



# 'Drivin' with your eyes closed': Results from an international, blinded simulation experiment to evaluate spatial stock assessments

Daniel R. Goethel<sup>1</sup>  | Aaron M. Berger<sup>2</sup>  | Simon D. Hoyle<sup>3</sup>  | Patrick D. Lynch<sup>4</sup>  |  
 Caren Barceló<sup>5</sup>  | Jonathan Deroba<sup>6</sup>  | Nicholas D. Ducharme-Barth<sup>7</sup>  |  
 Alistair Dunn<sup>8</sup> | Dan Fu<sup>9</sup> | Francisco Izquierdo<sup>10</sup>  | Craig Marsh<sup>11</sup> | Haikun Xu<sup>12</sup>  |  
 Giancarlo M. Correa<sup>13</sup>  | Brian J. Langseth<sup>14</sup>  | Mark N. Maunder<sup>12</sup> |  
 Jeremy McKenzie<sup>11</sup> | Richard D. Methot<sup>15</sup> | Matthew T. Vincent<sup>16</sup>  | Teresa A'mar<sup>17</sup>  |  
 Massimiliano Cardinale<sup>18</sup>  | Marta Cousido-Rocha<sup>10</sup>  | Nick Davies<sup>19</sup> |  
 John Hampton<sup>20</sup>  | Carolina Minte-Vera<sup>12</sup>  | Agurtzane Urtizberea<sup>21</sup> 

## Correspondence

Daniel R. Goethel, NOAA, Alaska Fisheries Science Center, 17109 Point Lena Loop Road, Juneau, AK 99801, USA.  
 Email: [daniel.goethel@noaa.gov](mailto:daniel.goethel@noaa.gov)

## Abstract

Spatial models enable understanding potential redistribution of marine resources associated with ecosystem drivers and climate change. Stock assessment platforms can incorporate spatial processes, but have not been widely implemented or simulated. To address this research gap, an international simulation experiment was organized. The study design was blinded to replicate uncertainty similar to a real-world stock assessment process, and a data-conditioned, high-resolution operating model (OM) was used to emulate the spatial dynamics and data for Indian Ocean yellowfin tuna (*Thunnus albacares*). Six analyst groups developed both single-region and spatial stock assessment models using an assessment platform of their choice, and then applied each model to the simulated data. Results indicated that across all spatial structures and platforms, assessments were able to adequately recreate the population trends from the OM. Additionally, spatial models were able to estimate regional population trends that generally reflected the true dynamics from the OM, particularly for the regions with higher biomass and fishing pressure. However, a consistent population biomass scaling pattern emerged, where spatial models estimated higher population scale than single-region models within a given assessment platform. Balancing parsimony and complexity trade-offs were difficult, but adequate complexity in spatial parametrizations (e.g., allowing time- and age-variation in movement and appropriate tag mixing periods) was critical to model performance. We recommend

For Affiliation refer page on 487

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Fish and Fisheries* published by John Wiley & Sons Ltd. This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA.

expanded use of high-resolution OMs and blinded studies, given their ability to portray realistic performance of assessment models. Moreover, increased support for international simulation experiments is warranted to facilitate dissemination of methodology across organizations.

#### KEYWORDS

fisheries management, mark-recapture, simulation, spatial ecology, spatial stock assessment, tag-integrated models

## 1 | INTRODUCTION

Sustainable exploitation of living marine resources requires scientifically informed management frameworks (Goethel, Omori, et al., 2023). In many instances, a stock assessment model is the primary tool for determining total allowable catch (TAC) quotas, and assessments have been widely touted for helping to rebuild many fish populations globally (Hilborn, 2012; Melnychuk et al., 2017). By simplifying real-world dynamics, stock assessments attempt to balance complexity and parsimony to adequately estimate temporal trends in abundance and current population status (Methot, 2009; Quinn & Deriso, 1999). However, determining adequate performance of an assessment application is difficult when the underlying truth is unknown. Thus, simulation modelling can be a useful tool for developing an understanding of assessment (i.e., estimation model, EM) robustness, because the true operating model (OM) dynamics (e.g., the coupled biological, fishery, and management system) are known (Deroba et al., 2015; ICES, 1993; Li et al., 2021).

With improved scientific understanding of the spatiotemporal nature of ecosystem drivers and the redistribution of marine species due to climate change, there has been an increase in the development of spatial assessment models to better represent spatial processes (Berger et al., 2017; Punt, 2019a, 2019b). Although spatially explicit simulation tools have demonstrated that spatial assessments are typically more robust than single-region or spatially implicit (i.e., areas-as-fleets, AAF) approaches when spatial dynamics are present, there remains ambiguity as to the conditions that necessitate implementing a spatial assessment (Bosley et al., 2022; Guan et al., 2019; McGilliard et al., 2015; Punt et al., 2017). Moreover, operational application of spatial assessments tends to be focused on highly mobile, wide-ranging large pelagic species (i.e., tunas; Punt, 2019a). Knowledge and dissemination of useful approaches for implementing spatial assessments tends to remain within associated regional fishery management organizations (RFMOs) that assess and manage these species (Goethel, Berger, et al., 2023). Thus, the assessment discipline would benefit from communication across RFMOs (e.g., through collaborative simulation experiments; Deroba et al., 2015; ICES, 1993; NRC, 1998) to aid dissemination of evolving spatial methodology.

However, designing and implementing simulation experiments to adequately portray the potential performance of an assessment model under real-world conditions is complicated, especially when

1.	INTRODUCTION	472
2.	METHODS	474
2.1.	Simulation experimental design	475
2.2.	Operating model	475
2.2.1.	Spatial dynamics	475
2.2.2.	Model conditioning	479
2.2.3.	Simulation	479
2.2.4.	Data aggregation and inputs	480
2.3.	Estimation models	480
2.4.	Model evaluation	481
3.	RESULTS	481
3.1.	Operating model dynamics	481
3.2.	Comparison across assessment spatial structures	481
4.	DISCUSSION	484
4.1.	Implications and potential drivers of estimation model performance	486
4.2.	Recommendations and future directions for collaborative, international simulations	486
4.3.	Conclusions	487
	ACKNOWLEDGMENTS	488
	DATA AVAILABILITY STATEMENT	488
	REFERENCES	488

undertaken as part of an international collaboration. For instance, there is an array of experimental design considerations to be addressed in a simulation experiment, including (Table 1): the number of estimation platforms to compare, the number of analysts to integrate in the experiment, the degree to which analysts should be informed of the underlying OM truth, the spatiotemporal resolution of the OM, how to condition the OM, the source and types of process error to include, the number of OM scenarios to simulate, and the extent of OM replication to undertake. Many decisions depend on the goals of the simulation experiment because strong trade-offs exist (see Table 1). For instance, when testing spatial assessments, key considerations include the resolution of the OM and the sources

TABLE 1 Major design considerations for simulation experiments that explore the performance of stock assessment models, including the settings chosen for the current experiment.

Design options	Alternatives	Benefits	Examples	Experiment setting
Number of estimation platforms	Single	Identify robustness of a given platform or determine robust parametrization for given OM scenarios	Goethel et al. (2021)	
	Multiple	Compare performance (e.g., given a common parametrization) across platforms or determine how unique model capabilities impact robustness	Li et al. (2021)	X
Operating model conditioning	Generic	Make generalizations based on common species or fishery characteristics	Goethel et al. (2019)	
	Data/species conditioned	Identify robust EMs for a given management application or need	Punt et al. (2017)	X
Process error source	Known misspecification (OM and EM frameworks match)	Diagnose impacts of a specific and known misspecification between OM and EM and identify best EM parametrization to deal with given process error	Bosley et al. (2022)	
	Misspecification not necessarily known (OM uses a unique framework from EM)	Develop a more realistic OM that better emulates real world data and dynamics where underlying population processes or sources of process error may not be completely replicated by EM	Marsh (2022)	X
Operating model resolution	Low resolution (non-spatial and/or cohort based)	Isolate potential drivers of bias by focusing on specific sources of error	Li et al. (2021)	
	High resolution (spatially explicit and/or agent based)	Develop more realistic OM that better emulates real world complexities	McGilliard et al. (2015)	X
Scenarios	Single	Reduce model run times and allow focused analysis on a single scenario that best emulates major sources of potential estimation error	Dunn, Hoyle, et al. (2020)	X
	Multiple	Explore robustness across an array of potential scenarios to allow generalization of conclusions	Bosley et al. (2022)	
Replication	Single	Reduce model run time	NRC (1998)	
	Multiple	Allow analysis of statistical significance across replicates	Punt et al. (2017)	X
Number of analysts	Single analyst or analyst group	Retain control across OM and EM runs to develop a deeper understanding of bias sources, robust parametrizations, and enable development of common parametrizations across platforms	Li et al. (2021)	
	Multiple analysts or analyst groups	Integrate wider expertise (e.g., across multiple platforms), encourage communication, and enable blind study designs	ICES (1993)	X
OM truth	Known	Develop robust EM parametrizations based on observed bias and implement alternate OM scenarios to address emerging issues	Goethel et al. (2021)	
	Blind	Emulate end-to-end model development processes to better understand how EMs would likely perform in real world scenarios	Deroba et al. (2015)	X

of process error. The use of high resolution OMs, which maintain an underlying structure that is independent of any given EM, can help produce more realistic levels of process and observation error, but deconstructing the sources of assessment misspecification can be more difficult (Fisch et al., 2021; Marsh, 2022; McGilliard et al., 2015; Saul et al., 2020).

Moreover, implementing blind study designs allows replication of the entire assessment development process (e.g., from data exploration to model diagnostics) with real-world uncertainty. However, implementing a blind study design is challenging due to the need to have multiple analyst teams (i.e., an organizing team to develop the OM and an analyst team to implement the EM without knowledge of the OM truth). Thus, to implement a blind simulation study, a collaborative approach is warranted. Unique logistical challenges then arise given the level of organization needed, the large time commitment required from each analyst group as well as the organizers, and the necessity for buy-in from national and international agencies to ensure support both financially and in terms of analyst time commitments (Deroba et al., 2015; NRC, 1998). Despite the challenges, integrating multiple analyst teams expands the number of platforms tested, increases the expertise (e.g., by incorporating experts for each platform), encourages communication and dissemination of methodology, and allows for the implementation of blind study designs (e.g., ICES, 1993; NRC, 1998).

Because the development of most spatial assessment applications remains insular within RFMOs, there is an inherent need to disseminate and share good practices in modelling spatial dynamics. Moreover, no study has attempted to compare how spatial dynamics are treated across assessment platforms or what the implications are for specific assumptions regarding spatial dynamics. To fill this research gap, researchers from the United States' National Oceanic and Atmospheric Administration (NOAA) and New Zealand's National Institute of Water and Atmospheric Research (NIWA)

organized an international simulation experiment to compare spatial assessment platforms. The goal of the experiment was to summarize the state of the science on developing spatially explicit stock assessment models, disseminate spatial model development methodologies and good practices across RFMOs, and to better understand the relative performance of single-region and spatial models when confronted with complex spatial structure and simulated data that reflected real-world applications. We summarize the primary findings from the blinded simulation approach by comparing results from single-region and spatial assessment models developed in an array of assessment platforms that are used globally. The experiment elucidated important differences in estimation performance between spatial and single-region models, which may not be observed when OM and EM structures are consistent (i.e., if a high-resolution OM had not been utilized). This article also assimilates feedback on the collaborative simulation process to provide recommendations for future multi-national simulation experiments.

## 2 | METHODS

For this study, a blinded, spatially explicit, cross-platform simulation experiment was implemented utilizing a high-resolution OM conditioned on Indian Ocean yellowfin tuna (*Thunnus albacares*; see Figure 1 and Table 1 for an outline of the experimental design and simulation settings). An international team of analysts with expertise using each of the major generalized stock assessment platforms with spatial capabilities was assembled. The study began in 2019 and concluded in early 2023. Results from various aspects of the study are presented across four journal articles:

- Summarizing the spatial capabilities of current generalized assessment platforms, which follows recommendations by Li et al. (2021)

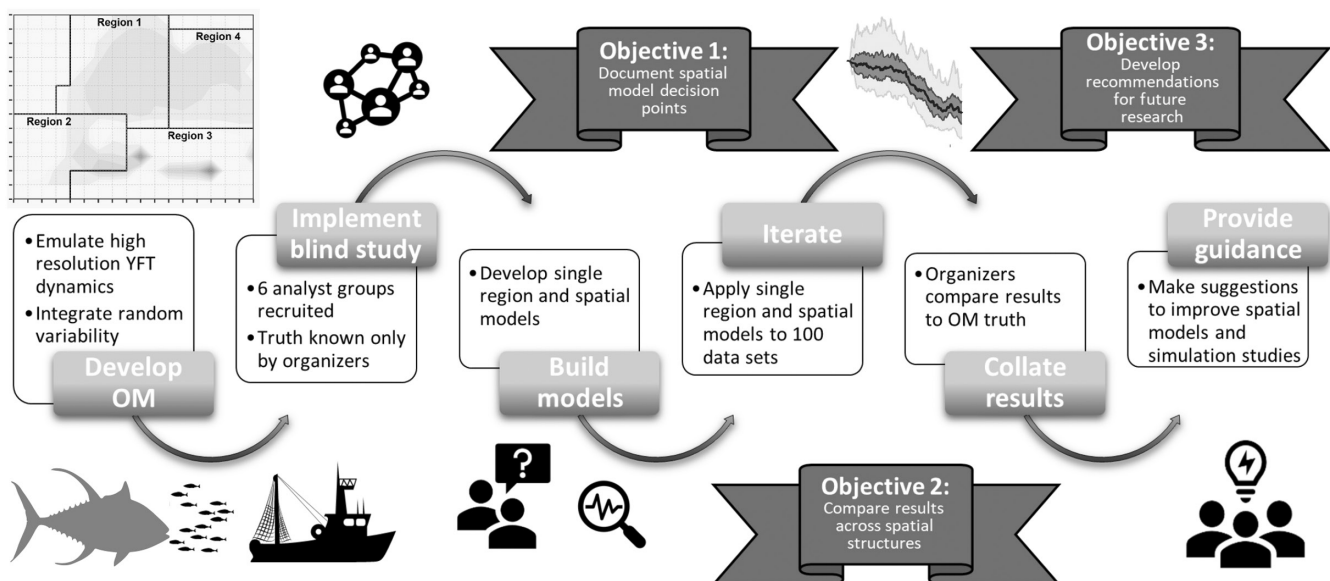


FIGURE 1 The simulation experimental design. The current manuscript focuses on objectives two and three.



that cross-platform simulation evaluations should begin with a comparison of source code and model features (Berger et al., [In press](#)).

- Description and demonstration of the spatially explicit, high-resolution, data-conditioned OM (Hoyle et al., [2024](#)).
- Results from the model development process for the spatial assessments produced by each analyst group (Berger et al., [2024](#)).
- Comparison of model outputs across spatial structures and lessons learned from implementing the simulation experiment (this article).

## 2.1 | Simulation experimental design

A centralized group organized the overall experiment, including the data simulation and dissemination to each analyst group, addressing any concerns or questions, hosting webinars and meetings, and collating results. A multi-national mixture of participants was convened and self-organized into six stock assessment analyst groups. The organizing group developed the OM and simulated data, which was then provided to the analyst groups. Each group developed a single-region and spatially explicit stock assessment in an assessment platform of their choice, and emulated the process that would be undertaken in a real-world assessment. They were requested to document the model building process, including data analysis, parametrization choices, model diagnostics, and model validation. Models were then applied to 100 replicates of the simulated data, and results were submitted to the organizing group for collation (see [Figure 1](#) for a summary of the study design; all material utilized in the experiment is available from the experiment GitHub site: <https://github.com/aaronmberger-nwfs/Spatial-Assessment-Modeling-Workshop>).

Because analyst groups may have had varying levels of prior knowledge regarding underlying population dynamics, cross-platform comparisons were not an explicit goal of the experiment. Thus, the simulation experiment was *not* designed to compare performance among platforms nor to identify a 'best' platform. Instead, the goal of the study was to provide insight into spatial model development with comparisons of model outputs intended to improve understanding of how spatial structure assumptions impacted model performance and the potential implications of ignoring spatial dynamics.

## 2.2 | Operating model

The Indian Ocean yellowfin tuna case study was chosen because it represented a high-profile species for which complex spatial dynamics are known to be a key source of uncertainty for assessment and management. Additionally, yellowfin tuna is a highly migratory species of considerable importance to numerous RFMOs, and the population dynamics reflect those of many worldwide tuna populations making findings generalizable (to a degree). The OM was developed to emulate the spatially explicit dynamics of yellowfin

tuna in the Indian Ocean by first conditioning the model on empirical data, observed biology, and expert judgment informing important ecological processes. Initial parametrization was based on the most recent spatial assessment model for Indian Ocean yellowfin tuna, which utilized a four-region Stock Synthesis 3 (SS3) model (Fu et al., [2018](#)).

The OM used the Spatial Population Model (SPM; see Dunn, Rasmussen, et al., [2020](#), for the user manual and underlying structural equations), which is a high-resolution, spatially explicit, quasi-estimation model that can be conditioned on observed data. Initial development of the data conditioned Indian Ocean yellowfin tuna operating model is described in Dunn, Hoyle, et al. ([2020](#)). The first iteration of the model was refined to meet the objectives of this study, optimize parametrization, and add sources of process error (i.e., stochasticity in cell-specific recruit apportionment and year-class strength). A complete description of the OM and data conditioning are provided in [Appendix S1](#) (with further descriptions provided in Hoyle et al., [2024](#)). The following highlights spatial drivers and data simulation pertinent to interpreting EM performance ([Table 2](#)).

### 2.2.1 | Spatial dynamics

The OM assumed a single population of yellowfin tuna that moved across and interacted within 221, 5° latitude × 5° longitude cells ([Figure 2](#)). Although fish in each cell underwent unique mortality and movement processes, biological parameters (i.e., maturity, growth, and natural mortality) were constant across all cells and externally derived from Fu et al. ([2018](#)). The model assumed a quarterly time step (i.e., 256 total time steps) and was both stage-structured (i.e., with two maturity partitions: immature and mature fish) and age-structured (ages 0–28+, in quarter ages, with the last age being a plus group). The SPM calculated abundance-at-age by cell and maturity stage based on forward calculations in a given time step from initial conditions (i.e., recruitment at age-1 and initial abundance-at-age in the first time step). Fish then moved among cells based on habitat preference functions and were subsequently removed due to fishing. Finally, a time step concluded with removals due to natural mortality and age- and maturity-transitions.

A global Beverton–Holt stock-recruit function with steepness of 0.8 was assumed where virgin (i.e., unfished) recruitment ( $R_0$ ) was specified based on Fu et al. ([2018](#)) to preserve scaling of the overall population. Recruitment occurred quarterly and was calculated as the product of the stock-recruit relationship (i.e., based on the total spawning stock biomass summed across all cells in the previous quarter), a time step-specific year-class strength multiplier, and the time step-specific apportionment layer that assigned recruitment to cells. Maturity-stage-specific movement rates were determined based on functions that integrated preference for sea surface temperature, chlorophyll-a concentrations, and distance among cells. The preference functions were probability density functions defining attraction to a given cell based on spatial attributes.

**TABLE 2** The general operating model (OM) settings used for the spatial population model (SPM) simulation of yellowfin tuna along with the major parametrization settings for the various estimation models (EMs). Note that spatially implicit areas-as-fleets models are abbreviated as 'AAF'.

Model type	Platform	Full name	Analyst team name	Model types implemented	Key parametrizations	Tagging data used	Convergence rate	Citation for platform
OM	SPM	Spatial Population Model	Organizers	Spatial (221 cells)	Global Beverton-Holt stock-recruit relationship with quarterly year class multiplier and apportionment to grid cell Movement based on preference functions (i.e., distance between cells, temperature, and chl-a) High-resolution, data conditioned OM with 7 fleets operating at different spatiotemporal dimensions	Yes, fit during data conditioning, then simulated at the cell level; tag mixing at the regional level was low, but higher for immature fish	-	Dunn, Rasmussen, et al. (2020)
EM	CASAL2	C++ Algorithmic Assessment Library (2nd Generation)	CASAL2	1-region	Full fleet structure aggregated to 1 region (7 fleets)	No	99%	Doonan et al. (2016)
				AAF	Full fleet structure for each region (i.e., selectivity estimated and catch/length composition data fit for each fleet and region combination)	No	94%	
				4-region	Full fleet structure fit at regional scale No movement	No	99%	
	MFCL	Multiple Length Frequency Analysis-Catch at Length	MFCL	1-region 4-region	Full fleet structure aggregated to 1 region (7 fleets) Full fleet structure by region with selectivity mirrored across regions Recruitment apportionment estimated for each region with deviations Movement estimated for each shared regional boundary (i.e., regions 1-2, 1-3, 1-4, 2-3, 3-4), with movement assumed to be age- and time-invariant but with deviations for quarter (i.e., movement parameters were estimated for each quarter, then held constant at that quarterly value for the entire time series)	Yes with 5 mixing periods Yes with 5 mixing periods	100% 100%	Kleiber et al. (2018)

TABLE 2 (Continued)

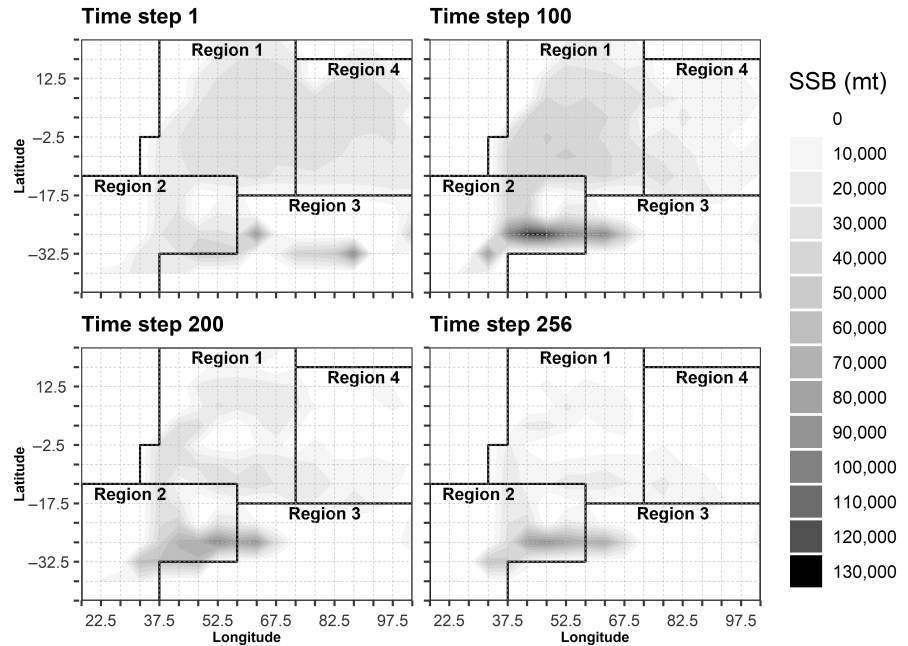
Model type	Platform	Full name	Analyst team name	Model types implemented	Key parametrizations	Tagging data used	Convergence rate	Citation for platform
SPASAM	SPASAM	Spatial Processes and Stock Assessment Methods	SPASAM	1-region	Full fleet structure aggregated to 1 region (7 fleets) Start model in time step 106	No	86%	Goethel et al. (2019, 2021)
				2-region	2 regions (1-2, 3-4) instead of 4 regions Fleets combined to reduce to 4 per region	Yes with complete mixing assumed (no latency period)	100%	
					Apportionment fixed at 99% to regions 1-2 and 1% to regions 3-4 Movement estimated among all regions, with movement estimated for 2 age groups (i.e., greater/less than age-9) in every other time step (i.e., 2 time step blocks)			
SS3	SS3	Stock Synthesis 3	SS3_A	1-region	Cell-specific CPUE analysed with spatiotemporal model to create CPUE index	Yes with 6 mixing periods and reporting rate estimated	95%	Method and Wetzel (2013)
				4-region	Full fleet structure aggregated to 1 region (7 fleets) Cell-specific CPUE analysed with spatiotemporal model to create CPUE index Full fleet structure by region with selectivity mirrored across regions for same fleets No recruitment in region 3, but apportionment deviations estimated for each time step Movement estimated among regions 1-2, 1-4, and 4-3, with movement estimated for two age groups (i.e., greater/less than age-16) and with no time-variation	Yes with 6 mixing periods and reporting rate estimated	70%	

(Continues)

TABLE 2 (Continued)

Model type	Platform	Full name	Analyst team name	Model types implemented	Key parametrizations	Tagging data used	Convergence rate	Citation for platform
			SS3_B	1-region	Full fleet structure aggregated to 1 region (7 fleets)	Yes with 4 mixing periods	100%	
				4-region	Apportionment estimated for each region without deviations	Yes with 4 mixing periods	100%	
					Movement estimated between regions 1–2, 1–4, and 3–4, with movement estimated for two age groups (i.e., greater/less than age-9) and with no time-variation			
			SS3_C	1-region	Cell-specific CPUE analysed with spatiotemporal model (VAST) to create CPUE index and associated length compositions	No	74%	
					Full fleet structure aggregated to 1 region (7 fleets)			
			AAF		Cell-specific CPUE analysed with spatiotemporal model (VAST) to create CPUE index and associated length compositions	No	87%	
					16 fleets based on region and fleet combinations, where purse seine fleet structure was identified using regression tree analysis			

**FIGURE 2** Distribution of spawning stock biomass for the representative simulation run (i.e., iteration four of 100) from select model time steps by operating model grid cell (dashed grey lines). The four regions utilized by the spatial assessment models (dark black lines) are also shown for reference. Note that cells not assigned to a region are not included in the operating model or estimation model dynamics (i.e., typically due to being land-based cells).



Fishing mortality was modelled as a cell-specific exploitation rate based on the real-world distribution and magnitude of catch from seven fishery fleets operating at various spatiotemporal scales. All fleets were assumed to have a double-normal selectivity function, aside from the primary longline fleet that was assumed to have a logistic selectivity function. Penalties on cell-specific catch levels were enabled to ensure that catch never exceeded biomass in a given quarter or cell. Length composition data were produced for each fishery in each cell and time step for which that fishery operated. The SPM is an age-based modelling framework, so the growth curve (with uncertainty) input from Fu et al. (2018) was internally converted to a distribution of length-at-age to produce length compositions. A longline fishery catch-per-unit effort (CPUE) index of relative abundance was also calculated.

Mark-recapture tagging data were included in the model to mimic the available data for yellowfin tuna (Fu et al., 2018). Tagged fish underwent an initial tag loss or mortality rate, but were then assumed to undergo the same dynamics as the untagged population with the addition of a chronic tag loss term (i.e., a certain proportion of tagged fish were removed due to tag loss). Tagged fish were assigned to a release cohort based on age of release, maturity stage, and release cell. The number of tag recaptures by time step, age, cell, and fishery fleet from a given release cohort were calculated as the product of fishery removals of tagged fish and the probability that a recaptured tag was reported (i.e., the fleet-specific reporting rate).

## 2.2.2 | Model conditioning

The OM was conditioned on empirical data using maximum likelihood estimation (MLE) to estimate realistic parameter values that were consistent with assumed parameter values. The spatially disaggregated data used for model conditioning were taken from the 2018

stock assessment (Fu et al., 2018) and binned to a model cell (see Appendix S1 for more details on available data and model conditioning). Data used for model conditioning included fleet-specific catch, longline fishery CPUE, fleet-specific length frequencies, environmental or habitat data (i.e., temperature, chlorophyll-a concentrations, and distance among cells), and tag releases and recaptures from the purse seine fleet (i.e., the only fleet that consistently reports tagged fish). The spatiotemporal coverage of tag releases was relatively limited with a total of 54,688 releases between 2005 and 2007, where most of the releases were of immature fish into Region 1 in the second and third quarters of 2006 (See Figure A2 in Appendix S1). The available tag recaptures for the purse seine fleet were fit by recapture cell and age-at-recapture assuming a reporting rate of 90%.

As noted,  $R_0$  was taken from Fu et al. (2018) and used to scale the model, while biological inputs were also taken from the recent assessment to ensure consistency (see Table A1 in Appendix S1). Similarly, year class strength multipliers were derived to ensure time step-specific recruitment generally followed the magnitude and trend from Fu et al. (2018). For conditioning, the recruitment apportionment layer was assumed to be time-invariant and based on the general distribution of juvenile fish observed in fishery length composition data (see Figure A1 in Appendix S1). The primary estimated parameters during the conditioning phase were selectivity for each fishery, parameters defining the habitat and ecosystem preference functions for each maturity partition (i.e., to define movement among cells), and a spatio-temporal invariant CPUE catchability coefficient. Parameter estimates and model diagnostics are reported in Dunn, Hoyle, et al. (2020).

## 2.2.3 | Simulation

Once the SPM was conditioned on the yellowfin tuna data, it was then run as a simulator from the MLE point estimates of the

estimated parameters. Thus, a single OM scenario was developed to emulate yellowfin tuna dynamics, which was designed to provide as realistic dynamics and data availability as possible. However, to ensure realistic simulation of uncertainty in primary population dynamics and observed data, observation and process error were included. Finally, to ensure adequate representation of assessment model performance, 100 simulation model replicates were produced to encapsulate variation and provide realistic estimates of uncertainty. Each model replicate was seeded with a different random number to ensure a unique realized population trajectory (e.g., due to a different set of year class multipliers and recruit apportionment layers) and set of data observations.

Similarly, observation error was included for each of the data sources at the cell-level (excluding catch and tag releases) based on the assumed error structure and associated variance as used to fit the data in the conditioning phase. Catch was assumed to be known without error and reflected the observed cell- and fleet-specific catch. CPUE data were simulated assuming log-normal deviations with a coefficient of variation of 0.2. Length composition data (i.e., converted from age compositions based on the input distribution of age-at-length) from each fishery were generated assuming a multinomial distribution with an effective sample size (ESS) of 5 for each quarter, cell, and fishery. The cell-specific ESS was chosen to produce compositional data with distributions that reflected the empirical data for yellowfin tuna, including patchiness in both length bins and spatial cells. Finally, tag release events matched the real-world tagging data described in the previous section. Simulated recaptures were based on the number of tags available in each quarter and cell along with the associated exploitation rate in the purse seine fishery. The final number of simulated recaptures was then determined assuming a binomial process of tag detection and a reporting rate of 90%.

### 2.2.4 | Data aggregation and inputs

Each analyst group was provided a document summarizing yellowfin tuna biology, population dynamics (based on observed data), and true values of the biological inputs (i.e., maturity, growth, weight-length relationships, and natural mortality). The document summarized the type of background knowledge that an assessment analyst would be given for the development of a real-world assessment application. Although more information was available than in a real-world example (e.g., the true biological parameters), the underlying truth (e.g., population trajectories) were known only by the organizing group.

The simulated data were provided to experiment participants at three levels of aggregation: the fully disaggregated, cell-specific data; aggregated to four regions matching the spatial resolution of the current stock assessment utilized for management advice (Figure 2); and a single region. Analysts were expected to explore the disaggregated simulated data to develop hypotheses regarding

distribution, movement, and tagging dynamics, then apply assessment models at the one- and four-region scale (or any other resolution desired). Aside from CPUE, simulated data (i.e., catch, length frequencies, tag releases, and tag recaptures) were simply summed across cells within each region for the various aggregated data sets. For the CPUE data, aggregation to four regions utilized regional scaling following Hoyle and Langley (2020) to ensure that regional indices were reflective of associated regional abundance. For tagging data, no explicit information on tag mixing rates or movement dynamics were provided.

## 2.3 | Estimation models

There were a total of six analyst groups, which applied four different model platforms, including Stock Synthesis 3 (SS3), Multiple Length Frequency Analysis Catch-at-Length (MULTIFAN-CL or MFCL), the C++ Algorithmic Assessment Library--2nd Generation (CASAL2), and the bespoke Spatial Processes and Stock Assessment Methods (SPASAM) model. Three analyst teams utilized the SS3 platform, but each undertook a unique approach to model development. Aside from SPASAM, which is primarily a research tool, the other three platforms have been widely used worldwide for applied stock assessments. All groups provided a single-region and multi-region model except the SS3\_C team, which developed an AAF model but not a spatially explicit model (see below). All models were applied to all 100 replicates of the OM. The primary model settings are provided in Table 2 and unique approaches or parametrizations are summarized briefly below. However, please see Berger et al. (2024) for complete details on each model development process and final model settings. As a general note, all models used the true biological inputs from the OM, so any differences in performance were not due to spatial heterogeneity (or associated misspecification) in growth, maturity, or natural mortality.

Generally, most of the spatial models utilized similar parametrizations (e.g., all but one spatial model assumed four regions) with slight differences in terms of the regions to which recruitment was apportioned, the regions among which movement was estimated, and the number of tag mixing periods (i.e., the number of quarters during which tag recaptures were ignored or removed; Table 2). Therefore, differences in performance were likely to be more influenced by fundamental discrepancies in modelling approach rather than slight variability in parametrizations. For instance, the SS3\_A team implemented (and fit the resulting index from) a spatiotemporal standardization model on the spatially disaggregated, cell-specific CPUE data using the R package INLA, while the SS3\_C team used a similar approach with the Vector Autoregressive Spatiotemporal (VAST) model. The SS3\_C team also computed the associated longline fishery length compositions in VAST with compositional data weighted by the CPUE in each cell (i.e., before aggregating to a region), while selectivity was separately specified for the longline index and the longline fishery. In addition, the



length compositions for the purse-seine fisheries were weighted by the catch in each cell, which was an important difference from other groups' models where length compositions were weighted by the number of samples in a given cell (i.e., before aggregating to a region). Moreover, the SS3\_C team's AAF model utilized a novel regression tree approach (Lennert-Cody et al., 2010, 2013) to define the fleet structure of the purse-seine fisheries using the associated length composition data. The CASAL2 team also implemented an AAF model in addition to the full complement of model spatial structures. For the spatial model, the CASAL2 team was the only one to ignore movement and not integrate tagging data. The SPASAM group implemented a unique spatial configuration with a simplified structure that aggregated to two regions (i.e., regions 1–2 and 3–4 were combined). Additionally, the time series in both the single-region and spatial SPASAM model configurations was shortened to start in time step 106 based on performance and run times.

Following the experiment, the organizing team developed an array of alternate EM parametrizations to explore the impact of model assumptions on resulting performance to highlight future research avenues. Model descriptions and results pertaining to the exploratory runs are provided in Appendix S2 (see Table B1 in Appendix S2 for a description of each exploratory run).

## 2.4 | Model evaluation

Because comparison across model platforms was not an explicit goal of the simulation experiment, a relatively simple approach to model evaluation was undertaken. Convergence rate was used as an initial measure of model stability, which was defined as a model having a positive definite Hessian matrix and a maximum objective function gradient component less than a pre-specified cut-off value (e.g., 0.001; each analyst group utilized slightly different thresholds, but chosen values were unlikely to vary greatly or have a strong impact on interpretation of model performance). Only converged runs of each EM were included in the results. Percent relative error (i.e., bias) in key model outputs typically used for management advice (i.e., SSB and depletion) was used to make general comparison across spatial structures. Bias was calculated as the estimated value from the EM minus the true value from the OM, which was then divided by the true value and converted to a percentage. Visual comparisons of bias were undertaken by plotting the time series of median bias along with the 75% and 95% intervals across all the converged simulation iterations for each quantity of interest, including both aggregated (i.e., across all regions) and region-specific quantities for spatial models. SSB was calculated as the weight of mature females in metric tons (mt). Depletion represented the ratio of SSB in a given year divided by the SSB in the first year of the model time series (i.e., the denominator was static) for a given model and spatial scale of interest (i.e., for the SPASAM models the first model year was time step 106, instead of time step 1 as was the case for all other platforms).

## 3 | RESULTS

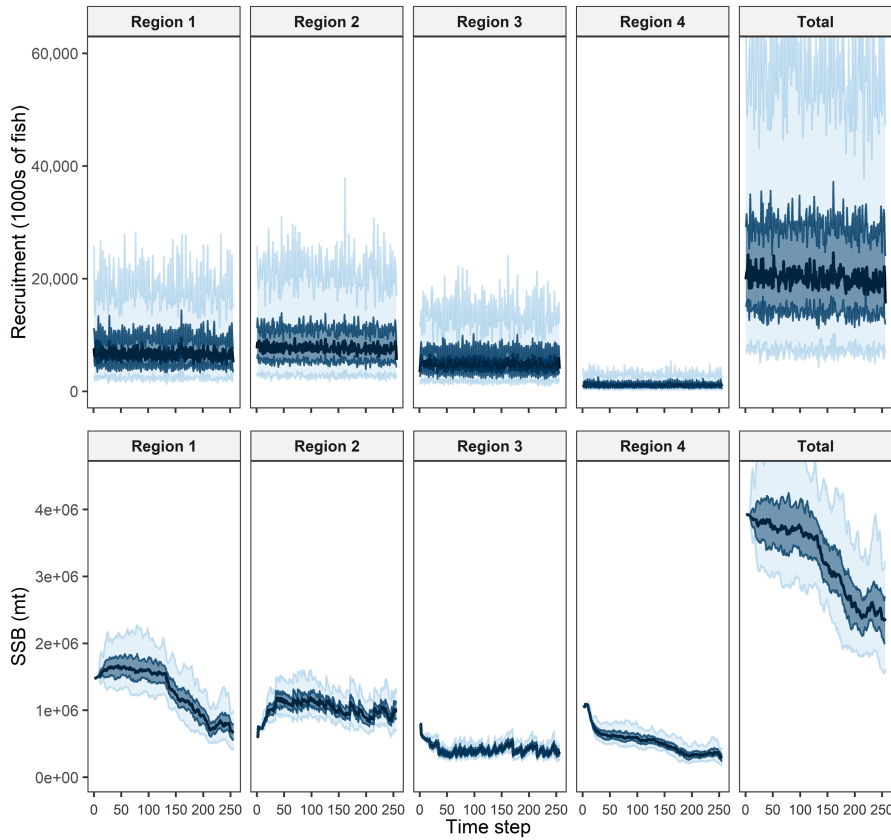
### 3.1 | Operating model dynamics

The full suite of parameter estimates and model diagnostics are reported in Dunn, Hoyle, et al. (2020) with detailed outputs provided in Appendix S1. The OM was generally able to emulate the most plausible population scale and trends (e.g., the most recent stock assessment; Fu et al., 2018, 2021) by region and across the entire population (Figure A9 in Appendix S1). The distribution of yellowfin tuna in the model was driven by the combined impacts of the spatial processes (i.e., recruitment location, movement based on environmental preference, and locations of high fishing pressure), which resulted in the dynamic distribution of cells with high population density (Figure 2). A large portion of the mature biomass began in the northern cells (e.g., regions one and four), but, due to moderate recruitment and the bulk of the fishing pressure being centred here, the SSB declined rapidly (Figures 2 and 3). Conversely, the mature biomass in the southern cells (e.g., region two) increased across the first 50 time steps before levelling off, likely driven by relatively higher recruitment and lower fishing pressure. Region three demonstrated a decline in SSB followed by a slight rebound due to low fishing pressure. In region four, the SSB trends downward for much of the time series due to moderate fishing and low recruitment. The biomass trend for the population generally mimics that of region one, while recruitment fluctuates with little trend across the time series (Figure 3). By the end of the time series, density had decreased considerably with the highest density areas located primarily in the southern parts of the domain, particularly in region two (Figure 2). Because mature yellowfin tuna tended to redistribute to areas with lower fishing pressure, there was potential for 'cryptic' biomass (i.e., unobservable by the CPUE abundance index and generally undisturbed by harvest) in region two and, to a lesser extent, region three.

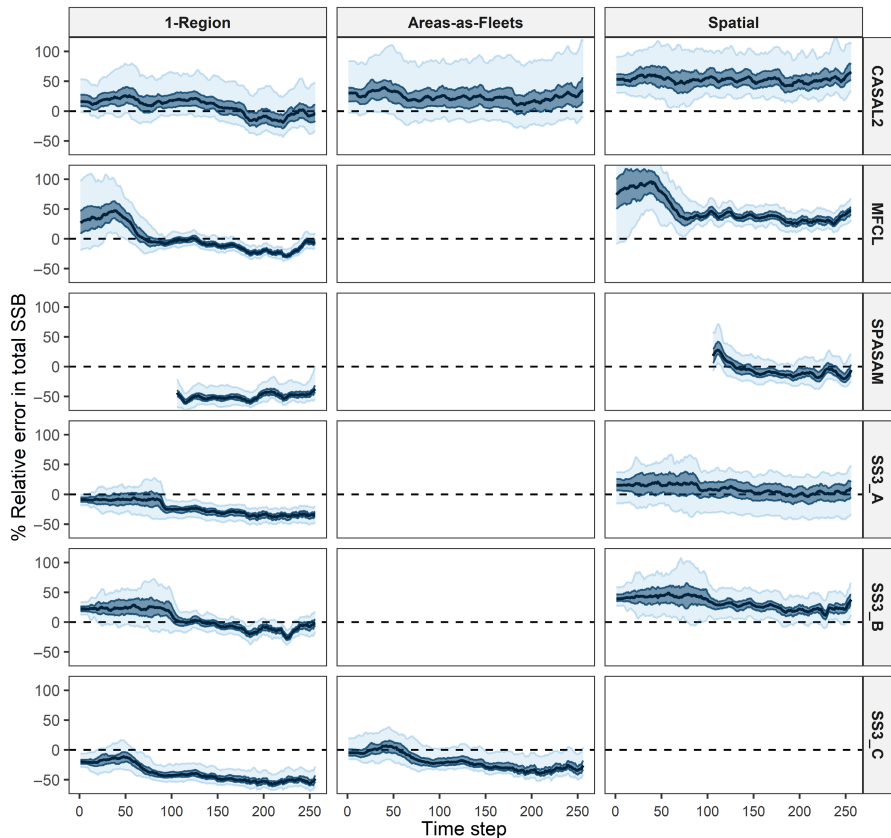
### 3.2 | Comparison across assessment spatial structures

Across platforms and spatial structures, convergence rates were satisfactory (i.e., greater than 80% except for two instances), indicating that models were generally stable and had likely converged to global solutions for the associated parametrization (Table 2). A slight pattern of lower convergence for single-region or spatially implicit (AAF) models compared to spatial models was present, but it was not consistent across all platforms. Analyst groups typically spent more time developing and analysing spatial models, which may have impacted the relative convergence rates.

The most prominent and consistent trend in results was that single-region models always estimated a lower population scale (i.e., SSB) relative to spatial models within a given platform (Figure 4). Moreover, the spatially implicit AAF models appeared to estimate population scales that were intermediate between single-region and



**FIGURE 3** Simulated regional (column) dynamics from the operating model, including recruitment (top row; 1000s of fish) and spawning stock biomass (bottom row; metric tons). The black line is the median, while the shaded regions represent the 95% (light blue) and 75% (dark blue) intervals across all 100 replicates of the operating model.



**FIGURE 4** Comparison of percent relative error in total (aggregated across regions for spatial models) spawning stock biomass across model spatial structure (columns; left is single-region models, centre is spatially implicit areas-as-fleets models, and right is spatial models) and assessment platform (row; see Table 2 for a description of model parametrizations). The true value from the operating model is represented by the dashed-line at zero. The solid dark line is the median percent relative error, while the shaded regions represent the 95% (light blue) and 75% (dark blue) intervals of percent relative error across all 100 simulation replicates. Only results from converged model runs are illustrated. Note that the SPASAM model started in time step 106.

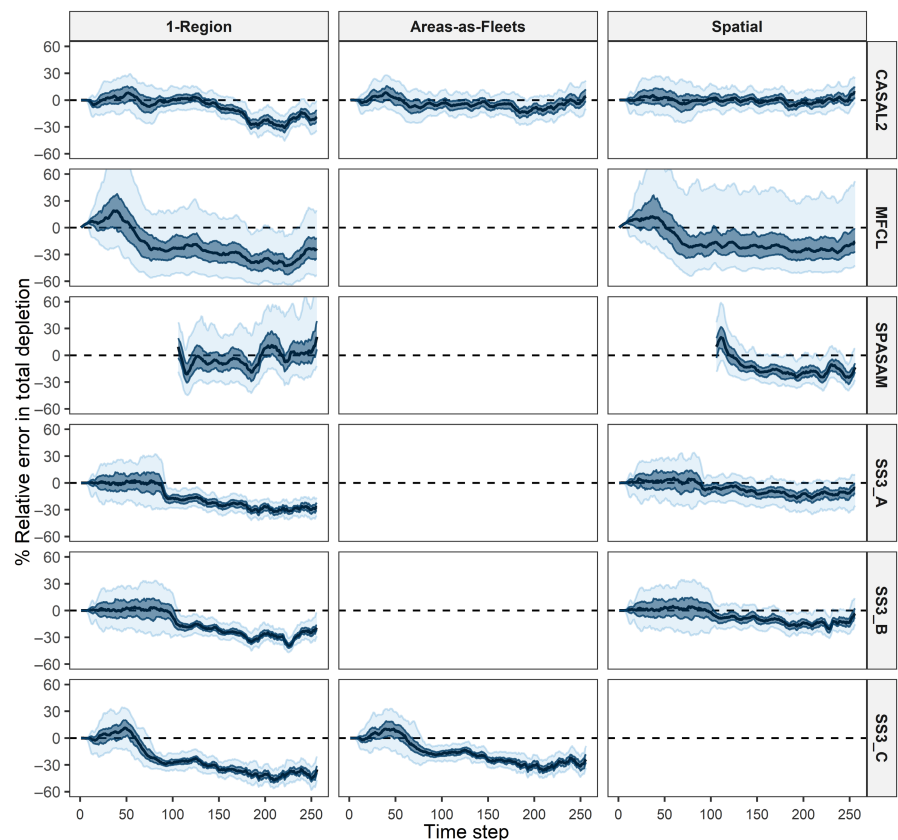
spatial models. However, there was only one platform (CASAL2) for which all three model types were implemented. The SS3\_C single-region and AAF models demonstrated a similar trend in population scaling as the corresponding CASAL2 models. All models were able to recreate the general declining trend in SSB across the time series (Figure C1 in Appendix S3).

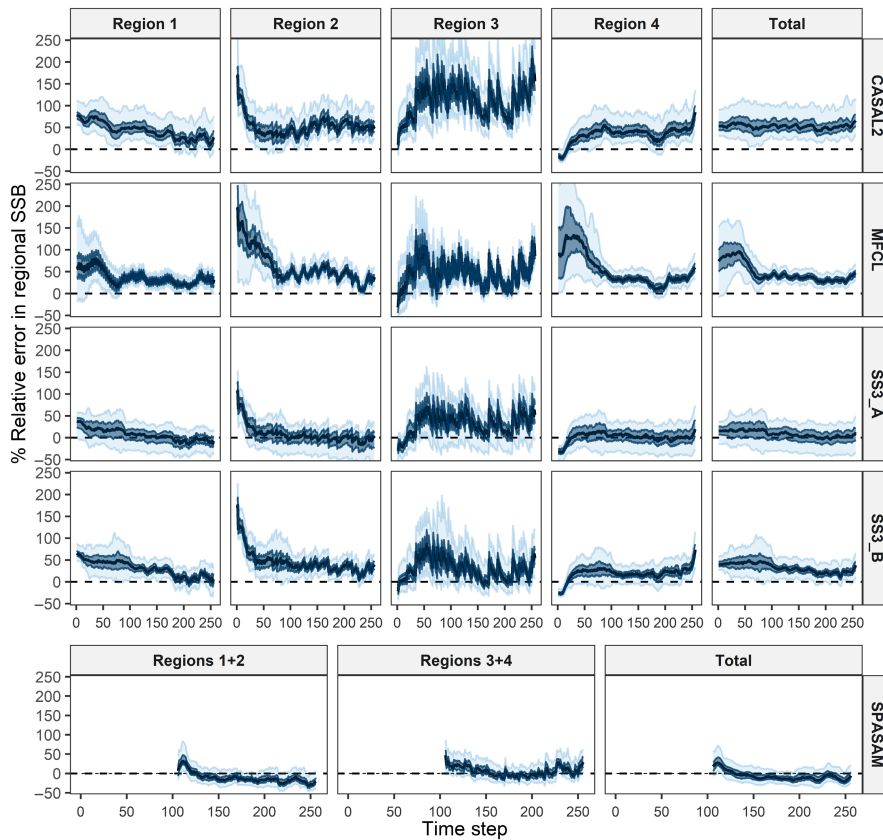
Across spatial structures, there were no consistent patterns in SSB bias, given that the initial scaling typically influenced the magnitude and direction of bias (Figure 4). For instance, the SS3\_A, SS3\_C, and SPASAM single region models tended to underestimate initial population scale and maintained negative bias for the entire time series (Figure 4). The corresponding SS3\_A and SPASAM spatial models were the least biased in regards to estimation of SSB, with moderate median bias (i.e.,  $< \pm 30\%$ ; Figure 4). The SS3\_C group did not have a spatial model, but their AAF model had reduced bias compared to the single region model, though, still with a general negative bias that exceeded median bias from the SS3\_A and SPASAM models. Conversely, the CASAL2 group's AAF approach was positively biased for the entire time series, while the associated single-region model demonstrated moderate median bias (i.e.,  $< \pm 30\%$ ), and the spatial model had increased positive bias compared to the AAF model. The SS3\_B and MFCL single region models demonstrated positive bias early in the time series with moderate median bias (i.e.,  $< \pm 30\%$ ) throughout the latter half of the time series. The corresponding spatial models demonstrated consistent positive bias throughout the time series. Aside

from the CASAL2 model (which maintained a consistent positive bias throughout the time series), the spatial assessments demonstrated a positive bias in the first half of the time series, which decreased through time (Figure 4).

As is expected given the definition of total population depletion (i.e., current year SSB divided by SSB in the first year of the model), it was accurately estimated early in the time series across all spatial structures and platforms (Figure 5). Accurate estimation of initial depletion is not surprising, given that it is a relative value (i.e., based on internal scaling not absolute scaling in comparison with the OM truth), and since all models were able to estimate the trend and scale of population declines (Figures C1 and C2 in Appendix S3). In general, the spatial assessments (except for the SPASAM model) provided more accurate estimates of depletion in the latter half of the time series compared to the single region models (Figure 5). In particular, the spatial models demonstrated reduced bias in estimates of terminal depletion, which is an important quantity for management decisions and harvest control rules (Figure 5). Moreover, the CASAL2 spatial model demonstrated low median bias ( $< \pm 10\%$ ) for depletion across much of the time series, while the SS3\_A and SS3\_B spatial models demonstrated a negative bias with slightly larger magnitude (i.e.,  $< \pm 20\%$ ; Figure 5). The SPASAM models demonstrated a unique pattern among spatial structures, where the single-region model tended to have less median depletion bias than the spatial model (Figure 5). Depletion bias for AAF structures was less than associated single region models.

**FIGURE 5** Comparison of percent relative error in total (aggregated across regions for spatial models) depletion of spawning stock biomass (i.e., relative to estimated spawning stock biomass in the first year of the estimation model) across model spatial structure (columns; left is single-region models, centre is spatially implicit areas-as-fleets models, and right is spatial models) and assessment platform (row; see Table 2 for a description of model parametrizations). The true value from the operating model is represented by the dashed line at zero. The solid dark line is the median percent relative error, while the shaded regions represent the 95% (light blue) and 75% (dark blue) intervals of percent relative error across all 100 simulation replicates. Only results from converged model runs are illustrated. Note that the SPASAM model started in time step 106.





**FIGURE 6** Comparison of percent relative error in regional (columns) spawning stock biomass for spatially explicit model structures across assessment platforms (row; see Table 2 for a description of model parametrizations). The true value from the operating model is represented by the dashed line at zero. The solid dark line is the median percent relative error, while the shaded regions represent the 95% (light blue) and 75% (dark blue) intervals of percent relative error across all 100 simulation replicates. Only results from converged model runs are illustrated. Note that the SPASAM model started in time step 106 and only modelled two regions (i.e., regions 1 and 2 were aggregated along with regions 3 and 4).

Relative regional scaling as well as regional population trends were generally consistent with the truth from the OM for the spatial models (Figure 6 and Figure C3 in Appendix S3). Estimates for regions one and four were generally the least biased across platforms (Figure 6). On the other hand, initial estimates from region two were more strongly biased, while estimates for region three demonstrated the highest bias across the time series (Figure 6). Results from the SPASAM model are difficult to compare to the other spatial models, given that only two regions were modelled. Bias in SSB for each of the combined regions was relatively low with a slight tendency to underestimate the SSB in the combined region that included regions one and two with overestimation in the combined region that included regions three and four (Figure 6 and Figure C3 in Appendix S3).

In terms of regional depletion, the spatial models demonstrated slight negative bias in region one, strong negative bias in region two, and strong positive bias for regions three and four (Figures C4 and C5 in Appendix S3). The MFCL spatial model differed slightly with a negative bias in depletion for region four (and an associated positive bias in SSB). Interestingly, despite demonstrating less bias in depletion estimates for most regions compared to the other spatial models, the MFCL spatial model had increased bias for total depletion compared to most other platforms (Figure 5). Generally, the spatial models matched the trends and depletion levels for the regions with the greatest contrast (i.e., regions one and four), but had difficulty estimating the dynamics for the smaller and less heavily fished regions (i.e., regions two and three).

## 4 | DISCUSSION

By implementing a high-resolution, spatially explicit OM in conjunction with a blinded experimental design, the current study provides a realistic demonstration of potential stock assessment performance and bias. Moreover, the experiment demonstrates the first use of a blinded experimental design for spatially explicit assessment models, emulating previous studies for non-spatial models (e.g., Deroba et al., 2015; ICES, 1993; NRC, 1998). Given the inherent complications and uncertainty presented by the study design, it was not surprising that none of the platforms or spatial structures were able to provide completely unbiased estimates of SSB or depletion. However, across all platforms and spatial structures, assessments were generally able to recreate the true population trends. The most consistent finding across spatial structures was that single region models always estimated a lower population scale compared to spatial models (within a given platform), while spatially implicit AAF models tended to estimate a population scale intermediate to the two extremes. Additionally, spatial assessments (aside from the SPASAM platform) were generally better able to estimate recent and terminal depletion compared to single region models.

Moreover, the spatial models were able to match the general population trends by region. Regional dynamics tended to be better estimated for the larger and more heavily exploited (see Figures A3–A5 in Appendix S1 for spatiotemporal trends in removals), and therefore better sampled, regions (e.g., region one). Because the dynamics within regions with less biomass (or smaller population

units) are often more difficult to differentiate, independently modelling and estimating associated region-specific parameters can be extremely difficult (Goethel et al., 2019; Vincent et al., 2017). Thus, careful delineation of regional boundaries is needed, particularly when no strong biological population structure exists (e.g., when a single population unit is distributed across multiple regions; Berger et al., 2021). Similarly, aggregating across regions with similar dynamics should be considered where appropriate (e.g., the SPASAM approach in this experiment), while ways to improve sampling (e.g., for compositional data and tag recaptures) from regions with lower biomass or fishing intensity need to be identified (Goethel, Berger, et al., 2023).

The general results are consistent with the conclusions of Deroba et al. (2015) in that model structure (i.e., surplus production compared to age-structured models in that study and single-region compared to spatial models in our study) appears to have the most important influence on population scaling, as opposed to different parametrizations within a given structure. Therefore, the decision to aggregate across important biological or fishery dynamics, whether by age or in space, is likely to be the most influential decision during the development of a stock assessment model. Careful deliberation and exploration (e.g., through analysis of disaggregated data) of model structure is merited to identify an adequate balance between parsimony and complexity. It is also interesting to note that, in the Deroba et al. (2015) experiment, the highest bias in model cross-tests often occurred in the terminal years, whereas that was not the case in the current experiment (i.e., bias was typically higher in initial scale rather than terminal SSB). Closed loop simulation and management strategy evaluation (MSE) would be necessary to identify which type of bias is more detrimental to the establishment of robust catch advice.

Two notable model parametrizations led to important improvements in model performance and merit further evaluation. First, allowing for time- and age-varying movement may improve spatial model performance, an approach recommended by Goethel et al. (2021) when limited knowledge exists as to the primary drivers of movement. For example, the SPASAM model (and the *Alt\_Move + App\_Rave\_Spat* exploratory run, see Appendix S2), which allowed time- and age-varying movement parameters, demonstrated limited bias in total and regional SSB. Further simulation testing is needed to determine if the added complexity outweighs the potential bias associated with simplifying movement dynamics. Second, utilizing a spatiotemporal CPUE standardization approach may better elucidate regional scaling (e.g., as observed with the performance of the SS3\_A and SS3\_C models). Therefore, high-resolution data analysis and preprocessing should be a first step in the development of any spatial assessment because it can inform all structural aspects of the assessment.

Moreover, decisions during the initial data aggregation stage of model development (e.g., whether to weight length compositions by samples per cell, as was the default for most modelling approaches in the experiment, or by cell-specific catch) may have important and unexpected influences on model results. For instance, the SS3\_C

team demonstrated that the approach used to aggregate purse-seine length compositions (i.e., catch or sample weighted) was more influential on the estimates of total SSB than model structure (i.e., single-region compared to AAF; Figure C6 in Appendix S3). An important aspect of developing future assessment good practices will be defining a more prescriptive, instead of subjective, approach to data processing and assessment decision-making.

For any model, there is a trade-off between the number of partitions that can be included and the associated complexity that can be integrated for each process, especially given data constraints. When spatially-explicit models are intractable, the results of the simulation experiment suggest that AAF models are likely to outperform single-region models, indicating that implicitly accounting for spatial processes may be preferred to completely ignoring them. Again, careful data analysis and aggregation could be critical for implementing adequate AAF approaches (e.g., using the SS3\_C group's novel regression tree approach, based on Lennert-Cody et al., 2010, 2013, to help identify and delineate fleet structure). For spatial models, the parsimony versus complexity balance will be unique to each application and must weigh data availability, primary spatial drivers, and management goals (Goethel, Berger, et al., 2023). For example, the SPASAM model included only two regions, yet performed well for estimating parameters in those regions. Similarly, the CASAL2 spatial model provided unbiased estimates of total depletion, despite ignoring movement and not integrating tagging data. Thus, appropriate delineation of regional or population boundaries may be an adequate first step towards spatial model development (Cadrin et al., 2019). Moreover, there is likely an interplay between spatial parametrization (i.e., the ability to estimate time- and age-variation in movement) with the number of regions modelled.

The treatment of tagging data is yet another aspect of spatial models that also merits further investigation. Tagging data are likely needed to adequately estimate movement rates, but there are important trade-offs when it is integrated into assessments. For example, including tagging data appeared to improve overall scaling, as indicated by the *No\_Tag* exploratory run results (see Figure B1 in Appendix S2). However, the difficulties of dealing with tag mixing predominated the model development process of many analyst groups (i.e., the MFCL, SS3\_A, and SS3\_B groups explored tag mixing extensively; see Berger et al., 2024). Analysing tag mixing rates is complex because mixing depends on the movement dynamics and dispersal potential of fish, the size of the modelled regions, and the distribution of tag releases (particularly in relation to regional boundaries). The simulated yellowfin tuna dynamics suggest that tag mixing rates were probably relatively low at the ocean-basin scale, but immature fish (i.e., most tagged fish were immature) were likely able to mix within and across regions before reaching maturity. Because most of the spatial assessment models assumed four regions (Figure 2), it is likely that intermediate tag mixing periods (e.g., four to eight quarters) would be adequate to allow tags to fully mix with the untagged population. Given the number of processes that tagging data can inform (i.e., movement, mortality, and distribution), further work to identify tag mixing periods (e.g.,



Kolody & Hoyle, 2015) and integrate new tag types (e.g., Thorson et al., 2021) should be a priority.

#### 4.1 | Implications and potential drivers of estimation model performance

The observed differences among spatial structures could be due to the relative distribution of fishing effort and SSB, where lower fishing pressure in more southern cells in the OM led to distributional hotspots and the potential for cryptic biomass (Figure 2). However, the lack of overlap between the fishery and biomass may be a model artefact due to the use of real-world catch locations with simulated recruitment and movement dynamics, and may not reflect the real-world situation for yellowfin tuna in the Indian Ocean.

Because single-region assessment models assume uniform dynamics and homogenous distribution, there may be a tendency to underestimate biomass when spatial structure exists and fishing pressure is heterogeneous (e.g., Guan et al., 2013). Conversely, spatial models are able to explicitly account for varying fishing pressure among regions, but rely on adequate and reliable data to discriminate among movement, recruitment, and mortality processes. When only fishery-dependent data are available (i.e., as is the case in the current study and for many tuna assessments), preferential sampling and lack of data from areas of low fishing pressure are likely to lead to increased bias compared to situations where more holistic sampling from fishery-independent surveys is available (Marsh, 2022).

The exact mechanisms that might lead to the pattern of spatial models estimating higher population scaling remain uncertain. One hypothesis that was introduced during the experiment workshop centred on the possibility that the catch and CPUE data create minimum biomass levels for each spatial region (i.e., 'biomass floors') to support the removals, which could increase overall scale when aggregated to the population level. Although untested, the hypothesis warrants further examination, likely through development of alternate OM scenarios with varying regional catch and data availability or quality. The treatment of abundance data could also have an influence on regional estimates, given that models utilizing CPUE from spatiotemporal standardizations performed well. Thus, there may be some advantages to using more sophisticated spatiotemporal CPUE standardization approaches to better account for spatial processes and autocorrelation, which merits additional investigation.

Multiple real-world tuna assessments have demonstrated a similar pattern of higher scaling in spatial compared to single-region models. For instance, an assessment of bigeye tuna (*Thunnus obesus*) in the western and central Pacific Ocean (Ducharme-Barth et al., 2020) estimated higher population scale and more optimistic levels of depletion for a nine-region spatial model compared to a single-region AAF model. A similar pattern was observed for an eight-region spatial assessment for skipjack tuna (*Katsuwonus pelamis*) in the western and central Pacific Ocean compared to a five-region model (Vincent et al., 2019). Hypothesized drivers for higher

population scaling of the more complex spatial models were similar to potential drivers in the current study. For instance, build-up of cryptic biomass in large, lightly fished temperate regions was proposed to limit information on regional scaling from CPUE indices. Moreover, compartmentalized regions of high fishing mortality in spatial models were hypothesized to lead to large-scale refuge from fishing pressure (i.e., compared to lower resolution models that assumed a more homogeneous distribution of fishing mortality).

Using a high-resolution, spatially explicit OM, McGilliard et al. (2015) observed a similar pattern in SSB estimation bias between spatial and single-region models (i.e., single-region models estimated a lower scale compared to the spatial models), too. However, the spatial models were generally unbiased, while the single-region models were negatively biased. McGilliard et al. (2015) included a fishery-independent survey, which likely aided estimation of regional recruitment, movement among regions, and fishing pressure. The improved performance of the spatial model in the exploratory run where a de facto fishery-independent survey was integrated (*Survey\_All\_Yrs\_Spat*) illustrates that relying solely on fishery data may be an important driver of results (see Figure B1 in Appendix S2). However, the general performance of spatial models relative to single-region counterparts is likely dependent on the spatial dynamics present (Goethel, Berger, et al., 2023; Guan et al., 2019), and further evaluation with high-resolution simulation frameworks is recommended.

We emphasize, though, that interpreting the experimental results must be done with care, given that it was impossible to ensure equality in resources and prior knowledge of system dynamics across analyst groups. For instance, there was a wide disparity across groups in terms of having worked on tuna assessments (or even the emulated yellowfin tuna assessment) before. However, it was unlikely that prior knowledge provided a large benefit in the context of the experiment. Conversely, time and resources devoted to model development and validation were likely unequal amongst teams, and the amount of time that a team was able to devote to the project probably influenced performance to an unknown extent. The experiment was also hampered by uncertain timelines induced by the COVID-19 pandemic. The one certainty was that no team was able to spend as much time as they would have preferred developing their spatial assessment.

#### 4.2 | Recommendations and future directions for collaborative, international simulations

Using a high-resolution OM provided a unique opportunity to better emulate real-world data scenarios, but it proved more difficult to decipher drivers of EM performance given the potential for more extreme misspecification. Future blinded simulation experiments could be improved by implementation of a more systematic investigation of EM performance following the group analysis stage of the experiment. Once the blinded experiment has been completed by each analyst group and the code base is available for translating OM outputs to EM inputs for each platform, then



a single lead could begin simultaneous investigations across platforms. For instance, the approach might start with implementing consistent versions of each platform (e.g., as was done by Li et al., 2021), then make systematic changes to identify how and why platform outputs diverged. Also, development of alternate OMs could help identify specific drivers of bias or data types that could improve performance.

In the future, high-resolution simulations could also be combined with more traditional approaches where the OM and EM had the same resolution or utilized the same framework (e.g., Deroba et al., 2015). Thereby, aspects of observed EM performance that were due to using the high-resolution OM could be isolated by comparing performance due to a specific or known misspecification from a lower resolution OM. Moreover, stepwise addition of complexity (i.e., building from low to high-resolution OMs) could be conducive to more thoroughly understand EM-OM interactions and performance. Conversely, to improve OM realism, we also recommend integration of high-resolution fishing effort dynamics models (e.g., Fisch et al., 2021; Saul et al., 2020). Moreover, utilizing agent-based OMs might help to better simulate mark-recapture and compositional (i.e., age and length) data that reflects real-world information content and uncertainty (e.g., Marsh, 2022; Scutt Phillips et al., 2018).

Implementing the blinded study design proved to be the most difficult aspect of the study. Ensuring that all analyst teams are starting with similar prior knowledge and resource allocations (i.e., time and personnel available for model development) would greatly aid the ability to make broader generalizations. Although the blinded design can be challenging to organize and implement, it should be more widely utilized, given that it enables recreating the entire assessment process (i.e., including high resolution data analysis) and helps implement more realistic uncertainty.

Our primary recommendation regarding simulation design is that consistent and sufficient funding should be secured at the start of the project, which covers expenses for the duration of the experiment associated with dedicated personnel to organize and facilitate the experiment, a primary research team devoted to OM development and refinement, and analyst time to implement EMs. Moreover, resources associated with file storage, virtual webinar hosting, and in person workshops should also be considered. A team approach and reproducible workflows are essential for sharing workloads across the organizational team and to account for attrition. Similarly, sufficient cloud storage will help with data sharing, backup, and collating results, given the large amounts of data that will be produced. Implementing realistic timelines may be one of the hardest aspects of a collaborative, international simulation, but realization of milestones can be aided by having personnel whose primary task is facilitating the study.

### 4.3 | Conclusions

As stock assessment methodology rapidly evolves, it is becoming more difficult to keep pace with new approaches, which are often

not documented in the primary literature. Thus, collaborative, international, and cross-platform simulation studies are increasingly important for the dissemination of stock assessment good practices (Deroba et al., 2015). Moreover, with increasing demands on fisheries managers to address climate change, ecosystem considerations, and marine spatial planning, utilization of high-resolution OMs as well as spatial assessment approaches will be necessary. The spatial simulation experiment provided a useful demonstration and stepping stone for future iterations of complex, spatially explicit simulation testing of fisheries models and management paradigms. We expect that the existing code base and public GitHub repository will provide a useful starting point for future simulation experiments, and we encourage researchers to make use of the existing resources from this experiment.

The results of the simulation experiment do not necessarily provide any generalizable answers to the question of whether single-region or spatial models perform best. Whether more complex spatial assessments are warranted for a particular fisheries management application will be context dependent and influenced by the goals of management, the data availability, biological understanding, and fishery dynamics (Goethel, Berger, et al., 2023). When spatial models are being pursued, it is recommended that single-region models, including AAF models, always be developed in tandem with the spatial model. Developing and presenting both single-region and spatial models may lead to synergistic improvements in management advice, instead of aiming to present a single 'best' model (e.g., the spatial model might be utilized to understand regional depletion and partition catch among regions). Moreover, it is recommended that future work with spatial assessments incorporate feedback control loops (i.e., utilize MSE) to explore the robustness of catch advice as opposed to only the bias in estimated quantities (e.g., Punt et al., 2017).

The simulation experiment provided a forum to share and disseminate spatial model building approaches across many of the world's fisheries organizations, which instigated numerous new collaborations and research agendas. Thus, we encourage RFMOs to pursue funding to support continuation of similar simulation experiments in the future, while working across institutional boundaries to improve and develop next-generation assessment platforms. The spatial complexities that must be confronted by fisheries assessment and management will increase as living marine resources redistribute in warming oceans, thereby, crossing regional and jurisdictional boundaries (Liu et al., 2023). Similarly, expansion of the blue economy will further complicate partitioning of the marine environment among competing sectors. Thus, earmarking resources to support increased international collaborations and development of high-resolution fishery models is imperative to ensure sustainable, scientifically informed management advice in the coming years.

### AFFILIATIONS

<sup>1</sup>NOAA, Alaska Fisheries Science Center, Juneau, Alaska, USA

<sup>2</sup>NOAA, Northwest Fisheries Science Center, Newport, Oregon, USA

<sup>3</sup>National Institute of Water and Atmospheric Research Ltd (NIWA), Nelson, New Zealand

- <sup>4</sup>NOAA, Office of Science and Technology, Silver Spring, Maryland, USA  
<sup>5</sup>Puget Sound Institute, University of Washington-Tacoma, Tacoma, Washington, USA  
<sup>6</sup>NOAA, Northeast Fisheries Science Center, Woods Hole, Massachusetts, USA  
<sup>7</sup>NOAA, Pacific Islands Fisheries Science Center, Honolulu, Hawaii, USA  
<sup>8</sup>Ocean Environmental Ltd., Wellington, New Zealand  
<sup>9</sup>Indian Ocean Tropical Tuna Commission, Victoria, Seychelles  
<sup>10</sup>Instituto Español de Oceanografía (IEO, CSIC), Centro Oceanográfico de Vigo, Vigo, Pontevedra, Spain  
<sup>11</sup>National Institute of Water and Atmospheric Research Ltd (NIWA), Auckland, New Zealand  
<sup>12</sup>Inter-American Tropical Tuna Commission, La Jolla, California, USA  
<sup>13</sup>School of Aquatic and Fishery Sciences, University of Washington, Seattle, Washington, USA  
<sup>14</sup>NOAA, Northwest Fisheries Science Center, Seattle, Washington, USA  
<sup>15</sup>NOAA, Office of the Science Director, Northwest Fisheries Science Center, Seattle, Washington, USA  
<sup>16</sup>NOAA, Southeast Fisheries Science Center, Beaufort, North Carolina, USA  
<sup>17</sup>National Institute of Water and Atmospheric Research Ltd (NIWA), Wellington, New Zealand  
<sup>18</sup>Department of Aquatic Resources, Institute of Marine Research, Swedish University of Agricultural Sciences, Lysekil, Sweden  
<sup>19</sup>Te Takina Ltd., Whangarei, New Zealand  
<sup>20</sup>Pacific Community, Noumea Cedex, New Caledonia  
<sup>21</sup>AZTI, Marine Research, Basque Research and Technology Alliance (BRTA), Pasaia, Gipuzkoa, Spain

## ACKNOWLEDGMENTS

Many of the ideas for this paper were spurred by webinars, presentations, and discussions associated with the simulation experiment. We acknowledge the many participants for contributing ideas that were utilized in this manuscript. Also, we would like to thank each of the reviewers that have improved the manuscript, including Kristen Omori, Ben Williams, Chris Lunsford, two anonymous reviewers, and the handling editor. 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 organizations.

## DATA AVAILABILITY STATEMENT

All data simulated for this experiment as well as all OM and EM base model configurations are available from the project GitHub (<https://github.com/aaronmberger-nwpsc/Spatial-Assessment-Modeling-Workshop>) or can be obtained from the lead author upon request.

## ORCID

- Daniel R. Goethel  <https://orcid.org/0000-0003-0066-431X>  
 Aaron M. Berger  <https://orcid.org/0000-0002-1408-7122>  
 Simon D. Hoyle  <https://orcid.org/0000-0002-1787-6565>  
 Patrick D. Lynch  <https://orcid.org/0000-0001-7121-6181>  
 Caren Barceló  <https://orcid.org/0000-0002-3336-5052>  
 Jonathan Deroba  <https://orcid.org/0000-0003-4269-6790>  
 Nicholas D. Ducharme-Barth  <https://orcid.org/0000-0003-2772-9195>  
 Francisco Izquierdo  <https://orcid.org/0000-0002-0781-1354>  
 Giancarlo M. Correa  <https://orcid.org/0000-0003-0682-1152>  
 Brian J. Langseth  <https://orcid.org/0000-0002-9901-6146>

- Matthew T. Vincent  <https://orcid.org/0000-0003-3636-0932>  
 Teresa A'mar  <https://orcid.org/0000-0002-8634-1142>  
 Massimiliano Cardinale  <https://orcid.org/0000-0003-1870-3882>  
 Marta Cousido-Rocha  <https://orcid.org/0000-0002-4587-8808>  
 John Hampton  <https://orcid.org/0000-0003-3006-8797>  
 Carolina Minte-Vera  <https://orcid.org/0000-0002-0537-1519>  
 Agurtzane Urtizberea  <https://orcid.org/0000-0002-0306-6094>

## REFERENCES

- Berger, A. M., Barcelo, C. M., Goethel, D. R., Hoyle, S. D., Lynch, P. D., McKenzie, J., Dunn, A., Punt, A. E., Methot, R. D., Hampton, J., Porch, C., McGarvey, R., Thorson, J. T., A'mar, T., Deroba, J., Elvarsson, B., Holmes, S. J., Howell, D., Langseth, B. J., ... Rasmussen, S. (In Press). Synthesizing the spatial functionality of contemporary stock assessment software to identify future needs for next generation assessment platforms. *Fisheries Research*.
- Berger, A. M., Deroba, J. J., Bosley, K. M., Goethel, D. R., Langseth, B. J., Schueller, A. M., & Hanselman, D. H. (2021). Incoherent dimensionality in fisheries management: Consequences of misaligned stock assessment and population boundaries. *ICES Journal of Marine Science*, 78(1), 155–171. <https://doi.org/10.1093/icesjms/fsaa203>
- Berger, A. M., Goethel, D. R., Hoyle, S. D., Lynch, P. D., Barcelo, C., Day, J., Langseth, B. J., Minte-Vera, C., Xu, H., Izquierdo, F., Fu, D., Ducharme-Barth, N., Vincent, M., Gruss, A., Olmos, M., Deroba, J., Correa, G., McKenzie, J., Marsh, C., ... Mace, P. (2024). 'Building the (im)perfect beast': Lessons-learned for developing spatial stock assessment models from an international, blinded simulation experiment. Manuscript in preparation.
- Berger, A. M., Goethel, D. R., Lynch, P. D., Quinn, T., II, Mormede, S., McKenzie, J., & Dunn, A. (2017). Space oddity: The mission for spatial integration. *Canadian Journal of Fisheries and Aquatic Sciences*, 74, 1698–1716. <https://doi.org/10.1139/cjfas-2017-0150>
- Bosley, K. M., Schueller, A. M., Goethel, D. R., Hanselman, D. H., Fenske, K. H., Berger, A. M., Deroba, J. J., & Langseth, B. J. (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.12616>
- Cadrin, S. X., Goethel, D. R., Morse, M. R., Gay, F., & Kerr, L. A. (2019). "So, where do you come from?" The impact of assumed spatial population structure on estimates of recruitment. *Fisheries Research*, 217, 156–168. <https://doi.org/10.1016/j.fishres.2018.11.030>
- Deroba, J. J., Butterworth, D. S., Methot, R. D., Jr., de Oliveira, J. A. A., Fernandez, C., Nielsen, A., Cadrin, S. X., Dickey-Collas, M., Legault, C. M., Ianelli, J., Valero, J. L., Needle, C. L., O'Malley, J. M., Chang, Y. J., Thompson, G. G., Canales, C., Swain, D. P., Miller, D. C. M., Hintzen, N. T., ... Hulson, P. J. F. (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(1), 19–30. <https://doi.org/10.1093/icesjms/fst237>
- Doonan, I., Large, K., Dunn, A., Rasmussen, S., Marsh, C., & Mormede, S. (2016). Casal2: New Zealand's integrated population modeling tool. *Fisheries Research*, 183, 498–505. <https://doi.org/10.1016/j.fishres.2016.04.024>
- Ducharme-Barth, N., Vincent, M., Hampton, J., Hamer, P., Williams, P., & Pilling, G. (2020). Stock assessment of bigeye tuna in the western and central Pacific Ocean. WCPFC-SC16-2020/SA-WP-03. Online. 11–20 August 2020. Western and Central Pacific Fisheries Commission. <https://meetings.wcpfc.int/node/11693>
- Dunn, A., Hoyle, S., & Datta, S. (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. Online. October 2020. [https://www.researchgate.net/publication/349548134\\_Development\\_of\\_spatially\\_explicit\\_operating\\_models\\_for\\_yellowfin\\_tuna\\_populations\\_in\\_the\\_Indian\\_Ocean](https://www.researchgate.net/publication/349548134_Development_of_spatially_explicit_operating_models_for_yellowfin_tuna_populations_in_the_Indian_Ocean)
- Dunn, A., Rasmussen, S., & Mormede, S. (2020). *Spatial population model user manual, SPM 2.0.4-2021-09-27. Ocean environmental technical report* (p. 233). Ocean Environmental Ltd. <https://github.com/alistairdunn1/SPM/releases/download/2.0.4-2021-09-27/SPM.pdf>
- Fisch, N., Camp, E., Shertzer, K., & Ahrens, R. (2021). Assessing likelihoods for fitting composition data within stock assessments, with emphasis on different degrees of process and observation error. *Fisheries Research*, 243, 106069. <https://doi.org/10.1016/j.fishres.2021.106069>
- Fu, D., Ijurco, A. U., Cardinale, M., et al. (2021). *Preliminary Indian Ocean yellowfin tuna stock assessment 1950–2020 (Stock Synthesis)*. IOTC–2021–WPTT23–12. Indian Ocean Tuna Commission, Working Party on Tropical Tunas. Online. 2 October 2021. [https://www.researchgate.net/publication/364813230\\_PRELIMINARY\\_INDIAN\\_OCEAN\\_YELLOWFIN\\_TUNA\\_STOCK\\_ASSESSMENT\\_1950-2020\\_STOCK\\_SYNTHESIS\\_IOTC-2021-WPTT23-12](https://www.researchgate.net/publication/364813230_PRELIMINARY_INDIAN_OCEAN_YELLOWFIN_TUNA_STOCK_ASSESSMENT_1950-2020_STOCK_SYNTHESIS_IOTC-2021-WPTT23-12)
- Fu, D., Langley, A., Merino, G., & Ijurco, A. U. (2018). *Preliminary Indian Ocean yellowfin tuna stock assessment 1950–2017 (Stock Synthesis)* (p. 116). IOTC–2018–WPTT20–33. Mahé, Seychelles. 29 October – 3 November, 2018. Indian Ocean Tuna Commission.
- Goethel, D. R., Berger, A. M., & Cadrin, S. X. (2023). Spatial awareness: Good practices and pragmatic recommendations for developing spatially structured stock assessments. *Fisheries Research*, 264, 106703. <https://doi.org/10.1016/j.fishres.2023.106703>
- Goethel, D. R., Bosley, K. M., Hanselman, D. H., Berger, A. M., Deroba, J. J., Langseth, B. J., & Schueller, A. M. (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., Bosley, K. M., Langseth, B. J., Deroba, J. J., Berger, A. M., Hanselman, D. H., & Schueller, A. M. (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(1), 141–160. <https://doi.org/10.1111/faf.12510>
- Goethel, D. R., Omori, K. L., Punt, A. E., Lynch, P. D., Berger, A. M., de Moor, C. L., Plagányi, É. E., Cope, J. M., Dowling, N. A., McGarvey, R., Preece, A. L., Thorson, J. T., Chaloupka, M., Gaichas, S., Gilman, E., Hesp, S. A., Longo, C., Yao, N., & Methot, R. D. (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>
- Guan, W., Cao, J., Chen, Y., & Cieri, M. (2013). Impacts of population and fishery spatial structures on stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences*, 70, 1178–1189. <https://doi.org/10.1139/cjfas-2012-0364>
- Guan, W., Wu, J., & Tian, S. (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>
- Hilborn, R. (2012). The evolution of quantitative marine fisheries management 1985–2010. *Natural Resource Modeling*, 25, 122–144. <https://doi.org/10.1111/j.1939-7445.2011.00100.x>
- Hoyle, S. D., Dunn, A., Gruss, A., Mormede, S., Barcelo, C., Goethel, D. R., Berger, A. M., & Lynch, P. D. (2024). 'Worlds of our own': Developing high spatial resolution operating models for an international, blinded simulation experiment to evaluate spatial stock assessments. Manuscript in preparation.
- Hoyle, S. D., & Langley, A. D. (2020). Scaling factors for multi-region stock assessments, with an application to Indian Ocean tropical tunas. *Fisheries Research*, 228, 105586. <https://doi.org/10.1016/j.fishres.2020.105586>
- ICES (International Council for the Exploration of the Seas). (1993). Reports (3) of the working group on "Methods of fish stock assessment". Report no. 191. p. 249. <https://doi.org/10.17895/ices.pub.4607>
- Kleiber, P., Fournier, D. A., Davies, N., Bouye, F., & Hoyle, S. (2018). MULTIFAN-CL user's guide. 1 September 2018. p. 197. [https://mfcl.spc.int/index.php?option=com\\_jdownloads&view=viewcategory&catid=3&Itemid=116](https://mfcl.spc.int/index.php?option=com_jdownloads&view=viewcategory&catid=3&Itemid=116)
- Kolody, D., & Hoyle, S. (2015). Evaluation of tag mixing assumptions in western Pacific Ocean skipjack tuna stock assessments. *Fisheries Research*, 163, 127–140. <https://doi.org/10.1016/j.fishres.2014.05.008>
- Lennert-Cody, C. E., Maunder, M. N., Aires-da-Silva, A., & Minami, M. (2013). Defining population spatial units: Simultaneous analysis of frequency distributions and time series. *Fisheries Research*, 139, 85–92. <https://doi.org/10.1016/j.fishres.2012.10.001>
- Lennert-Cody, C. E., Minami, M., Tomlinson, P. K., & Maunder, M. N. (2010). Exploratory analysis of spatial-temporal patterns in length-frequency data: An example of distributional regression trees. *Fisheries Research*, 102, 323–326. <https://doi.org/10.1016/j.fishres.2009.11.014>
- Li, B., Shertzer, K. W., Lynch, P. D., Ianelli, J. N., Legault, C. M., Williams, E. H., Methot Jr, R. D., Brooks, E. N., Deroba, J. J., Berger, A. M., Sagarese, S. R., Brodziak, J. K. T., Taylor, I. G., Karp, M. A., Wetzel, C. R., & Supernaw, M. (2021). A comparison of four primary age-structured stock assessment models used in the United States. *Fishery Bulletin*, 119, 149–167. <https://doi.org/10.7755/FB.119.2-3.5>
- Liu, O. R., Ward, E. J., Anderson, S. C., Andrews, K. S., Barnett, L. A. K., Brodie, S., Carroll, G., Fiechter, J., Haltuch, M. A., Harvey, C. J., Hazen, E. L., Hervann, P. Y., Jacox, M., Kaplan, I. C., Matson, S., Norman, K., Pozo Buil, M., Selden, R. L., Shelton, A., & Samhour, J. F. (2023). Species redistribution creates unequal outcomes for multispecies fisheries under projected climate change. *Science Advances*, 9, eadg5468. <https://doi.org/10.1126/sciadv.adg5468>
- Marsh, C. (2022). *Spatial methods for improved estimates of abundance indices for preferentially or systematically sampled data*. PhD dissertation, University of Auckland.
- McGilliard, C. R., Punt, A. E., Methot, R. D., & Hilborn, R. (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>
- Melnichuk, M. C., Peterson, E., Elliott, M., & Hilborn, R. (2017). Fisheries management impacts on target species status. *Proceedings of the National Academy of Sciences of the United States of America*, 114(1), 178–183. <https://doi.org/10.1073/pnas.1609915114>
- Methot, R. D., Jr. (2009). Stock assessment: Operational models in support of fisheries management. In R. J. Beamish & B. J. Rothschild (Eds.), *The future of fisheries science in North America* (pp. 137–165). Springer.
- Methot, R. D., Jr., & Wetzel, C. R. (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>
- NRC (National Research Council). (1998). *Improving fish stock assessments*. The National Academies Press. <https://doi.org/10.17226/5951>
- 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. <https://doi.org/10.1016/j.fishres.2017.08.021>
- Punt, A. E. (2019b). Spatial stock assessment methods: A viewpoint on current issues and assumptions. *Fisheries Research*, 213, 132–143. <https://doi.org/10.1016/j.fishres.2019.01.014>

- Punt, A. E., Haddon, M., Little, L. R., & Tuck, G. N. (2017). The effect of marine closures on a feedback control management strategy used in a spatially aggregated stock assessment: A case study based on pink ling in Australia. *Canadian Journal of Fisheries and Aquatic Sciences*, 74, 1960–1973. <https://doi.org/10.1139/cjfas-2016-0017>
- Quinn, T. J., II, & Deriso, R. B. (1999). *Quantitative fish dynamics* (p. 560). Oxford University Press.
- Saul, S., Brooks, E. N., & Die, D. (2020). How fisher behavior can bias stock assessment: Insights from an agent-based modeling approach. *Canadian Journal of Fisheries and Aquatic Sciences*, 77, 1794–1809. <https://doi.org/10.1139/cjfas-2019-0025>
- Scutt Phillips, J., Gupta, A. S., Senina, I., van Sebille, E., Lange, M., Lehodey, P., Hampton, J., & Nicol, S. (2018). An individual-based model of skipjack tuna (*Katsuwonus pelamis*) movement in the tropical Pacific Ocean. *Progress in Oceanography*, 164, 63–74. <https://doi.org/10.1016/j.pocean.2018.04.007>
- Thorson, J. T., Barbeaux, S. J., Goethel, D. R., Kearney, K. A., Laman, E. A., Nielsen, J. K., Siskey, M. R., Siwicke, K., & Thompson, G. G. (2021). Estimating fine-scale movement rates and habitat preferences using multiple data sources. *Fish and Fisheries*, 22(6), 1359–1376. <https://doi.org/10.1111/faf.12592>
- Vincent, M. T., Brenden, T. O., & Bence, J. R. (2017). Simulation testing the robustness of a multi-region tag-integrated assessment model that exhibits natal homing and estimates natural mortality and reporting rate. *Canadian Journal of Fisheries and Aquatic Sciences*, 74, 1930–1949. <https://doi.org/10.1139/cjfas-2016-0297>
- Vincent, M. T., Pilling, G. M., & Hampton, J. (2019). Stock assessment of skipjack tuna in the western and central Pacific Ocean. WCPFC-SC15-2019/SA-WP-05-Rev2. Pohnpei, Federated State of Micronesia. 12–20 August 2019. <https://meetings.wcpfc.int/node/11230>

## SUPPORTING INFORMATION

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

**How to cite this article:** Goethel, D. R., Berger, A. M., Hoyle, S. D., Lynch, P. D., Barceló, C., Deroba, J., Ducharme-Barth, N. D., Dunn, A., Fu, D., Izquierdo, F., Marsh, C., Xu, H., Correa, G. M., Langseth, B. J., Maunder, M. N., McKenzie, J., Methot, R. D., Vincent, M. T., A'mar, T. ... Urtizberea, A. (2024). 'Drivin' with your eyes closed': Results from an international, blinded simulation experiment to evaluate spatial stock assessments. *Fish and Fisheries*, 25, 471–490. <https://doi.org/10.1111/faf.12819>