Response to Reviewers

Alex A. Saoulis* Afonso Loureiro Maria Tsekhmistrenko
Ana M.G. Ferreira

Introduction

We have submitted a revised manuscript file alongside a marked-up manuscript with the changes highlighted with `latexdiff`. Please note, some minor bugs with `latexdiff` has meant that the marked-up file is missing Fig. S7 (and references to Fig. S7), and the changes to the abstract are not highlighted. The revised manuscript is complete, however.

We thank both reviewers for their constructive and comprehensive feedback. While our main conclusions remain the same as before, we have made major revisions to the manuscript, which we believe have made our goals and motivations clearer, as well as presenting more evidence for the strength of our approach. We summarise the key changes here:

- We have substantially revised the introduction in order to make our contribution more clear, and explicitly contrast it with prior work. We have also made a number of changes here to improve the clarity and flow.
- We now provide a more comprehensive presentation and analysis of the synthetic feature generation procedure in the Supplemental Materials.
- We have run a number of extra architecture exploration experiments in order to quantitatively report the performance of each architecture for both the instrument resonance and whale vocalisation classes. This provides evidence for the claims regarding architecture presented in the Results section.
- We manually inspected model predictions of whale vocalisations covering 200 hours of OBS data across the UPFLOW array. We used this to produce quantitative estimates of the whale call false positive rate, as well as explore cases of systematic false positives.
- We performed bootstrapping analysis of all of our validation sets to explore the stability of our evaluation metrics in light of the small dataset sizes. We found that our dataset sizes are comfortably large enough to draw the conclusions we do regarding relative model performance.
- We expanded the analysis of the instrument resonances. In order to aid interpretation of Fig. 8, we present the raw, unfiltered time series of the two variables (resonances and currents) in the Supplementary Materials. This shows periods of strong correlation between the two variables. In addition, we present a more in-depth analysis of the resonances across the entire frequency range, and find very narrow frequency band that appears to be activated by tidal currents.

*Corresponding author: `a.saoulis@ucl.ac.uk`

- We have added a limitations section to the discussion section to temper our conclusions. We have elsewhere restated several times that the example applications are proof-of-concept.

Finally, we would like to highlight that Seismica has a strict 10 figure and 10000 word limit in the main manuscript. We have therefore confined the additional 5 figures and 4 tables to the supplementary materials, and are restricted in presenting significantly more analysis in the main text. All our review replies are given below.

Reviewer 1

The manuscript, "Semantic segmentation for feature detection in ocean bottom seismometer data", presents an application of semantic segmentation to time–frequency representations of seismic and bioacoustics data from ocean-bottom seismometers (OBS). The study addresses an important challenge in automating the detection of rare and diverse signals, such as instrument resonances and blue whale calls, and explores the use of synthetic pre-training to enhance model performance. The results are promising and offer a rich dataset for further exploration. However, several aspects—ranging from dataset size and representativeness, clarity of methodology, model evaluation, and framing of novelty—would benefit from additional clarification or more cautious interpretation. The comments below are intended to provide constructive guidance to strengthen the rigor, reproducibility, and contextualization of the work.

Major Comments

1. The manuscript claims to be the first to apply semantic segmentation to time–frequency seismic data, but the framing of novelty could be more cautiously framed. There are closely related works in both bioacoustics and seismology that have used spectrogram-based segmentation or object detection. The paper should provide a more careful comparison to these studies and articulate what is genuinely new here (pixel-level annotations? synthetic pre-training strategy? or cross-domain generalization across instrument types?).

   We thank the reviewer for this comment. We believe this claim to be accurate; no prior work has performed semantic segmentation on time-frequency representations of seismic data. However, we agree that our contribution needs to be much more explicitly articulated. The suggestion to directly compare with prior work is very helpful and helps better frame the manuscript. We have rewritten this paragraph to provide a clearer survey of prior work, and explain exactly the novelty and advantages of our approach (see lines 69-79 of the revised manuscript). We have also made our claim slightly more cautious: we write "we introduce", as opposed to "this is the first" (see lines 11,76,604 of the revised manuscript).
   Regarding bioacoustics, we agree that there is some related work, notably one prior example of ML-based pixel-level segmentation of bioacoustic signatures. We now cite this work, which nevertheless explores quite different signals and datasets to our study, and that is why it was not included in the original version of our manuscript. Given the emphasis of our work on **seismic** instruments and signals, we feel that our claims and emphasis are fair and proportionate (see lines 69-82 of the revised manuscript).

2. The core dataset (500 annotated spectrograms from a single station) is relatively small, and it is unclear whether it is representative enough of the variability across the entire UPFLOW array or across different OBS deployments. The validation strategy is limited (50 spectrograms each from UP34 and RR40) and does not convincingly demonstrate robustness. The

manuscript should explicitly acknowledge the limits of statistical power here, and ideally, expand the evaluation to include more independent stations.

We appreciate this perspective. We agree that the core dataset is relatively small, though it is worth noting that segmentation annotations are much more information dense than image level labels. As you suggest below in the review, we have decided to perform bootstrapping / resampling of all three validation sets to explore the stability of our model performance. We also perform a two-fold test by reshuffling the UP05 dataset and repeating training. We report the results of these experiments in the supplementary materials (see Table ST3, and lines 390-392). We also re-computed results for the whale call segmentation models over three separate train / validation re-shuffles, and updated the Table 2 in the main text.
We find that for the instrument resonances, the model performance stabilises well before our chosen 50 spectrogram limit. At 50 spectrograms, the standard error in IoU is around 10^{-3} and so does not impact any of our conclusions regarding architectures. This indicates that the dataset size is sufficient to provide a reliable, low variance estimator of model performance at these stations. In addition, we find very little difference in model performance having retrained and re-evaluated the models on a new train / validation split. We believe this provides satisfactory evidence that the validation dataset sizes are more than enough to demonstrate robustness for a given station.
In addition, the model performance is similar across all three randomly selected stations, which vary in instrument type, spatial location and temporally. As explained in the main manuscript, we selected these stations for their diversity. In addition, the results of the proof-of-concept applications (particularly Figs. 7 & 9) show that not only is model performance relatively stable across different instruments, but that the models produce reliable results for downstream applications. We have also significantly extended the analysis and discussion of false positives, using data from across the entire deployment, to add weight to this conclusion, discussed below (see lines 535-546, Section S5, and Table ST5 of the revised manuscript). We believe that these comprehensive tests are sufficient to demonstrate the robustness of the model.

3. The fact that resonances were annotated far more frequently than whale calls suggests that the model evaluation could be influenced by annotation variability. While the synthetic pre-training is intended to address this, the performance claims (e.g., 90–100% improvement) must be interpreted with caution. The discussion should temper the strength of these conclusions.

We agree that annotation variability could impact model evaluation somewhat. Since the dataset size for the whale calls is very small, as you rightly point out, we have decided to perform further experiments. We run 9 independent training runs in total, with 3 repeats for 3 reshuffled train / validation splits, and report the results in the whale call segmentation results table (Table 2).
We find that the standard error of performance metrics only falls slightly, indicating quite a high natural variance in both training and evaluation. This is expected, as you point out, given our very small dataset size. However, we also found that with more repeats the mean transfer learning performance was actually significantly better than the two alternatives in relative terms than we first reported. Specifically, supervised learning gave IoU $= 0.100 \pm 0.025$ while transfer learning gave IoU $= 0.299 \pm 0.027$.
In light of these new results, which show between a 100% (conservative) to 200% (optimistic) relative improvement in performance, we feel it is defensible to quote this lower bound on performance improvement. We have added some discussion about the high variance in the model performance (see lines 429-438 of the revised manuscript).

4. The synthetic feature generation is central to the results, but it is described only heuristically. Key details (mathematical form of the whale call generator, distribution of parameters, validation of realism) are deferred to supplementary code. Without a clear description in the main text, it is difficult to assess how representative the synthetic features are. Moreover, there is no quantitative validation of how closely synthetic features match real examples. A more

systematic evaluation (e.g., comparing feature distributions of synthetic vs. real annotations) is needed.

We thank the reviewer for this perspective. We agree that we need to provide more details and examples of the synthetic features. We have added another section to the Supplementary Materials (section S2) describing the simple, heuristic approach generating both the synthetic resonances and blue whale calls. This is introduced in the main text on lines 278-283.
We have also added two more figures to the supplementary materials (Figs. S2 and S3). The first more clearly demonstrates the procedure for generating these synthetic features: we define some template for generating individual signals, sample from this template stochastically (also varying the spectral position and amplitude of the signals), and then distribute the samples across a background spectrogram. All of these steps were visually tuned, as we stress in the text.
The second figure provides a detailed statistical breakdown of the synthetic features compared to the annotations for each feature class. This covers the feature occurrence rates and amplitudes against frequency, as well as distributions of the height, width, and area of features in each approach. This figure shows that the general distribution of the features are relatively well modelled by our synthetic feature generators. However, there are also some small discrepancies (e.g. underestimation of the feature amplitudes, and incorrect or offset feature frequencies). This is a result of the fact that we performed no quantitative tuning of the feature distributions. Yet, this is not problematic for our study because we show pre-training on our synthetic features still yielded substantial performance gains. Our results demonstrate that simple, heuristic generation of the features is sufficient to provide pre-training benefits.

5. The authors acknowledge label ambiguity and suggest that IoU and recall are more informative than precision, but this raises the issue that model evaluation may be biased toward forgiving errors. If annotations are inconsistent, then improvements in IoU may not reflect genuine gains in detection quality. The manuscript should provide inter-annotator agreement statistics to support the reliability of the labels.

We thank the reviewer for this suggestion. We agree that presenting inter-annotator agreement is a good idea. We have added a table (Table ST4) and short discussion to the supplementary materials (section S4 in the supplementary materials) documenting the inter-annotator agreement statistics for the portions of the training dataset that had annotator overlap.
While annotation inconsistency introduces variability, we do not believe it leads to a systematically biased IoU toward overstating performance. Instead, as you suggest, inter-annotator agreement provides a realistic ceiling. Improvements in IoU are meaningful and straightforwardly interpretable as improvements while they approach, but do not exceed, this upper bound. As discussed in lines 382-384 and 1125-1128, we now demonstrate that we are in this regime.

6. False positives and false negatives are underexplored. For example, the blue whale detection task is reported to have "low false positive rate," but there is no rigorous quantitative assessment, particularly in out-of-season periods when calls should not exist. This makes it difficult to fully assess the generalization claim.

Thank you for this suggestion. We have significantly expanded the analysis for the blue whale vocalisation detection application in order to present a more comprehensive evaluation of the model false positives. The results of this effort are now summarised in the main text (see lines 535-546 and detailed in the supplementary materials (section S5, table ST5, lines 1130-1162).
We manually verified the model call detections for around 200 hours of spectrogram data sampled uniformly across the UPFLOW array. We found a very low false positive rate outside the expected vocalisation season ($\sim$ 5 false positives per station per week), corroborated both by the total number of detections in this period as well as our manually inspected dataset.
Interestingly, we found a significantly higher false positive rate during the active call season. We interpret this as being partially caused by the repeating nature of whale calls, which the model learns but then incorrectly interprets other signals as whale calls during highly active call periods. We also find that earthquakes and broadband transient signals are the most common source of false positives, effectively leading to systematic false positives. However, the overall precision is still very high during the period (0.98), indicating that false positives make up only $1 - 3\%$ of the total detections in our catalogue.
As we discuss below, false negatives are more challenging to define objectively because their quantification depends on a floor Signal-to-Noise-Ratio (SNR). This would require a substantial amount of work well beyond the scope of this study; we added text explaining this to the revised manuscript (see lines 1130-1134).

7. The resonance–tide relationship (Section 4.4.1) is intriguing, but the analysis is very preliminary. The scatter plots are complex and not quantitatively linked to oceanographic drivers. Without direct current-meter validation, the interpretation remains speculative. The authors should either (i) scale back the claims, or (ii) include more robust validation with independent current data.

As stated in the discussion, we agree that direct prediction of current from resonances is not possible without calibrating with current-meter data. We have revised the results section to be explicit on this point, and scaled back claims in this section (see lines 500-503).
Nonetheless, Fig. 7 quantitatively demonstrates the link between the resonances and ocean drivers. We have now also included a year-long time series of current against resonant energy which shows clear periods of correspondence between the tidal signal and the resonances (Fig. S5). As mentioned in the text, Godin et al. [2024] used current meters to demonstrate that (complex) regimes are expected in the data. In addition, while the results are clearly complicated to interpret and do not allow for direct measurement of currents, we believe there is scientific value in presenting these results — e.g. for other groups aiming to achieve similar things.

8. The blue whale tracking (Section 4.4.3) is similarly overstated. The agreement with Dréo et al. (2019) is presented as validation, but the method uses relatively simple assumptions (constant velocity, coarse time resolution). The observed eastward bias could indicate systematic errors, yet the discussion downplays this. The authors should either expand the localization analysis with a more realistic acoustic propagation model or present this as a proof-of-concept rather than a robust application.

We thank the reviewer for this helpful feedback. We agree that our localisation analysis should be framed as a proof-of-concept rather than as validation of the model. We stated this at the start of the Results section but it is worth re-stating. In the revised manuscript, we have clarified this point at the start of Section 4.4.3 and have softened our conclusions (see lines 567-569, 597-602). In particular, we now describe the analysis as an exploratory, proof-of-concept demonstration. We believe these changes make the scope and intent of this section clearer.

9. The conclusion suggests that the method could be applied to volcanic tremor, earthquake foreshocks, or deep earthquakes. These are fundamentally different signal classes from instru-

ment resonances or whale calls, and the extrapolation could be framed more cautiously. A more cautious outlook would strengthen the credibility of the conclusions.

We have revised the conclusion to frame the potential application to other seismic phenomena more cautiously. In particular, we now emphasise that such extensions are speculative and suggest them only as possible directions for future work, rather than as direct applications of our current approach (see lines 621-625 of the revised manuscript).
We also made clearer that these suggestions stem from the fact that our approach is grounded in a time–frequency representation, which is particularly well suited to signals with distinctive spectral and temporal patterns (see lines 622-623, 630-637 of the revised manuscript). This contrasts with more common time-domain detection approaches used in seismology, whose performance often degrades under very minor changes in background noise frequency content, etc.

10. While the manuscript points to a GitHub repository, it is not clear whether the full annotation set (500 spectrograms, plus validation data) will be made available. Without access to the labelled data, it will be difficult for others to reproduce the results. The authors should clarify data availability and consider at least releasing a subset of annotations if the embargoed raw data limits sharing.

Thank you for this suggestion. While the UPFLOW data cannot be made available until the end of the data embargo period, we are happy to release the manual annotation dataset via Zenodo (see line 701 of the revised manuscript).

11. The manuscript cites a very large number of works, but some appear only tangentially relevant (e.g., inclusion of multiple ML references without a direct link to seismology). Consider tightening references to those most directly connected to the study.

We agree that we failed to connect some of the ML works cited to our motivations and approaches in the manuscript, particularly at the start of the Section 3.2. We have re-written this paragraph to shorten it and make clear the connection to the approach of transfer learning utilised here (see lines 188-196).
However, we would also like to emphasise that many of the design decisions made in this manuscript were directly inspired by ML works without a direct link to seismology. We feel that there is still a lot to learn and adopt from various "application" areas of ML (particularly e.g. audio processing); we added some text about this to lines 79-82, 631-637 of the revised manuscript.

12. The manuscript evaluates segmentation models primarily against themselves (with vs. without synthetic pre-training). There is no comparison to simpler or established methods for event detection, such as threshold-based spectrogram detection, or classical machine learning approaches (e.g., random forests, SVM on spectrogram features), or existing bioacoustics tools (e.g., Pamguard, spectrogram correlation methods). Without such baselines, it is unclear whether deep learning is truly necessary or whether simpler methods would perform equally well on the limited dataset.

We agree that comparison with existing techniques is desirable, and indeed we point in that direction in the discussion. However, we feel that this request is out of scope for the following reasons:

- Our approach performs pixel-level time-frequency feature segmentation. This is rare in the literature (non-existent in seismology, a handful of examples in bioacoustics) and enables qualitatively new automated analysis compared to almost all traditional techniques (including in e.g. Pamguard). For instance, extracting feature frequency ranges and time durations, and isolating them relative to the background (as done in this work), is not something that can be achieved by the vast majority of existing approaches.
- The flexibility of ML-based approaches is relatively unique compared to the few existing classical "pixel-level" detection approaches in bioacoustics (e.g. [Roch et al., 2011, Gillespie et al., 2013]), which are highly tailored to individual signal types and require significant pre-processing and domain knowledge to design.
- There is limited space in the manuscript. We have already introduced three separate machine learning approaches and benchmarked them against one another on two separate tasks, as well as elucidating network design over a wide range of architectures. We feel that this is largely sufficient for the purpose of this study and very useful to Seismica's readers.

Following your suggestions throughout the review, we have rewritten parts of the introduction (see lines 69-82) and discussion (lines 677-681) to more clearly describe the novelty and usefulness of our approach. Hopefully these improvements make the first point above clearer.

13. The study implicitly assumes that a model trained on UPFLOW can generalize to RHUM-RUM. However, these deployments differ in noise environment, bathymetry, and instrument setup. No domain adaptation techniques (e.g., fine-tuning, transfer learning, normalization strategies) are attempted or even discussed.

On the first point, we respectfully disagree. We did not assume that the model trained on UPFLOW could generalize to RHUM-RUM. In fact, the reason we curated two extra datasets for two different OBSs (UP34, different instrument type, same deployment; RR40, different instrument type, different deployment) was to probe to what extent these complications would degrade performance. As demonstrated in Table 1, we find that model performance is remarkably consistent between datasets, particularly given the factors you mention (changing noise environment, bathymetry, and instrument setup). Our results explicitly demonstrate that our approach generalizes relatively well. To this end, our results suggest that the approach taken in this work produces a model that can be applied in a "zero-shot" fashion to new, unseen OBS data. There are obviously caveats to this statement, and we do not expect it to hold for all OBS data. However, the point is that our results suggest that *we do not need* added domain adaptation strategies to achieve acceptable performance on new, varied OBS data. We have amended the text to make this more explicit (see lines 168-173, 396-399 of the revised manuscript).
On the second point, thank you for this suggestion. We agree that discussing domain adaptation techniques such as fine-tuning or latent representation alignment is worthwhile, especially as it further highlights the strength of our "zero-shot" adaptation performance. We have updated the discussion to mention this (lines 613-618).

14. The segmentation operates on individual spectrograms, but many features (e.g., whale calls, resonances) are inherently temporal sequences. Ignoring continuity across time reduces robustness: a call spanning multiple windows may be fragmented or misclassified. A stronger approach would incorporate temporal context (e.g., sequential models, sliding-window consistency checks). At a minimum, the authors should discuss this limitation and potential improvements.

Thank you for this comment. We agree that evaluating the model on overlapping windows, or performing sliding window consistency checks, could improve performance. We have now added this to the limitations to highlight it as future work (lines 645-648).

15. The manuscript treats instrument resonances as a well-defined feature, but their origin and acoustic signature remain poorly constrained. Without an independent physical model (e.g., laboratory tests or known instrument responses), it is not clear whether the model is detecting true resonances or just clusters of spectral energy. This weakens the geophysical interpretations. The authors could clarify: How were resonances distinguished from background noise in annotation? Could the "resonance detections" simply reflect frequency-dependent site effects? Without a physical anchor, downstream analyses (e.g., resonance–tide link) may not be meaningful.

Previous studies such as by Godin et al. [2024] clearly showed that instrument resonances can be used as a proxy for seafloor currents. In addition, the studies by Stähler et al. [2016] and Corela et al. [2023] also presented instrument resonances as well-defined signals. As explained above, we now added an additional figure to the supplementary materials (Fig. S5) clearly showing a link between resonances and tides.

16. Several results are reported as "X% improvement" without presenting absolute performance values. For example, a "90% improvement" in whale call recall may correspond to going from 5% to 9.5% — may still have limited practical utility. The paper must present absolute detection rates, counts of true/false positives, and error bars to give a realistic sense of model effectiveness.

All the tables report these quantities in absolute values and so absolute performance values were already presented in our original manuscript. We have now updated the text to restate the absolute IoU values alongside the percentage improvement (see lines 396, 432-437 of the revised manuscript). We would also like to emphasise that we explicitly demonstrate and discuss the very great qualitative improvement a 100% performance improvement leads to. We have therefore argued that there is significant practical utility in the doubling of IoU in the context of whale call *detection* (see lines 422-428, 414-516, Fig. 6, Fig. S7).

17. The paper does not discuss the computational cost of training or inference. This is important for practical deployment on long-term OBS datasets (which can run to terabytes of continuous audio). How long does it take to process one month of data on typical hardware? Could the models be deployed in near-real time or only retrospectively?

Thank you for this suggestion. We have now expanded the details regarding computational cost of inference, and present it all at the start of the applications section. We provide some very brief commentary about the fact it would be easy to deploy these models in near-real time on consumer-grade hardware (lines 443-450). Of course, this would not be feasible on current OBSs, which do not typically provide near-real time data routinely yet, but could be applied to networked stations.

18. The manuscript does not fully specify how spectrograms were generated. Like, what are the window size, overlap, FFT parameters, and normalization for spectrograms? Whether the data bandpass filtered before spectrogram creation? Whether spectrogram intensity was log-scaled or left linear. These choices strongly affect feature appearance and hence model learning. They should be documented in the main text, not just in supplementary code.

Thank you for this comment. We apologise for this oversight. We have now provided the data preprocessing parameters and spectrogram computation parameters in the supplementary materials (see Section S1

19. The manuscript emphasizes successes but could benefit from more examples illustrating model limitations. For example: Are there systematic false positives (e.g., ship noise mistaken for whale calls)? Do weak whale calls get consistently missed? Are resonances misdetected under certain current conditions?

Thank you for this suggestion. The point regarding systematic false positives is important and worthwhile exploring. We have significantly expanded the analysis for the blue whale vocalisation detection application in order to present a more comprehensive evaluation of the model performance. The results of this effort are now summarised in the main text (lines 535-546), and detailed in the supplementary materials (Section S5, Table ST5, lines 1130-1162).
We manually verified the model call detections for around 200 hours of spectrogram data sampled uniformly across the UPFLOW array. We found a very low false positive rate outside the expected vocalisation season ($\sim$ 5 false positives per station per week), corroborated both by the total number of detections in this period as well as our manually inspected dataset.
Interestingly, we found a significantly higher false positive rate during the active call season. We interpreted this as being partially caused by the repeating nature of whale calls, which the model learns but then incorrectly interprets other signals as whale calls during highly active call periods. We also find that earthquakes and broadband transient signals are the most common source of false positives, effectively leading to systematic false positives. However, the overall precision is still very high during the period (0.98), indicating that false positives make up only $1 - 3\%$ of the total detections in our catalogue.
We chose the blue whale vocalisation class as false positives are much clearer and easy to identify in an objective way than resonance false positives. In addition, resonances were so common relative to other signal classes that any systematic false positives caused by our other signal classes would be relatively minor.

—

0.1

Minor Comments

0.1.1

Abstract

- Line 18: "and an enhancement of over 90% for rare features" → revise to "and an improvement of over 90% for rare features."

Thank you, we have made this change.

- Line 20: What are earlier OBS deployments?

We have amended this for clarity.

- Avoid overuse of percentages without context — absolute numbers should be given.

See our answer above in point 16.

0.1.2

Introduction

- Line 27: "Ocean bottom seismology offers ...." – too generic opening. This can be made more specific as to why OBS is uniquely valuable compared.

  We prefer to keep the generic opening as Seismica has a wide audience.

- Line 29-30: "the exposure ..... for a seismic analysis" – change this sentence as in this manuscript, the seismic counterpart is not discussed.

  We have slightly re-written this sentence.

- Line 31: "One scientist's noise, though, is another's signal"; rephrase as: "What is considered noise for one scientific objective may represent signal for another."

  Thank you for this suggestion - we made an amendment along the lines you suggest.

- Line 32-37: The list of noise-related applications (ship tracks, ice calving, storms, marine life, currents) is excellent, but the references are a bit overwhelming. Consider grouping by theme (anthropogenic, cryosphere, oceanographic, biological) to improve readability.

  Thank you, we have tried to improve the readability of this paragraph by splitting it into several shorter sentences and grouping the phenomena by theme, as you suggest.

- Line 37-39: "modern machine learning methods..." jumps abruptly from literature survey to the current study. Add a transition explaining the gap: prior work has shown noise can be informative, but manual identification is inefficient → hence ML.

  Thank you for this comment. We agree and have added a transitional statement to motivate the ML techniques we introduce.

- Line 40: change "much prior" to "most of the prior."

  We have made this change.

- Line 45-46: "...treat all noise as equal" – too simplistic phrasing. Recast as: "Such approaches generally suppress noise without distinguishing between its diverse physical sources."

  We have made this change.

- Line 40-51: The discussion of ML in seismology is accurate but too broad. Several lines are written on arrival-time datasets (ISC-GEM, STEAD), which aren't directly tied to spectrogram-based segmentation. Streamline this so it doesn't distract from the main contribution of this study.

  The key aim of these lines is to point out that the successes of ML models in seismology have heretofore relied on very large annotated datasets. This was not conveyed well enough before.
  We have re-written this paragraph and restructured the introduction to make this clearer, and integrated this limitation within our overall motivation (see lines 83-89).

- Line 50-51: "...these often work in the time-domain (with a few notable exceptions...)" – This is good but make explicit that novelty is extending segmentation to spectrograms. Right now, it's implied but not emphasized.

  Thank you for this comment. This has been helpful in improving our introduction. We have re-written much of it to make direct comparison with prior work and explicitly state the novelty introduced here (see lines 68-79).

- Line 56-58: What are the certain conditions where the vortices' frequency matches with eigen-frequency? Any examples or a reference?

  We have included references for this statement, which give an in-depth physical explanation for these processes (lines 56-58).

- Authors introduce fin whales in line 59 but later focus on blue whales in line 69. This may confuse readers — clarify that the study concentrates on blue whale calls but acknowledges other baleen species.

  We have added a sentence to make explicit we focus on blue whale vocalisations here (line 67).

- Line 62: How low is the "low-frequency" and how high is the "high amplitude"? Please add approximate dB levels or propagation ranges to strengthen the claim.

  Thank you for this comment. We have added an indication of what we mean by high frequency and high amplitude, and now cite two relevant studies (see lines 62-63).

- Line 67: UPFLOW is mentioned here first. Please provide its extended name.

  We have made this change (see line 90).

- Line 68: "...train a range of ML algorithms..." – better to specify here that the core method is semantic segmentation on spectrograms. "Range" suggests multiple architectures, but later authors only emphasize segmentation.

  We have amended the text to make it clearer we are referring to different training strategies rather than different architectures / ML algorithms (see lines 91-93).

- Line 68: Maybe change "regions of OBS data" to "different signals recorded on OBS data."

  We have made this change (see line 92).

- Line 69: What is the maximum frequency these OBS recorded? Sampling rate?

  The sampling rate of the OBS used in this study is either 100 Hz or 250 Hz, depending on the instrument. We have amended the text to add this detail (see line 109). The exact breakdown is presented in the UPFLOW data paper [Tsekhmistrenko et al., 2025] currently in review, which we do not repeat here since it should not have a significant impact on the signals analysed in this work.

- Line 71-72: "...to the best of our knowledge, this work is the first..." – could be risky claim. Authors cite Choi et al. (2024) who did related work. Maybe phrase more cautiously: "This

study extends prior work by applying semantic segmentation directly to spectrograms for OBS data."

> Thank you for this comment. As mentioned elsewhere, we have significantly reworked the introduction to make explicit comparisons with prior works. This has allowed us to present the novelties of our work with greater clarity.
> While Choi et al. [2024] trains on 2-D spectrogram data, in fact it is just projecting 1-D first arrival probability time-series into 2-D, meaning that it is functionally the same as a 1-D first arrival picker that takes spectrograms as input. We have amended the text to make this clear (lines 71-75). We have also rephrased this claim throughout to "we introduce semantic segmentation for...", which we believe to be accurate (lines 11,76,604).

- Line 79-84: Remove redundancy: "presented... presents" is repeated.

> Thank you, we have made this change.

- Some sentences are long and wordy; break them into shorter, clearer statements.

> We followed the reviewer's suggestion and broke down the manuscript's sentences whenever possible.

0.1.3

Data

- Line 88: "Atlantic ocean" change to "Atlantic Ocean."

> Thank you, we have made this change.

- Line 92: Change "faulty" to "malfunctioning."

> We have made this change.

- Line 94: What are the station IDs for OBSs having instrument response failures.

> After some detailed revision of UPFLOW's instrument responses carried out as part of the study by Tsekhmistrenko et al. [2025], we solved the issues with one OBS that initially seemed to have response problems. Hence, no UPFLOW OBSs had instrument response failures and no associated station IDs are given.

- Mark these malfunctioning OBSs in Figure 1 with different colors.

> Thank you, we have made this change.

- Line 96: "...data quality was mostly high." Mention the percentage of stations having good quality data or noise level comparison.

> UPFLOW's data quality and noise levels are extensively investigated and reported in the data paper by Tsekhmistrenko et al. [2025] and hence we refrain from discussing it in detail in this study (also for conciseness).

- Line 96-97: Include reference(s) for vertical data component data showing lower long-period noise levels in previous experiments.

Similar to the previous point, we refer to the study by Tsekhmistrenko et al. [2025].

- Line 109: word "event" is misleading here.

  We have made this change.

- Line 110: authors can include some examples of auxiliary components.

  We have added a few examples of auxiliary components that can cause these current-induced sources of noise: e.g., antennas, flags, ropes, and floats (see lines 127-128).

- Line 116: What is the range of high-frequency signals?

  We have now explicitly stated that these signals occur above 1 Hz.

- Line 117-118: Include a reference to the resonant state?

  We have made this change.

0.1.4

Methodology

- Line 126: What method is used to generate a spectrogram? STFT? Welch? Or Windowing function?

  Thank you for this comment. We apologise for the oversight of not providing the precise parameters used for producing the spectrogram data. We used a STFT with a Hann windowing functon. We have added an extra section in the supplementary materials giving the precise parameters used to compute the spectrograms (section S1, lines 1064-1074).

- Line 126: What is the rationale behind selecting 15 min window for spectrograms.

  We have now expanded the text to explain our rationale. We are now more explicit that our primary goal in designing this data representation was to adequately represent the instrument resonance feature class (lines 147-153).
  We found a 15 minute window length struck a good balance between good representation of the desired signal classes, against the manual annotation effort and computational burden that higher resolutions or longer durations would have required.

- Line 133: "Other" class is underspecified — what kind of signals ended up there? Did the authors exclude ambiguous cases to reduce noise in training labels?

  The "Other" class comprised signals that we did not recognise and could not assign a physical mechanism to. It is probable that occasionally one of the designated signal classes was incorrectly assigned to "Other" (or vice versa), but most of the time they did not resemble any of the named signal classes.

- Line 138-139: Diminishing returns from adding more training data is claimed, but no evidence is given (no curve/figure). This statement sounds anecdotal; include quantitative support or drop.

This is a fair point. We have decided that since this is a fairly common observation (model performance generally scales sub-linearly with dataset size), we will drop this statement. We have simplified this section further since the points originally made here are not particularly important (see lines 154-158).

- Line 141: The rationale for 50 spectrograms each is unclear. Why not balance more across stations/types?

This is a good point to address and clear up in the manuscript. We agree that broader balancing across stations is one possible strategy. Our motivation in choosing 50 spectrograms per station/type was instead to allow controlled comparisons: first between two instruments in the same deployment, then between different instruments in different deployments. This design would have allowed us to more directly attribute differences in model performance to specific factors (instrument vs. deployment context), which would be harder to disentangle if we had distributed annotations more broadly across many stations (and therefore diluted statistical power to make such causal inferences). We have now amended this section to make this point explicitly (lines 170-172).
In practice, however, we did not observe significant model degradation between station types. As such, we did not feel the need to spend much more time discussing this aspect of our experiment design.

- Line 147: Scaling [0,1] is standard, but mention whether scaling was done per spectrogram (min–max normalization) or across the dataset (global normalization).

We thank the reviewer for pointing out this oversight. We have now amended this snippet to state that the data were clipped between a reasonable range (-150 dB to -250 dB [m^2/Hz], now stated in the supplementary materials), and scaled between [0,1] globally.

- Line 148: Padding from (60,399) → (64,416) seems arbitrary. Why 64 and 416? Was it to ensure divisibility by powers of 2 for CNN pooling layers? If yes, then 416 seems arbitrary.

This is not arbitrary, as one of the vision transformer models experimented with in the early stages of this project required input image dimension divisibility by 32, i.e. 2^5.

- Line 151: What are the references for "very limited annotated data is available"?

Our statement was not intended as a reference to prior studies, but rather to motivate the problem setting: in many practical scenarios there is little or no annotated data available for training. We have revised the text to make this clearer (see lines 179-180).

- Line 154-167: A lot of this is generic ML history (ImageNet, foundation models) — it could be tightened. It looks like a filler unless tied directly to seismic spectrograms.

We appreciate this comment. In retrospect, we agree that the framing and tone of this paragraph was too closely tied to the history of ML. We have therefore simplified and tidied up the discussion, keeping only the key details that are relevant to this study (see lines 188-196).
However, we believe that this section is relevant to the study insofar that it provides the background and explanation for the use of both CNNs and transfer learning. In addition, it presents some intuition for why a strategy of specialised synthetic pre-training may be useful (shared feature sets). In our opinion, it is still worth connecting our application to the relevant body of ML literature as it directly motivated this approach.

- Line 170-171: Authors mention "a huge amount of unlabelled data" but do not quantify. How many hours/days of spectrograms were available in total?

> The total amount of high-quality unlabelled data available is that recorded across the UPFLOW array, i.e. 43 stations $\times \sim 365$ days. We assume you are instead asking about what unlabelled data we used in practice. We have updated the methods section on semi-supervised learning to include that we use 5000 randomly sampled unlabelled spectrograms (see lines 261-262). This is a data volume 10 times larger than the training dataset.
> We did not use data from other stations as we wanted to keep this the same across all approaches to better evaluate zero-shot generalisability. We have now updated the Discussion to add that future work could address this for both the semi-supervised and transfer learning approach (i.e. background spectrograms from different stations; see lines 654-657).

- Line 179: Are the input spectrograms in gray color scale or RGB?

> Thank you for pointing out this missing detail. The input spectrograms were input directly as single channel arrays, effectively as greyscale images. We have amended this line and restated this point elsewhere (see lines 176-177, 200).

- Line 180: What is the encoder model depth here? (ResNet-18 or ResNet-50?)

> We have updated this snippet to explicitly state we use the ResNet18. As detailed in the Results section (lines 201, 345-349), we experimented with several different ResNet depths and found them to perform similarly.

- Line 192-196: State how many models were trained per approach (one per feature class $\times$ 3 approaches = ?).

> We have now amended the text to state that we repeat training several times for each approach (lines 223-224). Since the exact number of repeats varies for each signal class, we leave the details for the Results section (where we provide additionally commentary on the need for repeating training more for the whale call segmentation models, see captions of Tables 1 & 2, and lines 389, 429-431).

- Line 192-196: Clarify whether hyperparameters (learning rate, optimizer, epochs, batch size) were identical across strategies for comparability.

> We have expanded this section to state that the initial architecture optimisation was performed for the supervised learning approach (see lines 221-223). Once this architecture was optimised, we then optimised additional parameters (e.g. learning rates, training duration, batch sizes, as well as the approach-specific hyperparameters) for each approach separately.

- Line 205: How is the null class weight chosen? Grid search or heuristic?

> We have amended this section to state that w is a hyper-parameter (see lines 230-232). We leave the details exactly how this (and all the other hyperparameters) were chosen to the Results section (lines 334-338, 356-363), which seems a clearer way to organise this point.

- Line 261: Background data selection is unclear: how confident are authors that "very low probability" spectrograms are truly resonance-free? Provide a misclassification risk discussion.

Thank you for this comment. We have added a sentence to this section to state that we visually inspected these spectrograms to ensure that there were very few visible resonances (see lines 288-289).

- Line 264: Authors suggest IoU precision/recall because annotations are imperfect — valid, but a more rigorous way is also to include inter-annotator variability as an upper bound.

As discussed above, we have now added inter-annotator variability statistics to the Supplementary Materials.

- Close Section 3 with a sentence tying methods back to the focus of this study: e.g., "These methods aim to robustly distinguish resonance noise from biological signals across diverse OBS deployments."

Upon re-reading our revised manuscript, we found that actually adding such sentence would not flow well with the rest of the text and would feel a bit random. Hence, we prefer to not add such sentence.

0.1.5

Results

- Line 289: Training time (2–10 min) is reported in wall-clock terms. How many epochs does this correspond to?

We thank the reviewer for spotting this. Indeed, we forgot to state the number of epochs we trained the supervised learning approached for in our original version of the manuscript. We have now added that information to this paragraph (lines 326-332).
As discussed in the paragraph, the concept of an epoch is less meaningful for the semi-supervised and synthetic pre-training approaches. We present wall-clock time as a simple and concrete indication of the computational time required for model training.

- Line 295–296: Batch size of 10 is reported as optimal. Was this chosen due to GPU memory constraints or after systematic tuning?

Batch size of 10 performed better than larger batch sizes, and we were not restricted by GPU memory.
We presented results for larger batch sizes in the supplementary materials. We have now restructured this paragraph to make this clearer (see lines 357-358).

- Line 296-297: "Low weight decay improved results" — what value was used? Was this improvement consistent across multiple runs, or marginal relative to variance?

We now state the weight decay used for all experiments. As could be seen from Table ST1, removing weight decay (IoU $= 0.430 \pm 0.002$) performed worse than having weight decay during both stages of training (IoU $= 0.441 \pm 0.002$) at a significant level relative to the variance.

- Line 300–303: Why were Vision Transformers and DeepLabV3+ rejected solely on "scale" grounds? Could patch-based ViT approaches have been adapted for spectrograms?

We attempted to train ViTs and DeepLabV3+ on the resonance task and found very poor performance. We now report the performance of these experiments in the Supplementary Materials (see Table ST1). We attribute this performance to the heavy downsampling (x8 - x16 downsampling) of the images that each model performs. In the former case, this is due to the approach of tokenising patches of size 16x16 pixels, which makes assigning classes to individual pixels within each patch very challenging. In the latter case, DeepLabV3+ is preset to perform output strides of x8, which makes the default settings model impossible to use for fine-grained annotations.
Our aim in this work was to take off-the-shelf model architectures (which had ideally been pretrained on large natural image datasets, as discussed in the main text), and apply them to our small datasets. Of course, these architectures could have been adapted to our task, but this conflicts with one of our aims. In addition, we believe that the core components of these two architectures: patch-based tokenisation and output strides, are inherently poor choices for our task of resonance segmentation. We therefore did not expect significant performance gains even if we tailored these architectures for our task, and so did not attempt this.
We have now run a range of experiments for these different architectures and report the results in the Supplementary Materials (Table ST1, discussed in the Results section, lines 339-355).

- Line 304: ResNet18 performed as well as deeper models — was this true across all tasks (resonance + whale)? A small table showing performance vs. depth would support this claim.

We reported the results using a ResNet50 encoder for the instrument resonance task in Table ST1 (now Table ST2). We have added a new table exploring the various architectures discussed in the main text on both tasks for the supervised learning approach that supports this claim (now Table ST1).

- Line 310–312: Binary classification per-task outperformed multiclass. Is this due to limited data? Would multitask training help with larger annotated sets?

This is completely possible but would be speculative at this stage. We have mentioned this in the added limitations section (lines 658-663).

- Line 313–316: Null class weights ($w = 0.5$ and $w = 0.1$) appear arbitrarily tuned. Was a grid search used, or only trial-and-error? Why is the whale task unstable at higher weights?

The null class weights were chosen after a short, manual search between $w \in [0.01, 1]$, where it was quickly noticed that performance degraded significantly at the limits of this range. We have updated the text to state this explicitly (lines 334-337). We reported some examples of different w of these in the Supplementary Materials (now Table ST2), showing that performance fell either side of $w = 0.5$. We repeated this search for the whale vocalisation task, and found that after a maximum w value the model would only predict the null class. We believe this instability is a consequence of the extreme class imbalance in the whale dataset, leading to this failure mode of model collapse. We have updated the text to better convey this (lines 361-363).
We note that our goal was not an exhaustive hyperparameter search to optimise a best-performing architecture, but rather to demonstrate a method that is flexible across signal-classes and relatively easy to tune.

- Line 318–319: Ramp-up of λ_cons to 0.5 after 20k spectrograms — why 20k? Any evidence that this schedule was optimal?

In a similar vein as above, we performed a short manual search of schedules and found only a very minor impact on performance. This was documented in the supplementary materials (Table ST2). We therefore felt there was little need for an extensive optimisation of this parameter.

- Line 320–321: Why use 10 labeled vs. 100 unlabeled spectrograms per batch? Was this ratio tuned?

> This was the ratio between the labelled to unlabelled datasets, ensuring all the unlabelled data was seen each epoch. We found only minor sensitivity to this parameter so did not expand on it in the first submission. It was tuned like all other parameters. We have now added two rows to the supplementary table showing this (Table ST2).

- Line 325–327: The dataset split (80/20) yields 400 training and 100 validation spectrograms. Was k-fold cross-validation considered?

> We have now added a table exploring the stability of the model performance on the validation set, showing it to be relatively robust (see table ST3). As a result of this comment, we also performed a 2-fold cross-validation to demonstrate that our methodology yields results that are largely consistent with different train / validation splits.
> Future work could explore k-fold cross-validation to heavily optimise the model, particularly if there are significant gains to be had by further optimisation. Similarly to the above reply, we emphasise that the goal of this manuscript was not to produce exhaustively optimised models.

- Line 337–339: Authors claim models sometimes detect resonances missed by human annotators. Were these validated independently, or could they be systematic false positives?

> This statement is about resonances that were validated independently, such as the resonance in Fig. 5 that we discuss.

- Line 341–343: Only three training repeats were used to compute means/SE. Would the value change with more training repeats?

> The SEs were relatively low across the board, indicating that the mean is already stable at a precise enough level to draw the conclusions we did. In addition, the new supplementary materials table (Table ST3) exploring the stability of the model performance on bootstrapped subsets of the validation set show a very clear downward trend in the standard error. We believe this is sufficient evidence that the mean and SE estimates are reliable enough to draw the conclusions we do.

- Line 355: RR40 dataset includes only 50 spectrograms. Was model performance stable under bootstrapping/resampling?

> Yes. This is now addressed in the added supplementary table on bootstrapping the evaluation datasets (Table ST3).

- Line 361–362: Authors claim transfer learning raises the "skill ceiling." Could the gain instead reflect implicit data augmentation (larger synthetic set size)?

> On reflection "skill ceiling" is a clumsy expression and not particularly descriptive. What we intended to convey is that, for this dataset, the supervised model consistently plateaued at around IoU $\sim$ 0.42, even after extensive architecture experimentation and hyperparameter tuning (see Tables ST1 & ST2). Synthetic pre-training, however, provided an immediate performance gain ($\sim$ 0.44 IoU). We agree that one plausible interpretation is that pre-training acts as implicit data augmentation by increasing the effective dataset size.
> We have amended the text to more precisely state what we intended (lines 408-412).

- Line 364–366: Authors mention whale calls appear in "10% of spectrograms." How many total positive examples does this correspond to (absolute count)?

We have now added absolute count statistics of all the feature classes to the supplementary materials (see Fig. S1).

- Line 366–368: The threshold for whale calls is set to $\tau = 0.8$. Was this value optimized on a validation curve, or chosen heuristically, as it looks different from that one used for instrument resonance segmentation?

  As stated in the text, this value was chosen heuristically to optimise for separate objects in the mask for each whale call. We have slightly amended the text to make this clear; see lines 418-420.

- Line 367: remove using (got repeated).

  We have made this change.

- Line 370–372: Authors note "stark difference between the quality of annotations." Does it refer to inter-annotator disagreement, inconsistent labeling, or inherent call ambiguity?

  This was incorrectly phrased on our part. We meant "stark difference in model predictions", rather than annotations. The point is to highlight that the different training strategies lead to very different model predictions for the whale call segmentation task. We have amended the phrasing.

- Line 376–377: Identifying individual calls is framed as "essential for source location inversions." This is an important applied point; authors could expand on how segmentation accuracy translates to location accuracy.

  To be clear, the point here is that getting call arrival timestamps is a precondition to performing any call association. This was only possible with a segmentation model that produced individual, isolated prediction regions for each call.
  The location accuracy itself is then much more determined by the resolution of our spectrograms and inaccuracies in the propagation model. Nonetheless, we used a conservative error estimate of the arrival time ($\sigma = 30$ s), which leads to large uncertainties in the location accuracy. We have rewritten parts of this section to make the approach clearer (lines 586-592).

- Line 385: The conclusion here implies synthetic pretraining is broadly superior for rare classes. Authors should note whether this generalizes to other whale species or is specific to blue whales at ∼17 Hz.

  We believe that in principle this could be extended to other classes of signals, including other types of blue whale vocalisations or indeed calls from other species. However, this is speculative and would need to be tested explicitly in future work. We therefore only alluded to this in the conclusions (lines 619-625, 670-672).

- Line 393–395: Runtime performance is reported (2 h for 43 stations). Is the runtime linear with station count?

  Yes, the runtime is linear with spectrogram count and therefore (to a very good approximation, less slight differences in station operation durations) linear in station count. We have amended the text to explicitly state this (lines 443-450).

- Line 396–398: Background power estimation method (mean over 0.75 Hz neighborhood). Why 0.75 Hz? What was the sensitivity to window size?

The 0.75 Hz neighbourhood was chosen to ensure there were adequate non-resonant pixels within each band to provide a reliable estimate of the background noise level. The resonance width generally varied between 1 - 4 pixels, corresponding to 0.125 - 0.5 Hz. The chosen neighbourhood was therefore large enough to ensure non-resonant pixels could be included in the estimate, while still allowing a fine-grained characterisation of the frequency dependence of the background noise. We have amended the text to state this motivating rationale (lines 457-465).
We visually inspected a number of spectrograms to check that the background power estimation method produced reasonable results; the estimated power (visually) matched the background power outside resonant bands, and the estimated power varied relatively smoothly in frequency. This indicating that there were no high variance regions (i.e. the band was wide enough). We did not explore sensitivity to this window size (in part because, lacking some "ground truth" to compare to, any comparison would be qualitative in nature).

- Line 415–417: Both time series are normalized to [0,1]. Is this per-station normalization or global across the array?

This is per-station and per-series normalisation. We have sightly reworded this snippet in the text and the caption to make this clear (see line 489, Fig. 8).

- Line 435–436: Over 1000 calls at most stations. Are these detection numbers uniform across all stations used or any pattern followed?

Thank you for pointing out this paragraph did not discuss the spatial pattern of call detections. We had left the bulk of this discussion to the following paragraphs discussing the animation.
We agree that a quick description of the spatial pattern of detections in Fig. 9 is helpful, so we have amended the text to include this (lines 517-524). More detailed analysis is presented in the following paragraphs.

- Line 452–454: False positive rate is described as <100/week in summer. How does this compare to false negatives during peak season?

We have now performed a more systematic analysis of the false positives, as discussed above. However, false negatives are much more challenging to analyse as they are dependent on defining an SNR floor. We leave this more complex task to future work where estimating false negatives is essential, such as, e.g., for density estimation.

- Line 483–484: Claim that UPFLOW station spacing is too wide is reasonable — but what is the expected timing precision (in seconds) vs. required for association?

Of course, for sophisticated association algorithms, improved precision would be necessary. However, in the examples discussed in this work the precision of the whale call detection is not the primary issue. At 150 km spacing, call arrival time differences (> 60 s) approaches and surpasses the inter-call intervals ($75 - 100$ s), which makes the very simple association approach used in this work infeasible. We have now amended the text to make this clear (lines 577-580). We now mention the need for more precise arrival times for more sophisticated association here (lines 571-576) and in the limitations section, and have expanded the motivation for our approach at the start of this section.

- Line 493–495: Equal differential time likelihood assumes homogeneous sound speed (1.5 km/s). How would the results change if a layered velocity model is used (e.g., using regional CTD profiles)?

As already indicated in the text, our simplistic velocity model likely leads to inaccuracies in the inferred track. We therefore implicitly expect results to change with added realism, but cannot speculate exactly as to how. We feel that more involved analysis using a realistic velocity model is beyond the scope of this work.

0.1.6

Discussion and conclusions

- Line 506–507: "first application of semantic segmentation for time-frequency representations of seismic data". Suggest softening: "among the first applications. . . " or explicitly contrasting with existing related work (e.g., Choi et al., 2024; Cotillard et al., 2024).

  Following the changes made in the introduction, we have softened this statement. In addition, we followed your earlier suggestions above by amending the Introduction to make explicit comparisons to prior work (see lines 71-75). Hopefully the novelty of our contribution is now much clearer from the Introduction text, and so we refrain from mentioning these prior works here again (also for conciseness).

- Line 512–514: Authors note flexibility of simulated→real transfer. Could they briefly mention failure cases (e.g., which classes transferred poorly)?

  We did not find an example where classes transferred poorly. Both the examples in this study show improvements in performance. We refrain from speculating about other classes that we did not explore.

- Line 518–522: Good motivation for rare signals (tremor, foreshocks, deep EQs). Could authors cite relevant examples where data scarcity limited ML progress?

  Following a suggestion above, we have reworked the introduction to better introduce our motivations for working with small datasets. We have now added a range of examples where data scarcity has led to issues with ML model performance for a range of different signal types (e.g. LF events, foreshock sequences, etc.), including some work which explicitly demonstrates deep ML models can perform worse than simpler alternatives in this regime (see lines 623-625). We think that grouping both the motivations and motivating examples in this section of the introduction is a sensible solution.

- Line 524–525: Unique contribution of pixel-level annotations is strong. Recommend emphasizing interpretability (duration, frequency bandwidth) as a differentiator over binary classifiers.

  We slightly amended this text to make direct comparison to the classification approach and now note the improved interpretability, per your suggestion (see lines 626-630). We have also re-stated that we have demonstrated these benefits in the work, as well as highlighting future avenues.

- Line 544–546: Authors highlight potential for multi-class segmentation. Suggest noting whether computational cost scales linearly with classes or if memory is a practical limitation.

We have already noted that multi-class segmentation would enable a significant improvement in computational efficiency. We feel a more detailed discussion of the scaling costs would be misplaced in the manuscript since it is a standard observation in ML multi-class prediction. For completeness, we provide a technical justification for the reviewer here.
In standard segmentation architectures, extending from binary to multi-class segmentation typically only increases the number of output channels in the final prediction head (one channel per class). The resulting extra memory and computational cost are therefore confined to the final head and are negligible compared with the dominant cost of the shared encoder–decoder backbone. In practice, adding a small number of classes (e.g. going from 1 to 2, or a few more) produces only a very small memory overhead. So the computational and memory cost scale very much sub-linearly, and are therefore not practically important. We therefore feel the section as currently written is accurate and further discussion would dilute the conclusions.

- Line 566-570: Last paragraph is wordy and repetitive; streamline.

We have re-written this paragraph to streamline our final conclusions (lines683-688). This now simply emphasises our key contributions.

0.1.7

Figures and Tables

- **Figure 1:**
  - Mark the malfunctioning OBSs in different colors.
  - Labels are too small to read in panel a).
  - Provide a bathymetry scale distance scale in panel a).

We have amended the figure to make these changes.

- **Figure 3:**
  - X and Y axis labels and ticks are missing.
  - Is the intensity color bar of the same values for all sub-figures?

We have amended this figure to add ticklabels. We have also ensured the same intensity colorbar is used in all subfigures.

- **Figure 7:**
  - Try to make this figure colorblind free (avoid usage of green and red lines together).

We have removed the green lines in these figures.

- **Figure 9:**
  - The labels are too small in the first panel.
  - In Panels i)-iii), avoid green and red colors together.

Thank you for these suggestions; we have amended the figure to make these changes.

- **Figure 10:**

- Some of the labels are overlapping.
- Latitude and longitude labels are of a different notion compared to Figure 1 and Figure 9, consider making all of them consistent.

  We have fixed the overlapping labels and made the latitude and longitude units consistent with the other plots.

- **Table 1, 2 and ST1:**
  - Some values are shown with 3 significant decimal points and others with 2; explain the reasons behind this discrepancy or make them consistent.
  - Abbreviations like IoU, UPFLOW, RR40 should be spelled out at first mention in each table.

    All done, except that for conciseness we preferred not to spell out UPFLOW in each table and, instead, just spell it out the first time it appears in the manuscript.

0.1.8

GitHub repository is mentioned, but:

- Clarify whether annotation files will be made available.

  We have made the manual UP05 annotations available on GitHub and Zenodo in COCO format. We have also added an example demonstrating how to load the annotations in Python.

- State version numbers of key packages (PyTorch/TensorFlow, etc.).

  We now state the key version number of PyTorch in the Data Availability section.

- Vaibhav Vijay Ingale (Scripps Institution of Oceanography)

1 Reviewer 2

This study proposes a 2D segmentation method of the time-frequency representation of OBS signals in a low annotated data context. It then shows concrete applications for bottom ocean current and baleen whale migration analyses. This is a solid work that deserves publication in *Seismica*. Yet the manuscript lacks a bit of clarity on some aspects of the methods and about some advantages compared to the literature. I think the first application could be pushed a bit further. And maybe the second application could also further insist on what could be done more compared to existing whale detectors. But overall, I think the authors could solve that with minor to moderate reviews.

I do not have that many general comments, more a list of specific points. Good luck with the revision!

Line-by-Line Comments

- L126-129: For reproducibility, you could make clearer what the original sampling rate is, the sliding window length (30 seconds?), tapering (Hanning?), the overlap chosen (50%?), and which is the first and second dimension for the (60, 399) pixels (time, frequency?).

  We have now provided the data pre-processing parameters and spectrogram computation parameters in the supplementary materials (Section S1). We have also slightly modified the sentence to clarify the order of the time-frequency dimensions relative to the pixels (see lines 146-147).

- L130-135: You did not manually select sections of interest? How many earthquakes and whales did you get? How many files with only noise? I think clearly showing the data imbalance would be interesting for the reader. Also, you then often talk about this concept of having a lot or not of each example in the training dataset.

  We agree that it would be very useful to explicitly show the data imbalance. We have therefore added a figure to the Supplementary Materials showing the number of pixels belonging to each class, as well as the number of objects belonging to each class (Fig. S1). We highlight the class imbalance in the main text on lines 161-163 and refer to the exact breakdown in the supplementary materials (Fig. S1).

- L130: How much time did it take?

  We have now updated the text to give a rough indication of the number of hours taken to complete the annotation of the main annotated dataset (lines 159-160).

- L136-139: You consider equal manually annotated data and automatically trained one? Meaning that the model learns from itself? Is this something conventional?

  We had utilised a trained model to produce a first pass of annotation to aid the annotator. The annotator then went through and modified the first pass, as well as removing and adding extra annotations where the model had failed. However, this had a very minor effect as significant manual modifications had to be made to the spectrograms. So, the model was not learning from itself; it was just an experimental way to speed up annotations.
  Given this comment and those made by another reviewer, we have decided to streamline this section and remove the mention of this step. We feel it did not affect our methodology and results enough to discuss, and detracts from the overall clarity of the method.

- L147: Say a word about that [0,1] normalization (min/max?) and about the different sampling rates that might lead to different original spectrogram sizes?

  We apologise for these oversights. We have now expanded this section to be clearer about the normalisation and spectrogram processing. We first clip all spectrograms globally within a suitable dynamic range, and then scale these between [0,1] (now stated in the main text, lines 174-175). We have also made it clear that the raw spectrograms are linearly interpolated onto a target grid with fixed dimensionality, which standardises the size of all spectrogram data regardless of the original sampling rate. As stated above, we have now also added a section to the supplementary materials with all this information alongside the precise spectrogram computation parameters (see Section S1).

- L187-189: If you perform binary segmentation, shouldn't you have only one output between [0, 1]? Then how do you attribute each pixel to one class? Do you use the highest probability? Also, how do your target labels look? Do you replace other classes with noise? Maybe you

could try using a unique model that outputs several binary masks with each a separate sigmoid activation rather than softmax. I think this is what is done by EQTransformer for example.

> We realise that our original wording was unclear. Each of our models is trained in a binary segmentation setting (target feature vs. background). The model outputs a single probability mask with values in [0, 1] indicating the likelihood that each pixel belongs to the target class. Final label predictions are obtained by thresholding this probability map.
> To avoid confusion, we have revised the text to make this clearer (see lines 207-211 of the revised manuscript).
> Conceptually, this is similar to 1-D segmentation approaches such as EQTransformer, but applied in two dimensions to spectrograms: each spectrogram pixel is assigned a probability of belonging to the target class. Other signal classes are treated as background (i.e. the null class) during training. While this can be extended to a multiclass setting with multiple output masks (as the reviewer suggests), in practice we found that training separate binary models provided better performance.

- L256 & L269: I think you should explain here or in the supplementary how this is done. Providing a link to some code does not really help the reader. The 'README.md' of the repo does not clearly contain the useful information. In Figure 4 you could show the background, the synthetic features, and the combination of both; this would make the process clearer.

> We appreciate this perspective. We understand that more detail regarding the synthetic feature generation was required.
> We have now added a new section to the supplementary materials, which describes in more detail our approach for generating these synthetic features (Section S2). We have also added two more figures to the supplementary materials (Figs. S2 and S3). The first is directly inspired by your suggestion here: we break down the synthetic feature generation procedure into the different steps, hopefully making the process clearer for the interested reader.
> The second figure then presents a statistical comparison between the features in the annotated datasets against the synthetic feature generators we used for pre-training. This shows both the strengths and limitations of our very heuristic feature generating approach.

- L272-273: By curiosity, why not using the F1 score? Maybe just say a word if it makes sense. Probably around L284-286.

> Thank you for the comment. We chose not to report the F1 score because in semantic segmentation tasks IoU is generally the headline metric and is widely used for benchmarking. IoU is mathematically related to F1 and thus conveys essentially the same information, while being more interpretable in terms of spatial overlap. We also report precision and recall separately (see Tables 1), as these are more directly interpretable than their harmonic mean, particularly when recall is prioritised in scenarios such as missed manual annotations.
> We have added a brief aside to the manuscript to note these factors (see lines 309-311).

- L326-329: Did you make sure that in the validation set you had stations that the model never saw while training? Do you expect your model to generalize over novel stations? You probably already said it in the data section but as you remind this here you could also reclarify it here. When I read back the data section it refers to 500 samples in dataset 1 and 2x50 samples for dataset 2 and 3. Here you talk about 400 samples for training. Could you make it clearer?

We have re-written (and moved) this section to make it clearer (see lines 321-325). Only data from UP05 were ever seen during training, and the two extra annotated datasets were only used for model evaluation. We are explicit that the 400 to 100 training - validation split only refers to the 500 UP05 annotated spectrograms.
We hoped that the model would be usable for stations not seen in training, but we did not expect the model to generalise over the novel stations as well as it did. We have added some more discussion regarding this in both the Methods section as well as the Discussion section (see lines 168-172, 613-618). We make it clear that these extra stations were chosen to explicitly evaluate model degradation (probing different instruments, and different noise backgrounds).

- L385: What about earthquakes? Did you have a too small number of them to measure metrics?

Yes, there were relatively fewer earthquakes. The main reason we did not analyse them here was that earthquakes are not the focus of this paper and we have limited space (lines 41-43). We prioritised under-explored signals that lacked large datasets, or existing models for detection. We intend to apply our methodology to earthquakes in the future.

- L393-408: It could be interesting to see a plot of that year-long time-series, maybe in the supplementary? Otherwise on that part, it is not clear what the model is useful for. Does it only help to estimate the background noise and remove it? Probably the method explanation is not super clear. Are you only using the 0.75Hz band? Are you stacking energy of all detected resonance? Does the model here allow to nicely separate energy from resonance to other signals? What would happen if you just selected the energy of the most dominant resonance band and just use that without any further processing to make your yearly series? I think this part should more clearly explain the gain of using the trained model. Maybe by comparing with a naive approach that does not use it?

We have re-written this paragraph to make the method explanation clearer (see lines 457-465).
Our model predicts which pixels correspond to resonances (and, by complement, the non-resonant pixels). We could then directly estimate the resonant energy of these pixels, but this could be biased by a time-varying, independent background noise. So we therefore estimate the background power spectrum using the negative of the masks, and subtract that from the resonant pixel energies. The "band" just allows us to more reliably estimate the background in-case an entire row (frequency bin) is resonant.
We agree that we left the analysis of the instrument resonances underexplored. We have now significantly extended the analysis and added extra figures to the supplementary materials (Figs. S5 and S6). We have added a comparison between the "naive" approach of averaging the total energy in the spectrogram and computing the PSD. We see that our approach significantly amplifies the tidal peaks identified in Fig 7.
In addition, we have added an extra figure to the supplementary materials where we have more directly explored the resonance properties at different frequency bins (Fig. S6). In the course of this analysis, we found a resonance at 7 Hz that seemed to be directly excited by the tidal current flow. We demonstrate that a very clear tidal peak can be seen from resonances in this very narrow band, enabled by our time-frequency segmentation approach.

- L418-426 & Fig. 8: Those scatter plots are quite messy. Most of the time you have no correlation between both variables. The Resonant energy seems to be triggered only in certain conditions while the projected current velocity is always oscillating. Maybe you could plot the correlation against time (compute it on sliding windows) this could be more informative. This part is intriguing but as a reader, I would expect the authors to try to make the data a little bit more interpretable.

Inspired by this suggestion, we have now produced an example of the year-long time series, shown in the Supplementary Materials (Fig. S5). This shows clear correspondence between the resonant energy and the tidal currents for long periods of the deployment, and complements the scatter plots in the main text (lines 493-496).
We experimented with plotting correlation over time or a Hilbert amplitude envelope scatter, but found these to be harder to interpret than a simple scatter plot. We believe that the comparison of the two series suffices to show periods of clear dependency between the resonances and the currents.

- L429-430: Do you mean that you only kept detection in a given frequency band around 17Hz? What band exactly?

We have clarified in the text that detections were restricted by multiplying the predicted segmentation mask with a binary frequency mask spanning 16–18 Hz around the 17 Hz blue whale call band (lines 507-511). This post-processing step not only reduces false positives but also illustrates the flexibility of the spectrogram segmentation approach, which allows prior expert knowledge to be incorporated in a straightforward way.

- L432-433: "We leave an investigation of the sensitivity of the model performance to this threshold for future work." Will you? I would rather say that those parameters provided satisfactory results.

Thank you for your encouragement. We have amended the text to explain why this threshold provided satisfactory results, rather than intimating at future work (see lines 513-514).

- L455-456: Again, it sounds a bit as if you were lazy to do that. What kind of analysis do you have in mind that would require to be postponed for another study? Maybe you could better explain the current limitations and, in the perspective part of the article, mention it as a possible improvement.

Per another reviewer's suggestion, we have extended the false positive rate analysis by manually validating around 200 hours of whale call predictions across the UPFLOW array. We have provided further analysis at this point (lines 535-546), and detailed results are presented in the supplementary materials (Section S5).
We have given some indication of the current limitations and requirements for future work. In particular, estimating whale density requires more in-depth analysis of spatio-temporal false positive rates, as well as quantification of false negative rates as a function of call SNR.

- L461: "Future work will explore..." Do you really commit yourself to do that? Up to you but I would be more evasive.

Thank you again. We have decided to amend this to "further work would be required".

- L479-482: Standard arrival-time techniques are equally demanding on the bathymetry and so on than the other models. You do not have a three-component method nor the resolution for multipath analysis.

We agree that our original explanation was inaccurate. As you note, standard arrival-time localisation methods are also dependent on (and can be biased by) bathymetry and propagation effects. The point we intended to make is that, given our coarse temporal resolution (15 s spectrogram frames), we cannot resolve multipath arrivals or finer time differences, so applying more advanced propagation models would not provide much additional benefit or more interpretable results.
We have rewritten the paragraph to make this explicit and explain that our simplified approach is just a pragmatic one (lines 571-576).

References

Sugi Choi, Bohee Lee, Junkyeong Kim, and Haiyoung Jung. Deep-learning-based seismic-signal P-wave first-arrival picking detection using spectrogram images. *Electronics*, 13(1):229, 2024.

Carlos Corela, Afonso Loureiro, José Luis Duarte, Luis Matias, Tiago Rebelo, and Tiago Bartolomeu. The effect of deep ocean currents on ocean-bottom seismometers records. *Natural Hazards and Earth System Sciences*, 23(4):1433–1451, 2023.

Douglas Gillespie, Marjolaine Caillat, Jonathan Gordon, and Paul White. Automatic detection and classification of odontocete whistles. *The Journal of the Acoustical Society of America*, 134 (3):2427–2437, 2013.

Oleg A Godin, Tsu Wei Tan, John E Joseph, and Matthew W Walters. Observation of exceptionally strong near-bottom flows over the Atlantis II seamounts in the northwest Atlantic. *Scientific Reports*, 14(1):10308, 2024.

Marie A Roch, T Scott Brandes, Bhavesh Patel, Yvonne Barkley, Simone Baumann-Pickering, and Melissa S Soldevilla. Automated extraction of odontocete whistle contours. *The Journal of the Acoustical Society of America*, 130(4):2212–2223, 2011.

Simon C Stähler, Karin Sigloch, Kasra Hosseini, Wayne C Crawford, Guilhem Barruol, Mechita C Schmidt-Aursch, Maria Tsekhmistrenko, J-R Scholz, Alessandro Mazzullo, and Martha Deen. Performance report of the RHUM-RUM ocean bottom seismometer network around La Réunion, western Indian Ocean. *Advances in Geosciences*, 41:43–63, 2016.

Maria Tsekhmistrenko, Ana M.G. Ferreira, Miguel Miranda, Samaneh Baranbooei, Roberto Cabieces Diaz, Mafalda Carapuço, Carlos Corela, José Luis Duarte, Henrique Ferreira, Wolfram Hartmut Geissler, Katrina Harris, Stephen P. Hicks, Kasra Hosseini, Kuan-Yu Ke, Frank Krüger, Dietrich Lange, Afonso Loureiro, Peter Makus, Augustin Marignier, Marta Neres, Luís Ramos, Theresa Rein, Alex Saoulis, David Schlaphorst, Mechita C. Schmidt-Aursch, and Frederik Tilmann. Performance of the 2021-2022 UPFLOW large ocean bottom seismometer array in the Azores-Madeira-Canary islands region, Atlantic Ocean. *Seismica*, 2025.