Dear Editor and Reviewers,

We would like to thank the reviewers for their helpful and constructive comments and have included a point-by-point response below. Reviewers' original comments are included in italic text and our responses to the reviewers are included in plain text. We have highlighted text changes that were incorporated into the manuscript in blue and included the line numbers (associated with the clean version) for ease of review. The additional tracked-changes pdf version shows text that has been removed as being crossed-out, red-colored text and newly inserted text in blue.

Reviewer A:

This paper makes a valuable contribution to the continued monitoring and understanding of whale habitat usage and migration patterns in the LSLS. The application of the Plourde and Nedimović, 2022 [PN22] methodology is utilised with minor tuning, and a convincing catalogue of Fin and Blue whale calls is produced. Quantitative analysis is performed on the catalogue, and a Machine Learning (ML) algorithm is implemented and validated using the catalogue. Finally, this spatial and temporal distributions of the catalogue are interpreted, particularly with reference to the earlier catalogue devised in PN22. Here, the paper presents a number of interesting findings, making several compelling arguments that draw on the literature to interpret the spatial and temporal variations discovered.

The discussion section of this paper appears sound and draws some very interesting conclusions when analysing the newfound detections. As far as I can tell, this main contribution is incremental, applying PN22 on new data and exploring the results: a valuable contribution in its own right. The main significant limitation of the paper in its current form is the treatment and discussion of the ML techniques used.

It would be helpful to set out a clear motivation for utilising ML from the outset. In particular, what do you hope to gain compared the PN22 method? As far as I can tell from the paper, you train the ML algorithm on a dataset purely constructed using PN22. The accuracies you report, while positive, seem to suggest that PN22 outperforms the ML method in identifying whale calls. If this is so, what is the benefit of the ML algorithm? Is it significantly more efficient, or does it generalise to more difficult to detect whale calls? If ML is performing better by some metric (which I may be misinterpreting from your results), make how explicit in your results section.

To explore these questions, a more in-depth study of the limitations of the two algorithms would be desirable. How does each algorithm perform at different noise levels? You report accuracy of the ML algorithm trained on the catalogue, but do you know the false positive rate of the

catalogue itself? I assume you have checked that it is high quality. Table 2 indicates that some filtering is required to ensure good performance of the ML model, but is this because the ML algorithm can't deal with low SNR calls or because not filtering for high SNR leads to the inclusion of lots of false positives? I think an exploration of the regions where ML fails (could just be in the supplement) would be useful. I can see several potential benefits, such as reducing reliance on repeating calls for detection; it would be great to explain and demonstrate such benefits empirically and quantitatively.

Answer: The Machine Learning (ML) algorithm presented in the first manuscript submission is still in the prototypical stage. We agree with the comments by both reviewers that further tests are needed for the ML algorithm to be fully functional thus useful for the community. As such, we decided to remove ML from the methods and results sections, but focused our last discussion section on the limitations of the current method used in this study and the motivation and potential to construct a whale call dataset annotated by human experts and develop a ML model trained on it , which will be the focus of a future study.

In section 4.5, L. 348-366 we added: "The Plourde and Nedimović (2022) detection algorithm used in this study has several limitations. First, this algorithm cannot be evaluated properly due to the lack of ground-truth labels. For instance, it is impossible to know its accuracy and whether the algorithm predicts a true positive or false positive and hence its recall and precision. These metrics, as well as their derivatives (e.g. F1-score), are fundamental to our understanding of its performance and its application to real-world scenarios. To address this limitation, we propose constructing a whale call dataset annotated by human experts, which can serve as a benchmark to evaluate Plourde and Nedimovic (2022) algorithm as well as potential detection algorithms developed in the future. One possible way for dataset construction is to filter the data catalog from this study and Plourde and Nedimovic (2022) manually. This new dataset will also pave the way for deep learning-based systems. Second, the detection in Plourde and Nedimovic (2022) is made based on a group of individual whale calls with each call group defined as a detection. This prohibits the model from being applied to scenarios where only individual whale calls are available (e.g. other calls are lost due to data transmission issues). A deep learning-based model trained on call-level (in contrast to detection-level) data can mitigate this problem. Third, many automatic detection algorithms, including the one used in this study, for monitoring whale calls do not consider variable acoustic conditions (Madhusudhana et al., 2021). The LSLS land seismometer whale detection catalog, combining the results from this study and Plourde and Nedimovic (2022), consists of nearly 7 years of labeled stereotyped fin whale and blue whale calls. These whale call signals were detected over a wide spatial extent from the Estuary to the Gulf of St. Lawrence, throughout all seasons and over multiple years. The use of a deep learning-based model trained on this

dataset (with annotations) could be particularly useful to detect and classify whale calls exposed to different environmental conditions, or signal context."

The rest of the work, particularly the interesting conclusions drawn in sections 4.1 and 4.2 regarding interpretation of the results, speaks for itself as a valuable contribution to the literature, both in terms of the marine biology context specific to the LSLS, and for demonstrating the effectiveness of the PN22 method for whale call detection using land seismometers more generally.

Some more minor comments:

Some more details (again, can be left to the supplement) on the ML architecture should be included.

1. *How deep was the LSTM? Hidden layer sizes and number of layers, etc.*
2. *Did you perform any hyperparameter tuning? Was training stable for each of the datasets you explored?*
3. *How was it trained (LR / optimizer)*
4. *Did you try any other architectures, such as CNNs which are commonly used for processing spectrogram data (both in and outside of seismology)?*

Answer: As we have removed the ML algorithm from the main results (see reply above), we will not discuss the technical details in this paper. Here we provide answers based on our work on the prototype model.

1. The architecture consists of 3 layers of LSTM followed by two MLPs, one for classifying call types and one for call time regression. The hidden size is 128.
2. Yes. We performed hyper-parameter tuning. Here is the table of hyper-parameters investigated:

Hyper-parameter	Type	Range/Categories	Step	Optimal value	Additional Info
bidirectional	categorical	true, false	N/A	false	if use Bidirectional LSTM
num_layers	int	1 to 5	1	3	Depth of LSTM
hidden_dim	int	64 to 256	32	128	Hidden layer size
lr	float	0.0001 to 0.01	N/A	0.001	Learning rate Search in log space

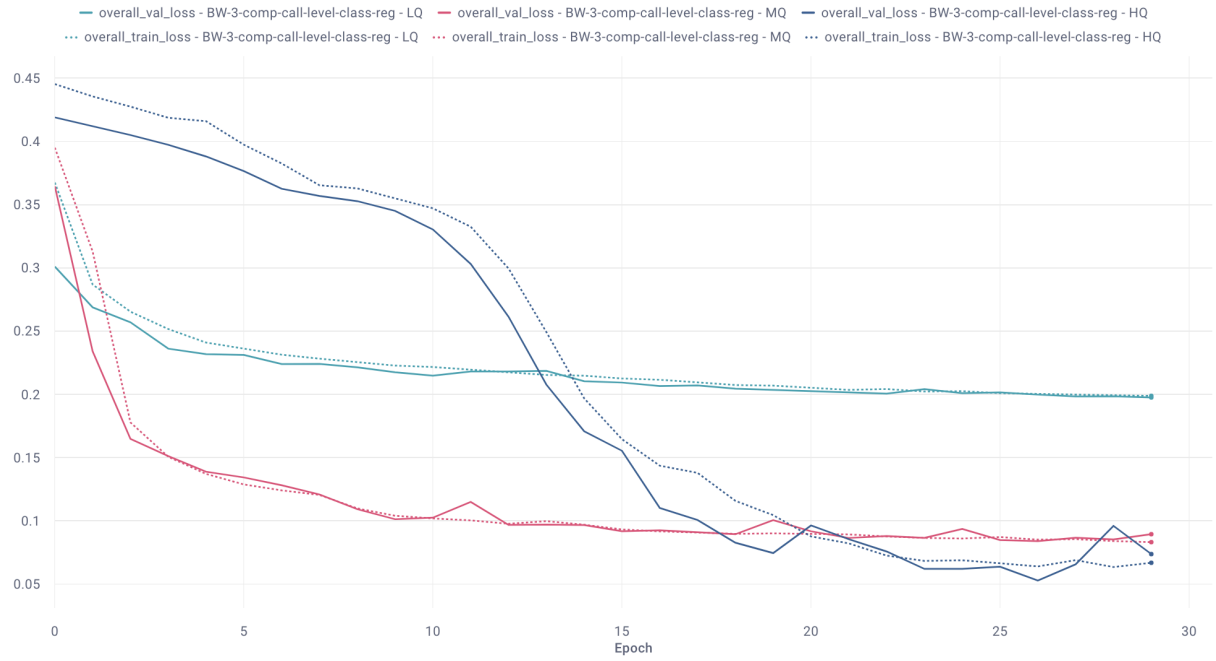
reg_loss_weight	float	0.1 to 0.5	0.1	0.5	λ in the loss function (see below)

The target is a joint task of a binary classification task on discriminating positive and negative examples and a regression task on predicting call time. The loss function to be minimized is:

$$L = L_{cls} + \lambda I_{y_{pred}=1} L_{reg}$$

where L_{cls} is the binary cross-entropy loss, L_{reg} is the L_1, $I_{ypred=1}$ is the indicator function and λ is the weighting factor.

The training is stable for each of the datasets. The following figure shows the training history of BW datasets with different quality:



3. The model was trained for 30 epochs with a batch size of 64 using AdamW optimizer with a OneCycle Learning rate scheduler for fast convergence (https://arxiv.org/pdf/1708.07120).
4. We investigated 1D-Unet, a CNN-based architecture originally designed for image segmentation and has been applied in seismograph data in previous studies (e.g. https://academic.oup.com/gji/article/216/1/261/5129142#123811673). **It models waveforms directly.** But we don't include the results as they are not competitive. We consider this could be due to two reasons:
    a. The model requires precise determination of whale call starting and ending time (similar to P/S arrival times picked by seismologist), which are not available in the current dataset. The PN22 method returns the center times of whale calls.
    b. The amount of data in this study is not big enough to train a model like U-Net. We have only ~10k and ~1k high quality FinWhale data and BlueWhale data while the study mentioned previously trained a 1D-UNet using ~800k earthquake samples.

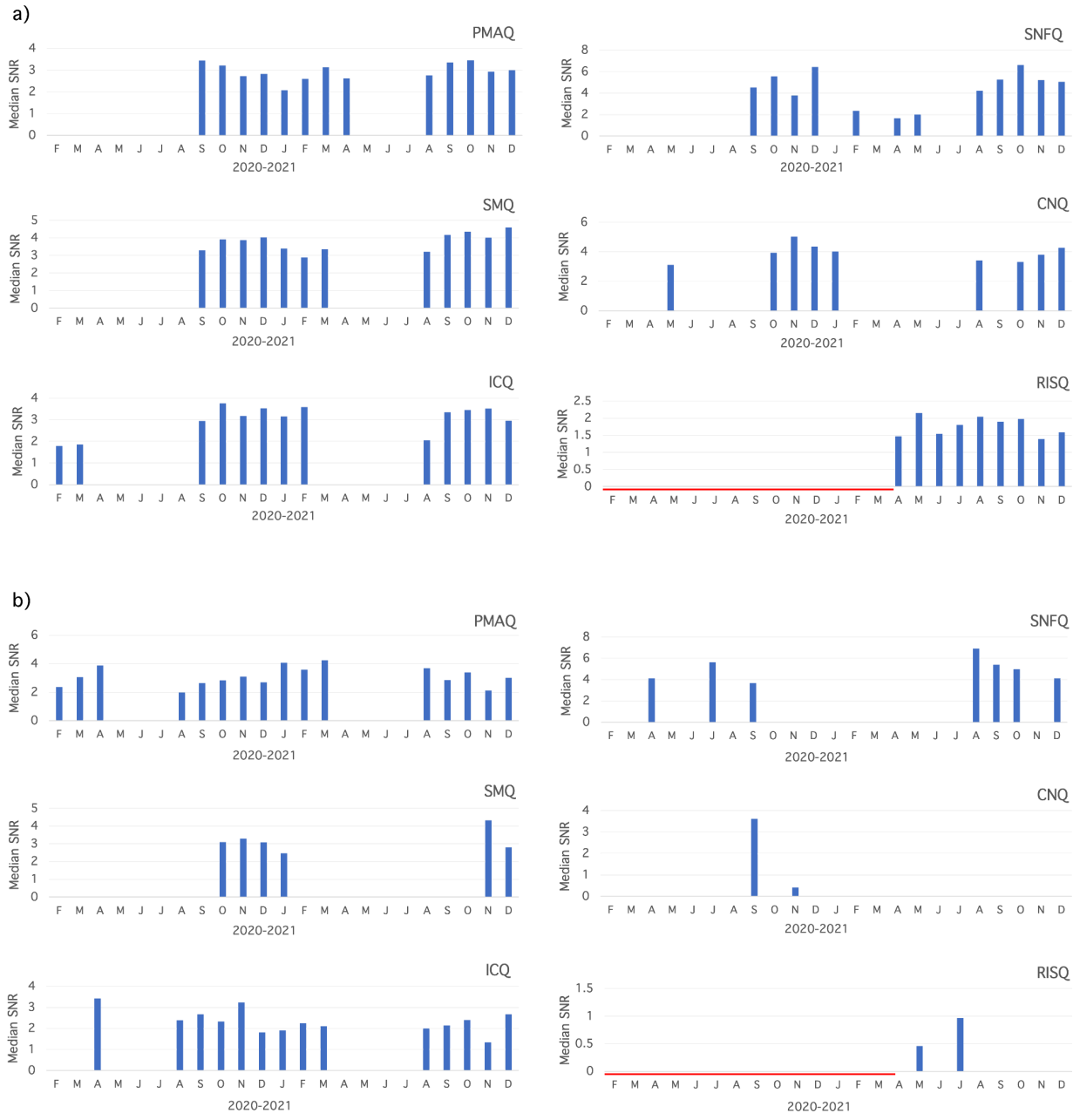
We don't consider vanilla CNNs on spectrograms as we consider longer temporal dependency is important to predict whale call time, while CNN only focuses on local context within a short time window.

More general comments:

You make sure to interpret the effect of background noise on the PN22 method in your discussion sections, but it would be reassuring to explain this effect when presenting your spatial / station-wise results – otherwise readers may interpret differences in station statistics as directly measuring whale activity. Maybe this could be included in Fig. 4 somehow?

Answer: To support the effect of background noise on the PN22 method in our discussion sections, we have included a monthly time series of the median SNR values of the "active" whale calls plotted in Figure 4, for each station. We think this might be best to be included in Supplementary Materials (Figure S3) to not overwhelm the results section of the paper. Figure S3 shows there is no significant variation in the SNRs at the Gulf (left column) and Estuary (right column) stations during our study period.

Figure S3 Monthly median SNR values of "active day" a) fin whale calls and b) blue whale calls plotted in Figure 4. Stations located in the Gulf are positioned on the left side and those in the Estuary are on the right. Note RISQ is only active starting in April 2021.



Can the hydrophone results provide more evidence re. the spatial distribution of calls, as discussed in 4.1?

Answer: In this study, we only have hydrophone data over 2 months (Aug-Sep) at the one location between Forestville and Rimouski (marked as "MARS" in Figure 1). Without localizing the calls, these can potentially be produced from whales up to ~30-40 km from the station, quite a wide spatial extent. From the hydrophone results and low detections on the Rimouski side of St. Lawrence (from RISQ seismometer) we suggest that the spatial distribution is likely skewed towards the Forestville side. This is because Forestville is close to the Saguenay Fjord, a

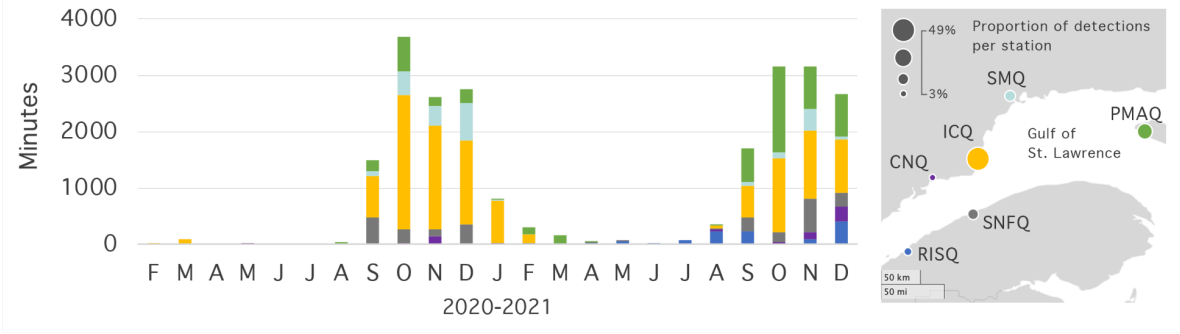
known biologically productive area due to steep bathymetry and winds enhancing upwelling of prey. However, the Forestville seismometer (FORQ) was discontinued in 2019, therefore we do not have land seismometer data to directly compare the whale call detections across the river. So based on the limited spatial and temporal coverage, the hydrophone catalog may be combined with the land seismometer catalog to infer limited information on spatial distribution of calls.

Figures:

- *4: could this information also be somehow represented by overlaying the statistics (maybe splitting by season) on the geometry of Fig 1 ? This may aid interpretation – maybe show activity within the detection radiuses of each of your instruments.*

Answer: Yes, we agree that including the Figure 1. map in the presentation of the detection results will help orient the readers spatially. To address this, we decided to insert a simplified representation of Figure 1., where the location of each seismometer is indicated with a circle, the circle size is proportional to the number of detections recorded and color matches the bar plot legend of the original Figure 4, at each station respectively. We do not think that splitting the detections at each station by season (summer/winter) would necessarily be helpful when displaying these results spatially, since detections are near zero at all stations in the summer. However, seasonal (summer/winter) and interannual (2015-2021) variations are shown for detections combined at all stations (Figure 5, S5).

a)



b)

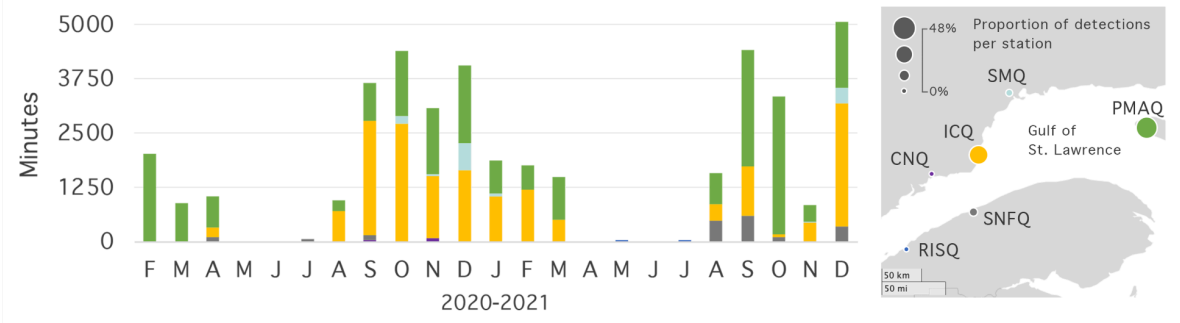
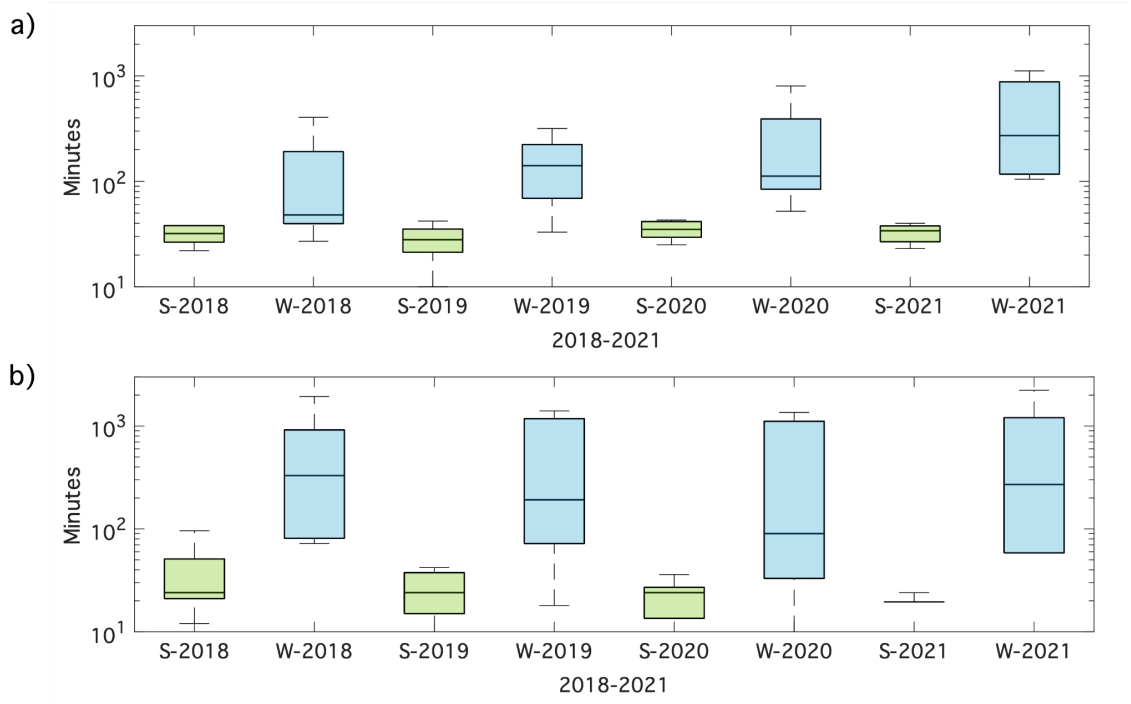


Figure 4 (Left) Monthly distribution of active minutes of whale calls, and (right) proportions of detections per station represented spatially, for a) fin whales and n) blue whales between February 2020 and January 2022.

- *5: Does the change in parameters between your work and the PN22 catalogue explain some of the differences here? Do you use the exact same stations to build both whisker plots, or could these be affected by the change in available stations?*

Answer: We do not change any of the detection parameters from PN22. Over the nearly 7 years of land seismometer detections in the LSLS between this study and PN22, there are changes in the number of available stations. As several north shore stations were decommissioned in 2019 (see Figure 1), in this study we used 5 stations between February 2020 and January 2022 and 1 station between April 2021 and January 2022. Note when a station was not operational, it was omitted from the median calculation. We create a supplementary plot using the 5 common stations in both studies (CNQ, SMQ, ICQ, PMAQ, SNFQ) between June 2018 and January 2022. The trend of high detection rates in the winter is still apparent.

Figure S5 Summary of winter (September-April) and summer (May-August) median detections across all 5 common stations (CNQ, SMQ, ICQ, PMAQ, SNFQ) with continuous data between June 2018 and January 2022 (from Plourde and Nedimović (2022) and this study) for a) fin whales and b) blue whales. The central line on each box refers to the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the maximum and minimum data points.



- 6: Is duration the most useful comparison metric here, given the differences in detection methods? I understand picking individual calls is not reasonable, so maybe just include a comment addressing this.

Answer: Ideally, we would be comparing the times of individual calls detected from the hydrophone and the seismometer. Since both of these studies were first conducted independently, our detection methodologies are slightly different and duration is likely the most effective comparison, without additional analysis of the hydrophone dataset. We include the following comment to address this point in L. 218-220: "Fin whale and blue whale songs can persist up to many hours, therefore we do not think the spectrogram windows of slightly different length analyzed by each study used to declare a detection impacts our comparison results significantly."

Grammar suggestions:

L195: consisted -> constituted ?

L200: detected(- space missing

L232: species, -> species

L278: Azore -> The Azores, Icelanic -> Icelandic

Answer: These grammar suggestions have been corrected in the revised manuscript, thanks!

Alex Saoulis, UCL

Recommendation: Revisions Required

Reviewer C:

This paper is a follow-up of the study by Plourde and Nedimović (2022) that extends the analysis to the Feb. 2020-Jan. 2022 period (initial work between Oct. 2015 - Feb. 2020). The conservation motivations for the study are well documented. However, I find the definition of the objectives/research question and associated development unclear. Are the goals of this manuscript:

1. *To analyze more years than Plourde and Nedimović, 2022, update the call library and investigate blue and fin whale spate-temporal behavior in the LSLS for the entire period (2015-2022)?*
2. *Or to create a deep-learning model for the detection of blue and fin whale stereotyped calls?*

These are both interesting topics; however, they should not be addressed in a unique paper. I'd recommend that the authors pick one.

Answer: Thanks for the comment. The focus of our current study is on topic 1). While we are also interested in exploring 2), we understand that the DL model presented is prototypical and still needs further work before its publication as a useful detection tool. We have removed the DL model part from the methods and results section and added a discussion point on the limitations of current detection method and the potential of DL model development. A similar comment was also raised by Reviewer A. Please see our response above.

My impression is that most of the paper is geared towards topic 1) and the remainder of my comments will be to help clarify that message and results, but in case the authors decide on 2), here are my overall suggestions:

- *It is important to improve the background and literature on machine and deep learning in bioacoustics as well as to mention other pipelines specific to fin & blue whale detection, e.g., but not limited to:*
    - *https://doi.org/10.7717/peerj.13152/*
    - *https://doi.org/10.1002/rse2.297*
    - *https://doi.org/10.1098/rsif.2021.0297*
- *Regarding the model design and training, getting more details on the architecture/training process (data preparation, optimization, hyperparameters, batch sizes, etc) and the evaluation metrics would be helpful.*
- *It is usually recommended to use independent train test data. For example, a useful split would be to use the first dataset from Plourde and Nedimović (2022) as train data and examples of the 2020-2022 dataset as test data.*
- *It is crucial to evaluate the performance of the model. Can you show precision-recall curves? It would also be interesting to compare those performances to the comparison with the "rhythm" detector from Plourde and Nedimović (2022).*
- *The final part of this paper could be the application to the new (or entire) dataset.*

- *It would also be important and helpful to discuss the risks of using the outputs of detection methods to train a DL model — what could be the pitfalls?*

Answer: Thanks for these valuable suggestions. We will take these into consideration in our next step in the development and performance evaluation of the ML algorithm.

Independently for the choice of topic 1 or 2, I'd suggest editing the title. The use of "tracking" is misleading as it generally refers to estimating a succession of positions of a moving source and trajectory, which is not the topic of this manuscript. A more descriptive title would be helpful (e.g., with species names, years, etc).

Answer: Yes, we agree that our initial title could be misleading. We have changed to a more descriptive title: “Spatiotemporal Variability of Fin Whale and Blue Whale Calls Detected by Land Seismometers along the Lower St. Lawrence Seaway”.

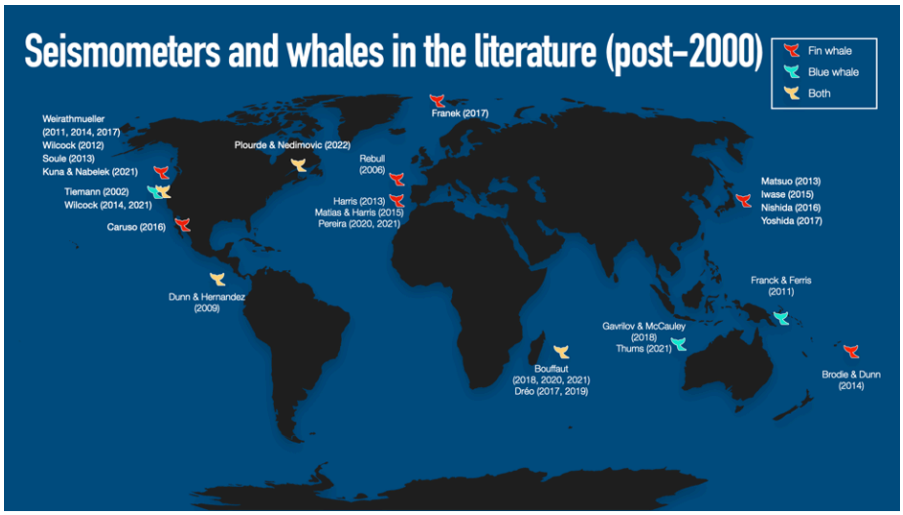
Detailed comments ——————

For clarity, I'd suggest adding subsections to Section 2, e.g., 2.1. Data collection, 2.2 Detection method, etc.

Answer: These subsections have now been added.

Seismometers

- *L. 58-59: there is quite an extensive literature related to (mostly Ocean Bottom) Seismometers & whales (I attached a map I worked on a year or so ago, next page), not only in the Pacific Ocean! I don't think it is necessary to provide a full review, but I'd rephrase to suggest that similar work has been conducted in almost all oceans*.



Answer: Thank you for attaching the map with associated papers! This is a really nice graphic. We would like to reference it, if possible? We correct the sentence and rephrase it in L. 69-71. "Several studies have successfully used ocean bottom seismometer (OBS) networks to detect whale calls and/or track whales globally (Dunn and Hernandez, 2009; Wilcock, 2012; Kuna and Nábelek, 2021; Franek et al., 2017; Mathias and Harris, 2015; Bouffaut, 2020; Dréo, 2019; Iwase, 2015; Brodie and Dunn, 2015; Tary et al., 2024)."

- *It would also be helpful to give a quick technical description of the seismometers that were used, e.g., model, bandwidth, if the response is flat in the frequency band of interest, etc.*
- *I don't think I was able to find that information in the text. Out of the three components, which one(s) is/are used in this study? Could the seismometer orientation impact the detection range (mentioned later)? And if it is not uniform, could this be indicated on the map?*

Answer: We provide more technical characteristics of the seismomters that were used in L. 99-105: "Stations ICQ, PMAQ, RISQ, SNFQ use a Nanometrics Trillium 120 Sec Posthole sensor that measured the three-component velocities. The other two instruments, at CNQ and SMQ, use a Geotech S-13 short-period sensor that only measured the vertical component. We used all three components at stations ICQ, PMAQ, RISQ, SNFQ. Whale calls usually have very similar vertical and horizontal amplitudes at onshore seismometers, so we expect that summing all three-components, to slightly improve SNR. At stations CNQ and SMQ, we only used the vertical component. All instruments have frequency responses that are flat within the 10-32 Hz bands of data processing"

- *A complete reference to IRIS is missing; see https://www.earthscope.org/how-to-cite/*

Answer: Thanks. We have added the reference to IRIS in the recommended format. In the Data and code availability section, we included in L. 392-395: "All seismic data were downloaded through the EarthScope Consortium Web Services (https://service.iris.edu/), including the following seismic network: (1) CN (Natural Resources Canada, 1975)."

The reference has also been added in L. 451: "Natural Resources Canada. Canadian National Seismograph Network. 1975. doi: 10.7914/SN/CN."

Whale acoustics

There is a need to establish more information on blue and fin whale acoustics in the introduction. For example, it is essential to establish from the beginning that this study focuses on song (or stereotyped vocalizations/calls) instead of social vocalizations, which are the expected target signals for a survey of feeding grounds. I understand that blue whale D-calls and fin whale 40 Hz calls are not captured in the land seismometers' bandwidth, but it should be acknowledged by the authors from the beginning. Besides, giving a brief overview of the frequency range of these signals could help justify the choice of the bandpass filters.

I believe that some references to fin whale and blue whale Inter-Call/Pulse/Note-Intervals and associated literature in the North Atlantic would be helpful, both in the introduction and discussion, e.g., https://doi.org/10.7554/eLife.83750, and should be used to interpret the obtained IPIs.

Answer: In the introduction, we provide more background information on the vocalization characteristics of fin whales and blue whales that frequent the LSLS, in L. 39-52. "Whales produce many acoustic signals often associated with either social or foraging functions (Romagosa et al., 2021). Northwest Atlantic fin whales produce songs consisting of a series of individual call units ranging in frequency from 18-21 Hz, lasting 1 second, and repeated every 10-15 seconds (Roy et al., 2018). The interval between consecutive call units is referred to as the internote interval (INI). Similarly, Northwest Atlantic blue whales also produce songs consisting of a series of individual tonal A-call units, at slightly lower frequencies ranging from 16-18 Hz, lasting approximately 8 seconds, with an INI of 68-78 seconds (Mellinger and Clark, 2003). In some cases, a secondary B-call is produced following the A-call (Simard et al., 2016). These songs with stereotyped repetition are believed to act as social/mating displays that are only produced by males (Romagosa et al., 2024; Širovic and Oleson, 2022). In addition, fin whales and blue whales produce intermittent audible 40 Hz calls downsweeping from 75-40 Hz and D-calls from 90-25 Hz, respectively, on feeding grounds (Simard et al., 2016; Romagosa et al., 2021; Širovic and Oleson, 2022). The INI and frequency range of whale songs are distinctive characteristics that vary spatially and are used to differentiate stocks and populations (Romagosa et al., 2024). Moreover, acoustic recordings from hydrophones deployed in the water has allowed biologists to better understand geographic ranges of whale populations and their habitat usage (Watkins et al., 2000; Stafford et al., 2007)."

In the discussion section, we have added a section on the limitations of the PN22 methodology and potential for DL. Here we include a few sentences about the results of some studies that have observed changes in INIs, frequency ranges of whale songs and implications on our methodology. L. 338-347 (section 4.5): "Within a single population, INIs and frequency limits of calls can change over time (Rice et al., 2022; Romagosa et al., 2021). Previous PAM datasets from 1998-2001 have shown that fin whale INIs used to be about 7 seconds longer than the

current 12 seconds, in the central and eastern North Atlantic Ocean (Romagosa et al., 2021). This INI shift occurred over four years, a relatively short amount of time, with the 12 second INI becoming dominant as of 2004 (Romagosa et al., 2021). Additionally, the peak frequency of both fin whale and blue whale stereotyped songs have been decreasing in nearly all ocean basins, since first recorded in the 1960s (Rice et al., 2022; Weirathmueller et al., 2017). The potential variability of INI and peak frequency of whale songs is important to note when using the characteristic recurrence method since it relies on these features for detection. However, there is typically a transitional period associated with these changes and the parameters of our detection algorithm can be adjusted as needed."

The hydrophone data is available and contains D-calls (see Figure S1). While this would require additional analysis, could the author label D and 40 Hz calls for the available period (2 months shown) to give a sense of how much whale activity is missed because of the seismometer's bandwidth? That could inform the discussion in 4.2.

Answer: We agree that this would be interesting to investigate and decided to perform additional visual analysis of spectrograms from the hydrophone dataset to label D-calls over August and September 2021, to determine the amount of higher frequency calls missed by the seismometer due to its lower sampling rate. In section 3.3 of the revised manuscript, we add in L. 220-224: "The hydrophone dataset contains higher frequency whale vocalizations that were not recorded by the seismometer due to its lower sampling rate. Over the two month recording period, there were no instances of blue whale D-calls vocalized without A-calls. D-calls were identified in only 25% of 5-minute time windows with the stereotyped songs. Therefore, during this time period, the nearby land seismometer would not have missed any whale activities due to its bandwidth limit."

These findings support part of section 4.3 of the revised manuscript in L. 277-281: "The higher frequency vocalizations produced by fin whales and blue whales are dominant on feeding grounds during the summer, with low vocalization rates outside these months (Romagosa et al., 2021; Širović and Oleson, 2022). From the two-month (August-September 2021) hydrophone catalogue presented in this study, we observe that blue whale D-calls are still present, although they are rare and of short duration compared to stereotyped songs (Fig. S6)."

Detection method

- *For clarity, I suggest separating the method's theory from its empirical application, with values specific to each call type.*

Answer: This paper does not present any new theory about the method. We thought it may be a more direct approach to demonstrate to the readers how the PN22 (Plourde and Nedimović,

2022) methodology is applied. Table 1 shows the values specific to the detection of each call type.

- *This method contains several user-defined values. It is important to describe the reasoning behind these decisions as it would facilitate applying the algorithm to other signal types (e.g., why 120 s /12 min window as input? why W(t)>3? See other suggestions in "whale acoustics")*

Answer: We have added our reasoning behind the choices of the 120s/720s time windows as input in L. 123-125: "Within 120 s and 720 s time windows multiple individual whale calls (usually 7-10) are present if a fin or blue whale is vocalizing, respectively. This provides sufficient seismic data for the detection method to recognize the energy peaks within the frequencies of interest over the recurrence intervals". In L. 135-138 we explain that "the threshold W(t) was chosen to maximize the ratio between the standard deviation and mean detections per day following Plourde and Nedimović (2022), hence to retain the maximum number of detections while minimizing the noise contamination (false positives)".

We further clarify in L. 152-155: "These user-defined values are not necessarily optimal and need to be fine tuned for applications in the detection of other types of whale calls and/or in other regions. We retain the values chosen by Plourde and Nedimović (2022) as they have been demonstrated to work optimally for the fin and blue whale call detections in the Lower St. Lawrence Seaway."

- *References to similar detection methods should be given.*

Answer: In L. 130-132 we refer to two studies that used related detection methods: "The whale call index is an energy detector similar to those described in Sirovic et al. (2015) in offshore Southern California and Pilkington et al. (2018) in the Canadian Pacific waters."

These citations have been added in L. 501-502 and L. 452-454 of the References section:

"Sirovic, A., Rice, A., Chou, E., Hildebrand, J.A., Wiggins, S.M. & Roch, M.A. (2015) Seven years of blue and fin whale call abundance in the southern California bight. Endangered Species Research, 28, 61–76. https://doi.org/10.3354/esr00676

Pilkington, J.F., Stredulinsky, E.H., Abernethy, R.M. & Ford, J.K. (2018) Patterns of fin whale (Balaenoptera physalus) seasonality and relative distribution in Canadian Pacific waters inferred from passive acoustic monitoring. DFO Canadian Science Advisory Secretariat Research Document. p. 032."

- *I'd also like to note that the continuous notation doesn't match the method description's "algorithmic" style, but I'll leave this decision up to the authors.*

Answer: We chose to leave the existing notation as is, the same way it was first presented in PN22.

- *Reporting the methods' performances, e.g., precision/recall curves, before setting a detection threshold is standard and expected. Can the authors elaborate on their performance evaluation method and reasoning for setting the Power ratio W(t) and SNR thresholds?*

Answer: The method's performance was evaluated when first presented in PN22. Since we follow the same parameter choices as in PN22 in this paper, we expect the method performance to be highly similar, if not identical, to that in PN22. However, we agree with the reviewer that it is necessary in this paper to briefly re-iterate the false positive rate that PN22 estimated when introducing the method.

We have added in L. 155-163. "Plourde and Nedimović (2022) created an additional detection algorithm, using a 20 second recurrence interval, targeting the 18-21 Hz band and a W(t) = 3 threshold. The primary spectrograms are computed for 1.5 second windows and the secondary spectrogram for 180 second windows. To be considered active, they require four detections (between the fin/blue whale thresholds used in the main algorithms) on a given day. The purpose of this "20 second period test" is to estimate the number of false fin whale and blue whale detections in the catalogue. They estimated a false positive rate of approximately 8.5% for fin whales and 4.8% for blue whales, by comparing the proportion of incorrectly designated active day detections from the 20 second period test and the total amount of active day detections. Since we follow the same parameter choices, we expect the method performance to be highly similar, if not identical."

- *This is my understanding: a positive detection is considered across a 120-second window for fin and 12 minutes for blue whales. However, the SNR is measured at the individual call level. What are the motivations for analyzing the detector's outputs at such a coarse level, while it seems like individual call detections are available?*

Answer: The motivation for using a coarse level for detection is now briefly mentioned at the beginning of the methods section in L. 114-120: "Northwest Atlantic fin whale 20 Hz calls and blue whale A calls are known to have relatively consistent intervals (INIs) between individual call units, with songs lasting up to hours (Roy et al., 2018; Simard et al., 2016). Waveforms at land seismometers can pick up on a lot of external noise, and whale calls received at these stations often have a relatively low amplitude. Due to this, if we were to evaluate detections on the

temporal scale of an individual call, especially in the case of fin whale calls which last only ~1 second, we think that surrounding noise would trigger detections and increase the likelihood of false positives. As such, we choose to rely on the recurrence intervals of whale calls, rather than individual call detections."

- *Why is a different detection method applied to the hydrophone with a distinct temporal resolution (5 min)?*

Answer: The MARS hydrophone audio data was recorded continuously, by 5-minute files. Manual annotation of presence / absence per file was initially performed independently from the present study, using the original lengths (5 minutes) of the data segments. However, as continuous data were analyzed in both the seismometer and hydrophone data processing, the different choices of data segments do not affect the comparison results .

- *Remove all sections related to the DL method and results.*

Answer: We have removed the DL sections in the methods and results. We insert a paragraph in the discussion about the limitations of our current whale detection methodology (as used in PN22), and why DL might be useful for whale call detection at land seismometers and future ocean bottom seismometers in the LSLS.

Figure 2: *What is the amplitude scale in the waveforms and what is the dB reference on the spectrograms? Please also add the spectrogram parameters to the caption.*

Figure 3: *Same comments as Figure 2*

Answer: We used raw waveforms with units in count in this study (as in PN22), which are presented in Figure 2 and 3. Alternatively, our detection algorithm can also be applied to waveforms after instrument response is removed and the detection results are not affected. We recently noticed that the y-axis on the spectrograms are on a linear scale, instead of a log scale. These have been corrected to log scale and now display the right frequency ranges for the fin whale pulses (18-21 Hz) and blue whale A calls (16-18 Hz). The spectrogram parameters (number of points and overlap) have been added to both figure captions.

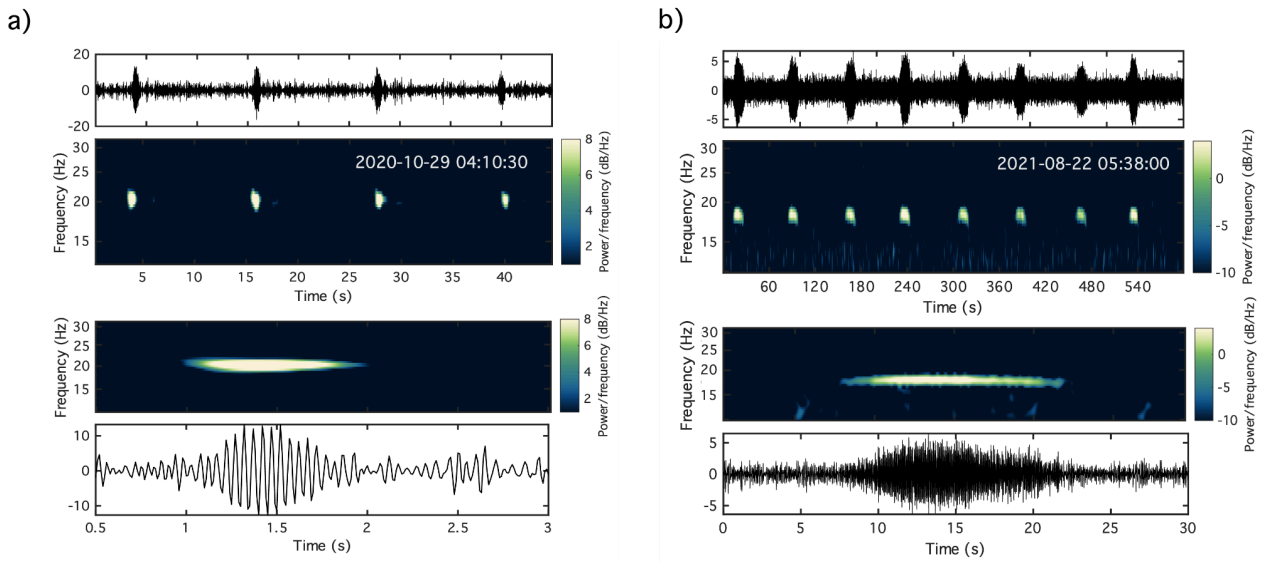


Figure 2 Characteristic waveform and spectrogram of a) fin whale calls at station ICQ (spectrogram parameters: STFT at 48 points with a 85% window overlap, 12-32 Hz filtered) and b) blue whale type A calls recorded at station SNFQ (spectrogram parameters: STFT at 48 points with a 70% window overlap, 10-32 Hz filtered). Bottom panels show the zoom-in of the first call within each series (spectrogram parameters: STFT at 48 points with a 95% window overlap). Note the y-axis unit for the waveforms is in count. Instrument response was not removed.

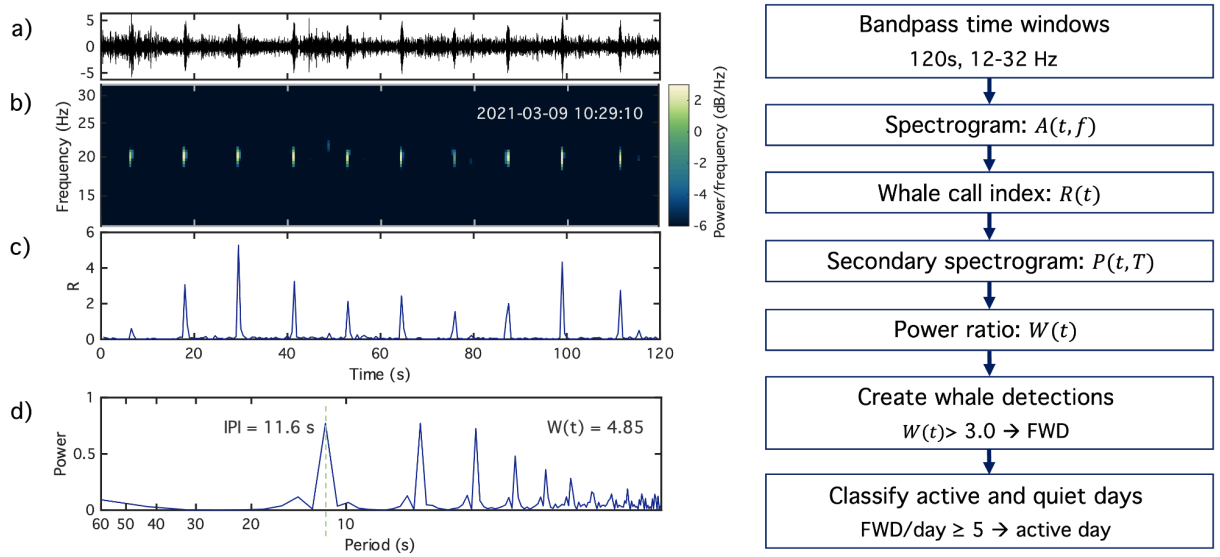
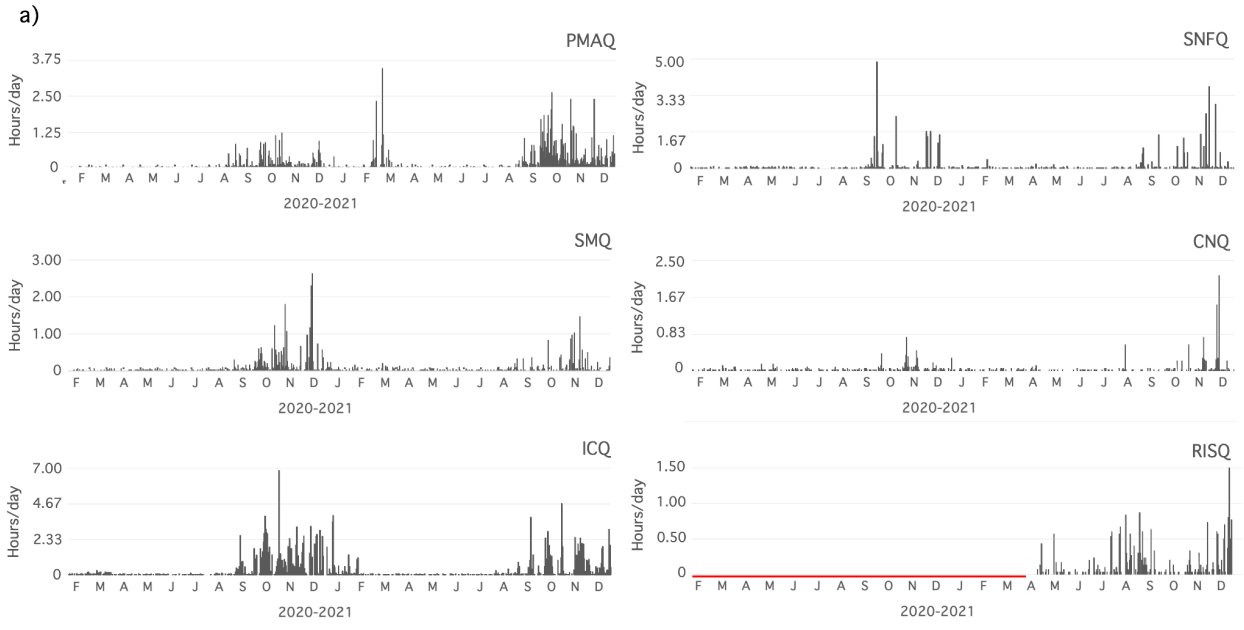


Figure 3 Example of a fin whale call detection procedure, as developed by Plourde and Nedimović (2022), using station PMAQ. a) bandpassed waveform segment, b) associated spectrogram (spectrogram parameters: STFT at 48 points with a 50% window overlap, 12-32 Hz filtered), c) whale call index (equation 1), d) periodogram of c) and power ratio (equation 2).

Figure 4: *The minutes per month are quite unusual and pretty coarse. What was the motivation? Could the authors show an additional representation with, if available, the number of calls per day, or the number of hours with detected calls per day, or even daily presence/absence?*

Answer: The demonstration of monthly data was done to characterize the broad interannual presence/absence of fin whale 20 Hz and blue whale A calls in the LSLS. This representation allows the reader to quickly and clearly observe key patterns (most detections in Northwest Gulf from fall to early-mid spring, and nearly no detections at all stations in summer) that we use as the basis in our discussion sections. We do have the more precise daily data and understand that other researchers concerned with higher temporal resolution might be interested in this. As such, we have included additional time series with the number of hours with detected calls per day, for each whale and station in the Supplementary Materials (Figure S4). The detailed dataset is also included in the Zenodo repository https://doi.org/10.5281/zenodo.10028774.

Figure S4 Hours per (active) day of a) fin whale calls and b) blue whale calls. Stations located in the Gulf are positioned on the left side and those in the Estuary are on the right. Note RISQ is only active starting in April 2021.



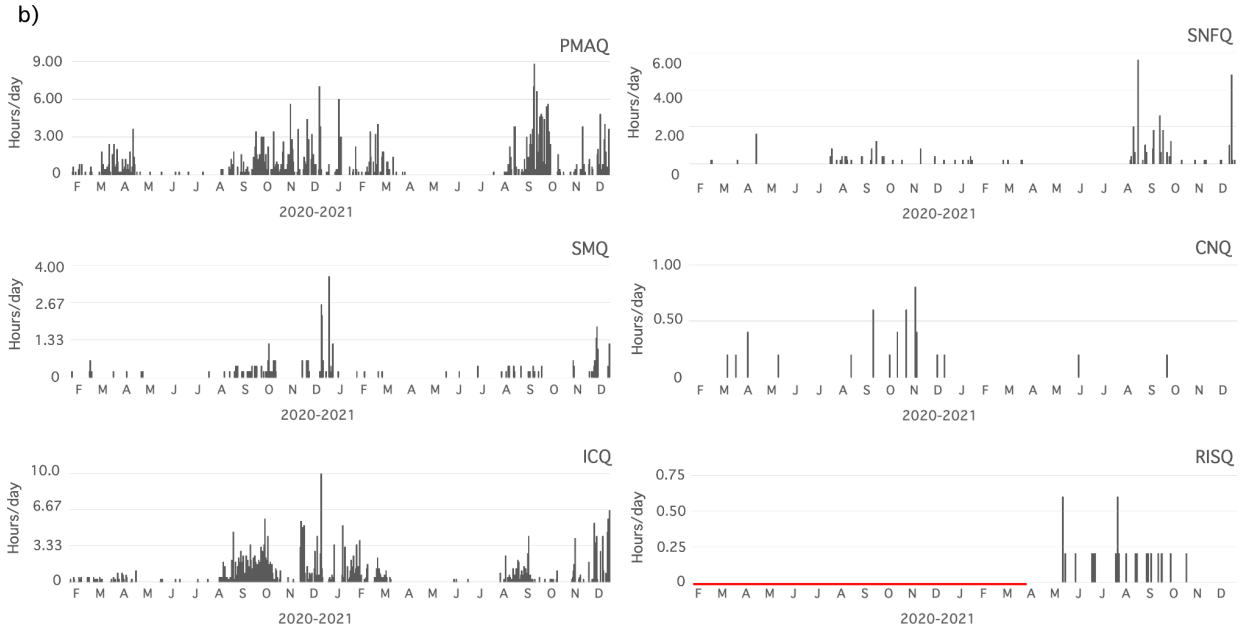


Figure 5: *The variations along the y-axis of this graph are challenging to read. Could they be converted to a log scale?*

Answer: Yes, we agree changing the y-axis to a log scale makes the figure more clear. Below is the revised Figure 5:

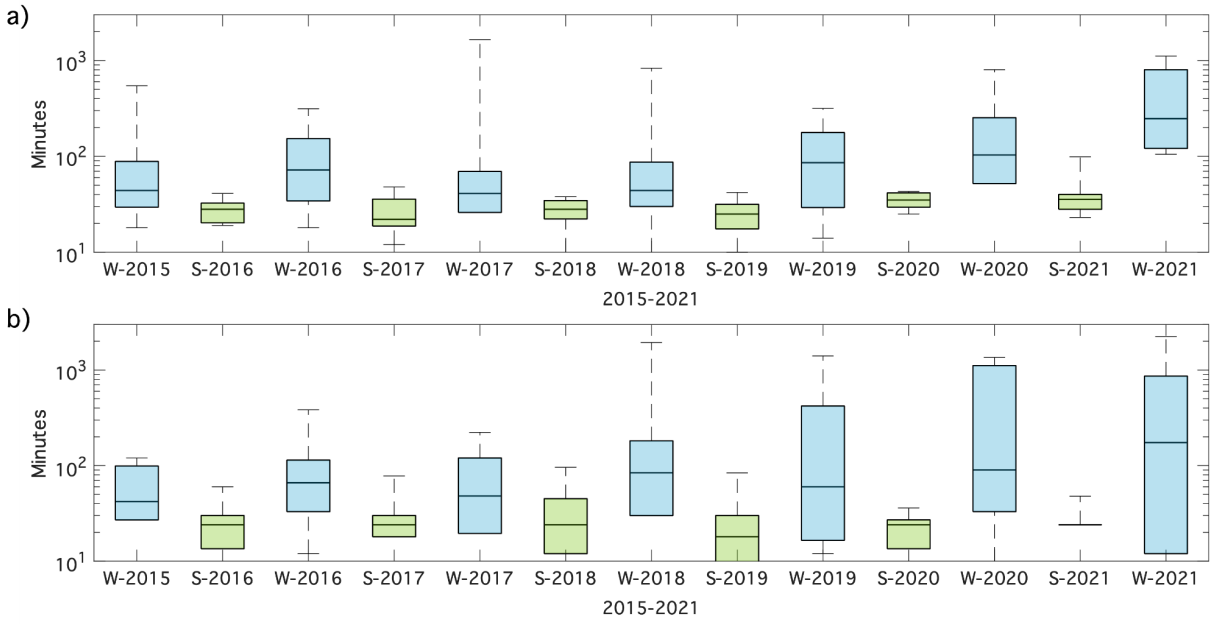


Figure 5 Summary of winter (September-April) and summer (May-August) median detections across all available stations between October 2015 and January 2022 (from Plourde and Nedimović (2022) and this study) for a) fin whales and b) blue whales. The central line on each box refers to the median, and the

bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the maximum and minimum data points.

Figure 6: *I find the combination of log scale y-axis and stacked bar plot hard to read. Could the instruments' results be juxtaposed?*

Answer: The initial Figure 6 displayed the seismometer and hydrophone results in a stacked format. We adjusted this plot, displaying the values side by side, with a small gap to distinguish each day. Changing the y-axis to log-scale should also make the side-by-side comparison clear.

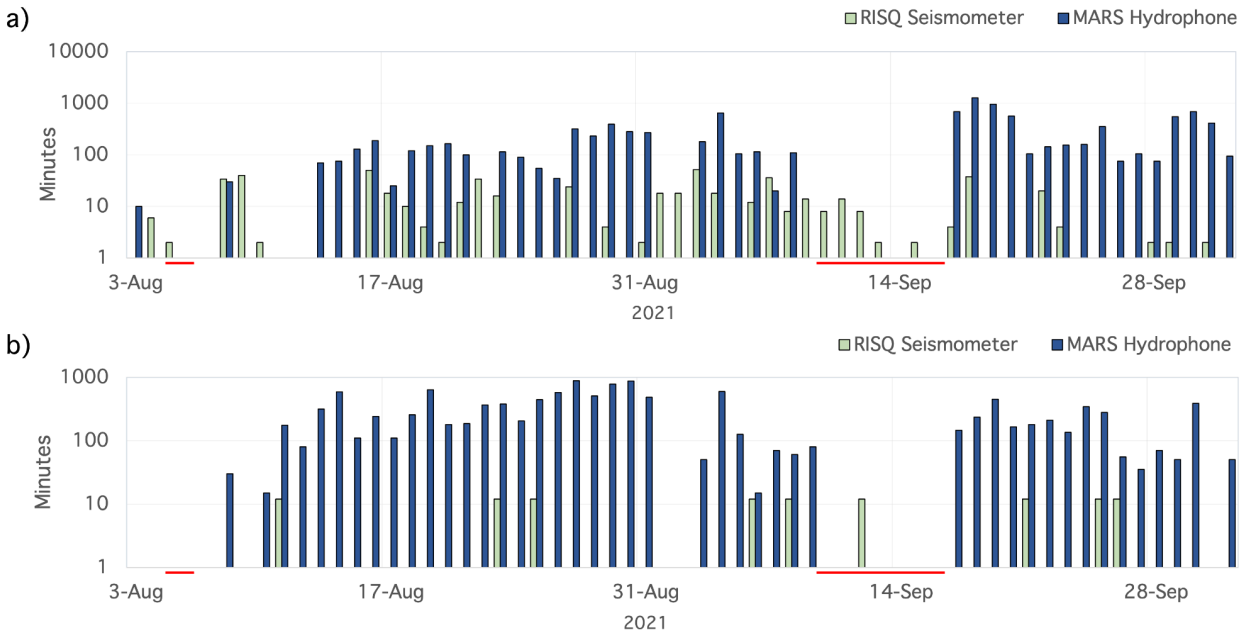
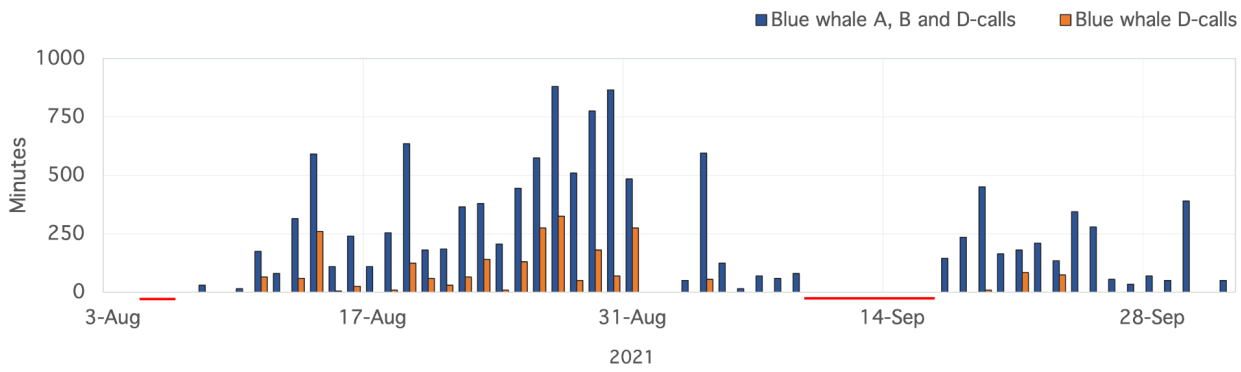


Figure 6 Distribution of minutes recorded at RISQ seismometer and MARS hydrophone during August and September 2021, for a) fin whales and b) blue whales. Red lines indicate inactive periods of the hydrophone.

Since segments with D-calls are all within already identified 5-minute windows that are shown in Figure 6b), we kept it as is. We created a supplementary plot (Fig. S6) to show the portion of blue whale D-calls with respect to blue whale stereotype A songs.

Figure S6 Distribution of blue whale presence and type of vocalization recorded at MARS hydrophone during August and September 2021. Red lines indicate inactive periods of the hydrophone.



Results

- *A few text portions should be moved to the discussion (L. 167-176; 185-191 + figure 5). If the authors want to integrate the data from the previous study into the analysis, this should be mentioned in the method section.*

Answer: We would like to directly compare our results to that of previous years, as shown in the original manuscript. In the revised manuscript, we now mention in the methods section, in L. 164-166: "We compare annual spatiotemporal trends in fin whale and blue whale detections from our catalogue (February 2020 to January 2022) with that of previous years (October 2015 to February 2020) reported by Plourde and Nedimović (2022). "

- *Similarly, the discussion of the detection range (L. 197-209) should be moved to the correct section (4.3).*

Answer: We agree and merge those lines from section 3.3 of the original manuscript to the Discussion section 4.2 in the revised submission.

- *I recommend adding that L. 197-198 is true for a specific recorder setup and a call type.*

Answer: Great point. We add that in L. 300-301: "For a given recorder setup and type of whale call, the detection area of a hydrophone is mainly dependent on the local bathymetry, water mass characteristics and ship traffic in the location of the station (Simard et al., 2016)."

- *L. 202-204: please rephrase; there may be many eigenpaths between a source and a receiver, including several reflections in underwater propagation.*

Answer: Yes, we rephrase in L. 306-307 to clarify that we are referring to the first wave recorded at a station. "When a whale produces a call, the first acoustic wave that reaches the receiver is usually a direct wave, travelling in the water column from the whale to the hydrophone."

Discussion

- *A first discussion section is missing on the performances of the detection method.*

Answer: We add a first discussion section on the performance of the detection method in L. 227-234. "The characteristic recurrence power ratio methodology used in this study appears to produce robust results, following similar spatiotemporal whale detection patterns noted by previous studies in the LSLS (Roy et al., 2018; Simard et al., 2016). The minimum number of detections within a day to be considered "active" serves as a first way to eliminate likely false detections. Out of the total number of detections labelled using this method, 86.4% and 84.1% of blue whale and fin whale detections respectively were classified as "active". Plourde and Nedimović (2022) estimated a false positive rate of 8.5% for fin whales and 4.8% for blue whales, respectively for active day detections. Since we do not change any of the detection parameters, we assume this estimate is likely very similar for this 2020-2021 catalogue. "

- *The discussion on the detection range should come in second, as it impacts all of the results of the spatio-temporal analysis.*
    - *What is the impact of potential differences between stations on the detection range?*

Answer: The potential different detection ranges of stations could have some impact on our results. The smaller the detection range of a station, the less whale calls it has the potential to detect. At the moment, we cannot quantitatively say by how much the detection range varies across the stations used in this study.

    - *The detection range will change throughout the year because of changes in water propagation. How could that impact seismometer detection ranges?*

Answer: We are not sure about the exact meaning of "water propagation" in this question - we interpret water propagation to refer to the flow/velocity of the estuary. Given that the detection ranges appear at least in some cases to be quite short (~a few km), we imagine contrasts in surficial geology and/or bathymetry are much more relevant than the water flow.

- *L. 267-282 should be the first part of the biological discussion, as it strongly impacts the interpretation of the results.*

Answer: We agree. L. 267-282 of the original manuscript has been moved to the beginning of the biological discussion (now section 4.3, L. 315-330).

Data/Code repositories

It is great that the data are available. However, considering the FAIR principles, I suggest that the data, code, and models be published with an associated DOI (this is possible in GitHub: https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content) and cited accordingly in the paper, with thorough metadata.

As a suggestion, here are examples of data in a format that is easily reusable by other bioacousticians & data scientists:

- *https://zenodo.org/doi/10.5281/zenodo.7078498*
- *https://zenodo.org/doi/10.5281/zenodo.7018483*

Answer: In the data and code availability section, the supplementary material that was originally linked to a google drive folder has now been uploaded onto Zenodo (https://doi.org/10.5281/zenodo.10028774), along with the MATLAB whale detection code we used and a .mat file with all labelled whale call center times (new Table S1).

Recommendation: Revisions Required