https://github.com/cordmaur/WaterDetect.

## Funding

This research received no external funding.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The authors would like to acknowledge the technical and financial support from the French Spatial Agency CNES (OBS2CO and SWOT-AVAL projects) for the realization of this study.

## References

GDAL/OGR contributors, 2020. GDAL/OGR Geospatial Data Abstraction Software Library (Open Source Geospatial Foundation).

Acharya, T.D., Lee, D.H., Yang, I.T., Lee, J.K., 2016. Identification of water bodies in a Landsat 8 OLI image using a J48 decision tree. Sensors 16, 1075. https://doi.org/10.3390/s16071075.

Acharya, T.D., Subedi, A., Lee, D.H., 2018. Evaluation of water indices for surface water extraction in a Landsat 8 scene of Nepal. Sensors (Basel) 18. https://doi.org/10.3390/s18082580.

Al-Bayati, M., El-Zaart, A., 2013. Automatic thresholding techniques for SAR images, in: Computer Science & Information Technology. In: Presented at the First International Conference on Computational Science and Engineering, Academy & Industry Research Collaboration Center (AIRCC), pp. 75–84. https://doi.org/10.5121/csit.2013.3308.

Ansper, A., Alikas, K., 2018. Retrieval of chlorophyll-a from Sentinel-2 MSI data for the European Union water framework directive reporting purposes. Remote Sens. 11, 64. https://doi.org/10.3390/rs11010064.

Baetens, L., Desjardins, C., Hagolle, O., 2019. 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 Sens. 11, 433. https://doi.org/10.3390/rs11040433.

Bangira, T., Alfieri, S.M., Menenti, M., van Niekerk, A., 2019. Comparing thresholding with machine learning classifiers for mapping complex water. Remote Sens. 11, 1351. https://doi.org/10.3390/rs11111351.

Barbosa, C.C.F., Novo, E.M.L. de M., Martins, V.S, 2019. Introdução Ao Sensoriamento Remoto De Sistemas Aquáticos: Princípios E Aplicações (Instituto Nacional de Pesquisas Espaciais).

Bukata, R.P., 2013. Retrospection and introspection on remote sensing of inland water quality: Like Déjà Vu All Over Again. J. Great Lakes Res. Remote Sensing Great Lakes Other Inland Waters 39, 2–5. https://doi.org/10.1016/j.jglr.2013.04.001.
Buma, W.G., Lee, S.-I., Seo, J.Y., 2018. Recent surface water extent of Lake Chad from multispectral sensors and grace. Sensors 18, 2082. https://doi.org/10.3390/s18072082.
Caliński, T., Ja, H., 1974. A dendrite method for cluster analysis. Commun. Stat. 3, 1–27. https://doi.org/10.1080/03610927408827101.
Condé, R.C., Martinez, J.-M., Pessotto, M.A., Villar, R., Cochonneau, G., Henry, R., Lopes, W., Nogueira, M., 2019. Indirect assessment of sedimentation in hydropower dams using MODIS remote sensing images. Remote Sens. 11, 314. https://doi.org/10.3390/rs11030314.
Delegido, J., Tenjo, C., Ruiz-Verdú, A., Peña, R., Moreno, J., 2014. Modelo empírico para la determinación de clorofila-a en aguas continentales a partir de los futuros Sentinel-2 y 3. Validación con imágenes HICO. Rev. Teledetec 37. https://doi.org/10.4995/raet.2014.2295.
Dinh Ngoc, D., Loisel, H., Jamet, C., Vantrepotte, V., Duforêt-Gaurier, L., Minh, C., Mangin, A., 2019. Coastal and inland water pixels extraction algorithm (WiPE) from spectral shape analysis and HSV transformation applied to Landsat 8 OLI and Sentinel-2 MSI. Remote Sens. Environ. 223, 18. https://doi.org/10.1016/j.rse.2019.01.024.
Donchyts, G., Schellekens, J., Winsemius, H., Eisemann, E., Van de Giesen, N., 2016. A 30 m resolution surface water mask including estimation of positional and thematic differences using landsat 8, srtm and OpenStreetMap: a case study in the Murray-Darling basin, Australia. Remote Sens. 8, 386. https://doi.org/10.3390/rs8050386.
Du, Y., Zhang, Y., Ling, F., Wang, Q., Li, W., Li, X., 2016. Water bodies' mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the SWIR band. Remote Sens. 8, 354. https://doi.org/10.3390/rs8040354.
Feng, Q., Gong, J., Liu, J., Li, Y., 2015. Flood mapping based on multiple endmember spectral mixture analysis and Random Forest classifier—the case of Yuyao, China. Remote Sens. 7, 12539–12562. https://doi.org/10.3390/rs70912539.
Feng, M., Sexton, J.O., Channan, S., Townshend, J.R., 2016. A global, high-resolution (30-m) inland water body dataset for 2000: first results of a topographic–spectral classification algorithm. Int. J. Digital Earth 9, 113–133. https://doi.org/10.1080/17538947.2015.1026420.
Feyisa, G.L., Meilby, H., Fensholt, R., Proud, S.R., 2014. Automated Water Extraction Index: a new technique for surface water mapping using Landsat imagery. Remote Sens. Environ. 140, 23–35. https://doi.org/10.1016/j.rse.2013.08.029.
Firdaus, S., Uddin, A., 2015. A survey on clustering algorithms and complexity analysis. Int. J. Comp. Sci. Issues 12, 24.
Frampton, W.J., Dash, J., Watmough, G., Milton, E.J., 2013. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. ISPRS J. Photogramm. Remote Sens. 82, 83–92. https://doi.org/10.1016/j.isprsjprs.2013.04.007.
Gascoin, S., Grizonnet, M., Bouchet, M., Salgues, G., Hagolle, O., 2019. Theia Snow collection: high-resolution operational snow cover maps from Sentinel-2 and Landsat-8 data. Earth System Science Data 11, 493–514. https://doi.org/10.5194/essd-11-493-2019.
Hagolle, O., Huc, M., Pascual, D.V., Dedieu, G., 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. Remote Sens. Environ. 114, 1747–1755. https://doi.org/10.1016/j.rse.2010.03.002.
Hollstein, A., Segl, K., Guanter, L., Brell, M., Enesco, M., 2016. Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in Sentinel-2 MSI images. Remote Sens. 8, 666. https://doi.org/10.3390/rs8080666.
Ji, L., Zhang, L., Wylie, B., 2009. Analysis of dynamic thresholds for the Normalized Difference Water Index. PE&RS 75, 1307–1317. https://doi.org/10.14358/PERS.75.11.1307.
Jiang, W., He, G., Long, T., Ni, Y., Liu, H., Peng, Y., Lv, K., Wang, G., 2018. Multilayer perceptron neural network for surface water extraction in Landsat 8 OLI satellite images. Remote Sens. 10, 755. https://doi.org/10.3390/rs10050755.
Kaplan, G., Avdan, U., 2017. Object-based water body extraction model using Sentinel-2 satellite imagery. Eur. J. Remote Sensing 50, 137–143. https://doi.org/10.1080/22797254.2017.1297540.
Ko, B.C., Kim, H.H., Nam, J.Y., 2015. Classification of potential water bodies using Landsat 8 OLI and a combination of two Boosted Random Forest classifiers. Sensors 15, 13763–13777. https://doi.org/10.3390/s150613763.
Kordelas, G.A., Manakos, I., Aragonés, D., Díaz-Delgado, R., Bustamante, J., 2018. Fast and automatic data-driven Thresholding for inundation mapping with Sentinel-2 data. Remote Sens. 10, 910. https://doi.org/10.3390/rs10060910.
Kordelas, G.A., Manakos, I., Lefebvre, G., Poulin, B., 2019. Automatic inundation mapping using Sentinel-2 data applicable to both Camargue and Doñana biosphere reserves. Remote Sens. 11, 2251. https://doi.org/10.3390/rs11192251.
Lins, R.C., Martinez, J.-M., Motta Marques, D.D., Cirilo, J.A., Fragoso, C.R., 2017. Assessment of chlorophyll-a remote sensing algorithms in a productive tropical estuarine-lagoon system. Remote Sens. 9, 516. https://doi.org/10.3390/rs9060516.
Markert, K.N., Chishtie, F., Anderson, E.R., Saah, D., Griffin, R.E., 2018. On the merging of optical and SAR satellite imagery for surface water mapping applications. Results Phys. 9, 275–277. https://doi.org/10.1016/j.rinp.2018.02.054.
Martinez, J.M., Guyot, J.L., Filizola, N., Sondag, F., 2009. Increase in suspended sediment discharge of the Amazon River assessed by monitoring network and satellite data. CATENA 79, 257–264. https://doi.org/10.1016/j.catena.2009.05.011.
Martinis, S., Twele, A., Voigt, S., 2011. Unsupervised extraction of flood-induced backscatter changes in SAR data using Markov image modeling on irregular graphs. IEEE Trans. Geosci. Remote Sens. 49, 251–263. https://doi.org/10.1109/TGRS.2010.2052816.
McFeeters, S.K., 1996. The use of the normalized difference water index (NDWI) in the delineation of open water features. Int. J. Remote Sens. 17, 1425–1432. https://doi.org/10.1080/01431169608948714.
Mishra, K., Prasad, P.R.C., 2015. Automatic extraction of water bodies from Landsat imagery using perceptron model. J. Computat. Environ. Sci. 2015, 1–9. https://doi.org/10.1155/2015/903465.
Mueller-Wilm, U., Devignot, O., Pessiot, L., 2019. Sen2cor Configuration and User Manual.
Nandi, I., Srivastava, P.K., Shah, K., 2017. Floodplain mapping through support vector machine and optical/infrared images from Landsat 8 OLI/TIRS sensors: case study from Varanasi. Water Resour. Manag. 31, 1157–1171. https://doi.org/10.1007/s11269-017-1568-y.
Nielsen, F., 2016. Hierarchical Clustering, in: Introduction to HPC with MPI for Data Science. Springer International Publishing, Cham, pp. 195–211. https://doi.org/10.1007/978-3-319-21903-5_8.
Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybernetics 9, 62–66. https://doi.org/10.1109/TSMC.1979.4310076.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., 2011. Scikit-learn: machine learning in Python. JMLRsentine 12, 2825–2830.
Pekel, J.-F., Vancutsem, C., Bastin, L., Clerici, M., Vanbogaert, E., Bartholomé, E., Defourny, P., 2014. A near real-time water surface detection method based on HSV transformation of MODIS multi-spectral time series data. Remote Sens. Environ. 140, 704–716. https://doi.org/10.1016/j.rse.2013.10.008.
Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422. https://doi.org/10.1038/nature20584.
Pu, F., Ding, C., Chao, Z., Yu, Y., Xu, X., 2019. Water-quality classification of Inland Lakes using Landsat8 images by convolutional neural networks. Remote Sens. 11, 1674. https://doi.org/10.3390/rs11141674.
Qiu, S., Zhu, Z., He, B., 2019. Fmask 4.0: improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. Remote Sensing Environ. 231, 111205. https://doi.org/10.1016/j.rse.2019.05.024.
Santiago, P.L., 2019. CNES ALCD Open water masks (Version 1.1) [Data set]. https://doi.org/10.5281/zenodo.3522069.
Shen, X., Wang, D., Mao, K., Anagnostou, E., Hong, Y., 2019. Inundation extent mapping by synthetic aperture radar: a review. Remote Sens. 11, 879. https://doi.org/10.3390/rs11070879.
Souza, C., Kirchhoff, F., Oliveira, B., Ribeiro, J., Sales, M., 2019. Long-term annual surface water change in the Brazilian Amazon biome: potential links with deforestation, infrastructure development and climate change. Water 11, 566. https://doi.org/10.3390/w11030566.
Suhet, Hoersch B., 2015. Sentinel-2 User Handbook.
Toming, K., Kutser, T., Laas, A., Sepp, M., Paavel, B., Nõges, T., 2016. First experiences in mapping Lake water quality parameters with Sentinel-2 MSI imagery. Remote Sens. 8, 640. https://doi.org/10.3390/rs8080640.
Verpoorter, C., Kutser, T., Seekell, D.A., Tranvik, L.J., 2014. A global inventory of lakes based on high-resolution satellite imagery. Geophys. Res. Lett. 41, 6396–6402. https://doi.org/10.1002/2014GL060641.
van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., 2014. Scikit-image: image processing in Python. PeerJ 2, e453. https://doi.org/10.7717/peerj.453.
Wang, X., Xie, S., Zhang, X., Chen, C., Guo, H., Du, J., Duan, Z., 2018. A robust multi-band water index (MBWI) for automated extraction of surface water from Landsat 8 OLI imagery. Int. J. Appl. Earth Obs. Geoinf. 68, 73–91. https://doi.org/10.1016/j.jag.2018.01.018.
Wang, G., Wu, M., Wei, X., Song, H., 2020. Water identification from high-resolution remote sensing images based on multidimensional densely connected convolutional neural networks. Remote Sens. 12, 795. https://doi.org/10.3390/rs12050795.
Wessel, P., Smith, W.H.F., 1996. A global, self-consistent, hierarchical, high-resolution shoreline database. J. Geophys. Res. 101, 8741–8743. https://doi.org/10.1029/96JB00104.
Wieland, M., Martinis, S., 2019. A modular processing chain for automated flood monitoring from multi-spectral satellite data. Remote Sens. 11, 2330. https://doi.org/10.3390/rs11192330.
Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int. J. Remote Sens. 27, 3025–3033. https://doi.org/10.1080/01431160600589179.
Yadav, S., Yamashiki, Y., Susaki, J., Yamashita, Y., Ishikawa, K., 2019. Chlorophyll estimation of lake water and coastal water using Landsat-8 and Sentinel-2A satellite. Int. Arch. Photogramm. Remote Sens. Spatial. Inf. Sci. XLII-3/W7, 77–82. https://doi.org/10.5194/isprs-archives-XLII-3-W7-77-2019.
Yang, X., Zhao, S., Qin, X., Zhao, N., Liang, L., 2017. Mapping of urban surface water bodies from Sentinel-2 MSI imagery at 10 m resolution via NDWI-based image sharpening. Remote Sens. 9, 596. https://doi.org/10.3390/rs9060596.
Yang, X., Qin, Q., Grussenmeyer, P., Koehl, M., 2018. Urban surface water body detection with suppressed built-up noise based on water indices from Sentinel-2 MSI imagery. Remote Sens. Environ. 219, 259–270. https://doi.org/10.1016/j.rse.2018.09.016.
Yepez, S., Laraque, A., Martinez, J.-M., De Sa, J., Carrera, J.M., Castellanos, B., Gallay, M., Lopez, J.L., 2018. Retrieval of suspended sediment concentrations using Landsat-8 OLI satellite images in the Orinoco River (Venezuela). Comptes Rendus Geoscience, Rivers of the Andes and the Amazon Basin: Deciphering global change

from the hydroclimatic variability in the critical zone 350, 20–30. https://doi.org/10.1016/j.crte.2017.08.004.

Yousefi, P., Jalab, H.A., Ibrahim, R.W., Noor, N.F.M., Ayub, M.N., Gani, A., 2018. Water-body segmentation in satellite imagery applying modified kernel kmeans. Malaysian J. Comput. Sci. 31, 143–154. https://doi.org/10.22452/mjcs.vol31no2.4.

Zhai, K., Wu, X., Qin, Y., Du, P., 2015. Comparison of surface water extraction performances of different classic water indices using OLI and TM imageries in different situations. Geo-spatial Information Science 18, 32–42. https://doi.org/10.1080/10095020.2015.1017911.

Zhang, F., Li, J., Zhang, B., Shen, Q., Ye, H., Wang, S., Lu, Z., 2018. A simple automated dynamic threshold extraction method for the classification of large water bodies from landsat-8 OLI water index images. Int. J. Remote Sens. 39, 3429–3451. https://doi.org/10.1080/01431161.2018.1444292.

Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028.

Zhu, Z., Wang, S., Woodcock, C.E., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. Remote Sens. Environ. 159, 269–277. https://doi.org/10.1016/j.rse.2014.12.014.