Scan Line Void Filling of Airborne LiDAR Point Clouds for Hydroflattening DEM

Wai Yeung Yan, *Senior Member, IEEE*

***Abstract*—Generation of LiDAR-derived digital elevation model (DEM), particularly for hydrologic and shore environments, poses a continuous challenge. The presence of laser dropouts found on the water bodies causes data voids/holes in the airborne LiDAR data point clouds. Unnatural huge triangular artifacts may appear in these regions when a DEM is generated, resulting in not only unpleasant visual effect but also inaccurate terrain analyses. The United States Geological Survey has stressed the need of having a hydro-flattened DEM in the LiDAR Base Specification. Different forms of water bodies should be represented by a flat surface. Existing approaches mainly rely on the use of ancillary data or manual intervention during the hydroflattening process. In this study, an automatic data processing workflow is proposed to: 1) classify land and water data points collected by a topographic airborne LiDAR system based on the scan line intensity-elevation ratio; 2) perform scan line void filling of data points in close-to-nadir region and at both swath edges; 3) generate a virtual water surface based on the classified water data points; 4) perform hydroflattening on the DEM. The proposed workflow was examined using five datasets collected by topographic airborne LiDAR on the inland ponds and lakes, inland rivers, nontidal boundary water bodies, tidal water bodies, and islands, as addressed in the LiDAR Base Specification. The results showed that the proposed workflow can successfully generate hydroflattened DEMs and overcome the drawback of existing approaches.**

***Index Terms*—Airborne LiDAR, hydroflattened digital elevation model (DEM), hydroflattening, laser dropouts, scan line intensity-elevation ratio (SLIER), void filling.**

I. Introduction

AIRBORNE LiDAR data have received a favorable attention over traditional remote sensing imagery due to its capability of collecting 3-D topographic information together with the backscattered laser energy (i.e., intensity). One of the common LiDAR applications is to record the instantaneous topographic features in order to generate a digital elevation model (DEM) [1]. The derived DEM not only serves the purpose of 3-D visualization and topographic mapping, but also supports different thematic analyses, such as urban flood risk estimation [2], snow depth measurements [3], hydrogeomorphological assessment [4], and other ecosystem analyzes. The mature development of ground filtering algorithms, such as the recently introduced cloth simulation modeling [5], can further aid in delineating the on-ground data points and off-ground data points so that the bare earth, or the so-called digital terrain model (DTM), can be derived [6], [7]. Subsequently, the off-ground data points can be further normalized using the DTM. These normalized data points can be used to extract different land cover classes [8].

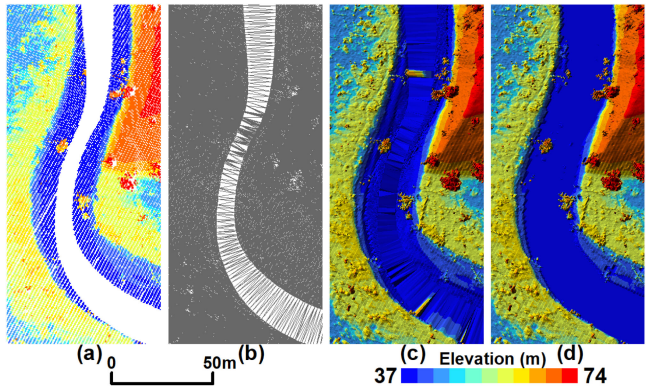

Fig. 1. (a) Airborne LiDAR data with void along an inland river. (b) and (c) Resulted TIN and DEM with undesired triangular meshes, respectively. (d) Hydroflattened DEM.

Despite the mature development of DEM ground filtering algorithms, there are several research areas that require further efforts. A prime example is the important and continuous challenge of generating DEM for hydrology and oceanography [9]. Laser pulses commonly follow specular reflection, energy absorption, and bragg scattering when backscattering from the water surface [10]. The combined effects result in laser dropouts found at large incident angles and extreme peak returns in close-to-nadir region [11], [12]. Such a phenomenon is especially notable when near-infrared (NIR) laser (e.g., 1064 or 1550 nm) is in operation. These anomalies appear as data holes or voids of different sizes on water bodies, and thus, present serious issues in DEM generation. Importantly, unpleasant triangular meshes and unnatural surface artifacts may appear within these water regions due to the appearance of these data voids [see Fig. 1]. These defects may cause a negative impact on the hydrological connectivity, i.e., computation of terrain's slope and aspect or flow accumulation [13]. Therefore, the United States Geological Survey (USGS) has highlighted the need of having the

Manuscript received December 22, 2020; revised April 20, 2021 and May 27, 2021; accepted June 8, 2021. Date of publication June 15, 2021; date of current version July 14, 2021. This work was supported in part by the FCE Start-up Fund (BE2U) of the Hong Kong Polytechnic University, and in part by the Early Career Scheme under Project 25213320 by the Research Grants Council of the Hong Kong Special Administrative Region. *(Corresponding author: Wai Yeung Yan.)*

The author is with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, and also with the Department of Civil Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: waiyeung.yan@polyu.edu.hk; waiyeung.yan@ryerson.ca).

Digital Object Identifier 10.1109/JSTARS.2021.3089288

hydroflattened DEM for hydrologic and oceanographic environments in the LiDAR Base Specification Version 1.0 [14].

The USGS LiDAR Base Specification has listed guidelines for the hydroflattening process of airborne LiDAR data. It specifically mentions that "*the goal is to produce topographic DEM that, with respect to water surfaces, resemble DEM derived from traditional photogrammetric methods and, to the degree practical, are free of unnatural triangulation effects.*" The specification covers five typical water and hydrologic environments, including inland ponds and lakes, inland streams and rivers, nontidal boundary water bodies, tidal water bodies, and islands. The common requirement for these typical environments mainly refers to the need of having a flattened water body represented by a flat and level water surface. The latest version of LiDAR Base specification [14] further supplies the detailed specifications, such as the width, area, diameter, etc., of these five types of water bodies that have to undergo the hydroflattening process. Prior to hydroflattening, it is necessary to have a precisely classified land and water data points within the collected LiDAR data. With the determined water data points and their instantaneous water surface, all the water data points can be flattened to a consistent elevation prior to generating the triangulated irregular network (TIN) or DEM.

Several initial attempts can be found on generating a hydroflattened DEM based on airborne LiDAR point clouds or rasterized dataset. Early research mainly relied on the use of existing breaklines or manual digitization of breaklines [15], [16] to locate the water regions. Coleman *et al.* [17] proposed to integrate the bathymetric LiDAR data with depth sounding data obtained from the Electronic Nautical Charts to fill the data holes. Deshpande and Yilmaz [18] proposed a semiautomatic approach by first creating a bare ground surface (i.e., separating the ground and nonground points) and then a virtual water surface based on the existing river centerline. The derived virtual surface was intersected with the bare ground surface to generate the bank shorelines and subsequently the hydroflattened DEM. Zheng *et al.* [19] proposed a raster-based approach to extract river network based on a least-cost path approach. The detected river centerlines were used to check the network connectivity and ensure nonnegative curvature in order to generate a filtered mask for the subsequent hydroflattening process. Czuba *et al.* [20] even reported the use of a boat survey to determine the instantaneous water surface to generate the hydroflattened DEM.

Most of these existing studies focus on the environment of inland streams and rivers, while the other four environments pointed out by the USGS (i.e., inland lakes/ponds, tidal and nontidal water bodies, and islands) are not well researched or addressed. In addition, the existing approaches heavily rely on either manual intervention, existing topographic data (e.g., river centerline, breakline, or tidal data) or field survey to aid in the task. Also, most of the studies focus on grid-based rasterized LiDAR datasets, in which details may already be lost from the interpolation of the original LiDAR point clouds. This study proposes a fully automatic approach to generate a hydroflattened DEM based on the use of the raw point cloud data strips collected by airborne LiDAR system. The approach is built upon the author's previously proposed algorithm, i.e., scan line intensity-elevation ratio (SLIER) [12], to automatically determine the water surface. In addition, two new void filling algorithms are proposed to compensate the laser dropouts found at the swath edges and in the close-to-nadir region. With all the possible voids being filled, the approach can overcome the drawback of having unpleasant triangular meshes while generating the hydroflattening DEM. Also, the proposed method was examined with the five typical water environments as addressed by the USGS, unlike other studies only researched on specific water/shore types.

II. Method

The proposed method is applicable to the point clouds collected by airborne LiDAR systems operated with oscillating mirror or rotating prism, which are usually found in commercial topographic airborne LiDAR systems, such as Leica Geosystems, RIEGL, and Teledyne Optech. Original LiDAR data strips collected by any of these systems, which are operated with NIR laser (i.e., 1064 or 1550 nm), can serve as an input for the proposed data processing. It is assumed that the collected LiDAR data strips have the following fields, including x, y, z, I, s, a, t, which represent x-coordinate, y-coordinate, z-coordinate, intensity, scan flag, scan angle, and time, respectively.

The original time-tagged LiDAR data in LAS format are first used to compute the SLIER, which reads the LiDAR scan line one by one. SLIER can be computed by having the standard deviation of intensity divided by the standard deviation of elevation for each scan line so that higher SLIER values usually represent water regions [12]. Once LiDAR points with high SLIER values are identified, their respective mean elevation can be computed so that it acts as the virtual water surface. Every LiDAR data point is either labeled as land or water based on the virtual water surface. To compensate the data holes/voids found in the LiDAR data strips, two void filling algorithms are proposed to fill the laser dropouts at the swath edges and in the close-to-nadir water region. Finally, all the water data points, including the original LiDAR data and void filled data, are flattened with the elevation of the virtual water surface. The processed LiDAR data strips can be used to generate a hydroflattened DEM. Fig. 2 shows the overall workflow of the hydroflattening.

A. Scan Line Intensity-Elevation Ratio

SLIER is an airborne LiDAR ratio index, which is proposed for automatic water surface mapping [12]. When a laser beam interacts with the water surface, extreme peak intensity returns are commonly recorded in the close-to-nadir region while laser dropouts appear at the swath edges, i.e., large incident angles. SLIER is equipped to capture these unique effects and is defined as the standard deviation of intensity (I) divided by the standard deviation of elevation (z) along each scan line (l) for the LiDAR data (L), as shown as follows:

$$\text{SLIER} = \frac{\sigma_I}{\sigma_z}, \quad \forall\, l \in L. \tag{1}$$

With this setting, water bodies are usually associated with high SLIER values, and nonwater bodies (e.g., land) usually capture

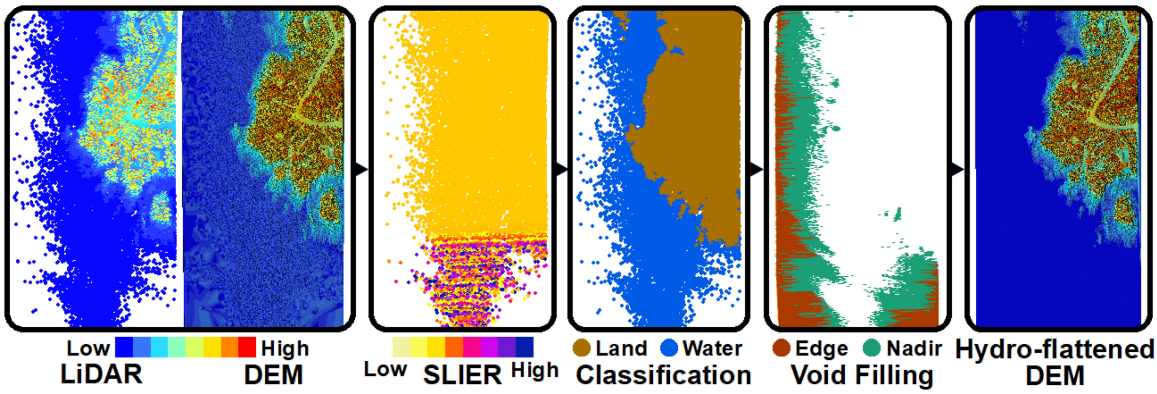

Fig. 2. Process of automatic generation of hydroflattened DEM.

low SLIER values. An enhanced version of SLIER can further inject the properties of laser dropouts at large incidence angles; readers are referred to [12]. After computing the SLIER value for each scan line, LiDAR data points with very high SLIER values can be labeled as water data points. According to the previous study of [12], top 10% of the SLIER is considered to be an optimal cutoff to pick up the water data points. The respective mean elevation of these selected data points can be treated as a virtual water surface. Any data point of the original LiDAR data strips with elevation below the virtual water surface should be labeled as water, whereas LiDAR data points with elevation above the virtual water surface should be labeled as land. This, thus, provides an initial delineation of land and water regions. However, the use of SLIER has certain limitations. It requires the LiDAR data being collected having a portion of the water region entirely covered by the swath. If such a requirement cannot be fulfilled, the turnaround solution can be found in [11], where a Gaussian mixture model (GMM) can be fitted on the LiDAR intensity/elevation histogram to look for the water surface elevation. The abovementioned mechanism can be adopted to initially label the land and water bodies. If improvements on the classification results are desired, one can consider treating the above labeled data points as training data and performing a land-water classification based on different LiDAR-derived features. A comprehensive land-water classification workflow can be found in [11].

B. Scan Line Void Filling

Even after all the LiDAR data points are labeled as either land or water, hydroflattening may not commence until data voids/holes are filled with artificial points. The filling process can help avoid generating unpleasant triangular meshes in the resulting DEM. Since the proposed workflow focuses on the use of original LiDAR data strips, one can rely on the properties of swath width and time difference between two consecutive points to fill the data voids along each scan line. Here, two void filling algorithms are proposed to tackle the cases where data voids are found in the close-to-nadir region and at both swath edges. Fig. 3 illustrates the appearance of the data voids under these

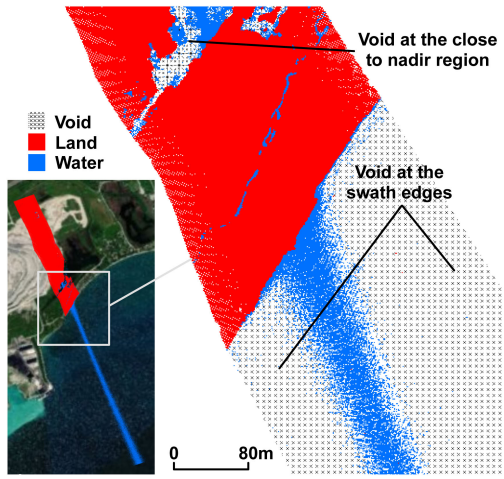

Fig. 3. Examples of data voids found on the water bodies.

two scenarios, and Fig. 4 shows the proposed workflow of the two scan line void filling algorithms.

1) Void Filling at Swath Edges: When an airborne LiDAR system is operated with an oscillating mirror or a rotating prism, a scan line pattern should appear on the ground with the data points collected from left to right (with respect to the flight direction) and then right to left direction. In the case when the the airborne LiDAR system surveys on the coastal water region where the entire swath width covers the water region, laser dropouts would occur at the swath edges. To compensate the voids found at the swath edges, Algorithm 1 is proposed. The algorithm requires an input of 2-D point spacing m that can be estimated within the data strip L. Subsequently, the algorithm starts by first estimating the swath width w of the LiDAR data strip and selecting the maximum length of the scan line l within the LiDAR data strip L. Then, the algorithm handles the k number of scan line l individually. It first skips the first scan line to avoid an incomplete swath scan. The two edge points, i.e., p_l and p_r, of the second scan line are stored as p'_l and p'_r so that they act as a constraint for the next scan line. Afterward, the algorithm looks for the approximate nadir point, p_c, along each l. Specifically, this task can be achieved by finding all data points

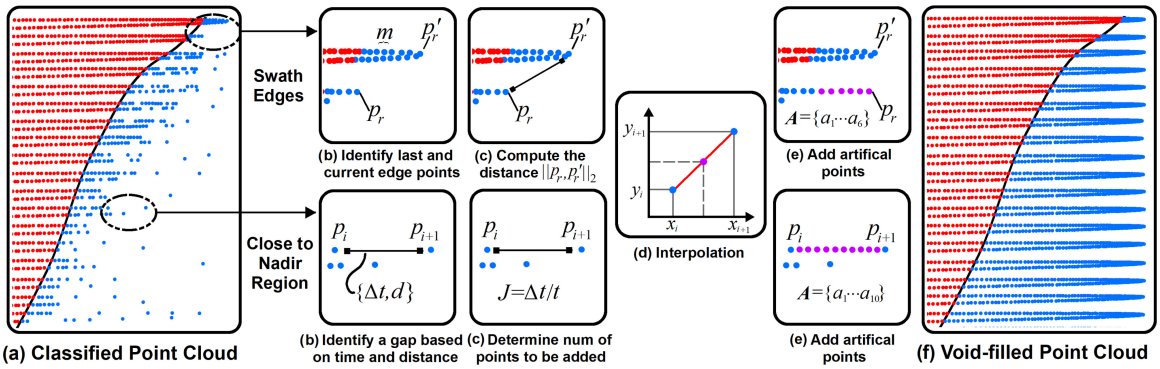

Fig. 4. Overall workflow of the two scan line void filling algorithms at swath edges and in close-to-nadir region (red points = ground and blue points = water).

with scan angle equivalent to zero. Then, the mean value of these points is computed. The specific data point that is close to this mean value is treated as the nadir point (i.e., p_c in Algorithm 1). In addition, the length of the scan line is estimated as d. If the length d is less than or equal to the maximum swath width w, then this scan line suffers from laser dropouts at the swath edges. In this case, Algorithm 1 attempts to add artificial point(s) at either or both of the swath edge(s) separately. The x and y coordinates of the artificial point can be determined by first constraining its 2-D direction with respect to the nadir point ($p_c = \{x_c, y_c\}$) that should be equal to the 2-D direction between, for instance, the right edge point ($p_r = \{x_r, y_r\}$) and p_c as shown as follows:

$$\frac{y - y_c}{x - x_c} = \frac{y_r - y_c}{x_r - x_c}. \tag{2}$$

Also, the offset distance between, for instance, the left edge point and the artificial point is constrained by the mean point spacing m, as shown as follows:

$$\sqrt{(x - x_r)^2 + (y - y_r)^2} = m. \tag{3}$$

Since (2) and (3) have two unknowns, the simultaneous equations can simply be solved algebraically. One should bear in mind that the set of simultaneous equations has exactly two solutions (a_1, a_2). A specific condition is made to determine the needed artificial point (a) from the two solutions: it is the one that is farther from the nadir point (p_c). The newly added artificial point is thus being treated as an updated edge point. The above process is identical for the left edge point by simply replacing p_r with p_l. After the artificial point a is estimated, a specific condition is placed here to check whether the newly estimated a is closer to the last edge point p_r' than the current edge point p_r. If this condition is fulfilled, the newly estimated point a is added to the array A.

The abovementioned artificial point adding process is repeated for the left edge point p_l. The entire process iterates until the length of the current scan line d, which is updated with the newly added artificial points at both swath edges, is less than or equal to the maximum swath edge w (line 25 of Algorithm 1). The updated edge points can act as p_r' and p_l' for the next scan line. Once the abovementioned process is accomplished, the array A is combined with the data strip L, where L' indicates the void filled LiDAR data strip.

Algorithm 1: Void Filling at Swath Edges.

Input : A data strip L with k number of scan lines (l)
Output: A void filled data strip L'

1 Initialize array A, M;
2 $w = \underset{l \in L}{\text{argmax}}\,|l|$; // Maximum swath width
3 **for** $i \in n \wedge i \neq n$ **do**
4 $\quad M[i] \leftarrow ||p_i, p_{i+1}||_2$;
5 $m \leftarrow \overline{M}$; // 2-D mean point spacing
6 **for** $i \in k \wedge i \neq 1$ **do**
7 $\quad \{p_l \cdots p_c \cdots p_r\} == l$; // p_l and p_r are the far left and right points
8 $\quad$ **if** $i \leftarrow 2$ **then**
9 $\quad\quad p_l' = p_l$; // skip the 1st scan line
10 $\quad\quad p_r' = p_r$; // store the two edge points of the 2nd scan line
11 $\quad$ **else**
12 $\quad\quad p_c \leftarrow \tilde{l}$; // estimate nadir point
13 $\quad\quad d \leftarrow ||p_l, p_r||_2$; // scan line length
14 $\quad\quad$ **do**
15 $\quad\quad\quad a_1, a_2 \leftarrow$ **solve** Eqs. 2 & 3 with p_r;
16 $\quad\quad\quad$ **if** $||a_1, p_c||_2 > ||a_2, p_c||_2$ **then**
17 $\quad\quad\quad\quad a = a_1$;
18 $\quad\quad\quad$ **else**
19 $\quad\quad\quad\quad a = a_2$;
20 $\quad\quad\quad$ **if** $||a, p_r'||_2 < ||p_r, p_r'||_2$ **then**
21 $\quad\quad\quad\quad p_r \leftarrow a$;
22 $\quad\quad\quad\quad A \leftarrow a \cup A$; // add new point a to the A array
23 $\quad\quad\quad$ **repeat** line 12 to 19 for p_l;
24 $\quad\quad\quad d \leftarrow ||p_l, p_r||_2$;
25 $\quad\quad$ **while** $d \leq w$
26 $\quad\quad p_l' = p_l$;
27 $\quad\quad p_r' = p_r$;
28 **return** $L' \leftarrow L \cup A$; // add the list of artificial points in L

2) Void Filling in Close-to-Nadir Region: The above algorithm can successfully compensate the data voids found at the swath edges; however, it is unable to cater those water bodies that are located within the close-to-nadir region, having a smaller

Algorithm 2: Void Filling in Close-to-Nadir Region.

Input : A data strip L with n number of points (p)
Output: A void filled data strip L'

1 Initialize array A, M, T;
2 **for** $i \in n \wedge i \neq n$ **do**
3 $\quad T[i] \leftarrow t_{i+1} - t_i$; $M[i] \leftarrow ||p_i, p_{i+1}||_2$;
4 $t \leftarrow \overline{T}$; $m \leftarrow \overline{M}$; // mean time interval and spacing between two points
5 **for** $i \in n$ **do**
6 $\quad \Delta t \leftarrow t_{i+1} - t_i$;
7 $\quad d \leftarrow ||p_i, p_{i+1}||_2$;
8 $\quad$ **if** $(d > \Omega \cdot m) \wedge (\Delta t > t)$ **then**
9 $\quad\quad J \leftarrow \Delta t/t$;
10 $\quad\quad$ **for** $j \leftarrow 1$ **to** J **do**
11 $\quad\quad\quad x_j \leftarrow x_i + d \cdot j/J$;
12 $\quad\quad\quad y_j \leftarrow y_i + d \cdot j/J$;
13 $\quad\quad\quad A \leftarrow \{x_j, y_j\} \cup A$; // add a new artificial point to A
14 **return** $L' \leftarrow L \cup A$; // add the artificial points in L

dimension than the swath width, e.g., inland rivers, streams, ponds, pools, etc. To cater to this scenario, another void filling algorithm (i.e., algorithm 2) is proposed. The algorithm first starts by looping the entire data and computing the average time interval t and mean 2-D point spacing m between two consecutive points. It is assumed that the entire LiDAR data strip L is sorted with respect to the data collection sequence. Any data void or laser dropout can be identified along the scan line l if the following conditions are fulfilled (see line 8 of Algorithm 2). First, the time difference (Δt) between the two adjacent points (i.e., p_i and p_{i+1}) is greater than t. Second, the 2-D distance (d) between the two adjacent points should be greater than the 2-D mean spacing (m) with a factor of Ω. Additional criteria, such as the two adjacent points have to be the last returns, can also be added.

Such tight conditions can avoid those multiple data returns (e.g., tree branch and ground) having a notable void in 2-D (and time) yet not belonging to the gap caused by a laser dropout. The use of Ω (e.g., $\Omega = 3$) can also ensure that the data void/gap is large enough so that it can avoid any over-insertion of artificial points along the scan line. The proposed approach prefers missing the insertion of an artificial point between the two adjacent points, rather than adding artificial points between the adjacent points that are too close. Once all these conditions are fulfilled, artificial points are inserted between p_i and p_{i+1} through interpolating the xyz coordinates between the two points. The number of artificial points (J) required to be inserted depends on the time difference between these two points. Once the artificial points are added, they are labeled as water (i.e., class 9 according to LAS format).

C. Hydroflattening

As mentioned in Section II-A, the use of SLIER can aid in labeling the data points into either land (class = 2) or water (class = 9). With all those data points being classified as water, the average elevation can be computed so that it can be treated as the virtual water surface. Such value is also assigned as the new elevation for all the artificial points being estimated from Algorithms 1 and 2, and all the data points classified as water in the original data strip. All the artificial points are then combined together with the original LiDAR data strip so as to generate a hydroflattened DEM in TIN data structure. In the following experiments, the hydroflattened DEM is generated using the delaunay triangulation method.

III. Experimental Dataset

The USGS Base Specification [14] specifically mentions that five water and hydrologic environments are required to follow the hydroflattening process. They are as follows:

1) inland ponds and lakes;
2) inland streams and rivers;
3) nontidal boundary water bodies;
4) tidal water bodies;
5) permanent islands.

Since most of the existing studies examine one or two of the abovementioned water environments, the experiment in this study focuses on all the five types of water bodies in order to assess the capability of the proposed workflow and algorithms. Table I summarizes the data configuration of the five datasets. Since these datasets were collected by different LiDAR systems and scanning configurations, they should be sufficiently representative to test the proposed hydroflattening process. Also, the experiments ignored the ground filtering process on the land and instead focused on the hydroflattening on the water bodies.

A. Inland Ponds and Lakes

A total of seven multispectral LiDAR data strips were collected for a research forestry area located at Ontario, Canada. The research forest, i.e., Petawawa Forest, was surveyed by a multispectral airborne LiDAR system, i.e., Optech Titan equipped with three laser channels, where laser channels 1 to 3 are operated with wavelengths of 1550, 1064, and 532 nm, respectively. The inland lake covered in this dataset is the Cartier Lake with an approximate area of 2.7 km^2. The flight mission was accomplished on July 20, 2016. The Optech Titan sensor was operated with scan angle $\pm 20^\circ$, pulse repetition frequency 375 kHz, scan frequency 40 Hz, and flying attitude 1100 m above sea level. The total number of data points in channels 1 to 3 are 1.28, 1.34, and 0.39 billion, resulting in a 78 GB file storage. With these settings, the mean point density yields approximately 11.9, 12.4, and 4.8 points/m^2 for the channel 1, channel 2, and channel 3, respectively. The multispectral LiDAR dataset was used to study the plot level forest attributes, and now the LiDAR data of channel 2 was used to map the inland ponds and lakes.

B. Nontidal Boundary Water Bodies

Great Lakes are commonly treated as nontidal boundary water bodies, since Great Lakes have less effects of water level changes caused by the gravitational forces. In this study, five

TABLE I
SUMMARY OF DATA CONFIGURATION OF FIVE LIDAR DATASETS

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Environment	Inland ponds and lakes	Non-tidal water bodies	Tidal water bodies	Permanent islands	Inland streams and rivers
Location	Ontario, Canada	Tobermory, Canada	Tallahassee, USA	Santorini, Greece	Vaihingen, Germany
Date	July 20, 2016	July 3, 2015	June 12, 2016	May 16, 2012	July 24, 2008
System	Optech Titan	Optech Titan	Optech Titan	Leica ALS50	Leica ALS50
Wavelength	1064 nm	1064 nm	1550 nm	1064 nm	1064 nm
PRF	375 kHz	225 kHz	200 kHz	94.7 kHz	Nil
SF	40 Hz	Nil	35 Hz	58.2 Hz	Nil
Flying Speed	Nil	140 knots	140 knots	135 knots	Nil
Flying Height	1,100 m	500 m	375 m	1,100 m	500 m
MPD	∼12.4 pts/m^2	∼10 pts/m^2	∼5 pts/m^2	∼5 pts/m^2	∼7.6 pts/m^2
FOV	40°	40°	44°	24°	45°
Data Strips	7	5	3	12	5
Area	14.818 km^2	3.691 km^2	0.737 km^2	12.516 km^2	0.931 km^2
Number of Points	∼79.3 million	∼51.3 million	∼3.8 million	∼35.2 million	∼6.6 million
Projection	UTM 18N	UTM 18N	UTM 17N	UTM 35N	UTM 32N

*FOV = Field of View, MPD = Mean Point Density, PRF = Pulse Repetition Frequency, and SF = Scan Frequency.

parallel airborne LiDAR data strips collected nearby the shore of Lake Huron, Tobermory, Ontario, Canada were chosen as a testing dataset for nontidal water bodies. The LiDAR data were collected by an Optech Tian with pulse repetition frequency of 225 kHz, flying speed 140 knots, FOV 40° and flying height of approximately 500 m, resulting in a mean point density of 10 points/m^2. The dataset has been reported in [12], which was originally used to test the capability of SLIER. The land-water classification accuracy consistently yielded over 98%, regardless of the use of different multispectral LiDAR data channels and their derived feature sets. Specifically, the south-west area of the data strips, covering a rocky shore, is intentionally selected to examine the scan line void filling algorithms, because notable sizes of voids can be found in the close-to-nadir region and at the swath edges of all the five airborne LiDAR data strips.

C. Tidal Water Bodies

The study area is located at the northeastern region of the Gulf of Mexico and south of Tallahassee, Florida, USA, covering a shore region of Apalachee Bay. Three parallel airborne LiDAR data strips were collected by Optech Titan on June 12, 2016. Specifically, the dataset covers the Peler Creek channel, and the open water region to the Apalachee Bay. The flying height, flying speed, pulse repetition frequency, and scan frequency are 372 m, 140 knots, 200 kHz, and 35 Hz, respectively. The study area has a mild elevation difference between the shore and the water region; it thus raises certain challenges in delineating the virtual water surface. Although Optech Titan operates with three laser channels, the dataset collected by the 1550-nm laser channel is used since the previous two experimental datasets have already adopted the LiDAR point clouds collected by the 1064-nm laser channel.

D. Permanent Islands

Regarding the scenario of permanent islands, a free LiDAR dataset was acquired from figshare [21] covering the Kameni islands (Nea Lameni and Palea Kameni), located at Santorini, Greece. According to [22], the airborne LiDAR dataset was collected on May 16, 2012 by a Leica ALS50 system with a flying height of approximately 1100 m, pulse repetition frequency 94.7 kHz, and scan frequency 58.2 Hz. The final dataset has more than 40 million data points, collected by 12 data strips and a mean point spacing of approximately 5 points/m^2. This dataset was intentionally collected to investigate the submarine volcanic structure through analyzing the historic lava flows using a fusion of the LiDAR and bathymetry approach [22]. Since the two distinct islands have a total area approximately 3.9 km^2, this dataset thus fulfills the requirement as listed in the USGS LiDAR Base Specification, where hydroflattening process is required for permanent islands with an area of 0.4 ha (or 72 m in diameter) or larger.

E. Inland Streams and Rivers

This dataset includes five LiDAR data strips collected by Leica ALS50 on July 24, 2008. It is also a ISPRS bench mark dataset provided for object detection and 3-D building reconstruction [23]. The study area is located in Vaihingen, Germany, and includes the river Enz. The width of the river is greater than 35 m, and thus, it fulfills the requirement of hydroflattening as stated in the USGS LiDAR Base Specification [14]. The entire study area is within 1 km^2 with an approximately 6.6 million points. The flying altitude is 500 m with a FOV coverage of 45°, and the mean point spacing is 7.6 points/m^2. Within this dataset, the entire river Enz had a total absence of laser returns. Therefore, the scan line void filling was directly implemented without computing SLIER for land-water classification.

IV. RESULTS AND ANALYSIS

A. Extraction of Water Data Points

All the datasets underwent the process of SLIER computation, except dataset 5, due to a lack of LiDAR data points found along the river Enz. Therefore, dataset 5 was implemented with the void filling algorithm directly to add artificial points found along the river channel. For datasets 1 to 4, all of them were classified into either land (class label = 2) or water (class label = 9) after computing the SLIER. Fig. 5 shows the elevation of the labeled LiDAR data points after computing SLIER, and the boxes in each subfigures represent the interquartile range (IQR) of the data (i.e., 25th to 75th percentiles) of either land

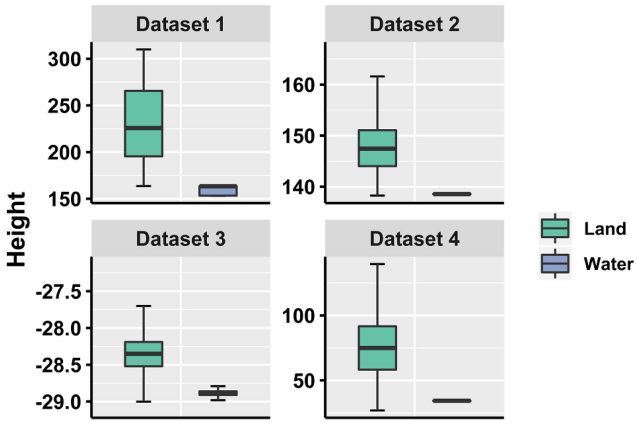

Fig. 5. Elevation of classified LiDAR point clouds (land versus water) in the first four datasets.

TABLE II
EXTRACTION OF WATER DATA POINTS: OVERALL ACCURACY (OA), RECALL, AND PRECISION

Data	OA	Recall	Precision
Dataset 1 (Lake)	98.05%	98.92%	73.83%
Dataset 2 (Non-tidal)	99.36%	99.56%	98.06%
Dataset 3 (Tidal)	98.33%	77.73%	99.48%
Dataset 4 (Islands)	98.45%	97.79%	99.43%
Dataset 5 (River)	Nil	Nil	Nil

or water. The elevation difference between the two classes is obvious. A large IQR can be found in the land features in the four datasets, and a small IQR can be observed on the water bodies in terms of the height. For instance, a steep terrain situated next to the Cartier Lake has an elevation difference of approximately 100 m in dataset 1, which causes a large IQR found on the land's elevation. Among the four datasets, the IQR boxes between the two land and water do not have any overlap. This indicates that the water bodies have an obvious height difference comparing to the land in the respective study area. As a result, automatic extraction of water bodies can be achieved through land-water classification without considering those special cases, such as land depression as reported in [11].

Accuracy assessment was conducted on the classification results derived based on a set of ground truth dataset which was digitized with reference to the aerial photos and the LiDAR datasets itself. Through intersecting the classification results with the ground truth reference dataset, the overall accuracy, recall, and precision were computed through the confusion matrix. Table II shows the results of accuracy assessment. The overall accuracy of all the four datasets was higher than 98%, similar to the experiments reported in [11] and [12] that adopted the SLIER approach. Although the Datasets 1 and 3 reported to have a recall and precision of 73.83% and 77.73%, respectively, the rest of the experiments all produced an accurate extraction of water bodies with close to or higher than 98% in terms of these two values.

Figs. 6 and 7 show the labeled land and water data points after classification in datasets 1 to 4. It is obvious that data voids are found to be particularly serious on the water bodies in datasets 1 to 3 (i.e., inland lake, nontidal water bodies, and tidal water bodies) in the close-to-nadir region and at the swath edges. With an accurate extraction of water data points, it facilitates the subsequent process of void filling where the data voids are assumed to appear on the water bodies.

B. Void Filling With Artificial Points

Table III summarizes the number of artificial points being generated to fill the voids in the close-to-nadir region and at the swath edges after implementing the proposed void filling algorithms on the five datasets. Depending on the instantaneous water conditions, different sizes of void regions appear in each of the datasets. Thus, these affect the number of artificial points being added both in the close-to-nadir region and at the swath edges.

Dataset 3 (i.e., tidal water body) has significant data voids among the five datasets. As a result, an additional 30.58% and 9.09% of artificial points were added in the close-to-nadir region and at the swath edges, respectively, in the original LiDAR data strips to fill the voids caused by laser dropouts. As shown in Fig. 6, the void filling algorithms successfully generate the artificial points along the open water region and small inland streams in the Apalachee Bay, Gulf of Mexico.

Similar to dataset 3, the rest of the four datasets all have more artificial points created in the close-to-nadir region than at the swath edges. For instance, datasets 1, 2, and 3 had more than or close to 3-million artificial points generated in the close-to-nadir region, resulting in approximately 4.35%, 6.2%, and 8.07% of their respective original LiDAR data points. All these three datasets had only less than 1% of artificial points filled at both of the swath edges. In dataset 5, there was an absence of LiDAR data points found along the river channel. The entire river body was filled with an additional 5.12% and 0.13% of artificial points in the close-to-nadir region and at the swath edges.

Regardless of the experimental dataset, most of the artificial points were added in the close-to-nadir region, and less than 1% of artificial points were generated in both side of the swath edges, except for dataset 3. In this particular case, the tidal water body along the shore of Apalachee Bay had notable voids caused by laser dropouts. As a result, the proposed algorithms added an additional 40% of artificial points in the close-to-nadir region and at the swath edges.

C. Effects of Triangular Facets

As those large triangular facets on the DEM are one of the drawbacks caused by the data voids, the statistics of the triangles located on the water regions were computed before and after the hydroflattening in each dataset (see Table IV). Since the two void filling algorithms added artificial points on the water region in the original point clouds, this resulted in a significant increase of number of triangles after the hydroflattening process.

In dataset 1 (Lake), the number of triangles increased from 418 276 to 793 778, resulting in a 89.77% of addition triangles covering the voids found within the Cartier Lake. This also led to a reduced mean area of triangles from 7.091 to 3.965 m^2 ($\downarrow$ 44.08%) covering the lake. The corresponding standard

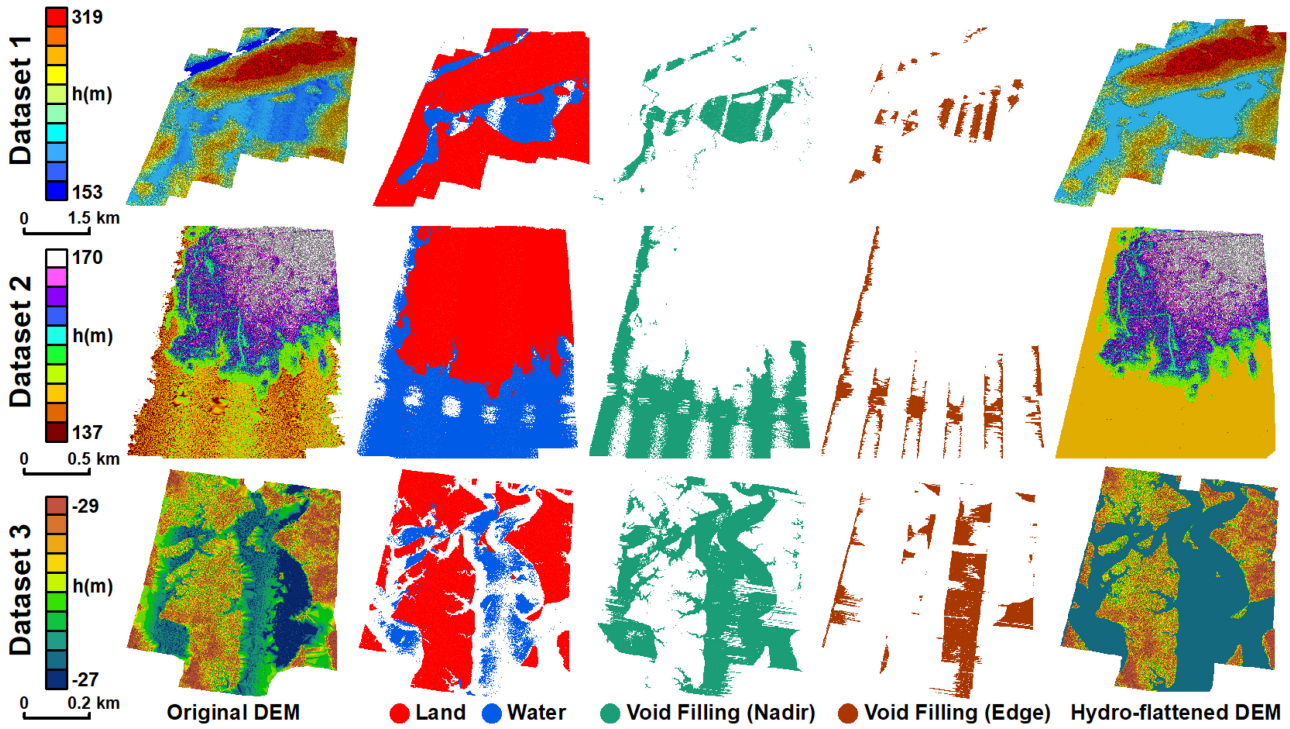

Fig. 6. Dataset 1 (Inland lake), 2 (nontidal water body) and 3 (tidal water body) (From left to right): Original DEM, result of land-water classification, void filled in close-to-nadir region, void filled at swath edges, and the hydroflattened DEM.

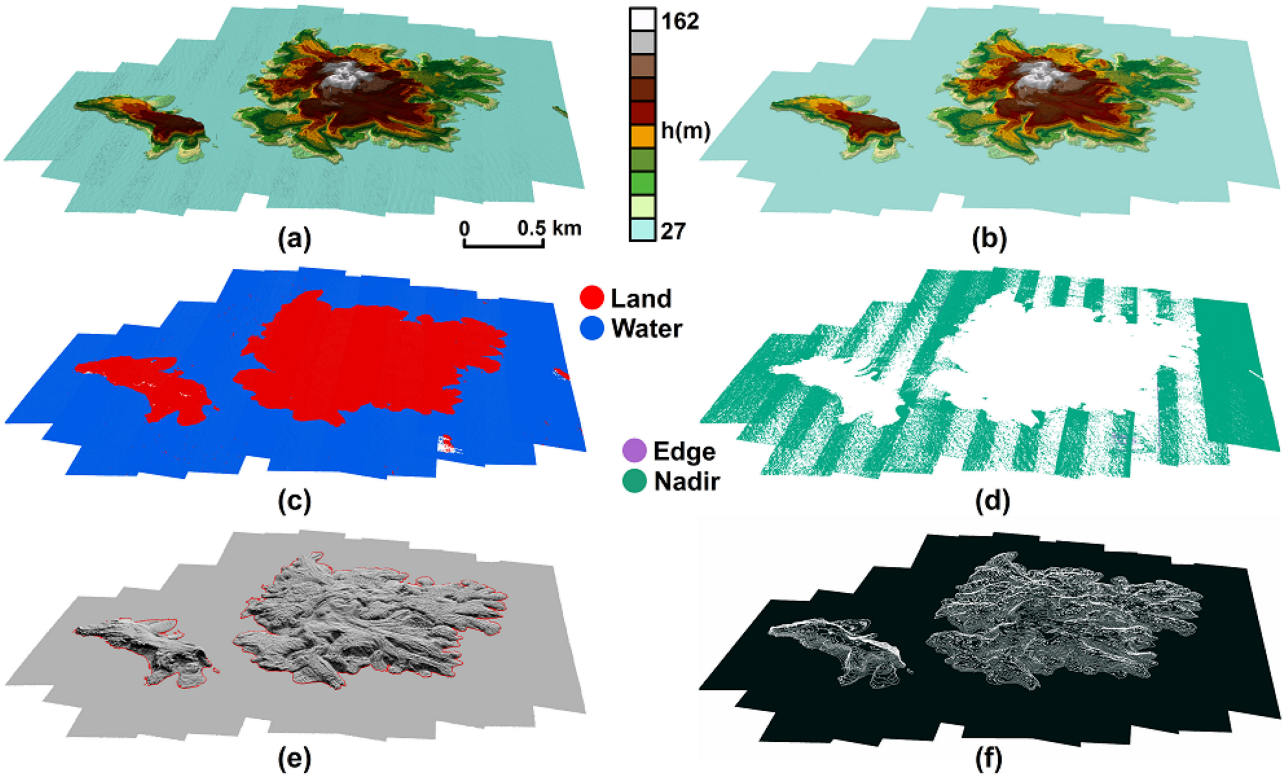

Fig. 7. Dataset 4 (Permanent Island). (a) Original DEM, (b) hydroflattened DEM, (c) classification result, (d) void filling result, (e) extracted shoreline (in red), and (f) extracted contour lines (5-m interval).

TABLE III
STATISTICS OF ORIGINAL LIDAR DATA AND ARTIFICIAL POINTS ADDED AFTER VOID FILLING

Data	Original data	Num of artificial points		% of artificial points added	
		Close-to-nadir	Swath edges	Close-to-nadir	Swath edges
Dataset 1 (Lake)	79,257,099	3,448,123	638,367	4.35%	0.81%
Dataset 2 (Non-tidal)	51,264,611	3,176,379	178,264	6.20%	0.35%
Dataset 3 (Tidal)	3,788,320	1,158,378	344,422	30.58%	9.09%
Dataset 4 (Islands)	35,210,238	2,839,771	58,644	8.07%	0.17%
Dataset 5 (River)	6,635,530	339,721	8,798	5.12%	0.13%

TABLE IV
NUMBER AND AREA (MEAN AND STANDARD DEVIATION) OF TRIANGLES COVERING THE WATER BODIES BEFORE AND AFTER HYDROFLATTENING

Data	Number of Triangles		Mean Area (m^2)		Standard Deviation (m^2)	
	Before	After	Before	After	Before	After
Dataset 1 (Lake)	418,276	793,778	7.091	3.965	40.267	5.583
Dataset 2 (Non-tidal)	415,470	681,224	2.533	1.611	8.211	1.716
Dataset 3 (Tidal)	613,660	3,588,313	0.593	0.101	8.572	0.097
Dataset 4 (Islands)	1,956,723	2,237,469	4.166	3.644	4.936	4.304
Dataset 5 (River)	12,091	745,549	5.588	0.098	22.672	0.074

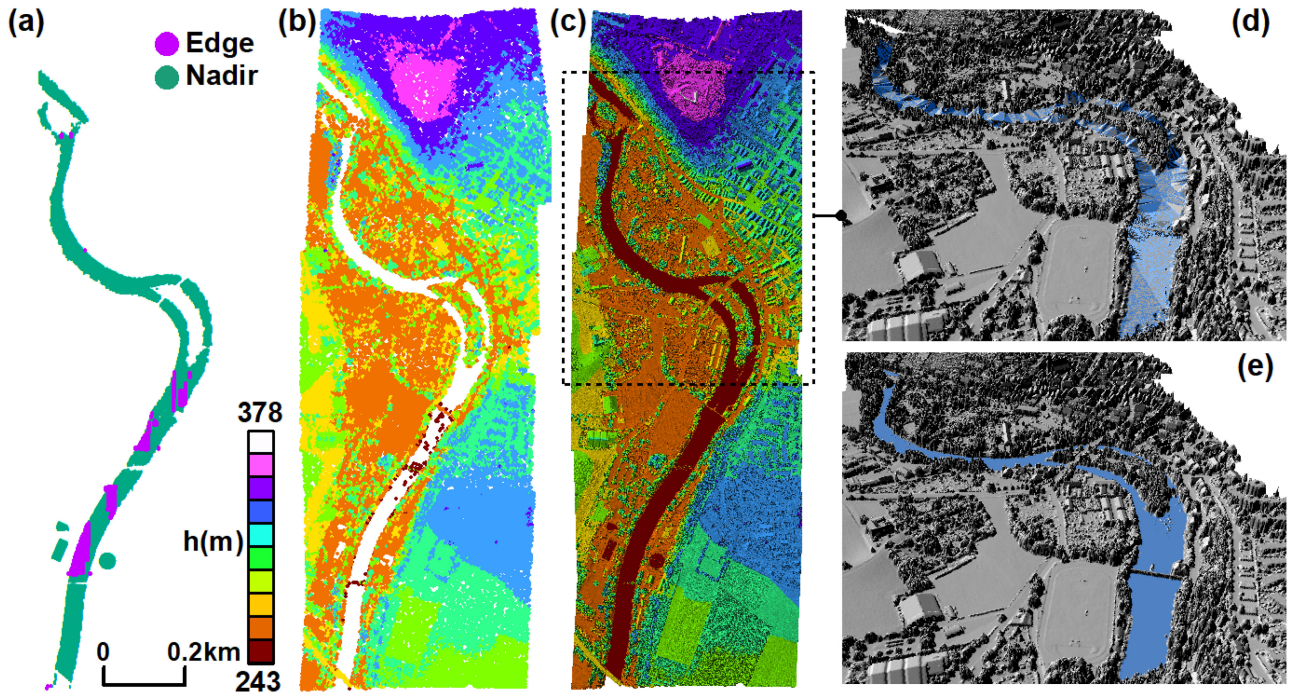

Fig. 8. Dataset 5 (Inland river). (a) Void filling result, (b) original LiDAR data, (c) hydroflattened DEM, (d) 3-D view of the original DEM with unpleasant triangular artifacts along the river, (e) 3-D view of the hydroflattened DEM.

deviation also significantly reduced from 40.267 to 5.583 m^2 (↓ 86.14%). In datasets 2 and 4, the effects of laser dropouts were not obvious. Therefore, the proposed algorithms only added additional 63.96% and 14.35% of triangles on the water regions on Lake Huron (nontidal water bodies) and nearby the Kameni islands, respectively.

In the case study of dataset 2, the mean area of triangles and the corresponding standard deviation decreased by 36.4% and 79.1% after the hydroflattening. On the other hand, the datasets 3 and 5 both have notable data voids (see Figs. 6 and 8), therefore, the number of triangles covering the water region raised 4.8 times in the tidal water body along the shore of Apalachee Bay and 60.6 times along the river Enz after implementing the proposed algorithms. The mean area of triangles covering the water regions in these two cases reduced to close to 0.1 m^2, i.e., a respective 82.97% and 98.25% reduction. The standard deviation of the triangles' area found on these water regions also largely decreased in datasets 3 and 5, from 8.572 to 0.097 and 22.672 to 0.074, respectively.

In short, the number of triangles in the resulted TIN found on the water regions increase in the five case studies after adding the artificial points. The average size of the triangles and the

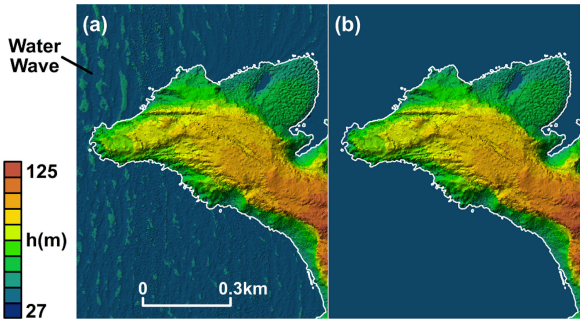

Fig. 9. Dataset 4 (Permanent Island). (a) Original DEM with water wave effect on the TIN, and (b) hydroflattened DEM.

respective standard deviation all significantly reduce. This can statistically prove that those unpleasant triangular facets found on the original DEM should be resolved after the hydroflattening process.

D. Visual Inspection

It is obvious from visual inspection that those huge triangular facets found on the original DEM are removed after the hydroflattening process. For instance, in Fig. 8(d), the river Enz in dataset 5 was covered by unnatural triangular artifacts in the original DEM. This was attributed to the lack of water data points [see Fig. 8(b)] causing the delaunay triangulation generated based on the data points found along the river bank or shore only. After filling with artificial points in the close-to-nadir region and at the swath edges [see Fig. 8(a)], the river was represented by a flat and level water surface [see Fig. 8(e)]. As shown in Fig. 8(c), the elevation of land found on both sides of the river was considerably consistent; therefore, a gradient downhill/downward water surface was not applied.

Apart from removing those huge and unwanted triangular facets, the hydroflattening process also aids in smoothing the water surface. For instance, in dataset 4, the original DEM had an obvious effect of ocean ripple and water wave in the Mediterranean Sea, [see Fig. 9(a)]. The hydroflattening process completely removed all these small undulations by assigning a flat, virtual water surface [see Fig. 9(b)]. Also, the hydroflattening process can help improve the hydrological connectivity of the terrain and extraction of complex shoreline profiles and topographic products. Fig. 7(e) shows the extracted shorelines of the Nea Lameni and Palea Kameni islands. They were derived by fitting a convex hull on the labeled land data points with slight manual adjustment. Fig. 7(f) shows the contour lines generated on the labeled land data points of the respective islands in Greece. Other examples of shorelines can be found in Fig. 10. Fig. 10(a) and (b) shows the 3-D view of datasets 1 (lake) and 3 (tidal water body), respectively, suffering from the huge unnatural triangular facets in the original TIN. After hydroflattening, the water surface is leveled to a flat elevation and the shorelines of the respective islands/marsh can be identified.

V. Discussion and Limitations

Although the above five reported case studies demonstrate the void filling process, leading to a successful generation of hydroflattened DEMs, the proposed approach is built based on a number of assumptions that should be clearly stated. First, the scan line void filling is tailor made for linear-mode topographic airborne LiDAR systems operated with an oscillating mirror or a rotating prism. Therefore, the proposed approach is not applicable for photon-counting LiDAR or airborne laser bathymetry operated with Fresnel wedge. Second, the floating points backscattered in the atmosphere indeed affect the interpolation process of the artificial points. Therefore, a 3-D spatial filter should be implemented prior to the void filing process so as to remove noisy data returns. Also, both void filling algorithms assume that the data voids are caused by the appearance of water bodies. However, objects with low spectral reflectance (e.g., asphalt road or building roof) may also result in null returns. Therefore, end users can select an appropriate value of Ω in Algorithm 2 to identify only those large voids found on the water regions.

From an operational perspective, the following procedures should be followed. Since the void filling mechanism requires to first estimate the swath width and the corresponding edge points; therefore, the LiDAR data strip should have the first couple scan lines covering the land region before heading to the water region. It is also recommended to sort the LiDAR data in a sequence based on the GPS time prior to the data processing so as to make sure the interpolation of void filling works properly (see Algorithm 2). One may notice the void filled LiDAR data strip may not result in straight swath edges after running the Algorithm 1, yielding a zig-zag pattern. This can be attributed to the imprecise estimation, or in some cases missing, of the nadir point p_c that is derived based on the mean position of data points with scan angle equal to zero. Since most of the LiDAR data strips may not bundle with the trajectory information, the proposed void filling algorithm at swath edges thus adopts a data-driven approach to estimate the nadir points. If the LiDAR data are provided with the trajectory file, the p_c can be replaced by the using the trajectory file. In this case, the precision of determining the data points of the swath edges should be improved. The proposed algorithms assume the data voids are located with adjacent data points being labeled as water (see Fig. 4). Nevertheless, this is not always the case, particularly in datasets 3 and 5 (see Fig. 6 and 8), where data voids may be found between two distinct pieces of land without any water data point in between. Therefore, the algorithm should offer an option to let end users to select whether the algorithms should generate artificial points if the previous/next data point is water or not. Also, the assumption may not stand if the data voids are caused by the occlusion/shadow of cloud or smoke between the aircraft and the topography.

VI. Conclusion

This study presents two algorithms for automatic scan line void filling of topographic airborne LiDAR data at swath edges

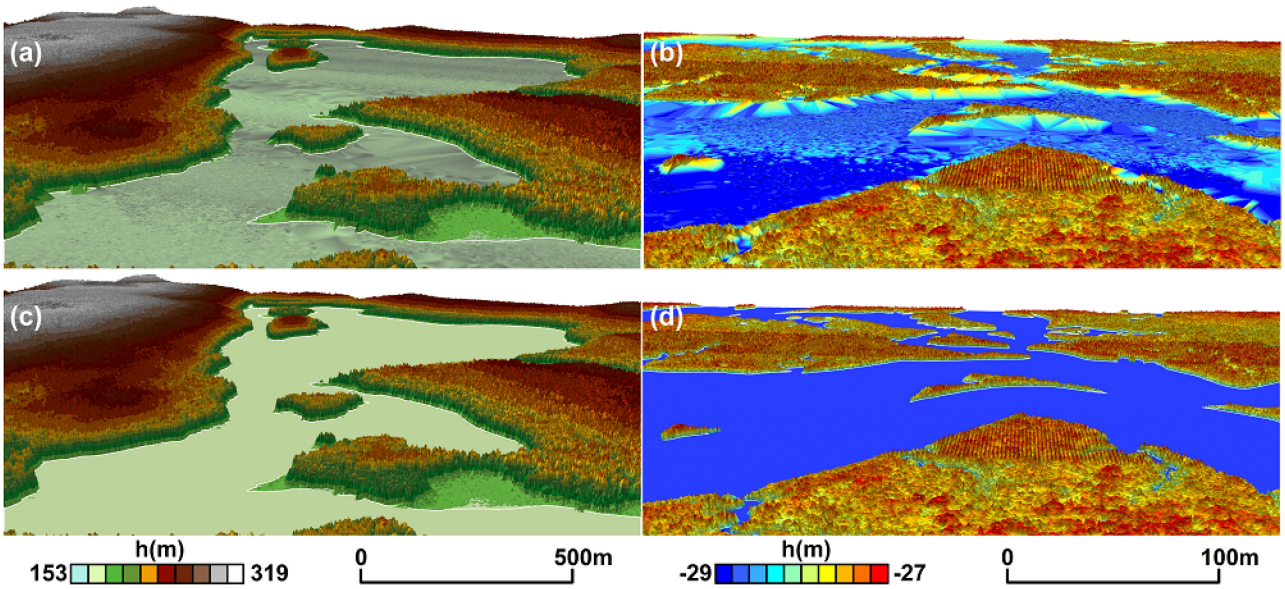

Fig. 10. Original DEM with unpleasant triangular artifacts in (a) dataset 1 and (b) dataset 3, and (c) and (d) hydroflattened DEM of the respective dataset.

and in close-to-nadir region, with the overall objective of facilitating hydroflattening DEM. The proposed algorithms were examined on five types of water and hydrologic environments, including inland ponds and lakes, inland streams and rivers, nontidal water bodies, tidal water bodies, and permanent islands, which are highlighted in the USGS LiDAR Base Specification. The proposed approach overcomes the drawbacks of existing approaches and does not require any ancillary data (e.g., tidal data, shoreline, etc.) for extraction of water bodies and hydroflattening.

Based on the experimental tests, the use of SLIER first yielded an accurate extraction of water data points with up to 98% in terms of overall accuracy. Once the water data points were determined, the two void filling algorithms added 1%–9% of artificial points at swath edges and 4%–31% of artificial points in close-to-nadir region within the water regions. These artificial points together with the labeled water data points were combined to generate a virtual water surface with a flat elevation for hydroflattening. The average area of resulted TIN triangles covering the water region reduced by 13%–98% and the corresponding standard deviation also decreased by 13%–99%. Those unnatural, huge triangular facets were no longer evident after the hydroflattening process. Also, the water waves found on the open water bodies were also removed in the resulting DEM. As a beneficial by-product, the proposed workflow also aids in generating the shore lines and contour lines, as demonstrated in the case study of permanent islands.

Detection of data voids based on the use of time difference and point offset may not be sufficiently robust. Future work can explore the use of persistent homology [24] to locate the data voids through analyzing the simplicial complex at different spatial scales. In addition, the hydroflattening process adopted in this study assumes a uniform elevation of the water points. Therefore, further effort should be made to retain the gradient information of water elevation, such as the downstream of a river reach. An improved terrain sampling strategy can also be implemented to reduce the number of artificial points representing the water surface while retaining the hydroflattening effect.

Acknowledgment

The author would like to thank all the below parties for their support and providing valuable datasets. The LiDAR dataset 1 was obtained through the AWARE project (NSERC File: CRDPJ 462973 - 14, Grantee: N.C. Coops, FRM, UBC), in collaboration with Canadian Wood Fibre Centre (CWFC), FP-Innovations, and Tembec. The dataset 5 was a ISPRS benchmark dataset provided by Prof. F. Rottensteiner. The author would like to thank Dr. K. van Ewijk and Prof. P. Treitz for providing the dataset, and Prof. A. Shaker, Mr. V. Ramnath, Dr. P. E. LaRocque, and Dr. A. P. Kersting for the discussion of the research problem. The author also would like to thank Mr. L. Kwan for his assistance in data preprocessing and Dr. E. Ho for proofreading the manuscript.

References

[1] X. Liu, "Airborne LiDAR for DEM generation: Some critical issues," *Prog. Phys. Geography*, vol. 32, no. 1, pp. 31–49, 2008.
[2] C. Arrighi, M. Brugioni, F. Castelli, S. Franceschini, and B. Mazzanti, "Urban micro-scale flood risk estimation with parsimonious hydraulic modelling and census data," *Natural Hazards Earth Syst. Sci.*, vol. 13, no. 5, pp. 1375–1391, 2013.
[3] J. S. Deems, T. H. Painter, D. C. Finnegan, "Lidar measurement of snow depth: A review," *J. Glaciol.*, vol. 59, no. 215, pp. 467–479, 2013.
[4] P. M. Biron, G. Choné, T. Buffin-Bélanger, S. Demers, and T. Olsen, "Improvement of streams hydro-geomorphological assessment using LiDAR DEMs," *Earth Surface Processes Landforms*, vol. 38, no. 15, pp. 1808–1821, 2013.
[5] W. Zhang *et al.*, "An easy-to-use airborne LiDAR data filtering method based on cloth simulation," *Remote Sens.*, vol. 8, no. 6, 2016, Art. no. 501.
[6] G. Sithole and G. Vosselman, "Experimental comparison of filter algorithms for bare-earth extraction from airborne laser scanning point clouds," *ISPRS J. Photogrammetry Remote Sens.*, vol. 59, no. 1/2, pp. 85–101, 2004.

[7] X. Meng, N. Currit, and K. Zhao, "Ground filtering algorithms for airborne LiDAR data: A review of critical issues," *Remote Sens.*, vol. 2, no. 3, pp. 833–860, 2010.
[8] W. Y. Yan, A. Shaker, and N. El-Ashmawy, "Urban land cover classification using airborne LiDAR data: A review," *Remote Sens. Environ.*, vol. 158, pp. 295–310, 2015.
[9] Z. Chen, B. Gao, and B. Devereux, "State-of-the-art: DTM generation using airborne LIDAR data," *Sensors*, vol. 17, no. 1, 2017, Art. no. 150.
[10] S. Tamari, J. Mory, and V. Guerrero-Meza, "Testing a near-infrared Lidar mounted with a large incidence angle to monitor the water level of turbid reservoirs," *ISPRS J. Photogrammetry Remote Sens.*, vol. 66, no. 6, pp. S 85–S91, 2011.
[11] A. Shaker, W. Y. Yan, and P. E. LaRocque, "Automatic land-water classification using multispectral airborne LiDAR data for near-shore and river environments," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 94–108, 2019.
[12] W. Y. Yan, A. Shaker, and P. E. LaRocque, "Scan line intensity-elevation ratio (SLIER): An airborne LiDAR ratio index for automatic water surface mapping," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 814.
[13] B. B. Worstell, S. Poppenga, G. A. Evans, and S. Prince, "Lidar point density analysis: Implications for identifying water bodies," U. S. Geological Surv., Reston, VA, USA, Sci. Investigations Rep. 2014-5191, 2014.
[14] H. K. Heidemann, "Lidar Base Specification version 1.3 (ver. 1.3, February 2018)," in U. S. *Geological Surv. Techn. Methods*, *Book* 11, 2018, ch. B4, p. 101.
[15] L. Shea Rose *et al.*, "Challenges and lessons from a wetland LiDAR project: A case study of the Okefenokee swamp, Georgia, USA," *Geocarto Int.*, vol. 28, no. 3, pp. 210–226, 2013.
[16] L. P. Rampi, J. F. Knight, and C. F. Lenhart, "Comparison of flow direction algorithms in the application of the CTI for mapping wetlands in minnesota," *Wetlands*, vol. 34, no. 3, pp. 513–525, 2014.
[17] J. B. Coleman, X. Yao, T. R. Jordan, and M. Madden, "Holes in the ocean: Filling voids in bathymetric Lidar data," *Comput. Geosci.*, vol. 37, no. 4, pp. 474–484, 2011.
[18] S. Deshpande and A. Yilmaz, "A semi-automated method to create a LiDAR-based hydro-flattened DEM," *Int. J. Remote Sens.*, vol. 38, no. 5, pp. 1365–1387, 2017.
[19] X. Zheng, L. Godbout, J. Zheng, C. McCormick, and P. Passalacqua, "An automatic and objective approach to hydro-flatten high resolution topographic data," *Environ. Model. Softw.*, vol. 116, pp. 72–86, 2019.
[20] J. A. Czuba, S. R. David, D. A. Edmonds, and A. S. Ward, "Dynamics of surface-water connectivity in a low-gradient meandering river floodplain," *Water Resour. Res.*, vol. 55, no. 3, pp. 1849–1870, 2019.
[21] D. Pyle, M. Parks, T. Mather, and P. Nomikou, "2012 Santorini LiDAR Data," 2014. Accessed: Jul. 17, 2019. [Online]. Available: http://dx.doi.org/10.6084/m9.figshare.1138718
[22] P. Nomikou *et al.*, "The emergence and growth of a submarine volcano: *Kameni islands*, Santorini (Greece)," *GeoResJ*, vol. 1, pp. 8–18, 2014.
[23] F. Rottensteiner, G. Sohn, M. Gerke, J. D. Wegner, U. Breitkopf, and J. Jung, "Results of the ISPRS benchmark on urban object detection and 3D building reconstruction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 93, pp. 256–271, 2014.
[24] A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete Comput. Geometry*, vol. 33, no. 2, pp. 249–274, 2005.

Wai Yeung Yan (Senior Member, IEEE) received the Ph.D. degree in civil engineering from Ryerson University, Toronto, ON, Canada, in 2012.

He is currently an Assistant Professor with the Department of Land Surveying and Geo-Informatics, the Hong Kong Polytechnic University, Hong Kong, and an Adjunct Professor with the Department of Civil Engineering, Ryerson University. His research interests include improving LiDAR data quality, point cloud processing, laser scanning, and remote sensing.