RESEARCH ARTICLE

Semi-automated seal detection on the Western Antarctic Peninsula: an unsupervised machine learning approach for detecting ice seals in aerial survey data

Claire McGinnity[1], Connor C.G. Bamford[1] (iD), Nathan Fenney[1] (iD), Andrew Fleming[1], Jaume Forcada[1], Michael S. Tift[2], Luis A. Hückstädt[3,4], Daniel P. Costa[4,5] & Peter T. Fretwell[1]

[1]British Antarctic Survey, Natural Environmental Research Council (NERC), High Cross, Madingley Road, CB3 0ET, Cambridge, UK
[2]Department of Biology and Marine Biology, University of North Carolina Wilmington, Wilmington, North Carolina, USA
[3]Centre for Ecology and Conservation, Faculty of Environment, Science and Economy, University of Exeter, Penryn Campus, Cornwall TR10 9FE, UK
[4]Institute of Marine Sciences, University of California, 115 McAllister Way, Santa Cruz California, 95060, USA
[5]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, California, USA

Abstract

Over the past 25 years, the Western Antarctic Peninsula (WAP) has experienced dramatic shifts in sea ice extent. This change has coincided with rapid alterations in ice-dependent ecosystems, including those supporting crabeater seals —the most abundant Antarctic seal and one of the largest mammalian consumers of krill. Despite their ecological importance, population estimates for ice seals remain scarce due to the difficulty of surveying large-scale, remote, ice-covered habitats. In 2023, during an abnormally low sea ice year, we conducted aerial surveys over Crystal Sound and Marguerite Bay during the end of the breeding season, flying over 1000 km of transects. Seals were extremely sparse in the resulting imagery—occupying less than 1% of the surveyed area. This posed a significant challenge for both manual annotation and automated detection. Here, we present a semi-automated, rule-based image analysis pipeline to substantially reduce human annotation time. Our method leverages hierarchical clustering with just two tuneable parameters, avoiding the computational burden and opacity of deep learning models. Using this method, we identified 758 seals within an $\sim 350 \text{ km}^2$ survey subset, achieving a test recall of $79\% \pm 9.1\%$. In the absence of concurrent tagging data to estimate haul-out corrections, we refrain from extrapolating to a population estimate. However, the low observed densities highlight the urgent need for continued monitoring. Our improved data processing pipeline is a key step in facilitating the large-scale analysis required to inform conservation strategies for this key species.

Introduction

Since the peak in Antarctic sea ice extent in 2014, subsequent years have experienced dramatic, nonlinear declines (Diamond et al., 2024; Ionita, 2024; Parkinson, 2019). These declines deviate from IPCC projections (Gilbert & Holmes, 2024) and coincide with atmospheric warming far exceeding the global average (Cook et al., 2016; Henley et al., 2019; Moffat & Meredith, 2018; Siegert et al., 2019; Turner et al., 2005). Together, these changes have caused profound impacts on the physical environment of the Antarctic Peninsula.

In terms of sea ice coverage and seasonality, sea surface temperatures have risen (Meredith & King, 2005; Smith &

Polvani, 2017; Vaughan et al., 2003); seasonal sea ice thickness has decreased, and its persistence has shortened by 3 months (Henley et al., 2019; Parkinson & Cavalieri, 2012). Glaciers have thinned (Cook et al., 2016), freshening coastal waters (Azaneu et al., 2013; Dotto et al., 2016), which in turn has triggered biological responses. Notably, there has been a shift towards gelatinous zooplankton assemblages (Schofield et al., 2017), along with a southward redistribution of Antarctic krill (*Euphausia superba*; hereafter, krill) biomass (Atkinson et al., 2017, 2019; Hückstädt et al., 2020). There has also been a significant reduction in suitable habitat, particularly pack ice, for ice-dependent penguin and seal species (Clucas et al., 2014; Forcada et al., 2012; Trivelpiece et al., 2011).

Among the ice-dependent seal species on the Western Antarctic Peninsula (WAP), crabeater seals (*Lobodon carcinophaga*) are by far the most abundant, with an estimated population exceeding 1.8 million in the WAP (Forcada et al., 2012), the highest known densities of this species anywhere on the continent. Despite their name, crabeater seals rely on krill for > 88% of their diet (Hückstädt et al., 2012), making them a key indicator species (Hazen et al., 2019) as one of Antarctica's most specialized foragers and one of the most important air-breathing consumers of krill (Hewitt & Lepsky, 2009; Hill et al., 2006; Hückstädt et al., 2012; Laws, 1977). However, data collection on crabeater seals is challenging due to their year-round residence within the pack ice (Burns et al., 2004; Forcada et al., 2012), a notoriously difficult habitat to sample without disturbance, and the fact that they do not aggregate at breeding colonies nor haul out at predictable locations. Crabeater seals are also sighted at sea, and in large numbers, adding to the complexity of data acquisition. As such, our understanding of their population is coarse and out of date, an issue amplified by the fact that the ice conditions on the WAP have changed rapidly in recent years.

Population metrics are essential to inform and validate theoretical models, which in turn guide management and conservation strategies (Torney et al., 2019). Remote sensing overcomes many of the traditional difficulties of surveying pack-ice species by ship (Bester et al., 2021), but presents new challenges, namely the processing of the drastically increased volume of data. Satellites and aircraft (both piloted and remotely operated) are the main platforms currently used for animal detection (Xu et al., 2024). However, these platforms, which are often photo-based, produce data volumes that are impractical to process manually (Charry et al., 2021; Stapleton et al., 2014). To overcome processing demands, researchers have implemented various methods including citizen science campaigns to perform a 'search area

reduction' of > 260 000 km² of fast ice (LaRue et al., 2021) and developing convolutional neural networks (CNNs) to detect pack-ice seals (Gonçalves et al., 2022), wildebeest (*Connochaetes taurinus/gnou*) (Wu et al., 2023) and whales (Borowicz et al., 2019). However, deep learning approaches such as CNNs inherently lack transparency, interpretability and explainability due to the model structure (Perry et al., 2022) and may fail to generalize (Wu et al., 2023) outside their original, often narrow, training environment. This may in part have led to lower uptake of such methods, particularly with the aforementioned species and within the ecological field.

Deep learning is a powerful tool but requires large volumes of labelled training data that are both specific to and representative of the survey environment (Whang & Lee, 2020). Published models are often highly case-specific and typically require retraining for each new application (Wu et al., 2023). Another major challenge is the scarcity of publicly available training datasets for wildlife detection (Xu et al., 2024); those that do exist are often narrowly focused on a particular species, geographic region, or datatype (Cubaynes & Fretwell, 2022). Consequently, deploying deep learning in new conditions demands extensive and tedious manual data labelling, a process that is acknowledged not to produce perfect ground truth on which to train and evaluate a model, with errors in the training labels significantly reducing the accuracy of neural net classifiers (Bishop, 2014; Heaton, 2018; Hestness et al., 2017; Mnih & Hinton, 2012; Zhang et al., 2021). Models with higher capacity, with parameter count as a common proxy for model capacity, require disproportionately more data to avoid overfitting (Bishop, 2014; Heaton, 2018; Hestness et al., 2017). High-capacity models can even memorize noisy labels (Zhang et al., 2021), underscoring why label quality and quantity remain a fundamental constraint for training large models.

An especially challenging attribute of aerial surveys for pack-ice seals is the sparsity of targets (i.e., ice seals) within the vast background. The categorization of binary classification tasks for which a single class represents <15% of the distribution is regarded as highly unbalanced (Alsheri et al., 2022). However, for seals in pack-ice environments, individuals are found to occupy <1% (Forcada et al., 2012; LaRue et al., 2020) of the background, a trend similarly reflected in our data (Fig. 1).

Our classification problem is therefore a case of extreme imbalance, which presents challenges for algorithms and human annotators alike. Imbalance is known to bias the predictive performance of many classification algorithms towards the majority class (Hasanin
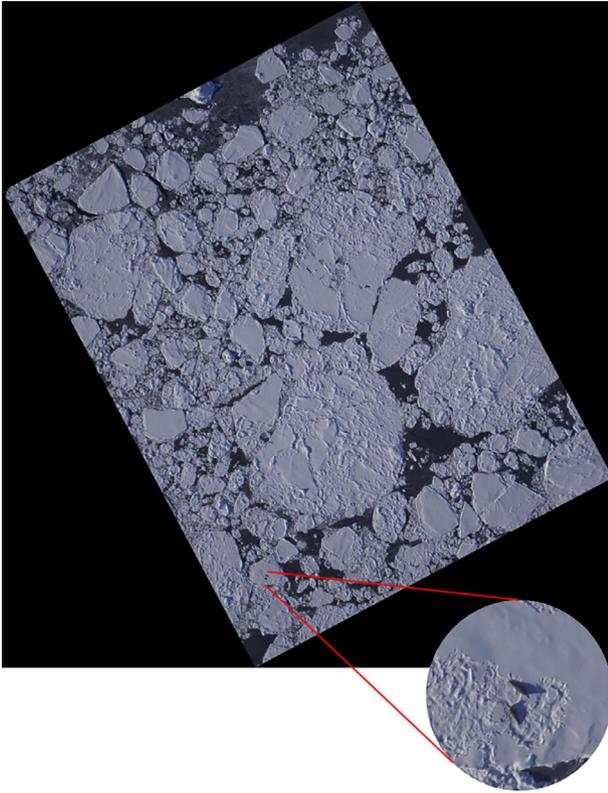
Figure 1. Example of a single aerial image captured during the 2023 survey of the Western Antarctic Peninsula (WAP). Seals occupy less than 1% of the surveyed area, posing a significant challenge for classification algorithms as imbalanced data sets often skew predictive performance towards the majority class, increasing the risk of seals going undetected.

et al., 2019), which in our case would be negative. Manual annotation suffers from similar performance problems; for instance, Rožanec et al. measured a 3-fold increase in precision of human annotators when labelling a balanced dataset over an imbalanced set, linking this to improved participant attention (Rožanec et al., 2022). Additionally, the quality of human labelling is known to degrade over time with fatigue (Rožanec et al., 2023), which presents a further concern for the processing of vast quantities of data produced in wildlife surveying.

In the face of the challenges posed by data sparsity, class imbalance and the limitations of both deep learning and human annotation, we adopt a parsimonious approach to analyse data from a set of aerial surveys flown over Crystal Sound and Marguerite Bay in 2023 (Fig. 2). Rules-based algorithms have been successfully deployed to significantly reduce researcher workload. For example, the count time for humpback whales was reduced from 24 h to $\sim$30 min by filtering the satellite data using just a shape criterion (Laws et al., 2003).

Building on this rationale, we present a semi-automated, rules-based image analysis pipeline that leverages unsupervised machine learning, using hierarchical clustering to streamline data processing with minimal parameter tuning to significantly reduce the manual annotation burden, making the processing of regular and large-scale wildlife surveys more feasible.

In this study, we apply such methods to introduce and evaluate a semi-automated image processing pipeline. We then review previous seal survey efforts on the WAP, highlighting methodological inconsistencies that have hindered long-term comparisons. Finally, we argue for standardized survey design and concurrent tagging efforts as essential components of a sustainable monitoring strategy. Together, these steps offer a practical framework for tracking population trends in a dynamic polar ecosystem.

Materials and Methods

Data collection and aerial photography

In 2023, dedicated surveys were flown with a De Havilland Canada DHC-6 Series 300 Twin-Otter aircraft operated by the British Antarctic Survey out of Rothera Research Station (67.6° S 68.1° W, Fig. 2). These flights took place in mid-November at the end of the breeding season and targeted pack-ice habitat typically associated with crabeater seals, with the chosen embayments anecdotally known for high densities of this species (Forcada et al., 2012). However, 2023 was a year of record sea ice lows, with a maximum extent 10% below the average and a shortening of the sea ice season by 3–4 months in most regions (Clem et al., 2024). It is unknown how this sea ice volatility may have impacted ice seals in the surveyed region, as reference baselines are limited.

Data were collected using a Phase One iXU 150 medium format aerial camera (16-bit 3-band, RGB), with a sensor resolution of 50 megapixels and a 55-mm Schneider lens. The surveys were undertaken at a target altitude of 500 m, aiming to achieve a ground sampling distance (GSD) of 4.5 cm and a swath width of 400 m. A total of 1041.6 km was flown over Crystal Sound and Marguerite Bay on the 18 November 2023 (Fig. 2). The Phase One iXU 150 was configured to capture an image every second, producing a forward overlap of around 80%. The camera system lacked an integrated IMU (inertial measurement unit), having only a GNSS (Global Navigation Satellite System) receiver to record the position of the camera in space. Without an IMU, collecting precise orientation information of the system during image capture was not possible. Instead, the average heading for each flight line was calculated and used to correct the orientation of the imagery. Using Socet GXP, each
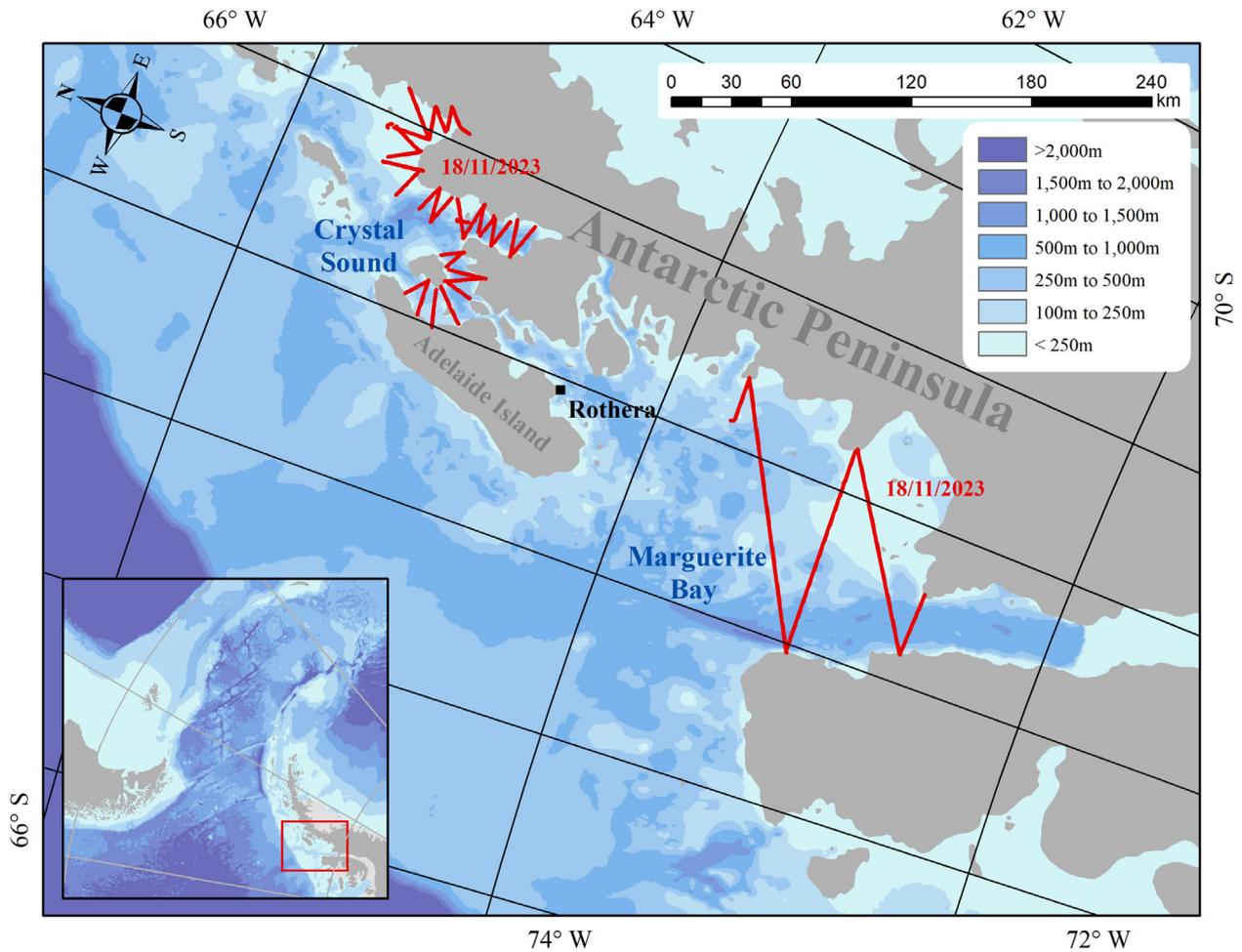
Figure 2. Surveys flown on the 18 November 2023 over Crystal Sound (560.2 km) and Marguerite Bay (481.4 km). Flights were conducted using a twin-otter aircraft equipped with a Phase One iXU 150 medium format aerial camera (3-band, RGB). These regions were selected based on the typical prevalence of pack-ice and the known association of the target species with this habitat.

image was then orthorectified against a flat surface with an elevation of 0 m and exported as a georeferenced TIF, using a geographic coordinate system (EPSG:4326—WGS 84). A subset of the orthorectified image tiles spanning $\sim 350 \text{ km}^2$ of ice were then passed to a semi-automated pipeline, described below, aimed to improve the efficiency of this process.

Seal detection

Seals in each image tile were identified by applying an adaptation of a pipeline developed to process aerial magnetic surveys (McGinnity et al., 2024). This process leverages real-world knowledge and unsupervised machine learning to drastically reduce human review time while also improving recall by significantly mitigating the imbalance problem. We apply this new method to process > 5500 images from the surveys flown on 18 November 2023.

Of the three main species of ice seals on the WAP, each of them typically grow to lengths in excess of 2 m, whether that be crabeater (2.05–2.4 m; Bengston, 2009; Laws et al., 2003), Weddell (*Leptonychotes weddellii*, 2.5–3.3 m; Thomas & Terhune, 2009) or leopard (*Hydrurga leptonyx*, 3.3–3.8; Rogers, 2009). Their dark pigmentation contrasts with the largely homogeneous Antarctic environment, making these species easily identifiable in aerial imagery. Initial experiments determined that the contrast between the seals and ice was most significant in the blue channel of the RGB imagery, with their low saturation in the Hue Saturation Value (HSV) decomposition aiding their differentiation from shadows cast by geological features. This is intuitive given the colour profile of the Antarctic environment.

Our method utilizes these features as domain priors—empirically derived knowledge of seal characteristics—to guide unsupervised detection. These are not Bayesian

priors in a probabilistic sense, but rather heuristic constraints based on observable properties of seals in aerial imagery. We produce a model with just two tuneable parameters: (i) a darkness threshold on the blue channel of the RGB image and (ii) a threshold on the saturation channel of the HSV decomposition. Models with fewer parameters are generally less prone to overfitting, which can improve generalizability to new datasets (Montesinos López et al., 2022). Our approach therefore drastically reduces the necessary amount of training data, mitigates the risk of overfitting and has improved interpretability compared with a deep learning model.

Our model's main stages (Fig. 3) include data preprocessing, contouring, HSV conversion, filtering and clustering. The original image is blurred to reduce noise (Gonzales & Woods, 2018), a common preprocessing step in image analysis to reduce noise and improve stability of subsequent operations (Gonzales & Woods, 2018). In the contouring stage, the model removes areas that satisfy the darkness threshold but are obviously too small or large to be a crabeater seal. Areas smaller than 1 m in both dimensions or larger than 5 m in either are discarded. This eliminates large bodies of water and large portions of brash ice. To further eliminate irrelevant dark areas that meet the shape criteria but are not seals, we filter out regions of high saturation, typical of shadows cast on the ice by terrain.

Hierarchical clustering is then performed on the remaining pixels that satisfy the darkness threshold, using the single linkage method with distance metric. Clusters made up of less than 20 pixels (<0.05 m^2) are immediately discarded as noise and the remaining are processed into a series of clipped images, centred on potential seals, ready for human review.

Figure 3 shows examples of the output of the automatic pipeline. Alongside the close-up image, the reviewer is presented with a wider view to employ context clues, such as the presence of tracks or ice holes, when performing the final classification. This also acts to catch and remove duplicated seals present due to the forward overlap in the aerial imagery. Even after tuning our parameters, we encounter a set of images, typically of complex fields of brash ice, that our algorithm fails to handle satisfactorily, flagging an excess number of potential seals. We subject these outlier images to traditional manual review, removing the top 5% of images ranked by total number of proposed seals.

Model tuning

In an effort to estimate a ground truth to enable evaluation of model performance, $\sim 1/3$ of the images were first manually reviewed by a human expert. The unaided reviewer identified 125 seals across ~ 1600 images, with a review time in excess of 25 h. Additionally, a second expert performed a subset review of ~ 800 of these images, aided by a coarse filter and detected 63 additional seals missed in the initial review. The coarse filter was applied with conservative parameter settings, generating over 30 000 clipped images from these 800 full-size images. This step supplemented the unaided manual review to derive an improved ground truth estimate for evaluating model parameters. Reviewing the clipped images required a similar amount of time as the unaided review.

Clipped images were standardized in size, with two views provided to the reviewer: a zoomed-in square window with side length 3.25 m and a wide-view square window with side length 12.5 m. These dimensions were chosen empirically through trial and error to capture the seals while providing sufficient context.

We took clippings around these 188 seals and supplemented them with representative background images, creating a thinned dataset. We implemented a 10-fold cross-validation procedure, randomly splitting the dataset into 10 subsets of approximately equal size. In each fold, 9 subsets were used for training and 1 subset for testing. All imagery was collected under highly consistent daylight conditions, so a random split is appropriate and alternative partitioning strategies (e.g., temporal splits) would not introduce meaningful variation. It should be noted that these clippings represent an estimated ground truth; some seals may have been missed during manual review or the semi-automated selection process, so reported recall and precision are relative to the detected seals rather than a comprehensive census. We performed a parameter grid search to tune our two parameters, and we computed the recall (Equation 1; proportion of actual positive cases correctly identified by the model) and precision (Equation 2; the proportion of predicted positive cases that are actually correct) of the automatic portion of our pipeline for all combinations of the saturation and darkness thresholds

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

To select the final thresholds, we quantify the trade-off between recall and precision of the candidate values by computing the F_β score (Goutte & Gaussier, 2005) for a range of β (Equation 3)

$$F_\beta = (1 + \beta)^2 \times \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}} \quad (3)$$

The thresholds that maximize the mean F_β score across the 10 folds are optimal for a given β. Here, β controls

INPUT

ALGORITHM STAGE 1 - PROCESSING

Contour

Add HSV Conversion

Filter

ALGORITHM STAGE 2 - CLUSTERING

Output:
(x,y) location of cluster
centroid taken as potential seal

HUMAN REVIEW STAGE

Reviewer presented with full zoom and wider context view to aid final classification

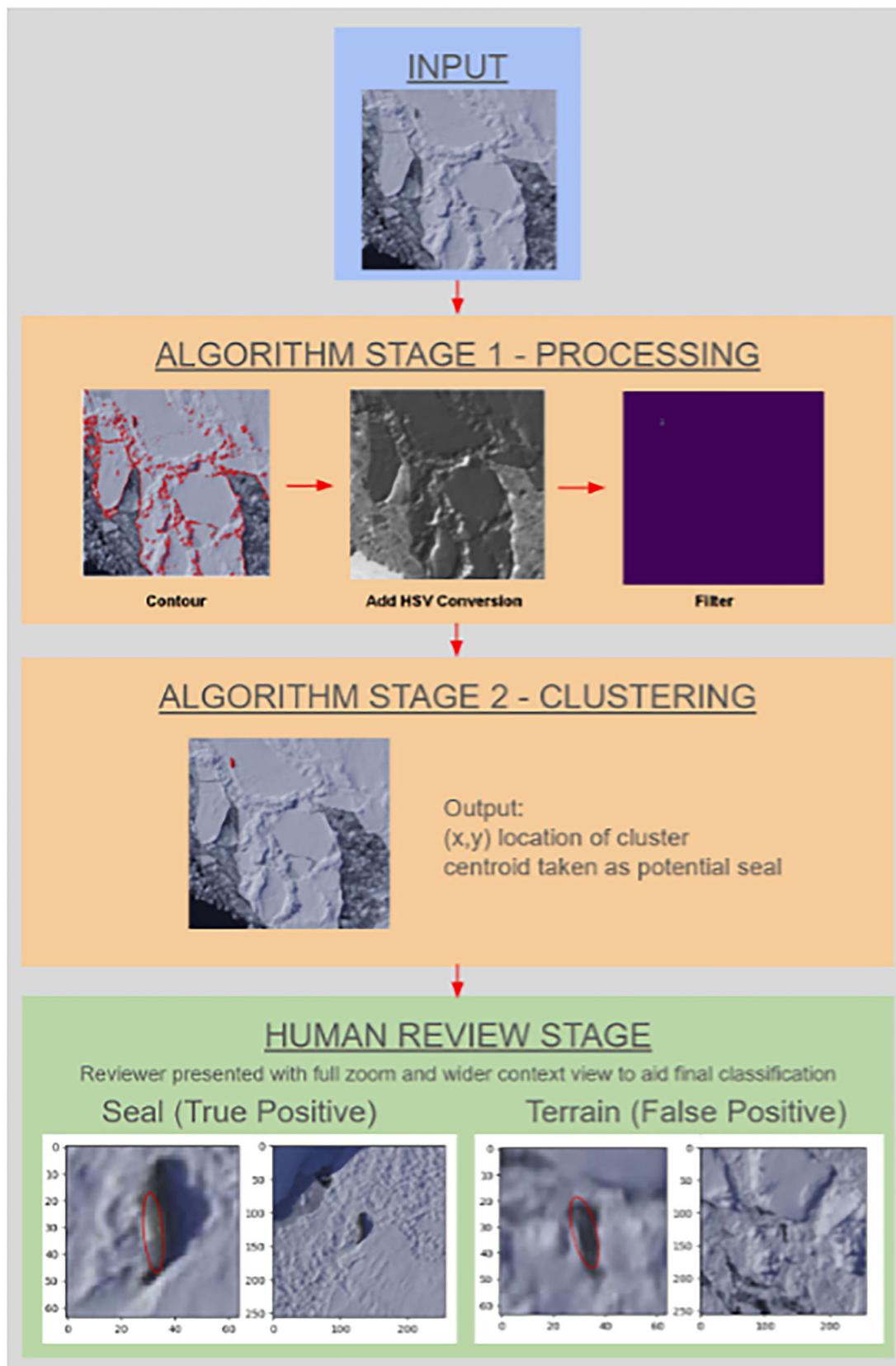Seal (True Positive)

Terrain (False Positive)

Figure 3. The pipeline exploits visual domain priors related to seal size, shape and colour to streamline detection. In the first processing stage of our algorithm, the original data are blurred to reduce noise. Contouring then eliminates large bodies of dark water and other irrelevant dark areas that do not match the seal shape criteria. A HSV decomposition is used to distinguish seals by their low saturation compared with other dark regions, such as shadows. Filter thresholds reflect the selected parameters from the model training. In the second algorithm stage, remaining pixels are grouped into potential seals using hierarchical clustering. The cluster centroids are brought forward to the human review stage. The reviewer is shown cropped candidate seals alongside wider context imagery. This dual-view format aids the human reviewer in distinguishing seals from similar features (e.g., shadows or terrain artefacts). Example outputs from the semi-automated detection pipeline are shown. The left images contain a false positive candidate seal that the reviewer ought to discard. When terrain elements closely resemble seals, the importance of the wider contextual review increases in improving classification accuracy.

the relative weight of recall to precision, with higher values emphasizing recall. Importantly, β is not treated as a tuning parameter in the optimization process. Instead, we compute F_β scores across a range of β values to explore the trade-offs between recall and precision, and to illustrate the resource implications of different operating points. While $\beta = 4$ yields higher recall, it also results in a significantly higher human review burden. β is a design choice dependent on both the desired application of the model and the available resources. Note that reported recall values are specific to the dataset and conditions used in this study and should be interpreted within this intended domain.

Species identification and density estimation

Species identification of the processed image was conducted by an experienced observer who used differences in body shape, length, colouration, presence of aggregations and associated habitat type/feature to differentiate species (Table 1). Only ice-obligate seals were classified to species level, with any ice-tolerant species, namely Antarctic fur (*Arctocephalus gazella*) and southern elephant seals (*Mirounga leonina*), were pooled with those where the clarity of the imagery prevented identification. Ice seal counts were used to estimate density $\widehat{D}$ by dividing the number of observed seals by the ice area surveyed according to Equation (4). The total number of pixels was summed over the ~ 5500 processed images, excluding

pixels corresponding to bodies of water or other non-ice areas. No adjustments were made for forward overlap between images in the density estimation. Instead, the total number of detected seals, divided by the total image area. We acknowledge that the exact surveyed area could not be obtained because IMU/georeferencing data were not available. While this calculation ignores potential overlap and duplicate detections, it provides a reasonable approximation of relative seal density (Casella & Berger, 2024). For a large number of images, the law of large numbers implies that the ratio of total detected seals to total image area converges to the true mean density (Casella & Berger, 2024). Since the primary objective of this study is to develop and validate the detection pipeline rather than to derive precise population estimates, this approximation is adequate for our purposes.

$$\widehat{D} = \frac{\#\text{seals observed}}{\#\text{pixels} \times \text{area per pixel}} \times \frac{1}{\text{model recall}} \quad (4)$$

As with all surveys of a species that are influenced by availability bias (Marsh & Sinclair, 1989), the number of seals observed on the ice represents only a subset of the total number present. Accurate population estimates therefore require correction for the proportion of animals that are not hauled out at the time of the survey. This is typically achieved using bio-logging data to parameterize haul-out probabilities (Forcada et al., 2012). In the absence of tagging data for the season (spring), we do not attempt such extrapolation here.

Table 1. Morphological and contextual characteristics used to distinguish between Antarctic ice seal species (crabeater, Weddell and leopard seals) in aerial imagery.

Species	Crabeater	Weddell	Leopard
Length	2.05–2.4 m	2.5–3.3 m	Up to 3.8 m
Body shape	Slender; pointed snout; obvious scarring	Bulky fusiform shape	Elongated; bulky
Colouration	Medium brown to silver, fading through the year	Dark, almost black with grey streaks; spotted belly	Grey/silver
Aggregations	Solo to large aggregation possible	Small groups	Rare
Habitat cues	—	Proximity to breathing holes	—
Sources	Bengston (2009)	Thomas & Terhune (2009)	Rogers (2009)

Note: Traits include body size and shape, colouration, group behaviour and proximity to habitat features. These distinctions support manual species classification following initial automated detection.

As with all surveys of a species that are influenced by availability bias (Thomas & Terhune, 2009), the number of seals observed on the ice represents only a subset of the total number present. Accurate population estimates therefore require correction for the proportion of animals that are not hauled out at the time of the survey. This is typically achieved using bio-logging data to parameterize haul-out probabilities (Forcada et al., 2012). In the absence of tagging data for the season (spring), we do not attempt such extrapolation here.

Results

Model development

The recall and precision response surfaces for varying thresholds are shown in Figure 4. Our final model achieved a recall of 79% ± 9.1%, surpassing the ∼50% recall from unaided manual review of the 2023 data while also reducing the survey area requiring manual review by ∼94%.

The nonlinear relationship between the threshold parameters (Fig. 4) indicates that relaxing a threshold does not necessarily result in the detection of more seals. The scatter plot of precision and recall for all threshold combinations (Fig. 5) provides insight into the trade-off between these metrics. The final threshold choice will fall along the red frontier, with combinations beneath the curve being strictly suboptimal. The frontier's gradient captures how quickly the recall declines while seeking greater precision. This provides a visual understanding of the effect of changing the value of β. The F-score response surface reveals the optimal threshold choice for a given β as the surface global maximum. Figure 5 shows the response surface when $\beta = 4$.

Table 2 presents the training and test recall for $\beta = 1$, 2, 3 and 4 across the thinned dataset, where a higher β places increasing importance on recall. As a more insightful metric than precision in our semi-automated system with expert review, we present the search area reduction for the human reviewer across the dataset of > 5500 images (as precision was close to 100% for experts reviewing these high-resolution images). The search area consists of the cumulative area covered by the clipped images of potential seals, as well as any complex images flagged for traditional manual review.

The human review stage of this semi-automated pipeline was completed by an expert for the $\beta = 2$ case, processing an equivalent of ∼6% of the total survey area.
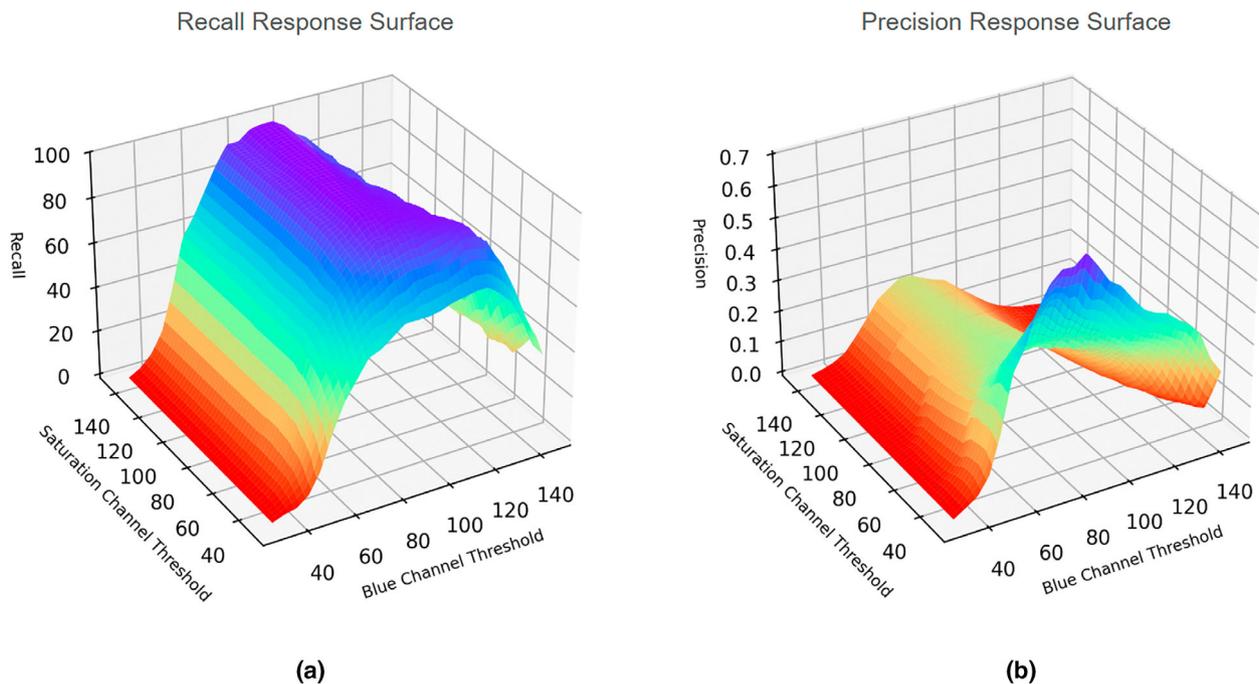

Figure 4. Response surfaces illustrate how model recall (a) and precision (b) vary as a function of threshold combinations applied to the blue channel and saturation channel of the RGB aerial imagery. The irregular shape of these surfaces indicates a non-linear relationship between the parameters. Due to this non-linearity, strictly increasing a threshold is not guaranteed to capture more seals; therefore, thresholds were varied systematically across two dimensions, with a coarser grid in regions of low performance and a denser grid in regions approaching optimality. N.B. Colour-scale only serves as a visual aid.
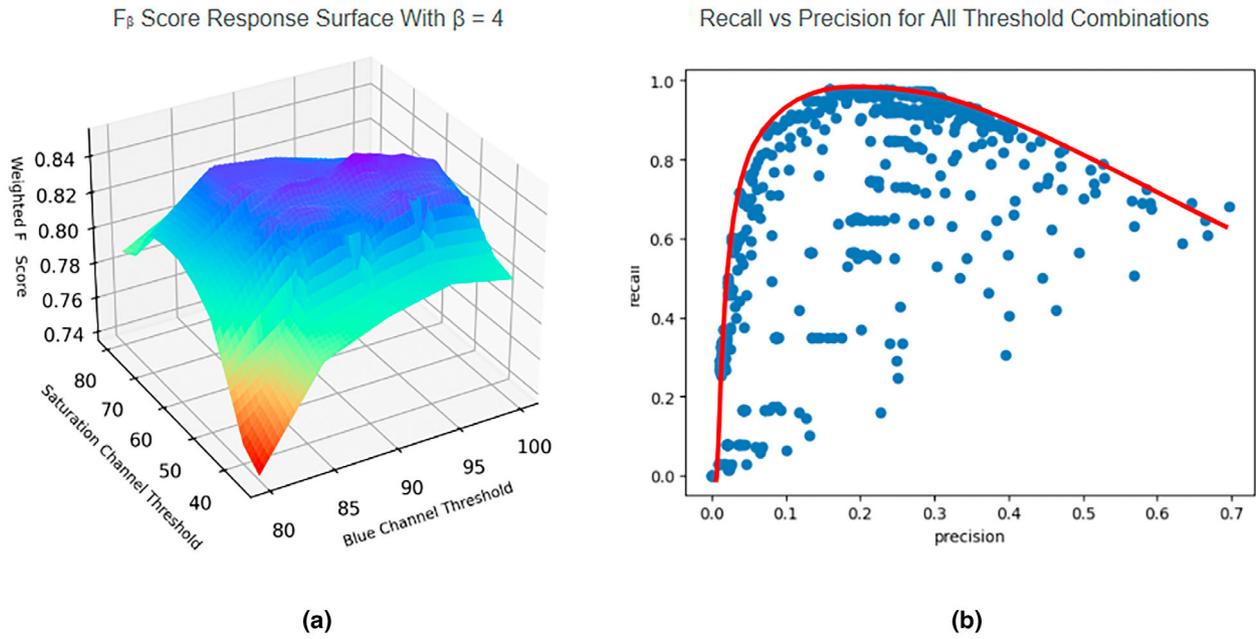
(a)

(b)

Figure 5. The *F*-score response surface (a) illustrates the balance between recall and precision for each model evaluated in the parameter grid search, highlighting the optimal threshold choice for a given β. The *F*-score combines precision and recall into a single value using β as a weighting factor, where $\beta > 1$ gives greater importance to recall. Precision, recall and *F*-score were computed for every combination of our two tuneable parameters, evaluated over our image subset. The subset consists of the 188 images containing seals supplemented with representative background images. The global maximum on the surface in (a) corresponds to the optimal threshold when $\beta = 4$. Similar surfaces can be generated for other β values. While $\beta = 4$ yields higher recall, it also results in a significantly higher human review burden. We do not treat β as an internal tuning parameter in this process; rather, it is a final design choice dependent on both the desired application of the model and the available resources. (b) shows a scatter plot created by calculating the recall and precision for all threshold combinations. This provides an additional decision-making tool, aiding in the selection of the desired β based on constraints such as resource availability and required performance. Moving past the peak of the red frontier, the gradient captures how quickly the recall falls off as you seek greater precision. Threshold choices beneath the frontier are strictly suboptimal.

The combined recall of the automated and human review stages across the 800 manually reviewed images remained 79% $\pm$ 9.1%.

Table 2. Model performance metrics for different β values, showing mean and standard deviation of recall on training and test datasets across 10-fold cross-validation, and the percentage of total image area reduced by the automated detection step.

β	K fold mean test recall (%)	STD (%)	Area reduction (%)
4	94	6.0	90
3	92	7.0	92
2	79	9.1	94
1	64	10.5	95

Note: Higher β values favour recall over precision. Recall and area reduction move inversely: as recall increases, area reduction typically decreases, meaning more image area must be inspected manually (e.g., a 95% area reduction corresponds to only 5% of the images remaining for manual inspection). We selected $\beta = 2$ for the final model to balance detection sensitivity with human review effort across the >5500 images in the 2023 WAP survey dataset. Recall values are calculated for the conditions and dataset used in this study and may not generalize to substantially different survey conditions.

Density estimation

The density estimates by species are summarized in Table 3. We used the presented algorithm to analyse a > 5500 image subset of the imagery captured in 2023 along the track lines in Figure 2. Using Equation (4), we computed a $\widehat{D}$ of 2.61 ice seals per km². We identified a

Table 3. Total seal counts, species-level distribution and estimated on-ice densities derived from the semi-automated analysis of 350 km² of 2023 aerial imagery.

Species	Count	Percentage	Density (seals/km²)
Crabeater	236	31	0.81
Weddell	222	29	0.76
Leopard	12	2	0.04
Unidentifiable	288	38	0.99
Total	758		2.61

Note: Crabeater and Weddell seals made up the majority of identifiable individuals. Densities were estimated using a correction for model recall (Equation 4). Unidentified seals reflect cases where image resolution precluded confident classification.

total of 758 unique seals. Of these 758 seals, manual species identification was carried out by an experienced observer, with the species split between crabeater and Weddell's being approximately even in cases where species could be confidently determined.

Discussion

Semi-automated pipeline

The semi-automated pipeline aims to reduce the human review time while improving recall by mitigating the inherent imbalance in the dataset via search area reduction. The unaided expert reviewer, who conducted traditional manual reviews over entire images, detected 63 seals in 800 images. The second expert, aided by a coarse filter, detected 63 additional seals. These 126 seals form a ground truth estimate, with an unknown number of seals potentially undetected by either method. The true recall for this unaided manual review is at best 50%, which forms the baseline against which we compare our new model performance. The manual review performance is reflective of the imbalance in the dataset (Rožanec et al., 2022) as well as the difficulty in identifying such small objects among a vast background (Fig. 1); a skew similarly reflected in other studies (Forcada et al., 2012; LaRue et al., 2020). Our model's improved recall of $79\% \pm 9.1\%$ stems in part from the re-framing of the question posed to the reviewer into an easier, more balanced task (Rožanec et al., 2022). In traditional review, the reviewer must answer the question 'How many seals are in this large landscape?', a task which combines object localization and object classification; in our method, the reviewer only classifies whether a bite-sized image is a seal or not.

It should be noted that the recall and precision of the model's automated stage are distinct from the recall and precision of the overall method, which includes human review of the clipped images. The human can only discard clipped images. Thus, the overall method recall is less than or equal to the recall in the automated stage, with the difference depending on the human reviewer's false negative rate. Likewise, the human quickly discards false positives served by the algorithm, increasing the method's overall precision. Expert review of the high-resolution clipped images had near perfect recall and precision. However, the decision to perform expert review in the $\beta = 2$ scenario reflects a bottleneck in expert availability to process the downstream clipped images. In future research, we intend to investigate the potential of both crowd sourcing this final review stage and completely replacing the human review with a second classification algorithm (e.g., de Zarzà et al., 2023). We

expect degradation of model performance when replacing expert review with either. However, outsourcing the final review stage to crowds or algorithms facilitates a higher β selection, which could potentially offer improved overall recall than our current $79\% \pm 9.1\%$.

Our model systematically failed to detect seals in areas of dark shadow on the ice, likely due to both the lower contrast between the seal and its background and the elimination of large dark shadows that may contain these seals in the contouring stage. This was a rare occurrence in the labelled dataset with just three instances of seals in shadow. Methods to overcome this blind spot could include the incorporation of adaptive contrast enhancement (Mahmood et al., 2019) or other preprocessing techniques (Chang et al., 2018; Kansal et al., 2018). However, with insufficient data points, there is a higher risk that efforts to recapture these seals introduces overfitting. As more data is gathered, efforts to handle these edge cases can be more robustly developed.

Figure 6 displays a mosaic of seal proposals before the human review stage, with arrows added manually indicating true seal detections. Within these images, those not annotated with arrows show examples of false positives, primarily stemming from geographical features in the ice. Investigating additional methods to remove these features could improve the precision of the automated stage and further reduce the necessary human input. One previous avenue that has been explored was the addition of infrared imagery to aid in the detection of penguins and pinnipeds in overhead drone imagery (Hinke et al., 2022). However, the addition of this data stream was not shown to increase model performance in their test case. Although our developed model will not generalize to other animals, environments, or conditions, the principle of taking a priors-based approach to assist in processing large datasets is sound (Gülçehre & Bengio, 2016). For each new problem, new domain priors must be defined and tuned, but the architecture can be recycled. For example, because the current dataset lacks imagery from different times of day or lighting conditions, we would expect model performance to degrade under twilight or other atypical conditions, and retuning of the relevant parameters would likely be required. Because the method encodes known domain rules, a small set of representative images would likely be sufficient. This makes the approach both computationally efficient and flexible for future applications. Finally, it should be again noted that the ground truth in these datasets can only be estimated, as there is no way to guarantee seals are comprehensively detected. Some seals may have been missed during both manual and semi-automated review, so reported recall and precision values represent best estimates relative to observed detections. Such potential blind spots highlight
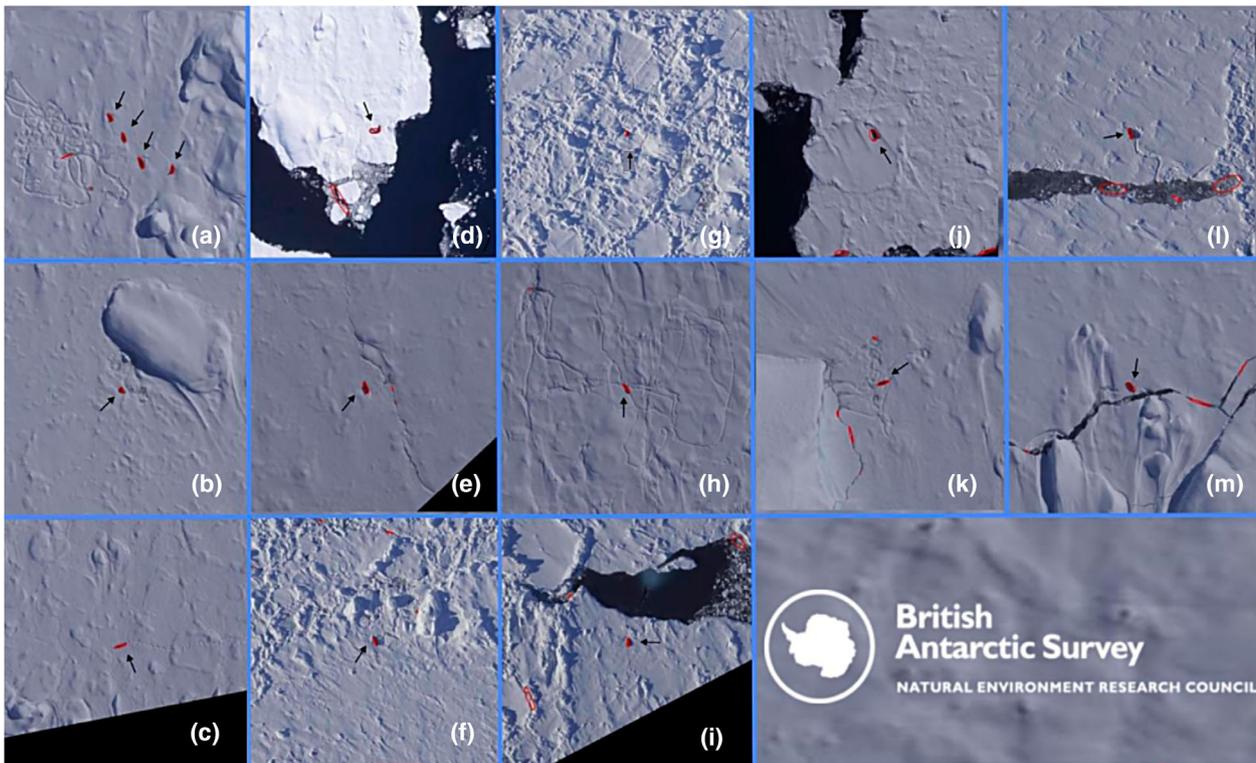
Figure 6. A collection of images (a-m) from the 2023 survey showing typical model outputs. The areas circled in red have been identified by the model as potential seals. All true seal detections are indicated by a black arrow. All remaining model detections are examples of false positives, which must then be discarded by the human reviewer. False positives typically follow terrain features. Developing additional methods to remove terrain features could improve model precision.

the broader challenge of evaluating model performance when ground truth itself is inherently uncertain.

Survey efforts in previous years

In 1999, observer-based aerial surveys were conducted as part of the UK's contribution to the Antarctic Pack Ice Seals (APIS) surveys (Forcada et al., 2012; Southwell, 2012). These surveys were flown between the 22 and 29 January 1999, during the moult period, where seals are hauled out on ice floes. These surveys specifically targeted areas of pack ice and, to date, provide the most comprehensive published assessment of the density and abundance of pack-ice seals on the Antarctic Peninsula. The stated crabeater seal density in the Marguerite Bay area was the highest of the areas observed at 8.428–20.258 seals/km^2, with an abundance estimate of 373 132. The flight lines conducted during these surveys are shown in Figure 7 with red lines. However, the 1999 surveys differ in both scope and methodology from those conducted in 2023. Firstly, the 1999 survey covered a larger set of study areas, with data collected in real time by human

observers with unknown recall and without comprehensive, permanent imagery that can be referred back to. Secondly, their estimates for seal density include corrections for seal haul-out proportions, a factor highly dependent on both environmental and behavioural conditions. In the absence of bio-logging data corresponding to our own survey dates in November, we cannot directly compare recorded densities.

The next most significant aerial survey effort within the Crystal Sound and Marguerite Bay area consisted of 1041.6 km of flight lines in late October/early November 2015; Figure 7, black lines. These data are available on request from the UK Polar Data Centre. In the scope of this study, areas of sea ice (282.7 km^2) in the acquired imagery were manually counted as ice seals by a single observer by loading the imagery into ArcMap v10.3 and placing shapefile points on top of every identified seal. Analysis of this imagery data produced a total of 800 seals, corresponding to a density of 2.83 seals per km^2. Although no bio-logging data was utilized in this survey, we are again unable to compare our density estimates directly without introducing bias. Furthermore, as the
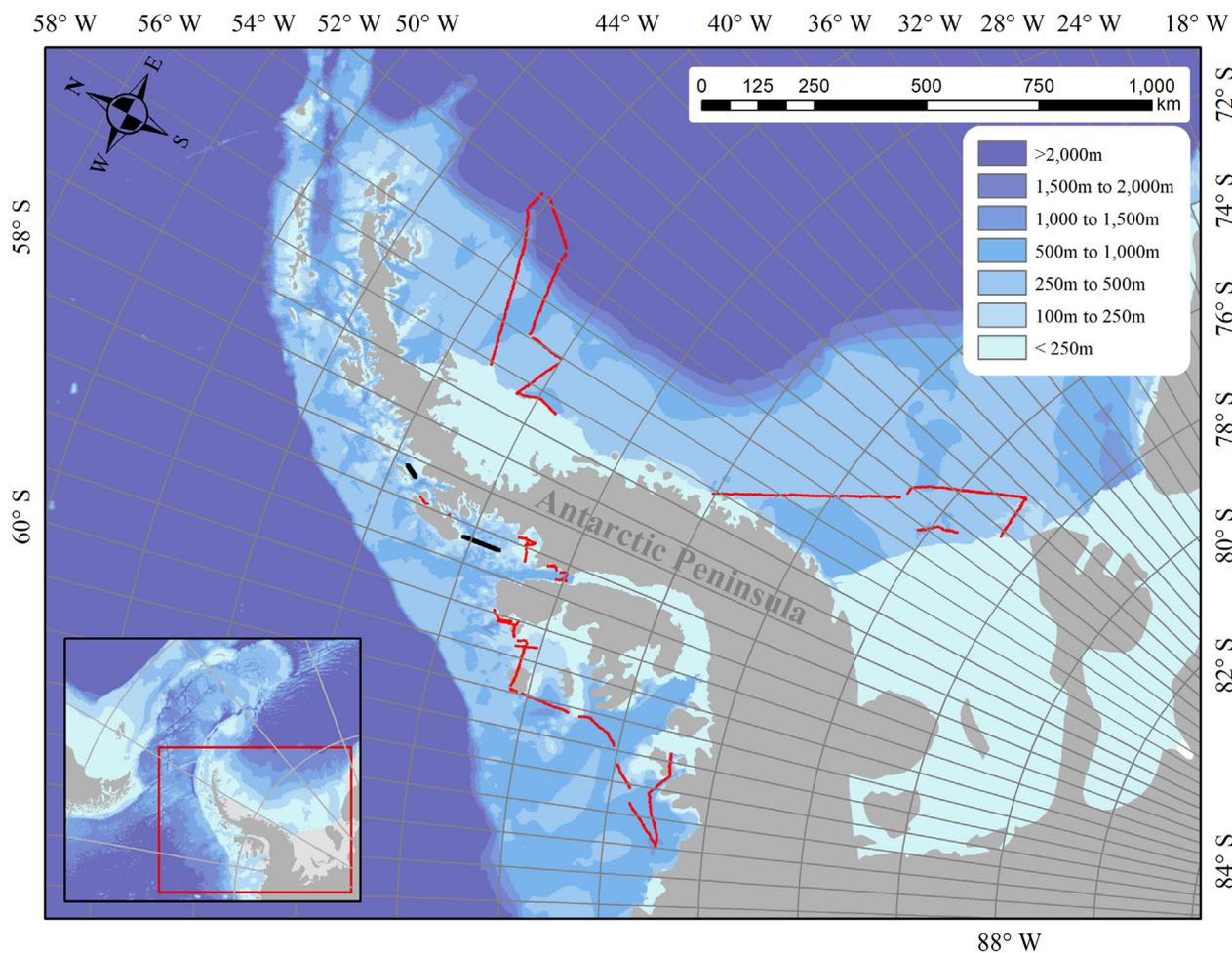
Figure 7. Previous ice seal aerial surveys conducted on the Antarctic Peninsula. Surveys conducted as part of the UK's contribution to the Antarctic Pack Ice Seals (APIS) surveys in 1999 (de Zarzà et al., 2023; Forcada et al., 2012) are depicted in red, with BAS aerial surveys flown in 2015 in black.

imagery collected in 2015 was panchromatic, and therefore it lacked the full spectral range of information needed to apply our semi-automatic pipeline.

Historically, the substantial time and effort required to manually process aerial imagery has limited the scope and regularity of aerial surveys targeting ice seals. As a result, survey efforts have been infrequent, methodologically inconsistent, or opportunistic, which has limited the availability of reference baselines. Table 4 outlines key differences in survey design, and Figure 7 shows the

Table 4. Key methodological differences between the 1999, 2015 and 2023 seal surveys in the Western Antarctic Peninsula.

	1999	2015	2023
Survey method	On-board observer	Panchromatic camera	RGB camera
Time of year	January	Late Oct/early November	Mid November
Concurrent tagging data	Y	N	N
Uncorrected counts available	N	Y	Y
Area calculation	Unknown	Ice only, water excluded	Ice only, water excluded
Suitable for our pipeline	N	N	Y

Note: Variations in timing, imaging platforms, tagging data and image formats affect data comparability across years. Only the 2023 survey used a full-colour RGB camera compatible with the current semi-automated detection pipeline, underscoring the need for standardized future survey protocols. However, the lack of concurrent tagging data prevents extrapolation of on-ice densities into population estimates.

variation in spatial coverage across years. Together, these undermine the ability to detect true trends in seal populations. Overcoming these constraints is essential for generating robust, repeatable population estimates.

Our development of a semi-automated image analysis pipeline addresses this bottleneck, reducing the area requiring manual review by more than 95% and making it feasible to conduct standardized, high-resolution surveys at regular intervals. By enabling consistent data collection, this approach lays the groundwork for meaningful long-term monitoring of ice-obligate seal species in the WAP, or further afield. We strongly recommend the concurrent deployment of satellite-linked bio-logging instruments, where feasible, to complement surveys and enable abundance estimates rather than being limited to haul-out densities alone. Trans-dermal tagging campaigns, which mitigate against tag loss during the annual moult, are essential to quantify haul-out probabilities and correct for detection biases, providing the missing link needed to convert on-ice counts into robust population estimates.

The abnormally low sea ice conditions on the WAP in 2023 (Clem et al., 2024) may have contributed to unusually low observed densities of ice seals. While these trends could indicate future patterns if sea ice remains scarce, they should be interpreted cautiously until verified by further surveys and further sea ice monitoring. The unusually low sea ice across the WAP region may have prompted ice seals to redistribute away from historically abundant areas; yet the presence of dense pack ice in the surveyed areas, which would typically attract seals, suggests the possibility of a more concerning decline.

Conclusion

Monitoring ice seal populations is crucial, as through a more developed understanding of their population dynamics, we can gain a more developed picture of the wider health and functionality of the WAP ecosystem. This is particularly important in light of climate change and increased krill fishery pressure on the WAP (Nicol et al., 2012). However, the processing of increased volumes of primarily imagery-based survey data in line with the fast-paced advances in remote sensing places onerous demands on researchers. Here, we present a semi-automatic model developed to process a > 5500 image subset of the imagery captured in 2023 during our aerial survey campaign over Marguerite Bay and Crystal Sound. By using a priors-based approach and unsupervised machine learning, we sidestep several key barriers for deep learning: (i) insufficient volume of readily available labelled data; (ii) the imbalanced nature of the classification problem; (iii) and low quality ground truth

estimates due to human error exacerbated by data volume and imbalance. We tuned our model using the F-score with $\beta = 2$ and achieved a recall of $79\% \pm 9.1\%$ after expert review of automatically generated image clippings, outperforming the $\sim 50\%$ recall for unaided manual review, while reducing the survey area, and therefore human input, by $\sim 94\%$. In future research, we will evaluate both crowd sourcing and a secondary classifier to replace the expert review, enabling a higher β selection which could further improve the model's overall recall.

While previous aerial surveys have offered valuable snapshots of seal presence along the WAP, methodological inconsistencies have limited their utility for assessing long-term trends. By providing a scalable, reproducible and efficient framework for image-based survey processing, higher-frequency, standardized surveys become feasible. When paired with concurrent tagging campaigns, we can uncover population trends for key indicator species and take a critical step towards more responsive and evidence-based polar conservation policy.

Acknowledgements

Author Contributions

Claire McGinnity: Methodology; formal analysis; writing – original draft; software; investigation; writing – review and editing; validation; visualization. **Connor C.G. Bamford:** Writing – original draft; writing – review and editing; conceptualization; methodology; formal analysis; project administration; validation; visualization. **Nathan Fenney:** Data curation; methodology; resources; writing – original draft; writing – review and editing. **Andrew Fleming:** Writing – review and editing; data curation; supervision. **Jaume Forcada:** Supervision; methodology; writing – review and editing; funding acquisition. **Michael S. Tift:** Writing – review and editing; funding acquisition; project administration. **Luis A. Hückstädt:** Project administration; writing – review and editing; funding acquisition. **Daniel P. Costa:** Funding acquisition; supervision; writing – review and editing. **Peter T. Fretwell:** Writing – review and editing; funding acquisition; data curation; project administration.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. The code accompanying this research is openly available in GitHub.

References

Alshehri, J., Stanojevic, M., Dragut, E. & Obradovic, Z. (2022) On label quality in class imbalance setting—a case study. In: Proceedings—21st IEEE International Conference on Machine Learning and Applications, ICMLA 2022, Nassau, Bahamas. https://doi.org/10.1109/ICMLA55696.2022.00256

Atkinson, A., Hill, S.L., Pakhomov, E.A., Siegel, V., Anadon, R., Chiba, S. et al. (2017) KRILLBASE: a circumpolar database of Antarctic krill and salp numerical densities, 1926–2016. *Earth System Science Data*, **9**(1), 193–210. https://doi.org/10.5194/essd-9-193-2017

Atkinson, A., Hill, S.L., Pakhomov, E.A., Siegel, V., Reiss, C.S., Loeb, V.J. et al. (2019) Krill (*Euphausia superba*) distribution contracts southward during rapid regional warming. *Nature Climate Change*, **9**(2), 142. https://doi.org/10.1038/s41558-018-0370-z

Azaneu, M., Kerr, R., Mata, M.M. & Garcia, C.A.E. (2013) Trends in the deep Southern Ocean (1958–2010): implications for Antarctic bottom water properties and volume export. *Journal of Geophysical Research. Oceans*, **118**, 4213–4227. https://doi.org/10.1002/jgrc.20303

Bengston, J.L. (2009) Crabeater seal Lobodon carcinophaga. In: Perrin, W.F., Wursig, B. & Thewissen, J.G.M. (Eds.) *Encyclopedia of marine mammals*, 2nd edition. Academic Press (Elsevier), San Diego.

Bester, M.N., Lübcker, N., Haddad, W., Bornemann, H. & Wege, M. (2021) Antarctic pack ice seal observations during spring across the Lazarev Sea. *Polar Record*, **57**, e12. https://doi.org/10.1017/S003224742100005X

Bishop, C.M. (2014) Bishop—pattern recognition and machine learning—springer 2006. *Antimicrobial Agents and Chemotherapy*, **58**(12), 738.

Borowicz, A., le, H., Humphries, G., Nehls, G., Höschle, C., Kosarev, V. et al. (2019) Aerial-trained deep learning networks for surveying cetaceans from satellite imagery. *PLoS One*, **14**(10), e0212532. https://doi.org/10.1371/journal.pone.0212532

Burns, J.M., Costa, D.P., Fedak, M.A., Hindell, M.A., Bradshaw, C.J.A., Gales, N.J. et al. (2004) Winter habitat use and foraging behavior of crabeater seals along the Western Antarctic Peninsula. *Deep Sea Research Part II: Topical Studies in Oceanography*, **51**, 17–19. https://doi.org/10.1016/j.dsr2.2004.07.021

Casella, G. & Berger, R.L. (2024) *Statistical inference*, 2nd edition. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9781003456285

Chang, Y., Jung, C., Ke, P., Song, H. & Hwang, J. (2018) Automatic contrast-limited adaptive histogram equalization with dual gamma correction. *IEEE Access*, **6**, 11782–11792. https://doi.org/10.1109/ACCESS.2018.2797872

Charry, B., Tissier, E., Iacozza, J., Marcoux, M. & Watt, C.A. (2021) Mapping Arctic cetaceans from space: a case study for beluga and narwhal. *PLoS One*, **16**(8), e0254380. https://doi.org/10.1371/journal.pone.0254380

Clem, K.R., Raphael, M.N., Adusumilli, S., Amory, C., Baiman, R., Banwell, A.F. et al. (2024) Antarctica and the Southern Ocean. *Bulletin of the American Meteorological Society*, **105**(8), S331–S370. https://doi.org/10.1175/BAMS-D-24-0099.1

Clucas, G.V., Dunn, M.J., Dyke, G., Emslie, S.D., Levy, H., Naveen, R. et al. (2014) A reversal of fortunes: climate change 'winners' and 'losers' in Antarctic Peninsula penguins. *Scientific Reports*, **4**(1), 5024. https://doi.org/10.1038/srep05024

Cook, A.J., Holland, P.R., Meredith, M.P., Murray, T., Luckman, A. & Vaughan, D.G. (2016) Ocean forcing of glacier retreat in the Western Antarctic Peninsula. *Science*, **353**(6296), 283–286. https://doi.org/10.1126/science.aae0017

Cubaynes, H.C. & Fretwell, P.T. (2022) Whales from space dataset, an annotated satellite image dataset of whales for training machine learning models. *Scientific Data*, **9**(1), 245. https://doi.org/10.1038/s41597-022-01377-4

de Zarzà, I., de Curtò, J., Hernández-Orallo, E. & Calafate, C.T. (2023) Cascading and ensemble techniques in deep learning. *Electronics (Basel)*, **12**(15), 3354. Available from: https://www.mdpi.com/2079-9292/12/15/3354

Diamond, R., Sime, L.C., Holmes, C.R. & Schroeder, D. (2024) CMIP6 models rarely simulate Antarctic winter sea-ice anomalies as large as observed in 2023. *Geophysical Research Letters*, **51**(10), e2024GL109265. https://doi.org/10.1029/2024GL109265

Dotto, T.S., Kerr, R., Mata, M.M. & Garcia, C.A.E. (2016) Multidecadal freshening and lightening in the deep waters of the Bransfield Strait, Antarctica. *Journal of Geophysical Research: Oceans*, **121**, 3741–3756. https://doi.org/10.1002/2015jc011228

Forcada, J., Trathan, P.N., Boveng, P.L., Boyd, I.L., Burns, J.M., Costa, D.P. et al. (2012) Responses of Antarctic pack-ice seals to environmental change and increasing krill fishing. *Biological Conservation*, **149**(1), 40–50. https://doi.org/10.1016/j.biocon.2012.02.002

Gilbert, E. & Holmes, C. (2024) 2023's Antarctic sea ice extent is the lowest on record. *Weather*, **79**(2), 46–51. https://doi.org/10.1002/wea.4518

Gonçalves, B.C., Wethington, M. & Lynch, H.J. (2022) SealNet 2.0: human-level fully-automated pack-ice seal detection in very-high-resolution satellite imagery with CNN model ensembles. *Remote Sensing*, **14**(22), 5655. https://doi.org/10.3390/rs14225655

Gonzales, R.C. & Woods, R.E. (2018) *Digital image processing*, Vol. **1**, 4th edition. Pearson, New York.

Goutte, C. & Gaussier, E. (2005) A probabilistic interpretation of precision, recall and F-Score, with implication for evaluation. *Lecture Notes in Computer Science*, **3408**, 345–359. https://doi.org/10.1007/978-3-540-31865-1_25

Gülçehre, Ç. & Bengio, Y. (2016) Knowledge matters: importance of prior information for optimization. *Journal of Machine Learning Research*, **17**, 1–32.

Hasanin, T., Khoshgoftaar, T.M., Leevy, J.L. & Bauder, R.A. (2019) Severely imbalanced Big Data challenges: investigating data sampling approaches. *Journal of Big Data*, **6**(1), 107. https://doi.org/10.1186/s40537-019-0274-4

Hazen, E.L., Abrahms, B., Brodie, S., Carroll, G., Jacox, M.G., Savoca, M.S. et al. (2019) Marine top predators as climate and ecosystem sentinels. *Frontiers in Ecology and the Environment*, **17**, 565–574. https://doi.org/10.1002/fee.2125

Heaton, J. (2018) Ian Goodfellow, Yoshua Bengio, and Aaron Courville: deep learning. *Genetic Programming and Evolvable Machines*, **9**(1–2), 305–307. https://doi.org/10.1007/s10710-017-9314-z

Henley, S.F., Schofield, O.M., Hendry, K.R., Schloss, I.R., Steinberg, D.K., Moffat, C. et al. (2019) Variability and change in the west Antarctic Peninsula marine system: research priorities and opportunities. *Progress in Oceanography*, **173**, 208–237. https://doi.org/10.1016/j.pocean.2019.03.003

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., et al. (2017) Deep learning scaling is predictable, empirically. *arXiv:1712.00409*. Available from: http://arxiv.org/abs/1712.00409

Hewitt, R. & Lepsky, J.D. (2009) Krill and other plankton. In: Perrin, W.F., Wursig, B. & Thewissen, J.G.M. (Eds.) *Encyclopedia of marine mammals*, 2nd edition. Academic Press (Elsevier), San Diego.

Hill, S.L., Murphy, E.J., Reid, K., Trathan, P.N. & Constable, A.J. (2006) Modelling Southern Ocean ecosystems: krill, the food-web, and the impacts of harvesting. *Biological Reviews*, **81**(4), 581–608. https://doi.org/10.1017/S1464793106007123

Hinke, J.T., Giuseffi, L.M., Hermanson, V.R., Woodman, S.M. & Krause, D.J. (2022) Evaluating thermal and color sensors for automating detection of penguins and pinnipeds in images collected with an unoccupied aerial system. *Drones*, **6**(9), 255. https://doi.org/10.3390/drones6090255

Hückstädt, L.A., Burns, J.M., Koch, P.L., McDonald, B.I., Crocker, D.E. & Costa, D.P. (2012) Diet of a specialist in a changing environment: the crabeater seal along the western Antarctic Peninsula. *Marine Ecology Progress Series*, **455**, 287–301. https://doi.org/10.3354/meps09601

Hückstädt, L.A., Piñones, A., Palacios, D.M., McDonald, B.I., Dinniman, M.S., Hofmann, E.E. et al. (2020) Projected shifts in the foraging habitat of crabeater seals along the Antarctic Peninsula. *Nature Climate Change*, **10**(5), 472–477. https://doi.org/10.1038/s41558-020-0745-9

Ionita, M. (2024) Large-scale drivers of the exceptionally low winter Antarctic sea ice extent in 2023. *Frontiers in Earth Science*, **12**, 1333706. https://doi.org/10.3389/feart.2024.1333706

Kansal, S., Purwar, S. & Tripathi, R.K. (2018) Image contrast enhancement using unsharp masking and histogram equalization. *Multimedia Tools and Applications*, **77**(20), 26919–26938. https://doi.org/10.1007/s11042-018-5894-8

LaRue, M., Salas, L., Nur, N., Ainley, D., Stammerjohn, S., Pennycook, J. et al. (2021) Insights from the first global population estimate of Weddell seals in Antarctica. *Science Advances*, **7**(39), eabh3674. https://doi.org/10.1126/sciadv.abh3674

LaRue, M.A., Ainley, D.G., Pennycook, J., Stamatiou, K., Salas, L., Nur, N. et al. (2020) Engaging 'the crowd' in remote sensing to learn about habitat affinity of the Weddell seal in Antarctica. *Remote Sensing in Ecology and Conservation*, **6**(1), 70–78. https://doi.org/10.1002/rse2.124

Laws, R.M. (1977) Seals and whales of the Southern Ocean. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, **279**, 81–96.

Laws, R.M., Baird, A. & Bryden, M.M. (2003) Size and growth of the crabeater seal Lobodon carcinophagus (Mammalia: Carnivora). *Journal of Zoology*, **259**(1), 103–108. https://doi.org/10.1017/S0952836902003072

Mahmood, A., Khan, S.A., Hussain, S. & Almaghayreh, E.M. (2019) An adaptive image contrast enhancement technique for low-contrast images. *IEEE Access*, **7**, 161584–161593. https://doi.org/10.1109/ACCESS.2019.2951468

Marsh, H. & Sinclair, D.F. (1989) Correcting for visibility bias in strip transect aerial surveys of aquatic fauna. *Journal of Wildlife Management*, **53**, 1017–1024. https://doi.org/10.2307/3809604

McGinnity, C., Kolster, M.E. & Døssing, A. (2024) Towards automated target picking in scalar magnetic unexploded ordnance surveys: an unsupervised machine learning approach for defining inversion priors. *Remote Sensing*, **16**(3), 507. https://doi.org/10.3390/rs16030507

Meredith, M.P. & King, J.C. (2005) Rapid climate change in the ocean west of the Antarctic Peninsula during the second half of the 20th century. *Geophysical Research Letters*, **32**(19), L19604. https://doi.org/10.1029/2005GL024042

Mnih, V. & Hinton, G. (2012) Learning to label aerial images from noisy data. In: Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland.

Moffat, C. & Meredith, M. (2018) Shelf–ocean exchange and hydrography west of the Antarctic Peninsula: a review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **376**(2122), 20170164. https://doi.org/10.1098/rsta.2017.0164

Montesinos López, O.A., Montesinos López, A. & Crossa, J. (2022) Overfitting, model tuning, and evaluation of prediction performance. In: *Multivariate statistical machine learning methods for genomic prediction*. Springer, Cham. https://doi.org/10.1007/978-3-030-89010-0_4

Nicol, S., Foster, J. & Kawaguchi, S. (2012) The fishery for Antarctic krill–recent developments. *Fish and Fisheries*, **13**(1), 30–40. https://doi.org/10.1111/j.1467-2979.2011.00406.x

Parkinson, C.L. (2019) A 40-y record reveals gradual Antarctic sea ice increases followed by decreases at rates far exceeding the rates seen in the Arctic. *Proceedings of the National Academy of Sciences*, **116**(29), 14414–14423. https://doi.org/10.1073/pnas.1906556116

Parkinson, C.L. & Cavalieri, D.J. (2012) Antarctic sea ice variability and trends, 1979–2010. *The Cryosphere*, **6**(4), 871–880. https://doi.org/10.5194/tc-6-871-2012

Perry, G.L.W., Seidl, R., Bellvé, A.M. & Rammer, W. (2022) An outlook for deep learning in ecosystem science. *Ecosystems*, **25**(8), 1700–1718. https://doi.org/10.1007/s10021-022-00789-y

Rogers, T.L. (2009) Leopard seal (*Hydrurga leptonyx*). In: Perrin, W.F., Wursig, B. & Thewissen, J.G.M. (Eds.) *Encyclopedia of marine mammals*, 2nd edition. Academic Press (Elsevier), San Diego.

Rožanec, J.M., Križnar, K., Montini, E., Cutrona, V., Koehorst, E., Fortuna, B. et al. (2023) Predicting operators' fatigue in a human in the artificial intelligence loop for defect detection in manufacturing. *IFAC-PapersOnLine*, **56**(2), 7609–7614. https://doi.org/10.1016/J.IFACOL.2023.10.1157

Rožanec, J.M., Zajec, P., Keizer, J., Trajkova, E., Fortuna, B., Brecelj, B., et al. (2022) Enhancing manual revision in manufacturing with AI-based defect hints.

Schofield, O., Saba, G., Coleman, K., Carvalho, F., Couto, N., Ducklow, H. et al. (2017) Decadal variability in coastal phytoplankton community composition in a changing West Antarctic Peninsula. *Deep-Sea Research. Part I, Oceanographic Research Papers*, **124**, 42–54. https://doi.org/10.1016/j.dsr.2017.04.014

Siegert, M., Atkinson, A., Banwell, A., Brandon, M., Convey, P., Davies, B. et al. (2019) The Antarctic peninsula under a 1.5°C global warming scenario. *Frontiers in Environmental Science*, **7**, 102. https://doi.org/10.3389/fenvs.2019.00102

Smith, K.L. & Polvani, L.M. (2017) Spatial patterns of recent Antarctic surface temperature trends and the importance of natural variability: lessons from multiple reconstructions and the CMIP5 models. *Climate Dynamics*, **48**(7), 2653–2670. https://doi.org/10.1007/s00382-016-3230-4

Southwell, C. (2012) A review of data on abundance, trends in abundance, habitat use and diet of ice-breeding seals in the Southern Ocean.

Stapleton, S., LaRue, M., Lecomte, N., Atkinson, S., Garshelis, D., Porter, C. et al. (2014) Polar bears from space: assessing satellite imagery as a tool to track arctic wildlife. *PLoS One*, **9**(7), e101513. https://doi.org/10.1371/journal.pone.0101513

Thomas, J.A. & Terhune, J. (2009) Weddell seal. In: Perrin, W.F., Wursig, B. & Thewissen, J.G.M. (Eds.) *Encyclopedia of marine mammals*, 2nd edition. Academic Press (Elsevier), San Diego.

Torney, C.J., Lloyd-Jones, D.J., Chevallier, M., Moyer, D.C., Maliti, H.T., Mwita, M. et al. (2019) A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, **10**(6), 779–787. https://doi.org/10.1111/2041-210X.13165

Trivelpiece, W.Z., Hinke, J.T., Miller, A.K., Reiss, C.S., Trivelpiece, S.G. & Watters, G.M. (2011) Variability in krill biomass links harvesting and climate warming to penguin population changes in Antarctica. *Proceedings of the National Academy of Sciences*, **108**(18), 7625–7628. https://doi.org/10.1073/pnas.1016560108

Turner, J., Colwell, S.R., Marshall, G.J., Lachlan-Cope, T.A., Carleton, A.M., Jones, P.D. et al. (2005) Antarctic climate change during the last 50 years. *International Journal of Climatology*, **25**(3), 279–294. https://doi.org/10.1002/joc.1130

Vaughan, D.G., Marshall, G.J., Connolley, W.M., Parkinson, C., Mulvaney, R., Hodgson, D.A. et al. (2003) Recent rapid regional climate warming on the Antarctic Peninsula. *Climatic Change*, **60**(3), 243–274. https://doi.org/10.1023/A:1026021217991

Whang, S.E. & Lee, J.G. (2020) Data collection and quality challenges for deep learning. *Proceedings of the VLDB Endowment*, **13**(12), 3429–3432. https://doi.org/10.14778/3415478.3415562

Wu, Z., Zhang, C., Gu, X., Duporge, I., Hughey, L.F., Stabach, J.A. et al. (2023) Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape. *Nature Communications*, **14**(1), 3072. https://doi.org/10.1038/s41467-023-38901-y

Xu, Z., Wang, T., Skidmore, A.K. & Lamprey, R. (2024) A review of deep learning techniques for detecting animals in aerial and satellite images. *International Journal of Applied Earth Observation and Geoinformation*, **128**, 103732. https://doi.org/10.1016/j.jag.2024.103732

Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. (2021) Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, **64**(3), 3072. https://doi.org/10.1145/3446776

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix Table. Model recall performance for each β value on both train and test sets across 10-fold cross-validation. The means and standard deviations across the 10 folds was summarized above in Table 2. Recall that as β increases, so too does the relative importance of recall as compared to precision. Train recall thus increases with the trade-off that area reduction is decreased. We find that performance is largely consistent between train and test sets across folds, especially for higher values of β.