Title:
An Automatic Method for Subglacial Lake Detection in Ice Sheet Radar Sounder Data

This paper appears in: IEEE Transactions on Geoscience and Remote Sensing
Authors: Ana-Maria Ilisei, Mahdi Khodadadzadeh, Adamo Ferro, Lorenzo Bruzzone
Publication date: 27 December 2018
DOI: 10.1109/TGRS.2018.2882911

An Automatic Method for Subglacial Lake Detection in Ice Sheet Radar Sounder Data

Ana-Maria Ilisei, Mahdi Khodadadzadeh, Adamo Ferro, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—During the past decades, radar sounder (RS) instruments have been effectively used to detect subglacial lakes (SLs). SLs appear as flat, smooth and bright reflectors in RS radargrams. The visual interpretation has been the main approach to SL detection in radargrams. However, this approach is subjective and inappropriate for processing large amounts of radargrams. While the analysis of RS data for understanding the subglacial hydrology has recently received increased attention, the literature on the development of automatic methods specifically designed for SL detection is still limited. In order to fill this gap, in this paper we propose a novel automatic technique for SL detection. The technique is made up of two steps, i.e., i) feature extraction, and ii) automatic detection. In the first step, we define and extract three families of features for discriminating between lake and non-lake radar reflections. The features model locally the basal topography, the shape of the basal reflected waveforms, and the statistical properties of the basal signal. In the second step, we provide the features as input to a support vector machine classifier (SVM) to perform the automatic SL detection. The proposed technique has been applied to radargrams acquired over two large regions in East Antarctica and Siple Coast. The obtained results, which are validated both quantitatively and qualitatively, confirm the robustness of the features and their capabilities to effectively characterize SLs. Moreover, they prove the potentiality of the method to process large amounts of radargrams and update the current SL inventory.**

***Index Terms*—subglacial lakes, ice sheet, radar sounder, automatic detection, remote sensing,**

I. Introduction

DURING the past half century, the identification of subglacial lakes (SLs) has been a matter of great scientific interest [1–9]. This interest is motivated by the key role of SLs in glaciology, e.g., SLs constrain geothermal flux [6], affect the dynamics and evolution of the ice sheet [10], represent a potential (although very extreme) habitat for microbial life [11, 12], and may contain ancient climate records [13–15].

The fourth and most recent SL inventory in Antarctica, which is dated 2012 [16], reports 379 lakes, of which ≈30% have been detected by analyzing ice surface elevation changes in altimeter data [7]. Such lakes are also called active, to highlight their observed dynamic behavior as a total or partial periodical discharge of their water [14]. The remaining ≈70% of inventoried Antarctic SLs have been detected in data acquired by airborne radar sounder (RS) instruments [15]. Recent analyses of RS data have also evidenced the presence of two SLs in Greenland [9] and a hypersaline SL complex in the Canadian Arctic [17]. Hereafter, we will focus our attention on the detection of SLs in RS data.

RSs are nadir-pointing instruments specifically designed for imaging the ice-sheet cross-section. They emit low-frequency electromagnetic waves and measure the power reflected by subsurface mechanical and thermal discontinuities, from the surface to the basal interface below the RS platform along a predefined path. These measurements are recorded in 2D matrices called radargrams. So far, for monitoring the Earth polar regions, glaciers and ice caps, RSs have been operated in dedicated airborne campaigns, thus providing a large amount of radargrams. This offered the possibility to study the basal conditions and identify SLs over wide areas. In the future, the amount of RS data is expected to drastically increase, since currently there are ongoing studies for the design of RS instruments for Earth Observation from space (e.g., [18]). Such missions could enable the detection of other SLs in areas unexplored during past and present RS surveys.

The detection of SLs in RS data has been mainly carried out by visual analysis (e.g., [1], [3], [5]). While this approach is effective in qualitatively describing the features of SL radar signatures (e.g., radar coherence, brightness, flatness, smoothness [19]), it is subjective and time consuming, thus unsuitable for the detection of SLs on large RS datasets. This motivated the development of automatic techniques for SL detection. As a first attempt, in [6] the authors present an automatic technique for SL detection and classification that uses the hydraulic flatness condition [20] to identify candidate SL interfaces. The candidate interfaces are classified into four SL classes, i.e., definite, dim, fuzzy and indistinct, by using a 2-step approach. In the first step, the basal reflection coefficients are estimated by inverting the radar equation assuming literature subsurface attenuation models. In the second step, the SL candidates are classified by imposing different constraints (thresholds) on their specularity and brightness (absolute or relative to the surroundings). Several other studies have greatly contributed to our understanding of the basal conditions and subglacial hydrology. They focused on the development of techniques for the analysis of the basal interface, the discrimination between dry and thawed interfaces and the modeling of the ice sheets. For instance, the technique presented in [21] relies on subsurface attenuation estimates, the study of the radar waveforms and their statistical characterization. A technique based on manual digitization and reflectivity analysis, derived from attenuation- and path-corrected bed echo power, is presented in [22]. In [23], a method for estimating the subglacial water geometry, independently on basal depth and subsurface attenuation, is described. It uses radar bed echo specularity derived from focusing the RS data with two different antenna apertures. In [24], this method is complemented with the analysis of the bed trailing echo in order to derive the distribution of basal water between Antarctic SLs. The analysis of the basal roughness as a consequence for radar scattering and basal water discrimination is presented in [25].

The literature also presents extensive reviews on the advances in understanding the subglacial environment [26], [27], [14]. An important outcome of such studies is the fact that, in most cases, active lakes are not visible in RS data [27], some exceptions being Lake Whillans [28] and a couple of active lakes in the Byrd glacier catchment [29]. On the other side, due to their static nature as isolated water bodies [14], the majority of lakes visible in RS data do not show changes in ice surface elevation; thus they cannot be detected by analyzing satellite altimeter data. This requires the development of novel automatic methods that can detect SL in RS data in an efficient and objective way.

In this paper we propose a novel automatic technique specifically designed for the detection of SLs in radargrams. The main novelty of the technique with respect to the related literature is the use of a pattern recognition approach based on machine learning to SL detection. This is an advantage with respect to existing methods, since it reduces the amount of human interaction, thus enabling an unbiased, objective and repeatable SL detection also on large amounts of RS data. The method is made up of two main steps, i.e., 1) basal interface feature extraction, and 2) automatic SL detection. The first step is the main contribution of this work and consists in defining and extracting a set of discriminative features of the basal interface for characterizing lake and non-lake interfaces. The definition of the features relies on several characteristics of the basal interface observed and reported in the literature. Based on such observations, we propose extracting three families of features that characterize locally the basal interface, i.e., i) topographic features (which depend on the topographic variations of the basal interface), ii) shape features (which model the shape of radar basal returns), and iii) statistical features (which model the statistical properties of the radar signals reflected by the basal interface). In the second step, the extracted features are given as input to a support vector machine (SVM) classifier [30] to perform the automatic SL detection. SVM is a supervised parametric classifier, meaning that in order to learn the properties of the classes, it requires to input a set of labeled data (in our case lake and non-lake basal samples) along with a set of features that properly describe them. This implies a minimum initial human interaction in the training phase of the SVM. The advantage is that, once the training is performed and the SVM model parameters are estimated, applying the estimated SVM model to classify new data is completely automatic. Moreover, the results obtained automatically are coherent on all data and thus not affected by possible different interpretation as in manual analysis of large data carried out by different scientists. The output of the SVM is analyzed to derive the degree of uncertainty associated with the SL detection. It is worth mentioning that this approach refines, extends and generalizes the method recently proposed in [31], which was our first attempt to automatically detect SLs in RS data.

In order to assess the validity of the proposed method, we applied it to two RS datasets acquired by the MultiChannel Coherent Radar Depth Sounder (MCoRDS) [32]. The two datasets have been acquired over large areas of the Lake District in East Antarctica and Siple Coast in West Antarctica [33].

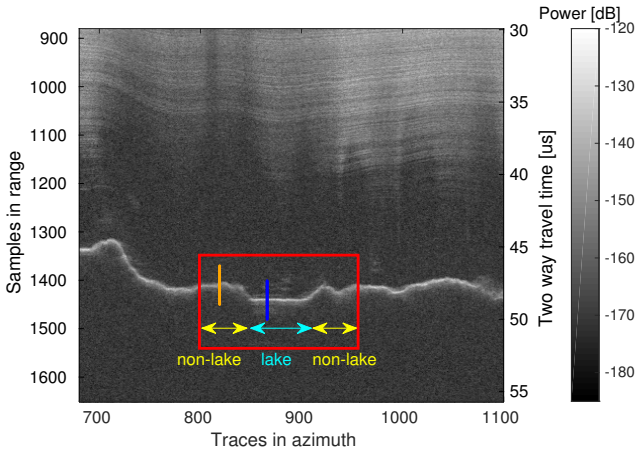


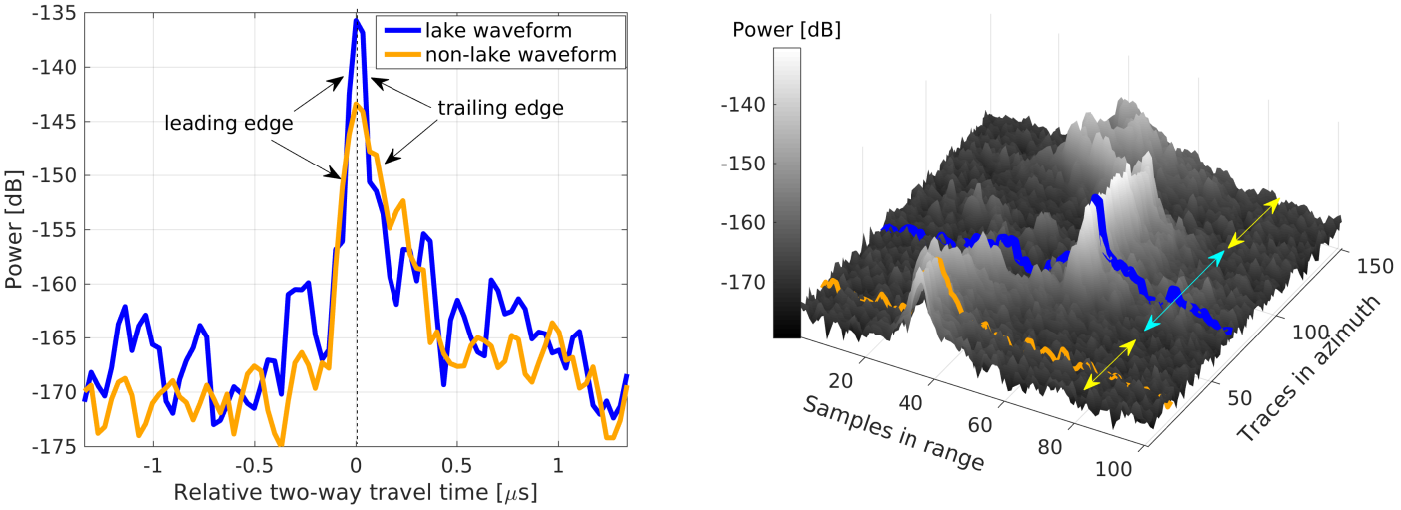
Fig. 1. Portion of radargram showing the ice sheet subsurface. The basal interface is the deepest scattering area, visible at about 45-50μs. In the red rectangle, both lake and non-lake reflections are present. The lake visualized in this radargram is lake number 71 reported in [16]. The blue and orange vertical lines are examples of lake and non-lake reflections, respectively. The waveforms of these reflections are illustrated in Fig. 2(a), whereas the whole region enclosed in the red rectangle is reported in Fig. 2(b). The radargram was acquired by MCoRDS in the East Antarctic Ice Sheet in 2013 [33].

We validated the method both qualitatively and quantitatively according to different metrics. The obtained results prove the effectiveness of the proposed features in characterizing lake interfaces for a wide range of lake depths and the usefulness of the method in discriminating lake from non-lake interfaces.

The remaining of the paper is organized as follows. Section II provides a review of the characteristics of the basal reflections in RS data. The proposed automatic method for SL detection is described in details in Section III. Section IV provides and discusses experimental results obtained by applying the proposed technique to real RS data. Finally, Section V draws the conclusion of this work and proposes ideas for future developments.

II. Basal interface characterization in RS data

A review of the literature points out that the basal interface in ice sheet RS data can be visually identified based on two main properties: its position in the radargram and its reflected power. The basal interface is the deepest subsurface target and induces a higher radar reflection compared to the surroundings (e.g., see Fig. 1). These properties are due to two main reasons: i) a higher dielectric permittivity ε of the basal material (ε_{BI} in the range $\approx [4, 80]$) with respect to the overlaying ice ($\varepsilon_{\text{ice}} \approx 3.15$), and ii) a higher conductivity ς of the basal material (ς_{BI} in the range $\approx [0.01, 3000]\text{mS/m}$) compared to the above ice ($\varsigma_{\text{ice}} \approx 0.01\text{mS/m}$) [34], [35]. On the one hand, $\varepsilon_{BI} \gg \varepsilon_{\text{ice}}$ implies a higher power reflection coefficient at the basal interface compared to the interfaces made by the deepest ice layers. For this reason, the basal interface generally appears brighter than the closest ice layers. On the other hand, $\varsigma_{BI} \gg \varsigma_{\text{ice}}$ implies a higher absorption through the basal material compared to the ice column, which causes a fast power decay as a function of wave travel time [35]. For this reason, the basal returns are typically the deepest subsurface



(a) (b)

Fig. 2. Lake versus non-lake basal interfaces in the range and azimuth directions. (a) The lake and non-lake waveforms are highlighted in Fig. 1 in blue and orange, respectively. The waveforms are aligned with respect to the peak power in order to emphasize the steepness of the lake waveforms compared to that of the non-lake waveforms. (b) 3D view of the power values enclosed in the red rectangle in the radargram shown in Fig. 1.

structures visible in radargrams. While these general properties are common to all kinds of basal interfaces, different types of basal materials (e.g., water, sediments, rock) reflect the radar waves differently, thus showing a radar signature characterized by different properties. Since the aim of this work is the detection of SLs, in the following, we will investigate and compare basal lake interfaces and basal non-lake interfaces.

Qualitative analyses of the basal interface in radargrams acquired over well known SLs (e.g., Vostok lake [19], Horseshoe lake [6], Komsomolskoe lake [5]) point out peculiar characteristics of lake reflections, both in the range and azimuth directions. These characteristics are highly related to the geophysical properties of SL, i.e., topography, dielectric permittivity, and interfacial roughness. As reported in [6], due to the absence of basal shear stress and under the assumption that the water supports all overburden pressure, lakes are exceptionally flat (i.e., lakes have a low topographic variability). The dielectric permittivity of water ($\varepsilon_{\text{water}} = 80$) is greater than that of any other subglacial material (e.g., bedrock, sediments, soil) [35]. Moreover, lakes are perceived as smooth at the wavelength scale [6] (i.e., lakes have an interfacial roughness comparable or larger than the typically used wavelengths λ). The much higher dielectric permittivity along with the smoothness of lake interfaces yield narrow basal waveforms characterized by high peak power, and steep leading and trailing edges, as derived in [21] in the case of wet basal interfaces. This can be seen in Fig. 2(a), which shows an example of lake waveform. Moreover, the flat topography and smoothness of lake interfaces imply that in radargrams lakes appear as flat interfaces characterized by a high degree of correlation on several consecutive traces in the flightline direction. An example of lake interface is shown in Fig. 1 at the azimuth location highlighted in cyan. As expected, in the radargram the lake appears as a flat interface in the azimuth direction. Fig. 2(b) illustrates the 3D view of the portion of radargram enclosed in the red rectangle in Fig. 1, highlighting both the lake waveform steepness in the range direction and its flatness in the azimuth direction. Moreover, the high degree of correlation of lake waveforms on several consecutive traces is also evident in this figure.

The reflections of basal non-lake interfaces are, in general, qualitatively different from those of lake interfaces at similar depths. Compared to lake waveforms, in the range direction non-lake waveforms are usually wider and characterized by lower peak power and moderate leading and trailing edges [see the orange waveform in Fig. 2(a)]. This is due to both the fact that $\varepsilon_{\text{non-water}} \ll \varepsilon_{\text{water}}$ and to the usually higher roughness of non-lake interfaces [36], [6]. Indeed, rough surfaces scatter the radar wave in different directions causing a decrease in peak power and introducing a strong non-coherent component determining moderate waveform edges [21], [37]. The higher roughness and usually more variable topography of non-lake interfaces also affect their azimuth signatures in radargrams. Accordingly, non-lake interfaces show less flat signatures and a lower degree of correlation on consecutive traces compared to lake interfaces [see, e.g., Fig. 1 and Fig. 2(b)]. This holds especially for the interior of East Antarctica, which is characterized by a generally rougher bedrock topography with frozen bed. Indeed, the contrast between lake and non-lake signatures in terms of flatness, correlation on consecutive traces, and waveform steepness may be lower in sedimentary regions (e.g., the Siple Coast in West Antarctica). However, in these areas, the power contrast between ice-water and ice-sediments interfaces can be used for the discrimination of SLs.

In this context, studies performed on radar data also pointed out the direct relationship between interfacial roughness and statistical properties of the reflected radar signal (e.g., [21, 38–40]). In [38], the author studied and discussed the differences in local statistics of reflecting surfaces with different roughness (air/sea, air/ice and ice/water) estimated from RS data acquired

over the Ross Ice Shelf in Antarctica. In [21], the shape of the statistical distribution of the relative intensity of the RS signal has been used as a proxy for validating the presence of dry and thawed subglacial interfaces.

The above analysis emphasizes that lake and non-lake interfaces show potentially measurable differences which could drive the automatic detection of SLs. Indeed, this analysis allowed us to define and extract three families of features, i.e., i) topographic, ii) shape, and iii) statistical (see Section III-A), which are used by the SVM classifier for the automatic SL detection (see Section III-B).

III. Proposed method

Let us denote a radargram as:

$$\mathbf{P} = \{P(x,y) \,|\, x \in X = [1,\dots,n_T], y \in Y = [1,\dots,n_S]\}, \tag{1}$$

where n_T and n_S are the number of traces in the azimuth direction and the number of samples in the range direction, respectively. $P(x,y)$ is the measured power in dB scale. The proposed method takes as input the radargram and aims to provide, for each trace x_0, a label $q_{x_0} \in \{-1,+1\}$ for the basal interface to belong to either a non-lake ($q_{x_0}=-1$) or to a lake interface ($q_{x_0}=+1$). Here we consider that the input radargram has been already corrected according to standard methods in order to remove the distortions due to the aircraft elevation variations.

The proposed method consists of two main parts, i.e., i) basal interface feature extraction, and ii) automatic detection of SLs. In the first part, a set of features that quantitatively capture the properties of the basal interface are extracted. In the second phase, the features are given as input to an SVM classifier that automatically performs the SL detection and estimates their probability. The description of the processing steps within each part is provided in details in Section III-A and Section III-B, respectively.

A. Basal Interface Feature Extraction

The extraction of discriminant features for SL detection is based on the analysis presented in Section II. Accordingly, we extract three families of features, i.e., 1) topographic features, 2) shape features, and 3) statistical features. In order to capture the local characteristics of the basal interface and the higher azimuth correlation of lake interfaces, all the features are extracted considering sequences of consecutive basal waveforms of azimuth length N_x. In particular, N_x is given by a trade-off analysis that considers the expected minimum lake dimension, as well as constraints on the number of samples to ensure a significant statistical analysis and sufficient feature discrimination capabilities. In the range direction, we focus on the power measurements belonging to the main lobe of lake basal waveforms, which has a width denoted with N_y.

1) Topographic Features: In order to measure the local variability of the basal topography, we propose extracting the *root mean square height* (RMSH) feature, denoted ξ. The RMSH represents the standard deviation of the basal topography about a mean surface [39], [25], and is defined as (2):

$$\xi_{x_0} = \sqrt{\frac{1}{N_x-1}\sum_{x\in X_0}\left[T^{BI}(x) - \overline{T}^{BI}(x_0)\right]^2}, \tag{2}$$

where $X_0 = [x_0 - n_x, x_0 + n_x]$ represents a sequence of $N_x = 2n_x+1$ traces centered on trace x_0, with $n_x \in \mathbb{Z}^+$, and $T^{BI}(x)$ is the topography of the basal interface, computed as:

$$\begin{aligned} T^{BI}(x) = \left[Elv(x) - \frac{v_{\text{air}}\Delta_{\text{air}}}{2}\right] + \\ - \left[y^{SI}(x) - y^{BI}(x)\right]\cdot\frac{v_{\text{ice}}\delta_{\text{ice}}}{2}, \forall x \in X, \end{aligned} \tag{3}$$

where $v_{\text{air}} = 3\cdot 10^8$m/s and $v_{\text{ice}} = 1.69\cdot 10^8$m/s are the speed of the wave in air and ice, respectively. $Elv(x)$ is the elevation of the aircraft with respect to the WGS84 system, $y^{SI}(x)$ is the sample position of surface interface (air/ice) peak power, and $y^{BI}(x)$ is the sample position of the basal interface peak power on trace x. $y^{BI}(x)$ can be obtained by using automatic methods (e.g., [41]). $\Delta_{\text{air}}(x)$ is the measured 2-way travel time of the wave in the air and $\delta_{\text{ice}}(x)$ is the 2-way travel time corresponding to an ice subsurface sample. $\overline{T}^{BI}(x_0)$ is the mean basal topography inside the azimuth window, computed as:

$$\overline{T}^{BI}(x_0) = E\left\{T^{BI}(x) | x \in X_0\right\}, \tag{4}$$

where $E\{\cdot\}$ is the expectation operation. Since SLs are characterized by a flat topography, we expect $\xi_{x_0} \to 0$ for traces x_0 belonging to lake interfaces. On the other hand, the greater the basal topographic variation, the greater ξ_{x_0}, which is expected in the case of non-lake basal interfaces.

As already mentioned, the topography and the interfacial roughness influence the correlation of basal reflected waveforms. In order to quantify the *local waveform correlation*, we propose extracting the following feature:

$$\zeta_{x_0} = E\left\{\frac{cov(w_{x_0}, w_x)}{\sigma_{w_{x_0}}\sigma_{w_x}} \,|\, x \in X_0, x \neq x_0,\right\} \tag{5}$$

where $cov(a,b)$ denotes the covariance between waveforms a and b, and σ_a denotes the variance of signal a. w_x is a basal waveform (column vector) containing the measurements of power on trace x belonging to a neighborhood $\mathcal{B}_{x_0}$ that includes $N_x = 2n_x+1$ consecutive traces, defined as:

$$\begin{aligned} \mathcal{B}_{x_0} = \{P(x,y) \,|\, x \in X_0, \\ y \in [\min_{x\in X_0}(y^{BI}(x)) - n_y, \dots, \max_{x\in X_0}(y^{BI}(x)) + n_y]\}, \end{aligned} \tag{6}$$

where $n_y \in \mathbb{Z}^+$ and $2n_y + 1 = N_y$. From (6) one can see that the neighborhood $\mathcal{B}_{x_0}$ is a bounding box, whose size in the azimuth direction is constant and equal to $N_x, \forall x_0 \in X$. However, in the range direction, the size of the bounding box varies as a function of the local topographic variability, with the condition that the main lobe of the basal waveforms is always inside the bounding box, on all N_x traces. According to this reasoning, we expect that in the case of large topographic variations, the bounding box $\mathcal{B}_{x_0}$ has a range size greater than N_y and contains, along with waveform mainlobes, also sidelobes and noise measurements [see Fig. 3(a)]. In the case

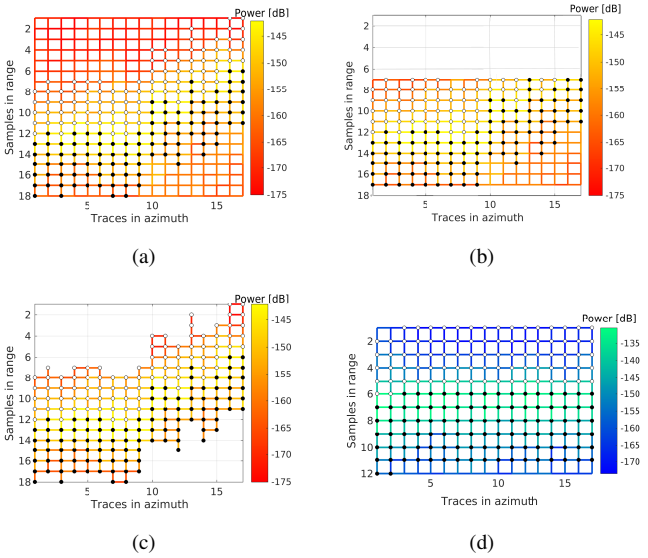


Fig. 3. Schematic representation of (a) the bounding box $\mathcal{B}$, (b) the compact box $\mathcal{C}$, and (c) the adaptive box $\mathcal{A}$ of a non-lake sequence, (d) a lake sequence, for which $\mathcal{A} \equiv \mathcal{B} \equiv \mathcal{C}$, approximately. The white and black dots correspond to the leading and trailing edges, respectively.

of ideally flat interfaces, the bounding box has a range size equal to N_y and includes only the main lobe of all the N_x waveforms [see Fig. 3(d)].

Two main considerations can be derived from (5) and (6). First, the greater the local topographic variability of the basal interface, as in the case of non-lake interfaces, the less correlated the waveforms inside $\mathcal{B}_{x_0}$, i.e., $\zeta_{x_0} \to 0$. Second, ideally $\zeta_{x_0} \to 1$ for samples x_0 belonging to flat smooth lake interfaces, for which the bounding box mainly contains aligned main lobes of similar waveforms. On the basis of these observations, it is clear that the use of the bounding box $\mathcal{B}_{x_0}$ for the calculation of ζ_{x_0} results in a great discrimination capability of the feature. In contrast, the computation of the correlation inside, for instance, a rectangular (compact) window $\mathcal{C}_{x_0}$ of size $N_x \times N_y$ moved over the basal interface (as in [31]) is less effective. Indeed, when the rectangular window is moved over interfaces with large topographic variation, several reflected waveforms may fall outside the window in the range direction [see Fig. 3(b)], thus leading to estimated correlation values unrepresentative of the real basal scattering.

2) Shape Features: As highlighted in Section II, the shape of the waveforms at the basal interface, in terms of steepness of the leading and trailing edges in the range direction, are qualitative indicators for the presence of lake interfaces. Here, we quantify such qualitative indicators by extracting shape features. To this aim, let us first define $\mathcal{A}_{x_0}$ as a sequence of $N_x = 2n_x + 1$ consecutive basal waveforms centered on trace x_0, with the condition that each waveform in $\mathcal{A}_{x_0}$ has length $N_y = 2n_y + 1$ and is centered in $y^{BI}(x)$, $\forall x \in X_0$. Thus, the sequence $\mathcal{A}_{x_0}$ is defined as:

$$\begin{aligned}\mathcal{A}_{x_0} = \{P(x,y) \,|\, & x \in X_0, \\ & y \in [y^{BI}(x) - n_y, \ldots, y^{BI}(x) + n_y]\}.\end{aligned} \tag{7}$$

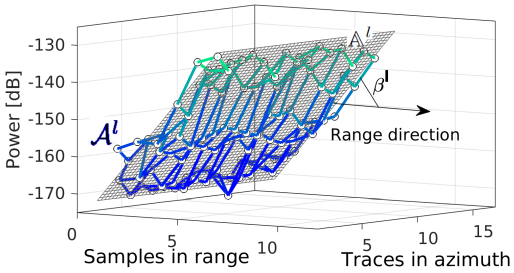


Fig. 4. Three dimensional schematic representation of a plane $\mathbb{A}^l$ fitted to a leading edge sequence $\mathcal{A}^l$. β^l is the slope of the plane with respect to the range direction.

Note that $\mathcal{A}_{x_0}$ is neither necessarily a compact rectangle $\mathcal{C}_{x_0}$ (see [31]), nor a variable size bounding box as in (6). Indeed, $\mathcal{A}_{x_0}$ adaptively follows the topography of the basal interface [see Fig. 3(c)] and contains a constant number of samples independently of its position.

For the extraction of the shape features, the sequence of waveforms $\mathcal{A}_{x_0}$ is split into leading edge sequences $\mathcal{A}^l_{x_0}$ and trailing edge sequences $\mathcal{A}^t_{x_0}$ [see Fig. 3(c)]. The leading and trailing edge sequences consist of the upper and lower half of the waveforms, respectively, both including the peak of the basal interface, such that $\{\mathcal{A}^l_{x_0} \cup \mathcal{A}^t_{x_0}\} = \mathcal{A}_{x_0}$ and $\{\mathcal{A}^t_{x_0} \cap \mathcal{A}^l_{x_0}\} = P(x, y^{BI}(x)), \forall x \in X_0$.

The key idea for estimating the *leading edge steepness*, denoted $\beta^l_{x_0}$, is to approximate $\mathcal{A}^l_{x_0}$ with a plane and to estimate its inclination with respect to the range direction. To this aim, we fit to $\mathcal{A}^l_{x_0}$ a plane $\mathbb{A}^l_{x_0}$ defined as:

$$\mathbb{A}^l_{x_0} = \alpha^l_{x_0} \cdot x + \beta^l_{x_0} \cdot y + \gamma^l_{x_0}, \tag{8}$$

where the coefficients $\alpha^l_{x_0}$, $\beta^l_{x_0}$ and $\gamma^l_{x_0}$ are estimated with the least squares criterion. Analogously, the *trailing edge steepness*, denoted $\beta^t_{x_0}$, is given by fitting to $\mathcal{A}^t_{x_0}$ the plane $\mathbb{A}^t_{x_0}$. Note that $\beta^l_{x_0}$ and $\beta^t_{x_0}$ are essentially the slopes of the fitted planes to the waveform edges in the range direction (see Fig. 4). For this reason, the extracted $\beta^l_{x_0}$ and $\beta^t_{x_0}$ features quantify the shape of the radar waveforms. As such, we expect lake reflections to be characterized by high values of $|\beta^l|$ and $|\beta^t|$. This is due to both the higher basal dielectric contrast, and reduced roughness over lake interfaces, which cause a narrow waveform with steep edges. In contrast, non-lake interfaces, which are characterized by lower dielectric contrast and greater roughness, yield wider waveforms (see also Section II). Accordingly, we expect both $|\beta^l| \to 0$ and $|\beta^t| \to 0$ for non-lake reflections.

An important observation regards the impact of the shape of $\mathcal{A}_{x_0}$ on the waveform steepness estimation. Besides the ideal case of x_0 belonging to perfectly flat interfaces, for which $\mathcal{A}_{x_0} \equiv \mathcal{B}_{x_0} \equiv \mathcal{C}_{x_0}$ [see Fig. 3(d)], the sequence $\mathcal{A}_{x_0}$ adaptively follows the position of the peak power of the basal interface. Thus, the leading/trailing edge sequence only contains the leading/trailing edge of the main lobe of all the basal waveforms on traces $x \in X_0$. The main advantage of using the adaptive box $\mathcal{A}_{x_0}$ against the bounding box $\mathcal{B}_{x_0}$ is that the fitting of the planes is performed on only relevant power measurements of the main lobe, i.e., without

side lobes and noise samples. On the other hand, the use of the compact box $\mathcal{C}_{x_0}$ can compromise the plane fitting on interfaces characterized by high topographic variability, since several reflected waveforms may fall outside the window in the range direction [see Fig. 3(b)].

3) Statistical Features: The importance of the statistical properties of the radar signal has been highlighted in several applications regarding the automatic analysis of remotely sensed radar data (e.g., [21], [38], [41], [42], [40]). To exploit these properties, in this paper we propose the extraction of the central moments of the basal reflected signal as features for the discrimination between lake and non-lake interfaces. Before performing the extraction of these features, it is worth noting that the use of $\mathcal{A}_{x_0}$ is preferred against the use of $\mathcal{B}_{x_0}$ and $\mathcal{C}_{x_0}$ for two main reasons. First, we aim to perform the statistical analysis only on the relevant basal backscattered power (i.e., the main lobe of the waveforms). Second, the number of samples inside the window should be constant irrespectively on the topography of the basal interface, to ensure the comparability of the extracted statistical features.

The first statistical feature we consider is the *mean adjusted basal peak power*, denoted by $\hat{\mu}_{x_0}^{BI}$ and defined as [43–45]:

$$\begin{aligned}\hat{\mu}_{x_0}^{BI} = E\{&P(x, y^{BI}(x)) + \\ &+2[2(Alt(x) + D\sqrt{\varepsilon_{\text{ice}}})]_{dB} + \\ &+2D[Ar]_{dB}\},\end{aligned} \tag{9}$$

where $[\cdot]_{dB}$ indicates power on a dB scale, $Alt(x) = v_{\text{air}} \cdot \Delta_{\text{air}}/2$ is the aircraft altitude with respect to the ice surface, D is the ice thickness and Ar is the one-way depth-averaged attenuation rate, which can be estimated using literature approaches, e.g., [46], [43], [47], [44], [45]. According to (9), the adjustment of the measured basal peak power $P(x, y^{BI}(x))$ is done in terms of both spreading losses in air and ice subsurface, and subsurface attenuation effects. Thus, $\hat{\mu}_{x_0}^{BI}$ is the power measured as if the basal interface was just below the aircraft. It is a function of roughness-modulated basal reflectivity [39], [38], system parameters and birefringence [44]. By neglecting the contribution of the system parameters and birefringence [45], it follows that in an ideal homogeneous englacial environment (i.e., with constant attenuation rate), $\hat{\mu}_{x_0}^{BI}$ contains information directly proportional to the reflection coefficient of the basal interface modulated (reduced) by its roughness [39]. In this ideal case, since lake interfaces have a high reflection coefficient and are smooth at the wavelength scale (see Sec. II), we expect similar values of $\hat{\mu}_{x_0}^{BI}$ at all lake interfaces, and greater $\hat{\mu}_{x_0}^{BI}$ at lake interfaces than at non-lake interfaces, independently on the interface depth. The same holds in a more realistic heterogeneous englacial environment with variable attenuation rates if the attenuation rates are adequately estimated. Thus, $\hat{\mu}_{x_0}^{BI}$ represents a robust feature for SL characterization and detection.

The roughness of an interface is only related to the wavelength, thus it is independent on the depth of the interface. Since the roughness influences the scattering [48], and thus the statistical properties of the radar signal [42] we expect that typically smooth lake interfaces and rough non-lake interfaces can be well distinguished by extracting adequate parameters from the local statistical properties of the basal reflected radar signal. In particular, we propose extracting the *coefficient of variation*, *skewness* and *kurtosis* as other potential discriminant features for SL detection.

The coefficient of variation, denoted ν_{x_0}, is defined as the ratio between the standard deviation σ_{x_0} and the absolute mean reflected power μ_{x_0}, i.e.,:

$$\nu_{x_0} = \frac{\sigma_{x_0}}{|\mu_{x_0}|}. \tag{10}$$

Lake interfaces, which are strong scatterers, show higher power and larger standard deviation than non-lake interfaces at the same depth [49]. On the other hand, because of subsurface attenuation effects, deeper lakes reflect a smaller mean power and show a reduced standard deviation than shallow lakes. In this case, the ratio ν_{x_0} reduces the subsurface attenuation effects, thus yielding comparable the properties of different lakes at different depths. Hence, the coefficient of variation ν_{x_0} represents a robust feature for discriminating between lake and non-lake interfaces independently on depth. It is worth noting that this reasoning holds only if the mean power μ_{x_0} inside $\mathcal{A}_{x_0}$ (which is in dB scale) is a positive quantity. Thus, before extracting ν_{x_0}, one may need to tune the input data, by adding a constant value (in dB) to all the dataset, such that the minimum power value is greater than zero. Note that this operation is a data manipulation artifice that does not affect the classification results, rather it insures the meaningfulness of ν_{x_0}.

The skewness, denoted ψ_{x_0}, is defined as:

$$\psi_{x_0} = E\{(\mathcal{A}_{x_0} - \mu_{x_0})^3\}/\sigma_{x_0}^3. \tag{11}$$

The skewness quantifies the degree of symmetry of a distribution [50]. A distribution with skewness 0 is perfectly symmetric, whereas a distribution with negative (positive) skewness has a long tail towards left (right).

The kurtosis, denoted κ_{x_0} is defined as:

$$\kappa_{x_0} = E\{(\mathcal{A}_{x_0} - \mu_{x_0})^4\}/\sigma_{x_0}^4. \tag{12}$$

The kurtosis measures the heaviness of the tails of a certain distribution [51], thus indicating the relative amount of outliers with respect to the normal distribution. In particular, if a distribution has a kurtosis $\kappa < 3$ (> 3), it has more (less) outliers than a normal distribution and is called platykurtic (leptokurtic), whereas a distribution with $\kappa = 3$ has a number of outliers comparable to a normal distribution, and is called mesokurtic.

We expect the skewness over lake interfaces to be larger than the skewness over non-lake interfaces, since the skewness becomes larger as the amount of scattering increases [49]. On the other hand, as pointed out in [52], if there is a weak concentration of scattered values around the mean (i.e., high standard deviation), the distribution will be spread or platykurtic, whereas if there is a strong concentration around the mean (i.e., small standard deviation), the distribution is leptokurtic. According to this analysis, we expect the distributions of the lake interfaces to be leptokurtic. Moreover, as already mentioned, because of wave subsurface attenuation effects, we expect the distribution of samples belonging to lakes located at

different depths to have different mean power values. However, ideally, the degree of symmetry and number of outliers of these distributions should not be greatly affected by the depth of the lakes. Hence, the skewness and kurtosis represent other two robust features for SL characterization independently on depth.

Since higher order central moments tend to be more affected by the presence of outliers [53], we limit the statistical analysis at the fourth central moment, in order to avoid providing ambiguous features to the automatic classifier. Note that the set of proposed statistical features allows us to overcome the limitations of the approach proposed in [21], while still exploiting the potentiality of the statistical properties of the radar signal for SL discrimination.

Summarizing, the full feature vector of the basal interface at azimuth position x_0 contains $n_F = 8$ features [see (2), (5), (8), (9), (10), (11), (12)] and is defined as:

$$\mathbf{v}_{x_0} = \left\{\xi_{x_0}, \zeta_{x_0}, \beta^l_{x_0}, \beta^t_{x_0}, \hat{\mu}^{BI}_{x_0}, \nu_{x_0}, \psi_{x_0}, \kappa_{x_0}\right\}. \tag{13}$$

Finally, the feature vectors are normalized and given as input to the SVM classifier to perform the SL detection.

B. Automatic Detection of Subglacial Lakes

The objective of this step is to predict whether a basal sample $x_0 \in X$ belongs to a non-lake ($q_{x_0} = -1$) or a lake interface ($q_{x_0} = +1$). To this aim, any binary classifier could be used. In this paper we propose to use an SVM classifier, considering its high performance in binary classification problems [30] and its success in classifying remotely sensed data [54], [41]. In the following, we briefly recall the main principles of SVM, which are useful for understanding the output of the proposed technique.

Let us assume that $n_L = n^+ + n^-$ out of the n_T samples of the basal interface are labeled by an expert, such that i) n^+ samples are labeled with $q = +1$, meaning they belong to lake interfaces, and ii) n^- samples are labeled with $q = -1$, meaning they belong to non-lake interfaces. All these samples represent the labeled set $\mathcal{L}$, which is used for training and testing the classifier. $\mathcal{L}$ is defined by the couples $(\mathbf{v}_j, q_j)$ as in (14):

$$\mathcal{L} = \{(\mathbf{v}_j, q_j), j \in X_L\}, \tag{14}$$

where X_L contains the indexes of the labeled samples.

The aim of a supervised classifier is to learn the characteristics of the two classes by using a training set, which is a subset of the labeled set $\mathcal{L}$ (for which both the features and the labels are known), in order to classify unlabeled samples (for which only the features are known). To this aim, the SVM classifier searches for the optimal separating hyperplane between the classes in the feature space. The optimal hyperplane is the one that maximizes the distance between itself and the nearest samples (support vectors) from each of the two classes. Accordingly, it computes a decision function $g(\mathbf{v})$ such that $q_* = \text{sign}[g(\mathbf{v}_*)] = \pm 1$ can be used to predict the label of any test sample $\mathbf{v}_*$. In the estimation of the decision function $g(\mathbf{v})$, a kernel function that fulfills Mercer's condition can be effectively included. A Mercer kernel function is continuous, symmetric and nonnegative definite, e.g., linear, polynomial, Gaussian radial basis function (RBF). Consequently, $g(\mathbf{v})$ is typically defined as:

$$g(\mathbf{v}) = \sum_i^{n_{SV}} \chi_i q_i \mathcal{K}(\mathbf{v}_i, \mathbf{v}) + b, \tag{15}$$

where b is the bias, n_{SV} is the number of support vectors, $\chi_i, i = [1, \ldots, n_{SV}]$ are the Lagrange multipliers and $\mathcal{K}$ is the kernel function. For further details on SVM, the reader is referred to [30], [54], [55].

The crisp outputs of the SVM (i.e., $q_* = -1$ or $q_* = +1$) can be transformed in soft probabilities, i.e., $p_* \in [0; 1]$, in order to allow a better interpretation of the lake detection results provided by the described method. To this aim, we use Platt's algorithm [56]. This algorithm is based on a parametric model S that approximates the SVM posterior class probabilities $Pr(q = +1|\mathbf{v})$ (i.e., the probability that a sample belongs to class $q = +1$ given its feature vector $\mathbf{v}$). In particular, the parametric model is a sigmoid $S_{A,B}[g(\mathbf{v})]$ defined as:

$$S_{A,B}[g(\mathbf{v})] = \frac{1}{1 + \exp(A \cdot g(\mathbf{v}) + B)}, \tag{16}$$

where A, B are the parameters of the sigmoid, which are optimized using the maximum likelihood estimation approach. As such, the best fitting parameters $\hat{A}$ and $\hat{B}$ such that $Pr(q = +1|\mathbf{v}) \approx S_{\hat{A},\hat{B}}[g(\mathbf{v})]$, are determined by minimizing the negative log likelihood of the training data [56], i.e.,:

$$\begin{aligned}(\hat{A}, \hat{B}) &= \min_{A,B} F(A, B) = \\ &= -\sum_j \left(t_j \log s_j + (1 + t_j) \log(1 - s_j)\right)\end{aligned} \tag{17}$$

where

$$s_j = S_{A,B}[g(\mathbf{v}_j)], \quad t_j = \begin{cases} \frac{n^+ + 1}{n^+ + 2} & \text{if } q_j = +1 \\ \frac{1}{n^- + 2} & \text{if } q_j = -1 \end{cases} \quad j \in X_L.$$

IV. Experimental results

In order to prove the validity of the proposed method, we applied it to two RS datasets acquired by the MCoRDS instrument [32]. However, the technique is general and can be applied to data acquired by other instruments or using different instrument parameters and/or data processing algorithms, at the condition that the basal interface should be visible and clearly detectable. To ensure the consistency of the SVM model with respect to the extracted features, the SVM must be trained and applied to data obtained i) by the same instrument operated with the same parameters, and ii) using the same processing algorithms. This implies that the SVM must be retrained if one of these two conditions changes.

A. Data Description

The two MCoRDS datasets considered in this paper have been acquired on two large regions in Antarctica, i.e., Lake District and Siple Coast, in the autumn campaign in 2013. Both datasets have been processed with range and azimuth compression and with the minimum variance distortionless response (MVDR) algorithm for clutter reduction, to enhance

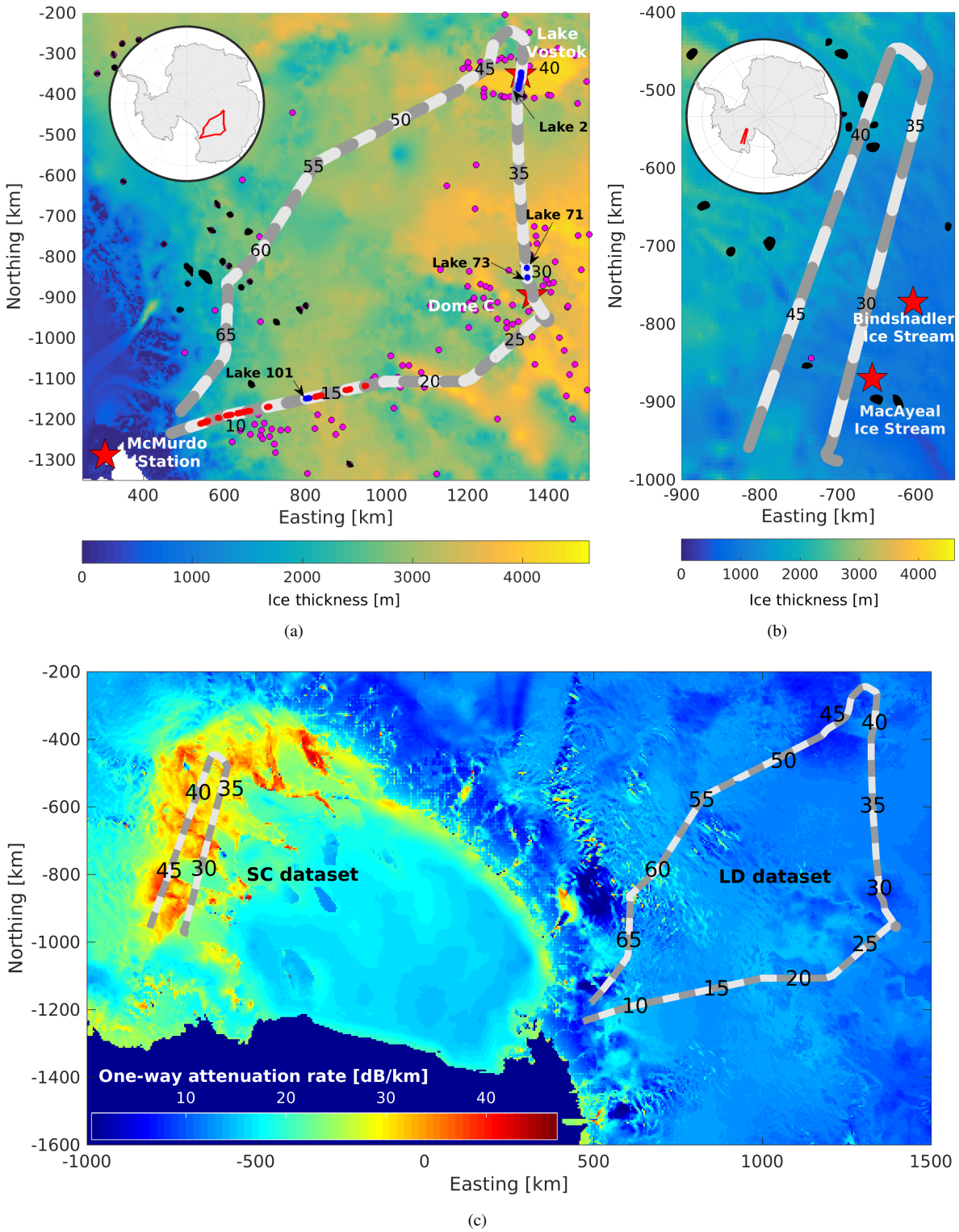


Fig. 5. Ice thickness map [57] of the (a) Lake District and (b) Siple Coast regions surveyed by MCoRDS in the airborne campaign held in autumn 2013. Overlapped are the locations of the investigated MCoRDS radargrams (from 7 to 69, and from 26 to 48, for LD and SC, respectively) with alternating gray and white colors. The locations of the radar-detected and active inventoried SLs in the two regions are represented with magenta dots and black polygons, respectively. In the LD dataset, the locations of the reference lake and non-lake interfaces are highlighted with blue and red, respectively. (c) Location of the LD and SC flightlines overlapped on the one-way depth averaged attenuation rate estimated in [43].

TABLE I
PARAMETERS OF THE MCORDS SYSTEM AND INVESTIGATED DATA.

Campaign	Lake District	Siple Coast
Number of radargrams	63	
Platform type	P3 aircraft	
Platform height above surface	≈500m	
Central frequency	195MHz	
Wavelength in ice	≈0.86m	
Pulse repetition frequency	12kHz	
Bandwidth	30MHz	
Range resolution in ice (pulse compressed)	4.3m	
Range spacing in ice	2.8m	
Azimuth resolution (SAR processed)	25m	
Azimuth spacing	30m	
Across-track resolution (pulse limited)	70m	
Total number of traces n_T	104559	38289
Total traveled distance	≈3000km	≈1125km
Basal interface depth (ice thickness)	247m - 4752m	737m - 2100m
Estimated mean one-way attenuation rate [43]	12dB/km	30dB/km

the basal signal [58]. The parameters of the RS system and the acquired data specifications are given in Table I. In particular:

- The analyzed Lake District (LD) dataset was acquired over a large part of the continent, with extremely variable ice thickness [see Fig. 5(a)] and relatively low attenuation rate (mean value equal to 12dB/km) [see Fig. 5(c)] [43]. The location of the flightline corresponds to regions where previous studies have reported several radar-detected SLs (e.g., [16]). The flightpath of the analyzed LD dataset is almost circular, starting from the McMurdo Station, passing over Dome C, twice over the Vostok lake and then coming back to the McMurdo Station. The whole LD dataset is composed of 77 radargrams, of which we discarded the first 6 and the last 8 radargrams, since they were acquired over the mountains nearby McMurdo Station. Thus, we analyze the remaining 63 radargrams (from radargram 7 to radargram 69), which cover a distance of about 3000km (see Fig. 5).
- The Siple Coast (SC) dataset was acquired over the MacAyeal Ice Stream going South through the Bindshadler Ice Stream towards the C1 tributary of Kamb Ice Stream and coming back towards the MacAyeal Ice Stream on a parallel track. The flightline, which is about 1125km long, crosses a region with several active SLs, but no radar-detected SL [16] except for a location near the C1 tributary of Kamb Ice Stream. There a recent work has claimed the presence of a basal distributed water sheet detected by analyzing RS data [24]. The ice has smaller thickness variability than in the case of the LD dataset [see Fig. 5(b)], but larger estimated values of attenuation rate, (mean value equal to 30dB/km) [see Fig. 5(c)]. The analyzed SC dataset contains 23 radargrams, from radargram 26 to radargram 48 of the full dataset acquired in the SC campaign, which contains 76 radargrams.

An initial analysis of the data pointed out a difference of ≈5dB between the mean noise power of the LD and SC datasets (data are not radiometrically calibrated), which we compensated before the feature extraction in order to yield comparable input-output data. Besides the MCoRDS datasets, we used the one-way ice attenuation rates of Antarctica estimated in [43] as input to the extraction of the mean adjusted peak power. This dataset is available at a horizontal spacing of 5km, which is much greater than the azimuth spacing of the RS data, i.e., 30m (see Table I). To slightly compensate for this difference, we upsampled the attenuation rate map with a factor of 5 (i.e., final horizontal spacing of 1km) using the bilinear interpolation method.

The most recent SL inventory [16] is dated 2012, i.e., one year prior to the LD and SC MCoRDS campaigns, which were held in 2013. For this reason, the datasets may contain SLs which are not inventoried. In order to train and test the classifier, we initially collected reference samples of both lake and non-lake interfaces, according to the following strategy. The lake samples were collected by crosschecking the locations and depths of the SLs inventoried in [16] with the location, depth and reflectivity of the basal interface in the MCoRDS dataset. By doing so we validated and thus confirm the presence of $N_{\text{ref}}^{\text{lakes}} = 4$ inventoried lakes in the MCoRDS dataset, i.e., lake 101 in radargram 14, lake 73 and lake 71 in radargram 30, lake 2 (Vostok) in radargrams 29, 40, 45, 46, located at depths of approximately 3420, 3020, 3570, and 4100m, respectively. Lake 169, at 2620m depth, is likely to be also present in the LD dataset. However, we do not consider the samples of lake 169 since the navigation data at its location report an aircraft roll larger than 15°, which may drastically reduce the reflected power, thus generating unreliable features. Moreover, given that the aircraft passed twice over lake Vostok, which is a very extended lake, the total number of samples of Vostok lake is very large compared to all the others, and may bias the classifier towards the detection of lakes with features similar to Vostok. In order to avoid this issue, in the experiments we used only a subset of Vostok samples. Thus, the total number of samples of all the reference lakes considered in the analysis is $n^+ = 1771$. The non-lake

samples were collected from locations that do not overlap with the coordinates of the inventoried SLs according to a careful visual interpretation of the basal signature in radargrams, i.e., sequences of traces with evident topographic variability and roughness. The total number of collected non-lake samples is $n^- = 3405$. Therefore, the total number of labeled samples is $n_L = n^+ + n^- = 5176$. Fig. 5 shows the locations of the lakes and non-lake interfaces used as reference in the experiments, with blue and red, respectively. It it is also worth mentioning that lakes 174, 57 and 106 are very close to the LD flightline, in radargrams 12, 31 and 60, respectively, at depths comparable to the estimated depth of basal returns in the radargram. However, from the visual analysis of the reflections in the radargrams it is not possible to confirm with high confidence the presence of these lakes in the LD dataset. Thus, these lakes were not included in the training of the SVM and in the quantitative analyses. However, they represent good canditates for qualitatively analyzing the behavior of the features and the output of the algorithm. Another important observation regards the fact that training samples are chosen from the LD dataset only, whereas the robustness of the algorithm is verified on both the LD and SC datasets.

B. Parameter Setting

The input parameters of the method are N_x and N_y, defining the azimuth and range size of the basal waveform sequences, and the parameters of the SVM.

The parameters N_x and N_y can be straightforwardly set based on the characteristics of the investigated data and targets:

- N_y should be set based on the width of the main lobe of lake reflected waveforms. To this aim, we performed an analysis of several waveforms reflected by labeled lake interfaces and derived that, in average, in the investigated MCoRDS dataset the width of the main lobe is $N_y = 11$ samples.
- N_x should be set based on a trade-off between algorithm constraints and minimum expected size of the SLs. On the one hand, regarding the algorithm constraints, N_x should be sufficiently large to i) ensure meaningful statistics in the extraction of the statistical features, and ii) guarantee sufficient discrimination capabilities of the features (for further details, see Fig. 8). On the other hand, an analysis of the distribution of the lengths of all known SLs in Antarctica has been performed in [15]. The analysis points out that the statistical distribution of the lengths of such lakes is positively skewed, with the bulk of SLs of less than 10km length and with the modal size being 5km (see Fig. 5 in [15]). However, there are several water bodies with dimension much smaller than 2km, among which a large part are about 500m long [36], [16]. Considering the azimuth spacing of the MCoRDS instrument (i.e., 30m, see Table I), 500m corresponds to $N_x = 17$ traces. This results in $N_x \cdot N_y = 187$ samples, which meets the above algorithm constraints.

For the classification, we used an RBF kernel for the SVM. This choice is motivated by the fact that the RBF kernel is typically more flexible than the linear kernel and it usually outperforms the polynomial kernel in convergence time [55]. Therefore, the SVM model parameters are the penalty error term and the width of the RBF kernel. In order to estimate them, we used a classical n-fold cross-validation approach [55] with $n = 10$ folds.

According to the above observations, the method depends on only four input parameters.

C. Analysis of the Extracted Features

In the following, we assess the effectiveness of the proposed features extracted from the MCoRDS datasets. Fig. 6 shows the features of the non-lake/lake sequences of $N_x = 17$ basal waveforms centered on the orange/blue traces in Fig. 2(b). In particular, Fig. 6(a) and 6(b) show a qualitative representation of the corresponding adaptive and bounding boxes for the non-lake and lake sequences, respectively; Fig. 6(c), 6(d), 6(e), 6(f) and 6(g) show the 3D view of the investigated sequences, and the fitting of the leading and trailing edges; Fig. 6(i) and 6(j) show the statistical distribution of the samples within the adaptive boxes. From the analysis of the figures and values of the features, one can see that the investigated lake sequence with respect to the non-lake sequence is characterized by:

- a smaller RMSH ($\xi_{\text{non-lake}} = 2.31\text{m} > \xi_{\text{lake}} = 0.33\text{m}$);
- a higher local correlation ($\zeta_{\text{non-lake}} = 0.73 < \zeta_{\text{lake}} = 0.97$);
- a higher waveform steepness of both the leading edge ($|\beta^l_{\text{non-lake}}| = 3.02 < |\beta^l_{\text{lake}}| = 7$) and the trailing edge ($|\beta^t_{\text{non-lake}}| = 1.89 < |\beta^t_{\text{lake}}| = 4.78$);
- a higher mean adjusted basal peak power ($\hat{\mu}^{BI}_{\text{non-lake}} = -1.75\text{dB} < \hat{\mu}^{BI}_{\text{lake}} = 11.10\text{dB}$);
- a higher coefficient of variation ($\nu_{\text{non-lake}} = 0.025 < \nu_{\text{lake}} = 0.044$);
- a higher skewness ($\psi_{\text{non-lake}} = -0.5 < \psi_{\text{lake}} = 0.09$);
- a smaller kurtosis ($\kappa_{\text{non-lake}} = 3.1 > \kappa_{\text{lake}} = 2.12$).

All these results confirm the expectations about the topographic, shape and statistical features, as derived in Sec. III-A.

In order to analyze the variability or coherency of the proposed features along successive basal sequences, Fig. 7(b), 7(c), 7(d) and 7(e) report the features extracted for the basal interface illustrated in Fig. 7(a), which contains the reflection of lake 71 [also illustrated in the radargram example in Fig. 2(b)]. The analysis of these figures i) confirms what expected in terms of feature values over lake and non-lake interfaces (see Sec. III-A), and ii) points out a greater coherence of the features over the lake reflection than outside it.

Another observation regards the fact that the values of the features tend to lose coherence towards the sides of the lakes. This is due to the employed sliding window approach which acts like a low-pass filter in the azimuth direction, thus smoothing the borders of the targets. In order to better analyze the effect of the size of the window in the azimuth direction (i.e., sequence length) on the extracted features, Fig. 8 illustrates the features extracted for the basal interface shown in Fig. 7(a) with fixed $N_y = 11$ and variable $N_x \in [9, \ldots, 33]$ (corresponding to an azimuth distance between ≈ 250m and 1000m). In particular, the lake reflections are between traces 50 and 106 and the values of the extracted features (for

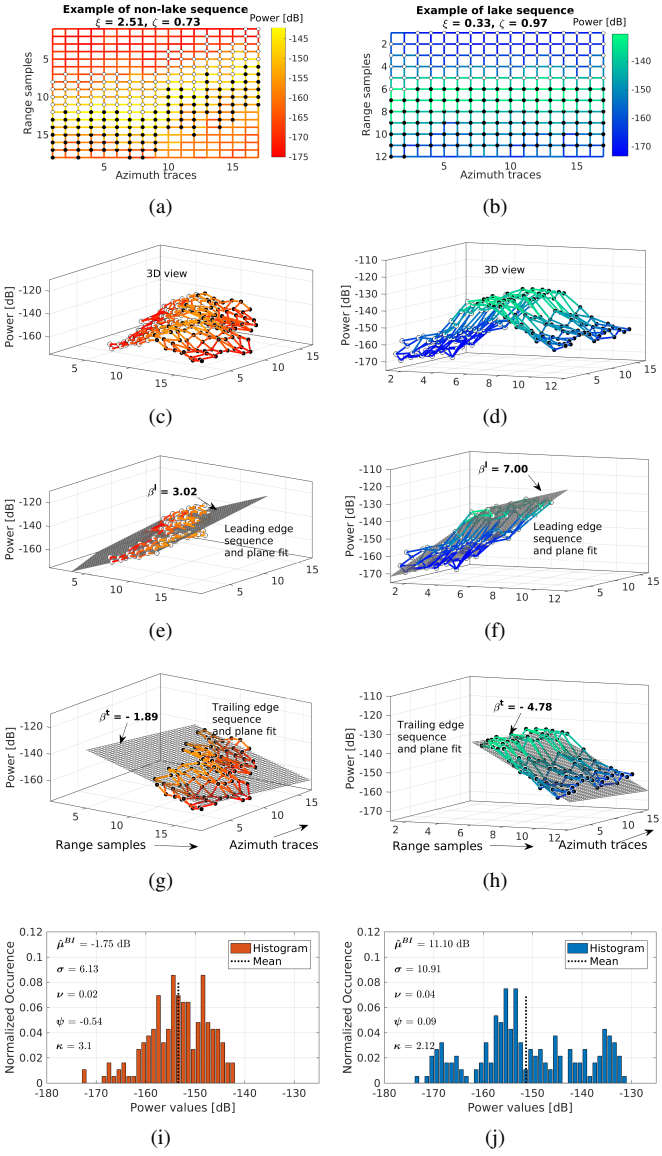


Fig. 6. Topographic, shape and statistical features for: (left) the non-lake basal sequence centered on the orange trace (i.e., trace 19) in Fig. 2(b), and (right) the lake basal sequence centered on the blue trace (i.e., trace 66) in Fig. 2(b). (a), (b) bounding box $\mathcal{B}$ (all pixels) versus adaptive box $\mathcal{A}$ (pixels marked with white and black dots) and values of the topographic (RMSH ξ and local correlation ζ) features; (c) and (d) 3D view of the sequences belonging to the adaptive box; (e) and (f) leading edge sequences (marked with white dots), and (g) and (h) trailing edge sequences (marked with black dots) along with the corresponding fitted planes and values of the leading and trailing edge steepness features, i.e., β^t and β^l; (i) and (j) histograms of the samples inside the adaptive box and values of the extracted statistical features (i.e., adjusted peak power $\hat{\mu}^{BI}$, coefficient of variation ν, skewness ψ and kurtosis κ).

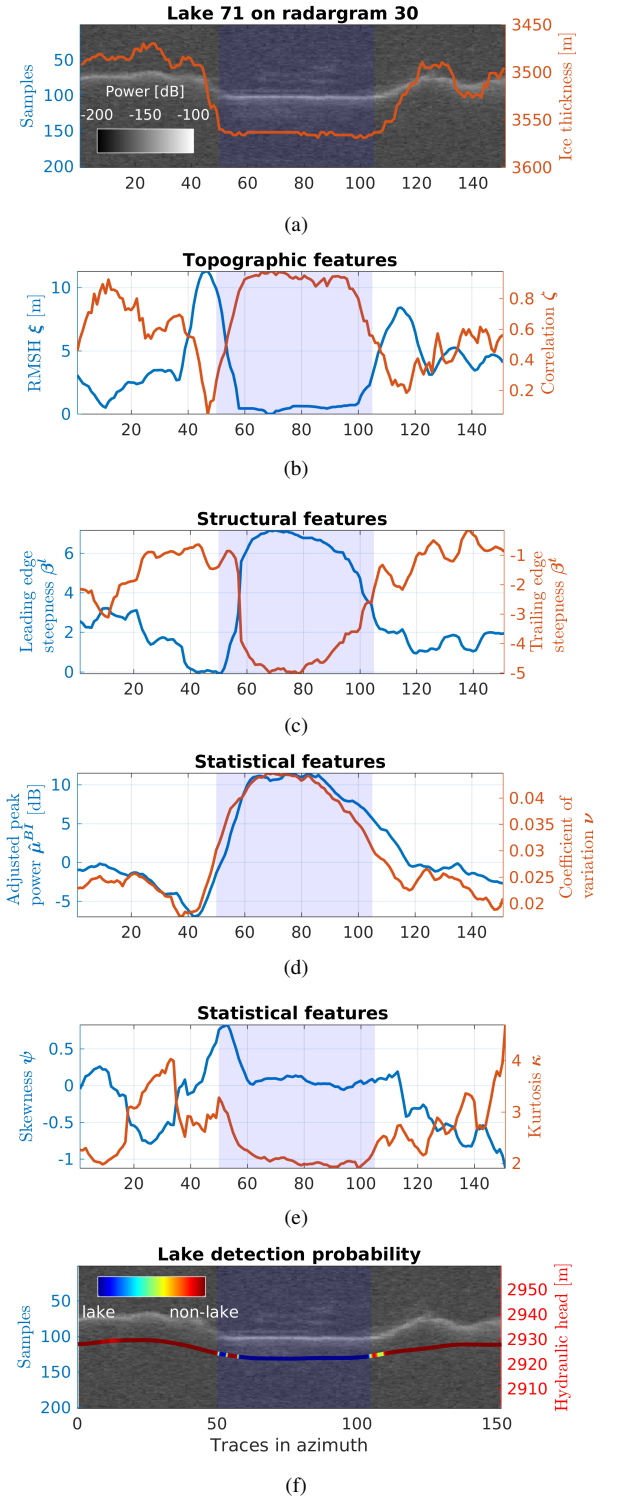


Fig. 7. (a) Portion of radargram showing lake and non-lake reflections [also illustrated in Fig. 2(b)] and corresponding (b) topographic features, (c) shape features, (d) and (e) statistical features, (f) lake detection probability. The features have been extracted by considering $N_x = 17$ and $N_y = 11$, as derived in Sec. IV-B.

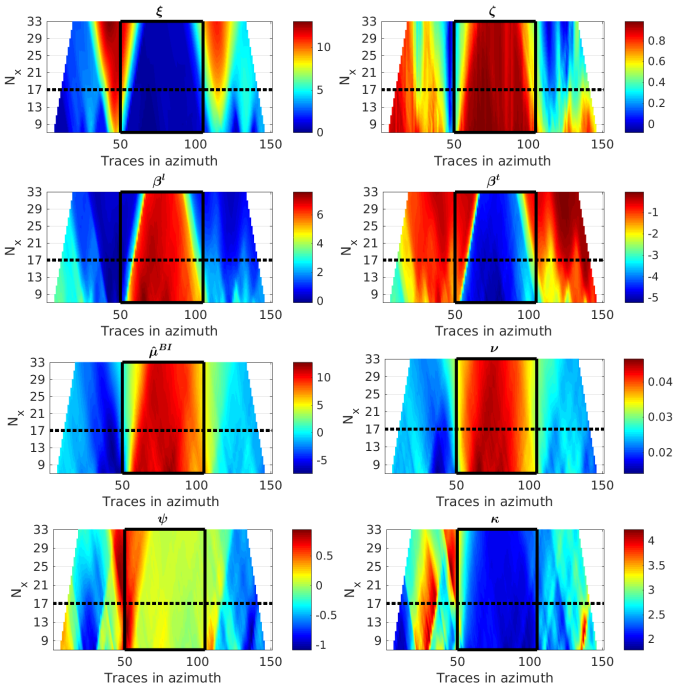


Fig. 8. Effect of azimuth window size N_x on the features.

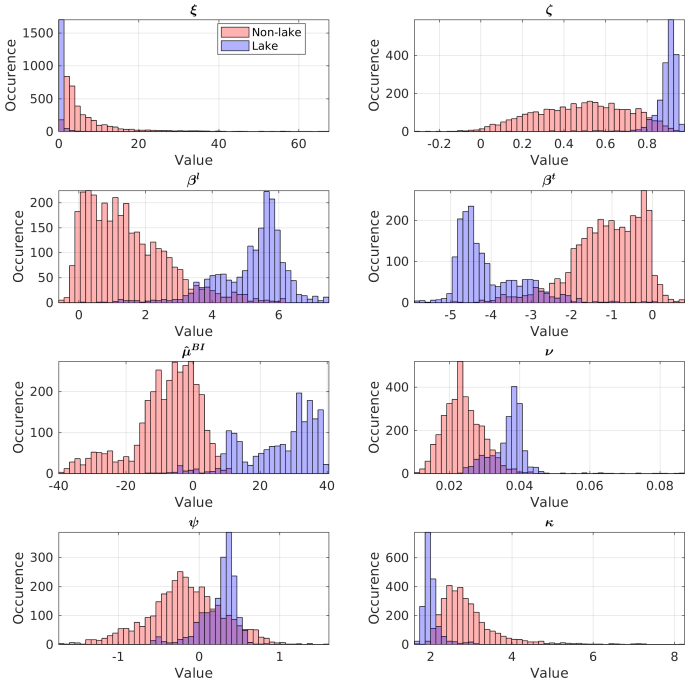


Fig. 9. Histograms of the values of the features for the non-lake labeled samples versus the lake labeled samples.

variable N_x) over the lake traces are enclosed in the black rectangle. The horizontal black dotted line marks the values of the features computed with $N_x = 17$ [which are thus equivalent to the features reported in Fig. 7(b), 7(c), 7(d) and 7(e)]. As one can see in Fig. 8, the effect of changing the azimuth sequence length N_x is as in the following:

- an increase of the sequence length, i.e., $N_x \gg 17$, tends to oversmooth the features at the borders of the lake interface. This is particularly evident in Fig. 8 for the correlation and the shape features. This may result in lower performance for the detection of lakes whose azimuth extent is comparable to the imposed detection limit (i.e., 500m).
- a decrease of the sequence length, i.e., $N_x \ll 17$, often implies the estimation of features with similar values for both lake and non-lake interfaces. For instance, this is evident in Fig. 8 for the correlation, skewness and kurtosis. Indeed, since the correlation depends mainly on the topographic variability, it is reasonable that in a small neighborhood the topography can be approximately constant also over bedrock interfaces. Moreover, a small N_x may not guarantee sufficient samples for a significant statistical analysis, thus leading to unstable estimates with greater oscillations.

The above analysis confirms that, given the characteristics of the MCoRDS datasets, $N_x = 17$ traces represents a good trade-off between feature discrimination capabilities and accurate SL detection over the greatest part of their length.

In order to understand the potentiality of the proposed features for discriminating lake from non-lake interfaces at a larger scale, in the following we analyze their statistical distribution for all the reference labeled samples. Fig. 9 reports the histograms of features extracted from the $n^+ = 1771$ lake samples (in blue) and the $n^- = 3405$ non-lake samples (in red). The analysis of the figure points out that the histograms of the features calculated on lake samples have either a small variance (e.g., for ξ, ζ, ν, ψ, κ), or are distant from the histograms obtained on non-lake samples (e.g., for β^l, β^t, $\hat{\mu}^{BI}$). The small variance indicates that the corresponding features are particularly suitable for characterizing SLs independently on depth. The distant modes suggest the capability of the features in discriminating lake from non-lake interfaces and their robustness to subglacial attenuation. The analysis of the figure also confirms that, as expected (see Section III-A), the lake interfaces have small topographic variability (i.e., the peak of the histogram of $\xi \to 0$), high correlation (i.e., the peak of the histogram of $\zeta \to 1$), high waveform edge steepness (i.e., higher values of $|\beta^l|$ and $|\beta^t|$), high adjusted peak power and high coefficient of variation. Another important observation that can be derived by analyzing Fig. 9 regards the shape of the statistical distribution of non-lake and lake basal sequences: i) the shape of the histograms of non-lake sequences is typically negatively skewed (the peak of the red histogram of ψ is at value <0) and mesokurtic (the peak of the red histogram of κ is at value ≈ 3) and ii) the shape of the histograms of lake sequences is typically positively skewed (the peak of the blue histogram of ψ is at value >0) and platikurtic (the peak of the blue histogram of κ is at value <3). The behavior of these features is in line with the theoretical expectations (see Section III-A3) and suggests that other potential SLs, currently not inventoried, can be described by similar values of the extracted features.

TABLE II
QUANTITATIVE RESULTS OBTAINED BY VARYING THE TRAINING SET SIZE. FOR EACH TRAINING SET SIZE, 10 RANDOM EXPERIMENTS ARE PERFORMED, AND THE MEAN VALUES (IN PERCENTAGE) AND STANDARD DEVIATION IN TERMS OF RECALL, SPECIFICITY, OVERALL ACCURACY AND PRECISION ARE PROVIDED.

			Training set size x% (x% randomly picked from 50% labeled lake and 50% labeled non-lake data)			
			25%	***50%***	***75%***	***100%***
Number of samples	Train	n^+	221	442	663	884
		n^-	426	851	1277	1702
		Total	*647*	*1330*	*1940*	*2586*
	Test	n^+	884	884	884	884
		n^-	1703	1703	1703	1703
		Total	*2587*	*2587*	*2587*	*2587*
Performance measures [%] on 10 random experiments	**Recall (= hit rate = 100 - miss rate)**	Mean	98.94	99.69	99.71	99.84
		Std	0.64	0.21	0.31	0.17
	Specificity (= 100 - false alarm rate)	Mean	98.44	99.08	99.34	99.46
		Std	0.79	0.41	0.23	0.25
	Overall Accuracy	Mean	96.97	98.22	98.73	98.96
		Std	1.57	0.81	0.46	0.48
	Precision	Mean	98.60	99.28	99.47	99.59
		Std	0.55	0.27	0.19	0.16

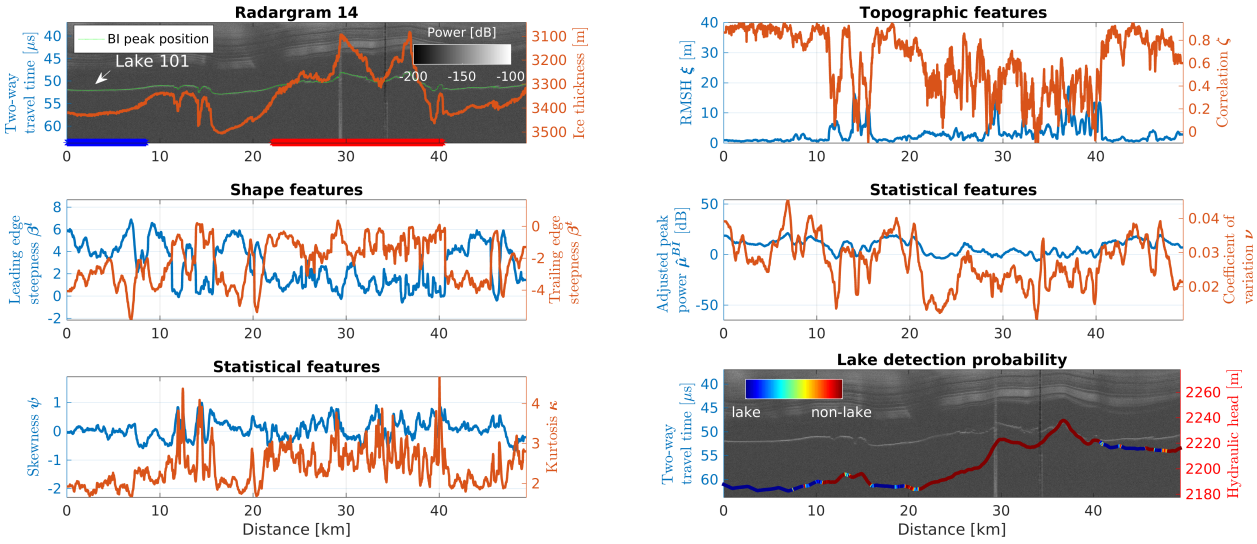
D. Classification Results

In order to asses quantitatively the performances of the method and analyze the stability of the results, we provide obtained classification results in terms of four commonly used metrics, i.e., recall, specificity, overall accuracy and precision. These have been computed by splitting the labeled dataset into training and test sets, according to the following approach. The test set consists of 50% samples picked randomly from all labeled lake and non-lake data, respectively. In the training phase, four training sets are formed with 25%, 50%, 75% and 100% (i.e., the maximum number) of the remaining lake and non-lake labeled data, respectively. Thus, this approach provides a means to understand the impact of the number of training samples on the final results. In order to asses the stability of the results at the random choice of the training samples, 10 random experiments have been performed by repeating the whole random selection procedure for each training/test set. Table II reports the number of lake and non-lake samples in the train and test sets, as well as the obtained results in terms of mean value and standard deviation of the four metrics for each training/test set generated according to the described validation approach. From the analysis of the results, the following observations can be derived:

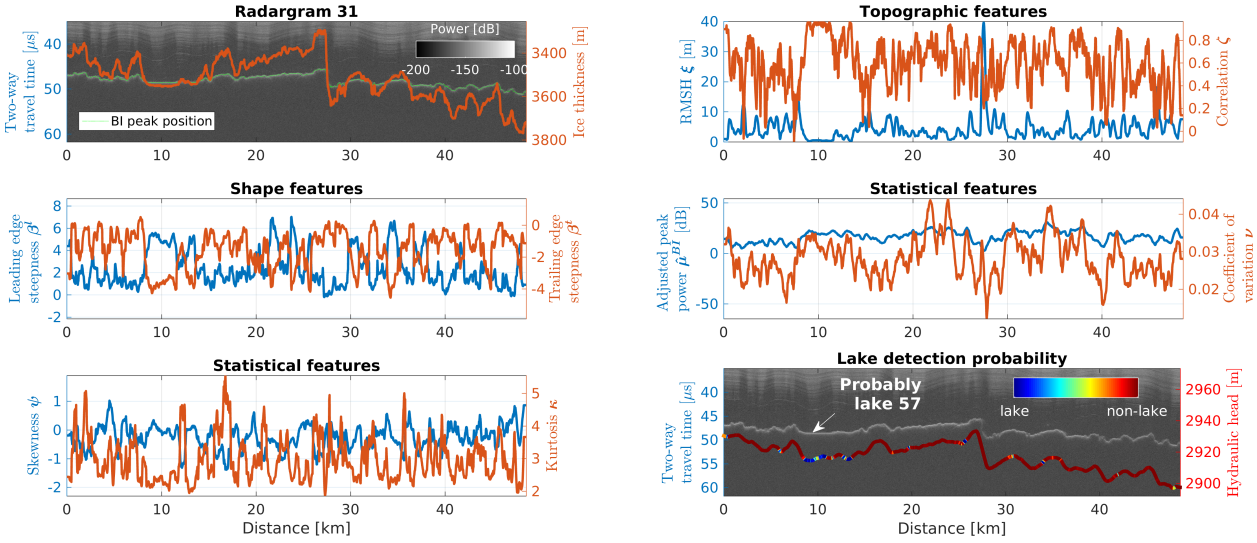
- The mean value of the considered measures increases by increasing the size of the training set, confirming the expected correlation of the results with the size of the training set. As such, the best results (i.e., performance greater than 98.96% according to all metrics) are obtained with the maximum number of possible training samples (i.e., column 100% in Table II);
- The mean values of the recall (= 100 - miss rate), specificity (= 100 - false alarm rate), overall accuracy and precision are greater than 98.94%, 98.44%, 96.97% and 98.60%, respectively. The standard deviation of these measures is always smaller than 1.57%, denoting a good stability of the obtained results. These values indicate that the proposed method can be confidently used to effectively detect SLs.

To further prove the generalization capabilities of the method, an additional experiment has been carried out. All the samples of one of the labeled lakes have been used only in the test phase, whereas the training has been performed with 50% random samples of the remaining lake and non-lake data. For this experiment we used for the test phase the samples of lake 101, since its medium size length (i.e., $\approx$ 8.7km corresponding to 289 samples) is sufficient for a meaningful statistical analysis. The mean $\pm$ standard deviation of the four accuracy metrics obtained by averaging the results of 10 random training experiments are: recall = $98.33 \pm 1.21\%$, specificity = $96.25 \pm 0.1\%$, overall accuracy = $96.48 \pm 0.05\%$ and precision = $76.93 \pm 0.71\%$. These results confirm the effectiveness of the method also when classifying samples of new lakes that are not considered in the training of the classifier.

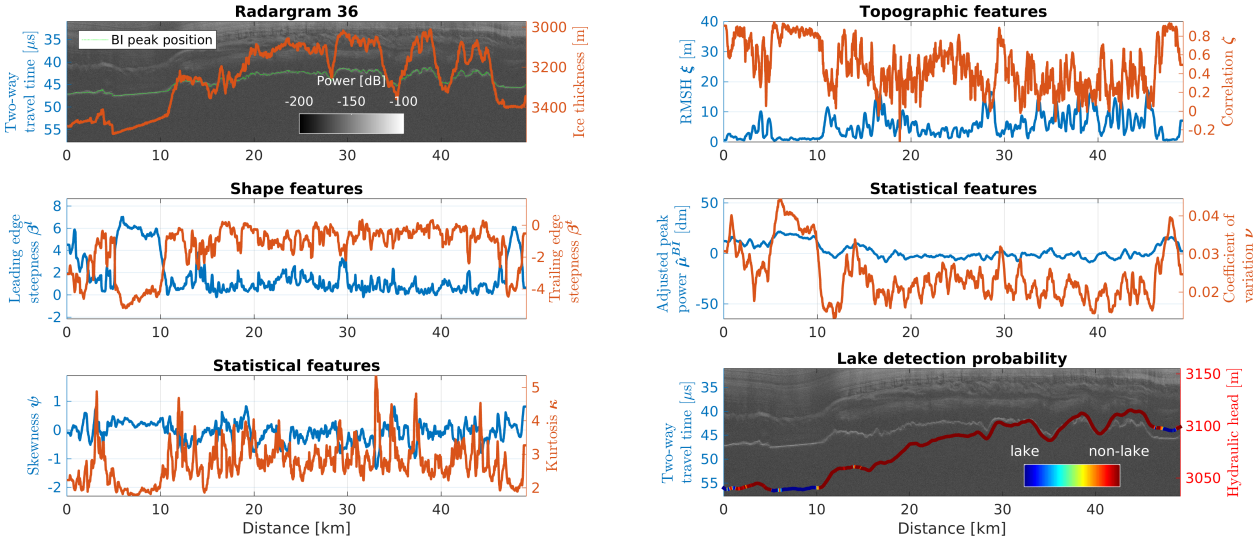
In order to qualitatively asses the performance of the method, the probability of SL presence on a subset of the test samples (i.e., on lake 71) is provided in Fig. 7(f). Fig. 10(a) shows another example of output, in which the reflections of lake 101 appear at an along-track distance between 0-10km. These results have been obtained by performing an experiment using 50% of all labeled samples randomly selected for training and the remaining samples for test. For this experiment, we obtained a recall of 99.88%, a specificity of 99.29%, an overall accuracy of 99.49% and a precision of 98.64%. As one can see, with few exceptions, the method provides a high/low probability of SL presence on the reference lake/non-lake samples, thus proving its effectiveness. The few exceptions are mostly at the borders of the lakes, where the output is more uncertain due to the sliding window approach. However,



(a)

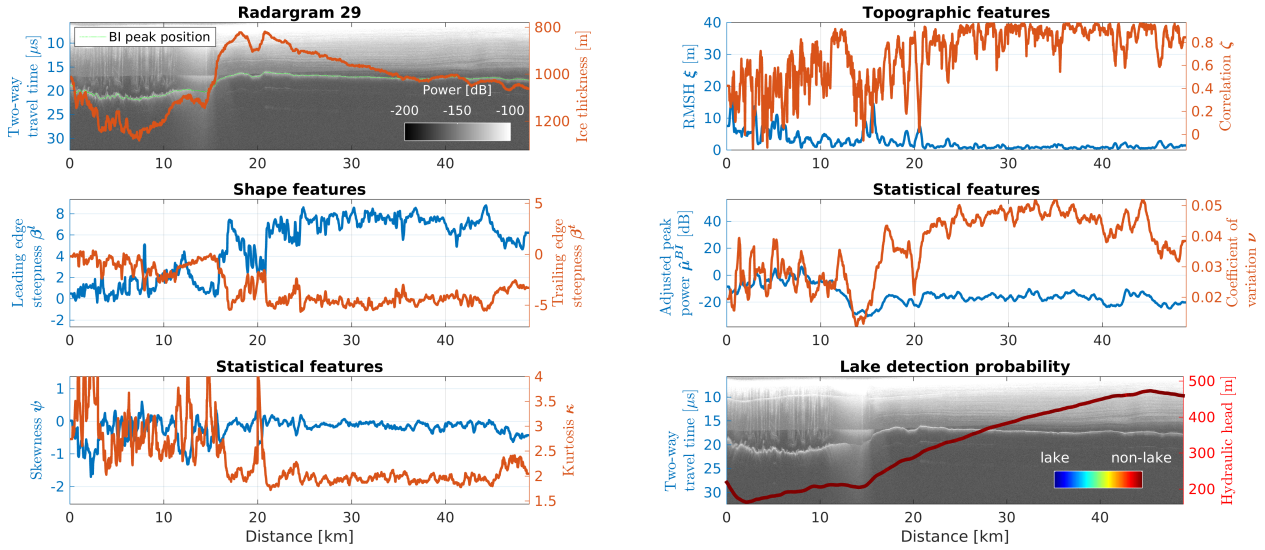


(b)

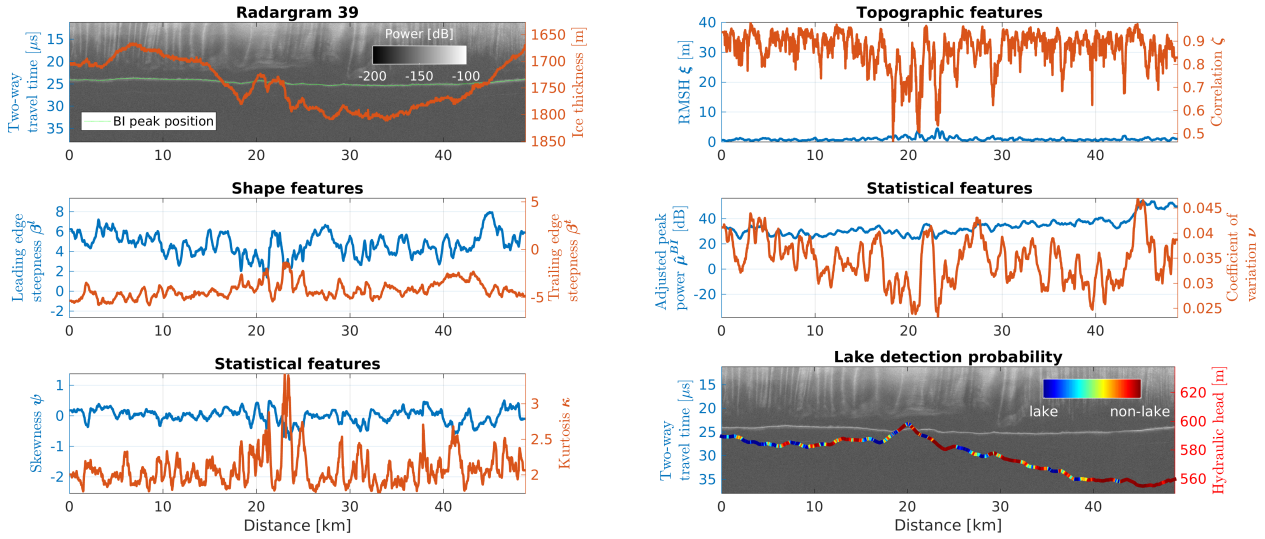


(c)

Fig. 10. Examples of LD radargrams, extracted features and estimated lake detection probability. Three radargrams are considered: (a) radargram 14 containing the inventoried lake 101, (b) radargram 31 passing close to lake 57, and (c) radargram 36 which does not pass close to any inventoried SL. Where present, the blue and red lines at the bottom of the radargrams correspond to azimuth locations of basal samples considered in the quantitative analysis, for training and testing the SVM.



(a)



(b)

Fig. 11. Examples of SC radargrams, extracted features and estimated lake detection probability. Two radargrams are considered (a) radargram 29, between the MacAyeal and Bindshadler Ice Streams, and (b) radargram 39 close to the C1 tributary of Kamb Ice Stream.

this is not critical for the goals of the proposed approach.

The high performance obtained on the reference samples suggest that the proposed method can be effectively used to confirm or accurately detect other potential SLs and update the current SL inventory. To prove this, Fig. 10(b) and Fig. 10(c) illustrate two examples of radargrams, along with the extracted features and the estimated SL detection probability. In particular, Fig. 10(b) shows a radargram acquired near lake 57, which is located at a depth of about 3574m [16]. The output of the method points out a SL, located at a depth of $\approx$ 3550m, thus potentially lake 57. Fig. 10(c) illustrates two relatively flat strong radar reflections, which extend over several consecutive traces, visually resembling the appearance of SLs. The behavior of the proposed features over these regions indicates a great coherence and similarity with the features over the reference lake interface shown in Fig. 7, further suggesting the presence of the SLs. The output of the SVM points out a very high probability of SL reflection, thus supporting the presence of the hypothesized SLs.

In order to further validate the presence of such lakes, we also performed an analysis of the results according to the water ponding condition [59], [6]. The water ponding condition relies on the Shreve hydrological model [20] that states that subglacial water flows down the gradient of hydraulic head H_h and ponds if the hydraulic head is nearly constant. This implies that water ponds only in hydraulically flat regions. Assuming that the water pressure is equal to the ice overburden pressure (e.g., [6], [22], [59]), H_h is defined as:

$$H_h = 0.917 \cdot D + T^{BI}. \tag{18}$$

For the LD dataset, the bottom-right panels in Fig. 10(a), 10(b) and 10(c) illustrate the estimated lake detection probabilities with blue-red colormaps as a function of hydraulic head (right axis). The proposed method reports a high probability of SL presence in correspondence of relatively flat hydraulic

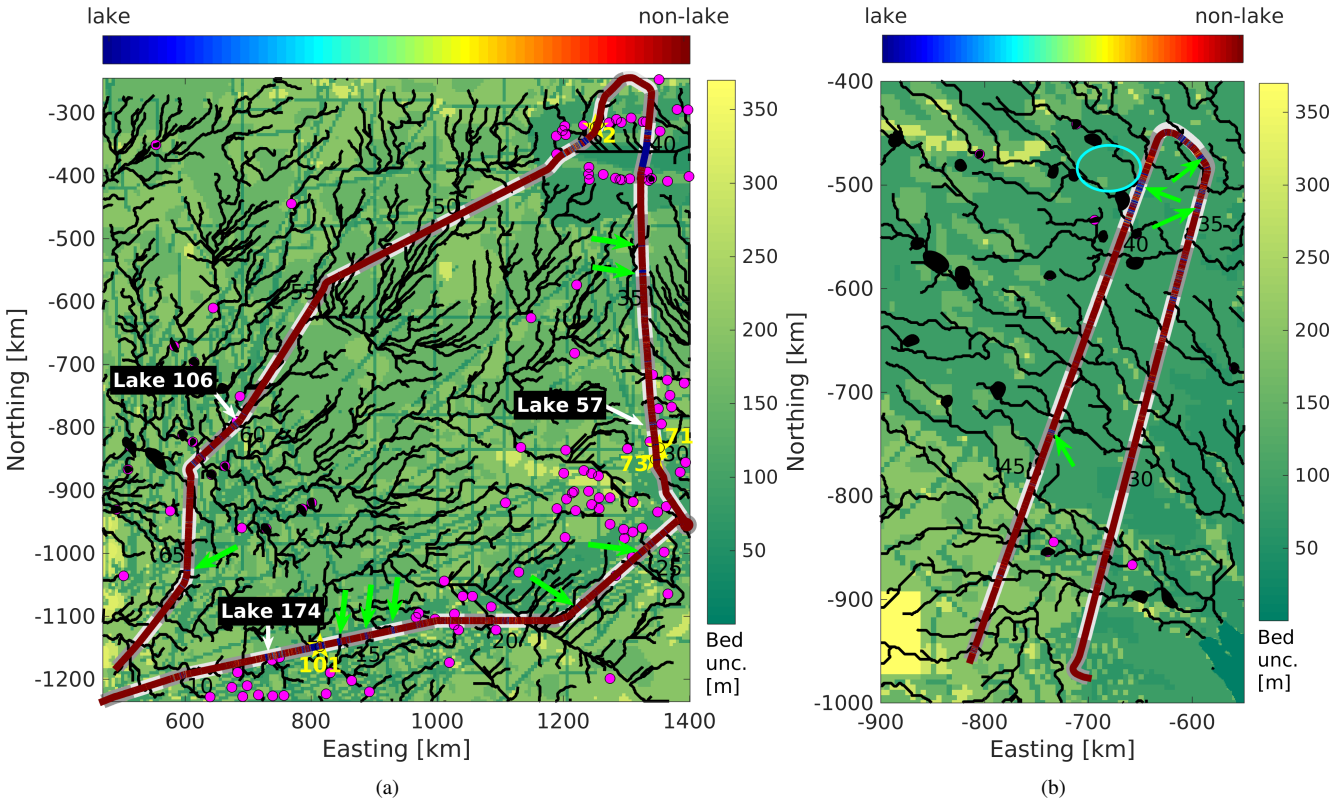


Fig. 12. Output of the proposed method (blue-red colormap) for the (a) LD dataset, and (b) SC dataset. Also provided are the location of the investigated radargrams (alternating white and grey colors), locations of the inventoried radar detected and active SLs (magenta dots and black polygons, respectively) and locations of the reference lakes (yellow stars) overlapped on subglacial water routes (black) computed using the BEDMAP2 dataset [whose uncertainty (from [57]) is depicted in green-yellow colormap]. Green arrows point the locations of other potential SLs detected by the proposed method. Annotated in (a) are also some SLs present in the SLI, which are not confirmed by the visual analysis of the reflections in radargrams, therefore not used in the training phase, but detected, thus confirmed by the method as interfaces with high probability to be SLs. The cyan ellipse in (b) marks a region, in proximity of the locations with a high estimated probability of lake presence, in which a recent work [24] confirmed the existence of a distributed subglacial water sheet.

head, thus sustaining the hypothesis of SLs presence. It is important to highlight that the water ponding condition is necessary, but not sufficient for detecting SLs. Accordingly, basal interfaces can obey the hydraulic flatness condition, without being lake interfaces (this is evident in the bottom panel in Fig. 10(b) which shows several hydraulically flat regions that are not associated with SLs, neither by visually analyzing their reflections in the radargram, nor by the output of the method). Nevertheless, the hydraulic flatness has been widely used in the literature for detecting candidate SL positions (e.g., [6]). Here we use it as an additional variable to assess the performance of the method. This type of analysis has been applied to both datasets. The output of the proposed method suggests the presence of other SLs in the LD dataset. Such results are not provided here for space constraints. However, the locations of these lakes are marked with a green arrow in Fig. 12(a).

For the SC dataset, two examples of radargrams are provided in Fig. 11(a) and Fig. 11(b), along with the extracted features and the estimated SL probability as a function of hydraulic head. In particular, Fig. 11(a) presents a radargram in which the basal interface appears as a flat, relatively shallow and strong reflector, which may be interpreted as a SL reflection. However, the method provides a very low probability of SL presence, which agrees with the hydraulic (non)flatness condition and with the expected geological setting in the region, i.e., the presence of unconsolidated sediments at the basal interface [60]. As another remarkable example, Fig. 11(b) illustrates a radargram acquired near the C1 tributary of Kamb Ice Stream. The hydraulic head is relatively constant and the output of the method suggests the presence of potential SLs, in agreement with [24]. These examples prove the robustness of the proposed method that, even if it was trained only with samples from the East Antarctic Ice Sheet (LD dasaset), it is able to accurately classify the basal interface in the West Antarctic Ice Sheet (SC dataset).

An important observation regards the probabilistic nature of the obtained results, which allows an objective assessment of potential SL locations based on the amount of similarity of the features with those of the reference data. This is evident by analyzing Fig. 10(a), 10(b) and 10(c) in which relatively flat, smooth, bright interfaces, characterized by a high adjusted power (similar to that illustrated in Fig. 7) have higher probability to be SLs than the interfaces with variable topography, increased roughness and lower adjusted peak power. This probabilistic nature of the method offers the possibility to experts analyzing large RS datasets to quickly retrieve only the basal locations with a certain probability (from 0 to 100%) to be SLs. This is extremely important for driving further analyses focused only on regions of interest.

The output of the proposed technique along the entire LD and SC flightlines is provided in Fig. 12(a) and 12(a), respectively. The black lines represent the subglacial water routes [61], estimated by using the BEDMAP2 data [57]. The water routes and the estimated probability of SL presence are overlapped on the estimated ice thickness uncertainty (from [57]). The analysis of the figure points out that the method provided a high probability of SL presence at several flightline positions that intersect the subglacial water routes or that are close to previously inventoried SLs in the LD dataset. There are also some flightline positions for which the method provided a high probability of SLs which do not intersect the water routes or are far from the inventoried lakes. However, these situations mainly correspond to locations for which the estimated BEDMAP2 data show large uncertainties (see Fig. 12), thus compromising the estimation of the water routing. It is also worth recalling that the spatial sampling of the BEDMAP2 data is 1km, whereas the spacing in the azimuth direction of the MCoRDS dataset is 30m. This large difference in spatial sampling prevents a fair comparison of the obtained results with the water routing algorithm. Therefore, this comparison can only provide an approximate qualitative interpretation of the results at large scale, while it is unsuitable for a precise quantitative validation of the method. However, the qualitative analysis points out that the method i) in the LD dataset confirms inventoried SLs and detects other SLs [see Fig. 12(a)], and ii) in the SC dataset confirms the presence of deformable sediments and of the subglacial distributed water sheet detected in [24] near the C1 tributary of Kamb Ice Stream [see Fig. 12(b)].

V. Conclusion

In this paper we presented an automatic technique for SL detection in RS data. The main novelty of the method is the use of a pattern recognition approach based on machine learning to SL detection. As an advantage with respect to existing methods, this approach requires less human interaction, thus resulting in more objectivity and efficiency, enabling its application to large RS datasets. The technique is made up of two main steps. In the first step, features for characterizing the basal interface and discriminating SLs from non-lake interfaces are extracted. In order to capture the peculiarities of the SLs, the feature extraction is performed locally on consecutive basal waveforms. In particular, the features model locally i) the basal topographic variability, ii) the shape of the reflected basal radar waveforms and iii) the statistical distribution of the reflected radar signal. In the second step of the technique, the features are given as input to an automatic classifier based on SVM to perform the SL detection and estimate the probability of SL presence. Remarkably, the method depends on four parameters only, i.e., the along-track and range dimensions of the basal sequences (N_x and N_y, respectively), and the two parameters of the SVM (penalty error term and width of the RBF kernel). The first two depend on the expected minimum SL extent, the discrimination capabilities of the features and the main lobe width of the lake reflected waveforms. The other two are estimated using a standard cross-validation procedure.

We applied the method to two different datasets acquired by MCoRDS in Antarctica, in Lake District (LD) and Siple Coast (SC) regions. Of all the inventoried lakes in the surveyed regions, we validated the presence of four SLs, i.e., lake 71, lake 73, lake 101, and lake Vostok, all clearly visible in the LD dataset. Samples from these lake interfaces and from other non-lake interfaces were used as reference data for training and testing the classifier. The qualitative analysis of the behavior of the extracted features on the reference data confirm the theoretical expectations, thus pointing out both their effectiveness in characterizing SLs and their robustness to subglacial attenuation effects.

The performance of the SVM for the automatic SL detection has been validated both quantitatively and qualitatively. By training the SVM with at least one quarter of the available refence samples, we obtained a classification performance greater than $\approx$97% according to four different metrics. These results are satisfactory and confirm the validity of the proposed approach. More importantly, they also suggest the effectiveness of the method in the glaciological context, i.e., for detecting other potential SLs and update the current SL inventory. Indeed, along the investigated RS flightline, besides the locations of the reference lakes, the method provided several locations with high probability of SL presence. The output of the method at these locations has been validated qualitatively with two approaches, i.e., by visually analyzing the basal reflections in radargrams, and ii) by verifying the hydraulic flatness condition. Both these approaches are in agreement with the output of the proposed method. Moreover, the method provided a high probability of SL presence at the locations of lakes 174, 57 and 106, which were not used in the training phase, thus further proving the effectiveness of the method. Interestingly, although the SVM has been trained with LD samples only, it provided reliable results also on the SC samples. In particular, near the C1 tributary of the Kamb Ice Stream, where a recent work claimed the presence of subglacial water bodies [24], the method provided several locations with a high probability of lake presence. A final remark regards the additional information embedded in the probabilistic output provided by the proposed method. Indeed, the method objectively assigns a probability of SL detection to each sample of the basal interface. This enables experts to easily retrieve locations of interest, with high probability of SL presence where to focus further dedicated analyses.

The obtained results suggest that the method can be an efficient processing tool for the analysis and systematic detection of SLs in large amounts of RS data. Thus, the method can support the glaciological community to i) confirm the presence of inventoried SLs in the analyzed datasets, ii) update the SL inventory from available and upcoming RS surveys, ii) renew previous estimates on the spatial distribution of SLs, their impact on ice sheet dynamics and the probability of microbial subglacial habitats.

As further developments, we aim to study the possibility of extracting other features for basal interface characterization, e.g., specularity content [23]. Moreover, we aim to further validate and test the generalization capabilities of the proposed method by training and testing the classifier with data acquired

in other regions in both Antarctica and Greenland.

Acknowledgement

This research has been funded by the Italian Space Agency (ASI). We acknowledge the use of data and/or data products from CReSIS generated with support from the University of Kansas, NSF grant ANT-0424589, and NASA Operation IceBridge grant NNX16AH54G. The authors would like to thank E. Dalsasso who supported the collection of reference data and K. Matsuoka who provided the attenuation rate of Antarctica [43].

References

[1] G. Robin, C. Swithinbank, and B. Smith, "Radio echo exploration of the Antarctic Ice Sheet," *IASH publication*, vol. 86, 1970.

[2] J. Ridley, W. Cudlip, and S. Laxon, "Identification of subglacial lakes using ERS-1 radar altimeter," *Journal of Glaciology*, vol. 39, no. 133, pp. 625–634, 1993.

[3] I. Tabacco, A. Forieri, A. Della Vedova, A. Zirizzotti, C. Bianchi, P. De Michelis, and A. Passerini, "Evidence of 14 new subglacial Lakes in the Dome C-Vostok area," *Terra Antarctica Reports*, 2003.

[4] R. Bell, M. Studinger, M. Fahnestock, and C. Shuman, "Tectonically controlled subglacial lakes on the flanks of the Gamburtsev Subglacial Mountains, East Antarctica," *Geophysical Research Letters*, vol. 33, no. 2, 2006.

[5] S. Popov and V. Masolov, "Forty-seven new subglacial lakes in the 0-110° E sector of East Antarctica," *Journal of Glaciology*, vol. 53, no. 181, pp. 289–297, 2007.

[6] S. Carter, D. Blankenship, M. Peters, D. Young, J. Holt, and D. Morse, "Radar-based subglacial lake classification in Antarctica," *Geochemistry, Geophysics, Geosystems*, vol. 8, no. 3, 2007.

[7] B. Smith, H. Fricker, I. Joughin, and S. Tulaczyk, "An inventory of active subglacial lakes in Antarctica detected by ICESat (2003–2008)," *Journal of Glaciology*, vol. 55, no. 192, pp. 573–595, 2009.

[8] S. Livingstone, C. Clark, J. Woodward, and J. Kingslake, "Potential subglacial lake locations and meltwater drainage pathways beneath the Antarctic and Greenland ice sheets," *Cryosphere*, vol. 7, no. 6, pp. 1721–1740, 2013.

[9] S. Palmer, J. Dowdeswell, P. Christoffersen, D. Young, D. Blankenship, J. Greenbaum, T. Benham, J. Bamber, and M. Siegert, "Greenland subglacial lakes detected by radar," *Geophysical Research Letters*, vol. 40, no. 23, pp. 6154–6159, 2013.

[10] E. Gudlaugsson, A. Humbert, T. Kleiner, J. Kohler, and K. Andreassen, "The influence of a model subglacial lake on ice dynamics and internal layering," *The Cryosphere*, vol. 10, no. 2, pp. 751–760, 2016.

[11] A. Kapitsa, J. Ridley, G. Robin, M. Siegert, and I. Zotikov, "A large deep freshwater lake beneath the ice of central East Antarctica," *Nature*, vol. 381, no. 6584, p. 684, 1996.

[12] S. Abyzov, I. Mitskevich, M. Poglazova, N. Barkov, V. Lipenkov, N. Bobin, B. Koudryashov, V. Pashkevich, and M. Ivanov, "Microflora in the basal strata at Antarctic ice core above the Vostok lake," *Advances in Space Research*, vol. 28, no. 4, pp. 701–706, 2001.

[13] J. Ellis-Evans and D. Wynn-Williams, "Antarctica-a great lake under the ice," *Nature*, vol. 381, no. 6584, pp. 644–646, 1996.

[14] M. Siegert, N. Ross, and A. Le Brocq, "Recent advances in understanding Antarctic subglacial lakes and hydrology," *Phil. Trans. R. Soc. A*, vol. 374, no. 2059, 2016.

[15] A. Wright and M. Siegert, "The identification and physiographical setting of Antarctic subglacial lakes: An update based on recent discoveries," 2011.

[16] ——, "A fourth inventory of Antarctic subglacial lakes," *Antarctic Science*, vol. 24, no. 06, pp. 659–664, 2012.

[17] A. Rutishauser, D. Blankenship, M. Sharp, M. Skidmore, J. Greenbaum, C. Grima, D. Schroeder, J. Dowdeswell, and D. Young, "Discovery of a hypersaline subglacial lake complex beneath Devon Ice Cap, Canadian Arctic," *Science Advances*, vol. 4, no. 4, 2018.

[18] L. Bruzzone et al., "STRATUS - SaTellite Radar sounder for eArth sUb-surface Sensing," 2016, Proposal in response to "Bando per attivita' preparatoria per future mission e payload di osservazione della terra".

[19] R. Bingham and M. Siegert, "Radio-echo sounding over polar ice masses," *Journal of Environmental & Engineering Geophysics*, vol. 12, no. 1, pp. 47–62, 2007.

[20] R. Shreve, "Movement of Water in Glaciers," *Journal of Glaciology*, vol. 11, no. 63, 1972.

[21] G. Oswald and S. Gogineni, "Recovery of subglacial water extent from Greenland radar survey data," *Journal of Glaciology*, vol. 54, no. 184, pp. 94–106, 2008.

[22] M. Wolovick, R. Bell, T. Creyts, and N. Frearson, "Identification and control of subglacial water networks under Dome A, Antarctica," *Journal of Geophysical Research: Earth Surface*, vol. 118, no. 1, 2013.

[23] D. Schroeder, D. Blankenship, K. Raney, and C. Grima, "Estimating Subglacial Water Geometry Using Radar Bed Echo Specularity: Application to Thwaites Glacier, West Antarctica," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 443–447, 2015.

[24] D. Young, D. Schroeder, D. Blankenship, S. Kempf, and E. Quartini, "The distribution of basal water between Antarctic subglacial lakes from radar sounding," *Phil. Trans. R. Soc. A*, vol. 374, 2016.

[25] T. Jordan, M. Cooper, D. Schroeder, C. Williams, J. Paden, M. Siegert, and J. Bamber, "Self-affine subglacial roughness: consequences for radar scattering and basal water discrimination in northern Greenland," *The Cryosphere*, vol. 11, no. 3, pp. 1247–1264, 2017.

[26] G. Flowers, "Modelling water flow under glaciers and ice sheets," *Proc. R. Soc. A*, vol. 471, no. 2176, 2015.

[27] S. Pattyn, F. Carter and M. Thoma, "Advances in modelling subglacial lakes and their interaction with the Antarctic ice sheet," *Phil. Trans. R. Soc. A*, vol. 374, no. 2059, 2016.

[28] K. Christianson, R. Jacobel, H. Horgan, S. Anandakr-

ishnan, and R. B. Alley, "Subglacial Lake Whillans — Ice-penetrating radar and GPS observations of a shallow active reservoir beneath a West Antarctic ice stream," vol. 331–332, 5 2012.

[29] A. Wright, D. Young, J. Bamber, J. Dowdeswell, A. Payne, D. Blankenship, and M. Siegert, "Subglacial hydrological connectivity within the Byrd Glacier catchment, East Antarctica," *Journal of Glaciology*, vol. 60, no. 220, pp. 345–352, 2014.

[30] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[31] A. Ilisei, M. Khodadadzadeh, E. Dalsasso, and L. Bruzzone, "Automatic detection of subglacial lakes in radar sounder data acquired in Antarctica," *Proc. SPIE 10427, Image and Signal Processing for Remote Sensing XXIII*, 2017.

[32] F. Rodriguez-Morales, S. Gogineni, C. Leuschen, J. Paden, J. Li, C. Lewis, B. Panzer, D. Gomez-Garcia Alvestegui, A. Patel, K. Byers, R. Crowe, K. Player, R. Hale, E. Arnold, L. Smith, C. Gifford, D. Braaten, and C. Panton, "Advanced multifrequency radar instrumentation for polar research," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2824–2842, 2014.

[33] CReSIS, "CReSIS Radar Depth Sounder Data." [Online]. Available: https://data.cresis.ku.edu/data/rds/2013_Antarctica_P3/CSARP_mvdr/20131127_01/

[34] S. Fujita, T. Matsuoka, T. Ishida, K. Matsuoka, and S. Mae, "A summary of the complex dielectric permittivity of ice in the megahertz range and its applications for radar sounding of polar ice sheets," *Phisics of Ice Core Records*, pp. 185–212, 2000.

[35] B. Hubbard and N. Glasser, *Field Techniques in Glaciology and Glacial Geomorphology*. Wiley, 2005.

[36] J. Dowdeswell and M. Siegert, "The physiography of modern Antarctic subglacial lakes," *Global and Planetary Change*, vol. 35, no. 3, pp. 221–236, 2002.

[37] A. Fung and K. Chen, *Microwave Scattering and Emission Models for Users*, ser. Artech House remote sensing library. Artech House, 2010. [Online]. Available: https://books.google.it/books?id=blMmAgAAQBAJ

[38] C. Neal, "Radio echo determination of basal roughness characteristics on the Ross ice shelf," *Annals of Glaciology*, vol. 3, no. 1, pp. 216–221, 1982.

[39] M. Peters, D. Blankenship, and D. Morse, "Analysis techniques for coherent airborne radar sounding: Application to West Antarctic ice streams," *Journal of Geophysical Research: Solid Earth*, vol. 110, no. B6, 2005.

[40] A. Pasmurov and J. Zinoviev, *Radar imaging and holography*. Institution of Electrical Engineers, 2005, vol. 19.

[41] A.-M. Ilisei and L. Bruzzone, "A system for the automatic classification of ice sheet subsurface targets in radar sounder data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3260–3277, 2015.

[42] C. Oliver and S. Quegan, *Understanding synthetic aperture radar images*. SciTech Publishing, 2004.

[43] K. Matsuoka, J. MacGregor, and F. Pattyn, "Predicting radar attenuation within the Antarctic ice sheet," *Earth and planetary science letters*, vol. 359, pp. 173–183, 2012.

[44] D. Schroeder, H. Seroussi, W. Chu, and D. Young, "Adaptively constraining radar attenuation and temperature across the Thwaites Glacier catchment using bed echoes," *Journal of Glaciology*, vol. 62, no. 236, pp. 1075–1082, 2016.

[45] D. Schroeder, C. Grima, and D. Blankenship, "Evidence for variable grounding-zone and shear-margin basal conditions across Thwaites Glacier, West Antarctica," vol. 81, pp. WA35–WA43, 01 2015.

[46] J. MacGregor, D. Winebrenner, H. Conway, K. Matsuoka, P. A. Mayewski, and G. Clow, "Modeling englacial radar attenuation at Siple Dome, West Antarctica, using ice chemistry and temperature data," *Journal of Geophysical Research: Earth Surface*, vol. 112, no. F3, 2007.

[47] J. MacGregor, J. Li, J. Paden, G. Catania, G. Clow, M. A. Fahnestock, S. Gogineni, R. Grimm, M. Morlighem, S. Nandi *et al.*, "Radar attenuation and temperature within the Greenland Ice Sheet," *Journal of Geophysical Research: Earth Surface*, vol. 120, no. 6, pp. 983–1008, 2015.

[48] A. Fung, K. Chen, and K. Chen, *Microwave scattering and emission models for users*. Artech house, 2010.

[49] Y. Jin, *Theory and approach of information retrievals from electromagnetic scattering and remote sensing*. Springer Science & Business Media, 2006.

[50] D. Moore, *The basic practice of statistics*, vol. 2.

[51] P. Westfall, "Kurtosis as peakedness, 1905-2014. RIP," *The American Statistician*, vol. 68, no. 3, pp. 191–195, 2014.

[52] A. El-Zaart and D. Ziou, "Statistical modelling of multimodal SAR images," *International Journal of Remote Sensing*, vol. 28, no. 10, pp. 2277–2294, 2007.

[53] M. Welling, "Robust higher order statistics," in *AISTATS*, no. 3, 2005, p. 7.

[54] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*. John Wiley and Sons, Inc, 2009.

[55] A. Ben-Hur and J. Weston, "A users guide to support vector machines," Tech. Rep., 2012.

[56] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*. MIT Press, 1999, pp. 61–74.

[57] P. Fretwell et al., "Bedmap2: improved ice bed, surface and thickness datasets for Antarctica," *The Cryosphere*, vol. 7, no. 1, pp. 375–393, 2013.

[58] J. Li, J. Paden, C. Leuschen, F. Rodriguez-Morales, R. Hale, E. Arnold, R. Crowe, D. Gomez-Garcia, and P. Gogineni, "High-altitude radar measurements of ice thickness over the Antarctic and Greenland ice sheets as a part of operation IceBridge," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 742–754, 2013.

[59] G. Clarke, "Subglacial processes," *Annu. Rev. Earth Planet. Sci.*, vol. 33, 2005.

[60] H. Engelhardt, N. Humphrey, B. Kamb, and M. Fahnestock, "Physical conditions at the base of a fast moving Antarctic ice stream," *Science*, vol. 248, no. 4951, pp. 57–59, 1990.

[61] C. Greene, D. Gwyther, and D. Blankenship, "Antarctic Mapping Tools for MATLAB," *Computers & Geosciences*, vol. 104, pp. 151–157, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0098300416302163