Fish and Fisheries

WILEY

FISH and FISHERIES

ORIGINAL ARTICLE OPEN ACCESS

‘Building the (Im)perfect Beast’: Strategies for Identifying Appropriate Spatial Stock Assessment Model Complexity From an International, Blinded High-Resolution Simulation Experiment

Aaron M. Berger[1] | Daniel R. Goethel[2] | Simon D. Hoyle[3,4,5] | Patrick Lynch[6] | Caren Barceló[7] | Alistar Dunn[8] | Brian J. Langseth[9] | Carolina Minte-Vera[10] | Jemery Day[11] | Haikun Xu[10] | Francisco Izquierdo[12] | Dan Fu[13] | Nicholas D. Ducharme-Barth[14] | Mathew Vincent[15] | Arnaud Grüss[16] | Jonathan J. Deroba[17] | Giancarlo M. Correa[18] | Jeremy McKenzie[19] | Will Butler[20] | Jie Cao[21] | Craig Marsh[2] | Teresa A'mar[16] | Valerio Bartolino[22] | Massimiliano Cardinale[22] | Claudio Castillo-Jordan[11] | Bjarki Þór Elvarsson[20] | John Hampton[11] | Andrea Havron[6] | Pamela Mace[23] | Arni Magnusson[11] | Mark Maunder[10] | Richard Methot[24] | Sophie Mormede[25] | Maria Grazia Pennino[26] | Alfonso Perez-Rodriguez[27] | Marta Cousido-Rocha[12] | Thomas Teears[11] | Agurtzane Urtizberea[28]

[1]Fishery Resource Analysis and Monitoring Division, Northwest Fisheries Science Center, NMFS-NOAA, Newport, Oregon, USA | [2]Auke Bay Lab, Alaska Fisheries Science Center, National Oceanic and Atmospheric Administration, Juneau, Alaska, USA | [3]National Institute of Water and Atmospheric Research, Nelson, New Zealand | [4]Department of Statistics, University of Auckland, Auckland, New Zealand | [5]Office Hoyle Consulting Ltd, Nelson, New Zealand | [6]Office of Science and Technology, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Silver Spring, Maryland, USA | [7]Puget Sound Institute, University of Washington-Tacoma, Tacoma, Washington, USA | [8]Ocean Environmental Ltd, Wellington, New Zealand | [9]Fishery Resource Analysis and Monitoring Division, Northwest Fisheries Science Center, NMFS-NOAA, Seattle, Washington, USA | [10]Inter-American Tropical Tuna Commission, La Jolla, California, USA | [11]The Pacific Community (SPC), Oceanic Fisheries Programme, Noumea, New Caledonia | [12]Instituto Español de Oceanografía (IEO, CSIC), Centro Oceanográfico de Vigo, Vigo, Spain | [13]Indian Ocean Tuna Commission, Victoria, Seychelles | [14]Fisheries Research and Monitoring Division, Pacific Islands Fisheries Science Center, NMFS-NOAA, Honolulu, Hawaii, USA | [15]Beaufort Laboratory, Southeast Fisheries Science Center, NMFS-NOAA, Beaufort, North Carolina, USA | [16]National Institute of Water and Atmospheric Research, Wellington, New Zealand | [17]Population Dynamics Branch, Northeast Fisheries Science Center, NMFS-NOAA, St. Woods Hole, Massachusetts, USA | [18]School of Aquatic and Fishery Sciences, University of Washington, Seattle, Washington, USA | [19]National Institute of Water and Atmospheric Research, Auckland, New Zealand | [20]Marine and Freshwater Research Institute (MFRI), Reykjavík, Iceland | [21]Department of Applied Ecology, North Carolina State University, Morehead City, North Carolina, USA | [22]Department of Aquatic Resources, Institute of Marine Research, Swedish University of Agricultural Sciences, Lysekil, Sweden | [23]New Zealand Ministry for Primary Industries, Wellington, New Zealand | [24]National Marine Fisheries Service, NOAA, Seattle, Washington, USA | [25]SoFish Consulting Ltd., Wellington, New Zealand | [26]Instituto Español de Oceanografía, Madrid, Spain | [27]Wageningen Marine Research, Ijmuiden, the Netherlands | [28]AZTI, Marine Research, Basque Research and Technology Alliance (BRTA), Pasaia, Gipuzkoa, Spain

Correspondence: Aaron M. Berger (aaron.berger@noaa.gov)

Received: 21 June 2025 | **Revised:** 20 November 2025 | **Accepted:** 1 December 2025

Keywords: fisheries management | movement | simulation experiment | spatially aggregated assessment models | spatially explicit assessment models | spatially explicit operating models

ABSTRACT

Despite their potential to inform sustainable regional harvest and climate-resilient fisheries management, spatial stock assessment models remain underused for management advice. To identify barriers that inhibit broader use of these methods, we conducted a blinded international simulation experiment mimicking real-world stock assessment development when confronting spatial complexity. Seven analyst teams built spatially aggregated and spatially explicit assessment models using data simulated from high-resolution operating models based on Indian Ocean yellowfin tuna and Ross Sea Antarctic toothfish dynamics. Each team documented how assessment software platform, data analyses, model building approach, and diagnostics influenced model complexity and realism. A consensus emerged on key assessment building approaches: (1) conduct high-resolution data analyses to identify appropriate spatial structure; (2) start with simplified models and incrementally add complexity; (3) iteratively evaluate diagnostics to determine necessary spatial complexity; and (4) maintain models with different spatial structures to aid interpretation. The experiment also revealed several valuable insights for parameterising assessments, including consideration of data pre-processing with spatiotemporal models to better inform data-sparse regions; regression trees to identify fleet and spatial structure; trade-offs in complexity between productivity and movement dynamics to achieve tractable and stable model structures; and ensemble modelling approaches to address structural uncertainty. Our findings demonstrate that international collaborations and simulation experiments are crucial for addressing challenges in implementing spatial stock assessments and for evaluating whether their added complexity is justified given management objectives. Broader collaborations are encouraged to foster innovation in fisheries management and to help recognise the practical trade-offs between model parsimony and complexity.

1 | Introduction

Space is a fundamental dimension of most ocean management decisions (van den Burg et al. 2019; Pittman et al. 2021). Sustainable harvest of living marine resources requires scientifically informed management decisions that integrate myriad dynamic biological and fishery processes (Goethel, Omori, et al. 2023) across a progressively demanding ocean-use landscape (Rea et al. 2017). Hence, there has been growing recognition that stock assessments, which provide the scientific basis to support fisheries management, must more explicitly incorporate spatial marine information into species harvest and ecosystem health decision processes. While there has been increased development of spatial stock assessment models and software applications to inform spatiotemporal management actions (Punt 2019b; Goethel, Berger, and Cadrin 2023; see Berger et al. 2024 for a description of the spatial capabilities of the most widely implemented assessment platforms), utilisation of spatial assessments within management frameworks remains scarce despite a clear need to ensure sustainable regional harvest and protect biocomplexity. Primary impediments to wider operational use of spatial assessments include increased data requirements, added model complexity, expanded model specification decisions (e.g., multiplicative effects of dimensionality and uncertainty; Evans et al. 2015), and institutional inertia (Berger et al. 2017; Punt 2019a, 2019b; Goethel, Berger, and Cadrin 2023).

The choice between spatially aggregated (single population), spatially implicit (e.g., fleets-as-areas), or spatially explicit (mechanistic account of spatial dynamics) models often hinges on data availability and the underlying spatial complexity of the modelled population (Berger et al. 2024). Ideally, spatial model complexity should be guided by simulation testing, which can evaluate variance-bias tradeoffs and assess performance related to management objectives. Previous studies have demonstrated that ignoring spatial processes in stock assessments can bias estimates of stock status and management quantities (McGilliard et al. 2015; Goethel et al. 2021; Bosley et al. 2022). However, spatially explicit models may perform worse than their spatially implicit or spatially aggregated counterparts when spatial dynamics are uncertain or unknown (Li et al. 2015; Lee et al. 2017; Guan et al. 2019). Despite the obvious recognition that model performance depends on how well it reflects population structure, the influence of the analyst and their decisions—prior knowledge, model development approach, and assessment platform choice—on model robustness remains poorly understood (e.g., NRC 1998; Deroba et al. 2015). Spatial structure increases the dimensionality and uncertainty associated with these issues. This underscores the need for more realistic simulation studies that emulate the real-world uncertainties associated with developing and implementing spatial assessment models (Goethel et al. 2024). In particular, studies where the true spatial structure is unknown to analysts (i.e., a blinded study design) can help identify biases arising from incorrect structural assumptions and evaluate whether incorporating spatial structure is necessary.

Simulation studies are an ideal tool to explore barriers and solutions to including spatial structure in stock assessments for management (Goethel et al. 2024; Punt et al. 2025). However, no previous studies have applied multiple spatial stock assessment software packages to the data simulated by spatially explicit operating models (OMs). To address this gap, we designed a collaborative, blinded, team-oriented simulation experiment to elicit how expert analysts navigate the assessment process when confronting spatial complexity, with emphasis on key successes, barriers, uncertainties, and novel approaches during model development. The objective of the experiment was to identify practical strategies to support broader adoption of spatial stock models in operational fisheries advice.

Multiple assessment analyst teams developed stock assessments based on the same high-resolution OM data, but were blinded to the true underlying spatial structure and dynamics. Here, we document and synthesise the analytical processes of each team, including workflows, model structures, and decision-making processes, to explore how spatial complexity was interpreted and addressed across teams. As spatial stock assessments gain traction in management contexts, this study provides a template for robust spatial simulation testing and identifies key considerations for integrating spatial dynamics into decision-making frameworks.

2 | Methods

A high-resolution, spatially explicit OM was used to simulate data conditioned on real-world biology and fishery dynamics for two case studies, Indian Ocean yellowfin tuna (YFT; *Thunnus albacares*) and Ross Sea region Antarctic toothfish (TOA; *Dissostichus mawsoni*; see Figure 1). The two exemplar stocks were identified as economically and internationally important case studies with spatial structure driven by unique processes. Yellowfin tuna is a highly migratory species with a broad-scale distribution that is harvested using multiple gear types. Antarctic toothfish is a deep-dwelling demersal species that exhibits ontogenetic migration patterns and is harvested by a single gear-type fishery in the Ross Sea region. The general population and fishery dynamics for these two species offered a way to simulate some real-world spatial population structure situations for this experiment. The study design is not intended to address specific management concerns for these species.

An international group of stock assessment analysts was convened to develop stock assessments for the simulated datasets using various software platforms with spatial modelling capabilities. We compared model-building processes across the analyst teams and assessment spatial structures. By emulating the data and knowledge gaps of real-world assessment applications, the simulation experimental design was able to represent the entire stock assessment process from high-resolution data analysis to model development and diagnostic testing. The primary goals of the experiment were to document how each analyst team approached the building of a spatially explicit stock assessment model, record how various modelling assumptions and parametrisations were decided, and then compare overall performance across model spatial structures. The experiment concluded with an in-person workshop to discuss and interpret the simulation results, summarise lessons learned, and provide recommendations for similar simulation studies in the future.

The design and results of the experiment are presented across four research products. A review manuscript summarised the spatial capabilities of current generalised assessment platforms (Berger et al. 2024). The utility and development of the spatially explicit, high-resolution, data-conditioned OMs are forthcoming (based upon earlier work by Dunn, Hoyle, and Datta 2020). The simulation results based on the comparison of model outputs across spatial structures for the yellowfin tuna case study, along with general conclusions from the simulation experiment, are provided in Goethel et al. (2024). In this article, we describe the development process undertaken by each team to implement a spatial assessment model, including the decisions and approaches that analyst teams confronted. We provide an overview of the experimental design, describe each OM and EM (assessment platform), and describe the criteria used to contrast each team's model building process and resultant model complexity.

2.1 | Experimental Study Design

The experiment organising committee developed and conditioned the OMs, simulated replicate data sets, disseminated data and guidance documents to each analyst team, and collated experiment results (Figures 1 and 2). Seven assessment analyst teams, spanning multiple countries and regional fisheries management organisations (RFMOs), were each tasked with developing a spatially aggregated and spatially explicit stock assessment model using an assessment platform of their choosing. Teams received a standard set of base instructions and guidance documents[1] to ensure consistent and identical information was available. Each team was requested to document the model building process, including data analysis, parametrisation choices, model diagnostics, and model validation approaches. Analyst teams could develop an EM for either or both case studies, but were encouraged to minimally develop a model for the yellowfin tuna case study first. EMs were built using a single representative dataset first and then applied to the remaining 99 replicate data sets for a total of 100 simulations for the given

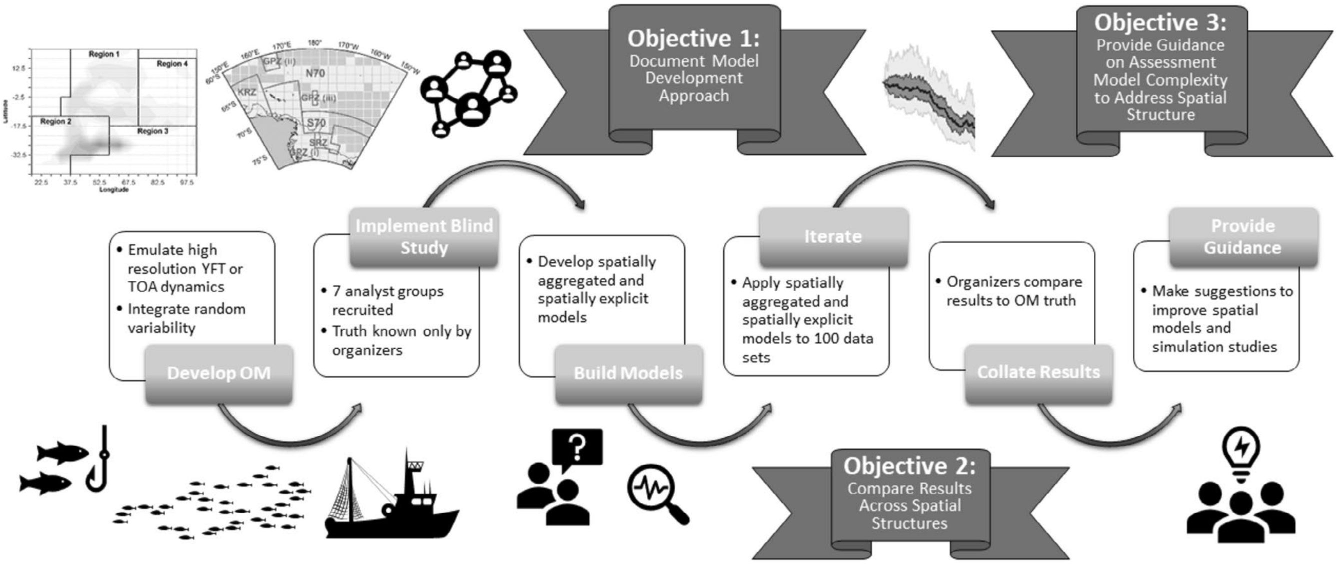
Region 1
Region 2
Region 3
Region 4
Latitude
Longitude
12.5
-2.5
-17.5
-32.5
22.5
37.5
52.5
67.5
82.5
97.5
150°E
160°E
170°E
180°
170°W
160°W
150°W
60°S
65°S
70°S
75°S
GPZ (ii)
N70
KRZ
GPZ (iii)
S70
SRZ
GPZ (i)
Objective 1:
Document Model Development Approach
Objective 3:
Provide Guidance on Assessment Model Complexity to Address Spatial Structure
• Emulate high resolution YFT or TOA dynamics
• Integrate random variability
Develop OM
Implement Blind Study
• 7 analyst groups recruited
• Truth known only by organizers
• Develop spatially aggregated and spatially explicit models
Build Models
Iterate
• Apply spatially aggregated and spatially explicit models to 100 data sets
• Organizers compare results to OM truth
Collate Results
Provide Guidance
• Make suggestions to improve spatial models and simulation studies
?
Objective 2:
Compare Results Across Spatial Structures

FIGURE 1 | A conceptual overview of the simulation study design where the focus of the current article is on objectives one and three (adapted from Goethel et al. 2024, which address objective 2).

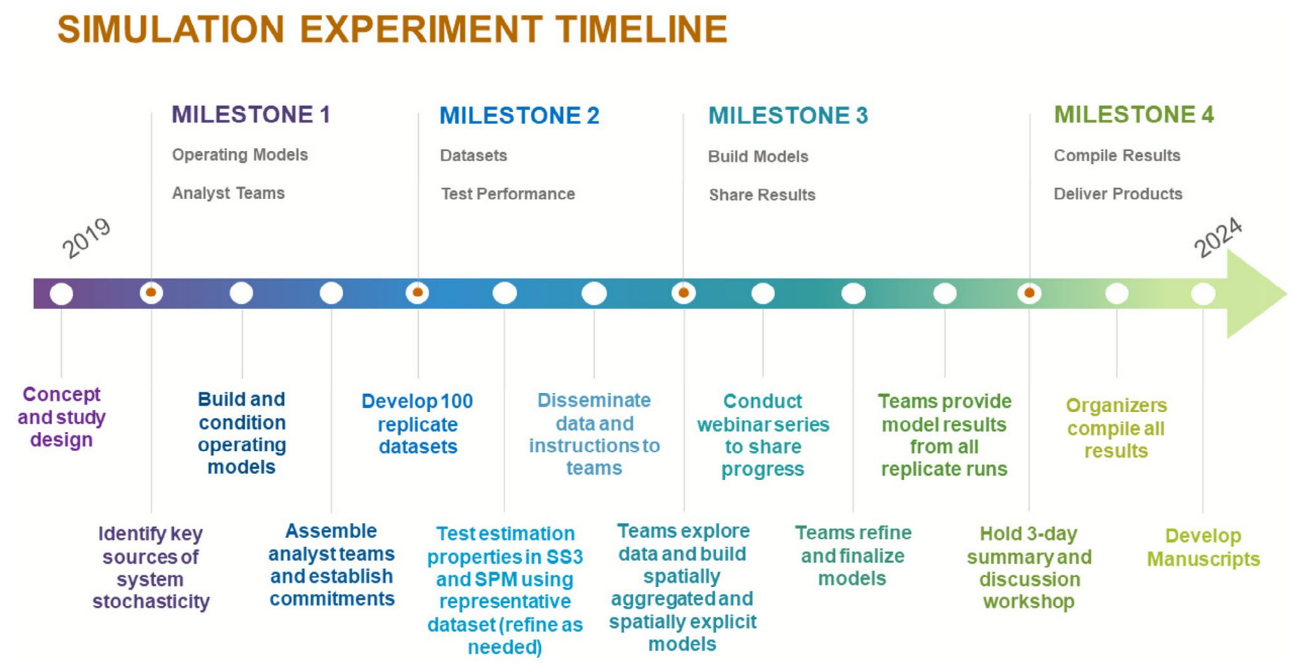
SIMULATION EXPERIMENT TIMELINE
MILESTONE 1
Operating Models
Analyst Teams
MILESTONE 2
Datasets
Test Performance
MILESTONE 3
Build Models
Share Results
MILESTONE 4
Compile Results
Deliver Products
2019
2024
Concept and study design
Identify key sources of system stochasticity
Build and condition operating models
Assemble analyst teams and establish commitments
Develop 100 replicate datasets
Test estimation properties in SS3 and SPM using representative dataset (refine as needed)
Disseminate data and instructions to teams
Teams explore data and build spatially aggregated and spatially explicit models
Conduct webinar series to share progress
Teams refine and finalize models
Teams provide model results from all replicate runs
Hold 3-day summary and discussion workshop
Organizers compile all results
Develop Manuscripts

FIGURE 2 | The key developments, products, and milestones associated with the simulation experiment study design.

case study. Results (including model outputs and descriptions of the model building process) were provided to the experiment organisers for synthesis and comparison (see Goethel et al. 2024, for a summary of model performance results; experiment materials and model outputs are available at the project GitHub[2] site). Further OM (SM.A and SM.B for YFT and TOA, respectively) and EM (SM.C and SM.D) details are provided in the Supporting Information.

2.2 | Operating Models

The OMs were developed with the Spatial Population Model (SPM; Dunn, Rasmussen, and Mormede 2020), a spatially explicit population modelling program written in C++ (source code is available[3]). SPM allows high-resolution modelling of populations with complex spatial structure and enables integration of environmental covariates to inform spatial distribution and movement. The SPM implementations for both exemplar species were conditioned on empirical data, observed biology, and expert judgment to inform important ecological processes.

The general structure (Table 1) and subsequent development of each OM followed several steps.

i. The SPM was tailored for a given species by emulating assumptions and dynamics from the most recent stock assessment.

ii. The SPM was then conditioned by running in estimation mode and fitting to the available empirical datasets (i.e., catch, catch-per-unit effort (CPUE), tagging, and length or age frequency data) and estimating key population and fishing parameters (e.g., fishery selectivity, catchability, fishing mortality, and parameters defining the movement preference functions).

iii. Model validation was conducted (see Dunn, Hoyle, and Datta 2020) to ensure tractable model performance, including fitting the simulated data with an independent stock assessment platform (Stock Synthesis 3, or SS3, for YFT; C++ Algorithmic Stock Assessment Laboratory, or CASAL, for TOA) to ensure that the data was informative and sufficient to allow model convergence and satisfactory estimation performance (e.g., estimated spawning stock biomass trends aligned with the OM reality).

iv. The conditioned OM was then evaluated across one hundred stochastic realisations in simulation mode to generate simulated (i.e., 'pseudo') replicate data sets that differed according to observation error (CPUE, maturity, compositional, and tagging data) and process error (recruitment).

v. For the purpose of developing and testing stock assessment models with different spatial structures, the pseudo-data for each OM realisation was aggregated and made available to analyst teams at three different spatial resolutions: grid cell-specific data (no aggregation) was made available for data exploration, a post-factum aggregation of the data to four regions, and a post-factum aggregation to a single region.

The OMs were uniquely parameterised to calculate abundance-at-age and biomass-at-age by grid cell, reproductive category (immature and mature), and tagged category (tagged, untagged) by accounting for spatially explicit recruitment, movement, fishing and natural mortality, and demographic processes (Figure 3). The primary spatiotemporal population and fishery dynamics for the YFT and TOA OMs are described next, while the full details of each OM are provided in Supporting Informations SM.A and SM.B (YFT and TOA, respectively) and on GitHub.[4,5]

TABLE 1 | The general settings used for the two (i.e., Indian Ocean yellowfin tuna, YFT, and Antarctic toothfish in the Ross Sea Region, TOA) operating models (OMs) as well as the primary parametrisation options used by each analyst team for their final spatial (or spatially implicit) assessment model (AM).

					Key model structure and paramertisation decisions					
Species	**Model type**	**Platform**	**Analyst team**	**Spatial population structure**	**Strata**	**Fleet structure**	**Recruitment**	**Movement**	**Tagging data use**	**Other**
YFT	OM	Spatial Population Model (SPM)	*Organisers*	Individual Grid Cells (highly migratory)	221	Seven fleets operating at different spatiotemporal dimensions across the model domain.	Global B-H SRR with quarterly year class multiplier and apportionment to grid cell based on distribuition of small (< 40 cm) fish.	Based on maturity state and habitat preference functions (distance between cells, sea surface temperature, and chlorophyll-a) by cell.	Yes, fit during data conditioning, then simulated at the cell level; tag mixing at the regional level was low, but higher for immature fish.	Data conditioned OM with high resolution outputs.
	AM	C++ Algorithmic Assessment Library, 2nd Generation (CASAL2)	*Casal2*	Areas-as-fleets	1	Full fleet structure assigned to four sub-regions.	Global B-H SRR with quarterly year class multiplier estimates.	None.	No.	Selectivity estimated and composition data fit for each fleet and region combination.
				Spatially Disaggregated	4	Full fleet structure fit at regional scale.	Regional B-H SRR, each with independent quarterly year class multiplier estimates.	None.	No.	Assumed four independent spatial populations with no movement.
		Multiple Length Frequency Analysis-Catch at Length (MFCL)	*MFCL*	Spatial heterogeneity	4	Full fleet structure by region with selectivity mirrored across regions for same gear-based fleets.	Global B-H SRR with quarterly year class multiplier estimates and apportionment estimated for each region with deviations.	Estimated for each shared regional boundary, with movement assumed to be age- and time-invariant except for a single, constant deviation for each season.	Yes, with 5 mixing periods (conducted sensitivity runs to alternative mixing periods).	An interim areas-as-fleets model was used to help specify the fleet structure used in the spatially explicit (spatial heterogeneity) model.
		Spatial Processes and Stock Assessment Methods (SPASAM)	*SPASAM*	Spatial heterogeneity	2	Fleets combined to reduce to four per region.	Global B-H SRR with quarterly year class multiplier estimates and apportionment fixed at 99% for regions 1–2 and 1% to regions 3–4.	Estimated among all regions, with movement estimated for two age groups (± age-9 quarters) in every other pseudo-year (half-year time blocks).	Yes, with complete mixing assumed (no latency period).	Two region structure as an aggregation of four regions (1–2, 3–4).
		Stock Synthesis 3 (SS3)	*SS3_A*	Spatial heterogeneity	4	Full fleet structure by region with selectivity mirrored across regions for some fleets.	Global B-H SRR with quarterly year class multiplier estimates and apportionment deviations estimated for each time step for three regions only. Assume no recruitment in region three.	Estimated between specific regions (1–2, 1–4, and 4–3) for two age groups (± age-16 quarters) and with no time-variation.	Yes, with 6 mixing periods and reporting rate estimated.	Cell-specific CPUE analysed with spatio-temporal model to create CPUE index; developed quarter and annual time step models.
			SS3_B	Spatial heterogeneity	4	Full fleet structure fit at regional scale.	Global B-H SRR with quarterly year class multiplier estimates and apportionment estimated for each region without deviations.	Estimated between specific regions (1–2, 1–4, and 3–4) for two age groups (± age-9 quarters) and with no time-variation.	Yes, with 4 mixing periods.	An interim areas-as-fleets model was used to help specify spatially explicit (spatial heterogeniety) model settings.
			SS3_C	Areas-as-fleets	1	Full fleet structure (sub-region and fleet combinations), where purse seine fleet structure was identified using regression tree analysis.	Global B-H SRR with quarterly year class multiplier.	None.	No.	Cell-specific CPUE analysed with spatiotemporal model to create CPUE index and associated length compositions.

(Continues)

TABLE 1 | (Continued)

					Key model structure and paramertisation decisions					
Species	Model type	Platform	Analyst team	Spatial population structure	Strata	Fleet structure	Recruitment	Movement	Tagging data use	Other
TOA	OM	Spatial Population Model (SPM)	*Organisers*	Individual Grid Cells (Ontogenetic distribution patterns)	189	A single fleet operating across the spatial domain.	Global B-H SRR with annual year class multiplier and equal apportionment to grid cells shallower than 800 m.	Based on life stage (ontogenetic), according to immature, mature, pre-spawning, spawning, and post-spawning status, and preference functions[a]	Yes, fit during data conditioning, then simulated at the cell level by age and life stage.	Data conditioned OM with high resolution outputs.
	AM	C++ Algorithmic Assessment Library, 1st Generation (CASAL)	*CASAL*	Areas-as-fleets	1	Four fleets, one assigned to each of four sub-regions that align with management areas.	Global B-H SRR with annual year class multiplier.	None.	Yes, used to estimate population size with complete mixing assumed (no latency period).	No CPUE time series was used.

Note: All teams built a single area model first (not shown, but see Goethel et al. 2024 for details), before building in alternative spatial model characteristics. Full fleet structure for YFT models indicates 16 fleets, which include all available gear types by four region combinations.

[a]Movement was based on habitat preference functions by cell according to depth (median and proportion in 450 to 2870 m depth zone), temperature at 500 m depth, presence of seamounts, and the Euclidian distance between cells or the presence of hills, depending on the reproductive category.

2.2.1 | Indian Ocean Yellowfin Tuna

Yellowfin tuna is a highly migratory species that ranges across the Indian Ocean, exhibiting complex spatial structure, movement, and fishery dynamics. The YFT OM assumes a single population modelled across a spatial grid of 221 equal-sized cells (5° latitude × 5° longitude; Figure 3). The model was single-sex and age-structured (ages 1 to 7+ years, modelled as 28 quarter-ages) with a quarterly time step (i.e., representing quarter-years from 1952 to 2015), and tracked both immature and mature fish independently. The quarterly time step was split into three phases: (1) recruitment of fish followed by tag releases and then fish movement among cells; (2) fishery operations to achieve cell-specific catch and all data observations; and (3) biological state-level transitions (including natural mortality and age, maturity, and growth increments). The YFT OM was primarily parametrised based on the recent operational spatial assessment built in Stock Synthesis 3 (SS3), which modelled four spatial regions (Fu et al. 2018).

Recruitment occurred quarterly (i.e., seasonally at 0.25 year increments) at the first age group (quarter age 1) and was the product of the stock-recruitment relationship (with a one quarter year lag from associated SSB), a year-class strength multiplier for each cohort, and a geographic apportionment that assigned total recruitment to individual grid cells. The base recruitment apportionment layer was defined by the average observed distribution of juvenile fish < 40 cm in the fishery. Movement was modelled using habitat-based preference functions that differed by maturity stage where grid cell preference included information about sea surface temperature, chlorophyll-*a* concentrations, and the distance between cells.

Seven unique fleets harvest YFT at varying spatiotemporal scales. These fleets were modelled assuming a time- and spatially-invariant double-normal selectivity function, except for the primary longline fleet, which was assumed to have a logistic selectivity function and time-invariant catchability (i.e., for CPUE calculations). Each fleet had a unique spatiotemporal distribution mimicking the real-world catch distributions (Fu et al. 2018). The amount and distribution of mark-recapture tagging data and release events also followed the actual data for yellowfin tuna (Fu et al. 2018), with the model adjusting for tag loss, tag-related mortality, and reporting rates. Many of the observed tag releases were of immature fish in a limited geographic region. Collectively, the spatial dynamics of the yellowfin tuna OM led to a temporally dynamic distribution of cells with high population density, exemplified by high initial spawning stock biomass (SSB) in northern cells that rapidly declined due to high fishing pressure (Goethel et al. 2024). The movement preference functions of immature fish, which indicated they were more likely to move, resulted in broader regional intermixing of juveniles.

Data used to condition the YFT OM included catch, CPUE, catch length composition, tag releases, and tag recaptures. For the OM simulations, data of the types used for conditioning were simulated assuming observation and process error, except that total catch was assumed known without error. Recruitment process error was driven temporally by quarter-specific year-class strength multipliers that assumed lognormally distributed deviations with a variance of 0.6 and

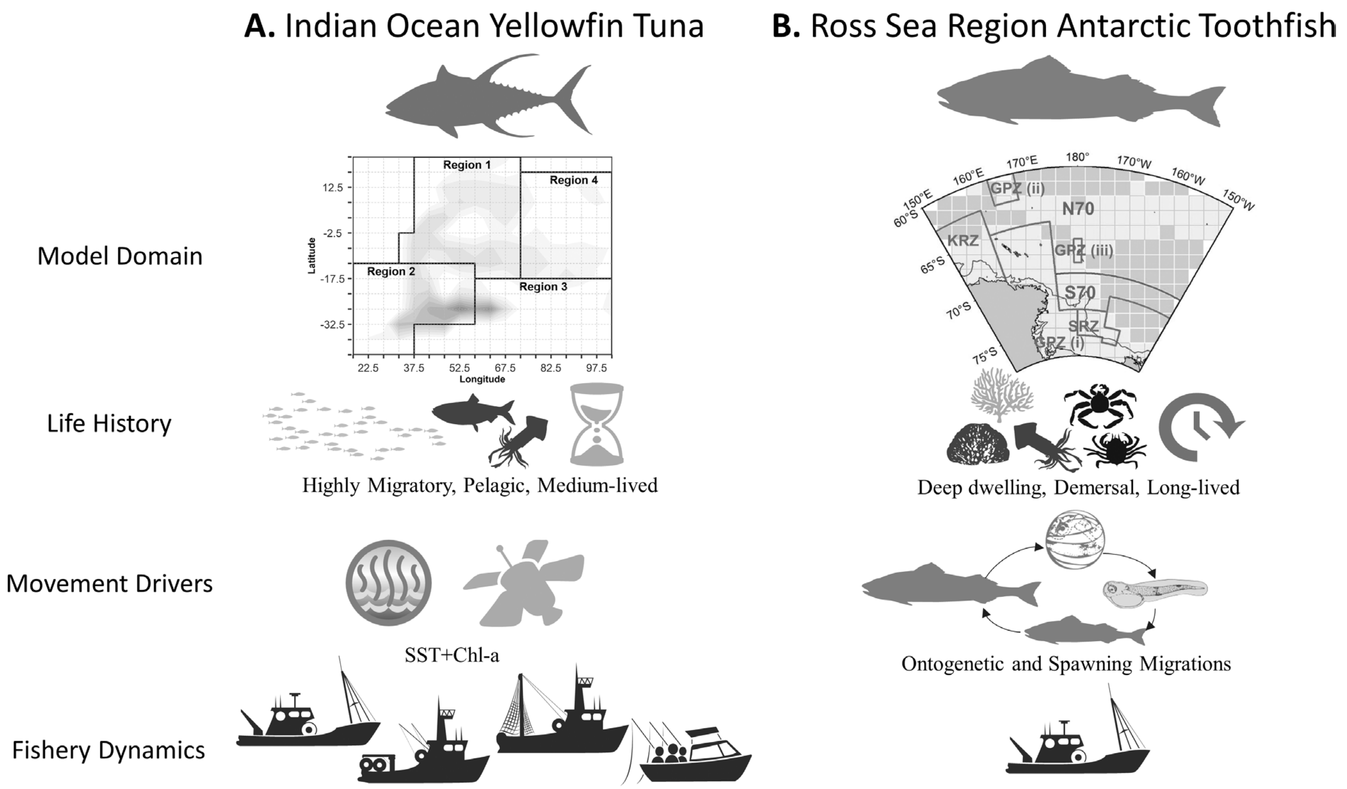

FIGURE 3 | Summary of the primary dynamics and spatial drivers of the Indian Ocean yellowfin tuna (A) and Ross Sea region Antarctic toothfish (B) case studies.

spatially by quarter-specific weighted random deviations in cell-specific apportionment using correlated Gaussian random fields (SM.A).

2.2.2 | Antarctic Toothfish

Spatial dynamics for Antarctic toothfish in the Ross Sea region primarily stem from ontogenetic movement patterns (where fish progressively move to deeper water as they age) and migrations to and from spawning grounds, while exploitation is due to a single longline fishery. Population dynamics result in spatial selectivity patterns due to differences in the availability of fish to the fishing gear at different locations according to age and season (Hanchet et al. 2008; Grüss et al. 2021). The TOA OM is based on work originally developed to better model toothfish spatiotemporal population and management dynamics (Mormede et al. 2014). Parametrisation of the OM was informed by an operational spatial assessment built in CASAL assuming four spatial regions (Dunn 2019).

The TOA OM assumes a single population modelled across a spatial grid of 189 equal-sized cells (156 km × 156 km; Figure 3). The model was single-sex and age-structured (i.e., ages 2 to 30+) with an annual time step (i.e., representing years 1995 to 2021). The life cycle of TOA was emulated by tracking fish in each grid cell by reproductive state that differed across ages and within years, including: immature, mature, pre-spawning, spawning, or post-spawning (Mormede et al. 2014). The yearly time step was split into three seasons, each allowing for specific population dynamics. The summer season included recruitment, fishing, half of natural mortality, tag releases, and the specification of reproductive state. In winter, the other half of natural mortality occurred as well as movement of immature fish, spawning migrations, reproductive transitions from pre-spawners to spawners, and spawning. Spring was solely used for reproductive transitions from spawners to post-spawners, post-spawning migrations, and the annual age increment.

Recruitment occurred at age 2 in all cells with a depth less than 800 m (i.e., spatially stationary), and total recruitment was the product of the stock-recruitment relationship and a year-class strength multiplier. Movement was modelled using habitat-based preference functions that differed by reproductive state, where grid cell preference was based on median depth, temperature at 500 m depth, proportion of preferred depth habitat in each cell, the presence of seamounts, and the distance between cells. The fishery was modelled as a single longline fleet operating across the model domain represented by a spatiotemporally invariant selectivity and catchability (i.e., for CPUE calculations). The amount and distribution of mark-recapture tagging data and release events also mimicked actual data for toothfish, where fish were tagged in proportion to the full-size distribution of the catch for each age and reproductive state. Tag loss, tag-related mortality, and non-reporting were all assumed to be negligible. The general spatial dynamics of the TOA OM was primarily based on the movement preference functions, which led to ontogenetic distribution patterns according to maturity and spawning states.

Data used to condition the TOA OM included catch, CPUE, catch age composition, tag releases, tag recaptures, and proportions spawning-at-age. The same data types used for conditioning the OM were simulated, assuming observation and process error, except total catch was assumed to be known without error. Process error in recruitment year class strength was implemented for each iteration, assuming lognormally distributed deviations with a standard deviation of 0.6.

2.3 | Assessment Models

Seven analyst teams independently developed stock assessment model(s) for the experiment using one of four different stock assessment platforms (see Table 1 for model, platform, and team descriptions). Six teams developed assessment models applied to data generated from the YFT OM. These included models built using the Casal2 (C++ algorithmic stock assessment laboratory 2nd Generation; Doonan et al. 2016), MFCL (Multiple Length Frequency Analysis Catch-at-Length; Fournier et al. 1990), SPASAM (Spatial Processes and Stock Assessment Methods; Goethel et al. 2019, 2021), and SS3 (Stock Synthesis; Methot and Wetzel 2013) stock assessment platforms. One team developed an assessment model applied to data generated from the TOA OM using the CASAL (Bull et al. 2012; the predecessor to Casal2) stock assessment platform. All platforms, at minimum, allow for a single stratum (panmictic population within a spatially aggregated assessment) and multiple strata (spatial heterogeneity among assessment regions) population structures. Each platform has many options (and some limitations) for integrating spatial data and parameterising spatial model structure, including choices about population structure, recruitment dynamics, connectivity, fleet dynamics, tagging dynamics, and environmental covariates. Berger et al. (2024) provide a detailed review of the spatial functionality and capabilities of common stock assessment platforms, including those used here. Detailed descriptions of each platform can be obtained from user manuals and associated web-based code management platforms (Table SM.C.1).

2.4 | Model Development and Comparison

Documented data analysis and model building steps, including specific methods, were considered the observational unit of the experiment. Specifically, stepwise choices associated with high-resolution data analysis, parameterisation of key demographics (e.g., recruitment and movement), spatial model approach and inference, and spatial diagnostics were compiled for each spatial assessment and OM pairing. These choices were then compared across teams to identify commonalities and differences in how spatial structure and model complexity tradeoffs were confronted (and ultimately dealt with) by the analyst teams. Qualitative summaries provide a deeper understanding of how analysts approach spatial stock assessment model building when confronted with real-world uncertainties, as well as how they address common decision points, particularly those associated with spatial dynamics.

3 | Results

Across the seven independent analyst teams, consistent patterns emerged in the process of developing spatial stock assessment models, highlighting both challenges and effective strategies for addressing spatial complexity. Regardless of software platform, each team converged on similar approaches, prioritising data exploration, incremental model building, diagnostic-driven complexity, and multiple model structures.

Each of the seven analyst teams implemented at least two assessments with varying spatial structures (i.e., a spatially aggregated and either a spatially implicit or spatially explicit assessment) and applied each to one of the OMs (Table 1). Typically, groups began with a high-resolution data exploration to understand the main spatial dynamics and identify data opportunities and limitations (e.g., sparsity in length or age compositions), while also aiding subsequent spatial model parametrisation decisions. All teams implemented a spatially aggregated assessment first, adding complexity in a stepwise fashion, as deemed necessary (Figure 4). Residual patterns in the fits of each model were used to help identify model misspecification, particularly for spatial processes. Thus, the general process for each team included incremental steps beginning with data exploration, developing conceptual models, implementing simple spatially aggregated models, adding complexity (e.g., fleet structure) within spatially aggregated models to address diagnostic issues, then, incrementally, adding (or simplifying) spatial complexity. Descriptions of each analyst team's assessment model development process are provided in the Supporting Information (SM.D).

3.1 | Data Explorations

The exploration of catch, effort, tagging, and composition data at the highest resolution available was common (Table 2; Figure 4). For example, team *Casal2* automated data summary procedures to rapidly adjust and evaluate data aggregations at different temporal and spatial scales that aligned with a realistic set of hypotheses, primarily related to population and fishery spatiotemporal structures. The exploratory analyses of catch, CPUE, and tagging data highlighted the need for seasonal time-steps and consideration of movement, while length composition data indicated potential scale-dependent sparsity issues leading to minimally sufficient spatiotemporal fishery definitions. Team *SS3_C* explicitly categorised data products using regression tree models (following Lennert-Cody et al. 2023; Maunder et al. 2022) to quantitatively define fleet structure, thereby removing the often discretionary (or arbitrary) nature of fleet definitions in stock assessment. Two teams used species distribution models to develop spatiotemporal CPUE indices of relative abundance and to inform regional abundance scaling parameters (following Hoyle and Langley 2020). One team additionally used species distribution models to develop spatiotemporal compositional data. Initial data explorations also informed subsequent modeling decisions, such as by providing information on tagging dispersion parameters, length composition effective sample sizes, and model dimensions (including candidate spatial structures).

3.2 | Model Development

Spatially explicit models were typically developed much later in the process and built iteratively. Each model update was incremental and focused on a specific spatial complexity, after which analysts investigated the resultant diagnostics to understand

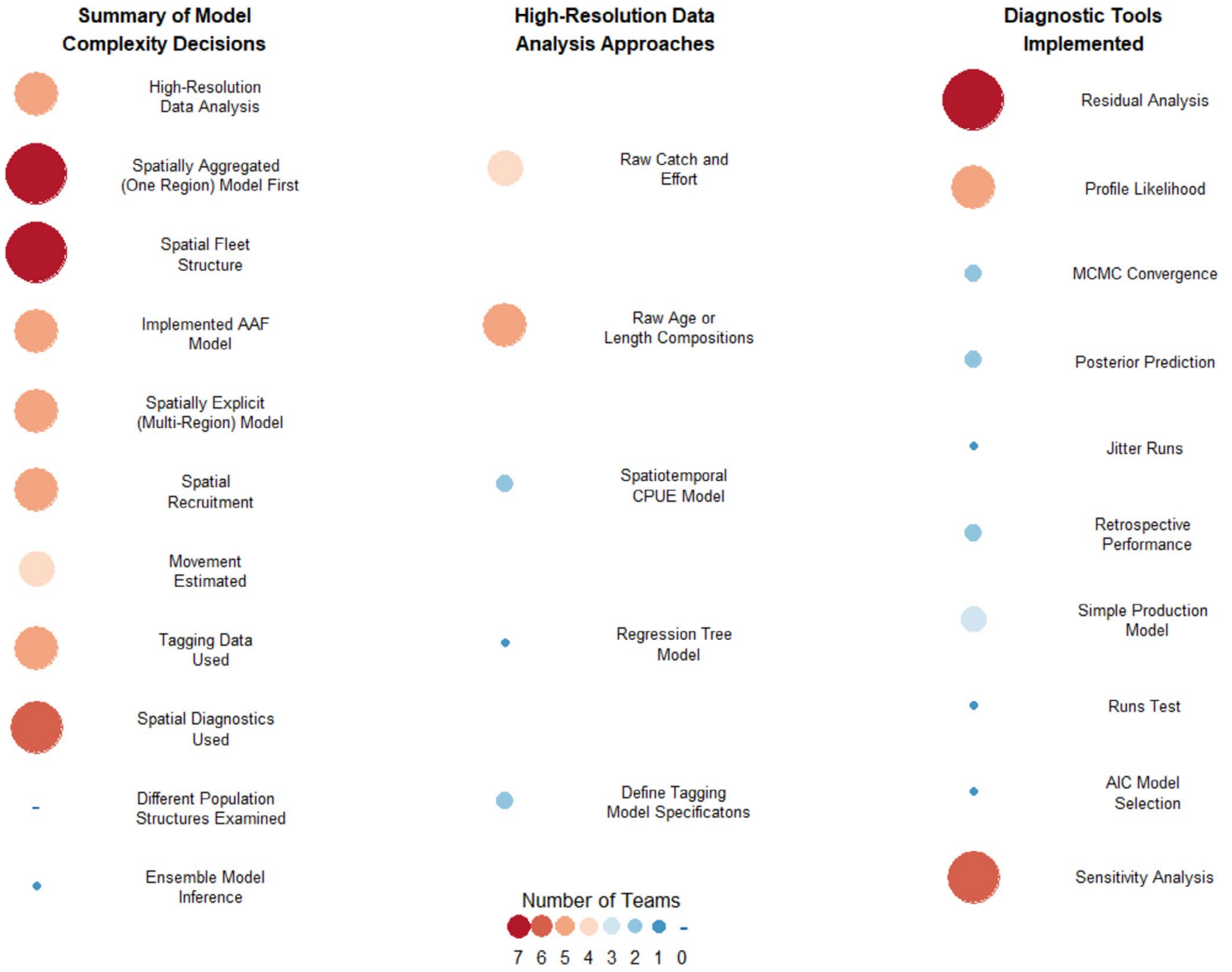

FIGURE 4 | A summary across analytical teams ($n = 7$) of key decisions made that informed final model complexity (first column), high-resolution data analyses utilised (second column), and diagnostic tools implemented (third column). The size and color of the bubble correspond to the number of teams that utilised a particular approach or analytical process.

model performance. Often, further data analyses would be explored based on insights gained or questions arising from a poorly performing spatial parametrisation. For instance, tagging analyses were often revisited once initial spatial models were running to better understand how best to handle tag nuisance parameters (such as the mixing period and reporting rate). Teams also undertook unique approaches to balance fleet and region structure, often focusing on parsimony to reduce the number of parameters related to movement, recruitment, and fleet selectivity (Table 2).

Residual fits and patterns were the primary diagnostics used by all teams to investigate model fit and adequacy (Table 2; Figure 4). In particular, persistent residual patterns during the development of spatially aggregated models were the driving force behind increases in model complexity. For example, the incorporation of tagging data was informed by analysts' interpretation of data representativeness and model fit diagnostics. Specifically, the lack of spatial, temporal, and age coverage of the tagging data (i.e., tagged YFT were mostly immature fish in a small portion of the domain during a relatively short time period) led some groups to omit it or develop hypotheses about time at liberty for proper tag mixing. Sensitivity analyses and profile likelihoods over influential parameters were also common tools used to identify alternative plausible model structures and within-model parameter identifiability. Other diagnostic tools included model stability metrics (i.e., jittered parameter runs and retrospective model performance), model selection criteria (i.e., AIC and runs tests), and implementation of simpler surplus production models. Two teams incorporated Bayesian diagnostics (e.g., MCMC convergence criteria, posterior trace plots, chain autocorrelation, and posterior predictions) to refine model parameterisations that were not identifiable or had poor convergence properties. Team *SPASAM* used run time as an implicit diagnostic to reduce fleet and temporal structure in the assessment model. In general, teams noted that they would have explored more model diagnostics if they had had more time or if spatial models were quicker to converge to a solution.

3.3 | Spatial Model Complexity

Nearly all assessment configurations that attempted to explicitly address spatial dynamics (six of eight) incorporated spatial

TABLE 2 | A summary of approaches used (common and distinctive) and major impediments encountered across all analyst teams during the model building and development process, with emphasis on those that helped inform the final spatial model complexity.

Spatial consideration	Common approaches	Distinctive approaches	Major impediments[a]
Data Analysis and Weighting	Evalute raw catch, effort, and composion data Start simple and add spatial complexity Iteratively reweight composition data (Francis method)	Explored alternative time steps Spatiotemporal model explorations (CPUE, composition data) Categorise data properties (regression tree) Reweight tagging dispersion parameter	Sparse effective sample sizes leading to lack of information Weighting of composition data relative to tagging data
Spatial Units	Single area with multiple fleets Single area with fleets-as-areas (FAAs) Pass on single area parameters as starting point for multi-area parameters Spatially explicit multi-area model	Use fleets-as-areas (FAA) model to define fleets in spatially explicit model Use four-area model diagnostics to build spatially explicit two-area model	Reconciling parsimonious fleet structure Increasing spatial areas lead to computing time bottlenecks and degrading diagnostics
Temporal Structure	Time step matching the OM as the main continuous time step	Aggregated seasonal dynamics to more coarse time steps Nested seasonal time steps within annual time steps Truncated beginning of the time series to shorten run times	Implementing temporal autocorrelation among seasonal time steps across years
Demographics/Biology	Specified per the structure of the operating model	Adjusted operating model specifications to align with alternative temporal structures	Estimates were sensitive to the method used to aggregate parameters (e.g., maturity, natural mortality, and growth) to alternative time steps
Recruitment Dynamics	Single (global) set of recruitment deviations estimated by time step matching the OM with stationary spatial apportionment to areas across time	Add time-varying apportionment of recruitment to areas Fixed apportionment based on simplier spatial model configuration estimates Reduce apportionment parameters to only primary breeding areas Constrain apportionment parameters according to strata-specific biomass distribution (catch-per-unit-effort; CPUE)	Only estimate recruitment deviations for periods with catch-per-unit-effort (CPUE) information
Fleet Structure	Single fleet across entire domain Single fleet per area with area-based selectivity parameters (FAA) initialised by the single area model Multiple fleets per area	Fleets with the same gear in different areas share selectivity parameters Alternative fleet structure models in single area model used as diagnostic to idenfity spatial model fleet structure External regression tree analysis conducted to inform optimal fleet structure	Identifying parsimonious fleet structure parameterisation (though see distinctive regression tree approach)

(Continues)

TABLE 2 | (Continued)

Spatial consideration	Common approaches	Distinctive approaches	Major impediments[a]
Movement	Reduce movement parameters by age, area, or time constraints for tractability Model movement rates for mature and immature fish separately	Only occurs during certain time steps (seasonality) Time-invariant between certain areas Only occurs in major tag data areas Imply movement through area-based selectivity parameters No movement parameters	Tag mixing period assumptions and sensitivity models Non-homogenous distribution of tagging data Overdispersion of tagging data Large number of movement parameters and subsequent modelling options
Tagging Dynamics	Tags used in one and four area models using Peterson[b] or Brownie[c] (or both) modelling methods Nusiance parameters either fixed or evaluated across multiple assumptions (sensitivity models)	Estimated tagging overdispersion parameter Used to inform abundance only (Peterson method) Used an ensemble of no tagging and tagging models with alternative mixing periods Tag data not used	Identifying appropriate tag mixing period and maximum at liberty model specifications as selected approach influenced results Lack of spatial coverage in tag release and recoveries
Diagnostics	Pearson residuals for composition and tagging data Likelihood profiles Sensitivity analyses Model convergence through positive hessian and gradients	Markov chain Monte Carlo (MCMC) trace plots, convergence diagnostics, and posterior prediction evaluations Use of Akaike information criterion (AIC) for nested model selection Model stability through model parameter jitter tests Use of simplier production models or runs tests	Large number of parameters to diagnose for multiple fleets and multiple area models

[a] As highlighted by analysts, not an exhaustive list.
[b] Based on the Lincoln-Peterson mark-recapture method (Lincoln 1930; Petersen 1896).
[c] Based on Brownie dead recovery mark-recapture methods (Brownie et al. 1985).

population and fleet structure in a unique way (Table 1), resulting in a range across teams of selected spatial model complexity (where definitions of relative model complexity follow fig. 2 in Goethel, Berger, and Cadrin 2023). The simplest spatial models (i.e., spatially implicit) used an areas-as-fleets (AAF) approach, whereby fleets of each gear type were defined for each region or further aggregated according to a machine learning algorithm (i.e., a regression tree approach). One team built a spatially disaggregated model, where independent assessment models were produced for each region (i.e., including spatial differences in recruitment but ignoring movement among regions) with fleets for each regional model defined for all gear types.

Four teams produced more complex, regionally linked models through the inclusion of explicit spatial recruitment and movement dynamics. Across spatial models, complexity differed according to the number of regions modelled and how fleets were defined (full complement of region and gear type versus simplifications by mirroring selectivity by gear types across regions). While all teams built up complexity from spatially aggregated single-region models to spatially explicit multi-region models, one team subsequently simplified its spatial assumptions (i.e., re-aggregating a four-region to a two-region spatial model). All assessment models produced during the experiment included region as a primary separator of fleets (and thus selectivity patterns). No team found the need to explore alternative population structures (e.g., due to regional demographic variation), primarily because demographics were specified for each species in the experiment documentation (i.e., each OM assumed a single population with spatial heterogeneity in fleet or population dynamics and no variability across space in growth, maturity, or natural mortality).

The primary axes for parametric exploration beyond fleet selectivity were related to recruitment apportionment and movement dynamics for the four assessment models that were explicitly spatial (Table 1). Recruitment from a global Beverton-Holt stock-recruitment relationship was apportioned by region in all four cases, but each team applied unique simplifying assumptions. The *MFCL* team used the least restrictive assumption by estimating time-varying recruitment apportionment deviation parameters for each region. Other teams reduced the number of apportionment parameters by restricting full time-variation to a subset of regions (*SS3_A*), including temporally stationary parameters for each region (*SS3_B*), and fixing apportionment for each region through time (*SPASAM*). Similarly, all four teams considerably restricted the number of estimated movement parameters across space and time (Table 1). The *SPASAM* team estimated movement parameters for all region combinations for two maturity-based age groups and allowed time variation assuming two season time blocks (i.e., movement was estimated for every other season in their truncated time period model). The *MFCL* team estimated movement rates between regions with shared boundaries and assumed that movement was age- and time-invariant except for a seasonal deviation that was invariant across all years. The two other teams considerably restricted movement to a reduced number of time-invariant region combinations for two maturity-based age groups based on initial explorations of tagging data. In general, teams took different approaches to balancing complexity in productivity and movement. Some elected for relatively high complexity in the parameterisation of movement at the cost of less complexity in recruitment, and vice versa (Table 1).

All teams that integrated tagging data to inform movement (YFT OM) did so following the Brownie tag-recovery method (Brownie et al. 1985), which required them to make tag mixing assumptions. Their assumed tag mixing periods ranged from 0 to 6 time-steps, which had an important influence on model fit and final model complexity. One team (*CASAL* for the TOA OM) elected to integrate tagging data to estimate absolute abundance following the Lincoln-Peterson mark-recapture method (Lincoln 1930; Petersen 1896), primarily in lieu of using a fishery-dependent CPUE index of relative abundance. The choice of how to use the tagging data (i.e., to inform movement or abundance estimates) was influenced primarily by case study dynamics. For example, the TOA OM included spatially representative tagging information, whereas the YFT OM included tagging of juveniles largely within a single region (see SM.B and SM.A, respectively). Given the quality of available tagging data, two teams elected not to use tagging data, ignore movement, and make simplifying population structure assumptions (i.e., assuming a single area with an AAF model or implementing spatially disaggregated assessment models without movement).

3.4 | Impediments and Unique Solutions

Several impediments to building and implementing spatial models were encountered, which also led to differing model building decisions across analyst teams (Table 2). For instance, teams found the realistic (i.e., spatiotemporally sparse) levels of data, in particular length-age composition and tagging data, to be a limiting factor for the degree of spatial complexity that could be integrated into a practical model. Resulting low effective sample sizes required aggregating composition data, which directly influenced decisions regarding the number of spatial strata to model and necessitated careful consideration of how to weight composition data (e.g., by catch, modelled CPUE, or sample sizes in each grid cell) to ensure representativeness of the data within and across regions. Similarly, the limited spatial and age coverage of tagging data in the YFT OM led to unresolved tag modelling assumptions (e.g., tag mixing and overdispersion) and necessitated simplifying assumptions (e.g., grouping movement in age or time blocks).

Collectively, analyst teams' explorations and decisions led to the evolution of unique solutions to manage model complexity (Table 1). For example, several teams reduced model complexity through the refinement of fleet, spatial, and temporal dimensions to enable tractable model run times while simultaneously addressing model fit and parsimony. One team elected to ignore the sparse tagging data and instead addressed population structure by developing multiple independent disaggregated models, where regional dynamics were captured through model-specific stock-recruitment functions. Similarly, another team chose to ignore the tagging data and instead developed spatiotemporal models that smoothed over sparse CPUE and length composition data by utilising spatial autocorrelations to produce representative population information across the entire domain (i.e., attempting to capture the collective end result of unobserved fine-scale dynamics at a single aggregated scale). Several teams considerably reduced the number of movement parameters by restricting movement among specific regions, aggregating to age and time blocks, or assuming age- and time-invariance. Besides movement, tagging data was also used by one team to estimate an abundance trend instead of relying on a fishery-dependent CPUE time series. Given the difficulty in deciding among different structural assumptions, particularly for tag mixing, one team pursued an ensemble modeling approach to implicitly account for structural uncertainty. However, time and resource constraints impeded a full implementation or adequate comparison of model ensemble outputs to single model results.

4 | Discussion

This study highlights how assessment model complexity when confronting spatial structure is not only influenced by biological or ecological considerations, but also by the analyst's background, platform constraints, and institutional priorities. Identifying the appropriate level of complexity in stock assessment models remains a persistent challenge. No comprehensive guidance is available (or likely possible) that can account for all potential scenarios, data limitations, and modeling contexts encountered in operational assessments (though see Punt, 2023 for general guidance). The challenge of balancing parsimony and complexity becomes even more difficult in the presence of persistent spatial structure or spatial dynamics in marine populations (Berger et al. 2017; Goethel et al. 2024). Although most generalised assessment platforms share similar underlying structure, the types of complexity that can be integrated typically depend on the species, region, or institution for which they were developed (Berger et al. 2024).

A notable outcome of this study was the strong 'analyst effect', where model development and final structure were affected by a combination of region (e.g., RFMO affiliation), software platform, individual expertise, and likely, career stage (e.g., student versus professional). For instance, platform choice was closely tied to RFMO affiliation, as RFMOs have often developed proprietary or preferred software. The platform in turn influenced the diagnostics, methods, and assumptions, particularly regarding movement, recruitment, or tagging, based on the options available within the platform. Moreover, several RFMOs viewed the experiment as an opportunity to test specific modelling approaches aligned with their research interests, which helped justify participation and funding, and shaped model development. Finally, resources available, particularly in terms of time dedicated to the project, depended on institutional structure, career stage, and ongoing commitments. Early career participants often dedicated more time and explored a wider array of approaches. Given the many decision points involved in spatial modelling, subtle platform differences, and the study's funding limitations, it is not surprising that this analyst effect was stronger than in previous blinded simulation experiments (e.g., Deroba et al. 2015).

Despite these differences, most analysts followed a broadly similar model development trajectory. However, unique approaches emerged that influenced final model complexity. Based on observations of the model building process and subsequent discussions, we offer general guidance for developing stock assessments when spatial structure is present (Figure 5). Many of the recommendations, such as the use of high-resolution data analysis, apply to both spatial and non-spatial contexts, since they are fundamental to understanding resource dynamics and the representativeness of the data. Moreover, the experiment demonstrated that common processes inform the development of both spatially aggregated and spatially explicit assessment structures, offering insight into when spatially implicit assessments might also be appropriate. Importantly, no single model structure or development approach consistently produced unbiased results (Goethel et al. 2024), emphasising that these recommendations are not necessarily 'best practices' but rather 'common practices' that can provide insight into underlying dynamics and model parametrisation.

4.1 | Data Analysis and Conceptual Model Development

The first step in any stock assessment is to compile an inventory of the available data. High-resolution spatiotemporal analyses should follow to evaluate data availability and quality, while also aiming to identify drivers of population dynamics and spatial structure that should inform parametrisation (e.g., resource distribution, fleet structure, connectivity patterns, and regional patterns in age or length). For example, mapping length composition and CPUE data were widely utilised to understand the distribution of the fleets and resources, while identifying data-sparse regions that might constrain spatial structure. Objective techniques such as regression trees were found useful for identifying fleet structure, and high-resolution mapping of tagging data helped identify representativeness and inform tag mixing assumptions. Flexible data structures are highly recommended to allow analysis at various spatiotemporal scales and to support model comparisons across different spatial resolutions (see next section).

Although mostly implicit in each group's approach due to time constraints, a well-developed conceptual model is essential. This model should describe all hypothesised population and spatiotemporal dynamics, and be informed by both the initial data analyses and a comprehensive literature review (Goethel et al. 2024; Minte-Vera et al. 2024; Cheng et al. 2025). Studies on growth, genetics, biomass flux, fishery dynamics, and management history can guide the model's spatial complexity. Population drivers can be prioritised by importance (e.g., primary drivers that must be addressed versus secondary or hypothesised drivers that should be explored) and certainty (e.g., well-documented versus hypothesised) to help focus the model development process, especially when time constraints exist.

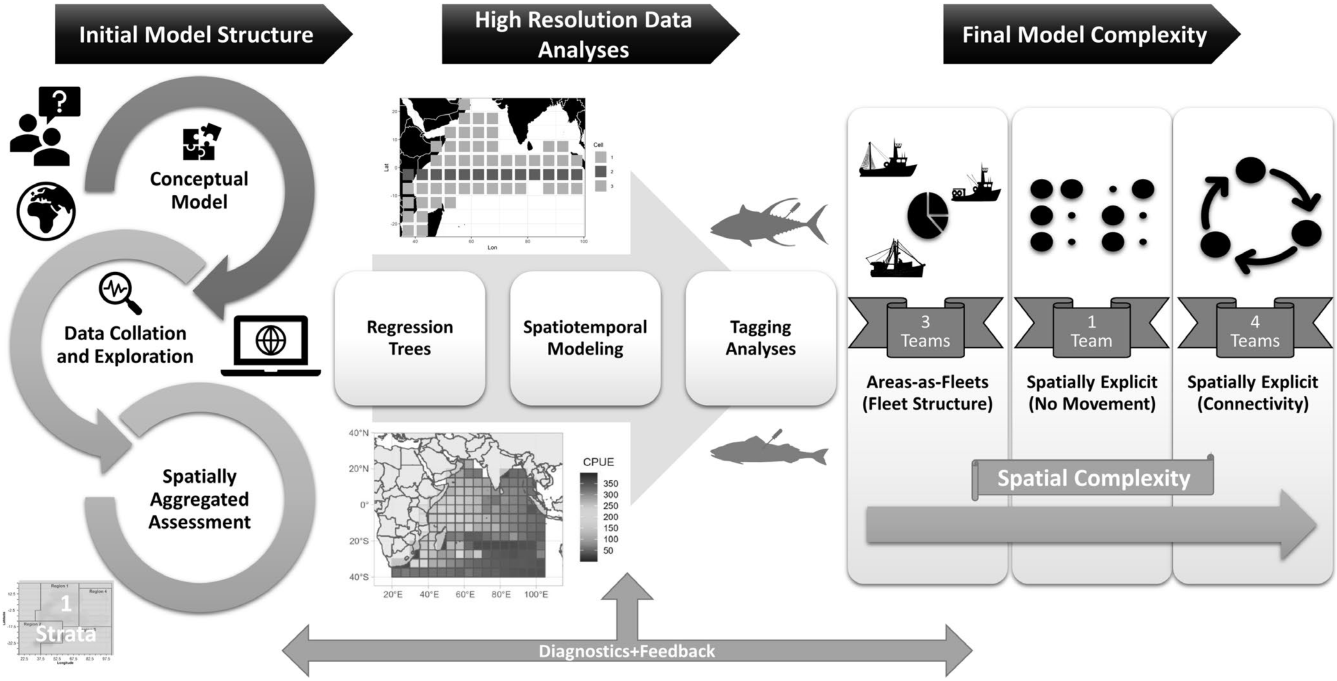

FIGURE 5 | Flow chart outlining a general approach for identifying appropriate stock assessment model complexity when confronting spatial structure. Final model complexity was achieved through the development of a conceptual model, data explorations, and implementation of a spatially aggregated model, which was then expanded to include spatial processes based on key hypotheses, high-resolution data analyses, model diagnostics, and feedback learning through iterative refinements. The presented steps are inclusive of the collective model development processes undertaken by the analyst teams.

4.2 | Identifying Model Complexity

Model development should proceed from the conceptual model, typically starting with a moderately complex approach that addresses the primary drivers of population dynamics (e.g., Cheng et al. 2025). Development should then be iterative and stepwise, adding or removing elements based on diagnostic feedback. Diagnostics allow you to evaluate whether: (1) critical elements are still missing and additional model features are needed, (2) elements are needed but data are lacking or specific processes are not understood, thus requiring future research, or (3) the model is performing adequately and no new elements are warranted. The process cycles through these steps until an adequate model is selected, often after the major data questions are addressed and future recommendations are identified.

A spatially aggregated model is often a useful starting point. These models benefit from increased sample sizes, simpler implementation with well-established modelling approaches, and, generally, high convergence rates. They also act as diagnostic tools in themselves: residual patterns and unrealistic population trajectories can indicate the need to add complexity. As complexity increases, earlier decision points should be revisited and may need to be adjusted due to interactions among processes (e.g., recruitment and movement; see next section).

Standard diagnostics can be used for both spatially aggregated and spatially explicit models (see Carvalho et al. 2021), but diagnostics that explicitly identify spatial process error remain a research priority. Common tools used in this experiment to refine spatial assessments included residual pattern evaluation, checking for convergence and parameter estimates near bounds, likelihood profiles or sensitivity runs to assess data conflicts, realism checks against expectations from the conceptual model (e.g., movement and spatial recruitment dynamics), self-tests (simulate data from an assessment model and rerun the assessment with the new data), fits to tagging data, and comparison of predicted and observed spatial biomass distributions. For example, analysts working with simulated YFT data used the non-homogenous spatial coverage of tag release and recapture data and tagging sub-model performance (e.g., model sensitivity to assumed tag mixing period and unreasonably large variance estimates) as justification to specifically alter model assumptions about movement. In some cases, this resulted in analysts aggregating tagging data across spatial areas, time scales, and life stages to improve tagging model performance and resulting movement estimates (Table 2). In other cases, it resulted in the removal of the tagging data altogether, including alternatively using a spatiotemporal CPUE model to implicitly capture seasonal movement. A few analyst teams also used less complex, biomass-based (no age-structure) surplus production models to diagnose and refine alternative recruitment parameterisations assumed in age-structured models. Analyst teams collectively identified further development of diagnostics to evaluate sufficient spatial model complexity as a research priority to better inform movement and productivity parameters and reduce confounding. Approaches such as examining profile likelihoods for movement with and without age variation, conducting cross-validation across spatial and temporal blocking schemes, and performing posterior predictive checks on regional tag flows and length- or age-compositions would complement this work.

When considering adding further complexity, analysts should assess whether the data are sufficient to support it. The level of model complexity should be determined by the quality, spatial resolution, and sampling distribution of the data (i.e., is there sufficient data to inform estimates within, and movement among, all regions?). Complexity should reflect biologically realistic expectations supported by conceptual models, which, ideally, would also match signals in the available data. However, determining whether model complexity is optimal likely requires testing models that are too complex. When added complexity does not improve results or convergence suffers (e.g., long run times, high gradients, poor Hessian behavior, or poor MCMC diagnostics), simpler models may be more appropriate. Multiple model spatial structures—both spatially aggregated and spatially explicit—should be maintained throughout the model-building process to aid comparisons among outputs, given that unique insight and support (validation) can be provided by each structure.

4.3 | Addressing Structural Uncertainty

Rather than identifying a single 'optimal' spatial structure, the goal should be to ensure that essential processes are appropriately captured. Models should adequately reflect removals (catch-at-age or length and by region), trends (CPUE or indices), and known biological processes. In a spatial context, this includes biomass flux and flows (e.g., movement and recruitment). Estimates of these processes are often highly correlated (Goethel et al. 2021), making deliberate and informed decision-making critical. Supporting evidence from data and literature should back each structural decision.

In some instances, particularly when data are limited or movement is generally ubiquitous, spatially aggregated or spatially implicit models may be sufficient. If the scale of the assessment aligns with the biological population unit and emigration is negligible, aggregated models can provide adequate management advice (Kerr et al. 2017; Cadrin 2020). Moreover, when spatial dynamics are primarily driven by fleet dynamics or differences in spatial availability to the primary fishing fleet, spatially implicit approaches such as areas-as-fleets (AAF) may be appropriate. As demonstrated in this study, AAF approaches combined with regression trees helped objectively identify fleet structure, while preprocessing data using spatiotemporal models accounted for other aspects of spatial variability in the data.

For spatially explicit models, the primary structural decisions relate to the number of regions, the fleet structure within each region, and the modelling of recruitment and movement across regions. These decision points are interconnected. For example, increasing complexity in regional structure may limit further complexity in recruitment and movement processes. As with any model, the complexity is limited by the data and prior information, given that greater partitioning (e.g., strata, fleets, ages, sexes) reduces sample sizes, often increases the number of parameters (e.g., recruitment, movement, and selectivity) that need to be estimated, and increases computation time. Techniques such as Bayesian priors, random effects, hierarchical models, parameter restrictions (e.g., Markovian movement), and parameter sharing can help reduce the effective number of parameters (Maunder et al. 2009; Thorson et al. 2021). However, one of the major remaining uncertainties is how best to incorporate tagging data, especially with uncertainty in appropriate tag mixing periods, to estimate movement and related parameters. Hybrid spatial modelling frameworks that bridge the extremes of spatial assessment modelling approaches (i.e., spatially stratified and spatiotemporal) could help address this uncertainty in the future by embedding high resolution spatiotemporal submodels (e.g., movement estimation from electronic tags) within a broader scale population model to help improve parameter estimates (e.g., movement; Thorson et al. 2021).

Finally, addressing structural uncertainty across spatial models may be best addressed by ensemble models or structural sensitivity analyses. In this experiment, many different model structural choices and assumptions were made by analysts using the same data, resulting in different estimated population dynamics (Goethel et al. 2024). The variability in model structure observed in this experiment suggests that complex population models, including spatial stock assessments, could gain robustness from ensemble modelling approaches, especially when extensive simulation testing cannot be conducted. Moving away from the 'best assessment' paradigm, ensembles allow integration of models with complementary strengths (Goethel, Omori, et al. 2023). Still, challenges remain in selecting ensemble members and defining appropriate weights for models with different spatial structures (Jardim et al. 2021; Adams et al. 2022). For instance, analysts must determine whether to include both spatially aggregated and spatially explicit models in the ensemble while also limiting the dimensions of primary modelling uncertainties to a tractable number. While these hurdles are surmountable (e.g., Adams et al. 2022), careful interpretation of ensemble results is essential.

5 | Conclusions

Our experiment did not identify a single optimal spatial structure or model building approach for stock assessments. However, it did emphasise the importance of deliberate and comprehensive data analysis to guide key structural decisions, while highlighting shared elements in the model development process. In particular, the co-development of models with alternative spatial structures proved essential to provide deeper insight into assessment performance and uncover potential spatial drivers. Using high-resolution data analyses was identified as a way to test for persistent spatial gradients before considering additional spatial structure and associated connectivity parameters. Although the need for a spatial model is context-specific—based on underlying spatial dynamics, data availability, and management needs—maintaining multiple model structures can help elucidate regional dynamics (e.g., spatially varying depletion) and validate model outputs (Li et al. 2025).

Adopting a reproducible and modular development process, such as the Transparent Assessment Framework (TAF; e.g., https://github.com/ices-taf), can facilitate model implementation and transition among assessment structures. TAF offers stock assessment workflow support and transparency by organising assessment data, methods, and results in an archival central hub. TAF also supports more flexible analysis of data at various spatiotemporal resolutions, thereby aiding decisions about model structure. In the coming years, increased research emphasis should also be placed on identifying best practices for data aggregation across assessment structures. Because all data are inherently spatial, some degree of aggregation is unavoidable regardless of model structure, and assumptions underlying how data are aggregated have important consequences (as observed in this experiment; Goethel et al. 2024). Moreover, diagnostics that can identify misspecification in relation to spatial processes, including how poor fit to aggregated data might be used to identify important spatial dynamics, remain underdeveloped and should be a future research priority.

Ultimately, overcoming institutional impediments to using spatial models in tactical decision-making requires improved communication across the science-policy divide. This includes open, frank discussions about whether a spatial model is needed, how it might inform management decision-making (e.g., distribution of regional catch limits based on regional abundance scaling parameters), and what trade-offs are involved. Similarly, improved collaboration and knowledge-sharing across RFMOs is needed to disseminate spatial modeling expertise and methods, via experiments such as this one and the publication of spatial assessments. There must also be clearer recognition that spatial assumptions are embedded in all assessments. A spatially aggregated model is not a neutral baseline—it represents an explicit decision that spatial processes are not an important driver of the dynamics of the resource. Thus, the assessment community should work together to understand the implications of ignoring spatial structure and to understand and address spatial drivers through improved data analyses, modeling approaches, and management strategy evaluation (MSE).

This experiment, particularly its blinded design and use of a high-resolution OM, highlighted a fundamental yet not widely acknowledged reality of stock assessment: all assessments, regardless of complexity or analyst experience, have the potential for bias. This insight suggests that increasing complexity alone is unlikely to resolve fundamental uncertainties. In addition to improving the stock assessments, priority should also be placed on developing and evaluating harvest control rules (HCRs) that are robust to both structural uncertainty and assessment bias (Evans et al. 2015). Future research should explore minimally complex, maximally robust HCRs, including empirical or quasi-empirical (e.g., linked to close kin mark-recapture estimates of absolute abundance) implementations (Goethel, Berger, and Cadrin 2023; Goethel, Omori, et al. 2023). Though challenging to implement, blinded simulation studies within high-resolution closed feedback (i.e., MSE) frameworks would be helpful to realistically evaluate assessment bias and support the design of resilient management strategies.

As stock assessments become increasingly multidisciplinary, spatial models provide a unique framework for integrating ecosystem, environmental, and socioeconomic drivers that act at various spatiotemporal dimensions (Goethel, Omori, et al. 2023). However, tradeoffs between parsimony and complexity must remain at the forefront of model development processes, particularly given the limitations of observed data for informing complex dynamics and associated parameters. Transparent communication about these limitations at the science-policy interface is critical, given rising expectations for model complexity and sophistication and the ever-present constraints on time, data, and resources.

Acknowledgements

This manuscript was the result of a global simulation experiment that began in May 2021. Many of the concepts were spurred by associated webinars, presentations, and discussions. We acknowledge the many participants for helping to contribute ideas to the manuscript. Finally, thanks to each of the reviewers that have helped improve various iterations of the manuscript, including Ole Shelton, Kristin Marshall, André Punt, and an anonymous reviewer. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect those of NOAA, the U.S. Department of Commerce, or any other author organisations.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request, given the storage space needed for these data.

Endnotes

[1] https://github.com/aaronmberger-nwfsc/Spatial-Assessment-Modeling-Workshop/blob/main/docs/PDFs/Analyst_guidance.pdf.

[2] https://github.com/aaronmberger-nwfsc/Spatial-Assessment-Modeling-Workshop.

[3] https://github.com/alistairdunn1/SPM.

[4] https://aaronmberger.nwfsc.github.io/Spatial-AssessmentModelingWorkshop/articles/OM_description_YFT.html.

[5] https://aaronmberger-nwfsc.github.io/Spatial-Assessment-Modeling-Workshop/articles/OM_Description_TOA.html.

References

Adams, G. D., K. K. Holsman, S. J. Barbeaux, et al. 2022. "An Ensemble Approach to Understand Predation Mortality for Groundfish in the Gulf of Alaska." *Fisheries Research* 251: 106303. https://doi.org/10.1016/j.fishres.2022.106303.

Berger, A. M., C. Barceló, D. R. Goethel, et al. 2024. "Synthesizing the Spatial Functionality of Contemporary Stock Assessment Software to Identify Future Needs for Next Generation Assessment Platforms." *Fisheries Research* 275: 107008. https://doi.org/10.1016/j.fishres.2024.107008.

Berger, A. M., D. R. Goethel, P. D. Lynch, et al. 2017. "Space Oddity: The Mission for Spatial Integration." *Canadian Journal of Fisheries and Aquatic Sciences* 74: 1698–1716. https://doi.org/10.1139/cjfas-2016-0335.

Bosley, K. M., A. M. Schueller, D. R. Goethel, et al. 2022. "Finding the Perfect Mismatch: Evaluating Misspecification of Population Structure Within Spatially Explicit Integrated Population Models." *Fish and Fisheries* 23: 294–315. https://doi.org/10.1111/faf.12584.

Brownie, C., D. R. Anderson, K. P. Burnham, and D. S. Robson. 1985. *Statistical Inference From Band Recovery Data: A Handbook*. 2nd ed. Resource Publication No. 156. U.S. Fish & Wildlife Service.

Bull, B., A. Dunn, A. McKenzie, et al. 2012. *CASAL (C++ Algorithmic Stock Assessment Laboratory) User Manual v2.30–2012/03/21 (NIWA Technical Report 135)*. National Institute of Water and Atmospheric Research.

Cadrin, S. X. 2020. "Defining Spatial Structure for Fishery Stock Assessment." *Fisheries Research* 221: 105397. https://doi.org/10.1016/j.fishres.2019.105397.

Carvalho, F., H. Winker, D. Courtney, et al. 2021. "A Cookbook for Using Model Diagnostics in Integrated Stock Assessments." *Fisheries Research* 240: 105959. https://doi.org/10.1016/j.fishres.2021.105959.

Cheng, M., C. Marsh, D. R. Goethel, et al. 2025. "Panmictic Panacea? Demonstrating Good Practices for Developing Spatial Stock Assessments Through Application to Alaska Sablefish (*Anoplopoma Fimbria*)." *Fish and Fisheries* 26: 825–847. https://doi.org/10.1111/faf.70002.

Deroba, J. J., D. S. Butterworth, R. D. Methot, et al. 2015. "Simulation Testing the Robustness of Stock Assessment Models to Error: Some Results From the ICES Strategic Initiative on Stock Assessment Methods." *ICES Journal of Marine Science* 72: 19–30. https://doi.org/10.1093/icesjms/fsu061.

Doonan, I., K. Large, A. Dunn, S. Rasmussen, C. Marsh, and S. Mormede. 2016. "Casal2: New Zealand's Integrated Population Modelling Tool." *Fisheries Research* 183: 498–505. https://doi.org/10.1016/j.fishres.2016.04.024.

Dunn, A. 2019. *Assessment Models for Antarctic Toothfish (Dissostichus mawsoni) in the Ross Sea Region to 2018/19*. WG-FSA-2019/08. CCAMLR.

Dunn, A., S. Hoyle, and S. Datta. 2020. *Development of Spatially Explicit Operating Models for Yellowfin Tuna Populations in the Indian Ocean*. IOTC-2020-WPT22(AS)-19. IOTC Working Party on Tropical Tunas 22.

Dunn, A., S. Rasmussen, and S. Mormede. 2020. *Spatial Population Model User Manual, SPM v2.0.3–2020-05-31*. Ocean Environmental Ltd.

Evans, K., J. N. Brown, A. S. Gupta, et al. 2015. "When 1+ 1 Can Be> 2: Uncertainties Compound When Simulating Climate, Fisheries and Marine Ecosystems." *Deep Sea Research Part II: Topical Studies in Oceanography* 113: 312–322. https://doi.org/10.1016/j.dsr2.2014.04.006.

Fournier, D. A., J. R. Sibert, J. Majkowski, and J. Hampton. 1990. "MULTIFAN: A Likelihood-Based Method for Estimating Growth Parameters and Age Composition From Multiple Length Frequency Data Sets Illustrated Using Data for Southern Bluefin Tuna." *Canadian Journal of Fisheries and Aquatic Sciences* 47: 301–317. https://doi.org/10.1139/f90-032.

Fu, D., G. Merino, A. Langley, and A. Ijurco. 2018. *Preliminary Indian Ocean Yellowfin Tuna Stock Assessment 1950–2017 (Stock Synthesis)*. IOTC 20th Working Party on Tropical Tunas.

Goethel, D. R., A. M. Berger, and S. X. Cadrin. 2023. "Spatial Awareness: Good Practices for Implementing the Continuum of Stock Assessment Approaches That Address Spatial Population Structure and Connectivity." *Fisheries Research* 264: 106703. https://doi.org/10.1016/j.fishres.2023.106703.

Goethel, D. R., A. M. Berger, S. D. Hoyle, et al. 2024. "Drivin' With Your Eyes Closed: Results From an International, Blinded Simulation Experiment to Evaluate Spatial Stock Assessments." *Fish and Fisheries* 25: faf.12819. https://doi.org/10.1111/faf.12819.

Goethel, D. R., K. M. Bosley, D. H. Hanselman, et al. 2019. "Exploring the Utility of Different Tag-Recovery Experimental Designs for Use in Spatially Explicit, Tag-Integrated Stock Assessment Models." *Fisheries Research* 219: 105320. https://doi.org/10.1016/j.fishres.2019.105320.

Goethel, D. R., K. M. Bosley, B. J. Langseth, et al. 2021. "Where Do You Think You're Going? Accounting for Ontogenetic and Climate-Induced Movement in Spatially Stratified Integrated Population Assessment Models." *Fish and Fisheries* 22: 141–160. https://doi.org/10.1111/faf.12510.

Goethel, D. R., K. L. Omori, A. E. Punt, et al. 2023. "Oceans of Plenty? Challenges, Advancements, and Future Directions for the Provision of

Evidence-Based Fisheries Management Advice." *Reviews in Fish Biology and Fisheries* 33: 375–410. https://doi.org/10.1007/s11160-022-09726-7.

Grüss, A., J. A. Devine, and S. J. Parker. 2021. "Characterisation of the Toothfish Fishery in the Ross Sea Region Through 2020/21." WG-FSA-2021/24. CCAMLR, Hobart, Australia, 37.

Guan, W., J. Wu, and S. Tian. 2019. "Evaluation of the Performance of Alternative Assessment Configurations to Account for the Spatial Heterogeneity in Age-Structure: A Simulation Study Based on Indian Ocean Albacore Tuna." *Acta Oceanologica Sinica* 38: 9–19. https://doi.org/10.1007/s13131-019-1485-4.

Hanchet, S. M., G. J. Rickard, J. M. Fenaughty, A. Dunn, and M. J. Williams. 2008. "A Hypothetical Life Cycle for Antarctic Toothfish (*Dissostichus mawsoni*) in the Ross Sea Region." *CCAMLR Science* 15: 35–53.

Hoyle, S. D., and A. D. Langley. 2020. "Scaling Factors for Multi-Region Stock Assessments, With an Application to Indian Ocean Tropical Tunas." *Fisheries Research* 228: 105586.

Jardim, E., M. Azevedo, J. Brodziak, et al. 2021. "Operationalizing Ensemble Models for Scientific Advice to Fisheries Management." *ICES Journal of Marine Science* 78: 1209–1216. https://doi.org/10.1093/icesjms/fsab010.

Kerr, L. A., N. T. Hintzen, S. X. Cadrin, et al. 2017. "Lessons Learned From Practical Approaches to Reconcile Mismatches Between Biological Population Structure and Stock Units of Marine Fish." *ICES Journal of Marine Science* 74: 1708–1722. https://doi.org/10.1093/icesjms/fsw188.

Lee, H.-H., K. Piner, M. Maunder, I. Taylor, and R. Methot. 2017. "Evaluation of Alternative Modelling Approaches to Account for Spatial Effects due to Age-Based Movement." *Canadian Journal of Fisheries and Aquatic Sciences* 74: 1832–1844. https://doi.org/10.1139/cjfas-2016-0294.

Lennert-Cody, C. E., J. Lopez, and M. N. Maunder. 2023. "An Automatic Purse-Seine Set Type Classification Algorithm to Inform Tropical Tuna Management." *Fisheries Research* 262: 106644.

Li, C., J. J. Deroba, A. M. Berger, et al. 2025. "Random Effects on Numbers-At-Age Transitions Implicitly Account for Movement Dynamics and Improve Stock Assessment and Management." *Canadian Journal of Fisheries and Aquatic Sciences*. https://doi.org/10.1139/cjfas-2025-0092.

Li, Y., J. R. Bence, and T. O. Brenden. 2015. "An Evaluation of Alternative Assessment Approaches for Intermixing Fish Populations: A Case Study With Great Lakes Lake Whitefish." *ICES Journal of Marine Science* 72: 70–81.

Lincoln, F. C. 1930. "Calculating Waterfowl Abundance on the Basis of Banding Returns." *United States Department of Agriculture Circular* 118: 1–4.

Maunder, M. N., H. J. Skaug, D. A. Fournier, and S. D. Hoyle. 2009. "Comparison of Fixed Effect, Random Effect, and Hierarchical Bayes Estimators for Mark Recapture Data Using AD Model Builder." In *Modeling Demographic Processes in Marked Populations*, edited by D. L. Thomson, E. G. Cooch, and M. J. Conroy, 917–946. Springer.

Maunder, M. N., H. Xu, and C. E. Lennert-Cody. 2022. *Developing Fishery Definitions for the Skipjack Tuna Stock Assessment in the EPO*. Inter-American Tropical Tuna Commission Report, SAC-13 INF-I. https://www.iattc.org/GetAttachment/855f6c5c-f2b1-4802-a713-8eaa9fe5d22d/SAC-13-INF-I_Developing-fishery-definitions-for-SKJ-stock-assessment-in-the-EPO.pdf.

McGilliard, C. R., A. E. Punt, R. D. Methot, and R. Hilborn. 2015. "Accounting for Marine Reserves Using Spatial Stock Assessments." *Canadian Journal of Fisheries and Aquatic Sciences* 72: 262–280. https://doi.org/10.1139/cjfas-2013-0364.

Methot, R. D., and C. R. Wetzel. 2013. "Stock Synthesis: A Biological and Statistical Framework for Fish Stock Assessment and Fishery Management." *Fisheries Research* 142: 86–99. https://doi.org/10.1016/j.fishres.2012.10.012.

Minte-Vera, C. V., M. N. Maunder, A. Aires-da-Silva, et al. 2024. "The Use of Conceptual Models to Structure Stock Assessments: A Tool for Collaboration and for "Modelling What to Model"." *Fisheries Research* 279: 10735.

Mormede, S., A. Dunn, S. J. Parker, and S. M. Hanchet. 2014. "Spatially Explicit Population Dynamics Models for Antarctic Toothfish in the Ross Sea Region." *CCAMLR Science* 21: 19–37.

National Research Council. 1998. *Improving Fish Stock Assessments*. National Academies Press. https://doi.org/10.17226/5951.

Petersen, C. G. J. 1896. "The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea." *Report of the Danish Biological Station* 6: 5–84.

Pittman, S. J., K. L. Yates, P. J. Bouchet, et al. 2021. "Seascape Ecology: Identifying Research Priorities for an Emerging Ocean Sustainability Science." *Marine Ecology Progress Series* 663: 1–29.

Punt, A. E. 2019a. "Modelling Recruitment in a Spatial Context: A Review of Current Approaches, Simulation Evaluation of Options, and Suggestions for Best Practices." *Fisheries Research* 217: 140–155.

Punt, A. E. 2019b. "Spatial Stock Assessment Methods: A Viewpoint on Current Issues and Assumptions." *Fisheries Research* 213: 132–143.

Punt, A. E., C. M. Dichmont, N. A. Dowling, et al. 2025. "Identifying Capacity Limitations and Training Needs Using a Stock Assessment Game." *Fisheries Research* 284: 107319. https://doi.org/10.1016/j.fishres.2025.107319.

Rea, A. W., A. W. Rea, and W. R. Munns. 2017. "The Value of Nature: Economic, Intrinsic, or Both?" *Integrated Environmental Assessment and Management* 13: 953–955.

Thorson, J. T., S. J. Barbeaux, D. R. Goethel, et al. 2021. "Estimating Fine-Scale Movement Rates and Habitat Preferences Using Multiple Data Sources." *Fish and Fisheries* 22: 1359–1376.

van den Burg, S. W. K., J. Aguilar-Manjarrez, J. Jenness, and M. Torrie. 2019. "Assessment of the Geographical Potential for Co-Use of Marine Space, Based on Operational Boundaries for Blue Growth Sectors." *Marine Policy* 100: 43–57.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Appendix S1:** faf70048-sup-0001-AppendixS1.docx.