Referee #2

Review of the article 'Review article: Feature tracing in radio-echo sounding products of terrestrial ice sheets and planetary bodies' by Hameed Moqadam and Olaf Eisen.

- We thank the reviewer for their effort to go through our manuscript and provide the feedback.

The paper proposes a review of the literature methods for detecting internal reflection horizons related to englacial stratigraphy and several glaciological applications to monitor the cryosphere. The structure of the paper is very complex and fragmented, making the sections contain a lot of repetitions from other sections. The English is good, but sometimes very informal, and there are typos, so it should be improved to match the journal's quality. Here are the general comments on the paper:

The paper's aim is unclear, whether to review the automatic methods for radargram analysis or to examine IRH/ice layer tracing. The paper focuses on the importance of IHR tracing (abstract, intro, and background sections). Still, it also presents methods for target detection and segmentation that are unrelated to edge/IHR/layer detection. It is also unclear what criterion is used to select the methods the paper analyzes (the abstract considers those applied to RES data, and sec 4 claims only those to analyze radargrams). What about methods for GPR data?

- Thank you for pointing this out. As the main objective of the paper is a review of methods that trace IRHs in radargrams, we have included methods that segment radargrams and find targets as well. This is due to the fact that these methods are quite similar to each other. Therefore, to keep the review article comprehensive, we have decided on including those. This is mentioned in Lines ~65, ~156 and in more details in Lines 414 – 418.
  RES (or radar) is an overarching terminology, which includes all airborne or ground-based active electromagnetic methods in the radio-frequency range. Details can be found in Schlegel et al. (2023, Ann. Glac.). We follow their terminology in our manuscript and, for the sake of brevity, will not discuss all possible namings of radar methods again.

The paper's headings are very confusing and do not give an overview of the section's content.

- Thank you for this comment, however it is not clear which of the headings are meant by this. In the revision we will revisit the headings and outline the sections and substructure more clearly.

General comments on Sec 3. The criterion used to divide the data analysis methodology into different sections (sec 3.1-3.12) is unclear. Considering the large number of methods, I would expect a clusteringwith a clear logic. Moreover, the writing should be improved to increase fluidity. Further, the sections miss an analysis of the pros and cons of the methodologies, e.g., which are the ideal and worst conditions, which are the method's core hypothesis, how should be the radar data preprocessed, how are the radar data non-ideality handled (e.g., speckle), which are the limitations, and on which data was the method validated (e.g., planetary, terrestrial, high or low spatial resolution, adaptability to other data).

- Thanks for this point. We will make the text on section 3 more simple to make it more readable.
  About stating the pros and cons, and best and worst conditions, considering both the scope and length of the paper, we have decided not to elaborate on those.
  About pre-processing methods of the radar data, as this is more dataset-dependent than method-dependent. We therefore do not consider it the scope of this review article.
  Which data were used for each of the studies are mentioned in section 4 in the individual description of the papers.

Moreover, I would expect more figures (e.g., flow charts) and formulas to understand each method's details better. Finally, I am very confused about whether the paper aims to show the methodologies for radargram automated analysis, IHR detection, or layer tracing. The paper focuses on the importance of IHR and related applications, while section 3 presents a list of automated methods used for analyzing radargrams.

- Considering the current length of the paper, adding figures for each (or even some) of the papers would make the article longer and less favourable for readers, without providing considerable added value in our opinion.
  The methods introduced in section 3 are the major method that are used for IRH tracing and radargram segmenting in the papers that are described in section 4. Nevertheless, we will consider in the revision to add a restricted number of additional figures to make the contents more clear.

General comment on Sec 4. It is unclear why Sec 3 and Sec 4 are divided and not merged, given the lack of details of Sec 3 and the repetition of concepts already explained in Sec 4. Moreover, consider using the taxonomy of sec 4 also for sec 3. Instead of Sec 3 and Sec 4, I suggest having three sections with the taxonomy of 4.1,4.2, and 4.3 that present the computer vision theory (an improved version of what is in sec3) and the application to the radar data (an improved version of what is in sec4). Moreover, the session lacks details, formulas, and images to understand how the methods work. In general, the methods' descriptions lack a discussion on the pros and cons, limitations, and if/how they can be improved. Also, the connection between paragraphs is missing, and the section looks like a list of methods without any critical analysis.

- The suggested taxonomy is very appealing and we will consider it in the revised version.
  Including figures and formulae, and stating pros and cons would make the article longer and are outside the scope of the paper, see above.
  Section 4’s aim is to summarize the research done on IRH tracing and radargram segmenting and not suggesting ways to improve each of the methodologies. In that context we do not aim for a textbook-style of article, but a comprehensive review of what has been tried and applied already to help other scientists to focus their efforts and avoid doubling

Further, the methods are applied to very different types of data (SHARAD, MCoRDS, HiCARS, accumulation data) without describing the properties of the data in terms of noise, resolutions, acquisition geometry, and so on. Consider adding information on the datasets used and their characteristics (maybe in sec 2.2). Further, the paper's aim is unclear (IRH detection or segmentation or automatic analysis of radar data). At the end of each section (i.e., 4.1, 4.2, and 4.3),

the methodologies should be discussed to understand the limitations that forced the development of novel techniques.

- Thank you for this accurate point. However, adding details about all the used datasets is out of the scope of this paper since the scope is not about datasets but the methodologies.
  The aim of the section 4 is stated at the beginning of the section.

General comment on Sec 4.3. This section moves from segmentation to layer detection. Consider fixing an aim to the paper and sticking to it. There are few indications on the data types used to validate the methods (and how the data-specific characteristics are handled, including noise and resolutions). Moreover, consider dividing the section into subsections (e.g., DL for layer detection, DL for semantic segmentation). The section lacks i) a critical analysis of the methods, ii) information on the type of data the methods are validated, and iii) information on the computational load (given the high computational cost of the DL algorithm).

- The datasets are mentioned for the summarized studies. Also pre- and post-processing methods that were used for each paper are also mentioned when it seemed important or specific.
  The scope of the section is not to analyze the methods but to give the reader an overview of the method implemented in each study.
  Dividing section 4.3 into DL for layer detection, DL for semantic segmentation is a good idea and we will consider that in the revised version.

Sec 5, discussion. To better understand the validity and differences of the methods, I would expect a qualitative/quantitative comparison.

- This is a very nice suggestion. But a qualitative/quantitative comparison of is way beyond the scope of a review article and is a good idea for writing another completely different study, as it would require an implementation of various methods to the same benchmark dataset. Obviously, that is beyond the scope of a review.

Sec 6. The paper's aim is changed to 'Consequently, this review aims to provide a contemporary overview of advancements in this field of research over the past two decades.' The section has a lot of repetition compared to sec 4 and 5. Consider merging the sections. Finally, the section misses the expected future directions in the IRH detection/ automatic analysis of radargrams.

- Thanks for the suggestion, we will consider that.

Below are the detailed comments (Pg stands for page and Ln/ln for line).

Pg 2, ln 55, Antarctica is larger than Greenland but poorer in terms of data. There are large areas without any acquisitions.

- Thanks for the suggestion, we will implement that.

Sec 2.1 repeats concepts already defined in the introduction. Given the length of the paper, consider removing the repetitions. Ln 96-100 lack details on the IHR formation and give partial information through examples. Consider improving this paragraph.

- Thanks for the suggestion, we will consider that.

Pg 4, ln 103. Consider adding references for z-scope radargrams, e.g., 'Schroeder, Dustin M., et al. "Radiometric analysis of digitized Z-scope records in archival radar sounding film." Journal of Glaciology 68.270 (2022): 733-740.'

- Thanks for the suggestion, we will consider that.

Sec 2.2, line 106. Consider removing the repetitions on the reflection generation: 'the aforementioned characteristics such as presence of impurities, acids, and changes in ice-crystal orientation cause reflections, and when they are laterally coherent, they appear as horizon'.

- Thanks for the suggestion, we will change that.

Sec 2.2, ln 107-110. Consider defining a radargram as a 2D matrix of N_T traces and N_S samples. The definition of the radargram pixels as indicating the power/amplitude is misleading. Consider that radargram may also be complex, meaning there is the amplitude and phase radargram or the imaginary and real radargram.

- Thanks for the suggestion, we will change that accordingly.

Figure 1.a and 1.b. Consider showing the figures in dB for a better visualization. Moreover, figs 1.a f lack the x and y axes ticks and labels. Is the x axis of fig 1.a valid also for fig 1.b-f? If so, clarify it in the caption. Moreover, clarify that the radargrams are presumed in the caption.

- They are the same radargram section, we will clarify this in the revised version.

Sec 2.3, ln 132-133. Repetition of the ice layer generation.

- Thanks for the suggestion, we will remove that.

The title of sec 2.4 is misleading as the section is about the information the englacial stratigraphy provides.

- Thanks for this important point. We will find a better title.

Sec3, ln 156. 'In this section, we briefly overview the methods applied to tracing IRH and segmenting radargrams.' This sentence is not in line with the previous section of the paper, as the focus was only on IHR and not on the identification of ice-sheet targets (i.e., segmentation). Consider improving the sentence.

- We will go thought the paper and make it more uniform in terms of these two applications.

Pg 7, ln 159. The sentence 'Given the versatile application of RES across various domains,' lacks a reference.

- Thanks for the suggestion, we will add a reference.

Sec 3 ln 160-165. Repetition of concepts described in section 2.4.

- Thanks for the suggestion, we will remove that.

Pg 7, ln 166. This sentence 'Constructing an automated tracing method for RES encounters a significant challenge when dealing with closely spaced layers.' lacks justification.

- This is an observation that we have. We will try to bring up better arguments for this.

Pg7, ln 179. What about folded or interrupted layers (e.g., those in the basal area)?

- This falls under the last one (complex englacial structures), but perhaps it is worth mentioning separately.

Pg 7, Ln 181. Horizontal or vertical resolution? Justification?

- We will add clarification for this.

Pg 7, Ln 182. Large SNR should be small SNR?

- Thank you for pointing it out, we will fix this mistake.

Pg 8, ln 163-164. Repetition on the motivations for tracing IHR.

- That is true, we will remove repetitions in the revised version.

Pg8, ln 183. 'We will give a short summary of the methods that have been utilized in mapping and segmenting radargrams. The provided method summaries are intended to aid understanding of the timeline of methodologies in section 4, to make readers more aware of the underlying components or procedures of each method.' I am unsure how this sentence relates to the previous part of the section that focuses on tracing IHR.

- This is a good point as the paragraph is not very coherent, we will improve that in the revised version.

Sec 3.1, cross-correlation and peak following. This section lacks the claim of the strong hypothesis that the ice stratigraphy is expected to be constant and horizontal (i.e., without abrupt changes in the steepness). What about the basal region, where shadows mask the reflections?

- Thanks for the suggestion, we will consider adding or clarifying the points.

Sec 3.2, filter. The reference Ilisei and Bruzzone does not refer to canny filtering (it's a statistical analysis of subsurface targets). The same is true for (Freeman et al., 2010), which uses morphological filters and thresholding). Consider removing the reference and description to canny filter and be more general about filtering and thresholding. Moreover, considering the complexity and hypotheses needed for thresholding, I expect at least a sentence discussing it. Finally, this section should at least refer to speckle and how it is tackled, given that speckle can be seen as an abrupt change of intensity of the pixels and thus very visible to canny filter.

- Thanks for the suggestion, we will consider a better description for Canny and include a mention of the thresholding and its significance.

Sec 3.3. improve the computer vision description of Snake.

- Thanks for the suggestion, we will consider that.

Section 3.4. What is this reference to 'cfd, 2019'? Missing justification to the sentence 'making it well-suited for the intricate analysis Radargrams'. It is unclear why this methodology is presented even if it has not been applied to radargrams. Consider removing the section.

- The reference name is a mistake and we will correct it.

Sec 3.5. Statistical analysis is not a method. This section should be improved as it does not detail how the analysis works. The section is also misleading as Rayleigh and Nakagami distributions are valid only under specific hypotheses (e.g., target analyzed and data type). Finally, this section concerns segmentation/target detection, not layer tracing. How is it related to the other sections?

- As previously stated, we are also considering target detection and segmentation methodologies. We will improve the text in 3.5.

Sec 3.6. Consider improving the English (too informal). I would not call a method based on an LPF and thresholding robust (… this method, although being robust...). What about the speckle? In general, the section lacks methodological details. Also, this is very similar to the method in 3.2; consider joining them or explaining the difference between the sections more clearly.

- Thanks for the suggestion, we will consider that.

Sec 3.7. There is no reference to a paper analyzing radargrams. Consider removing the section.

- Thanks for the suggestion, we will consider that.

Sec 3.8. Missing reference to Xiong, Siting, Jan-Peter Muller, and Raquel Caro Carretero. "A new method for automatically tracing englacial layers from MCoRDS data in NW Greenland." Remote Sensing 10.1 (2017): 43. And related works. Moreover, I expect formulas to help me better understand the methodologies.

- Thanks for the suggestion, we will add the references.

Sec 3.9. Most of the papers cited are not related to radargram analysis.

- In section 3, we describe the methods also in a general sense, that is why some of the references in this section are not related to radargram analysis.

Sec 3.10. The papers cited are not related to radargram analysis.

- In section 3, we describe the methods also in a general sense, that is why some of the references in this section are not related to radargram analysis.

Sec 3.11. Consider improving the section on SVM. There is plenty of work to be done for segmenting radargrams. Moreover, the description of SVM, which is very wrong (e.g., SVM is also multiclass), should be greatly improved. The meaning of this sentence 'SVMs are able to identify the optimal surface, mitigating overfitting during training' is unclear. What is the optimal surface? Also, overfitting mostly depends on the training sample numbers and representativity. Missing the motivations that pushed the community to move to deep learning (i.e., the necessity of manually designing the features).

- This is a good point and we will change this subsection in the revised version.

3.12. The section is confusing; for example, I would define deep learning before explaining representation learning. Consider improving the general description of deep learning and explaining why it is important for analyzing radargrams. In the sentence 'DL performs such tasks using multiple levels of non-linear modules, transforming these representations from raw to higher and more abstract level', what does 'representations' refer to? The sentence 'The advantage of semi-supervised learning is that it does not require a large amount of labeled data, but there is the danger

of learning irrelevant features' is not true. The problem is overfitting. In the sentence 'The third class is unsupervised learning. As the name suggests, the learning procedure is based on finding representations without help of known targets', known targets should be labeled images/data. Finally, this section lacks a description of the main disadvantage of DL, which is the overfitting and poor generalization capability of the network with small labeled datasets.

- This is a good point and we will change this subsection in the revised version.

Sec 3.12.1. Several complex concepts are not explained while describing the so-called general architecture of the neural network. In general, the paper shows inconsistent levels of detail in the description of computer vision concepts. Basic concepts (e.g., supervised, semisupervised, unsupervised) are described in detail, while the neural network layers are just mentioned. The most critical part of DL is the training and definition of the loss function. I would expect at least to mention how the network training works.

- We will change this subsection in the revised version.

Considering their usage in 3.12 and 3.12.1 representations and feature maps indicate the same concept. Consider being consistent.

- We will use the same terminology.

Sec 3.12.2. Consider removing repetitions about U net to improve the fluidity of the paragraph. As highlighted in the paper, autoencoders are trained to reconstruct the network's input, i.e., the network's output should be the same as the input. UNet is used to extract semantically meaningful features that are not like the input of the network (e.g., radargram, bio image). Consider being clearer on the paper and improving the section heading. In this direction, some sentences are misleading and lack a justification (e.g., 'In addition, since the basic idea of an autoencoder architecture is to have the same input and output dimensions, autoencoders are a good choice for segmentation task').

- We will change this subsection in the revised version.

3.12.3. This section lacks the network adaptation for the radargram characteristics (different from those of the computer vision/optical data), e.g., 1 channel instead of 3, and not additive noise.

- We will change this in the revised version.

Sec 3.12.4. 'Moreover, it uses side outputs compensating for the absence of deep supervision, which is a characteristic of fully convolutional neural networks.' How? The meaning of the sentence 'HED considers edge detection as a holistic problem (global image-to-image mapping).' is unclear. Missing details to understand the logic of the method.

- We will change this sub-chapter in the revised version.

Sec 3.12.5. It is unclear if the network extracts features at different scales or analyzes data acquired over the same area with different (spatial?) resolutions. How is it done?

- We will change this sub-chapter in the revised version.

3.12.6. 'Simply put, the input of each node is a combination of input and the hidden state of the same node from the previous time step (Goodfellow et al., 2016).' Too informal. From the paragraph, it is not clear how RNNs work.

- We will change this sub-chapter in the revised version.

Sec 4. The title is misleading. This paper aims to identify englacial stratigraphy. Consider being consistent and using the same terminology—are you identifying IHR, the englacial stratigraphy? What about segmentation and target detection?

- As said previously, other similar methods are considered as well.

Pg16, ln 401. What is feature referring to in the sentence 'Automatic feature detection methods'? Pay attention to that 'features'indicated the output of CNN.

- Noted, we will consider that.

Pg 16, ln 408. 'Such studies include a variety of approaches such as neural networks (Reichman et al., 2017)'. Which type of NN and training are used?

- We will modify the paragraph, as stated before.

Ln 410-412. I am not sure about the meaning of this sentence 'However, as a result of radar systems differing in frequencies and waveform characteristics (thus resolution and penetration depth), studies applied to GPR and RES systems over mediums other than ice, do not provide considerable insights.'.

- This is the reason why we do not include radar studies over mediums other than ice.

What is the difference between radargrams (acquired how? Spacecraft? aircraft?) and GPR and RES (radio echo sounding?)? Also, the abstract claims that 'we discuss a variety of methods which were developed or applied to RES data over the last decades, including image processing, statistical techniques, and deep learning approaches.'.

- The radargram differences fall under the categories of datasets, which is out of the scoepe of this review article.

Ln 414-418. 'In a number of studies radargrams were analysed to find different segments or subsurface targets (e.g. englacial boundaries, EFZ, basal units) and classes of events in each radargram (e.g., Donini et al., 2019; Goldberg et al., 2020; García et al.2021, 2023). Even though we focus on the methods for mapping englacial ice structure and tracing IRHs and/or layer boundaries, we also take a look at studies done to detect regions and targets in radar products since those are, in terms of methodology, in close vicinity to stratigraphy mapping endeavours.' If this is the case, it should be claimed in the abstract, title, introduction, etc.

- We will consider adding this.

Ln 422-424. Repetition. This should go into the motivation for IHR identification, not here.

- We consider this repetition and remove them.

Ln 426. I am not sure that I would call filtering and transformation computer vision-based methods, as DL and SVM are computer vision. Consider using classical/traditional machine learning methods or something similar. The heading of the following sections should be the same as the bullet list.

- We will consider better titles.

Ln 487. Feature is used as geological target/IHR. Pay attention to the fact that computer vision has a different meaning. Thus, trying to be consistent in the paper.

- As previously mentioned, we will make sure to be clear on this.

Ln 485 (Ferro and Bruzzone (2013)) I would specify that this work is on planetary data (sharad) given that most of the other works are on aircraft data (e.g., CREeSIS-MCords). Planetary data have different types of noise and radiometric characteristics than aircraft data.

- We will consider to add this note.

Ln 504. 'A user is required to determine the number of visible layers initially.' Manually?

- As mentioned in the original paper, in their training data, an expert counted the visible IRHs.

Ln 508, Panton 2014. The methods are not described.

- The method is mentioned, i.e. snake.

Ln 640. Consider adding the comparison to the paper 'The modification makes the method more appropriate for a wide range of radargrams, based on comparison with results of Crandall et al. (2012); Lee et al. (2014); Rahnemoonfar et al. (2017a).'

- We will consider to add this suggestion.

Ln 656.' After statistical analysis, different classes are represented by pdf. This is partially true. The authors designed a set of manual features (not only statistical features) that are extracted for each pixel in the radargram and then analyzed with SVM. The work was improved in Donini, Elena, et al. "An automatic approach to map refreezing ice in radar sounder data." Image and Signal Processing for Remote Sensing XXV. Vol. 11155. SPIE, 2019.

- We will consider this.

Ln 671, 'Going away from classification of regions and ice-base and ice surface IRHs to tracing internal IRH' is too informal.

- Thanks for pointing this out.

Ln 714. Foci -> focus?

- We will correct this.

Ln 713 'In this subsection, summaries and main points of the studies that used deep learning-based methods are presented. We would like to note that although the primary foci of some of the the works e.g. Donini et al. (2021, 2022c); Garcia et al. (2021); García et al. (2023); Ghosh and Bovolo (2022a, 2023b) lie in radargram region segmentation, they are included in this review because of their methodological relevance for the overall objective.' I would change the paper's objective to

methods for the automatic analysis of radargrams. Otherwise, explaining how to perform segmentation does not make sense if the paper's aim is layer detection.

- This point has been mentioned several times already and we will add this to the abstract and introduction.

Ln 755. The sentence 'Their initial stage is to remove the noise by using bilateral filtering.' Should emphasize that the method cannot handle the radar's noise characteristics and needs to apply strong preprocessing to mitigate speckle. This contrasts with the method Donini et al., and Garcia et al. used to manage the noisy properties within the network. The paper, in general, lacks this kind of critical analysis.

- The scope of this review article is giving the readers an overview of all the methods that haven been implemented for IRH tracing and radargram segmentation. The suggestions is quite detailed, but we will consider to add some information in this respect. However, our intention is to point interested readers to the published material so that they refer to those when they need more insights.

Ln 984. Some solutions have been proposed to overcome this (Cai et al., 2022; García et al., 2023; Moqadam et al., 2024). Also, Donini et al. propose a pre-training to set the network parameters to not random values.

- We will consider adding the paper as well. This paper has been cited other times in out manuscript as well.

Ln 929. The section on unsupervised segmentation of radargrams lacks the reference to Donini, Elena, et al. "Unsupervised semantic segmentation of radar sounder data using contrastive learning." Image and Signal Processing for Remote Sensing XXVIII. Vol. 12267. SPIE, 2022.

- We will consider adding this paper. However, as it is not open access it is difficult to find out more about the paper. Also we could not access it through our institution’s access either.