Delft University of Technology

MSc Geomatics Thesis Proposal

Comparison of remotely sensed and OpenStreetMap registered water reservoirs

Author:
Maria Moscholaki

Mentors:
Clara García-Sánchez
Balázs Dukai
Gennadii Donchyts
(Deltares)
Christine Rogers
(Deltares)

January 9, 2020

Contents

Preface

The Comparison of remotely sensed and OpenStreetMap registered water reservoirs is performed in partial fulfillment of the requirements for the degree of Master of Science in Geomatics. The research is assigned by Deltares institute for applied research in the fields of water, subsurface and and infrastructure. This document describes the graduation thesis proposal as submitted to the Delft University of Technology. The document starts with providing a short introduction to the problem definition of the defined topic (Chapter 1). Hereafter the related work is presented (Chapter 2), followed by the research objectives and the proposed methodology that is necessary to derive the desired results (Chapter 3,4). Finally the preliminary results and project timeplan are provided (Chapters 5,6).

1 Introduction

Water is one of the most vital elements on earth. It is of high importance for the preservation of all forms of life, humans, animals, and plants [20]. To manage these water resources, accurate maps that provide reliable information on the spatial distribution, interannual and annual changes of surface water are essential [40]. Satellite imagery has been used extensively for water detection purposes to support hydrological and ecological processes but also flood risk analysis, agricultural and industrial usage, food and health safety [29, 21].

Remote sensing techniques are based on the principle of measuring the reflected and emitted radiation from the Earth's surface by using several parts of the Electromagnetic spectrum that is not visible to the human eye, such as near-infrared, shortwave infrared, mid-wave infrared, and thermal. The reflectance of a surface depends on its material, and it varies with the wavelength of the electromagnetic energy, which is what makes it possible to identify Earth's surface features differently by analyzing their spectral reflectance signatures. Water detection is mainly based on its characteristic of significantly lower reflectance in the infrared part of the electromagnetic spectrum compared to other landcover types [15]. The foremost advantage of Remote sensing-based techniques is that they provide an effective way of monitoring the surface of Earth continuously on a global scale. This is due to the ease of data access that is offered freely and openly in different temporal and spatial resolutions [17, 14, 2]. However, there are some factors that affect the final accuracy and lead to the missclassification of water pixels (error of omission and commission). Optical Earth observation imagery is easily affected by cloud obstructions, terrain and cloud shadows, snow, ice and "dark" vegetation as they present similar spectral properties with surface water [41]. Moreover, other limitations lie within the remote sensing methods of extracting waterbodies. More specifically, the use of water indices, which are computed from two or more bands, to separate water from

non water features [15] are challenged by the need for an automated optimal threshold method [23]. On the other hand, supervised (with training samples) and unsupervised (without training samples) classification techniques are based on rules that are not easily formed and possibly not robust enough to be applied on a global scale [15].

Another supplementary source of information for the surface water extraction, are the Digital Elevation Models. DEMs are being used to eliminate the confusion that is created by cloud/terrain shadows [41]. Their most valuable characteristic is that it describes the Earth's surface, by indicating the elevation, i.e the height above/below a certain reference point [16]. The morphology of the terrain is also closely related to the flow path of water from higher to lower areas [15]. More specifically DEM derived drainage networks support the delineation of water reservoir areas by forming a flow accumulation map. DEM errors, however, related to the spatial resolution, the implemented algorithm, and the physical properties of the water features, can propagate into the drainage network models causing the need for error correction in the elevation data [22, 1, 13]. Therefore, DEM processing might lead to unrealistic results due to large depressions and subtle elevation differences that exist locally [6].

OpenStreetMap (OSM) is currently the biggest freely available geodata platform, that has been used in a wide range of Geographic Information Systems (GIS) and applications as an alternative or supplementary with other authoritative datasets [5]. It is based on the collection of Geographic Information gathered and updated by volunteers [4]. These data are provided from sources such as Global Positioning System (GPS) devices, cadastral data, through manual digitizing using medium and high-resolution satellite and aerial imagery or form knowledge about an area [12, 3]. The most significant advantage of this provider is its global coverage and up to date nature. Many studies, however, are questioning the OSM data quality [19], as they are created without any formal qualifications. This is the main reason why the use of these georeferenced data have not been extensively adopted by GIS professionals [30].

The dynamic nature of the water extend both in space and time along with the limitations mentioned above in the various water detection methods and datasets, make it very hard to create an accurate high-resolution waterbody map [44]. To overcome these problems and to be able to extract more accurate and precise water reservoir shapes, the fusion of different datasets is suggested by incorporating Earth observations with OpenStreetMap and DEMs. This way the status of registration of water reservoirs in OSM can be simulated in terms of completeness and positional accuracy by comparing it with the water features derived from the combination of different datasets.

2 Related work

Several surface water extraction methods have been developed in the past to separate water from non water features. McFeeters [25] created the NDWI (Normalised Difference Water Index) based on the difference of water reflectance values in NIR and green bands, followed later by the modified NDWI (MNDWI) of Xu (2006) where NIR was replaced by the SWIR band (short-wave infrared) [33, 43]. Most of the early developed optical Earth observation methods that used different water indices (NDWI, MNDWI, etc.) relied on the use of a single threshold segmentation [41]. In more recent studies Donchyts (et al., 2016) amongst other researchers, focused on the creation of an adaptive thresholding for more automated water detection algorithms [7, 42]. Supervised or unsupervised classification techniques were used as an alternative from Manavalan (et al., 1993) and Ozesmi and Bauer (2002) to create land cover maps in which water features were mapped.

To identify areas with a higher probability of water Renno et al. (2008) created the Height Above the Nearest Drainage (HAND). The HAND is a modified DEM that aims to exclude pixels that are falsely classified as water due to terrain shadows. Gallant and Downling (2003) introduced the Multi-resolution Valley Bottom Flatness (MrVBF) to identify valley areas, i.e topographic features of low gradient depressions. Donchyts (2018) and Huang et al. (2017) integrated in their implementations these DEMs to get more reliable water monitoring results. Satellite imagery has been also used in combination with DEM datasets for the delineation of watershed areas in flat terrains by Li et al. (2019).

Cloud obstructions are another significant problem when analyzing satellite imagery. Donchyts (2016) [8] and Hansen et al.(2013) proposed the creation of cloudless composite images that are based on average cloud-free reflectance values [41]. Another approach was introduced by Donchyts (2018), who uses multiple cloud-free images and a probability density function to accurately detect large-sized water reservoirs that present only small changes in their shape. The view angle of the satellite, and the position of the sun, have been used by Zhu and Woodcock (2014) for cloud shadow and snow detection. This technique was adapted by Tan et al. (2013) in combination with a DEM for terrain shadow detection.

Multi-temporal images are equally important with the analysis of higher resolution imagery for more accurate water body mapping. Mueller, N. et al. implemented an algorithm to map the surface water extend across Australia, (2016) by analysing 25 years of Landsat imagery using a decision tree classifier and logistic regression that compares the water classification results with ancillary datasets. This way it was possible to identify the areas where the occurrence of water is more persistent (e.g reservoirs) and where more temporal (e.g floodplains). Yamazaki et al. (2015) created a global 90 m resolution water body map from multi-temporal Landsat satellite images. Feng and Bai (2019)

created a global land cover map produced through integrating multi-source instead of mulitemporal, satellite imagery datasets [11].

OSM geographic information has assisted several mapping procedures. Yang et al. (2017) [45] combined OSM and Earth observations for landuse classification. The OSM data offering global coverage and up to date information were used as training samples for he classification procedure, making this way posssible the creation of a land use map at a large/regional scale. Integrating Openstreetmap Data in Object Based Landcover and Landuse Classification for Disaster Recovery was also proposed by Kato and Vedasto (2018) [19].

As OSM data have been criticized about their inherent variable quality amongst locations, several studies have put them to test to quantify the differences with authorative datasets. Brovelli et al. (2017) developed an automated comparison algorithm of OSM and authorative road datasets, and later in 2018 together with Molinari et al. Brovelli et al. (2018) performed also a map matching and similarity check analysis for buildings to estimate the completeness of building registrations in the OSM database [5]. Bhattacharya (2012) [4] attempted to find similarities and dissimilarities between the OSM and Dutch topographic map Top10 NL. The quality assessment and object matching was performed in terms of the positional accuracy and the shape of OSM polygons (e.g water polygons) however without evaluating the geometry of the boundary curves. Following the same mindset, Fairbairn et al. (2013) assessed the positional and shape quality (geometric similarity analysis) of OSM and other large scale data with ultimate goal to evaluate if the integration of these type of datasets is feasible in terms of accuracy and precision.

Jakovljevic et al. (2019) performed a waterbody mapping comparison of remotely sensed and GIS open data sources. As another application of waterbody mapping, Donchyts et al. (2016) produced a 30 m resolution surface water mask by using Landsat satellite imagery, Shuttle Radar Topography Mission (SRTM) DEM and OSM data. As a result of his comparison it was stated that 50% of the OSM linear water features agreed with the water extend extracted from Landsat 8 and the drainage network created by using the SRTM.

3 Research objectives

3.1 Problem definition

The prevailing majority of the methods developed in the studies mentioned above, try to either implement some type of waterbody mapping technique or to assess the completeness and shape fidelity of OSM features. However, the point of discussion shouldn't be not only that, but also the improvement of the geometry of these features. In the case of water reservoirs specifically, creating a permanent, more accurate waterbody database is of high importance. To do this, the different types of challenges need to be tackled, since even in situ observations, are only point-based and cannot give a representative idea of the spatial distribution of water in time and in large scale.

The amount of Earth observations and other geospatial information is constantly increasing, but instead of using every dataset separately it would be more beneficial to create a new higher level dataset by combining the strengths of various datasets. Therefore, we need a generic solution in higher resolution to get more detailed shapes of water reservoirs. Fusion of high resolution raster and vector datasets might be able to systematize the challenges and provide an automatic geometry specification. Since however all datasets contain uncertainties(e.g Figure 1) we an need objective confidence map of every water mask depending on topographic and other conditions. Combining high resolution EO with alternative sources could possibly gives us an indication of how confident we are that a specific pixel represents water. Consequently, the issue of better reconstruction of the actual surface dynamics e.g in case of floods, but also questions regarding the location and the geometric quality of water reservoirs is addressed, with ultimate goal the attempt to improve the existing database.

Figure 1: Example of OSM inconsistencies

3.2 Research question

The research question defines the scientific goals this study aims to explore and provide an answer to. The chosen general direction of this thesis on which the main focus will be, is the question:

To which extend can Remote Sensing and VGI be combined to accurately identify water reservoir features?

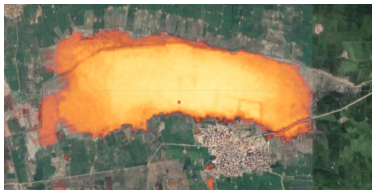
The main research question for this project is composed of the following sub-questions:

- How can multi-source datasets be combined to derive an optimal extend of water reservoirs?
- How complete is the registration of water reservoirs in the OSM database ?
- How valid are the OSM water vector data in terms of location and geometry?

3.3 Use cases

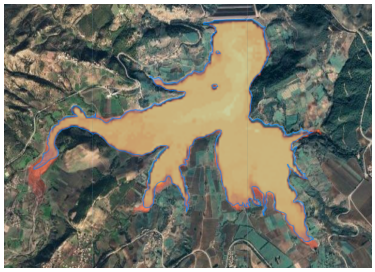
As a study area 10 distictive cases where choosen in total, which are considered sufficiently representative of the variability/diversity of errors in the estimation of the surface water extend (Figures:2,3,4).

(a) Laayoune, Algeria



(b) Changro Dandh, Pakistan

(c) Nai Gaj, Pakistan

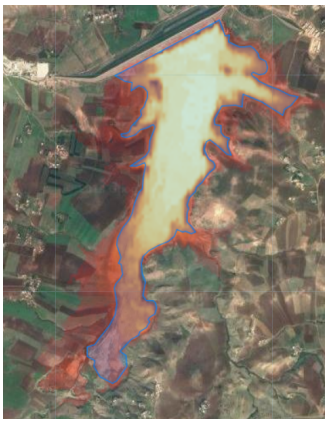


(d) Sidi, Algeria

Figure 2: Use Cases

(a) Hamal Lake, Pakistan

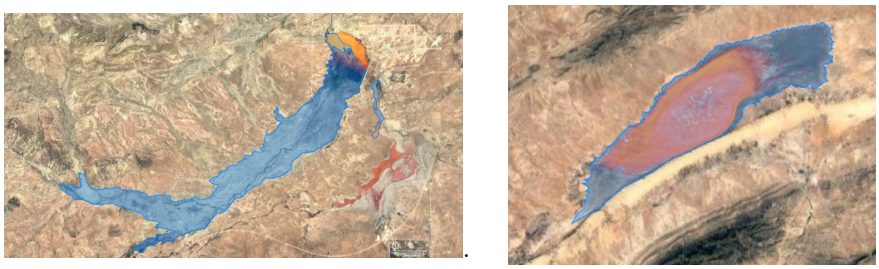
(b) Zeddine, Algeria



(c) Barrage Harraza, Algeria

(d) Lake Volta, Gana

Figure 3: Use Cases



(a) Barrage Boughzoul, Algeria

(b) Zehrez Chergui, Algeria

Figure 4: Use Cases

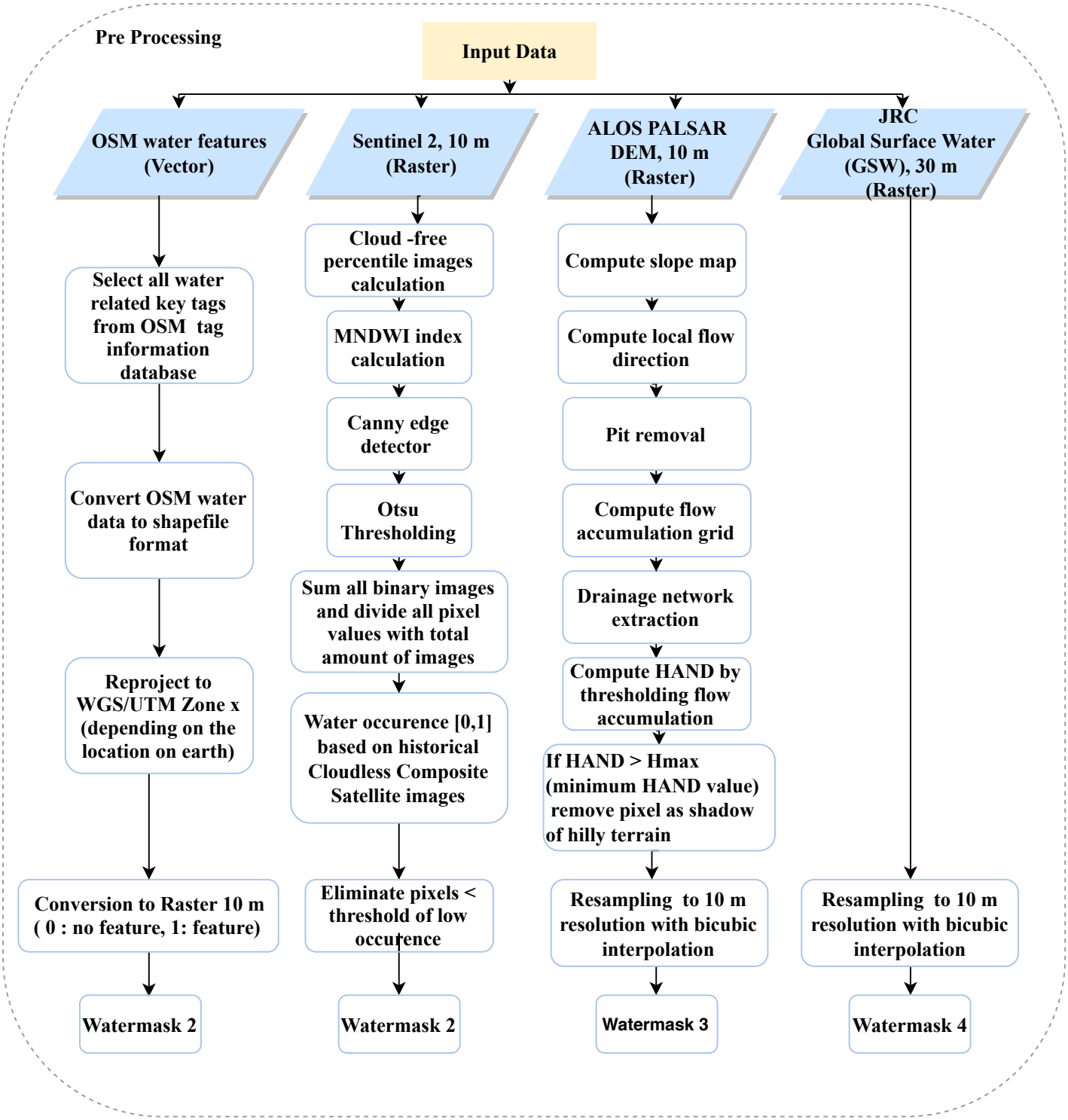
The selected use cases show the difference between the estimated JRC Global Surface Water occurence map based on 35 years of Landsat imagery and OSM surface water extend. Subfigures 2.a,c show presence of water reservoirs in the OSM database which were not detected with the JRC water occurence, whereas subfigure 2.b is not registered as an OSM water body although water was detected in a big extend with high frequency. The rest of the use cases show the discrepancies between the two datasets. With the combination of all the OSM, JRC, Sentinel 2 and ALOS PALSAR datasets to overcome false positive and false negative water classification results.

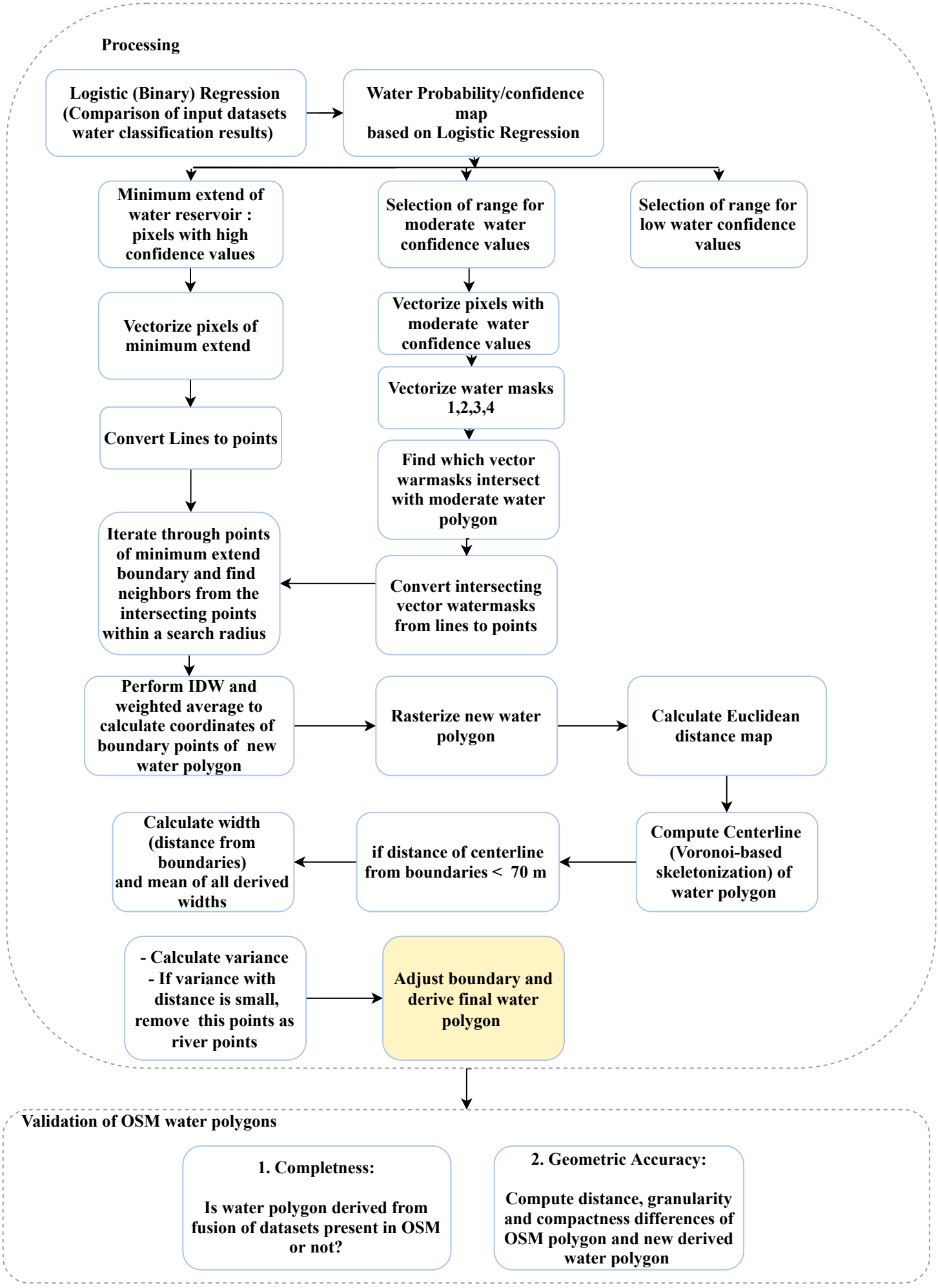
4 Methodology

The approach of the methodology combines different input datasets and analysis techniques to provide a water map for all uses cases by estimating the confidence level for the results. The methodology was based on the assumption that the accuracy of the chosen datasets is similar.

4.1 Workflow

The following workflow describes the steps that form the suggested methodology to create more accurate and precise waterbody estimates.


Pre Processing
Input Data
OSM water features (Vector)
Sentinel 2, 10 m (Raster)
ALOS PALSAR DEM, 10 m (Raster)
JRC Global Surface Water (GSW), 30 m (Raster)
Select all water related key tags from OSM tag information database
Convert OSM water data to shapefile format
Reproject to WGS/UTM Zone x (depending on the location on earth)
Conversion to Raster 10 m (0 : no feature, 1: feature)
Watermask 2
Cloud -free percentile images calculation
MNDWI index calculation
Canny edge detector
Otsu Thresholding
Sum all binary images and divide all pixel values with total amount of images
Water occurence [0,1] based on historical Cloudless Composite Satellite images
Eliminate pixels < threshold of low occurence
Watermask 2
Compute slope map
Compute local flow direction
Pit removal
Compute flow accumulation grid
Drainage network extraction
Compute HAND by thresholding flow accumulation
If HAND > Hmax (minimum HAND value) remove pixel as shadow of hilly terrain
Resampling to 10 m resolution with bicubic interpolation
Watermask 3
Resampling to 10 m resolution with bicubic interpolation
Watermask 4


Processing
Logistic (Binary) Regression (Comparison of input datasets water classification results)
Water Probability/confidence map based on Logistic Regression
Minimum extend of water reservoir : pixels with high confidence values
Selection of range for moderate water confidence values
Selection of range for low water confidence values
Vectorize pixels of minimum extend
Vectorize pixels with moderate water confidence values
Vectorize water masks 1,2,3,4
Convert Lines to points
Find which vector warmasks intersect with moderate water polygon
Iterate through points of minimum extend boundary and find neighbors from the intersecting points within a search radius
Convert intersecting vector watermasks from lines to points
Perform IDW and weighted average to calculate coordinates of boundary points of new water polygon
Rasterize new water polygon
Calculate Euclidean distance map
Calculate width (distance from boundaries) and mean of all derived widths
if distance of centerline from boundaries < 70 m
Compute Centerline (Voronoi-based skeletonization) of water polygon
- Calculate variance
- If variance with distance is small, remove this points as river points
Adjust boundary and derive final water polygon
Validation of OSM water polygons
1. Completness:
Is water polygon derived from fusion of datasets present in OSM or not?
2. Geometric Accuracy:
Compute distance, granularity and compactness differences of OSM polygon and new derived water polygon

4.2 Input Datasets

Dataset	Type	Resolution (m)	Notes
Sentinel-2	Raster	10 -60	Available from 23-06-2015
JRC Global Surface Water (GSW)	Raster	30	Water occurence based on Landsat satellite observations from 1984 to 2015
OpenStreetMap	Vector	1-100	Selection of all water related tags
ALOS PALSAR DEM	Raster	12.5	Drainage Network and HAND creation

Table 1: Input Datasets

4.3 Sentinel - 2 Water occurrence

Sentinel-2 being the latest temporal resolution and highest spatial resolution data available, is very useful for detailed water surface boundary extraction. The frequency of water presence in a pixel can be described with historical observations of the same area in different moments in time, namely water occurence. The approach is based on the sampling of different cloudless historical images, to compute a single image that presents where and how often water occured. The values of this composite image vary from 0 (no detection of water) to 1 (detection in all samples).

Cloudless composite images can be generated by employing percentile images to estimate the average cloud-free reflectance values. Cloud coverage is estimated by exploiting the statistical properties of the image and more specifically the reflectance property of clouds, i.e the brightness in SWIR band which is ideal for cloud detection. The main idea is that the more bright the pixel appears in this band, the more likely it is that it is covered by clouds. For this, a quantile analysis of the pixel distribution in the whole image can be performed, just to choose the first quantile that is considered cloud free [2]. The percentile images can be computed according to Donchtyts (2016) on a per pixel and per-band basis using top of atmosphere (TOA) reflectance values, to avoid the confusion created by different atmospheric correction algorithms of satellites [8]. To generate the water occurrence image, all cloudless binary images will be summed and the values of all pixels will be devided by the total amount of images [41].

4.3.1 Spectral indices

The water pixels present in the cloudless historical composites will be extracted with the help of the following water indices (Eq:1,2,3):

The Normalized Difference Water Index (NDWI) is found from the normalized difference between the green and near-infrared bands to assign each pixel a value between -1 and 1, calculated using the McFeeters (1996) equation [26]:

$$\mathbf{NDWI} = \frac{GREEN - NIR}{GREEN + NIR} \tag{1}$$

The Modified Difference Water Index (MNDWI) of Xu (2006) is considered more reliable in urban areas than the Normalized Difference Water Index (NDWI) [43]. However the limitation of MNDWI is that it cannot discriminate water from snow. It is expressed by the Eq:2:

$$\mathbf{MNDWI} = \frac{GREEN - SWIR1}{GREEN + SWIR1} \tag{2}$$

The resulting positive values represent the water features because of their higher TOA reflectance in GREEN and SWIR bands, while non-water features have negative values.

The Normalised Difference Vegetation Index will be used to exclude dark vegetated areas (Eq:3) with a high threshold of 0.3 [8]:

$$\mathbf{NDVI} = \frac{NIR - RED}{NIR + RED} \tag{3}$$

4.3.2 Canny edge detector

The Canny edge detector is a widely used method for accurate edge detection in images [9]. The edge filter can assist in the detection of boundaries between water and non water pixels, which helps further to reduce the extend of the area where the Otsu thresholding is applied. The detected sharp edges between water and land will be expanded with a buffer zone to make sure that all probable water and land pixels around the boundary are captured. This way we it will be possible to obtain a bimodal distribution which will assist the distinction of the two classes in the derived histogram of MNDWI values.

The algorithm consists of the following stages:

1. Image Smoothing: Edge detectors are prone to noise and therefore they are firstly smoothed with a square-sized Gaussian structural kernel usually of size 5×5 [23].

2. Gradient intensity calculation: The gradient direction defines the orientation of an edge, whereas the gradient magnitude indicates the intensity of a change in the reflectance values. High gradient magnitudes reveal the detection of an edge.

$$G = \sqrt{{G_x}^2 + {G_y}^2} \tag{4}$$

$$\theta = tan^{-1}\frac{G_x}{G_y} \tag{5}$$

   where G_x and G_y the x,y derivatives of the current pixel. θ is rounded to 0 (horizontally),45 (diagonally), 90 (vertically) or 130 (diagonally) degrees.

3. Non-maximum suppression: All pixels are checked to see if they are a local maximum in certain neighborhood. If they are not, they are suppressed, resulting in very thin edges.

4. Double thresholding: Removal of small pixel noises, based on two threshold values of the intensity gradient. It is applied to detect strong edges only.

5. Hysteresis thresholding: Pixels below a certain threshold are discarded. This way edges that are weak are suppressed.

4.3.3 Otsu Thresholding

In order to separate water from non water features, a threshold value for MNDWI will be estimated. Dynamic local thresholding will help avoiding errors in the surface water extraction procedure. Otsu thresholding [36] is based on a histogram of all MNDWI values in a certain area. The goal of this method is to create a binary image 0,1 of two different classes, white (no water) and black (water) pixels. The general idea of the method is to find the threshold that minimizes the weighted within-class variance which is equal to the weighted sum of variances of the two classes Eq:6. This is done by exploring all possible threshold values and calculating the variance of all pixels on each side of the threshold.

$${\sigma_w}^2(t) = {\omega_0}^2(t) * {\sigma_0}^2(t) + {\omega_1}^2(t) * {\sigma_1}^2(t) \tag{6}$$

where σ_w the intra-class variance, ω_0 and ω_1 the probabilities of the two classes separated by the threshold t and σ_0 and σ_1 variances of the two classes.

$$\omega_0 = \sum_{i=0}^{t-1} P(i) \tag{7}$$

$$\omega_1 = \sum_{i=t}^{L} P(i) \tag{8}$$

The class means μ_0, are given by the following equations:

$$\mu_0(t) = \frac{\sum_{i=0}^{t-1} iP(i)}{\omega_0(t)} \tag{9}$$

$$\mu_1(t) = \frac{\sum_{t}^{L-1} iP(i)}{\omega_1(t)} \tag{10}$$

Yousefi (2011) proved that maximizing the between-class variance, instead of minimizing the within-class variance has a higher performance [46]. Therefore,

$${\sigma_b}^2(t) = \sigma^2 - 2\sigma_w = \omega_0(t) * \omega_1(t) * (\mu_0(t) - \mu_1(t))^2 \tag{11}$$

4.4 Drainage network

Drainage networks will be extracted from the gridded elevation data DEMs [32]. Firstly the slopes will me computed providing a slope map, in order to estimate the flow directions of water (multi-D8 algorithm). Afterwards, the artificial pits will be removed by filling the depressions on the slope, to extract only the major drainage paths and compute the drainage accumulation areas. Then the drainage network will be extracted by means of a threshold for the flow accumulation [28].

The extraction of surface drainage features is however in problematic in flat reliefs as they create parallel lines during processing. Therefore, a need drainage direction in flat surfaces can be assigned by modifying the DEM elevations to enforce two gradients: one away from higher terrain, and one towards lower terrain.

4.4.1 Multi-D8 algorithm

The Multi-D8 algorithm described by Qin et. al will be used to determine the direction of the flow direction and accumulation of water on a terrain [37]. The flow direction is the direction the water would naturally flow towards, which is from each cell to its downslope neighbor or neighbors on the DEM [18]. To

describe this flow direction per DEM cell, eight discrete flow angle values towards the eight neighbours of the pixel (left, right, up, down and the diagonals [23] will be used. The flow accumulation at a given DTM cell can be estimated by the area that drains to it [34].

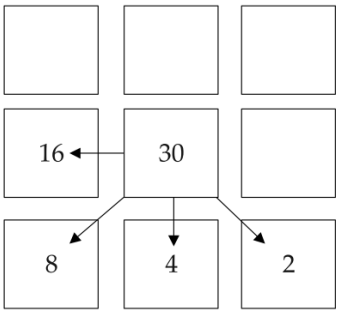


Figure 5: Multi-D8 algorithm [23]

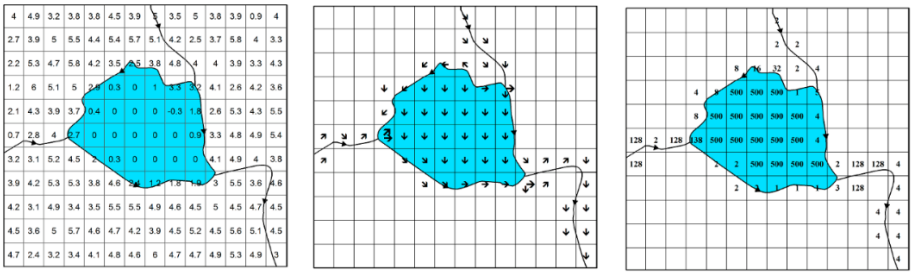


Figure 6: Flow direction and accumulation of water reservoir using multi-D8 [23]

4.4.2 HAND

The Height Above Nearest Drainage (HAND) dataset can be derived from the ALOS PALSAR Digital Elevation Model (DEM). It will be used as a topographic mask to detect and exclude pixels where potential errors occur due to the spectral similarity of terrain shadows to water especially in mountainous areas but also flood areas. HAND is a normalised version of the DEM, calculated based on the extracted drainage network. Basically it is a map that describes the vertical distance to the nearest drainage, i.e it the elevation difference of every pixel to its nearest drainage pixel. In order to extract HAND the flow accumulation of the drainage network needs to be thresholded with a maximum number of upstream cells. The HAND values, as proposed by Renno et al. (2008), can be estimated by classifying the pixels according to their drainage potential, as drainage or non drainage pixels [38]. Then the HAND values are

derived from the difference of the height of the pixel from the original DEM and the height of the drainage pixel that is closer to this (non drainage) pixel.

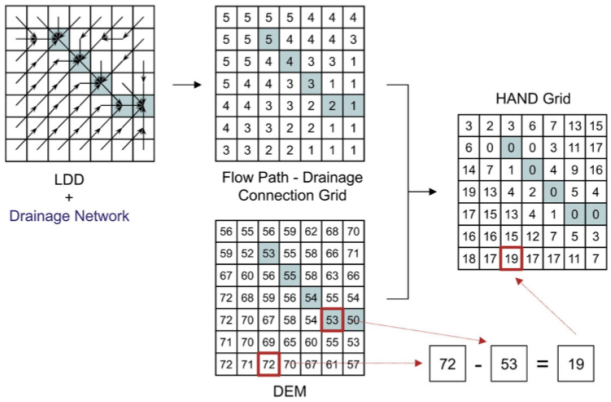


Figure 7: HAND value extraction based on Local Drainage Directions (LLD)[38]

4.5 Resampling with bicubic interpolation

Conversion from raster to vector might cause inconsistencies between the created polygons, due to the varying resolution of the datasets. A way to avoid this, is by resampling the raster image. Image resampling is a mathematical process of creating a new version of the raster cell grid with a with a different width and/or height in pixels. The value of each cell in the new raster will be computed by sampling or interpolating in a neighborhood of cells of this pixel in the original raster object [39]. Bicubic interpolation is considered to be slower in computation speed, but it is supposed to have better smoothing results when upsampling.

$$f(x,y) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_i j x^i y^j \tag{12}$$

where x and y the coordinates of the new location, f(x,y) the value of the pixel and $a_i j$ the 16 coefficients for the 16 neighbors.

4.6 Logistic Regression

Logistic Regression as described by Hestie et. al (2009) and Mueller et al. (2016) is a statistical method that describes the relationship between a dependent variable (Y) and one or more independent $(X_1, X_2, ..., X_N)$, in our case the input datasets [31]. The dependent variable takes only two values, 0 or 1. The goal of the logistic regression is to provide confidence intervals (probability scores) on the predicted values of 0 or 1. From the inpute datasets, OSM will be in

binary form , Sentinel-2 and JRC water occurence in the range [0,1], where 0 represents a negative response and the 1 represents a positive response (water or no water feature). The slope will be also considered in the process as a real value in the range [0,90]. The mean of the dependent variable (pixel) will be the proportion of positive responses. If P is the proportion of the input data with value 1, then 1-P is the probability of an outcome of 0. Therefore the logistic regression can estimate the log odds of the event that Y=1 from the Equation:

$$log\frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{13}$$

By exponentiating the log odds and with a simple algebraic manipulation we can derive the odds that a pixel represents water according to the input datasets:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n}} = \tag{14}$$

where β_0 the intercept from the linear regression equation and $\beta_1..n$ the coefficients that maximize the likelihood of predicting a high probability for training data belonging to class 1 while predicting a low probability for data actually belonging to class 0. More simply explained these coefficients are chosen to maximize the likelihood of observing the sample data that have been given as input.

Each of the input dataset (Sentinel water occurrence, JRC water occurrence, OSM, slope map from ALOS PALSAR, drainage network) will be used as indication of the reliability/confidence of water presence in a certain pixel p. For example, a pixel with high slope classified as water is probably incorrect. The probability of the pixel being an actual water pixel, will be defined by certain threshold value:

$$P(Y=1) > k \tag{15}$$

4.7 Inverse weighted Distance (IDW)

IDW is an interpolation method that uses distance to identify the neighbours of a point, and to assign to them weights. The specified neighborhood determines how far and where to look for the measured values to be used in the prediction. A 'searching circle' with a defined radius or a certain number of the closest neighbors can be used to select the data points involved in the interpolation at desired location [35].
The weight assigned to each point p at the interpolated location x is:

$$w_i(x) = |xp|^{-h} \tag{16}$$

where h is the power to be used, and $|xp|$ is the distance between the location x and the point p. The value of h determines how quickly/slowly the weight decreases with distance.

Afterwards the weighted average of the neighbors can be computed:

$$f(x) = \frac{\sum_{i=1}^{k} w_i(x) * a_i}{\sum_{i=1}^{k} w_i(x)} \tag{17}$$

where $w_i(x)$ is the weight of each neighbour p (with respect to the interpolation location x) and a_i the attribute of pi. A neighbour p is a sample point that is used to estimate the value of location x. In this implementation the attribute are the x,y coordinates of the neighbors of the interpolated location. This way the weighted average value of x and of y for the location x can be computed.

4.8 Euclidean Distance Transform MAP (EDM)

To create the Euclidean Distance Transform map the new optimal polygon needs to be rasterised. Afterwards, according to Meijster (2004) , the distance of a pixel to the closest boundary point can be calculated with the Euclidean Distance Transform for binary images [27]:

$$EDT(x, y) = min(x - i)^2 + G(i, y)$$

(18)

$$G(i, y) = min(y - j)^2 \; where \; F(i, j) = 0 \tag{19}$$

where F(i,j) the input image, i,j the rows and columns of the image array respectively.

The algorithm works as follows: The image is stored in an array of columns and rows. Afterwards, the algorithm iterates through all pixels from top to bottom and then bottom to top to, to compute the minimum distances in this dimension G(x,y) to the closest boundary pixel. Then the array G is scanned from left to right and right to left to calculate again the minimum distance to the closest boundary point [41].

4.9 Voronoi based skeletonization

The Voronoi diagram consists of cells of points p, in which all points are closer to the point p than to any other point. Considering this, the Voronoi Diagram can be extracted based on the the Euclidean Distance map, to compute the skeleton, i.e the medial axis of the water polygon. The approximation of the skeleton, will be done from the Voronoi diagram of the points sampled along

the boundary of the water polygon. The Voronoi diagram consists of convex polygon which are formed by vertices. By removing those vertices that intersect with the boundary of the water polygon, we can acquire the skeleton. As also presented by Thissen (2019) the derived centerline can be intersected with values of the EDM. This way we can acquire the centerline where every pixel has the information of the euclidean distance to the nearest boundary. The width can be obtained by multiplying by two [41] exploiting this way the equidistance property of the centerline from the boundaries.

4.10 Geometric Accuracy of OSM water polygons

To compare the OSM water polygons with the computed optimal water polygon, the surface distance and compactness of the polygons will be computed. The surface distance as described by Vauglin (1997) is [10]:

$$dS = 1 - \frac{S(A \cap B)}{S(A \cup B)}$$

(20)

where A, B the polygons to be compared, and dS takes values in the range of [0,1]. If distance is 0 then the polygons are equal and if 1 they are disjoint. The distance is calculated by dividing the intersection area of the polygons with their union.

Apart from the positional differences of the two polygons, also the geometric ones need to be evaluated. For this, the compactness of the polygons can be calculated according to MacEachren (1985) using the equation [24]:

$$C = 2\pi \times area/perimeter^2 \tag{21}$$

4.11 Tools

The following tools are considered useful to fulfill the current research:

1. Google Earth Engine cloud computing platform for Satellite data processing
2. Python for processing
3. ArcGIS for GIS operations
4. GDAL, Shapely, Fiona, rasterio libraries for processing of vector and raster data
5. ArcHydro for HAND extraction
6. Osmose and JOSM validator for detection of geometric, topological error dertection of OSM data

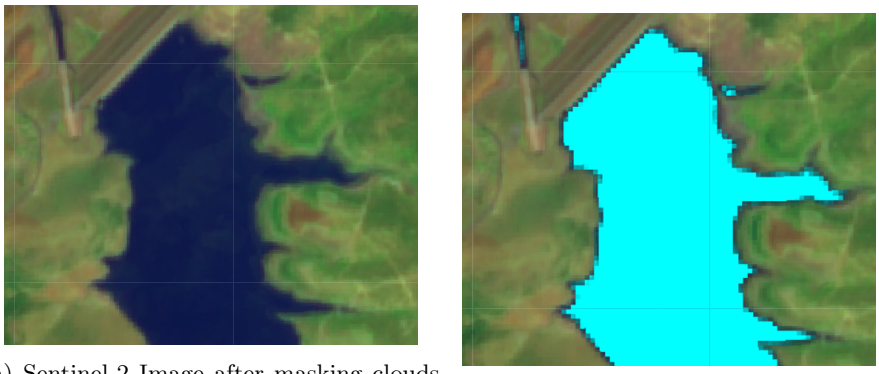
5 Preliminary results

The JRC water occurrence layers were scaled from [0,100] values to [0,1] and resampled with bicubic interpolation, resulting in an new image where output pixels appear more smooth relatively to the original image (Figure 8).

(a) (b)

Figure 8: JRC resampling 30 to 10 m resolution

Figure 9 shows a cloudless Sentinel-2 image, generated by masking clouds and cirrus over one year of data and afterwards by taking the median. For this Sentinel-2 TOA reflectance data were pre-filtered to get less cloudy granules. Then the Modified Normalized Difference Water Index (MNDWI) was calculated.



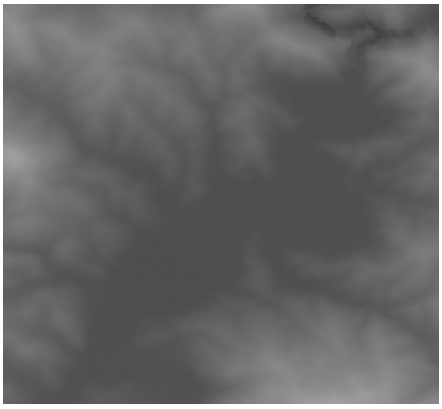
(a) Sentinel-2 Image after masking clouds and cirrus

(b) MNDWI index calculation

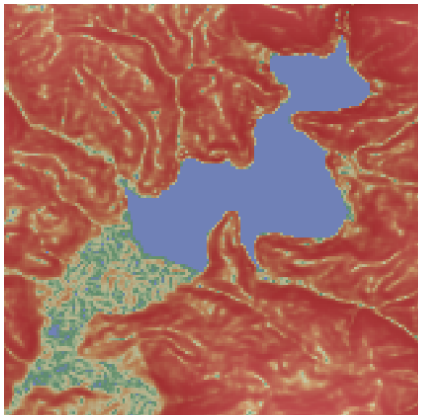
Figure 9: Surface water detection from Sentinel 2 Imagery

To compute the slope map, Digital elevation data from ALOS PALSAR were downloaded and processed to create slope values for every pixel in the raster image 10. The area in blue color represents the part with low slope values (<1 %), as an indication of accumulation of water.

(a) Water reservoir from Sentinel Imagery



(b) Digital elevation Model



(c) Extraction of slope map

Figure 10: Calculation of slopes from DEM

6 Timeplan

A GANTT chart is constructed to present how the project, and more specifically the time-related progress of each of phase of the project parts, the deadlines and the delivery of reports are organized (Page 27).

Time Planner

Deadline

ACTIVITY	Start	Duration	Period
			1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
P2 preparation	1	2	
P2 proposal & presentation	2	2	
Preprocessing of OSM and JRC	3	1	
Drainage Network and HAND	4	2	
Sentinel Water Occurrence	6	3	
Logistic Regression	9	3	
Optimal geometry extraction	11	3	
P3 Colloquium midterm	12	2	
Write Draft Thesis	3	17	
EDM map	14	2	
Veronoi skeletonization	16	2	
Width calculation	17	2	
Validation of OSM	18	2	
P4 preparation	18	2	
P4 formal process assesment	20	1	
Write final Thesis	20	6	
P5 preparation	23	2	
P5 presentation & final thesis	25	2	

References

[1] A.B. Ariza-Villaverde, F.J. Jiménez-Hornero, and E. Gutiérrez de Ravé. "Influence of DEM resolution on drainage network extraction: A multifractal analysis". In: *Geomorphology* 241 (2015), pp. 243–254.

[2] Nicolas Avisse et al. "Monitoring small reservoirs' storage with satellite remote sensing in inaccessible areas". In: *Hydrology and Earth System Sciences* 21.12 (2017), pp. 6445–6459.

[3] Christopher Barron, Pascal Neis, and Alexander Zipf. "A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis". In: *Transactions in GIS* 18.6 (2014), pp. 877–895. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12073`.

[4] Prajnaparamita Bhattacharya. "Quality assessment and object matching of OpenStreetMap in combination with the Dutch topographic map TOP10NL". In: (2012), pp. 23–88.

[5] Maria Antonia Brovelli and Giorgio Zamboni. "A new method for the assessment of spatial accuracy and completeness of OpenStreetMap building footprints". In: *ISPRS International Journal of Geo-Information* 7.8 (2018).

[6] John Nikolaus Callow, Kimberly P. Van Niel, and Guy S. Boggs. "How does modifying a DEM to reflect known hydrology affect subsequent terrain analysis?" In: *Journal of Hydrology* 332.1 (2007), pp. 30–39.

[7] Gennadii Donchyts. "Planetary-scale surface water detection from space". PhD thesis. Jan. 2018.

[8] Gennadii Donchyts et al. "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". In: *Remote Sensing* 8.5 (2016).

[9] Rui-ling Duan, Qing-xiang Li, and Yu-he Li. "Summary of image edge detection [J]". In: *Optical Technique* 3.3 (2005), pp. 415–419.

[10] Vauglin F. "Modèles statistiques des imprécisions géométriques des objets géographiques linéaires". PhD thesis. 1997.

[11] Min Feng and Yan Bai. "A global land cover map produced through integrating multi-source datasets". In: *Big Earth Data* 3.3 (2019), pp. 191–219.

[12] Marcus Goetz and Alexander Zipf. "The Evolution of Geo-Crowdsourcing: Bringing Volunteered Geographic Information to the Third Dimension". In: Jan. 2013, pp. 139–159.

[13] Laurence Hawker et al. "Perspectives on Digital Elevation Model (DEM) Simulation for Flood Modeling in the Absence of a High-Accuracy Open Access Global DEM". In: *Frontiers in Earth Science* 6 (2018), p. 233.

[14] Chang Huang et al. “An evaluation of Suomi NPP-VIIRS data for surface water detection”. In: *Remote Sensing Letters* 6.2 (2015), pp. 155–164.

[15] Chang Huang et al. “Detecting, Extracting, and Monitoring Surface Water From Space Using Optical Sensors: A Review”. In: *Reviews of Geophysics* 56.2 (2018), pp. 333–360.

[16] Ken Arroyo Ohori Hugo Ledoux Ravi Peters. “What is a digital terrain model ?” In: (2018), pp. 1–10.

[17] Gordana Jakovljević, Miro Govedarica, and Flor Álvarez-Taboada. “Waterbody mapping: a comparison of remotely sensed and GIS open data sources”. In: *International Journal of Remote Sensing* 40.8 (2019), pp. 2936–2964.

[18] Susan K Jenson and Julia O Domingue. “Extracting topographic structure from digital elevation data for geographic information system analysis”. In: *Photogrammetric engineering and remote sensing* 54.11 (1988), pp. 1593–1600.

[19] Lilian Vedasto Kato. “Integrating Openstreetmap Data in Object Based Landcover and Landuse Classification for Disaster Recovery”. In: (2018).

[20] Kamal Khodaei and Hamid Reza Nassery. “Groundwater exploration using remote sensing and geographic information systems in a semi-arid area (Southwest of Urmieh , Northwest of Iran)”. In: June (2008).

[21] J.P. Lacaux et al. “Classification of ponds from high-spatial resolution remote sensing: Application to Rift Valley Fever epidemics in Senegal”. In: *Remote Sensing of Environment* 106.1 (), pp. 66–74.

[22] Jing Li and David W.S. Wong. “Effects of DEM sources on hydrologic applications”. In: *Computers, Environment and Urban Systems* 34.3 (2010), pp. 251–261.

[23] Li, Yang, and Wu. “A Method of Watershed Delineation for Flat Terrain using Sentinel-2A Imagery and DEM: A Case Study of the Taihu Basin”. In: *ISPRS International Journal of Geo-Information* 8.12 (2019), p. 528. ISSN: 2220-9964.

[24] Alan M. Maceachren. “Compactness of Geographic Shape: Comparison and Evaluation of Measures”. In: *Geografiska Annaler: Series B, Human Geography* 67.1 (1985), pp. 53–67.

[25] S. K. McFEETERS. “The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features”. In: *International Journal of Remote Sensing* 17.7 (1996), pp. 1425–1432.

[26] S. K. McFEETERS. “The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features”. In: *International Journal of Remote Sensing* 17.7 (1996), pp. 1425–1432.

[27] Arnold Meijster. “Efficient sequential and parallel algorithms for morphological image processing”. English. Relation: https://www.rug.nl/ $date_submitted$: $2004Rights : UniversityofGroningen$. PhD thesis. 2004.

[28] Amnon Meisels, Sonia Raizman, and Arnon Karnieli. “Skeletonizing a DEM into a drainage network”. In: *Computers Geosciences* 21.1 (1995), pp. 187–196.

[29] Anil Kumar Misra. “Climate change and challenges of water and food security”. In: *International Journal of Sustainable Built Environment* 3.1 (2014), pp. 153–165.

[30] Peter Mooney, Padraig Corcoran, and Adam C Winstanley. “Towards quality metrics for OpenStreetMap”. In: *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. ACM. 2010, pp. 514–517.

[31] N Mueller et al. “Remote Sensing of Environment Water observations from space : Mapping surface water from 25 years of Landsat imagery across Australia”. In: 174 (2016), pp. 341–352.

[32] John F. O’Callaghan and David M. Mark. “The extraction of drainage networks from digital elevation data”. In: *Computer Vision, Graphics, and Image Processing* 28.3 (1984), pp. 323–344.

[33] Andrew Ogilvie et al. “Surface water monitoring in small water bodies: Potential and limits of multi-sensor Landsat time series”. In: *Hydrology and Earth System Sciences* 22.8 (2018), pp. 4349–4380.

[34] Ken Arroyo Ohori and Hugo Ledoux. “Applications : runoff modelling”. In: (2018).

[35] Ken Arroyo Ohori and Hugo Ledoux. “Spatial interpolation (1/2)”. In: (2018), pp. 1–10.

[36] Nobuyuki Otsu. “A threshold selection method from gray-level histograms”. In: *IEEE transactions on systems, man, and cybernetics* 9.1 (1979), pp. 62–66.

[37] C. Qin et al. “An Adaptive Approach to Selecting a Flow-Partition Exponent for a Multiple-Flow-Direction Algorithm”. In: *Int. J. Geogr. Inf. Sci.* 21.4 (Jan. 2007), pp. 443–458.

[38] Camilo Daleles Rennó et al. “HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia”. In: *Remote Sensing of Environment* 112.9 (2008), pp. 3469–3481.

[39] Jonathan Sachs. “Image Resampling”. In: (2001), pp. 1–14.

[40] Maurizio Santoro et al. “Strengths and weaknesses of multi-year Envisat ASAR backscatter measurements to map permanent open water bodies at global scale”. In: *Remote Sensing of Environment* 171 (2015), pp. 185–201.

[41] J.J.M. Thissen. “Automating surface water detection for rivers : the estimation of the geometry of rivers based on optical earth observation sensors”. In: 2019.

[42] Huan Xie et al. “Evaluation of Landsat 8 OLI imagery for unsupervised inland water extraction”. In: *International Journal of Remote Sensing* 37.8 (2016), pp. 1826–1844.

[43] Hanqiu Xu. “Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery”. In: *International Journal of Remote Sensing* 27.14 (2006), pp. 3025–3033.

[44] Dai Yamazaki, Mark A Trigg, and Daiki Ikeshima. “Remote Sensing of Environment Development of a global ˜ 90 m water body map using multi-temporal Landsat images”. In: *Remote Sensing of Environment* 171 (2015), pp. 337–351.

[45] Di Yang et al. “Geo-spatial information science open land-use map: a regional land-use mapping strategy for incorporating openstreetmap with earth observations open land-use map: a regional land-use mapping strategy for incorporating OpenStreetMap with earth observations”. In: *Online) Journal* (2017), pp. 1993–5153.

[46] Jamileh Yousefi. “Image Binarization using Otsu Thresholding Algorithm”. In: *University of Guelph, Ontario, Canada* (2011).