

ICES ANNUAL SCIENCE CONFERENCE
15-18 September 2026
Theme session G: Rethinking marine AI

Contents

Lessons on Limitations and Failure from Five Years of AI Open Hackathons (2020–2025)	2
When AI Meets the Ocean: Limitations and Lessons from Marine Image-Based Applications	2
Nephrops norvegicus population assessment through computer vision: (few) successes and (many) failures	3
Neph-flix: Practical Challenges in Developing Automated Nephrops Burrow Detection using Machine Vision	4
AI-assisted benthic image annotation in Mediterranean rocky reefs: opportunities and limitations	4
Estimating Sponges' Density in Deep-Sea Videos Using Automation	5
AI is a tool, not a sledgehammer – the challenges of universally applying AI and how to improve data transparency	6
From AIS pings to behaviour: Practical challenges for marine AI	6
Three limitations of an AI-assisted remote electronic monitoring (AI-REM) program for on-shore scientific sampling of fishing landings	7
Emblematic but challenging to work with: the case study of marine mammals' bycatch detection in the large-scale Electronic Monitoring program OBSCAMe+	8
Confronting the Illusion of Accuracy: Overcoming Spatiotemporal Autocorrelation to Operationalize Non-Extractive AI Monitoring in Fisheries.....	9
Bridging the Gap between Unstructured Marine Data and LLM Reliability using Agentic RAG	9
Evaluating limitations and operational challenges of deep learning-based analysis of annual growth increments on otolith images	10
A multi-perspective deep learning approach to ageing hoki (Macruronus novaezelandiae) from otolith images	11
Operationalizing AI for Salmon Age Estimation: Where Accuracy Is Not Enough	11
Evaluating how deep learning uncertainty propagates into stock structure, spatial composition, and growth estimates in Baltic herring.....	12
AI-assisted detection of maladaptation in Social-Ecological Systems: methodological challenges and Critical Reflections	13

Rethinking marine data architecture for the AI era: A framework for readiness without reinvention	13
Monitoring forced labor in fisheries with positive-unlabeled learning	14

Lessons on Limitations and Failure from Five Years of AI Open Hackathons (2020–2025)

Dr. Hassan Moustahfid

Artificial intelligence (AI) is increasingly promoted as transformative for marine science, yet the transition from prototype models to operational tools remains challenging. This paper reflects on five years (2020–2025) of AI Open Hackathons designed to accelerate applied AI in marine environmental contexts. Rather than highlighting successes alone, we focus on recurring limitations, failures, and barriers to adoption revealed through these intensive, multidisciplinary collaborations.

Across diverse case studies, models that performed well in controlled settings often failed to generalize to heterogeneous, real-world datasets. Persistent technical challenges included sparse observations, inconsistent metadata standards (i.e. AI data readiness), limited interoperability, and fragile preprocessing pipelines. Even with access to high-performance computing, scalability and reproducibility issues constrained deployment.

Organizational and cultural factors proved equally significant. Mistrust of opaque models, misalignment between developers and domain experts, limited post-hackathon ownership, and regulatory constraints frequently prevented uptake in decision-making contexts. Legal and ethical uncertainties around data access, transparency, and accountability further complicated implementation.

By foregrounding negative results and stalled transitions, this study identifies systemic gaps in data readiness, governance, and model interpretability. We propose a context-aware framework emphasizing co-design, transparent validation, and sustained collaboration to support robust, trustworthy, and operationally relevant marine AI.

When AI Meets the Ocean: Limitations and Lessons from Marine Image-Based Applications

Dimitris Politikos

Artificial intelligence, and machine learning in particular, are increasingly promoted as transformative tools for automating marine monitoring and assessment workflows. Despite promising results, most applications remain confined to pilot studies; their limitations are often underreported, and transition to operational deployment remains rare. This contribution examines barriers to deploying AI solutions in marine science, drawing on lessons learned from four imagery-driven use cases: fish age estimation from otoliths; vulnerable species detection in bottom-trawl catches, oyster habitat detection and population assessment from underwater videos; and marine litter detection from underwater surveys.

Across these applications, recurring technical and data-related constraints are identified, including limited spatial and temporal coverage, small and highly imbalanced training datasets, data quality issues, and a lack of standardized data collection protocols aligned with AI requirements. These constraints are compounded by fragmented data ownership and restricted data sharing, which hinder the aggregation of sufficiently large and diverse datasets required to develop robust and transferable models. As a consequence, models that perform well under controlled or localized conditions are rarely exposed to the ecological complexity

and operational variability necessary to support real-world marine applications at larger spatial and temporal scales.

Beyond technical issues, organizational and human factors further constrain operational uptake, including limited collaboration between marine scientists and machine learning experts, mistrust in automated outputs, and insufficient model interpretability to support assessment and management decisions. By explicitly reflecting on negative results and unmet expectations, this presentation argues that transparent reporting of limitations, stronger cross-disciplinary collaboration, and closer alignment between data collection practices and AI requirements are essential for developing trustworthy and operationally relevant AI applications in marine science.

Nephrops norvegicus population assessment through computer vision: (few) successes and (many) failures

Gabriele Pieri

This study proposes an automated methodology to support the assessment of *Nephrops norvegicus* population, a benthic species of major commercial importance in European fisheries. Population abundance is traditionally estimated through towed Underwater Television (UWTV) surveys, where trained analysts manually inspect the collected video footage to count the burrows, a process that is time-consuming and prone to human error. To address these limitations, the approach we propose integrates machine learning and computer vision techniques into the analysis pipeline, using a supervised model, trained on expert-annotated underwater videos, to automatically detect burrows. The method is designed to cope with challenging underwater conditions, including turbidity, uneven illumination, colour distortion, and ambiguous seabed features. Our approach uses data from the central Adriatic Sea, but it is potentially applicable across the species' distribution range, offering a scalable and objective alternative to conventional assessment methods.

Manual counting of *N. norvegicus* burrows, although standardised by ICES through training, reference footage, and agreement metrics such as Lin's CCC, is highly prone to subjective bias. On the other hand, introducing supervised machine learning methods requires large, high-quality expert annotations of burrows to ensure reliable model training and performance.

A custom convolutional neural network was designed to evaluate the automatic detection of *N. norvegicus* burrows using a novel training dataset, specifically prepared to fit this purpose. To reduce human workload, only a subset of the UWTV frames was annotated according to a set of guidelines developed with experts. To address the network requirements for single-label images and large amounts of training data, a derived dataset was obtained by partitioning the images into tiles. To mitigate the scarcity of labelled data, feature extraction and classification were decoupled, with an autoencoder trained to learn a suitable representation of the tiles and a multilayer perceptron trained to classify them.

Experimental results highlight some limitations of the proposed approach. Performance was evaluated at the tile level, although multiple tiles may correspond to the same burrow entrance, suggesting that entrance-level metrics would be more meaningful. Still, comparison experiments with general-purpose models (e.g. YOLOv8n) show that even state-of-the-art deep detectors perform poorly on this task, confirming the complexity of burrow recognition and the need for specialised, domain-adapted architectures. Nonetheless, the presented model should be regarded as a proof-of-concept to validate the annotation strategy rather than a complete detector.

These results are promising and set a path for future work aimed at assessing *N. norvegicus* abundance, including experimenting with different architectures, integrating spatial context, and using more diverse training data.

Neph-flix: Practical Challenges in Developing Automated Nephrops Burrow Detection using Machine Vision

Eva Linehan

Processing hundreds of hours of imagery to count *Nephrops* burrows remains a critical but resource-intensive component to estimating population abundance and density. Annual *Nephrops* Underwater TV (UWTV) surveys conducted by the Marine Institute sample more than 300 stations, each comprising approximately 7,500 images that undergo repeated expert review following extensive reference training. Across multiple reviewers, analysis of a single survey year exceeds 465 hours of expert assessment time. While this expert-driven approach delivers a robust long-term dataset, the workload is substantial. This project aims to develop a YOLOv8 computer vision model to detect burrow entrances and support automated burrow system identification, improving efficiency while preserving scientific integrity.

Significant barriers limit straightforward AI integration. Current UWTV data processing is not designed for machine learning, making the extraction of large, representative training datasets challenging. Burrow identification is inherently subjective, while bounding box annotation introduces inter-expert variability that may propagate into model training and evaluation. To address this, a temporally and spatially representative dataset has been curated, accounting for variation in burrow density, environmental conditions, and counter differences. A CLIP embedding framework supports cluster-based sampling to maximise semantic diversity while maintaining behavioural frequency distributions, with all annotations verified by at least two experts to strengthen ground-truth consistency.

Technical infrastructure presents an additional limitation. AI deployment requires coordinated systems for ingestion, annotation, training, validation and version control. A dedicated internal server hosting JupyterLab, the open source FiftyOne framework and CVAT has been established to enable collaborative workflows, along with formalised data management and retraining procedures to ensure reproducibility.

Operational integration introduces further challenges. Offshore survey conditions impose constraints, including limited computational resources and low-bandwidth connectivity. The model will therefore be deployed locally for real-time inference, with conservative detection thresholds and substantial expert quality control. A manual counting system will remain operational as a parallel safeguard. To maintain continuity in the survey data process, workflow modifications will undergo formal evaluation to preserve traceability and data integrity in line with the ISO-aligned Data Management Quality Framework. This early-stage work aims to operationalise the detection model within an existing survey process while retaining human oversight to support the responsible integration of AI in long-term marine monitoring.

AI-assisted benthic image annotation in Mediterranean rocky reefs: opportunities and limitations

Maria Zotou

AI-assisted annotation of underwater imagery offers a scalable solution for marine ecological monitoring, yet its transferability across software platforms, annotation strategies, and analytical workflows remains insufficiently tested. We evaluated the CoralNet platform—originally developed for coral reefs—using a large benthic image dataset from Mediterranean rocky reefs in the Aegean Sea (Greece), ecosystems strongly affected by overfishing and herbivore-driven macroalgal overgrazing. A key contribution was the migration of approximately 2,150 legacy annotated images from PhotoQuad to CoralNet using a custom conversion script,

enabling automated classification of an existing dataset that previously required labour-intensive manual annotation. This workflow transition demonstrates how historical image archives can be repurposed for cost-efficient, AI-supported monitoring.

Classifiers trained on morphofunctional group annotations (23 labels) were used to estimate the reef-EBQI index, which relies heavily on macroalgal cover as a proxy of ecological status. CoralNet-derived reef-EBQI values showed good agreement (87%) with indices based on fully manual annotations; however, this agreement depended on taxonomic resolution, training dataset composition, and workflow design. Overall classification accuracy remained moderate (67%), with misclassifications concentrated among visually similar groups and rare but ecologically important taxa, such as canopy-forming macroalgae. These patterns reflect class imbalance and label heterogeneity and can propagate to ecological indicators, occasionally depressing reef-EBQI scores when structurally important taxa are under-detected.

Using the same imagery and processing pipeline, we conducted a targeted trial to assess how AI performance varies with annotation strategy and analytical scope. Training datasets were reannotated across taxonomic resolutions, from near species-level identification (>100 labels) to broader morphofunctional groupings, and evaluated for both classification accuracy and their ability to reproduce reef-EBQI estimates of ecological status. Comparisons of reef-EBQI scores derived from different annotation approaches isolate the influence of taxonomic detail, annotation prioritisation, and user assumptions on AI-derived ecological indicators.

By holding data sources constant and using the same software, while varying annotation strategy, this study highlights a critical limitation of current marine AI applications: apparent classifier performance can mask substantial variability in management-relevant indicators when methodological priorities change. The results provide practical guidance for AI-assisted monitoring, emphasising the importance of context-aware training, transparent workflow transitions, and cautious interpretation of AI-derived ecological status metrics.

This work was supported by Biodiversa+, the European Biodiversity Partnership, within the BioBoost+ project.

Estimating Sponges' Density in Deep-Sea Videos Using Automation

Julia Grabiasz

Artificial intelligence (AI) is increasingly used in marine science to automate the analysis of underwater imagery, reducing the time, cost, and subjectivity associated with manual interpretation of large underwater datasets. In deep-sea environments, remotely operated vehicles (ROVs) generate extensive video transects that are essential for monitoring benthic ecosystems. Among these, deep-sea sponge grounds represent ecologically important habitats, yet estimating sponge density from continuous video remains labour-intensive and methodologically challenging. Automated detection and tracking systems offer a scalable approach for processing large volumes of ROV footage; their reliability for producing accurate, area-based density estimates under real survey conditions remains uncertain.

Recent advances in deep learning-based object detection and segmentation have significantly improved the localisation of benthic organisms in underwater imagery. Yet most existing approaches operate at the frame level and do not incorporate temporal continuity across video sequences. Environmental variability, including changes in illumination, turbidity, camera altitude, and complex seafloor morphologies, further reduces detection and tracking stability, leading to fragmented object trajectories and duplicate identifications of individual sponges. Moreover, inconsistencies in manual annotations, particularly when limited to bounding boxes, introduce label uncertainty and hinder model generalisability.

To address these limitations, this study develops an end-to-end pipeline for automated sponge density estimation that integrates segmentation-based labelling, temporal tracking, and spatial

scaling. Rather than relying on bounding-box detection, we introduce a human-in-the-loop segmentation framework that enables precise region-of-interest (ROI) annotation of individual sponges. These ROIs are then tracked across frames using modern multi-object tracking architectures, allowing for unique individual counting and continuity through time. Finally, spatial density estimates are derived by normalising tracked detections using ROV altitude and camera calibration parameters. This integrated pipeline directly links detection, tracking, and spatial metrics, bridging the gap between computer vision outputs and ecological density estimation.

This analysis highlights broader barriers to operational marine AI adoption, including data sparsity, limited temporal integration, insufficient spatial normalisation, lack of methodological standardisation, and limited reporting of uncertainty, using sponge density estimation as a focal example. The proposed framework aims to advance marine AI adoption by addressing key challenges in annotation robustness, temporal integration, and spatial normalisation—ultimately enabling more reliable, uncertainty-aware estimates of benthic organism density from deep-sea video surveys.

AI is a tool, not a sledgehammer – the challenges of universally applying AI and how to improve data transparency

Georgina Vickery

AI in fisheries monitoring is often presented as an immediate solution for improving efficiency and expanding data extraction from underwater video surveys. However, effective AI deployment depends on data quality and relevance to the task. Video from towed or drift cameras, remotely operated vehicles (ROVs) or other equipment existed long before AI, with many time series spanning decades. Therefore these data-rich video transects were optimised for manual processing, resulting in increased biases and uncertainties when coercing AI to extract additional information from these data sources.

This talk explores solutions to retrospectively applying AI to existing data, with discussion on current obstacles to the routine use of AI. The king crab stock assessment by the Institute of Marine Research in Norway was used as case study, whereby AI was applied to forward facing video to extract counts of additional taxa and crab carapace lengths. To address the issues emerging from this processing and facilitate automation of routine surveys, AI models to process acoustic data and imagery from autonomous underwater vehicles (AUVs) were developed. This increase in AI-derived data highlighted the challenge of integrating these species counts and lengths into existing database structures for wider use. Recommendations to facilitate data transparency include developing a systematic identifier, similar to a gear code. As a minimum, this identifier should incorporate model specifications, including detection confidence and evaluation metrics. This would increase interoperability of AI-derived data and facilitate dataset integration.

A final overlooked limitation is the range in methods used to calculate the area surveyed by underwater observation platforms, with the chosen method strongly impacting final density estimates. A standardised approach is presented in Vickery et al., 2026, which uniquely includes the associated uncertainty. Although not specific to AI, this step is fundamental to integrating AI-derived outputs into stock assessment frameworks.

From AIS pings to behaviour: Practical challenges for marine AI

Hjalte Danielsson

As machine learning is explored for fisheries monitoring, practical challenges emerge during system implementation. We share lessons from developing a Data Analysis Engine (DAE) to model vessel behaviour from AIS, VMS and sometimes catch data, with the goal of detecting potentially IUU-related activity.

In an initial prototype phase using AIS data from multiple fleet types within the Icelandic EEZ, two broad challenges became apparent: how to represent vessel behaviour from AIS transmissions, and how to model that behaviour once represented.

AIS data are irregular in timing and cadence, may contain missing fields, and can arrive out of order. Fixed window segmentation initially produced segments with too little information for trajectory features, or segments mixing distinct behaviours. We therefore adopted adaptive windowing to preserve sufficient context while maintaining temporal coherence.

More fundamentally, segmentation required explicit domain assumptions. Translating position reports into behavioural units meant, for example, deciding when a vessel had genuinely changed course rather than momentarily deviated. This required cumulative bearing-change thresholds and rule-based constraints. Segmentation has proven not a mere neutral preprocessing step but a modelling decision that shapes which patterns can later be detected. Once represented, behaviour modelling posed its own challenges. Single-algorithm clustering produced excessive "noise" clusters or failed to capture nuance. A multi-stage strategy emerged as a response to observed behaviour in the data: density-based clustering separated structured transit from less structured fishing activity, and Gaussian mixture modelling refined sub-phases within those behaviours.

Finally, we observed differences between offline development and real-time use. Offline, irregular data can be examined and interpolated in detail. In real-time, however, decisions must be made more quickly and often with less contextual information available at that moment. While correction and imputation remain possible, the constraints are narrower, requiring clearer thresholds and more cautious decision rules as we move from analytical development to live system use.

Our experience highlights three insights: (1) representation and segmentation choices can influence performance as much as algorithm selection; (2) in our case, unsupervised modelling required combining methods to capture different types of behaviour; and (3) AIS-based systems demand domain-informed iteration due to irregular transmissions and context-dependent behavioural definitions. By sharing these lessons, we aim to support more transparent and context-aware marine AI development.

Three limitations of an AI-assisted remote electronic monitoring (AI-REM) program for on-shore scientific sampling of fishing landings

David Espino García

The recent development of AI-based methods has opened new fields of application in marine sciences. One of these is the improvement of onshore scientific sampling programs of fishing landings, complementing traditional programs carried out by scientific observers (TPSO) in mechanized fishing ports where their access is restricted. Auctions are conducted using images projected onto television monitors of boxes of fish moving along conveyor belts. The implementation of a pilot project for AI-assisted remote electronic monitoring (AI-REM) for scientific sampling of landings in mechanized Spanish ports, although still incomplete, has already allowed us to identify some drawbacks that are crucial for future development. The first concern is taxonomic identification. Since the implementation of the European Data Collection Framework (DCF) in 2009, concurrent sampling has been established, which involves sampling all species present in each sampled trip. However, species identification

software requires a significant volume of images, covering a range of resolution levels and visual focus, to train its taxonomic classification algorithms. This means that the currently available software only covers 9% of the species covered by the Spanish TPSO, limiting its current applicability for comprehensive sampling.

The second limitation relates to biometric data. Unlike TPSO biometrics, where specimens are randomly selected upon landing and measured in the same way, biometrics collected by AI-REM must be inferred from specimens deposited in boxes, one on top of the other and often covered with ice flakes. This limits the number of photographed specimens suitable for extracting an accurate total length, so these must be supplemented with total lengths inferred from specific body parts based on species-specific biometric relationships, which further complicates software training.

The third challenge lies in the final integration of data from both sampling programs, TPSO and AI-REM. Even with the successful implementation of an AI-REM program in landing ports, that is, the taxonomic identification of all landed species and the collection of sufficient biometrics per box of fish, it would ultimately be necessary to integrate this data with that provided by TPSO when estimating the national scientific data required by fish stock assessment models. On one hand, the Spanish TPSO design is based on a two-stage sampling design where the primary sampling unit is the "visit" (port-day combination) and the secondary sampling unit is the fishing trip, randomly selected from those available on each visit. On the other hand, the AI-REM system could potentially census of all trips landed every day. This difference in sampling strategy, and the AI-REM's limited biometrics per box of fish, and mixed length types, poses an integration challenge.

Emblematic but challenging to work with: the case study of marine mammals' bycatch detection in the large-scale Electronic Monitoring program OBSCAMe+

Robin Faillettaz

Marine mammals are amongst the most well-known taxa worldwide: whales, dolphins, and seals have a millennial-long history of interactions with humans. In the Bay of Biscay, fishing activities have triggered a dramatic increase in strandings since 2017. A massive Electronic Monitoring (EM) program, OBSCAMe+, was launched in 2020, thoroughly recording all fishing operations of over one hundred vessels as of 2026. Despite their emblematic importance for ocean conservation, collecting and processing the data generated by such a program encountered numerous political, societal, and technical limitations for its implementation. The development of AI-based tools, somehow, significantly eased some aspects of the program thanks to a strict management of GDPR engagements (through the vessels' and fishers' anonymization) and to species-focused video processing that reduced the concerns raised by fishers about surveillance. We will here present the key aspects of the program by focusing on the philosophy that drove the AI developments, designed to balance three critical performance indicators: minimizing false negatives (to ensure no bycatch event nor fisher's face is missed), controlling false positives (to reduce manual review workload), and optimizing processing time (for cost-effective large-scale deployment), along with the main results (i.e. the models' performance for anonymization and bycatch detection) obtained with the processing of these sensitive data. Bound to the data challenges (i.e., heterogeneity in image quality and field of view, GDPR limitations preventing access to raw training data, etc.), we set up procedures to achieve replicability and versioning of the models, evaluation of the robustness and relevance when applied to new data, as well as large-scale deployment (i.e. to process around 10,000 h of video per month) and the—not so obvious—ultimate usefulness of AI versus human processing for this dataset. This case study appears as a great example of the considerable effort that has

to be put to achieve not just automatic AI-based but rather AI-relevant processing of such sensitive and challenging datasets.

Confronting the Illusion of Accuracy: Overcoming Spatiotemporal Autocorrelation to Operationalize Non-Extractive AI Monitoring in Fisheries

Kilian Huss

The integration of advanced computer vision into marine ecology offers a transformative pathway for fisheries science, enabling the transition from conservative relative abundance indices (MaxN) to absolute abundance estimation (Total N) via automated individual re-identification (re-ID). This shift supports non-extractive sampling, providing a viable mechanism to rigorously monitor Marine Protected Areas (MPAs) without the destructive impacts of scientific bottom trawling. The CodSpotter project investigates the feasibility of extracting robust biometric embeddings from spot patterns of Atlantic cod (*Gadus morhua*) using unconstrained Baited Remote Underwater Video systems (BRUVs).

Despite high theoretical potential, operationalizing re-ID in unconstrained marine environments reveals systemic methodological challenges. Standard validation metrics frequently yield artificially inflated accuracies, masking a profound vulnerability to high spatiotemporal autocorrelation. In densely sampled datasets covering limited temporal or spatial scales, networks exploit transient environmental correlates, such as turbidity, illumination, and specific backgrounds as computational shortcuts. Consequently, when subjected to biologically rigorous, "sighting-aware" evaluation protocols that mandate cross-temporal matching, representational robustness degrades significantly.

Addressing this generalizability bottleneck requires shifting away from traditional supervised paradigms toward self-supervised and semi-supervised representation learning. Modern Joint-Embedding Predictive Architectures (e.g., LeJEPa) offer a highly promising alternative by learning invariant, abstract representations of world states rather than relying on heavily curated, domain-specific contrastive pairs.

However, algorithmic refinement cannot unilaterally resolve what is fundamentally a data variance problem. We argue that realizing the promise of AI in fisheries necessitates a paradigm shift in experimental design and data curation. Overcoming the "illusion of accuracy" requires smart sampling strategies that deliberately decouple biological identity from environmental artifacts. This mandates the construction of long time-series datasets to capture natural spatiotemporal variability in individual appearance, alongside widespread cross-institutional collaboration to pool spatiotemporally disjointed data. Only by enforcing strict cross-temporal validation standards and aligning self-supervised architectures with ecologically representative sampling can we engineer reliable AI systems for sustainable fisheries policy.

Bridging the Gap between Unstructured Marine Data and LLM Reliability using Agentic RAG

Adeel Yusuf

Marine data holds immense value for understanding ocean ecosystems, supporting sustainable fisheries, and informing policy, yet much of it exists in unstructured or semi-structured, heterogeneous textual formats that make synthesis and interpretation challenging. While large language models (LLMs) offer new opportunities to unlock this information, concerns around numerical accuracy, provenance, and reproducibility limit their adoption in high-stakes decision environments. Bridging the gap between unstructured marine data and operational reliability therefore requires approaches that combine automated extraction with robust validation. Information contained within ICES advice sheets are used extensively by fisheries management in both devising actions and tracking impacts. Other stakeholders including non-governmental organisations and academics also make use of sheets in compiling reports. Extraction of

information from over one hundred PDF advice sheets is time consuming and risks transcription errors, both of which could be effectively addressed through automation

This study presents an agentic retrieval-augmented generation (RAG) framework designed to support reliable extraction of quantitative indicators from ICES advice on fishing opportunities documents. Using ICES advice as a real-world case study, we created a manually curated dataset covering 143 stocks across six years as ground truth. We then compare two complementary extraction approaches (i) a structured database extraction workflow, which retrieves relevant values from the ICES Advice Scenarios Database, Stock Assessment Graphs, and Stock Information Database via APIs and maps them to the benchmark dataset for evaluation; and (ii) a cloud-deployed agentic RAG workflow that retrieves and processes ICES advice PDFs obtained through the Figshare API. Outputs from both approaches were analysed to assess accuracy, completeness and consistency.

Results show that the structured database retrievals compared well with the baseline but required harmonisation and did not include all indicators needed for integrated analyses. The RAG workflow improves coverage and timeliness by retrieving information directly from source documents. Critically, the agentic architecture allows the system to autonomously plan and sequence multiple retrieval steps, select the most relevant document sections, and self-correct when initial extractions are incomplete or ambiguous. However, ensuring reliability of results requires continuous validation. We therefore position RAG as a decision-support layer that proposes values with traceable evidence, which are then cross-checked against curated benchmarks before acceptance.

The findings demonstrate that hybrid evidence pipelines can substantially improve scalability while maintaining the transparency and accuracy required in marine advisory contexts. This approach also provides a framework for integrating LLMs into other data-intensive ocean governance workflows.

Evaluating limitations and operational challenges of deep learning-based analysis of annual growth increments on otolith images

Mariia Tiurina

Age estimation and biochronological analysis based on otolith images remain key components of stock assessment for commercially important fish species. Manual reading of annual growth increments is time-consuming and requires highly trained experts, making the process both time- and resource-demanding. Consequently, a growing number of neural network-based approaches are being developed, either to fully automate readings or to support experts in a decision-support (“assistant”) mode, which in turn raises the need for robust calibration and evaluation of model predictions.

In this work, we present an approach based on deep learning models for automatic analysis of flatfish otolith images, designed to operate in two modes: (i) fully autonomous prediction of age and increment structure, and (ii) a smart-assistant mode supporting expert readings. Our pipeline follows a stepwise “think like an expert” strategy, sequentially running several models, mostly based on YOLO-family segmentation and pose architectures, dedicated to different tasks: detection of regions of interest, segmentation of growth increments, and extraction of increment widths along a defined reading axis (either defined automatically or specified by an expert in the AI-assistance mode).

We further explored methods to evaluate post-hoc the correctness of biochronological predictions using statistical models derived from expert data. Based on a biochronological dataset of expert-annotated otolith images, we apply a regression-based approach to flag potentially inconsistent model outputs and identify outliers.

Finally, we present comparative results illustrating how the AI assistant influences expert workload, both for age determination and for the production of detailed biochronological datasets. By critically evaluating both the benefits and limitations of the approach, we aim to contribute to a more transparent and context-aware integration of AI-based otolith analysis into operational stock assessment workflows.

A multi-perspective deep learning approach to ageing hoki (*Macruronus novaezelandiae*) from otolith images

Jeremy McKenzie

Age information underpins fisheries assessment and management. In recent years, the use of deep-learning Convolutional Neural Network (CNN) models has been posed as a cost-effective method for fish age estimation and has shown considerable promise for integration into routine ageing practices. Our presentation represents five years research investment by Earth Sciences New Zealand into developing AI otolith reading capability. We have reached the point at which our reading algorithms now achieve the accuracy and precision of the human readers over the common age classes for hoki (*Macruronus novaezelandiae*). ESNZ is in the process of integrating our AI system with its human process to provide second reader capabilities and QA validation for hoki. Over the last five years one of our biggest challenges has been in managing the unrealistic expectations of ESNZ managers and the external funding agencies. The recent hype around AI has reinforced a general belief that, after five years, we should have replaced human reading capability with AI systems for all NZ finfish stocks. In truth, at current investment levels, operationalisation of fully autonomous AI reading capability for one or more New Zealand fish stock is at least 10 years away.

Our AI mythology uses a series of single branch (i.e. single image) and multi branch (i.e., multiple image) CNN models, developed using the EfficientNet and ResNet architectures, to estimate the age of NZ hoki from images of sectioned otoliths. Hoki is assessed annually using an age structured model, with approximately 6000 otoliths aged each year to provide data to inform the assessment. In this study, the CNN models were trained using images of sectioned hoki otoliths from 1014 individual fish, validated using images from 125 fish, and evaluated against a standardised reference set of 480 individuals that is used by human readers. Model performance was significantly increased when a multi-branch model architecture was used, with the best multi-branch models achieving a CV of ~0.08 and precision and values of 0.77–0.79 and 0.83–0.84, respectively. Gradient visualisations (i.e., heat maps) generated using Grad-CAM revealed that the models used several attributes when deriving predictions, including the overall area and shape of the otolith section and its associated surrounding space, as well as patterns in opaque and translucent annual bands. The study also identified several areas for improvement, including poor performance of single-branch models and reduced predictive success for age classes with limited sample sizes across all models.

A key next step in the implementation of CNNs into routine ageing for hoki at a commercial scale is to converge this with the SmartDots software to enable a fully transparent and reproducible age estimation pipeline.

Operationalizing AI for Salmon Age Estimation: Where Accuracy Is Not Enough

Robert Ames

AI models in fisheries can show strong accuracy in controlled research settings; however, moving them into frameworks to affect management remains uncertain. In Phase I of the CHUMputer project, we developed an ensemble of convolutional neural networks to estimate chum salmon (*Oncorhynchus keta*) age from scale images. Using ~15,000 labeled scale images, representing fish originating across the North Pacific rim, the ensemble achieved >92% agreement with expert readers. This level of accuracy is consistent with trained human agers and suggests readiness for incorporation in aging protocols to aid in resource management. However, integrating the sample processing and analytical methodology into an operational ageing lab introduces additional challenges. Phase II, currently underway, is assessing whether this workflow can be integrated into existing systems in a production ageing laboratory supporting regulatory and in-season decisions. During development of the models, we identified limitations that must be addressed prior to operationalization. Rare age class predictions are less reliable because those ages are underrepresented in the training data. Performance varies across image sources, reflecting differences in image quality and scale condition. Automated preprocessing improves overall accuracy but, in some cases, can remove biologically meaningful structure. Ambiguous growth patterns (where experienced readers disagree) highlight that “ground truth” is often uncertain, complicating validation and performance metrics. A more difficult problem than the comprehensive dataset and analytical hurdles; however, is institutional. Fisheries management systems are based on human review, transparency, and legal defensibility. A deterministic but opaque model output, one that returns the same answer every time without explaining why, does not fit neatly into that structure. Operational use requires new workflows, QA/QC standards, auditing approaches, retraining schedules, and a shift in human effort toward oversight and quality assurance. This ongoing case study illustrates that technical readiness does not equate to institutional readiness. Whether these routines and this framework are adopted or remain experimental is unresolved. What is clear is that the outcome will depend less on model accuracy and more on whether they fit within existing management and regulatory structures.

Evaluating how deep learning uncertainty propagates into stock structure, spatial composition, and growth estimates in Baltic herring

Szymon Smoliński

As artificial intelligence becomes increasingly integrated into fisheries science, its performance is often evaluated only at the level of classification accuracy. However, the practical implications of AI-derived uncertainty for downstream biological inference and stock assessment remain poorly understood. Here, we critically evaluate how deep learning-based otolith classification propagates into key biological and spatial estimates used in fisheries science, using Baltic herring (*Clupea harengus*) stock components as a case study. We compared traditional expert-based stock component assignments with assignments derived from a convolutional neural network (ResNet-50) trained on otolith images collected during scientific surveys (2021–2024). Rather than treating AI outputs as definitive labels, we assessed how differences between expert and AI assignments affect subsequent estimates of (i) the spatial proportion of stock components across the Baltic Sea, resulting in contrasting spatial maps and stock contribution functions, (ii) age structure of the northern and southern components, and (iii) von Bertalanffy growth parameters estimated separately for each component. Our results demonstrate that even moderate disagreement between expert and AI-based classifications leads to measurable differences in inferred spatial composition, age distributions, and growth trajectories. These discrepancies highlight how uncertainty and bias introduced at the AI classification stage can propagate nonlinearly into biological parameter estimation, potentially influencing ecological interpretation and management-relevant

conclusions. By explicitly linking AI uncertainty to downstream biological estimates, this study moves beyond proof-of-concept applications and provides a rare, end-to-end evaluation of AI integration into fisheries workflows. Our findings underscore the need for transparent uncertainty handling, potentially hybrid expert–AI approaches, and caution when deploying AI outputs in operational stock assessment and decision-making contexts.

AI-assisted detection of maladaptation in Social-Ecological Systems: methodological challenges and Critical Reflections

Lucy Arnaud

Social-ecological systems are currently on the front lines of the climate crisis. Coastal communities are constantly evolving and adapting to absorb environmental shocks, often resulting in maladaptive outcomes, which increase their long-term vulnerability. Despite its importance, inconsistent terminology and fragmented data on maladaptation hinder large-scale synthesis.

To fill this gap, I employed *ClimateBert*, a transformer-based classification model, to conduct an AI-driven literature review of global empirical evidence of maladaptation. The goal was to investigate not only the outcomes of maladaptive trajectories, but also to identify their mechanisms: ecological traps, equity trade-offs, and time lags where a short-term "benefit" evolves into a long-term "loss".

While AI for literature synthesis faces significant challenges, the term "maladaptation" is a modern term emerging from an ancient human problem, making it difficult to gather relevant literature. Due to inconsistent terminology of maladaptation across scientific fields, the first challenge was to teach the model to recognize "invisible" maladaptation through adjacent concepts like negative feedback loops and unintended consequences of adaptation.

These linguistic barriers make it difficult to train a Natural Language Processing model that is both generalizable and performant to local context. This paper reflects these methodological pitfalls. Arguing the limits of AI to capture human and ecological complexity accurately. The result of this AI review demonstrates the need for hybrid approaches, such as "human-in-the-loop" to both benefit from AI tools and more traditional methods in academia, ensuring that automated syntheses on maladaptation do not overlook the complex, context-dependent failures of climate adaptation.

Rethinking marine data architecture for the AI era: A framework for readiness without reinvention

Nasrullah Iqbal

AI projects in marine science often fail to move beyond pilot stages due to outdated data infrastructure rather than poor model quality. Data systems specifically built for the generation of scientific knowledge were not designed with considerations for use by AI/ML. Based on exploratory work at ICES, we suggest a three-part framework to address this gap.

Part 1 of the framework is an AI Readiness Heatmap, a diagnostic tool that scores different data domains i.e. survey submissions, stock assessments, oceanographic observations against criteria including format standardization, API accessibility and metadata completeness. Rather than treating AI readiness as binary, the heatmap reveals variation across datasets within the institute, enabling AI readiness to be prioritised and targeted.

Part 2 of the framework is the Marine Data Contract, which sets clear requirements for data providers to ensure products are ready for AI use. Basing this on existing ICES data management best practices and FAIR principles, demands documented acquisition, defined data roles, clear

access and licensing under ICES data policy, and quality measured in timeliness, completeness, consistency, accuracy, uniqueness, and validity. These also align with the ICES Data Profiling Tool's evaluations, reframing governance as a foundation for AI readiness rather than creating new obligations. By moving accountability to the point of data sharing, the contract makes AI assumptions transparent and verifiable before they hinder AI related projects.

Finally, for implementation we advocate the Strangler Fig modernisation pattern incrementally layering modern API and data-quality interfaces over existing systems, allowing AI-ready workflows to grow without disrupting established processes. This approach suits organisations where pipeline changes carry governance implications across actors in the pipeline. We present this framework not as a polished success story but as a challenge offered in the spirit of transparent, collective learning.

Monitoring forced labor in fisheries with positive-unlabeled learning

Rocio Joo

The International Labour Organisation has identified forced labor at sea as a persistent problem in the global fishing industry. Identifying vessels that engage in this practice is challenging. Many of these vessels spend long periods of time at sea. When they do go to port, labor and human-rights inspections are rarely performed, and there are few opportunities for crew to speak up and receive protection.

To aid in the identification of fishing vessels with potential forced labor, we developed a model to find vessels with similar behavioral patterns to those with reported forced labor. We performed an extensive global search of reported cases of forced labor in the media, reports from NGOs and others, and identified 132 vessels from 17 flag states that we could match to GPS tracking from the Automatic Identification System (AIS), revealing their detailed movement patterns. Because the start and end dates of forced labor in the reported cases were not always clear, the model was trained to identify forced labor patterns at the scale of a year.

Traditionally, machine learning models use a classic positive-negative approach that uses training data from both positive and negative labeled cases. A challenge with our setting, however, is the lack of negative data, i.e., vessels known to be free of forced labor, to train and evaluate the model. We had access to a list of 66 vessels that were inspected and certified as meeting the criteria in the ILO C188 agreement of good working conditions. However, we could not guarantee that labor abuses were not committed outside the time of inspection. Therefore, we did not use these to train the model, but rather for post-hoc comparison.

To account for the lack of negative training data, we developed a positive-unlabeled learning model; that is, using our list of labeled positive cases (vessels with reported forced labor) alongside cases with an unknown class (all other vessels), we trained a model to predict whether vessels are either positive or negative based on their movement patterns and characteristics.

The model correctly identified 89% of the reported cases of forced labor, while 95% of the certified vessels were correctly classified as negative. From the unknown-class vessel, the model predicted 9,485 vessels as positive (similar patterns to those with forced labor), representing a total of ~261k people at risk of forced labor. Prevalence of forced labor was estimated as very high in the high seas and for vessels fishing in foreign EEZs, often referred to as the distant water fleet, with more than 40% at high risk.

In this presentation, I will talk about the model and decisions behind it, the data ethics analysis for assessing the consequences of releasing the predictions dataset, the difficulties in communicating the potential of the model while being transparent with its limitations, and the complications in rendering the data useful for reducing forced labor in the world.