Spatiotemporal Modeling of Spawner-Recruitment Relationships in Eastern Bering Sea Snow Crab (*Chionoecetes opilio*): From Structured Inference to Deep Learning

Nikolai Morokhovich

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2026

Committee:
Andre Punt, Chair
Ray Hilborn
Cody Szuwalski
James Thorson

Program Authorized to Offer Degree:

Quantitative Ecology and Resource Management

University of Washington

Abstract

Spatiotemporal Modeling of Spawner-Recruitment Relationships in Eastern Bering Sea Snow Crab (*Chionoecetes opilio*): From Structured Inference to Deep Learning

Nikolai Morokhovich

Chair of the Supervisory Committee:
Andre Punt
School of Aquatic and Fishery Sciences

Snow crab (*Chionoecetes opilio*) support one of the most economically important shellfish fisheries in the eastern Bering Sea, yet the spatial and temporal processes linking spawning adults to juvenile recruitment remain poorly understood. Spawner-recruit dynamics in this system constitute a biological teleconnection: adult spawning populations regulate the spatial distribution of juvenile recruits through a multi-year sequence of larval dispersal, settlement, and early benthic growth. Identifying this teleconnection requires methods capable of extracting predictive structure from a dataset that is high-dimensional in space but also severely limited in time. This thesis evaluates a gradient of methods for detecting and quantifying the snow crab spawner-recruit teleconnection, spanning classical dimensionality reduction to deep learning. The specific methods evaluated are three statistical approaches with station-level resolution: Empirical Orthogonal Function analysis with Generalized Additive Models (EOF-GAM), Coupled-Factor Generalized Linear Latent Variable Models (CF-GLLVM) with Linked and Spatially Varying Coefficient formulations, and spectral clustering with GAM-based prediction, and a machine learning approach (Vision Transformer, ViT) that is trained on spatially interpolated gridded fields, with Integrated Gradients attribution used to recover biologically interpretable structure from learned predictions. The methods are evaluated using held-out recruitment years spanning 2021-2023, a period of population collapse that provides a rigorous test of out-of-sample generalization. It is not possible to directly compare across methods quantitatively due to differences in spatial data representation. However, the CF-GLLVM achieved the highest station-level Spearman's correlation ($\rho = 0.733$) and F1 score (0.782), while the EOF-GAM produced the best hotspot identification (Jaccard = 0.409); spectral clustering performed at or below a climatological baseline across all metrics while the ViT achieved grid-level Spearman's correlations of 0.83-0.87 and reduced mean absolute error by 33-38% relative to the climatological baseline. Attribution analysis for the ViT revealed a multi-generational thermal signal, consistent with cold pool conditions shaping the thermal environment experienced by larval cohorts at the time of spawning. No single method dominated across all objectives, but together they provide converging evidence of a spatially structured spawner-recruit teleconnection that persists across a period of substantial population change. The framework developed here is broadly transferable to other ecologically and commercially important species where spatially structured leading populations regulate responses in recipient populations across space and time.

Chapter 0 Introduction

0.1 The ecological and economic stakes of snow crab recruitment

The eastern Bering Sea (EBS) snow crab (*Chionoecetes opilio*) fishery has historically ranked among the most economically important shellfish fisheries in the world, supporting coastal communities, processing industries, and export markets valued at hundreds of millions of dollars annually (Szuwalski et al. 2023).

Yet the collapse of the EBS snow crab stock and the subsequent need for a rebuilding plan demonstrated how rapidly environmental perturbations can overwhelm management frameworks. More broadly, sustainable harvest depends on the ability to anticipate recruitment: the process by which juveniles produced in one year eventually enter the harvestable population years later.

Recruitment in snow crab reflects the cumulative outcome of spawning location, larval dispersal across the EBS shelf, settlement into suitable habitat, and multi-year survival through a period of high vulnerability to temperature extremes and predation (Anger 2006; Sainte-Marie et al. 1995). The spatial distribution of recruits is itself ecologically and economically consequential, determining where future fishing effort will be productive and where spatial protections might most effectively shield rebuilding cohorts.

Spawner-recruit relationships are foundational to fisheries stock assessment, yet they remain notoriously difficult to model reliably (Szuwalski & Punt 2013). The core challenge is that the biological processes linking adult reproductive output to juvenile abundance operate across spatial and temporal scales that are difficult to observe simultaneously. In snow crab, a five-year ontogenetic lag separates the year during which spawning occurs from the year during which those offspring first appear reliably in the standard bottom trawl survey as 45-55 mm recruits (Sainte-Marie et al. 1995). During that interval, larval and early juvenile crabs are subject to ocean currents, temperature-dependent growth and mortality, and stochastic survival processes that decouple their eventual spatial distribution from that of their parents in ways that simple scalar spawner-recruit functions fail to capture (Anger 2006).

Conventional spawner-recruit approaches such as the Beverton-Holt and Ricker models collapse spatially explicit survey data into single annual abundance indices before fitting (Beverton & Holt 1957; Ricker 1954; Cadrin 2020). This aggregation is computationally convenient and statistically tractable given the short time series typical of fisheries data sets, but it discards potentially informative spatial structure. If certain spatial configurations of spawning adults are systematically more predictive of recruitment abundance and location than others, that signal is lost when observations are reduced to scalars. The result is that models that perform adequately on average may struggle precisely when spatial dynamics matter most: during regime shifts, range contractions, or spatial redistributions driven by environmental change (Szuwalski & Punt 2013).

The EBS snow crab dataset provides 35 years of spatially resolved bottom trawl survey data at 349 fixed stations on a 20 × 20 nautical mile grid (Markowitz et al. 2024). However, 35 annual observations, with a gap in 2020 due to the COVID-19 pandemic suspension of field operations, represent a severely data-limited setting for any statistical or machine learning model that must characterize relationships among hundreds of spatial locations simultaneously. This combination of high spatial dimensionality and few temporal observations defines the methodological challenge addressed in this thesis. It is not unique to snow crab: long-term ecological monitoring programs frequently lead to datasets with many spatial locations and relatively few temporal replicates, as a consequence of the practical limits of survey effort and the multi-decadal timescales over which meaningful ecological change unfolds (Warton et al. 2015). Methods that can extract predictive

signal from such data, while remaining robust to overfitting and generalizable across species and ecosystems, have value well beyond the specific system for which they are developed.

A teleconnection is a statistical or causal linkage between two geographically or temporally separated phenomena, such that conditions at one location or time predict conditions at another (Alheit & Bakun 2010). The term originates in atmospheric and oceanographic science, where large-scale climate modes such as the El Niño-Southern Oscillation (ENSO), the Pacific Decadal Oscillation (PDO), and the North Atlantic Oscillation (NAO) have long been recognized as sources of coordinated variability across ocean basins (Diaz et al. 2001; Yeh et al. 2018). In these environmental teleconnections, anomalies in sea surface temperature or atmospheric pressure in one region propagate through physical mechanisms to influence conditions thousands of kilometers away (Schwing et al. 2010). Biological teleconnections, in which a leading population at one location or time predicts a response population elsewhere or later, have received considerably less attention. Detecting and quantifying these linkages matters for fisheries management because they reveal hidden connectivity across seascapes and landscapes, can explain why populations fluctuate in synchrony or asynchrony across large distances, and provide early warning signals for populations that respond with a lag to conditions far away (Cardil et al. 2023; Serrão et al. 2025).

Snow crab spawner-recruit dynamics constitute a biological teleconnection. The spatial distribution of egg-bearing females in a given year is the leading population; the spatial distribution of recruits (quantified for snow crab by animals in the 45-55 mm size range) observed approximately five years later is the response. The intervening five years span a brief pelagic larval phase of several months (Anger 2006), followed by settlement and roughly four years of benthic juvenile growth before the cohort enters the survey selectivity window (Sainte-Marie et al. 1995). Both phases contribute to the spatial mismatch between spawner and recruit distributions: larvae are passively transported by currents during their pelagic period, while benthic juveniles redistribute through active movement and continued passive transport before reaching sizes at which they are selected by the survey gear. Combined with the five-year temporal lag, this multi-phase spatial decoupling means that no conventional spawner-recruit model can recover the relationship between spawners and recruits without explicitly accounting for its spatiotemporal structure.

Statistical models offer a flexible and generalizable alternative to mechanistic approaches for identifying biological teleconnections. Mechanistic models, such as regional ocean circulation models coupled to larval drift simulations, can represent the physical processes underlying dispersal in detail (Shchepetkin & McWilliams 2005; Sandvik et al. 2016). However, they are computationally intensive, highly sensitive to initial conditions and parameterization choices, typically system-specific in their design, and ill-suited to recovering delayed or indirect biological linkages that were not explicitly anticipated during construction (Cowen and Sponaugle 2009). Statistical models, by contrast, can identify the empirical signature of a teleconnection in observational data without requiring advance specification of the mechanism, can accommodate distributional assumptions appropriate for count-like or zero-inflated ecological data, and can propagate uncertainty in a way that is meaningful for management decisions.

0.2 Overall objective

This thesis addresses a single overarching question: can the spatiotemporal teleconnection between snow crab spawners and recruits in the EBS be identified and quantified in a way that is

predictively useful, biologically interpretable, and robust under conditions of severe data scarcity and unprecedented population change?

0.3 Basic approach and outline of thesis

Central to the statistical approaches explored in this thesis is the concept of dimensionality reduction: extracting lower-dimensional representations of high-dimensional spatial data that capture the dominant, predictively relevant patterns while discarding noise. Empirical Orthogonal Functions (EOFs) achieve this through singular value decomposition, representing the spatial field as a linear combination of orthogonal spatial patterns weighted by temporal coefficients (Lorenz 1956; Monahan et al. 2009). Generalized Linear Latent Variable Models (GLLVMs) embed similar dimensionality reduction within a formal statistical model that accommodates non-Gaussian distributions, spatial autocorrelation, and uncertainty quantification (Thorson & Kristensen 2024; Grüss et al. 2021). Clustering methods take a different approach, partitioning the survey domain into discrete regions of similar temporal behavior rather than decomposing continuous spatial fields (von Luxburg 2007). These approaches represent variations on a common strategy: making the relationship between hundreds of spatial locations measurable from a few dozen years of observations.

A natural extension of these methods leads to deep learning architectures, which can capture non-linear, long-range spatial dependencies without prior specification, but at the cost of requiring substantially more data than is typically available in ecological settings, and producing models that are difficult to interpret without additional analytical tools. The Vision Transformer (ViT) architecture (Dosovitskiy 2021) is a fundamentally different approach to the teleconnection problem: rather than pre-specifying the form of spatial dependence through covariance functions or orthogonal decompositions, its self-attention mechanism learns which spatial locations are mutually predictive directly from the data. This makes it well-suited to identifying long-range teleconnections whose spatial structure is unknown in advance. However, the black-box nature of deep learning models raises legitimate concerns about their usefulness in management contexts that require interpretable, scientifically defensible predictions. Integrated Gradients attribution (Sundararajan et al. 2017) is used to analyze the ViT's predictions and recover biologically interpretable structure.

Chapter 1 investigates the overall question using three statistical approaches operating at the scale of individual survey stations: an EOF-GAM pipeline, a Coupled-Factor GLLVM framework in both Linked and Spatially Varying Coefficient (SVC) formulations, and spectral clustering with GAM-based prediction. Chapter 2 approaches the same teleconnection from a deep learning perspective, adapting a Vision Transformer to predict gridded (50 × 50 km) recruit density fields from multi-channel historical inputs of spawner density, recruit density, and bottom temperature. The two chapters are evaluated using consistent holdout periods (2021-2023) that span a period of population collapse, providing a rigorous and ecologically meaningful test of each method's capacity to extrapolate beyond the conditions represented in training data.

A unifying theme across both chapters is the tension between model capacity and effective sample size. The EBS snow crab survey provides approximately 35 temporally independent observations, regardless of how many bootstrap replicates or spatial grid cells are used to represent each year. Methods that can extract robust, generalizable signal from this constraint, whether through explicit regularization, dimensionality reduction, or architectural restraint in deep learning, are the methods most likely to find application in similarly data-limited ecological systems.

It is important to note at the outset that the statistical and deep learning approaches operate on different data representations: station-level densities in Chapter 1, and Stochastic Partial Differential Equation (SPDE)-interpolated gridded fields in Chapter 2. This precludes direct quantitative comparison of performance metrics across chapters. Rather than treating the two chapters as competing approaches to the same problem, this thesis frames them as complementary lenses on the same problem.

Introduction References

Alheit, J., & Bakun, A. (2010). Population synchronies within and between ocean basins: Apparent teleconnections and implications as to physical–biological linkage mechanisms. Journal of Marine Systems, 79(3), 267–285. https://doi.org/10.1016/j.jmarsys.2008.11.029

Anger, K. (2006). Contributions of larval biology to crustacean research: A review. Invertebrate Reproduction & Development, 49(3), 175–205. https://doi.org/10.1080/07924259.2006.9652207

Cadrin, S. X. (2020). Defining spatial structure for fishery stock assessment. Fisheries Research, 221, 105397. https://doi.org/10.1016/j.fishres.2019.105397

Cardil, A., et al. (2023). Climate teleconnections modulate global burned area. Nature Communications, 14(1), 427. https://doi.org/10.1038/s41467-023-36052-8

Diaz, H. F., et al. (2001). ENSO variability, teleconnections and climate change. International Journal of Climatology, 21(15), 1845–1862. https://doi.org/10.1002/joc.631

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

Grüss, A., et al. (2021). Synthesis of interannual variability in spatial demographic processes supports the strong influence of cold-pool extent on Eastern Bering Sea walleye pollock (Gadus chalcogrammus). Progress in Oceanography, 194, 102569. https://doi.org/10.1016/j.pocean.2021.102569

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Lorenz, E. N. (1956). Empirical orthogonal functions and statistical weather prediction. https://cir.nii.ac.jp/crid/1370004237455296143

Markowitz, E. H., Dawson, E. J., Wassermann, S., Rohan, S. K., Charriere, N. E., & Stevenson, D. E. (2024). Results of the 2022 eastern and northern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate fauna. NOAA. https://repository.library.noaa.gov/view/noaa/50154

Monahan, A. H., Fyfe, J. C., Ambaum, M. H. P., Stephenson, D. B., & North, G. R. (2009). Empirical orthogonal functions: The medium is the message. Journal of Climate, 22, 6501–6514. https://doi.org/10.1175/2009JCLI3062.1

Ricker, W. E. (1954). Stock and recruitment. Journal of the Fisheries Research Board of Canada, 11(5), 559–623.

Sainte-Marie, B., Raymond, S., & Brêthes, J.-C. (1995). Growth and maturation of the benthic stages of male snow crab, Chionoecetes opilio (Brachyura: Majidae). Canadian Journal of Fisheries and Aquatic Sciences, 52(5), 903–924. https://doi.org/10.1139/f95-091

Sandvik, H., et al. (2016). Modelled drift patterns of fish larvae link coastal morphology to seabird colony distribution. Nature Communications, 7, 11599.

Schwing, F. B., et al. (2010). Climate change, teleconnection patterns, and regional processes forcing marine populations in the Pacific. Journal of Marine Systems, 79(3–4), 245–257. https://doi.org/10.1016/j.jmarsys.2008.11.027

Serrão, E. A. O., et al. (2025). The effects of teleconnections on water and carbon fluxes in the two South America's largest biomes. Scientific Reports, 15(1), 1395.

Shchepetkin, A. F., & McWilliams, J. C. (2005). The Regional Oceanic Modeling System (ROMS): A split-explicit, free-surface, topography-following-coordinate oceanic model. Ocean Modelling, 9(4), 347–404.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, 70, 3319–3328.

Szuwalski, C. S., et al. (2023). The collapse of eastern Bering Sea snow crab. Science, 382(6668), 306–310.

Szuwalski, C. S., & Punt, A. E. (2013). Regime shifts and recruitment dynamics of snow crab, Chionoecetes opilio, in the eastern Bering Sea. Fisheries Oceanography, 22(5), 345–354. https://doi.org/10.1111/fog.12026

Thorson, J., & Kristensen, K. (2024). Spatio-temporal models for ecologists. CRC Press.

von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and Computing, 17(4), 395–416.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So many variables: Joint modeling in community ecology. Trends in Ecology & Evolution, 30(12), 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Yeh, S.-W., et al. (2018). ENSO atmospheric teleconnections and their response to greenhouse gas forcing. Reviews of Geophysics, 56(1), 185–206.

Chapter 1: Statistical Methods for Predicting Teleconnections in Snow Crab (*Chionoecetes opilio*)

1.Abstract

Forecasting recruitment in spatially complex fisheries where observations are high-dimensional relative to the number of years sampled remains a persistent challenge, particularly when limited observations constrain the identification of meaningful ecological relationships. This research addresses the critical need for improved recruitment prediction methods for the Eastern Bering Sea snow crab (*Chionoecetes opilio*) fishery by investigating teleconnections between spawning stock biomass and recruitment patterns with a 5-year time lag. We introduced and compared three analytical approaches for identifying and leveraging spatial-temporal patterns: Empirical Orthogonal Function analysis with Generalized Additive Models (i.e., EOF or EOF-GAMs), Coupled-Factor Generalized Linear Latent Variable Models (CF-GLLVM) using either a relationship matrix (referred to as Linked CF-GLLVM) or spatially varying coefficients (SVC; referred to as SVC CF-GLLVM), and spectral clustering with time-series prediction. Each method was trained using an expanding window design and evaluated on out-of-sample predictions for 2021-2023 station-level density observations. Methods were assessed across multiple complementary metrics including Spearman's correlation, mean absolute error (MAE), root-mean squared error (RMSE), F1 score for presence-absence classification, and top-10% Jaccard score for hotspot identification. The two CF-GLLVM architectures achieved the highest Spearman's correlations (Linked CF-GLLVM: $\rho = 0.728$; SVC CF-GLLVM: $\rho = 0.733$) and F1 scores (Linked CF-GLLVM: 0.782; SVC CF-GLLVM: 0.767), while the EOF-GAM achieved the best hotspot identification (Jaccard = 0.409) and was competitive on magnitude accuracy (MAE = 1,701 vs. 1,555 for Linked CF-GLLVM). Spectral clustering performed comparably to the station-mean baseline across all metrics. Sensitivity analysis across 108 CF-GLLVM model configurations revealed that the presence of spatiotemporal random effects in the recruit model consistently improved spatial predictions, while more complex spawner configurations counterintuitively degraded performance, likely by absorbing predictive signal into in-sample residual fields. Using the spatiotemporal surface for the last year when making predictions was unstable, improving predictions in two of three holdout years but leading to large error for 2021. No single method dominated across all prediction objectives, and method selection should be guided by the specific management need: the EOF is most appropriate for computationally constrained settings requiring hotspot identification; the CF-GLLVM framework offers superior Spearman's correlation, continuous spatial fields, and formal uncertainty quantification at substantially higher implementation cost; and spectral clustering is best suited to exploratory regional characterization rather than operational forecasting. The spatial teleconnection signal proved robust across a period of unprecedented stock decline, suggesting that these methods can support adaptive management even during regime-shift conditions. More broadly, this work provides a practical methodological framework for detecting and quantifying biological teleconnections in other species and ecosystems where spatially structured leading populations regulate responses in recipient populations.

1.1 Introduction

Teleconnections, the phenomenon of connection or linkage between two geographically separated events, are a common phenomenon in ecology (e.g., Alheit and Bakun 2010; Thorson et al. 2021). With teleconnected distributions, we observe two distributions that may or may not overlap in time and space, where a leading distribution regulates a response distribution to some extent (Thorson

& Kristensen 2024). These teleconnected events, while common, can be difficult to quantify spatio-temporally.

Historically, the field of teleconnections has focused on environmental teleconnections, where large-scale climate processes influence atmospheric or oceanographic systems through indirect effects (e.g., El Nino-Southern Oscillation (ENSO), Pacific Decadal Oscillation (PDO), North Atlantic Oscillation (NAO); e.g., Diaz et al. 2001, Yeh et al. 2018). In contrast, biological teleconnections, teleconnections between organisms, have not been well studied. Despite this lack of literature, identifying biological teleconnections is critical for management of renewable resources as they provide insight into how populations are connected beyond their immediate surroundings, help explain asynchronous dynamics across landscapes or seascapes, and can reveal hidden structures in ecological responses to disturbance or climate variability (Alheit and Bakun 2010). Detecting these linkages can help inform spatial management strategies, improve recruitment forecasting, and offer a deeper understanding of ecosystem resilience in the face of environmental change (e.g., Cardil et al. 2023; Serrão et al. 2025).

Current approaches to modeling biological teleconnections often rely on mechanistic models, such as the Regional Ocean Modeling System (ROMS) and larval drift models (e.g., Shchepetkin and McWilliams 2005, Sandvik et al. 2016) to model the underlying physical processes behind biological teleconnections. While these approaches are capable of simulating physical processes in detail, they are computationally intensive, highly sensitive to small changes in initial conditions (e.g., release timing or location), and typically system-specific, limiting their broader applicability (e.g., Hawkes 2000). Moreover, mechanistic models struggle to account for temporal lags or interactions unless explicitly built into the model structure, which requires extensive parameterization and system-specific knowledge (Hawkes 2000; Shchepetkin and McWilliams 2005). In addition, much of the existing literature is focused on how climate influences ecological responses rather than how different components of a population influence one another (e.g., Rodionov 1995, Schwing et al. 2010, Mondal et al. 2025).

In contrast, statistical models offer a more flexible way to quantify biological teleconnections. They can capture delayed responses between leading and response variables, incorporate interactions using autoregressive or cross-lagged structures, and account for hierarchical spatial patterns (e.g., Thorson and Kristensen 2024). These models can be generalized across species and ecosystems, allowing for exploration of inter-population teleconnections rather than only environmental factors.

Furthermore, one of the many challenges statistical models address is high dimensionality, a common feature of ecological datasets where more spatial locations are sampled than time points. High dimensionality can lead to overfitting of data (where a model works well with the training dataset, but not the validation or test dataset), high variance, and computational strain. Statistical methods can mitigate this challenge by reducing the dimensionality of a dataset by extracting dominant patterns of variability (e.g., using Empirical Orthogonal Functions (EOFs; Lorenz 1956), or Coupled-Factor Generalized Linear Latent Variable Models (CF-GLLVMs; Thorson et al. 2020) or aggregating data over broader spatial or temporal scales (e.g., clustering). These approaches help preserve prediction power while reducing noise, allowing the analysis method to work better with new data.

In this chapter, we demonstrate how EOFs, CF-GLLVMs, and clustering techniques can be used to analyze biological teleconnections. EOFs were first developed in the atmospheric sciences to identify dominant modes of variability across space and time (Lorenz 1956). EOFs utilize concepts from linear algebra to summarize multi-dimensional datasets into a smaller number of

modes of variability, which can then be analyzed. However, EOFs have limitations, including: requiring complete data, difficulties in estimating total variance across all the calculations, and computational demands for non-normally distributed data (Thorson and Kristensen 2024). To address these limitations, we also explore the Coupled-Factor GLLVM framework, built on top of EOF-GLLVMs designed by Thorson and Kristensen (2024) and Grüss et al. (2021), which extends EOF to model responses as functions of both observed and latent variables, while accommodating different distributional assumptions. Lastly, we use a form of hierarchical clustering (spectral clustering) to group data that are similar in space and value, and Generalized Additive Models (GAMs) to detect potential teleconnections between these clusters of data.

These three methods were selected to span a gradient of statistical complexity, and to represent fundamentally different approaches to the same dimensionality reduction problem. The EOF serves as a well-established baseline: it is computationally efficient, conceptually transparent, and widely used in atmospheric and oceanographic sciences for spatiotemporal pattern extraction. The CF-GLLVM extends the EOF framework by embedding it within a formal statistical model that accounts for spatial autocorrelation, accommodates non-Gaussian distributions, and estimates spatial and temporal components jointly rather than sequentially. This allows testing of whether the added statistical rigor of model-based factor analysis improves predictive performance over a purely algebraic approach. Spectral clustering provides a non-EOF alternative that does not assume the data can be represented as a linear combination of orthogonal spatial modes. Instead, spectral clustering groups survey stations into discrete regions based on similarity in their temporal behavior, spatial proximity, and observed catch rates. Together, these three approaches explore whether prediction performance is best served by continuous orthogonal decomposition (EOF), model-based latent factor analysis (CF-GLLVM), or discrete spatial regionalization (clustering), providing a practical comparison framework for researchers facing similar teleconnection problems in other systems.

To demonstrate these methods on biological teleconnections, we apply them to the spawner-recruitment data for snow crab (*Chionoecetes opilio*) in the Eastern Bering Sea (EBS). Snow crabs are a commercially and ecologically important species distributed across the EBS (Szuwalski et al. 2023). After hatching from eggs, the larvae remain in the water column for several months, where they are subjected to ocean currents and atmospheric conditions (Anger 2006). The spatial distribution of juvenile crabs (recruitment) that arises from the larval dispersal phase, coupled juvenile survival rates caused by local bottom temperature, sea ice extent, and the spatial distribution of spawning crabs makes this a useful case-study for exploring teleconnections. By comparing EOFs, CF-GLLVMs, and spectral clustering in this system, this chapter provides a scalable, generalizable framework for detecting complex, nonlinear population linkages across space and time.

1.2 Methods

1.2.1 Data

The data used to demonstrate possible snow crab spawner-recruitment teleconnections were collected during the EBS bottom trawl survey for commercially important crab species conducted every year by NOAA (Zacher et al. 2024). The data covered the years 1988-2023 (missing 2020 due to covid). The EBS bottom trawl survey dataset contains information on the catch of snow crabs at 349 stations (Supplementary Figure S.1.1.1) during daylight hours. One 30-minute bottom trawl is conducted at each station. The survey grid of the stations is based on a systematic 20 x 20 nm grid and has been consistent since 1982. In addition to catch information (including sex, size,

etc.), the survey also records location, bottom temperature, depth, and other abiotic factors (Markowitz, et al. 2024).

The dataset was divided into two subsets: spawners (**S**) and recruits (**R**). Spawners include mature female crabs, defined as female crabs of any size carrying eggs (clutch size > 0). There is no direct maturity indicator in this dataset, so the presence of eggs serves as the proxy for maturity. Recruits are crabs with carapace width between 45-55mm, assumed to be 5 years old (Sainte-Marie et al. 1995). Because of this 5-year age assumption, the recruit data were time-lagged by 5 years to align with the spawner data from the year when those recruits were spawned. For example, crabs measuring 45-55mm in the 2021 survey are assumed to have been spawned in 2016, so are analyzed alongside spawner data from the 2016 survey. To determine station-level density, the total estimated catch in numbers (sample counts multiplied by their respective sampling/raising factors) was divided by the area swept during the trawl.

For the spawners, 39 stations (11.2% of total stations) were missing at least one year of data. The remaining data for those stations represented 4.1% of the total density across all years and stations. There were also 39 missing stations (11.2% of total stations) with incomplete time-series for the recruits. The remaining data for those stations represented 5.6% of the total density across all years and stations. Missing values were interpolated using the station-specific temporal mean. Given the small proportion of missing observations, alternative imputation strategies (e.g., nearest-neighbor interpolation or time series modeling) are unlikely to meaningfully affect results. This approach was necessary because the EOF method (see below) cannot handle incomplete datasets.

1.2.2 Prediction Methodology

The following sections present three spatiotemporal prediction methods, each designed to address the same fundamental challenge: reducing the high-dimensional spawner density data into interpretable patterns, and using these patterns to predict recruitment density five years into the future. All three methods follow a consistent prediction framework: they are trained on all available data up to year t-1, then they take spawner density data for year t and produce spatially explicit recruitment density estimates for year t+5, with outputs structured identically to enable direct comparison across methods. The EOF approach decomposes both spawner and recruitment spatial patterns using singular value decomposition, then models the relationship between the spawner and recruitment temporal components using GAMs. The CF-GLLVM method extends this framework by explicitly modeling spatial dependence using a stochastic partial differential equation approach, estimating latent spatial factors while accounting for spatial autocorrelation. The spectral clustering method takes an alternative approach by first identifying coherent spatial regions with similar temporal behavior, then modeling region-specific responses to historical spawner patterns. While these methods differ in their mathematical frameworks and computational requirements, they share the common objective of extracting lower-dimensional representations of complex spatiotemporal patterns to enable robust prediction of future recruitment.

1.2.2.1 EOF

In the Empirical Orthogonal Functions (EOF) methodology (Figure 1) (Monahan 2009, Alglave et al. 2024), an anomaly matrix (taken to be the centered and standardized data matrix) with rows as spatial location, and columns as time, is constructed. Singular value decomposition (SVD) is performed, which factorizes the anomaly matrix into orthonormal basis vectors, scaling the rows and columns by singular values that capture the matrix's dominant patterns and variation, i.e.:

$$\mathbf{S} = \mathbf{U\Sigma V}^{\mathrm{t}} \tag{1.1}$$

where **S** is the anomaly matrix with dimensions *m x n* (in this study, *m* represents stations, and *n* represents years sampled). **U** is a spatial loadings matrix with dimensions *m* x *l*, where each column of **U** is the spatial loadings or "map" for a factor *l*, and *l* is the number of factors. $l = \min(m, n)$. **Σ** is a singular value matrix with dimensionality (*l* x *l*), and entries on the diagonal only, representing the amount of variation explained by each factor *l*. **V** is a temporal loadings matrix with dimensions *n* x *l*, where each column of **V** is the temporal loadings or time-series of a given factor. Interpretation of each of these matrices is as follows: for a given factor *l*, **U** [·, *l*] (the entire column) is the spatial map/pattern, **Σ**[*l*, *l*] indicates how much variability that factor explains, and **V**[·, *l*] shows how prominent that pattern is present in each year, with positive values enhancing the pattern, and negative values suggesting the opposite pattern.

The EOF was modified in one key way for this work: instead of using centered and standardized anomalies, the SVD was applied to log-transformed raw density data (number/km^2), with a small constant (1e^{-4}) added to avoid zeros. In a standard EOF analysis, the data are centered by subtracting station means prior to decomposition, so the first mode captures the largest source of variation *around* the mean. When SVD is instead applied to the uncentered log-transformed data, the mean spatial pattern dominates the first singular vector. The remaining modes then capture progressively smaller deviations from that mean, similar to what the first few modes of a centered EOF would capture. This approach avoids a separate centering step and allows the full spatial structure, including the mean, to be represented within the same decomposition.

The full SVD produces *l* factors, but in practice only the first few (n_f) factors capture meaningful patterns (see Section S.1.2 for the approach used for selecting n_f). The number of factors were held fixed for the EOF and CF-GLLVM frameworks and holdout scenarios to ensure direct comparability.

The SVD method was applied to the spawners, **S**, and the recruits, **R**. This resulted in the spatial loadings, **U**, the singular value matrix **Σ**, and the temporal loadings, **V** for both datasets. A GAM was then used to model the relationship between spawner temporal loadings and recruit temporal loadings. All model permutations were analyzed, with the response variables of the GAMs as the recruit temporal loadings ($\mathbf{V}_1^{(R)}$,, $\mathbf{V}_2^{(R)}$, …, $\mathbf{V}_{nf}^{(R)}$) and the predictor variables being the spawner temporal loadings ($\mathbf{V}_1^{(S)}, \mathbf{V}_2^{(S)}, \ldots, \mathbf{V}_{nf}^{(S)}$), where the superscript denotes which dataset it came from, and the subscript denotes the factor. Each predictor was included as a penalized smooth term, and the GAM was fit using restricted maximum likelihood to estimate smoothing parameters. Model selection was based on the conditional Akaike Information Criterion (cAIC) (Akaike 1974; Wood 2007) and deviance explained.

The GAM was used to predict the recruitment temporal loading, $\mathbf{V}_t^{(R)}$, for a specific prediction year, *t* once the best GAM for each factor were applied. The newly predicted temporal loadings were multiplied by the singular value matrix, $\mathbf{\Sigma}^{(R)}$, and spatial loading matrix, $\mathbf{U}^{(R)}$, obtained via the SVD methodology (eq 1.1) to estimate the density spatial distribution (see Supplementary Section S.1.3 for a full breakdown of the prediction steps). As a final step, all predicted densities less than the minimum observed value in the actual dataset were set to zero. This was applied because the method struggles to predict true zeros and small non-zero values skew the Spearman's correlation calculation.

1.2.2.2 CF-GLLVM

1.2.2.2.1 CF-GLLVM EFA

The methods for the Cross-Factor Generalized Linear Latent Variable Model (CF-GLLVM) (Figure 2) are built upon EOF-GLLVM models described by Thorson and Kristensen (2024) and Grüss et al. (2021):

$$Y_{b,t} \sim f\left(\mu_{b,t}, \theta\right) \tag{1.2a}$$

where $Y_{b,t}(b,t)$ is the density of crabs (spawners or recruits) at station b, f is a measurement distribution with variance parameter θ; in this specific analysis, f is the Tweedie distribution (Tweedie 1984) given the large number of stations with zero observed density.

The Simplest linear predictor takes the form:

$$g\left(\mu_{s,t}\right) = \beta_0 + \omega_b + \sum_{l=1}^{n_f} \lambda_{t,l}\epsilon_{b,l} \tag{1.2b}$$

where β_0 is an intercept independent of location and time (can be thought of as a temporal constant), ω_b is a spatial average relative to β_0 for a given location, n is the total number of factors, which was set to n_f to stay consistent with the factor count selected for the EOF method, $\boldsymbol{\lambda}_{t,l}$ is the temporal loading associated with factor *l* and time *t*, $\boldsymbol{\epsilon}_{b,l}$ is the spatial loading associated with factor *l* and location *b*. Together the factor terms capture dominant spatiotemporal covariation in density, analogous to the **U** and **V** matrices in EOF decomposition. However, the constant intercept β_0 assumes that mean abundance is stable across years, which is unlikely given the interannual variability characteristic of snow crab recruitment. This motivates replacing β_0 with a time-varying intercept:

$$g\left(\mu_{b,t}\right) = \beta_t + \omega_b + \sum_{l=1}^{n} \lambda_{t,l}\epsilon_{b,l} \tag{1.2c}$$

where $\boldsymbol{\beta}_t$ is a vector of spatially independent intercepts that vary with time *t.* Neither Equation 1.2b nor 1.2c allows the spatial pattern of recruitment to vary across years. To relax this assumption, a full-rank spatiotemporal surface $\boldsymbol{\zeta}_{b,t}$ was added:

$$g\left(\mu_{b,t}\right) = \beta_0 + \omega_b + \sum_{l=1}^{n_f} \lambda_{t,l}\epsilon_{b,l} + \zeta_{b,t} \tag{1.2d}$$

$$g\left(\mu_{b,t}\right) = \beta_t + \omega_b + \sum_{l=1}^{n_f} \lambda_{t,l}\epsilon_{b,l} + \zeta_{b,t} \tag{1.2e}$$

where $\boldsymbol{\zeta}_{b,t}$ is a full-rank spatiotemporal surface for a given location *b* and time *t,* and captures year-specific deviations in spatial structure not explained by the factor terms, paired with either a constant (1.2d) or time-varying (1.2e) intercept. Comparing Equations 1.2(b-e) to the EOF methods, equation 1.1, β_0 and $\boldsymbol{\omega}$ are akin to centering the data, $\boldsymbol{\lambda}$ is ***V,*** and $\boldsymbol{\varepsilon}$ is ***U.*** In contrast to EOF, a distribution is assumed for the spatial variables (Thorson and Kristensen 2024), i.e.:

$$\boldsymbol{\epsilon}_{b,l} \sim \mathrm{MVN}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_{b,l}}\right) \tag{1.3a}$$

$$\boldsymbol{\omega}_b \sim \mathrm{MVN}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}_b}\right) \tag{1.3b}$$

$$\boldsymbol{\zeta}_{b,t} \sim \mathrm{MVN}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\zeta}_{b,t}}\right) \tag{1.3c}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_{b,l}}$ is the spatial covariance for spatial responses to the EOF indices, $\boldsymbol{\Sigma}_{\boldsymbol{\omega}_b}$ is the spatial covariance for average spatial patterns, and $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}_{b,t}}$ is the spatial covariance for year-specific spatiotemporal surface (Thorson and Kristensen 2024). These covariances are specified and estimated using a sparse precision matrix implemented using the Stochastic Partial Differential Equation (SPDE; Lindgren et al. 2011) method. $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_{b,l}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\omega}_b}$were estimated using the same Matérn parameters (τ, κ), and $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}_{b,t}}$ was estimated using separate Matérn parameters (τ_z, κ_z). Ecologically, this structure assumes that the persistent, time-invariant spatial patterns operate at a

similar spatial scale, whereas the year-specific spatiotemporal deviations ($\boldsymbol{\zeta}$) may operate at a fundamentally different scale. For more information on the SPDE method, see Supplementary Section 1.1. Broadly, the SPDE method involves creating a finite element analysis (FEA) mesh to condense the dataset into fewer points (interpolating to a mesh of knots specified by the analyst), calculating the sparse matrices and using them to calculate a precision matrix, **Q,** using Matérn equations in TMB. The fixed and random effects parameters at those knots in the mesh are then estimated and extrapolated to a prediction grid.

The Tweedie distribution, *f*, was parameterized using a dispersion parameter σ (estimated on the log scale as $\ln(\sigma)$) and a power parameter p, where $1 < p < 2$ corresponds to a compound Poisson-gamma distribution. The power parameter was estimated via the transformation $p' = 1 + \text{logit}^{-1}(p)$, where $\text{logit}^{-1}(x) = \exp(x) / (1 + \exp(x))$, which constrains p to the (1, 2) interval.
Equations 1.2 are referred to as an Exploratory Factor Analysis (EFA) because the estimation model does not use any prior information, physiological assumptions, or structural constraints (Suhr, 2006; Grüss et al., 2021). The method is "exploratory" because it allows the underlying factor structure to emerge solely from the observed data, as opposed to being guided by a predefined hypothesis (Suhr, 2006; Grüss et al., 2021).

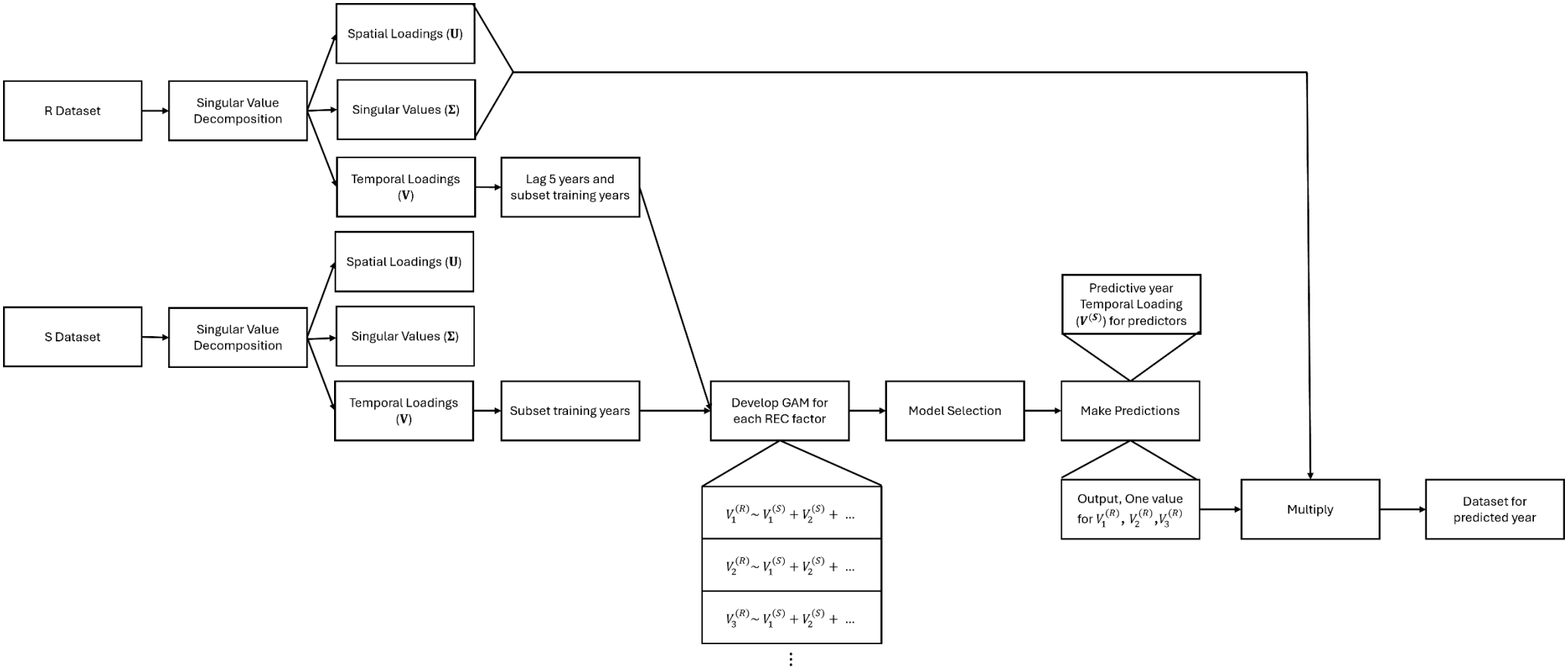


Figure 1.1. Flowchart of the Empirical Orthogonal Function (EOF) pipeline used to forecast snow crab recruitment. The spawner (**S**) and recruit (**R**) datasets undergo Singular Value Decomposition (SVD) to extract dominant spatial and temporal modes. Spawner temporal loadings are then used as predictors in Generalized Additive Models (GAMs) to forecast the recruit temporal loadings, which are subsequently recombined with the spatial basis to generate spatially explicit recruitment predictions.

1.2.2.2.2 Linked CF-GLLVM

The Linked CF-GLLVM structure was modified to follow a Confirmatory Factor Analysis (CFA) structure, which allows predictions of the spatial distribution of recruits to be made using the spawner temporal loadings, $\boldsymbol{\lambda}_{t,l}^{(S)}$. Unlike EFA, CFA utilizes predefined hypotheses to specify a model and confirm a hypothesis. This approach is used to identify teleconnections in the snow crab spawner-recruitment relationship by using the spawner crab temporal loadings to inform the recruit crabs model, i.e.:

$$Y_{b,t} \sim f\left(\mu_{b,t}, \theta\right) \tag{1.4a}$$

$$g\left(\mu_{b,t}\right) = \beta_0{}^{(R)} + \omega_b{}^{(R)} + \sum_{l=1}^{n_f}\left[\sum_{k=1}^{n_f} X_{l,k}\lambda_{t-5,l}{}^{(S)}\right]\epsilon_{b,l}{}^{(R)} \tag{1.4b}$$

$$g\left(\mu_{b,t}\right) = \beta_t{}^{(R)} + \omega_b{}^{(R)} + \sum_{l=1}^{n_f}\left[\sum_{k=1}^{n_f} X_{l,k}\lambda_{t-5,l}{}^{(S)}\right]\epsilon_{b,l}{}^{(R)} \tag{1.4c}$$

$$g\left(\mu_{b,t}\right) = \beta_0{}^{(R)} + \omega_b{}^{(R)} + \sum_{l=1}^{n_f}\left[\sum_{k=1}^{n_f} X_{l,k}\lambda_{t-5,l}{}^{(S)}\right]\epsilon_{b,l}{}^{(R)} + \zeta_{b,t}{}^{(R)} \tag{1.4d}$$

$$g\left(\mu_{b,t}\right) = \beta_t{}^{(R)} + \omega_b{}^{(R)} + \sum_{l=1}^{n_f}\left[\sum_{k=1}^{n_f} X_{l,k}\lambda_{t-5,l}{}^{(S)}\right]\epsilon_{b,l}{}^{(R)} + \zeta_{b,t}{}^{(R)} \tag{1.4e}$$

where $\boldsymbol{Y}_{b,t}$ is the density of recruits at station b and time t, f is a measurement distribution with variance parameter θ (again the Tweedie distribution), g is a link function, where $\beta_0{}^{(R)}$ is the intercept. (R) denotes the parameter belongs to the recruitment model and is estimated in this model, (S) denotes it comes from the spawner model using equations 1.2, $\boldsymbol{\beta}_t{}^{(R)}$ is a vector of spatially independent intercepts that vary with time t, $\boldsymbol{\omega}_b{}^{(R)}$ is the spatial average relative to β_0 or $\boldsymbol{\beta}_t$ for a given location, n_f is the number of factors, $\boldsymbol{\lambda}_{t,l}{}^{(S)}$ is the temporal loading associated with factor l at time t, and $\boldsymbol{\epsilon}_{b,l}{}^{(R)}$ is the spatial loadings of the recruitment model, $\boldsymbol{X}_{l,k}$ is a matrix of scalars that determines the strength of the interaction between the loadings from the model of spawners (estimated using equations (1.2) in the previous step) and the loadings from the recruitment in Linked CF-GLLVM model.

The key innovation is the $\boldsymbol{X}_{l,k}$ matrix, which determines how temporal patterns in spawners map onto those of recruits. This is a factor × factor matrix that is estimated during model fitting. $\boldsymbol{X}_{l,k}$ is constrained to be lower triangular (upper triangle set to zero) to ensure identifiability. For example, with n_f factors, $\boldsymbol{X}_{l,k}$ is a $n_f \times n_f$ matrix where $\boldsymbol{X}_{l,k}$ represents the weight connecting the spawner temporal factor l to the recruitment spatial factor k. This structure allows the model to learn which spawner patterns are most predictive of recruitment patterns.

In addition, a spatially varying coefficient (SVC) version of the model was developed to relax the assumption of a globally uniform spawner-recruit relationship. While the Linked CF-GLLVM assumes that the connection matrix $\boldsymbol{X}_{l,k}$ applies uniformly across all locations, the SVC formulation allows these linkages to change dynamically across space. This implies that if spawner factor l drives recruitment, it does so with the same relative strength everywhere in the EBS. However, the spatial heterogeneity of the EBS suggests that the influence of a given spawner temporal signal on recruitment may vary geographically. The SVC formulation relaxes this assumption by replacing the global $\boldsymbol{X}_{l,k}$ matrix and recruit spatial factors, $\boldsymbol{\epsilon}_{b,l}$, with a set of spatially varying coefficient fields $\boldsymbol{\delta}_{b,l}$, one per spawner factor. Each $\boldsymbol{\delta}_{b,l}$ field is a continuous spatial surface describing how strongly that spawner factor influences recruitment at each location, with its own SPDE prior enforcing spatial smoothness. This formulation is also more parsimonious due to $\boldsymbol{\delta}_{b,l}$ being able to be directly identifiable without any triangular constraint, whereas the Linked model (equations 1.4) requires a lower-triangular constraint on $\boldsymbol{X}_{l,k}$ to resolve the rotational ambiguity between $\boldsymbol{X}_{l,k}$ and $\boldsymbol{\epsilon}_{b,l}$.

$$Y_{b,t} \sim f\left(\mu_{b,t}, \theta\right) \quad (1.5a)$$

$$g\left(\mu_{b,t}\right) = \beta_0{}^{(R)} + \omega_b{}^{(R)} + \sum_{l=1}^{n_f} \delta_{b,l}{}^{(R)} \lambda_{t-5,l}{}^{(S)} \quad (1.5b)$$

$$g\left(\mu_{b,t}\right) = \beta_t{}^{(R)} + \omega_b{}^{(R)} + \sum_{l=1}^{n_f} \delta_{b,l}{}^{(R)} \lambda_{t-5,l}{}^{(S)} \quad (1.5c)$$

$$g\left(\mu_{b,t}\right) = \beta_0{}^{(R)} + \omega_b{}^{(R)} + \sum_{l=1}^{n_f} \delta_{b,l}{}^{(R)} \lambda_{t-5,l}{}^{(S)} + \zeta_{b,t}{}^{(R)} \quad (1.5d)$$

$$g\left(\mu_{b,t}\right) = \beta_t{}^{(R)} + \omega_b{}^{(R)} + \sum_{l=1}^{n_f} \delta_{b,l}{}^{(R)} \lambda_{t-5,l}{}^{(S)} + \zeta_{b,t}{}^{(R)} \quad (1.5e)$$

where the only change to equations 1.4 is the removal of the $\boldsymbol{X}_{l,k}$ matrix and $\boldsymbol{\epsilon}_{b,l}{}^{(R)}$, and the introduction of $\boldsymbol{\delta}_{b,l}{}^{(R)}$. $\boldsymbol{\delta}_{b,l}{}^{(R)}$ is a SVC: for each spawner factor *l*, there is a spatial field showing how strongly that spawner factor influences recruitment at each location *b*. Unlike the Linked model where a single scalar $\boldsymbol{X}_{l,k}$ governs the spawner-recruit connection globally, $\boldsymbol{\delta}$ allows this connection strength to vary continuously across space. $\boldsymbol{\delta}_{b,l}{}^{(\boldsymbol{R})}$ directly multiplies the spawner temporal loadings, so the contribution of spawner factor *l* to recruitment at location *b* in year *t* is $\boldsymbol{\delta}_{b,l}{}^{(\boldsymbol{R})} \cdot \boldsymbol{\lambda}_{t-5,l}{}^{(S)} \cdot \boldsymbol{\delta}_{b,l}{}^{(\boldsymbol{R})}$ also has its own SPDE prior:

$$\boldsymbol{\delta}_{b,l}{}^{(R)} \sim \text{MVN}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\delta}_{b,l}{}^{(R)}}\right) \quad (1.6a)$$

$$\boldsymbol{\omega}_b{}^{(R)} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}_b{}^{(R)}}) \quad (1.6b)$$

$$\boldsymbol{\zeta}_{b,t}{}^{(R)} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\zeta}_{b,t}{}^{(R)}}) \quad (1.6c)$$

As in the Linked model, $\boldsymbol{\delta}_{b,l}{}^{(\boldsymbol{R})}$ and $\boldsymbol{\omega}_b{}^{(R)}$ share the same Matérn parameters (τ, κ), while $\boldsymbol{\zeta}_{b,t}{}^{(R)}$ is estimated using separate Matérn parameters (τ_z, κ_z).

1.2.2.2.3 Identifiability Constraints

The temporal loadings matrix $\boldsymbol{\lambda}_{t,l}$ in the CF-GLLVM EFA models is constrained to be upper-triangular (below the diagonal fixed at zero) to ensure rotational identifiability. The estimated loadings are subsequently rotated using varimax-like Principal Component Analysis (PCA) rotation for interpretation. Varimax rotation redistributes variance across factors to produce a simpler structure where each spatial location loads strongly on as few factors as possible, improving interpretability without changing the overall fit of the model.

In the Linked model, the connection matrix $\boldsymbol{X}_{l,k}$ ($n_f \times n_f$) is constrained to be lower-triangular (upper triangle fixed at zero). This ensures identifiability of the mapping from spawner temporal factors to recruit spatial factors while allowing the model to learn which spawner patterns drive which recruitment patterns.

In the SVC model, no triangular constraint is needed because the spawner temporal loadings $\boldsymbol{\lambda}^{(S)}$ are passed as fixed data (not estimated), making the spatially varying coefficients $\boldsymbol{\delta}$ directly identifiable.

1.2.2.2.4 Model configuration design

Each of the three CF-GLLVM model families (EFA, Linked, and SVC) was tested under four structural configurations defined by a 2 × 2 factorial design (Table 1.1).

Configuration	Intercept	Spatiotemporal residual (ζ)
c1	Constant (β_0)	None
c2	Annual ($\boldsymbol{\beta}_t$)	None
c3	Constant (β_0)	Full-rank GMRF
c4	Annual ($\boldsymbol{\beta}_t$)	Full-rank GMRF

Table 1.1. The CF-GLLVM configurations

Configuration c1 serves as the baseline. Configuration c2 relaxes the intercept constraint by allowing a separate intercept for each year, which accommodates interannual variation in overall recruitment levels that is not captured by the EOF factors. Configuration c3 retains the constant intercept but adds a full-rank spatiotemporal residual field $\boldsymbol{\zeta}_{b,t}$, which can capture spatiotemporal variation not explained by the low-rank factor decomposition. Configuration c4 combines both relaxations and represents the most flexible model structure.

1.2.2.2.5 Predictions

To make predictions, the spawner temporal loadings from year *t*-5 are extracted from the fitted EFA model and passed through the estimated Linked and SVC parameters. For the Linked model, the prediction at station *b* for holdout year *t* is computed as:

$$g(\boldsymbol{\mu}_{b,t}) = \hat{\beta}_0{}^{(R)} + \widehat{\boldsymbol{\omega}}_b{}^{(R)} + \sum_{l=1}^{n_f}\left[\sum_{k=1}^{n_f}\widehat{\boldsymbol{X}}_{l,k}\hat{\boldsymbol{\lambda}}_{t-5,l}{}^{(S)}\right]\hat{\boldsymbol{\epsilon}}_{b,l}{}^{(R)} \quad (1.7a)$$

For the SVC model:

$$g(\boldsymbol{\mu}_{b,t}) = \hat{\beta}_0{}^{(R)} + \widehat{\boldsymbol{\omega}}_b{}^{(R)} + \sum_{l=1}^{n_f}\widehat{\boldsymbol{\delta}}_{b,l}{}^{(R)}\hat{\boldsymbol{\lambda}}_{t-5,l}{}^{(S)} \quad (1.7b)$$

These are the same as Equations 1.4b and 1.5b respectively, but applied to the holdout year. The intercept $\hat{\beta}_0{}^{(R)}$, spatial mean $\widehat{\boldsymbol{\omega}}_b{}^{(R)}$, spatial loadings $\hat{\boldsymbol{\epsilon}}_{b,l}{}^{(R)}$ (or $\widehat{\boldsymbol{\delta}}_{b,l}{}^{(R)}$), and the connection matrix $\widehat{\boldsymbol{X}}_{l,k}$ are all set at their training-period estimates. Only the spawner temporal loadings, $\hat{\boldsymbol{\lambda}}_{t,l}{}^{(S)}$, change, since the spawner data from year *t*-5 were observed during the training period. For exact details of the prediction steps, see Supplementary Section 1.3.

For configurations using annual intercepts (c2: eq 1.4c and eq 1.5c, and c4: eq 1.4e and eq1.5e), $\widehat{\boldsymbol{\beta}}_t{}^{(R)}$ is not estimated during training and must be extrapolated. Three methods for estimating $\boldsymbol{\beta}_t{}^{(R)}$ were explored: (a) the most recent value, (b) the mean of the last two recent values, and (c) a value estimating using an AR1 model of the intercept vector. All three of these versions for all of the models which used a time-varying intercept were run and included in the model selection process detailed below.

$\boldsymbol{\zeta}_{b,t}{}^{(R)}$ (eq 1.4d, 1.4e, 1.5d, 1.5e) is time-indexed and therefore has no defined value for future years; predictions are generated from the estimated structural components only (e.g., Anderson et al. 2025). We conducted a two-part sensitivity analysis to ensure no information was lost. First, the in-sample contribution of $\boldsymbol{\zeta}_{b,t}{}^{(R)}$ was quantified by comparing fitted values with and without the $\boldsymbol{\zeta}_{b,t}{}^{(R)}$ field for all c3 models (equations 1.4d and 1.5d) across the three prediction years. For each model, the percentage of variance in the log-scale fitted values attributable to $\boldsymbol{\zeta}_{b,t}{}^{(R)}$, the Spearman's correlation between fitted values with and without $\boldsymbol{\zeta}_{b,t}{}^{(R)}$, and the standard deviation of the terminal training year's $\boldsymbol{\zeta}_{b,t}{}^{(R)}$ field were computed. These diagnostics indicate how much spatiotemporal structure $\boldsymbol{\zeta}_{b,t}{}^{(R)}$ is absorbing relative to the low-rank factor components. Second,

we also tested whether the terminal training year's $\boldsymbol{\zeta}_{b,t}{}^{(R)}$ field could improve out-of-sample predictions under a persistence assumption, using $\boldsymbol{\zeta}_{b,t-1}{}^{(R)}$ as an approximation for $\boldsymbol{\zeta}_{b,t}{}^{(R)}$. For each holdout year, two sets of predictions from the Linked model were generated: (a) using only the structured components (e.g.: $\beta_0{}^{(R)} + \boldsymbol{\omega}_b{}^{(R)} + \boldsymbol{\epsilon}_{b,l}{}^{(R)} \times \boldsymbol{x}_{l,k} \times \boldsymbol{\lambda}_{t,l}{}^{(S)}$) and (b) adding the final training year's $\boldsymbol{\zeta}_{b,t}{}^{(R)}$ field projected to station locations. Both were evaluated against observed recruitment densities using Spearman's correlation, RMSE, and MAE.

All predicted densities less than the minimum observed value in the actual data were again set to zero because the method struggles to predict true zero and small non-zero values skew the Spearman's correlation calculation.

1.2.2.2.6 CF-GLLVM sensitivity design

The sensitivity analysis followed a full factorial design. For each of the three prediction years (2021, 2022, 2023), all 4 × 4 = 16 spawner-recruit configuration pairings (c1, c2, c3, c4) were tested for both the Linked and SVC model types. This yields 16 pairings × 2 model types × 3 years = 96 model fits. Each of the 16 pairings required a spawner EFA fit under the corresponding spawner configuration, producing 4 EFA configs × 3 years = 12 spawner EFA fits. In total, the sensitivity analysis encompasses 108 model fits. Combined with the three intercept extrapolation strategies tested for each annual-intercept configuration and the zeta persistence variants, this design enables systematic evaluation of how structural assumptions in both the spawner decomposition and the recruitment prediction model affect out-of-sample forecast performance.

Model selection for the CF-GLLVM Linked and SVC models involved using multiple complementary metrics computed on out-of-sample predictions for each holdout year (2021, 2022, 2023). Spearman's correlation measures the agreement between predicted and observed spatial rankings of crab density across stations, without assuming a linear relationship. Mean absolute error (MAE) measures the average magnitude of prediction errors in density units. A composite score was computed for each model in each holdout year to provide a single summary metric that balances spatial accuracy and magnitude accuracy. Within each holdout year, Spearman's correlation and MAE were normalized to [0, 1] using min-max scaling across all models tested (with Spearman's correlation coefficient scaled so that higher = 1, and MAE scaled so that lower = 1, via 1-normalized MAE). The composite score is then the geometric mean of the two normalized metrics:

$$\text{Composite score} = \sqrt{\check{\rho} \times \widetilde{MAE}} \tag{1.8}$$

where $\check{\rho}$ is the normalized Spearman's correlation value (between 0 and 1), and $\widetilde{MAE}$ is the normalized Mean Absolute Error (MAE) value (between 0 and 1). Mean composite scores across the three holdout years were used to rank models and select the optimal configuration.

1.2.2.2.7 Implementation

All models were implemented in Template Model Builder (TMB; Kristensen et al. 2016) using custom C++ objective functions compiled and linked via R. The spatial random effects ($\boldsymbol{\omega}$, $\boldsymbol{\varepsilon}$ or $\boldsymbol{\delta}$, and $\boldsymbol{\zeta}$ where applicable) were integrated out using the Laplace approximation (Tierney & Kadane 1986). Fixed-effect parameters were optimized using the nlminb optimizer (Gay 1990) with a maximum of 10,000 function evaluations and 10,000 iterations. Standard errors were obtained via the delta method using the TMB sdreport function (Kristensen 2016).

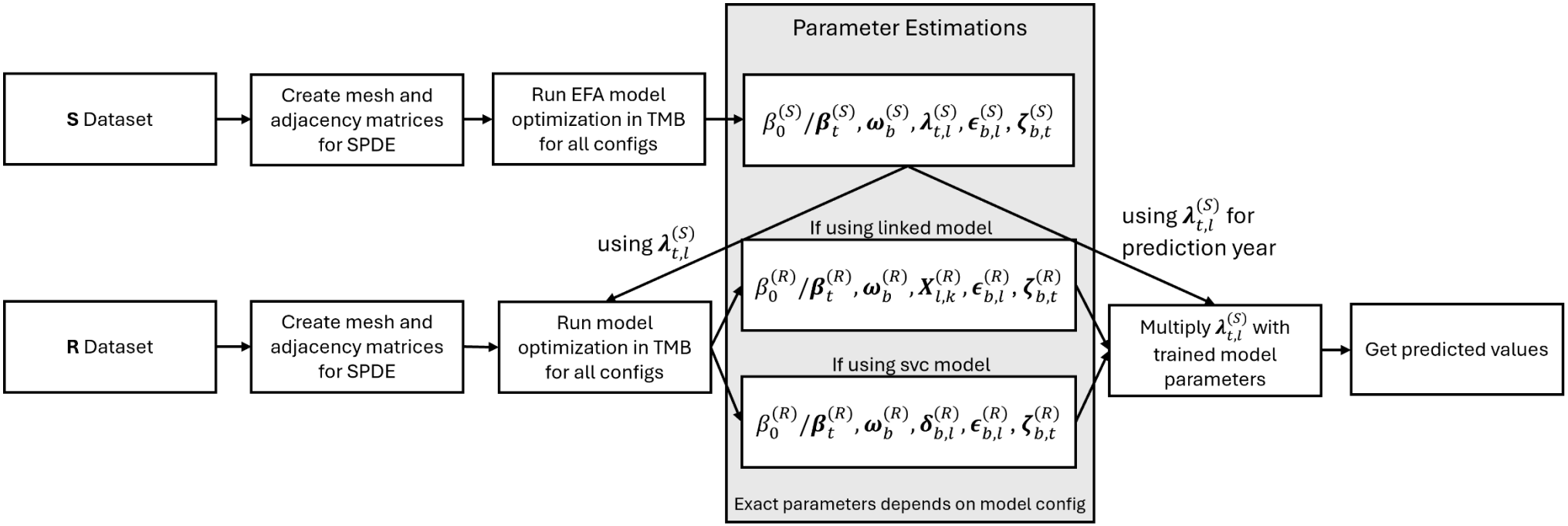


Figure 1.2. Flowchart of the CF-GLLVM Confirmatory Factor Analysis (CFA) pipeline. Spawner density data are first decomposed using an Exploratory Factor Analysis (EFA) to extract spatial modes ($\boldsymbol{\varepsilon}$) and temporal loadings ($\boldsymbol{\lambda}^{(S)}$). The temporal loadings are then passed as fixed data to the CFA recruitment model, which estimates the spawner-recruit connection (via the $\boldsymbol{X}$ matrix or spatially varying coefficients $\boldsymbol{\delta}$), recruit spatial modes, and the spatial mean ($\boldsymbol{\omega}$). Four structural configurations (c1–c4) are tested, defined by the presence or absence of annual intercepts ($\boldsymbol{\beta}_t$) and a full-rank spatiotemporal residual ($\boldsymbol{\zeta}$). For out-of-sample prediction, estimated parameters are held fixed and spawner loadings from t-5 are passed through the fitted prediction model (linked or SVC) to generate spatially explicit recruit density forecasts.

1.2.2.3 Clustering

Many common clustering methods exist (e.g. K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM)) but spectral clustering (von Luxburg 2007) was chosen for this analysis because it performs well in high-dimensional settings and can identify arbitrarily shaped clusters by focusing on connectivity between data points rather than their density or distance in the original space (von Luxburg 2007). Unlike regular K-means clustering, which assumes spherical clusters, spectral clustering first transforms the data into a lower-dimensional space that captures the underlying structure of pairwise similarities between observations before applying K-means, making it better suited to identifying clusters of irregular shape. The idea for this method was to take a fundamentally different approach to dimensionality reduction than the EOF-based methods. Rather than decomposing the entire spatial field into continuous modes of variability, spectral clustering first partitions the survey domain into discrete spatial regions where stations share similar temporal dynamics, then models the spawner-recruit relationship at the aggregated level. This serves two purposes: it tests whether the teleconnection signal is better captured by discrete spatial units than by continuous spatial modes, and it provides a comparison point from outside the EOF framework. If biologically meaningful spatial boundaries exist in the EBS then clustering may identify them more naturally than a global decomposition that assumes spatial patterns are linear combinations of orthogonal modes.

Clustering methods focus on grouping data points that share characteristics together, while separating points with little or no similarities. It does this through calculating pairwise distances or similarities between points. Three types of distances were calculated for each pair of stations (i, j) in this method: spatial distance, temporal distance, and magnitude distance. Haversine distance (accounting for Earth's curvature) between station coordinates in km was used for spatial distance, dynamic time warping (DTW) (Sakoe & Chiba 1978; Giorgino 2009) for the temporal distance between station time series, and Euclidean distances between three snow crab density summary statistics (mean density, median density, standard deviation) across years for magnitude distance. DTW measures shape similarity by allowing non-linear alignment of time series, making it robust to temporal shifts or compression (Giorgino 2009). This was preferable to correlation because DTW captures the temporal sequence of patterns, not just linear association.

Each distance matrix was normalized by its maximum value to ensure comparable scales, and then transformed to similarities using a Gaussian kernel. The kernel converts unbounded distance values into similarities on (0, 1], where nearby stations receive values near 1 and distant stations decay smoothly toward 0, i.e.:

$$\boldsymbol{\Omega} = \exp(-\boldsymbol{D^2}/2\phi^2) \tag{1.9}$$

where $\boldsymbol{\Omega}$ is the similarity matrix for a given metric, $\boldsymbol{D}$ is the distance matrix, and ϕ is the kernel bandwidth, which controls the rate at which similarity decays with distance. ϕ was set to the median of non-zero distances for each metric, a standard method (Gretton et al. 2012), which calibrates the kernel so that station pairs separated by a typical distance receive moderate similarity, while pairs substantially closer or farther receive high or low similarity respectively. The combined similarity matrix was the calculated using a weighted sum:

$$\boldsymbol{\Omega}^{Combined} = w^{spatial} \times \boldsymbol{\Omega}^{spatial} + \mathrm{w}^{temporal} \times \boldsymbol{\Omega}^{temporal} + \mathrm{w}^{magnitude} \times \boldsymbol{\Omega}^{magnitude} \tag{1.10a}$$

$$1 = w^{spatial} + \mathrm{w}^{temporal} + \mathrm{w}^{magnitude} \tag{1.10b}$$

where $\mathbf{\Omega}$ is a similarity matrix, and w is a weight scalar such that the sum of weights adds to one. Weights were determined using hyperparameter tuning (described below). The diagonal of the combined similarity matrix ($\mathbf{\Omega}^{Combined}$) was set to 1, reflecting that each station is perfectly similar to itself.

After computing the combined similarity matrix, the stations were partitioned into groups where within-group similarity is high and between-group similarity is low. The similarity matrix $\mathbf{\Omega}^{Combined}$ can be thought of as the adjacency matrix of a weighted graph, where each station is a node and the edge weight between stations *i* and *j* is their combined similarity. The objective is to divide the graph into *k* groups while minimizing the total weight of edges cut, essentially separating stations that are dissimilar while keeping similar stations together.

Solving this graph-cutting problem directly is not currently computationally feasible: finding the optimal partition requires evaluating a combinatorial number of possible groupings. However, by relaxing the requirement that each station be assigned to exactly one discrete cluster and instead allowing continuous-valued assignments, the problem can be solved efficiently through the eigendecomposition of the graph Laplacian (von Luxburg 2007). The continuous solution is then converted back to discrete cluster assignments in the final step.

The spectral embedding proceeds as follows:

1. Compute the degree matrix $\mathbf{D}$, a diagonal matrix where each entry $\mathbf{D}_{i,i}$ is the sum of the *i*-th row of $\mathbf{\Omega}^{combined}$: $\mathbf{D}_{i,i} = \sum_j \mathbf{\Omega}_{i,j}^{combined}$. The degree of a node measures how strongly connected it is to the rest of the graph. Stations in dense, high-similarity regions have high degree.
2. Construct the normalized Laplacian: $\boldsymbol{L}^{normalized} = \boldsymbol{I} - \mathbf{D}^{-0.5}\mathbf{\Omega}^{combined}\mathbf{D}^{-0.5}$. The normalization by $\mathbf{D}^{-0.5}$ accounts for differences in node degree, preventing highly connected stations from dominating the clustering solution. Without normalization, the algorithm tends to isolate single outlier stations rather than finding balanced partitions.
3. Compute the eigendecomposition of $\boldsymbol{L}^{normalized}$ and extract the *k* smallest eigenvectors (excluding the trivial eigenvector associated with eigenvalue 0). Each eigenvector assigns a continuous value to every station, and stations that belong to the same cluster receive similar values. By taking *k* eigenvectors, each station is re-represented as a point in a *k*-dimensional space where the cluster structure becomes geometrically apparent. This is analogous to how EOF extracts the most informative dimensions of a spatiotemporal field. In this case, the eigenvectors extract the most informative dimensions of the similarity graph.
4. Normalize the rows of the eigenvector matrix to unit length, which improves the separation between clusters in the embedded space.
5. Apply K-means clustering to the embedded points. Standard K-means is sufficient for the final assignment because the spectral embedding has already done the hard work of making the clusters separable.

Hyperparameter tuning involved a grid search over: (a) the number of clusters: $k \in \{7, 8, 9, 10\}$; (b) the spatial weight: $w^{spatial} \in \{0.7, 0.8, 0.9\}$; (c) the temporal weight: $w^{temporal} \in \{0.1, 0.2\}$; (d) the magnitude weight: $w^{magnitude} \in \{0.1, 0.2\}$; and (e) the kernel width: $\phi \in \{0.05, 0.1, 0.2, 0.3\}$. The spatial weight was set high (0.7-0.9) because the primary objective of clustering was to identify spatially contiguous regions, with temporal and magnitude similarity serving as secondary criteria. The cluster range (7-10) was guided by preliminary silhouette analysis (Rousseeuw 1987)

indicating that fewer than seven clusters merged ecologically distinct regions while more than ten clusters produced clusters with too few stations for reliable GAM estimation. Kernel width values spanned an order of magnitude to ensure the results were not sensitive to this choice. Each combination was evaluated using silhouette scores, which measure how well each station fits within its cluster relative to other clusters. The best-performing configuration was selected for each dataset (spawner and recruitment, for each training scenario).

The mean density for each cluster in each year was calculated and used to produce a cluster-density time series. GAMs were then used to predict recruitment cluster-density time series from spawner cluster-density time series, testing all possible combinations of spawner clusters as predictors (i.e., following the approach applied for the EOF GAM modeling). For each recruitment cluster, the best GAM was selected by comparing all possible subsets of spawner clusters as predictors. Models were ranked by deviance explained and AIC, with a parsimony criterion that penalizes additional predictors, favoring simpler models when the improvement from adding predictors was marginal.

Once the best model for each recruitment cluster-density time-series was selected, predictions were generated by applying the spawner cluster-density values for the out-of-sample year to the fitted GAMs, producing an estimated recruit cluster-density value for each cluster. The predicted cluster mean was then assigned to every station within that cluster. See Supplementary Section 1.4 for more information.

As a final step, predicted densities below the minimum observed non-zero value in the survey data were set to zero. This thresholding was necessary because the GAM predicts cluster-level mean densities, which are always positive, providing no natural mechanism for predicting zero-density stations. Without this correction, stations in regions with no crabs historically would receive small but non-zero predictions, inflating error metrics.

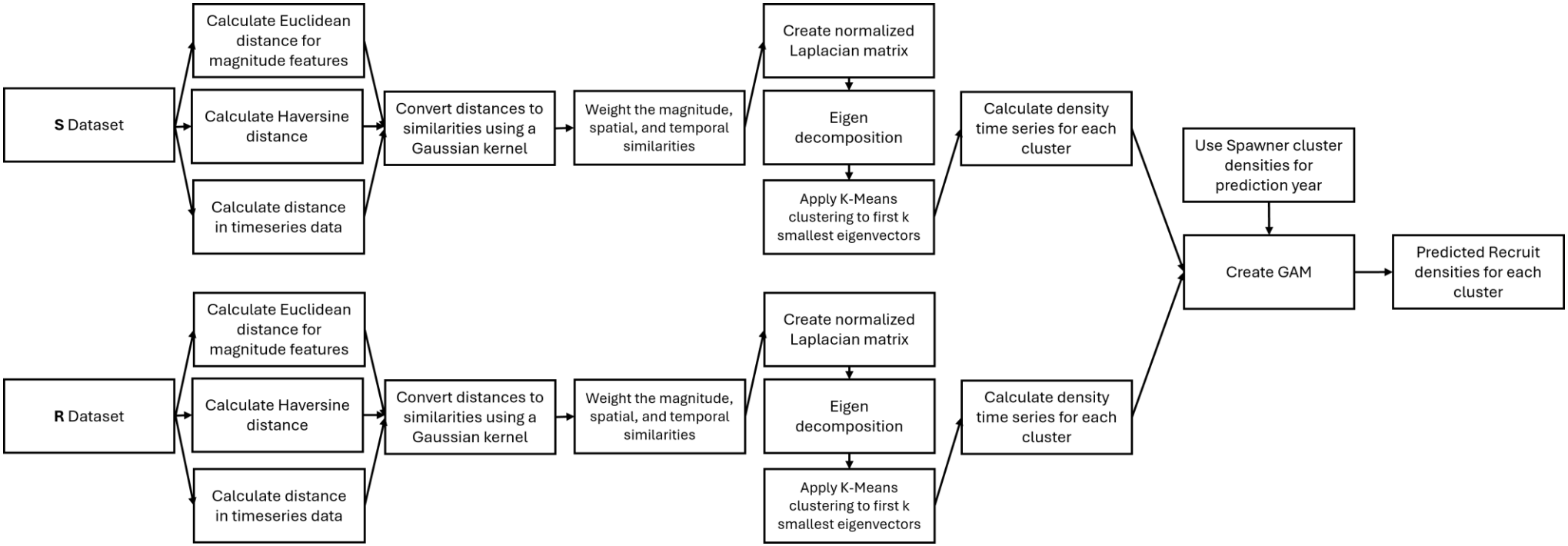


Figure 1.3. Flowchart of the spectral clustering pipeline used to forecast snow crab recruitment. Pairwise station distances (spatial, temporal, and magnitude) are converted to similarities via a Gaussian kernel and combined into a weighted similarity matrix. Spectral embedding via the graph Laplacian reduces stations to a lower-dimensional space where K-means assigns cluster labels. Mean densities per cluster produce spawner and recruit cluster-density time series, which are linked via GAMs to forecast recruit cluster means for the holdout year. Predicted cluster means are then assigned to all stations within each cluster to produce spatially explicit predictions.

1.2.3 Holdout scenarios and model comparisons

An expanding window validation approach with three prediction holdout scenarios was used to evaluate model performance for each of the methods discussed above. Each scenario predicts one recruitment year using all previous data:

- Scenario 1: Fit spawner decomposition on 1988-2016; fit spawner-recruit linking model on recruit years 1993-2019 (corresponding to spawner years 1988-2014); predict 2021 recruitment using 2016 spawner patterns.
- Scenario 2: Fit spawner decomposition on 1988-2017; fit spawner-recruit linking model on recruit years 1993-2021 (corresponding to spawner years 1988-2016); predict 2022 recruitment using 2017 spawner patterns.
- Scenario 3: Fit spawner decomposition on 1988-2018; fit spawner-recruit linking model on recruit years 1993-2022 (corresponding to spawner years 1988-2017); predict 2023 recruitment using 2018 spawner patterns.

The 2015 spawner data were excluded from training because they would correspond to 2020 recruitment data, which are unavailable. This expanding window approach mimics real-world forecasting where models are updated as new data become available.

No single metric fully captures prediction quality for spatially explicit density forecasts, because a model can correctly identify where crabs are concentrated (spatial pattern) while systematically over- or under-predicting how many are there (magnitude), or vice versa. The three methods were therefore evaluated across a suite of complementary metrics that quantify different aspects of predictive performance. Comparison of the methods was based on a number of metrics (see Table 1.2). Spearman's correlation was used to determine how well the predicted spatial distribution matches the observed spatial distribution. Mean Absolute Error (MAE) was used to compare the average magnitude of prediction error at the station level. Root Mean Square Error (RMSE) measures prediction error with greater sensitivity to large deviations than MAE, since squaring the errors before averaging disproportionately penalizes extreme mispredictions.

Comparing RMSE and MAE together reveals whether errors are distributed evenly across stations or concentrated at a few poorly predicted locations. Top-10% Jaccard measures how well a method identifies high-recruitment hotspot stations. Specifically, the metric calculates the overlap between the set of stations in the top 10% of observed densities and the set of stations in the top 10% of predicted densities. Presence-absence F1 score combines precision (the fraction of predicted zeros that are true zeros) and recall (the fraction of true zeros that are correctly predicted) into a single metric that is high only when both are high, measuring how well the method distinguishes stations with no observed crabs from stations with positive density.

Metric	Definition	Equation
Spearman's Correlation	Rank correlation between predicted and observed spatial density fields across stations; measures spatial pattern agreement independent of scale or magnitude	$1 - \frac{6\sum d_i^2}{n(n^2-1)}$ where d_i is the difference in ranks
MAE	Mean absolute error between predicted and observed recruit density across stations; measures average magnitude of prediction error with equal weight to all stations	$\frac{1}{n}\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$

RMSE	Root mean squared error between predicted and observed recruit density across stations; penalizes large errors more heavily than MAE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$
Presence-Absence F1	Harmonic mean of precision and recall for presence-absence classification, where a station is classified as occupied if predicted density exceeds the minimum observed non-zero value; high only when both false positives and false negatives are low	$\frac{2 \cdot precision \cdot recall}{precision + recall}$ where $precision = \frac{TP}{TP + FP}$ $recall = \frac{TP}{TP+FN}$
Top-10% Jaccard	Intersection-over-union overlap between the set of stations in the top 10% of observed densities and the set of stations in the top 10% of predicted densities; measures hotspot identification accuracy	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$

Table 1.2. Summary of evaluation metrics used to compare out-of-sample prediction performance across methods. All metrics were computed at the station level for each holdout year (2021, 2022, 2023) and averaged across years. For Spearman's correlation, F1, and Jaccard, higher values indicate better performance; for MAE and RMSE, lower values indicate better performance.

1.3. Results

1.3.1 Station Mean Baseline

The station-mean (Figure 1.4), which is the historical mean density at each station computed from training years, achieved a mean Spearman's correlation of 0.696 but with substantially larger errors: mean MAE of 4,207 and mean RMSE of 12,621. This reflects the baseline's inability to capture interannual variability in the spatial distribution of recruitment. All subsequent methods are evaluated against this standard.

1.3.2 EOF

The elbow point in the variance explained curve and silhouette scores (Thorndike 1953, Rousseeuw 1987) suggested that five factors should be retained for the spawner data and five factors for the recruitment data (Supplementary Figure S.1.2.1). This led to the retention of approximately 46-47% of the variance in spawner density and of 44-46% in variance of recruit density across the three holdout scenarios (holdout scenarios described in section 1.2.3). The first spawner factor consistently dominated, explaining roughly 16% of variance on its own, with subsequent factors each contributing between 5 and 11%. The first factor captures the mean spatial pattern rather than the largest deviation from it because SVD was applied to log-transformed uncentered data, with remaining factors capturing progressively smaller departures. The results for the recruit data were similar, with the first factor explaining 14-15% and the remainder declining gradually.

GAM selection identified strong nonlinear relationships between spawner and recruit factors across all three holdout years (Supplementary Table S.1.6.1). Factor 3 of the recruit field was consistently the most predictable, with deviance explained ranging from 85.0% to 86.2% across the 2021-2023 holdout years. Factor 1 was likewise well-captured, with deviance explained between 75.2% and 78.4%. Factor 5 of the recruit field was the hardest to predict, with deviance explained being only 39.0% for the 2023 holdout scenario. However, as the last factor retained, factor 5 also contributed the least total variance to the reconstruction, and its poor predictability has limited influence on spatial prediction accuracy.

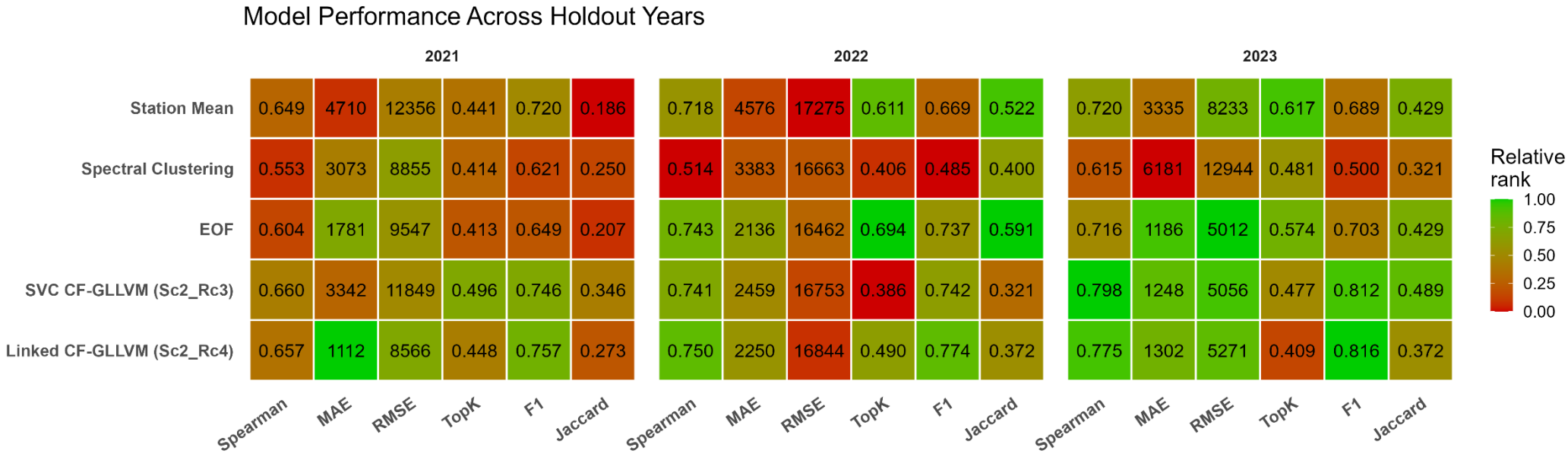


Figure 1.4. Heatmap of model performance across holdout years (2021, 2022, 2023) for the five methods. Each cell displays the metric value, with color scaled from 0 (red, worst) to 1 (green, best) using min-max normalization within each metric and year. Metrics shown are: Spearman's correlation (spatial pattern agreement between predicted and observed densities), mean absolute error (MAE; average magnitude of station-level prediction error), root mean square error (RMSE; magnitude error with greater sensitivity to large deviations), F1 score (accuracy of predicting presence versus absence of crabs at each station), and Top-10% Jaccard (overlap between the highest-density stations in the predictions and observations). Higher values indicate better performance for Spearman, F1, and Jaccard; lower values indicate better performance for MAE and RMSE. Normalization is applied so that 1 (green) always corresponds to the best-performing model for that metric and year.

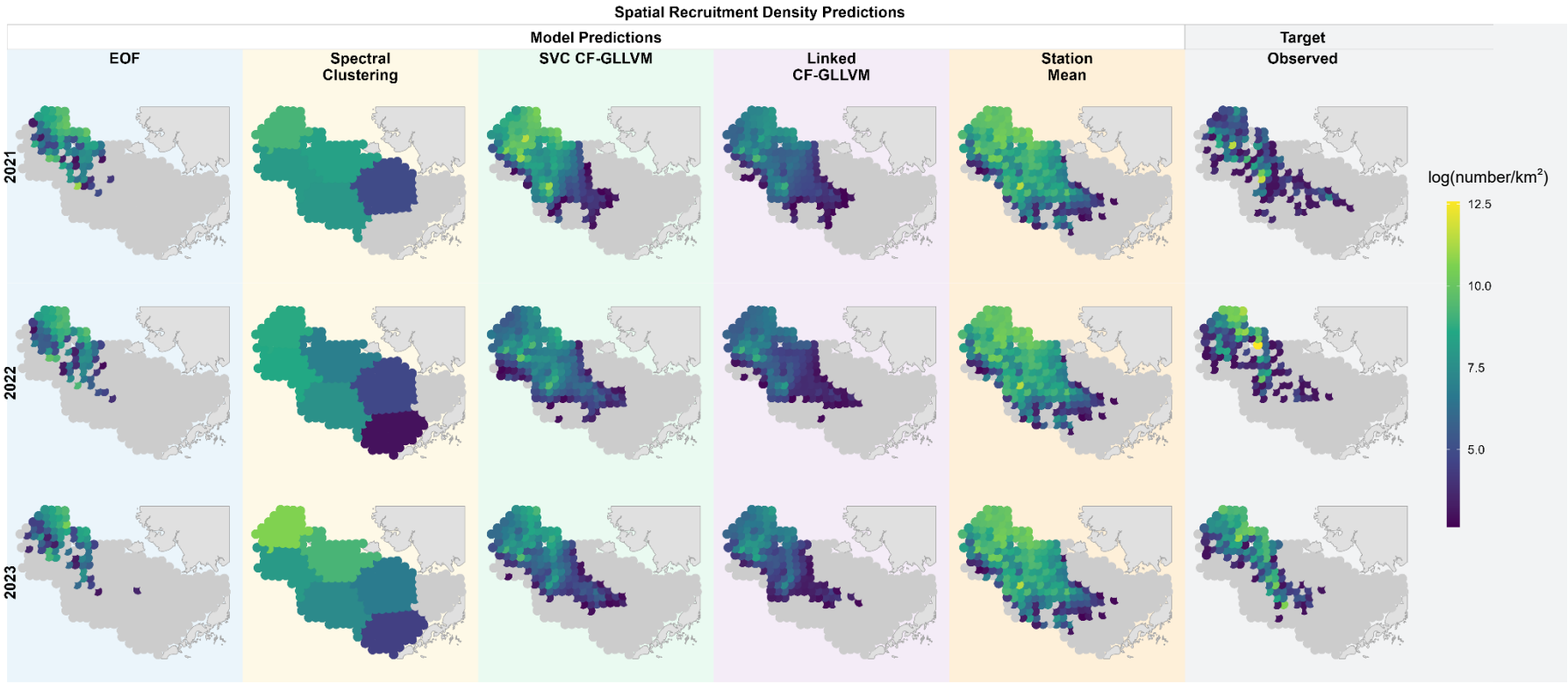


Figure 1.5. Spatial predictions of recruit density (number/km²) for each holdout year (rows: 2021, 2022, 2023) and each method (columns: EOF, Spectral Clustering, Linked CF-GLLVM , SVC CF-GLLVM, Station Mean). Color scale represents predicted density at each survey station (grey = zero density). The "target observed" column shows actual recruit densities from the EBS bottom trawl survey for comparison.

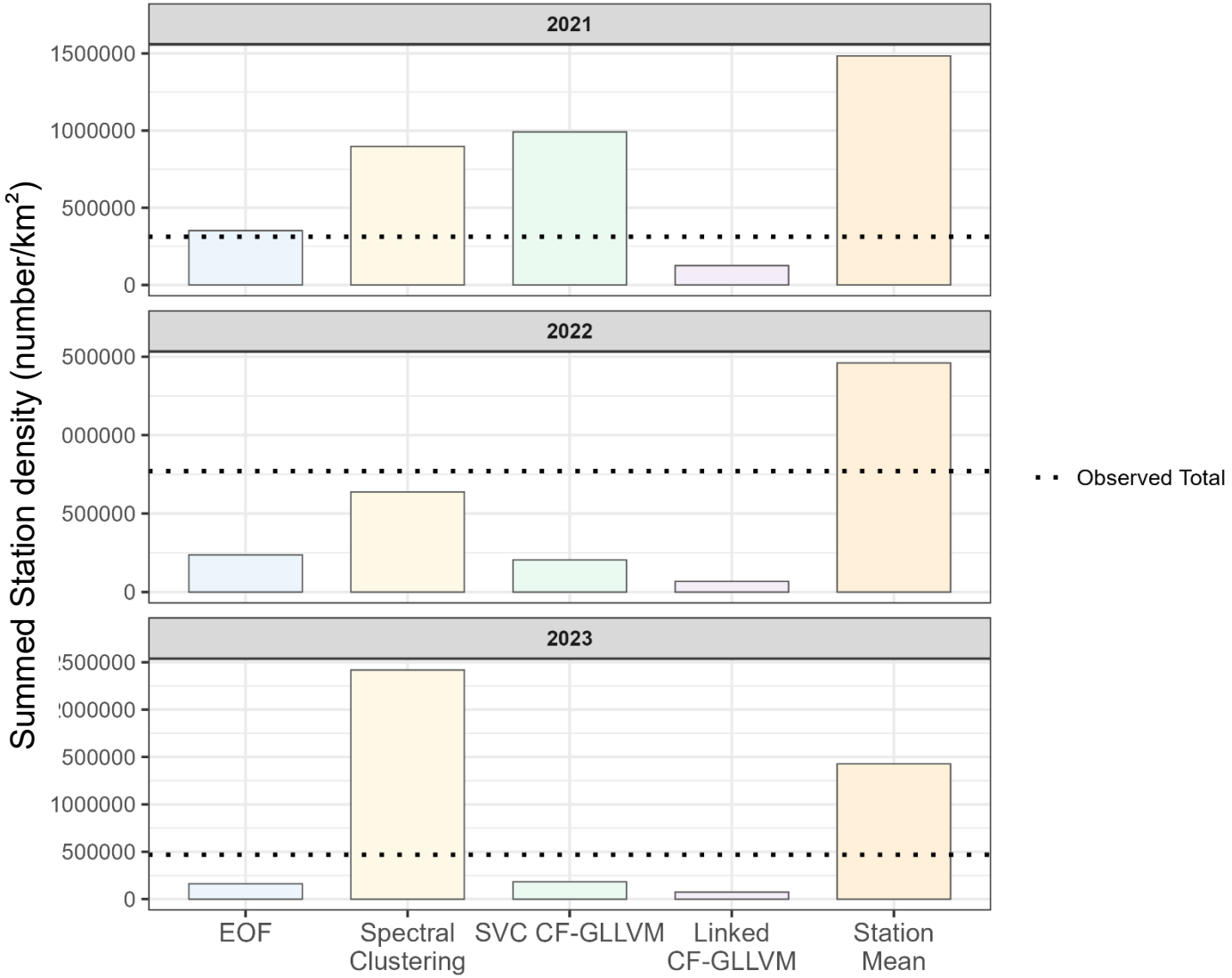


Figure 1.6. Total predicted recruit abundance (summed station density, number/km²) for each holdout year by method, compared to observed totals. Bars represent model predictions; the dashed horizontal line indicates the observed total for each year. Deviations from the observed line indicate systematic over- or under-prediction of total recruitment.

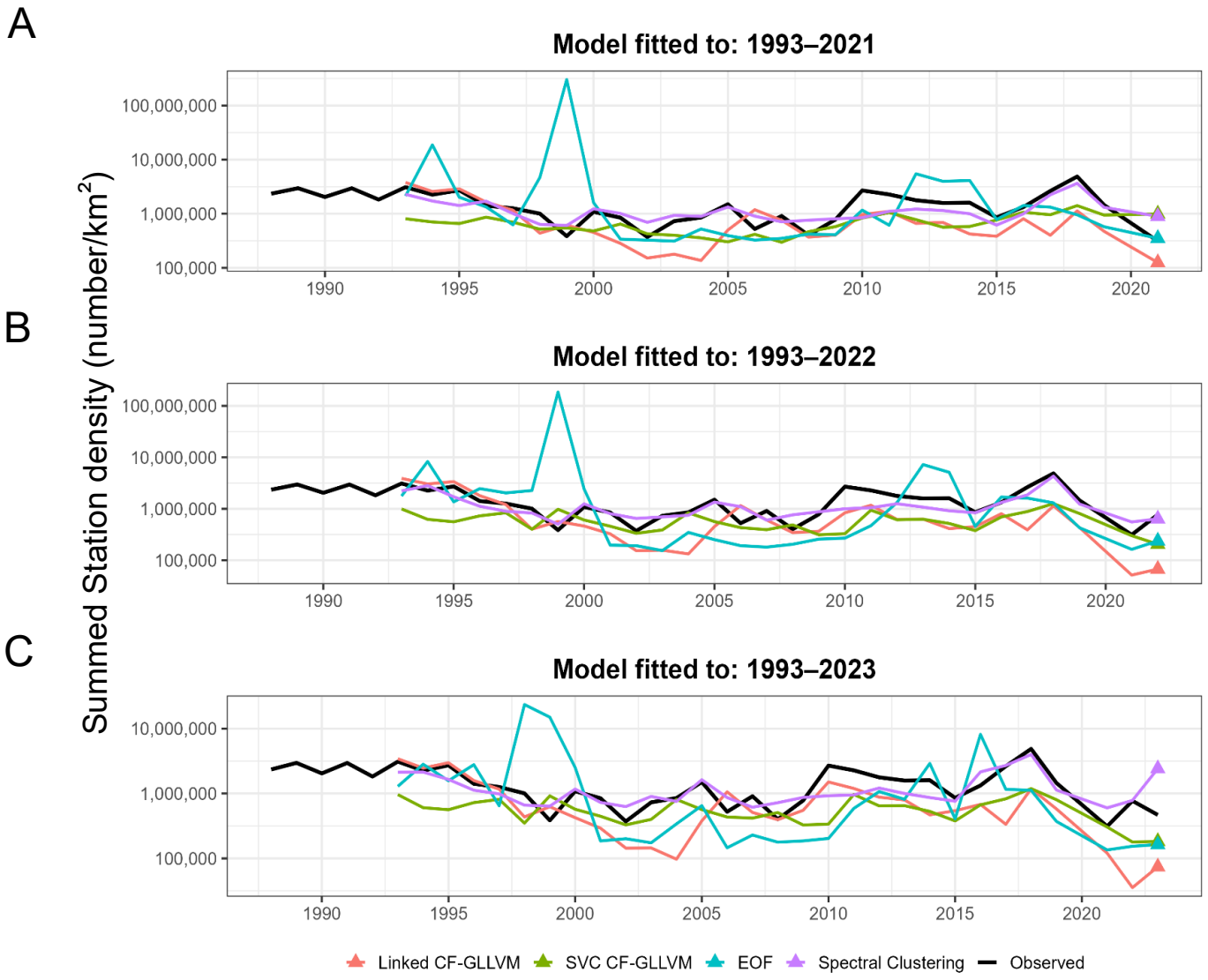


Figure 1.7. Time series of total predicted recruit abundance index (summed station density, number/km²) across holdout years (2021-2023) for each method. The observed total is shown in black; each model is shown in a distinct color.

Out-of-sample predictive performance of the EOF-GAM was moderate across the three holdout years (Figure 1.4). Spearman's correlation averaged 0.688, ranging from 0.604 in 2021 to 0.744 in 2022. MAE averaged 1,701, substantially below the MAE for the station-mean baseline (4,207). RMSE averaged 10,340 crabs km^{-2}, with a high 2022 value of 16,462. The top-10% Jaccard index averaged 0.409, indicating that roughly 40% of the highest-density stations were correctly identified. The F1 score averaged 0.696 across the three holdout scenarios.

1.3.3 CF-GLLVM

1.3.3.1 Factorial configuration effects

The CF-GLLVM framework was evaluated across a 4×4 factorial of spawner (c1–c4) and recruit (c1-c4) structural configurations (Table 1.1) applied to the Linked and SVC architectures, yielding 128 configuration-architecture combinations across the three holdout years, each evaluated for three beta extrapolation strategies when using the time-varying beta structure ("last", "mean2", and "AR(1)"), for a total of 384 evaluated predictions. Results varied substantially across configurations.

The structural configuration of the recruit model had the largest overall impact on out-of-sample Spearman's correlations. When evaluating the recruit models, including a free spatiotemporal residual field (ζ) consistently improved predictive accuracy. Averaged across all spawner configurations, the c3 recruit configuration (constant intercept plus ζ) achieved the highest mean Spearman's correlation (ρ = 0.724), followed by the c4 configuration (annual intercepts plus ζ, ρ = 0.704). In contrast, configurations c1 and c2 lacked the ζ field and relied solely on constant or annual intercepts, which degraded their average Spearman's correlation to approximately 0.660, comparable to the station-mean baseline.

Conversely, the spawner structural configurations demonstrated a counterintuitive pattern: simpler configurations produced better downstream recruit predictions. The baseline c1 spawner configuration (constant intercept, no ζ) was associated with the highest mean Spearman's correlation (ρ = 0.698), while the highly complex c4 spawner configuration yielded the lowest (ρ = 0.667). This performance drop likely indicates that complex spawner models are overfitting to the training data. By absorbing the underlying predictive signal into the in-sample residual fields, these models degrade the quality of the temporal loading vectors used for forward prediction, a dynamic explored further in Section 1.4.

For the intercept extrapolation, mean2 (average of the two most recent training-year intercepts) had a marginal advantage over last-year and AR(1) extrapolation for the Linked models (mean Spearman's correlation coefficients of 0.686 vs. 0.682 and 0.680), while all three strategies performed equivalently for the SVC models (all ≈ 0.682). For models without annual intercepts (c1 and c3), prediction used a fixed constant intercept with no extrapolation required. These configurations performed comparably to the annual-intercept models (c2 and c4) despite requiring no intercept extrapolation, suggesting that the extrapolation step adds complexity without consistent benefit, and/or that the model could absorb the annual variation in the factor analysis.

1.3.3.2 Best configurations and architecture comparison

Model selection across the factorial configurations was based on the composite score (Eq. 1.8; the geometric mean of the normalized Spearman's correlation and normalized MAE, balancing rank and magnitude accuracy). The results below report the top-performing configuration under this criterion. The top-performing Linked configuration used a c2 spawner (annual intercepts, no ζ) paired with a c4 recruit structure (annual intercepts + ζ), using the mean2 extrapolation strategy

("Sc2_Rc4, mean2"). This configuration achieved a mean Spearman's correlation of 0.728, a mean MAE of 1,555, a mean RMSE of 10,227, a mean top-10% Jaccard of 0.339, and a mean F1 of 0.782 across the three holdout years. Year-specific Spearman values for this configuration were 0.657 (2021), 0.750 (2022), and 0.775 (2023) (Figure 1.4).

The top-performing SVC configuration used a c2 spawner paired with a c3 recruit structure under the constant strategy (Sc2_Rc3, constant with ζ). Across the three holdout years, this configuration achieved a mean Spearman's correlation of 0.733, a mean MAE of 2,350, a mean RMSE of 11,219, a mean top-10% Jaccard of 0.386, and a mean F1 of 0.767. Per-year Spearman's correlation coefficients were 0.660 (2021), 0.741 (2022), and 0.798 (2023) (Figure 1.4).

Although the SVC top-performing configuration ranked first on the composite score, this ordering reflects the relatively compressed range of Spearman's correlation coefficients across all configurations (roughly 0.63-0.81), such that small differences in Spearman's correlation contribute less to the composite score than the larger differences in MAE. The Linked model's best configuration MAE of 1,555 represents a 33% improvement over the SVC's 2,350, which is a meaningful gain in magnitude accuracy even as the difference in mean Spearman's correlation coefficient (0.728 vs. 0.733) falls well within the noise of a three-year evaluation window.

Averaged across all 108 tested configurations, the SVC architecture had a slightly higher mean Spearman's correlation coefficient (0.685 vs. 0.684 for Linked) but substantially higher MAE (12,075 vs. 2,621 for Linked). The Linked architecture was generally more robust on point-accuracy metrics across all configuration choices, while the SVC architecture demonstrated higher maximum Spearman's correlation coefficients (best: 0.810 vs. 0.780) but at the cost of greater sensitivity to configuration choice.

The connection matrix $\boldsymbol{X}$, which links recruit factors to spawner temporal loadings, had a lower-triangular structure consistent with the hierarchical factor constraints imposed during estimation (Figure 1.8). Across the three holdout years, active connections were concentrated in the lower rows (factors 3-5), with near-zero loadings in rows 1 and 2 across all years. This pattern is notable because the model was not forced to zero out these upper rows but still chose this anyway. What emerged is effectively a form of implicit regularization: the model used only the factor connections it needed to explain the data, concentrating the spawner-recruit linkage in the lower-order factors while leaving the upper rows uninformative. Factors 1 and 2 of the recruit spatial decomposition contributed little to the teleconnection signal, though their precise spatial interpretation requires examination of the rotated factor maps (Supplementary Figure S.1.8.3).

Post rotation, recruit factor variance was heavily concentrated in the first one to three factors: the first factor alone accounted for 83.6% of recruit spatial variance for the 2022 holdout scenario, with the remaining 16.4% in factor 2. The 2021 and 2023 solutions were somewhat more distributed (60.4/27.7/12.0% and 71.7/22.5/5.8%, respectively), but in all cases three or fewer factors captured effectively all recruit spatial variance.

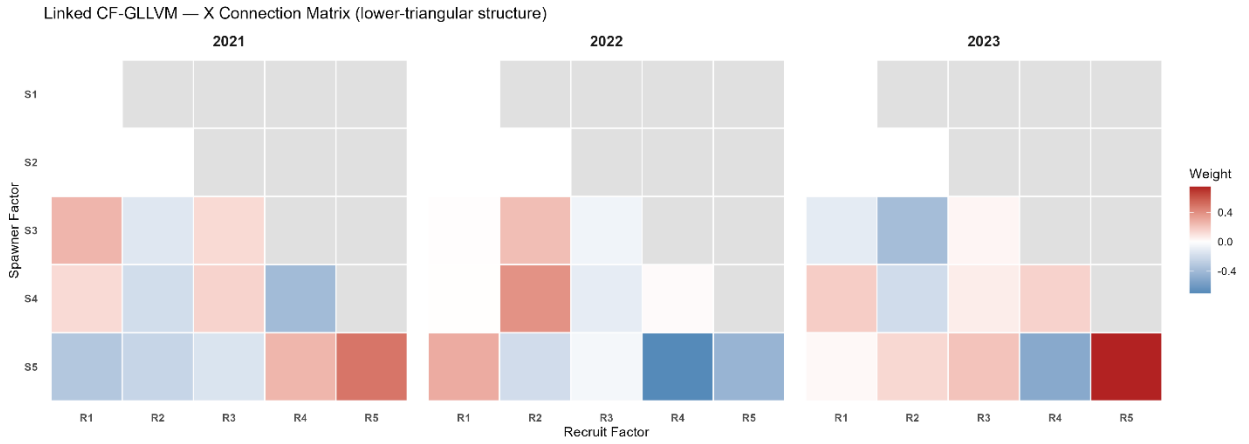


Figure 1.8. Estimated connection matrix ($\widehat{\boldsymbol{X}}$) from the best-performing CF-GLLVM configuration (spawner c2, recruit c4, mean-2 extrapolation) for each holdout year. Each entry $\widehat{\boldsymbol{X}}_{l,k}$ represents the estimated weight connecting spawner temporal factor *l* (columns) to recruit spatial factor *k* (rows). The matrix is constrained to be lower-triangular for identifiability (upper triangle fixed at zero). Non-zero entries indicate spawner temporal patterns that are predictive of recruitment spatial patterns five years later. Values near zero suggest weak or no teleconnection between that spawner-recruit factor pair.

1.3.3.3 Role of the free Spatiotemporal surface

The spatiotemporal surface ζ, estimated as part of configurations c3 and c4, accounted for a small fraction of in-sample log-density variance in the recruit models. Across the three holdout scenarios, ζ explained an average of 2.4-2.6% of log($\hat{y}$) variance for Linked, and 2.6-3.0% for SVC, compared to only 0.1% in the spawner EFA models. The terminal training year's ζ surface showed substantial spatial spread, with standard deviations ranging from 0.97 to 1.41 log-density units across model and holdout combinations and spatial ranges spanning roughly ±3 to ±4.5 log-density units.

Despite this in-sample influence, using the terminal ζ surface for the prediction year did not consistently improve out-of-sample performance. For example, adding the ζ surface (taken from the most recent training year) to the CFA c3 configuration improved the Spearman's correlation coefficient in 2022 (0.745 to 0.755) and 2023 (0.769 to 0.787), but substantially degraded 2021 performance, increasing RMSE from 13,327 to 55,913. Averaged across holdout years, the mean Spearman's correlation coefficient with the persisted surface was 0.722 versus 0.716 without, but the mean RMSE was 27,303 versus 11,736. Given this instability, predictions reported for the best GLLVM configurations in the final comparison exclude the persisted $\boldsymbol{\zeta}$ term.

1.3.4 Spectral clustering

The spectral clustering approach grouped stations into $k = 7$ clusters for all three holdout scenarios (Supplementary Figure S.1.9.1), with optimal hyperparameters converging consistently on high spatial weight (0.8), low temporal and magnitude weights (0.1 each), and ϕ = 0.2-0.3. Silhouette coefficients were modest but stable, ranging from 0.478 to 0.495 across spawner and recruit cluster solutions, suggesting that the spatial structure of the eastern Bering Sea survey domain supports a moderately well-separated 7-cluster partition. A summary of the best GAM chosen by model selection are listed in Supplementary Table S.1.6.2.

Predictive performance of spectral clustering was the weakest of all methods evaluated (Figure 1.4). The mean Spearman's correlation across holdout years was 0.561, compared to 0.688 for the EOF-GAM, with per-year values of 0.553 (2021), 0.514 (2022), and 0.615 (2023). The mean MAE

was 4,212 and the mean RMSE was 12,821, both comparable to the station-mean baseline (MAE = 4,207, RMSE = 12,621). The top-10% Jaccard averaged 0.324, the lowest of any method. These results suggest that the coarse spatial aggregation imposed by the cluster structure suppresses within-cluster variability and limits the method's ability to resolve fine-scale spatial heterogeneity in recruitment. F1 for spectral clustering was the poorest of any method (0.535), a direct consequence of the prediction mechanism since every station within a cluster receives the same predicted density value since the method has no way to distinguish stations at the high and low ends of the within-cluster density distribution, collapsing all within-cluster variation to zero.

1.3.5 Method comparison

Across all methods, the two CF-GLLVM architectures achieved the highest Spearman's correlations (Figure 1.4). SVC ranked first at 0.733, followed closely by Linked at 0.728. The station-mean baseline (0.696) ranked third on Spearman's correlation, narrowly ahead of the EOF-GAM (0.688), while spectral clustering performed much more poorly than the other methods (0.561).

The pattern reverses when point accuracy is considered. Linked achieved the lowest MAE (1,555) and RMSE (10,227), followed by the EOF-GAM (MAE = 1,701, RMSE = 10,340). CFA SVC improved on the station-mean baseline (MAE = 2,350 vs. 4,207; RMSE = 11,219 vs. 12,621) but performance was poorer than for Linked on magnitude accuracy. Spectral clustering performed similarly to the station mean on absolute error (MAE = 4,212, RMSE = 12,821).

F1 scores for presence/absence classification followed a similar ranking to the Spearman's correlations: Linked (0.782) and SVC (0.767) performed best, followed by EOF-GAM (0.696), station mean (0.693), and spectral clustering (0.535). The top-10% Jaccard, which measures the overlap between observed and predicted high-density stations, showed that the EOF-GAM performed best overall (0.409), followed by SVC (0.386), the station mean (0.379), Linked (0.339), and spectral clustering (0.324). The relatively lower Jaccard value for Linked despite its high Spearman's correlation and low MAE suggests that while the CF-GLLVM captures rank ordering well across the full density distribution, it is somewhat less precise at pinpointing the specific locations of peak recruitment than the EOF-GAM.

1.3.6 Spatial Prediction and Aggregate Abundance

Visualizing the spatial predictions (Figure 1.5) reveals additional insights into model performance beyond summary statistics. The CF-GLLVM architectures produced smooth, continuous spatial gradients that closely tracked the overall distribution of the observed data, though they tended to smooth over the sharpest localized peaks. Conversely, the EOF-GAM produced a highly variable spatial surface with sharper localized peaks, contributing to its higher Jaccard score for hotspot identification. Spectral clustering produced uniform, step-function density blocks across large regional clusters, failing to capture fine-scale gradients.

When these spatial predictions are aggregated into total predicted abundance (Figure 1.6), the instability of the point estimates becomes apparent. Predicted aggregate abundance varied substantially between methods and across holdout years. While Linked CF-GLLVM and the EOF-GAM were competitive on Mean Absolute Error (MAE), no single method consistently was closest to the true observed aggregate abundance across all three years.

Furthermore, back-modeling the full historical time series of aggregate abundance (1993-prediction year) exposed severe magnitude deviations in the in-sample hindcasts (Figure 1.7). While the models successfully captured the rank-order spatial distribution (as evidenced by the

Spearman's correlations), their ability to accurately reconstruct the raw total abundance over the historical training period was poor, characterized by very substantial over- and under-estimations depending on the method.

1.4 Discussion

The results confirm that all three methods detected a meaningful spatial teleconnection signal between snow crab spawners and recruits, yet diverged substantially in which aspects of that signal they captured well, a pattern that motivates the method-specific analysis that follows.

The three methods evaluated here were selected to span a range of statistical complexity and represent fundamentally different approaches to the same dimensionality reduction problem. The EOF serves as a well-established baseline that is computationally efficient and widely used in atmospheric and oceanographic sciences. The CF-GLLVM extends the EOF framework by embedding it within a formal statistical analysis that accounts for spatial autocorrelation and non-Gaussian distributions. Spectral clustering provides a non-EOF alternative that does not assume that the data can be represented as a linear combination of orthogonal spatial modes. Together, these approaches explore whether prediction performance is best served by continuous orthogonal decomposition, model-based latent factor analysis, or discrete spatial regionalization.

1.4.1 Data considerations

Before interpreting method performance, it is worth noting several features of the underlying data that complicate all the use of the three approaches. The recruit class analyzed here, crabs with carapace width 45-55 mm, is assumed to be 5 years old (Sainte-Marie et al. 1995), and using this age assumption to assign recruits to their spawning year is a necessary simplification. In reality, there is likely a non-trivial proportion of individuals in this size range that are not exactly 5 years old: slower-growing or faster-growing animals may enter this size window at 4 or 6 years, which means matching the spawner data to a given recruit cohort is an approximation. This introduces noise into the teleconnection signal that is irreducible given the available data, and places a practical ceiling on how well any method can perform for this dataset.

Furthermore, while crabs of this size class are largely benthic and less subject to passive dispersal than larvae, they are not sessile: adults can redistribute between survey years through active movement and passive transport by currents, meaning that the spatial pattern of spawners during year t is an imperfect record of where spawning actually occurred relative to where the resulting recruits eventually settle. These sources of spatial mismatch are inherent to the data structure and cannot be resolved without individual tracking data or more detailed larval drift modeling.

The analysis period also spans one of the most dramatic population fluctuations in the history of the EBS snow crab stock. Snow crab abundance declined sharply during the early 2020s, with the 2022 stock assessment (Szuwalski 2022) identifying a near-collapse relative to historical levels, attributed to a combination of warm bottom temperatures, reduced sea ice extent, and possible food limitation during the 2018-2019 marine heatwave (Szuwalski et al. 2023). The holdout years in this analysis, 2021, 2022, and 2023, therefore do not reflect just routine interannual variability but the start of what might be a regime shift. On one hand, this makes the prediction problem harder: models trained on data from a historically high-abundance period must generalize to a fundamentally different state. On the other hand, this is precisely the context in which accurate spatial recruitment forecasts matter most. Early detection of where recruits are concentrated, even when overall abundance is depressed, is critical for survey design and for

informing management decisions about spatial closures or harvest allocation. Spearman correlations of 0.60-0.80 across methods indicate that the spatial distribution of recruitment is more predictable than its magnitude, and that this distributional signal persists even under potential regime-shift conditions.. However, there is a decline in the performance of exact density prediction, reflected in the MAE values across holdout years. The 2022 holdout, which falls well within the collapse period, led to the highest MAE of any year across all methods, driven by the difficulty of predicting the magnitude of an unprecedented decline from training data that contained no comparable event. The spatial rank correlations, which are scale-independent, remained more stable across years, suggesting that the spatial structure of the teleconnection persisted even as total abundance collapsed.

1.4.2 EOF-GAM

Despite being outperformed on the Spearman's correlation coefficient and F1 metric, the EOF-GAM approach performed best in terms of the highest top-10% Jaccard metric of any method (0.409), indicating that it is the most reliable at correctly identifying the specific stations with the highest recruitment. The Jaccard metric has a natural ceiling given that any spatial mismatch between the predicted and observed top-10% sets reduces the score, and though a Jaccard score of 0.409 still represents meaningful hotspot overlap, more than half of true hotspot stations were not captured in the top-10% predictions. One reason may be that the SVD basis functions, by construction, span the dominant spatial modes of the observed recruit field directly, giving the reconstruction strong information about where high-density areas tend to occur. This ability to recover the spatial identity of hotspots, not just the overall rank ordering, is important for managers focused on identifying high-density patches for survey design or harvest allocation.

The EOF-GAM method is also the easiest of the three methods to implement. The full pipeline requires only linear algebra operations (available in any scientific computing environment), GAM fitting (a mature, well-documented class of models in R and Python), and matrix multiplication for reconstruction. On the hardware used here, the entire EOF fitting and GAM selection procedure ran in under a minute. The method is also easier conceptually: the decomposition into spatial modes and temporal loadings has a direct interpretation in terms of dominant patterns and their year-to-year prominence, and the GAM relationships between spawner and recruit temporal loadings can be visualized and explained intuitively.

The method's primary structural limitation is that it cannot handle missing data. Because SVD requires a complete matrix, the 39 stations with incomplete time series had to be imputed using station-specific temporal means prior to decomposition. This approach introduces a known bias: imputed stations contribute artificially reduced variability to the decomposition, which may compress the estimated spatial modes slightly (Little and Rubin 2019). A larger concern is that the method has no mechanism for spatial smoothing or uncertainty quantification. Predicted densities at each station are deterministic functions of the GAM-predicted temporal loadings and the fixed SVD basis, with no propagation of the uncertainty in the GAM predictions through to the spatial output. Finally, EOF factors can be difficult to interpret ecologically, particularly after varimax rotation, where the factors no longer have a simple relationship to the dominant axes of variance.

1.4.3 CF-GLLVM approach

The Linked and SVC architectures outperformed the EOF-GAM and the station-mean baseline on Spearman's correlation and F1 score, suggesting that model-based factor analysis recovers more of the interannual signal in the spawner-recruit relationship than algebraic SVD followed by GAM

regression. The CF-GLLVM fits a smooth spatial surface to the data during training, constrained by the SPDE prior to be spatially coherent. Prediction involves the spatial fields ($\boldsymbol{\omega}$ and $\boldsymbol{\varepsilon}$) acting as a form of spatial memory by pulling predictions toward patterns that were spatially consistent in the training data. By contrast, the EOF reconstructs the prediction by taking fixed spatial maps and scaling them by GAM-predicted numbers. If the GAM is slightly off on one temporal loading, that error gets multiplied across the entire spatial map with no correction. The CF-GLLVM is not subject to that problem because the spatial structure is estimated jointly and does not rely solely on the temporal loadings being correct.

An additional benefit of the SPDE framework is that model outputs are defined on a continuous spatial field, not just at the 349 survey stations. Predictions can be generated at any location within the mesh, including areas not sampled during a given year. This is particularly valuable for snow crab, where some stations are occasionally missed due to weather, gear issues, or access constraints, and where spatial interpolation between stations is needed for abundance index construction. The CF-GLLVM produces such interpolated fields as a natural byproduct of estimation, with uncertainty quantified using the Laplace approximation, rather than requiring post-hoc spatial smoothing.

These advantages come with real costs. Of the three methods compared here, the CF-GLLVM is by far the most technically demanding to implement. The approach requires custom C++ code compiled through TMB, careful construction and tuning of the SPDE mesh, management of multiple parameter classes (the intercept vector $\boldsymbol{\beta}$, the persistent spatial field $\boldsymbol{\omega}$, the factor spatial loadings $\boldsymbol{\varepsilon}$ or $\boldsymbol{\delta}$, the temporal loadings $\boldsymbol{\lambda}$, the connection matrix $\boldsymbol{X}$, and the spatiotemporal residual $\boldsymbol{\zeta}$), and post-hoc rotation of the estimated factors for interpretability. Fitting all 108 model configurations required approximately two days of computation. The parameter space is also difficult to navigate: mesh construction choices (cutoff distance, mesh extent) interact with the estimated Matérn range parameters in ways that are not always transparent, and poorly specified meshes can lead to convergence failures or implausible spatial range estimates. For practitioners without experience in latent Gaussian models or TMB, the barrier to implementation is substantial; though recent developments such as the `rtmb` package (Kristensen 2026), which allows TMB models to be specified entirely in R, lower this barrier meaningfully. Interpretability is also a challenge: the decomposition into $\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\varepsilon}, \boldsymbol{\lambda}, \boldsymbol{X},$ and $\boldsymbol{\zeta}$ produces rich output, but communicating what each component means to a fisheries manager or a non-specialist audience requires careful framing. One tradeoff of the SVC formulation is that while $\boldsymbol{\delta}$ directly encodes where each spawner factor influences recruitment spatially, it collapses the distinction between connection strength and spatial pattern into a single field. This means that the ability to ask how strongly spawner factor l drives recruitment overall is lost, since that global weight no longer exists as a separable scalar in the way that $\boldsymbol{X}$ does in the CFA model.

One of the more counterintuitive findings concerns the spawner structural configuration. When averaged across all recruit configurations, simpler spawner models produced better out-of-sample recruit predictions than the most complex spawner model: the c1 spawner (constant intercept, no $\boldsymbol{\zeta}$) was associated with the highest mean Spearman's correlation coefficient (0.698), while the c4 spawner configuration (annual intercepts plus $\boldsymbol{\zeta}$) led to the lowest (0.667). This is counterintuitive because c4 fits the spawner data better in-sample. The likely explanation is that the c4 spawner model, by including both annual intercepts and a free spatiotemporal residual $\boldsymbol{\zeta}$, absorbs interannual variability in the spawner field that is actually predictive of recruitment. When $\boldsymbol{\zeta}$ captures year-to-year spatial anomalies in spawner distribution, the temporal loading vectors passed to the recruit model carry less of that signal, reflecting only the low-rank factor structure,

with the remainder attributed to ζ. Simpler spawner models, which cannot offload variance to ζ, force more interannual structure into the temporal loadings, making them richer predictors when passed to the Linked model. Parsimony in the upstream model can be more important than goodness-of-fit to the training data when the goal is prediction rather than in-sample description. While the simplest spawner configuration (c1) performed best on average, the top-performing individual configurations for both Linked and SVC used a c2 spawner configuration, which suggests that a modest relaxation of the intercept constraint, without the addition of ζ, may strike the best balance between flexibility and predictive transfer.

The instability of persisting ζ to the prediction year (when performing the zeta-sensitivity analysis), which led to improved predictions in 2022 and 2023 but to a very substantial increase in RMSE for 2021, likely reflects the intersection of two unusual conditions. First, the 2020 survey was cancelled due to COVID-19, so the terminal training year's ζ field for the 2021 prediction was estimated from 2019 data, two years prior to the prediction target rather than one year. Second, 2019 preceded the population crash, meaning that the spatial pattern of recruitment density encoded in the 2019 ζ field was fundamentally incompatible with the 2021 spatial distribution, which reflected the early stages of the collapse. Carrying forward a ζ field from a high-abundance year into a crash year does not just fail to help, it actively misguides the prediction by imposing a spatial prior that is directionally wrong. This illustrates a general limitation of ζ persistence as a forecasting strategy, and why it was ultimately decided to not include it in the predictions: it implicitly assumes that the residual spatial anomaly from the most recent training year is a useful predictor of the next year's anomaly, an assumption that breaks down precisely when it is most operationally important (during rapid environmental or population transitions).

1.4.4 Spectral clustering

The poor performance of the spectral clustering approach is consistent with the hypothesis that the EBS recruitment field does not have sharp spatial boundaries that a discrete clustering approach can exploit. Compounding this, the prediction mechanism itself is fundamentally misaligned with station-level evaluation: because every station within a cluster receives the same predicted density value, the method produces a step-function spatial field that cannot recover within-cluster variability. This structural limitation directly impact performance on Spearman's correlation, inflates MAE at stations far from the cluster mean, and makes it essentially impossible to identify specific high-density stations for the Jaccard metric. The convergence on k=7 and high spatial weight across all three holdout years suggests the method found a stable partition, but the within-cluster homogeneity assumption suppresses the spatial variability that the other methods preserve. The EOF-based methods and CF-GLLVM produce spatially continuous predictions that are inherently better suited to a field such as prediction of crab recruitment, which varies smoothly across the survey domain. That said, spectral clustering is the most interpretable of the three methods in a qualitative sense. Results can be communicated as discrete spatial regions with associated density trajectories: when cluster A has high spawner density, cluster B tends to produce high recruitment five years later. This kind of region-to-region teleconnection narrative is accessible to non-specialists and may be more useful for stakeholder communication than the factor-loading structure of the EOF-based methods. Spectral clustering may therefore be most valuable as an exploratory or communicative tool rather than a primary predictive one.

1.4.5 Aggregation Bias and hindcasting limitations

The back-modeled time series (Figure 1.7) reveals a limitation in utilizing these spatial teleconnection models for total abundance estimation. Despite capturing the spatial rank-ordering effectively, the models struggled to hindcast the true historical aggregate abundance.

For the EOF-GAM, this magnitude error is a direct mathematical consequence of Jensen's inequality acting on log-transformed data. The pipeline applies SVD to log-transformed densities, truncates to five factors, reconstructs a log-space prediction, and then exponentiates to recover density. Because the exponential function is convex, positive errors in the log-space reconstruction are amplified disproportionately: a log-space overprediction of +2 produces a 7.4-fold inflation in density, while an underprediction of −2 only reduces density by a factor of 0.14. When summed across 349 stations, even a small number of stations with positive log-space errors can dominate the total, leading to systematic overestimation of aggregate abundance.

For the CF-GLLVM and spectral clustering methods, the hindcasting errors stem from spatial smoothing. These models capture the dominant modes of spatial variability and regional coherence, respectively, but in doing so they smooth over the extreme localized density peaks that can contain a disproportionate share of total abundance in a given year. Snow crab density is highly right-skewed where a small number of stations routinely account for the majority of the survey total, so even modest smoothing of these peaks substantially reduces the predicted aggregate. This is not a failure of the spatial prediction per se, but a fundamental tension between spatial pattern recovery and magnitude preservation for highly aggregated species.

Approaches that learn temporal dynamics directly, such as recurrent or transformer-based architectures trained on sequences of spatial fields, may be better suited for aggregate abundance estimation because they can learn the nonlinear mapping from spatial patterns to total abundance across multiple years, rather than predicting a single spatial field and summing it post hoc. However, such approaches face their own challenges in data-limited systems where the number of unique spatial patterns is small relative to model complexity (see Chapter 2).

1.4.6 Final comparison and advice

The tradeoffs across methods are summarized in Table 1.3. No single method dominates across all dimensions, and the appropriate choice depends on the research context and management objective. The methods with the highest interpretability (spectral clustering, station mean) performed the worst in terms of prediction, while the methods that perform best for prediction (Linked CF-GLLVM, SVC CF-GLLVM) required the greatest implementation investment. The EOF-GAM occupies a practical middle ground, offering competitive performance with substantially lower barriers to implementation.

Method	Performance	Implementation complexity	Computation time	Interpretability
EOF-GAM	High-Moderate	Low	< 1 min	Moderate
Linked CF-GLLVM	High	High	~1.5 hours per model	Low-Moderate
SVC CF-GLLVM	High	High	~1.5 hours per model	Low-Moderate
Spectral Clustering	Low	Moderate	Minutes	High
Station Mean Baseline	Moderate	Minimal	Seconds	High

Table 1.3. The performance, implementation, computation, and interpretability of each method.

The methods are not universally interchangeable, and each is best suited to particular research contexts, data characteristics, and management objectives. For example, the EOF-GAM approach is most appropriate when computational resources are limited, data are approximately complete,

and the primary goal is identifying which spatial modes of the spawner field are predictive of recruitment patterns. As an illustration, a researcher studying recruitment of Pacific sardine *Sardinops sagax* along the California Current, where survey coverage is extensive and reasonably complete, could apply the EOF-GAM to quickly characterize how upwelling-driven spatial patterns in adult biomass propagate forward to juvenile distribution, producing interpretable spatial maps and GAM response curves that are readily communicable to stock assessment teams.

The CF-GLLVM approach is most appropriate when the data are non-Gaussian or zero-inflated (as is typical for many species with patchy distributions), when spatial autocorrelation is known to be strong, when prediction at unsampled locations is needed, or when a rigorous probabilistic framework is required. The CF-GLLVM framework is also the most extensible of the three approaches. The distributional assumption, the spatial covariance structure, and the number and type of random effect components can all be modified for a specific application because the model is implemented through a custom objective function. For example, a researcher working with count data rather than continuous density could substitute a negative binomial likelihood, or one working in a domain with anisotropic spatial structure could modify the Matérn covariance accordingly. This flexibility comes at the cost of implementation effort but makes the framework broadly applicable across taxa and data types. A natural application would be a study of recruitment Pacific halibut *Hippoglossus stenolepis* in the Gulf of Alaska, where survey data are sparse in some areas and where abundance estimates at unsampled locations are needed for stock assessment purposes. The continuous spatial field and formal uncertainty quantification of CF-GLLVM make it well-suited to this kind of spatial interpolation problem, even though the implementation investment is substantial.

Spectral clustering is most appropriate when the goal is regional characterization rather than station-level prediction, when results need to be communicated to a broad audience, or as a first-pass exploratory analysis before applying a more complex method. For instance, a manager overseeing a multi-species groundfish fishery in the Bering Sea might use spectral clustering to identify coherent spatial regions where multiple species co-vary in abundance, providing a basis for spatially stratified management zones. The discrete cluster structure maps naturally on to spatially bounded management units in a way that continuous EOF modes do not.

1.4.7 Limitations and future directions

This analysis is constrained to three holdout years, which limits the precision of any performance comparison. Differences of a few hundredths of a Spearman unit should be interpreted cautiously, and a longer evaluation window would be needed to draw firmer conclusions about method performance. Factor count was fixed at five across all methods to enable direct comparison, but this choice may not be optimal for each method individually, and future work could explore method-specific factor selection.

The teleconnection framework assumes a fixed 5-year lag, consistent with the age-at-maturity assumption for the 45-55 mm recruit class. However, the effective lag likely varies across space and time in ways that this analysis cannot resolve given the proportion of individuals in this size class that are not exactly 5 years old, due to growth variability across the EBS. Incorporating a spatially or temporally varying lag structure, potentially informed by bottom temperature data (which is known to influence snow crab growth rates, Sainte-Marie et al. 2021), could improve the biological realism of the pairing between spawner and recruit data.

The methods developed here were motivated in part by the practical limitations of mechanistic larval drift models, which require extensive physical forcing data, careful parameterization of

release timing and location, and are computationally intensive and system-specific. The statistical approaches presented here provide a data-driven alternative that achieves meaningful spatial prediction skill without requiring explicit simulation of the physical dispersal process. A natural direction for future work is not to replace these statistical methods with drift models, but to ask whether physical transport information could be incorporated as structured prior information.

1.5 Conclusion

This study demonstrates that biological teleconnections between snow crab spawners and recruits in the Eastern Bering Sea are empirically supported and that relative spatial patterns of recruitment are meaningfully predictable across a range of methodological frameworks, even as accurate prediction of absolute recruitment magnitudes remains an open challenge.

The three approaches evaluated here: EOF-GAM, CF-GLLVM, and spectral clustering, each captured a meaningful signal in the spawner-recruitment relationship, yet differed substantially in which aspects of spatial prediction they did best.

The central practical finding is that increased model complexity does not always translate into improved forecast performance. The EOF-GAM, despite its simplicity, matched or exceeded the performance of CF-GLLVM architectures on hotspot identification and was competitive on magnitude accuracy, while requiring a fraction of the implementation effort and computation time. The CF-GLLVM framework offered genuine advantages in rank correlation, presence-absence classification, and spatial continuity of predictions, but these came with real costs in complexity, computation, and sensitivity to structural choices in the spawner and recruit models. Spectral clustering, while the most interpretable approach, was limited by its discrete prediction structure and performed no better than climatology on most metrics.

Several findings from the factorial sensitivity analysis carry implications beyond this specific application. The superior predictive performance of simpler spawner configurations suggests that when a model serves as input to a second model, its complexity should be calibrated for predictive transfer rather than in-sample fit. The instability of spatiotemporal residual persistence during the 2021 population crash illustrates the broader fragility of data-driven approaches during rapid environmental transitions, and underscores the value of multiple holdout years that span different regimes.

Perhaps most importantly, these methods produced meaningful spatial predictions during one of the most severe stock declines in the history of the EBS snow crab fishery. The persistence of the spatial teleconnection signal (Spearman's correlations of 0.60-0.80 even as total abundance collapsed) suggests that the underlying spawner-recruitment linkage is a structural feature of the system rather than an artifact of the high-abundance period used for training. This has direct implications for management: spatially explicit recruitment forecasts based on spawner distributions observed five years prior can inform survey design, spatial closure decisions, and harvest allocation even when the stock is in a depleted state.

Future work should extend the evaluation window beyond three holdout years, explore variable-lag structures that account for growth variability in the recruit size class, and investigate whether physical transport information from climate processes simulations can be incorporated as structured prior information in the CF-GLLVM framework. Together, the methods developed here provide a flexible and generalizable toolkit for detecting and quantifying biological teleconnections across marine ecosystems. These methods complement existing mechanistic approaches and can be adapted to other species and survey systems where spawner-recruitment relationships operate across space and time.

S.1 Supplementary Material

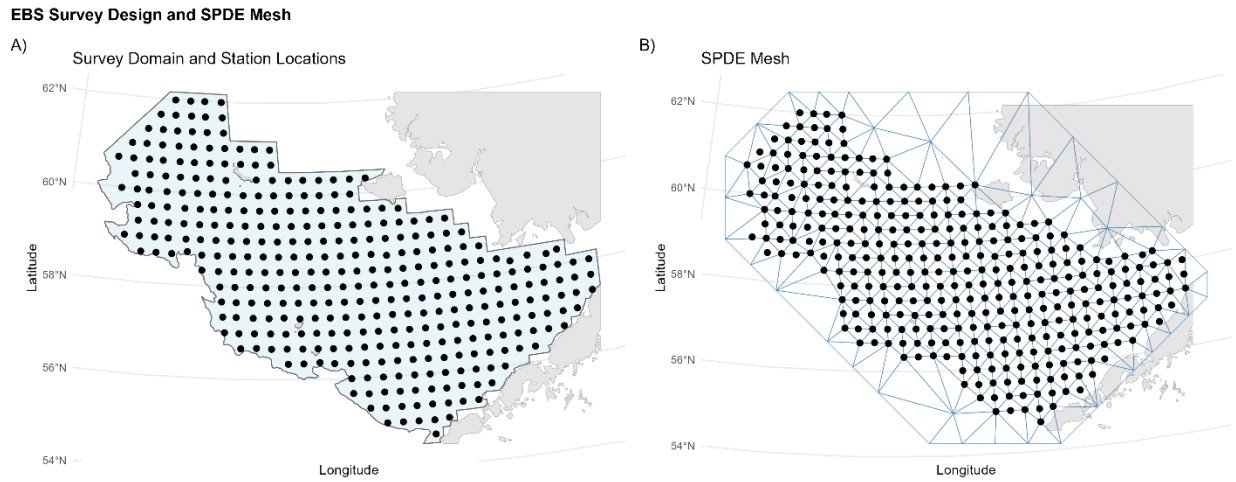
S.1.1 SPDE Mesh and Spatial Framework

The SPDE mesh used for all CF-GLLVM models was constructed using the fmesher package in R (Lindgren 2026). The mesh was built using all 349 survey stations as input locations with a cutoff distance of 40 km, which sets the minimum edge length of the triangulation and prevents overly fine mesh resolution near clustered stations. The resulting mesh contains 211 vertices. A prediction grid of 20 km × 20 km cells covering the EBS survey domain was used for spatial interpolation of model outputs. The mesh and prediction grid are shown in Supplemental Figure 1.

The SPDE approximation represents the continuously indexed Matérn Gaussian random field as a Gaussian Markov Random Field (GMRF) defined on the mesh vertices. The precision matrix $\boldsymbol{Q}$ is computed from three sparse matrices M_0, M_1, and $\boldsymbol{M}_2$ derived from the finite element basis functions, combined using the Matérn smoothness and range parameters (τ, κ) as:

$$\boldsymbol{Q} = \tau^2(\kappa^4\boldsymbol{M}_0 + 2\kappa^2\boldsymbol{M}_1 + \boldsymbol{M}_2) \quad \text{(S.1.1)}$$

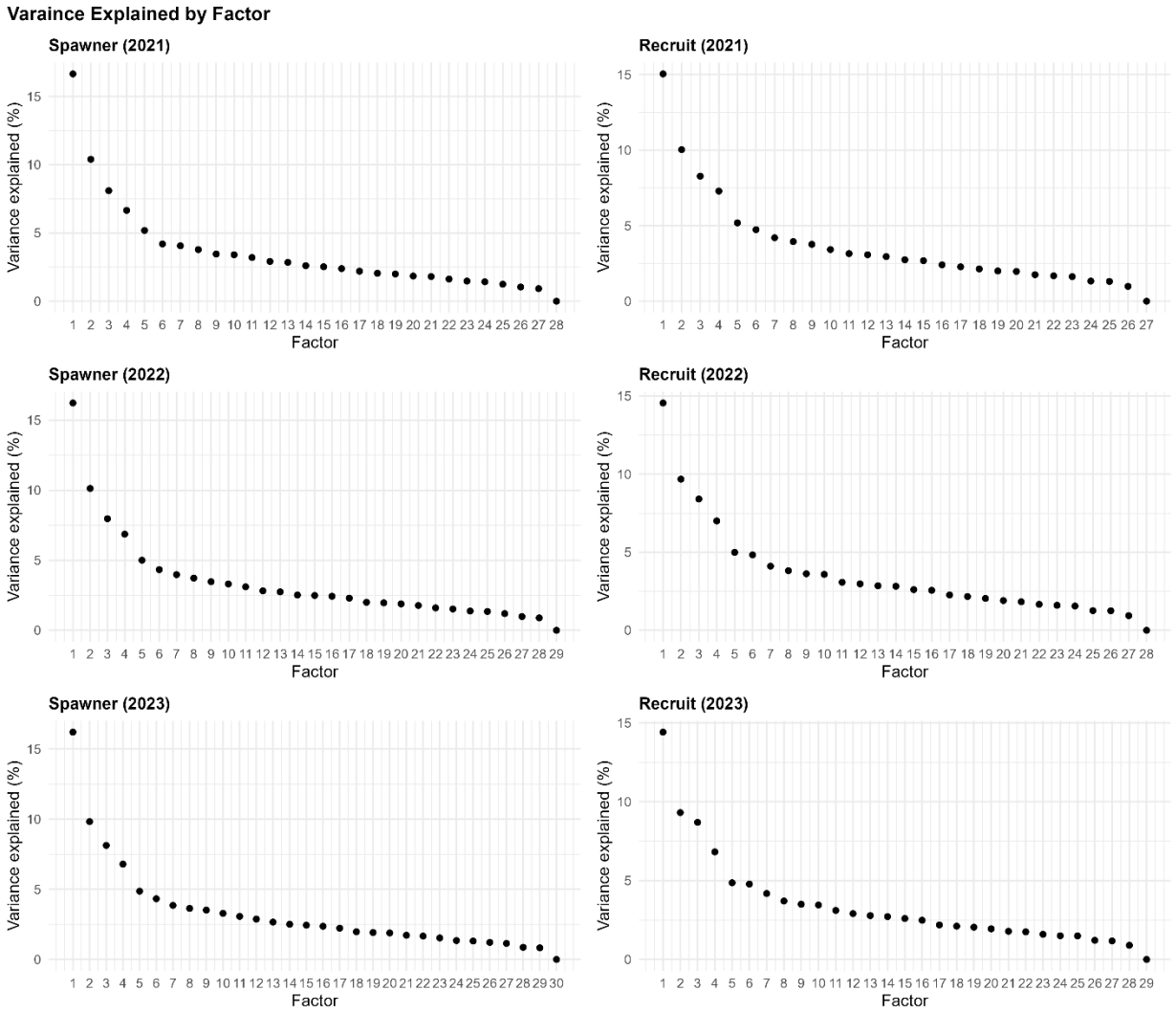
Fixed and random effects are estimated at the mesh vertices and projected to station locations and grid cell centroids using the bilinear interpolation matrices $\boldsymbol{A}_{is}$ and $\boldsymbol{A}_{gs}$ respectively.



Supplementary Figure S.1.1.1 (A) Eastern Bering Sea bottom trawl survey boundary with the 349 station locations used in this analysis. (B) Finite element mesh constructed using the fmesher package with a cutoff distance of 40 km and all survey stations as vertices. The mesh extends beyond the survey boundary to reduce edge effects in the SPDE spatial estimation.

S.1.2 Factor Selection

The factor count for all methods was determined using the elbow point in the variance explained curve (Thorndike 1953) and silhouette scores (Rousseeuw 1987) computed on the temporal loadings. For both the spawner and recruit decompositions, the variance explained curve showed a clear inflection after five factors, with subsequent factors each contributing less than 5% of the total variance. Silhouette scores were also maximized at five factors across all three holdout scenarios. To ensure direct comparability between the EOF and CF-GLLVM frameworks, five factors were retained in all model types and holdout scenarios. Supplementary Figure S.1.2.1 shows the variance explained curves and silhouette scores for all six decompositions (spawner and recruit × three holdout years).



Supplementary Figure S.1.2.1. Elbow plots of singular values from the SVD decomposition for spawner and recruit anomaly matrices across the three prediction scenarios (2021, 2022, 2023). The elbow point at five factors, where the rate of decrease in variance explained levels off, was used to determine the number of retained factors for all analyses.

S.1.3 EOF Full Prediction Steps

The following steps describe the complete prediction procedure for the EOF-GAM approach, corresponding to the methods described in Section 1.2.2.1.

1. Log-transform raw density data for both spawners and recruits, adding a small constant ($1e^{-4}$) to avoid zeros.
2. Apply SVD separately to the spawner matrix **S** (stations × years) and recruit matrix **R** (stations × years) to obtain **U**, **Σ**, and **V** for each.
3. Retain the first n_f=five factors from each decomposition based on the elbow in the variance explained curve (Supplemental Section and Figure S.1.2.1).
4. For each recruit factor l, fit a GAM with the recruit temporal loading $\mathbf{V}_l^{(R)}$ as the response and all subsets of spawner temporal loadings $\mathbf{V}_1^{(S)} \ldots \mathbf{V}_{nf}^{(S)}$ as candidate smooth predictors. Select the best GAM using AIC as a parsimony criterion.

5. For the holdout year *t*, extract the spawner temporal loadings from the fitted spawner SVD corresponding to year *t*-5.
6. Use the selected GAMs to predict the n_f recruit temporal loadings $\mathbf{V}_1^{(R)} \ldots \mathbf{V}_{nf}^{(R)}$ for year *t*.
7. Reconstruct predicted recruit log-density as: $\log(\hat{\mathbf{Y}}) = \mathbf{U}_{1-5}\, \mathbf{\Sigma}_{1-5}\, \mathbf{V}_{1-5}$
8. Back-transform to density scale.
9. Set all predicted values below the minimum observed non-zero density to zero.

S.1.4 CF-GLLVM Full Prediction Steps

The following steps describe the complete two-stage prediction procedure for the Linked and SVC CF-GLLVM models, corresponding to Section 1.2.2.2.2-3.

Stage 1 Spawner EFA CF-GLLVM:

1. Fit the EFA model (Eq. 1.2, depending on configuration) to spawner density data for all training years.
2. Extract the estimated temporal loadings matrix $\boldsymbol{\lambda}^{(S)}$ (years × factors) and apply varimax-like PCA rotation for interpretability.
3. Retain the rotated temporal loadings for all training years plus the spawner data year corresponding to the holdout recruit year (*t*-5).

Stage 2 Recruit Linked or SVC CF-GLLVM prediction:

1. Fit the Linked or SVC model (Eq. 1.4 or 1.5) to recruit density data for all training years, using the spawner temporal loadings from Stage 1 as fixed input data.
2. Set all estimated parameters ($\boldsymbol{\beta}$, $\boldsymbol{\omega}$, $\boldsymbol{\varepsilon}$ or $\boldsymbol{\delta}$, $\boldsymbol{X}$) at their training-period values.
3. Extract the spawner temporal loadings for year *t*-5 from the Stage 1 EFA output.
4. For annual-intercept configurations (c2, c4), extrapolate $\boldsymbol{\beta}_t$ to the holdout year using the selected strategy ("last", "mean2", or "AR1").
5. Compute the linear predictor at each station using Eq. 1.7a (Linked) or 1.7b (CFA SVC).
6. Back-transform to density scale by exponentiating the linear predictor.
7. Project from mesh vertices to station locations using $\mathbf{A}_{is}$.
8. Set all predicted values below the minimum observed non-zero density to zero.

S.1.5 Spectral Clustering Full Prediction Steps

1. Compute three pairwise distance matrices across all 349 stations: Haversine spatial distance, DTW temporal distance on log-transformed density time series, and Euclidean distance on density summary statistics (mean, median, standard deviation).
2. Normalize each distance matrix by its maximum value.
3. Convert to similarity matrices using the Gaussian kernel (Eq. 1.9) with the bandwidth φ set to the median of non-zero distances for each metric.
4. Combine similarity matrices using the tuned weights (Eq. 1.10).
5. Compute the normalized graph Laplacian and extract the *k* smallest eigenvectors.
6. Normalize eigenvector rows to unit length and apply K-means clustering to assign each station to one of *k* clusters.
7. Compute mean density each year for each recruit cluster to produce a cluster-level time series.
8. Fit GAMs predicting each recruit cluster time series from all subsets of spawner cluster time series; select best model by AIC with parsimony criterion.

9. Apply spawner cluster mean densities to the fitted GAMs to predict recruit cluster mean densities for the holdout year.
10. Assign the predicted cluster mean density to every station within that cluster.
11. Set all predicted values below the minimum observed non-zero density to zero.

S.1.6 GAM Selection Tables

Supplemental Table S.1.6.1 presents the best-fitting GAM for each recruit factor across the three holdout years for the EOF-GAM, including the model formula, deviance explained, AIC, and statistically significant spawner factor predictors ($p < 0.05$).

Year	Response	No. Predictors	Dev. Expl. (%)	AIC	Significant Terms ($p < 0.05$)	Formula
2021	$\mathbf{V}_1^{(R)}$	*4*	*78.4*	*-27.8*	$\mathbf{V}_1^{(S)}$; $\mathbf{V}_3^{(S)}$; $\mathbf{V}_4^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_4^{(S)}) + s(\mathbf{V}_5^{(S)})$
	$\mathbf{V}_2^{(R)}$	*3*	*62.2*	*-14.6*	$\mathbf{V}_3^{(S)}$	$s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_4^{(S)}) + s(\mathbf{V}_5^{(S)})$
	$\mathbf{V}_3^{(R)}$	*5*	*86.2*	*-41.2*	$\mathbf{V}_2^{(S)}$; $\mathbf{V}_4^{(S)}$; $\mathbf{V}_5^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_2^{(S)}) + s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_4^{(S)}) + s(\mathbf{V}_5^{(S)})$
	$\mathbf{V}_4^{(R)}$	*3*	*61.5*	*-26.0*	$\mathbf{V}_1^{(S)}$; $\mathbf{V}_3^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_2^{(S)}) + s(\mathbf{V}_3^{(S)})$
	$\mathbf{V}_5^{(R)}$	*4*	*78.3*	*-27.7*	$\mathbf{V}_1^{(S)}$; $\mathbf{V}_4^{(S)}$; $\mathbf{V}_5^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_4^{(S)}) + s(\mathbf{V}_5^{(S)})$
2022	$\mathbf{V}_1^{(R)}$	*4*	*75.3*	*-27.9*	$\mathbf{V}_1^{(S)}$; $\mathbf{V}_3^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_4^{(S)}) + s(\mathbf{V}_5^{(S)})$
	$\mathbf{V}_2^{(R)}$	*4*	*72.8*	*-20.5*	$\mathbf{V}_3^{(S)}$; $\mathbf{V}_5^{(S)}$	$s(\mathbf{V}_2^{(S)}) + s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_4^{(S)}) + s(\mathbf{V}_5^{(S)})$
	$\mathbf{V}_3^{(R)}$	*5*	*84.8*	*-42.1*	$\mathbf{V}_2^{(S)}$; $\mathbf{V}_3^{(S)}$; $\mathbf{V}_4^{(S)}$; $\mathbf{V}_5^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_2^{(S)}) + s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_4^{(S)}) + s(\mathbf{V}_5^{(S)})$
	$\mathbf{V}_4^{(R)}$	*4*	*62.9*	*-27.8*	$\mathbf{V}_1^{(S)}$; $\mathbf{V}_2^{(S)}$; $\mathbf{V}_3^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_2^{(S)}) + s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_5^{(S)})$
	$\mathbf{V}_5^{(R)}$	*4*	*67.6*	*-20.9*	$\mathbf{V}_1^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_4^{(S)}) + s(\mathbf{V}_5^{(S)})$
2023	$\mathbf{V}_1^{(R)}$	*3*	*75.2*	*-32.4*	$\mathbf{V}_1^{(S)}$; $\mathbf{V}_3^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_5^{(S)})$

	$\mathbf{V}_2^{(R)}$	*3*	*58.6*	*-18.0*	$\mathbf{V}_3^{(S)}$	$s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_4^{(S)}) + s(\mathbf{V}_5^{(S)})$
	$\mathbf{V}_3^{(R)}$	*3*	*85.0*	*-40.0*	$\mathbf{V}_1^{(S)}$; $\mathbf{V}_2^{(S)}$; $\mathbf{V}_5^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_2^{(S)}) + s(\mathbf{V}_5^{(S)})$
	$\mathbf{V}_4^{(R)}$	*4*	*66.6*	*-32.9*	$\mathbf{V}_1^{(S)}$; $\mathbf{V}_2^{(S)}$; $\mathbf{V}_3^{(S)}$; $\mathbf{V}_5^{(S)}$	$s(\mathbf{V}_1^{(S)}) + s(\mathbf{V}_2^{(S)}) + s(\mathbf{V}_3^{(S)}) + s(\mathbf{V}_5^{(S)})$
	$\mathbf{V}_5^{(R)}$	*1*	*39.0*	*-23.7*	$\mathbf{V}_4^{(S)}$	$s(\mathbf{V}_4^{(S)})$

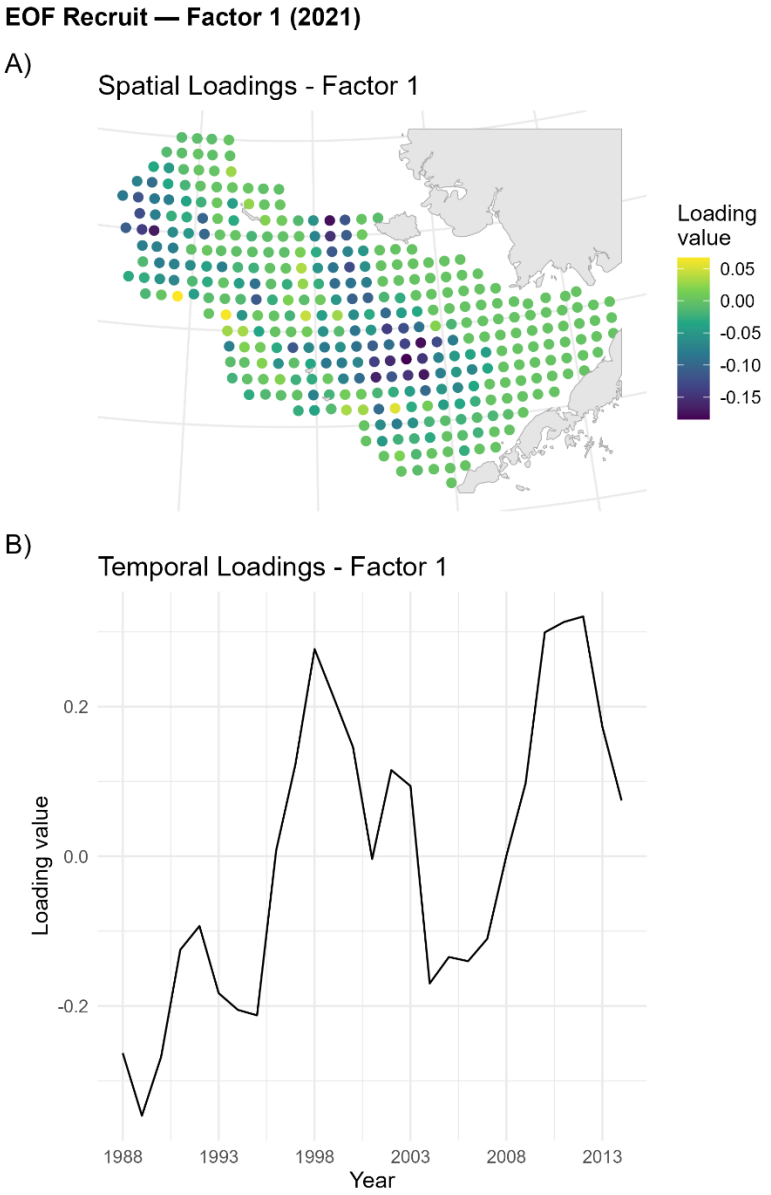
Table S.1.6.1. Best GAM for the EOF method, selected using AIC for each recruit temporal loading ($\mathbf{V}^{(R)}$) across the three prediction scenarios. Each predictor is a penalized smooth term s() of a spawner temporal loading ($\mathbf{V}^{(S)}$). Models were fit using REML. Significant terms are those with $p < 0.05$.

Supplemental Table S.1.6.2 presents the best-fitting GAM for each recruit cluster across the three holdout years for the spectral clustering approach, including the spawner clusters used as predictors, deviance explained, and AIC.

Scenario	Recruit Cluster	No. Predictors	Spawner Clusters Used	Dev. Expl. (%)	AIC
2021	1	3	1, 2, 7	74.3	47.8
	2	1	1	24.7	79.0
	3	1	4	16.1	70.6
	4	5	1, 2, 3, 4, 6	99.6	72.6
	6	3	1, 2, 4	64.7	61.9
	7	2	4, 6	54.5	116.0
2022	2	3	1, 4, 7	57.5	68.8
	3	2	1, 6	54.8	114.0
	4	4	1, 4, 6, 7	48.4	69.4
	5	3	3, 6, 7	88.2	154.0
	6	2	4, 5	66.1	70.7
	7	4	3, 4, 5, 7	77.0	68.6
2023	1	5	1, 2, 4, 5, 6	87.4	164.0
	2	4	1, 2, 4, 6	70.1	71.7
	3	4	1, 2, 6, 7	59.9	96.9
	4	3	4, 5, 7	55.7	70.8
	5	3	1, 5, 6	47.0	72.8
	6	3	5, 6, 7	59.0	121.0

Table S.1.6.2. Best GAM selected by AIC for each recruit cluster time series across the three prediction scenarios. Each predictor is a penalized smooth term s() of a spawner cluster mean density time series. Models were fit using REML. Clusters with no significant model (cluster 5 in 2021) are omitted.

S.1.7 EOF Spatial and Temporal Factor Maps

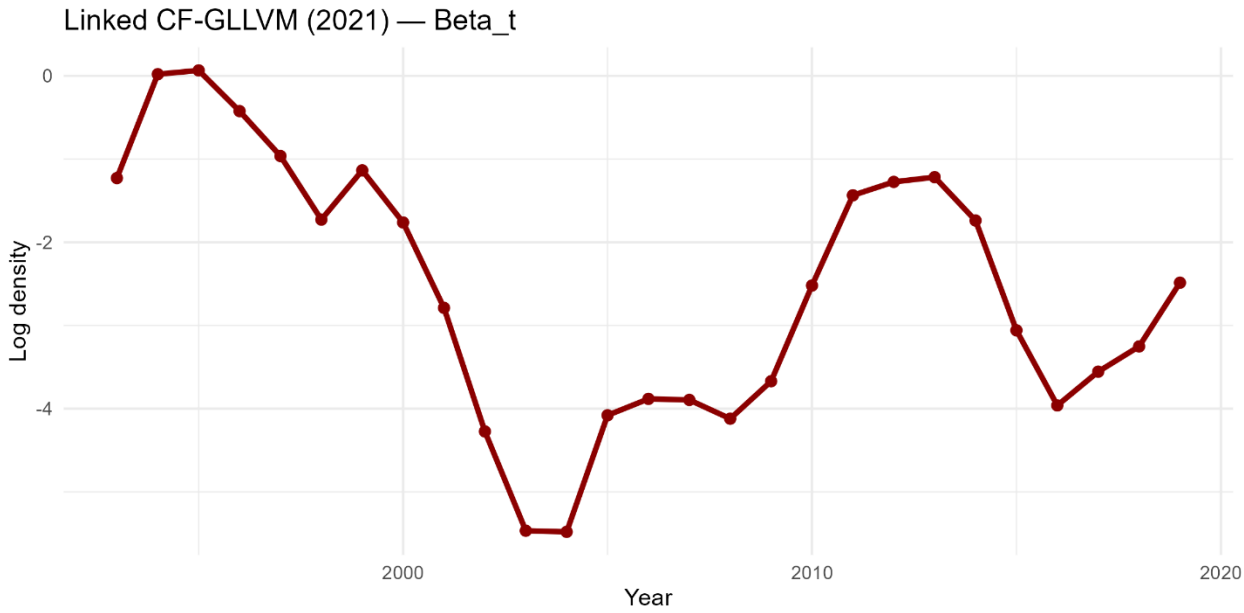


Supplementary Figure S.1.7.1. A) Example of the rotated spatial loadings (U, top) and B) temporal loadings (V, bottom) from the EOF decomposition of recruit density for the 2021 prediction scenario, factor 1. Positive spatial loadings (warm colors) indicate stations where that factor enhances the mean density pattern; negative values (cool colors) indicate the inverse. Temporal loadings show the year-to-year prominence of each spatial pattern, with positive values indicating years where that pattern was more strongly expressed. In this example, nonzero spatial loadings are concentrated in the central-southern EBS shelf, an area of high interannual variability in recruit density. The temporal loadings decline from positive values during early years to negative values during recent years, indicating that this spatial pattern (elevated density in the central-south) was more prominent in the earlier portion of the time series and has weakened in recent years, consistent with observed distributional shifts. The remaining nine factor maps (factors 2-5 for spawners and recruits) follow the same interpretation framework and are available upon request.

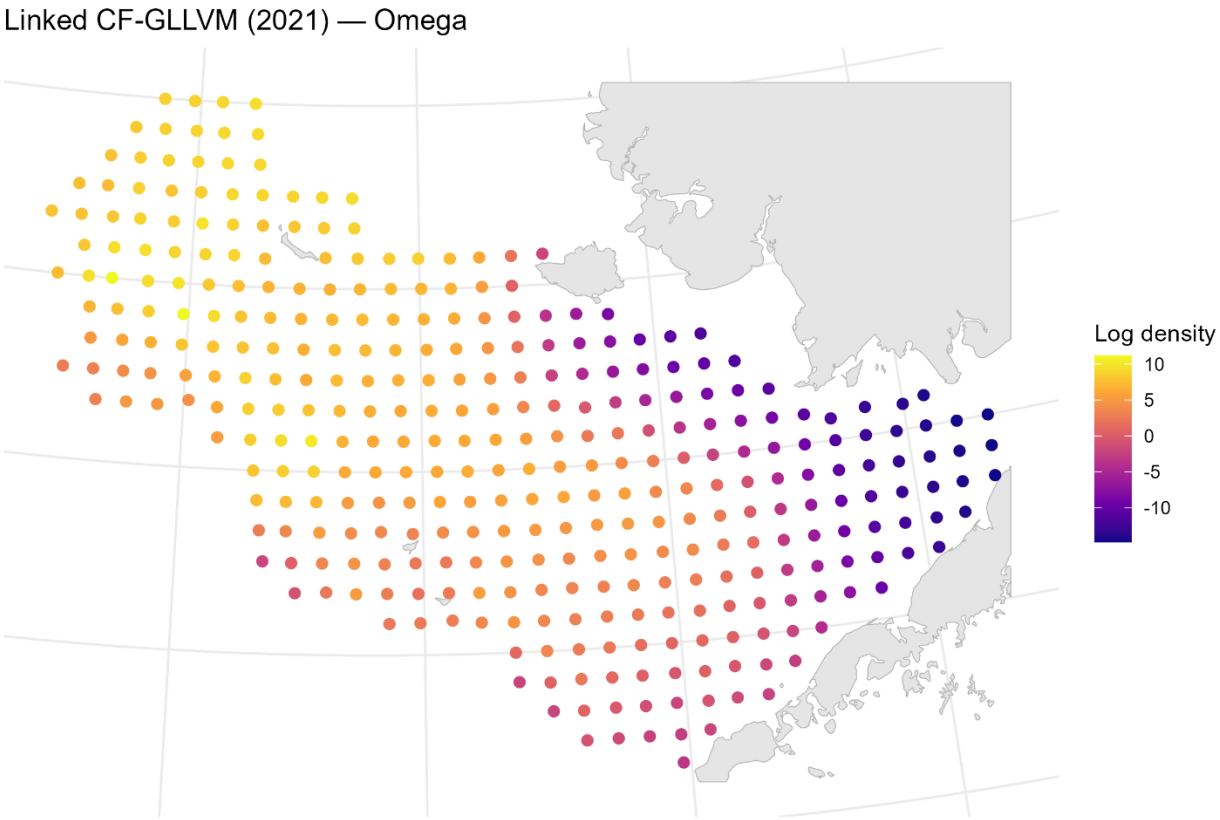
S.1.8 CF-GLLVM Spatial Factor Maps and Connection Matrices

The confirmatory factor analysis CF-GLLVM models produce a large number of estimated spatial fields and parameter matrices. Displaying all factors across all 108 model configurations is not

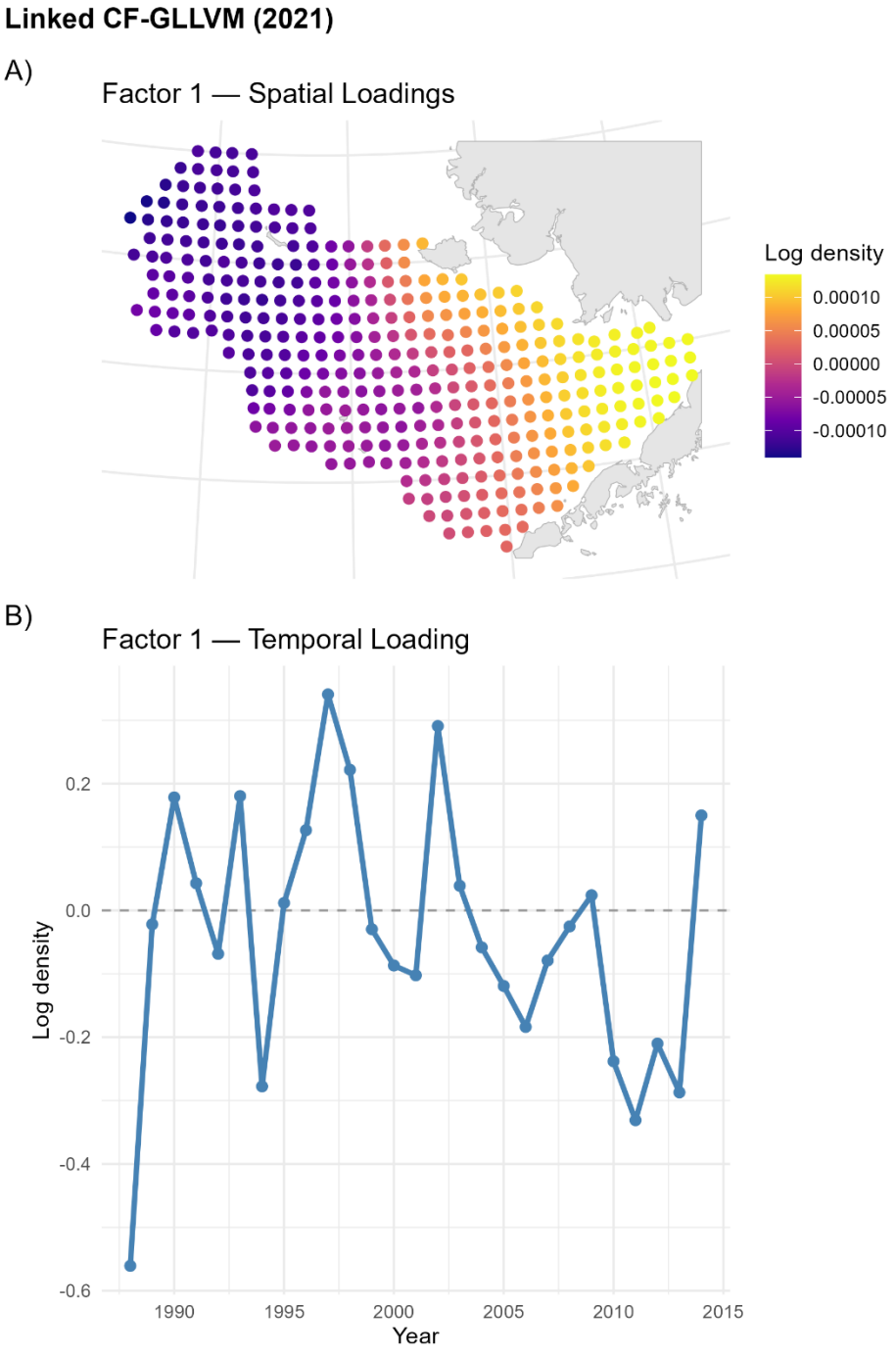
feasible. Instead, we present the full set of estimated parameters for one representative model: the Linked CF-GLLVM with configuration c4 (annual intercepts and spatiotemporal residual) to demonstrate interpretation.



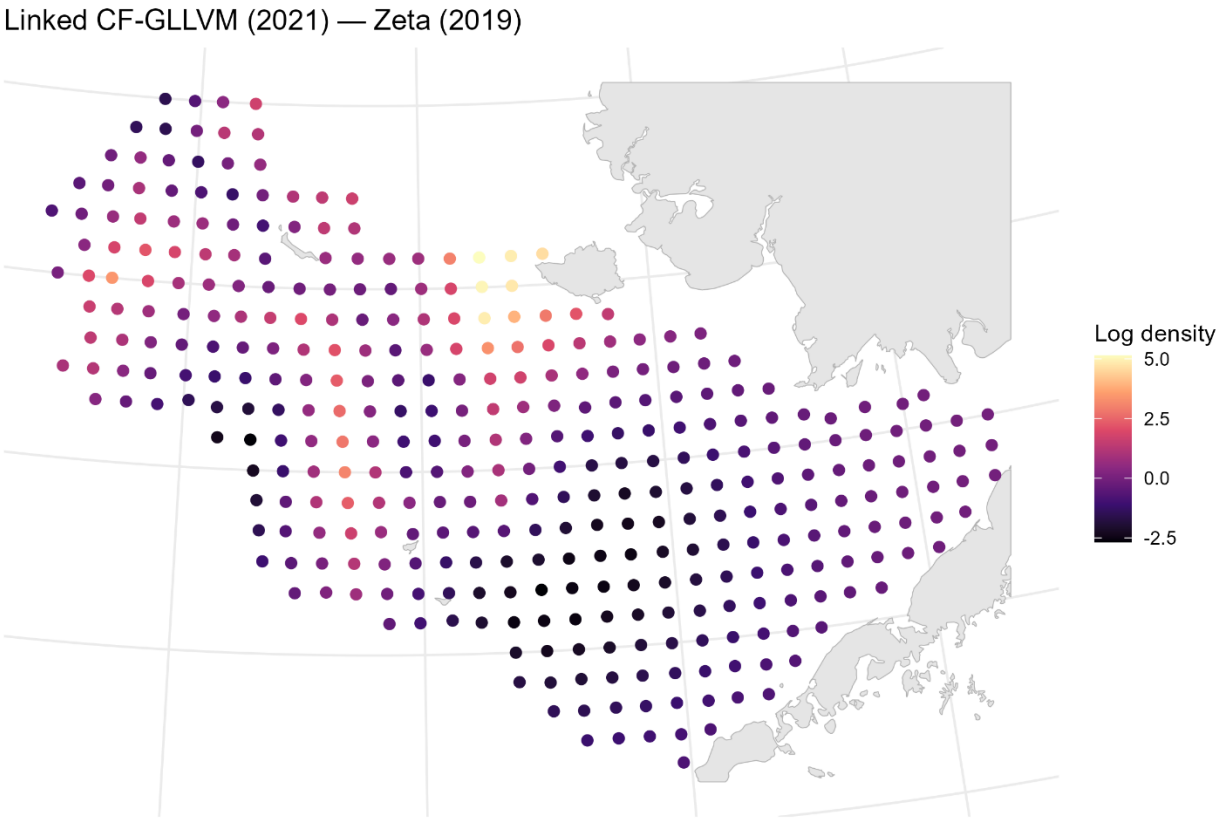
Supplementary Figure S.1.8.1. Estimated annual intercept vector ($\widehat{\boldsymbol{\beta}}_t$) for the Linked CF-GLLVM model, configuration c4, across training years. Each point represents the estimated intercept for one year, capturing interannual variation in overall recruit density not explained by the spatial factor structure.



Supplementary Figure S.1.8.2. Estimated spatial mean field ($\hat{\boldsymbol{\omega}}$) for the Linked CF-GLLVM model, configuration c4. This field represents the time-invariant average spatial pattern of recruit density, analogous to the station means removed during centering in the EOF approach.

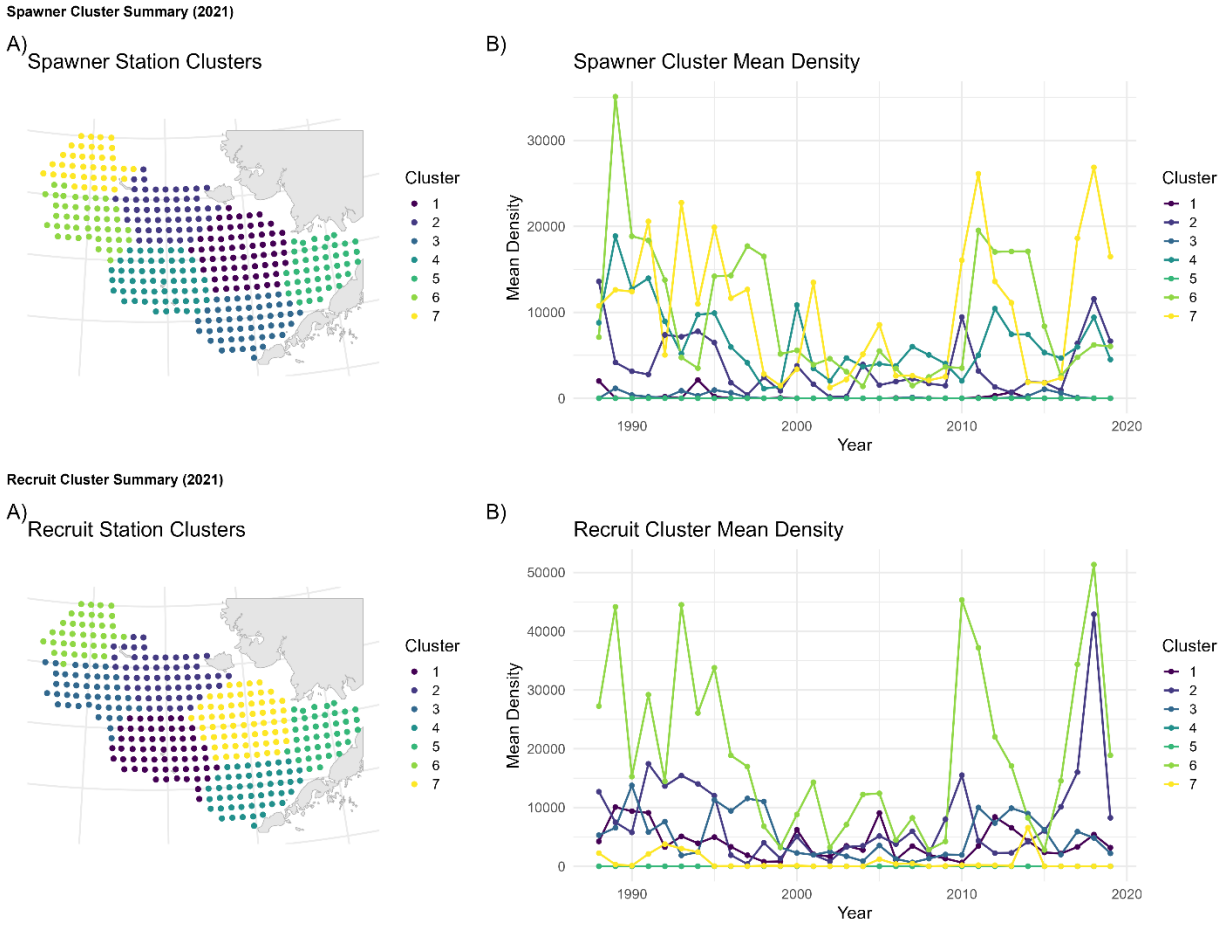


Supplementary Figure S.1.8.3. A) Estimated spatial factor maps ($\hat{\boldsymbol{\varepsilon}}$) and B) corresponding temporal loadings ($\hat{\boldsymbol{\lambda}}$) for the Linked CF-GLLVM model, configuration c4. The spatial map shows the recruit spatial response patterns estimated by the Linked CF-GLLVM model. The temporal loadings are taken from the spawner EFA decomposition and are treated as fixed data in the CFA estimation; they represent the year-to-year strength of each spawner spatial mode.



Supplementary Figure S.1.8.4. Estimated spatiotemporal residual field ($\hat{\xi}$) for the terminal training year (2019) from the Linked CF-GLLVM model, configuration c4. This field captures spatial variation in the final training year that is not explained by the low-rank factor decomposition or the spatial mean.

Supplementary Section S.1.9: Cluster Maps and Cluster-Density Time Series



Supplementary Figure S.1.9.1. Spectral clustering results for the 2021 prediction scenario. Left panels (panel A) show the spatial distribution of the seven clusters for spawners (top) and recruits (bottom), with each station colored by cluster assignment. Right panels (panel B) show the corresponding cluster-level mean density time series used as input to the GAM prediction step, where each line represents the mean density across all stations within a cluster for each survey year. Spawner cluster time series span the training period (1988-2016); recruit cluster time series span 1993-2019. The spatial configurations and time series for the 2022 and 2023 scenarios followed similar patterns and are available upon request.

Chapter 1 References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alglave, B., et al. (2025). Empirical orthogonal functions for ecology. HAL Open Science. https://hal.science/hal-04693871/

Alheit, J., & Bakun, A. (2010). Population synchronies within and between ocean basins: Apparent teleconnections and implications as to physical–biological linkage mechanisms. *Journal of Marine Systems*, *79*(3), 267–285. https://doi.org/10.1016/j.jmarsys.2008.11.029

Anderson, S. C., Ward, E. J., English, P. A., Barnett, L. A. K., & Thorson, J. T. (2025). sdmTMB: An R Package for Fast, Flexible, and User-Friendly Generalized Linear Mixed Effects Models with Spatial and Spatiotemporal Random Fields. Journal of Statistical Software, 115(2), 1–46. https://doi.org/10.18637/jss.v115.i02

Anger, K. (2006). Contributions of larval biology to crustacean research: A review. *Invertebrate Reproduction & Development*, *49*(3), 175–205. https://doi.org/10.1080/07924259.2006.9652207

Cardil, A., et al. (2023). Climate teleconnections modulate global burned area. *Nature Communications*, *14*(1), 427. https://doi.org/10.1038/s41467-023-36052-8

Diaz, H. F., et al. (2001). ENSO variability, teleconnections and climate change. *International Journal of Climatology*, *21*(15), 1845–1862. https://doi.org/10.1002/joc.631

Gay D. M. (1990). "Usage Summary for Selected Optimization Routines." Computing Science Technical Report 153, AT&T Bell Laboratories, Murray Hill, NJ.

Giorgino, T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, 31(7), 1–24. https://doi.org/10.18637/jss.v031.i07Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*, 723–773.

Grüss, A., et al. (2021). Synthesis of interannual variability in spatial demographic processes supports the strong influence of cold-pool extent on Eastern Bering Sea walleye pollock (*Gadus chalcogrammus*). *Progress in Oceanography*, *194*, 102569. https://doi.org/10.1016/j.pocean.2021.102569

Hawkes, C. (2000). Woody plant mortality algorithms: Description, problems and progress. *Ecological Modelling*, *126*(2), 225–248.

Kristensen, K., et al. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*, 1–21.

Kristensen K (2026). *RTMB: 'R' Bindings for 'TMB'*. R package version 1.9, https://github.com/kaskr/rtmb.

Lindgren, F., Rue, H., & Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(4), 423-498. DOI: 10.1111/j.1467-9868.2011.00777.x

Lindgren, F. (2026). *fmesher: Triangle meshes and related geometry tools* (R package version 0.7.0). https://github.com/inlabru-org/fmesher

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons. https://doi.org/10.1002/9781119482260

Lorenz, E. N. (1956). *Empirical orthogonal functions and statistical weather prediction*. https://cir.nii.ac.jp/crid/1370004237455296143

Markowitz, E. H., et al. (2024). *Results of the 2022 eastern and northern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate fauna*. NOAA. https://repository.library.noaa.gov/view/noaa/50154

Monahan, A. H., Fyfe, J. C., Ambaum, M. H. P., Stephenson, D. B., & North, G. R. (2009). Empirical orthogonal functions: The medium is the message. *Journal of Climate*, *22*, 6501–6514. https://doi.org/10.1175/2009JCLI3062.1

Mondal, S., et al. (2025). Teleconnection impacts of climatic variability on tuna and billfish fisheries of the South Atlantic and Indian Ocean. *Fish and Fisheries*, *26*(2), 240–256. https://doi.org/10.1111/faf.12876

Rodionov, S. N. (1995). Atmospheric teleconnections and coherent fluctuations in recruitment to North Atlantic cod (*Gadus morhua*) stocks. In R. J. Beamish (Ed.), *Climate change and northern fish populations* (Vol. 121, pp. 45–55). Department of Fisheries and Oceans Canada.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sainte-Marie, B., Ouellet, P., & Dionne, M. (2021). Direction, ontogeny, and mechanism of the temperature-size rule operating in a large marine crab, *Chionoecetes opilio*. *ICES Journal of Marine Science*, *78*(2), 557–574.

Sainte-Marie, B., Raymond, S., & Brêthes, J.-C. (1995). Growth and maturation of the benthic stages of male snow crab, *Chionoecetes opilio* (Brachyura: Majidae). *Canadian Journal of Fisheries and Aquatic Sciences*, *52*(5), 903–924. https://doi.org/10.1139/f95-091

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055

Sandvik, H., et al. (2016). Modelled drift patterns of fish larvae link coastal morphology to seabird colony distribution. *Nature Communications*, *7*, 11599.

Schwing, F. B., et al. (2010). Climate change, teleconnection patterns, and regional processes forcing marine populations in the Pacific. *Journal of Marine Systems*, *79*(3–4), 245–257. https://doi.org/10.1016/j.jmarsys.2008.11.027

Serrão, E. A. O., et al. (2025). The effects of teleconnections on water and carbon fluxes in the two South America's largest biomes. *Scientific Reports*, *15*(1), 1395.

Shchepetkin, A. F., & McWilliams, J. C. (2005). The Regional Oceanic Modeling System (ROMS): A split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Modelling*, *9*(4), 347–404.

Suhr, D. D. (2006). Exploratory or confirmatory factor analysis? *Statistics and Data Analysis: SUGI 31 Proceedings*, 26–29.

Szuwalski, C. (2022). *An assessment for eastern Bering Sea snow crab*. In Stock assessment and fishery evaluation report for the king and tanner crab fisheries of the Bering Sea and Aleutian Islands regions. North Pacific Fishery Management Council.

Szuwalski, C. S., et al. (2023). The collapse of eastern Bering Sea snow crab. *Science*, *382*(6668), 306–310.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393), 82-86.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276.

Thorson, J., & Kristensen, K. (2024). *Spatio-temporal models for ecologists*. CRC Press.

Thorson, J. T., et al. (2020). Defining indices of ecosystem variability using biological samples of fish communities: A generalization of empirical orthogonal functions. *Progress in Oceanography*, *181*, 102244.

Thorson, J. T., et al. (2021). Forecasting community reassembly using climate-linked spatio-temporal ecosystem models. *Ecography*, *44*(4), 612–625.

Thorson, J. T., et al. (2023). Spatially varying coefficients can improve parsimony and descriptive power for species distribution models. *Ecography*, *2023*(5), e06510.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proceedings of the Indian Statistical Institute golden jubilee international conference* (Vol. 579, pp. 579–604).

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4), 395-416

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press.

Yeh, S.-W., et al. (2018). ENSO atmospheric teleconnections and their response to greenhouse gas forcing. *Reviews of Geophysics*, *56*(1), 185–206.

Zacher, L. S., et al. (2024). The 2023 eastern and northern Bering Sea continental shelf trawl surveys: Results for commercial crab species. https://doi.org/10.25923/c5j1-4r46

Chapter 2: Applying Vision Transformer AI pipeline to eastern Bering Sea Snow Crab (*Chionoecetes opilio*) Spawner-Recruit Relationships

2.Abstract

Snow crab (*Chionoecetes opilio*) support one of the most economically important fisheries in the eastern Bering Sea. However, predicting the spatial distribution of recruitment remains a central challenge for stock assessment and harvest management. Conventional approaches collapse spatially explicit survey data to scalar abundance indices before conducting stock assessments, discarding potentially informative spatial structure. I present a Vision Transformer adapted for spatiotemporal ecological prediction, and evaluate it as a case study in applying deep learning to spatial recruitment forecasting under severe data constraints. The model takes multi-input spatial grids of historical spawner density, recruit density, and bottom temperature and predicts the 50 x 50 km gridded distribution of recruit density across the eastern Bering Sea shelf. Predictions and evaluation targets are both derived from SPDE-interpolated spatial fields rather than raw station observations, a design choice discussed in detail in the methods. The model was evaluated across three forecasting horizons: same-year now-casting, one-year-ahead forecasting without current-year survey data, and five-year-ahead prediction using temporally lagged inputs reflecting the approximate ontogenetic lag between spawning and recruitment. Model performance was also evaluated on three synthetic benchmark datasets with known spawner-recruit relationships spanning a range of lag structures and signal strengths. The model achieved Spearman's correlations between predicted and observed recruitment fields of 0.83 to 0.87 on held-out test data from 2021-2023, across three forecasting horizons. The one-year-ahead configuration performed nearly as well as the same-year now-cast despite withholding all current-year survey data, suggesting the model learned genuine multi-year temporal dynamics rather than propagating current-year spatial patterns. A temperature-only configuration, relying only on historical bottom temperature without any direct observation of crab density, ranked third overall, indicating that thermal habitat structure alone carries substantial spatial information about recruitment intensity. Attribution analysis via Integrated Gradients revealed that the model's learned structure was largely consistent with known features of snow crab biology: same-year spawner distributions dominated predictions when available, and historical temperature fields aligned with the cold pool boundary were consistently influential. Under the five-year lag configuration, temperature at a ten-year lag emerged as the dominant predictor, a somewhat unexpected result that may reflect a lagged thermal memory in the system, though whether this represents a true biological signal or an artifact of the lag structure remains uncertain. All configurations exhibited substantial overfitting on actual survey data relative to synthetic benchmarks, reflecting the fundamental constraint that there are only 24 independent spatial configurations available for training. Reduced model architectures and validation-based checkpoint selection were essential for extracting generalizable predictions. These results demonstrate that Vision Transformers can recover biologically interpretable spatiotemporal relationships using small ecological datasets, and provide a practical template for applying deep learning to spatiotemporal prediction in data-limited fisheries and ecological monitoring contexts.

2.1 Introduction

Snow crab (*Chionoecetes opilio*) support one of the most economically important fisheries in the eastern Bering Sea (EBS). However, predicting recruitment remains a central challenge for its stock assessment and harvest management. Recruitment is determined by complex biological and

physical processes that link spawning events to juvenile survival over multi-year timescales, and it varies substantially across both space and time.

A linkage across time and space is called a teleconnection, and while common, can be difficult to quantify spatiotemporally. Snow crab spawning and recruitment can be thought of as a teleconnection, with the distribution of spawning crab leading the spatial recruitment distribution. Accurately predicting when and where recruits will appear as a result of this teleconnection is critical for managers of this fishery. However, the spawner-recruit relationship for snow crab is difficult to model (Szuwalski & Punt 2013), and the spatial structure of that relationship is even more poorly understood. Conventional stock-recruitments collapse spatial fields to scalar abundance indices before fitting (Ricker 1954; Cadrin 2020), discarding potentially informative spatial structure and making it impossible to identify the geographic scale at which spawner-recruit dynamics operate.

Geostatistical and species distribution modelling approaches can represent spatial fields explicitly but typically require the analyst to pre-specify the form of spatial dependence rather than learning it from data (such as through a covariance function, a neighborhood structure, or a set of environmental covariates) (Cressie & Wikle 2011; Lindgren et al. 2011; Elith & Leathwick 2009). This requirement can be constraining when the relevant spatial relationships are unknown, as is the case for EBS snow crab. A method that can identify which spatial locations are predictive of recruitment elsewhere without being pre-specified would lead to a meaningful improvement in forecasting and the ability to characterize recruitment.

Machine learning and deep learning methods have seen rapidly growing application in ecology over the past decade, particularly for tasks involving image classification such as species identification from camera trap photographs, acoustic event detection, and remote sensing land cover mapping (e.g., Christin et al. 2019, Norouzzadeh 2018). These applications share a common structure, where the input is an image or spectrogram, and the output is a category label, a task for which convolutional neural networks (CNNs) have proven highly effective (e.g., Norouzzadeh 2018). Spatiotemporal prediction of continuous ecological processes such as spatial recruitment prediction is a fundamentally different problem. The input is a spatial field that evolves through time, the output is another spatial field, and the relationship between them may involve non-local dependencies that CNNs are poorly suited to capture given their model architecture. CNNs operate using small fixed-size filters (e.g., 3×3 or 5×5 kernels) that only "see" a small neighborhood of pixels at a time, termed the “locality problem” (Luo et al. 2016). A CNN must stack many layers so that information can propagate through successive local operations to capture relationships between distant parts of a spatial field, like teleconnections, and even then, the effective receptive field may not span the full domain, or the long-range signal is diluted through many intermediate processing steps (Luo et al. 2016). Recurrent architectures such as Long Short-Term Memory models (LSTMs; Hochreiter & Schmidhuber 1997) can represent temporal dynamics, however, they process spatial information through flattening or spatial pooling, losing the two-dimensional structure of the fields (Shi et al. 2015). Applications of deep learning to the class of ecological prediction problem represented by prediction of recruitment remain sparse (Kapur et al. 2026).

The transformer architecture, originally developed for natural language processing (Vaswani et al., 2017) and subsequently adapted for image analysis as the Vision Transformer or ViT (Dosovitskiy et al., 2021), offers a potential solution to the problem of modeling long range spatial relationships. Rather than scanning an image using a fixed-size filter as is the case for CNNs, a ViT divides the image into spatial patches and then uses a self-attention mechanism to learn which patches are relevant to predicting the representation of any other patch directly from the data. This

addresses the locality constraint of CNNs and allows every location in an input to directly influence every other location simultaneously. A ViT can consequently discover long-range spatial dependencies without pre-specification, making it well-suited to teleconnection problems where the relevant spatial scales are unknown. ViT has achieved competitive or state-of-the-art performance across a wide range of image analysis tasks and has recently been extended to spatiotemporal prediction in climate science and remote sensing (Gao et al. 2022; Nguyen et al. 2023), but to my knowledge has not previously been applied to spatial prediction in fisheries ecology or to spatiotemporal ecological prediction more broadly from ground-based survey data.

I present a ViT adapted for spatial recruitment prediction for EBS snow crab, and evaluate it as a case study in applying modern deep learning methods to spatiotemporal ecological prediction under severe data constraints. The specific contributions are threefold. First, I describe and motivate the architectural and training adaptations required to apply a ViT to small ecological datasets, including channel-wise patch embedding that respects the ecological meaning of different input types, a temporal memory bank encoding the multi-year historical context, and a masking mechanism that handles missing observations. Second, I evaluate the model on three synthetic benchmark datasets with known spawner-recruit correlations spanning a range of spatial correlation structures and temporal lags, providing a controlled assessment of what the model can and cannot recover. Third, I apply the model to actual EBS survey data for snow crab under three forecasting scenarios: a) same-year prediction, b) one-year-ahead forecasting without current-year spawner data, and c) five-year-ahead prediction in which spawner and temperature inputs are aligned with the estimated year of origin of the observed recruit cohort, and use Integrated Gradients attribution (Sundararajan et al. 2017) to examine the spatial and temporal structure of what the model has learned. I also conduct a sensitivity analysis across model sizes, loss functions, and input channel configurations to identify which design choices most affect performance under severe data scarcity.

This paper is as much a methodological investigation as an applied forecasting study. The EBS snow crab time series spans approximately 35 survey years, of which 24 were used for training after temporal splitting and exclusion of the 2020 pandemic year. Despite bootstrap resampling to augment sample size, the effective number of statistically independent training observations remains close to the number of survey years, well below the sample sizes typically required for deep learning. The goal of this work is therefore to assess the performance of the ViT method on spatial pattern and abundance predictions, and use the findings as a practical reference for ecologists considering deep learning approaches in similar data-scarce settings.

2.2 Methods

A spatiotemporal deep learning pipeline (Figure 2.1) was used to predict the spatial distribution of snow crab recruitment across the EBS using historical observations of spawner density, recruit density, and bottom temperature. The central prediction goal was to map juvenile (recruit) density for year t from multi-channel spatial inputs that encode the recent history of spawner abundance, recruit abundance, and environmental conditions (an annual index of bottom temperature obtained from the `coldpool` R package, Rohan et al. 2022) at every location. The pipeline is evaluated using actual survey data smoothed using the Stochastic Partial Differential Equation (SPDE; Lindgren et al. 2011) approach, which are then augmented using spatial bootstrap resampling to address data scarcity (see below), and separately on simulated data to quantify model performance when the spatiotemporal scenario is known. The core model is a ViT (Dosovitskiy et al., 2021) adapted for ecological spatiotemporal prediction. A ViT architecture was chosen because its attention mechanism allows it to learn non-local dependencies between spatially distant regions, making it

well equipped to investigate potential teleconnections in the snow crab spawner-recruit system and make predictions of future recruitment in space and time. Multiple forecasting scenarios were explored to assess performance across conditions by varying the temporal information available to the model at prediction time, ranging from same-year now-casting to one-year-ahead forecasting.

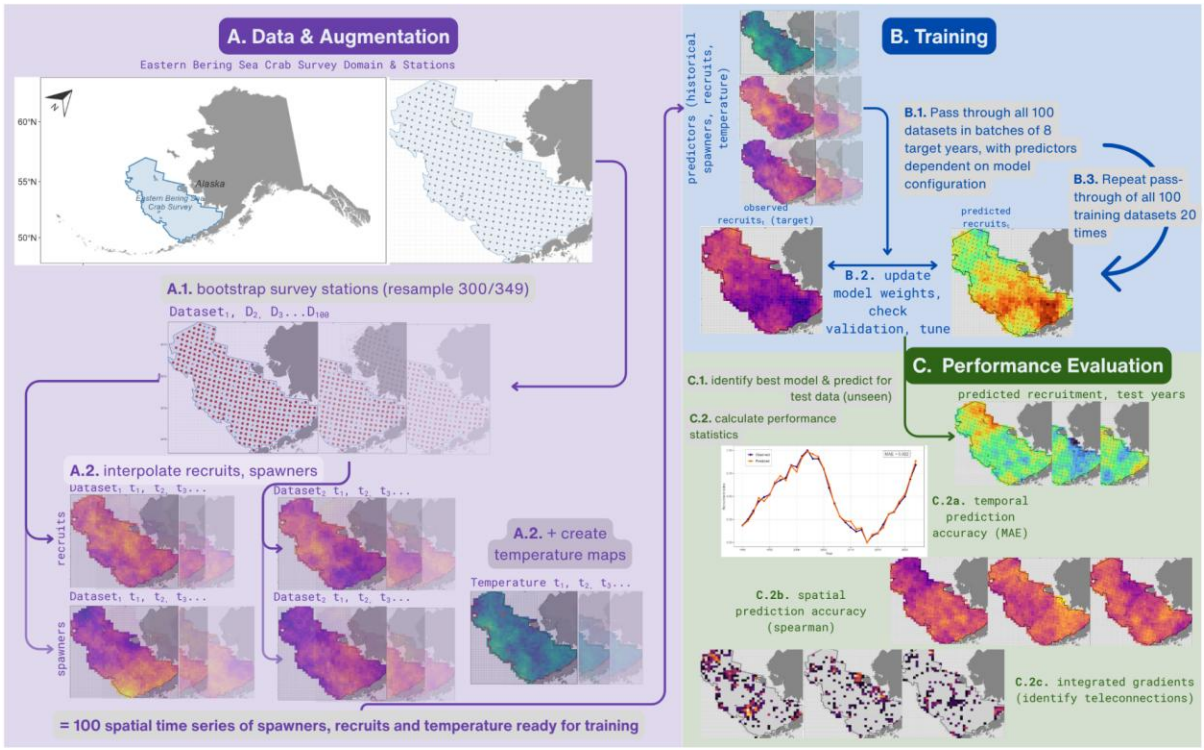


Figure 2.1. Overview of the ViT modeling pipeline. (A) Data preparation: survey station locations across the EBS shelf are used as inputs to the spatial interpolation procedure. (A.1) For each of 100 bootstrap replicates, 300 of the 349 available stations are drawn without replacement. (A.2) An SPDE model is fit to the subsampled stations for each survey year, projecting interpolated spawner and recruit densities and bottom temperature onto a 50 × 50 grid. (B) Model training: the full training dataset is passed through the ViT model in batches of eight samples. (B.1) Each batch is forward-passed through the ViT to generate predictions. (B.2) Loss is computed between predictions and observations, and model weights are updated via backpropagation. (B.3) This process repeats until all training samples have been seen once, constituting one "epoch"; training continues for 20 epochs, with validation loss monitored after each epoch to identify the best model checkpoint. (C) Model evaluation: the best checkpoint is applied to the held-out test set. (C.1) Predictions are generated for all test samples. (C.2) Performance is assessed using mean absolute error (C.2a), Spearman's correlation between predicted and observed spatial fields (C.2b), and Integrated Gradients attribution to identify which input channels and spatial regions most strongly influenced predictions (C.2c).

2.2.1 Study Region and Survey Data

Snow crab density data were obtained from the NOAA EBS bottom trawl survey, which has been conducted annually since the mid-1970s at approximately 349 fixed stations arranged on a regular grid spanning the EBS shelf (Markowitz et al., 2024; Figure 2.1A). Survey years with incomplete or absent coverage were excluded; specifically, the 2020 survey year was omitted due to the suspension of field operations during the COVID-19 pandemic. The survey dataset covers 1988-2023, and contains individual-level catch information including sex, carapace width, and clutch data, as well as station-level data on location, bottom temperature, depth, and other abiotic factors.

One 30-minute bottom trawl is conducted at each station during daylight hours on a consistent 20 × 20 nm systematic grid.

The dataset was divided into two subsets: spawners and recruits. Spawners were defined as female crabs carrying eggs (clutch > 0), using egg presence as a proxy for reproductive maturity in the absence of a direct maturity indicator. Recruits were defined as crabs with carapace width between 45-55 mm, assumed to be approximately five years old (Sainte-Marie et al., 1995). Station-level density was calculated by dividing the total estimated catch in numbers by the area swept during the trawl.

Bottom temperature data were obtained using the coldpool R package (Rohan et al., 2022), which provides interpolated rasters of EBS bottom temperature at 5 km resolution from 1982 to 2025, excluding 2020. Bottom temperatures were recorded at survey stations during the bottom trawl survey. Temperature surfaces were produced using ordinary kriging of bottom temperatures recorded at standard index stations. These rasters were downsampled on to the 50 × 50 prediction grid by extracting the mean bottom temperature within each grid cell (Figure 2.1, step A.2). Bottom temperature was used instead of other abiotic factors due to the data being readily available, and because there is thought to be a relationship between temperature and snow crab population dynamics (Fedewa et al. 2020).

Station-level density observations were projected on to a regular spatial grid of 50 × 50 cells, each covering approximately 23 × 23 km (cell size = 25 km; grid extent: 50 columns × 35 rows over the EBS region, with remaining rows padded to produce a square grid). The 50 × 50 grid dimension was selected to balance spatial resolution against computational tractability for the ViT architecture, which processes the grid as a sequence of patches. Spatial interpolation was carried out using an SPDE approach implemented in R-INLA (Lindgren et al., 2011), which approximates a Gaussian Markov Random Field (GMRF) and allows geostatistical interpolation of irregularly spaced point observations onto a regular lattice (Supplementary Section S.2.1). Each year was modeled independently with no temporal correlation structure; temporal dependencies were handled by the ViT memory bank (Section 2.2.6). Station-level density observations were projected on to the prediction grid for each survey year and bootstrap replicate independently (Figure 2.2, steps 2-4), and the resulting gridded fields were assembled into a time series of 50 × 50 density maps for both spawners and recruits (Figure 2.2, step 7). A binary spatial mask identifying cells within the EBS survey footprint was constructed to exclude out-of-region cells from all model inputs and loss calculations.

2.2.2 Bootstrap Data Augmentation

A fundamental challenge in applying deep learning to fisheries survey data is data scarcity: even a long time series of annual surveys yields only on the order of 30-40 independent spatial snapshots, far fewer than is typically required to train a neural network. To address this, the dataset was augmented using a spatial bootstrap resampling approach. For each of 100 bootstrap replicates, a random subsample of 300 of the 349 survey stations was drawn without replacement (Figure 2.2, step 2), an independent SPDE model was fit to each year's subsampled data (Figure 2.2, step 3), and the interpolated densities were projected on to the 50 × 50 grid (Figure 2.2, step 4). This procedure introduces realistic variability in spatial interpolation arising from measurement uncertainty and sampling variability, producing 100 plausible realizations of the spatial density field for spawners and recruits for each survey year. The procedure was applied separately to spawners and recruits (Figure 2.2, step 6), yielding 100 gridded time series for each variable (Figure 2.2, step 7). The full resampling and interpolation workflow is illustrated in Figure 2.2.

This augmentation increases the number of training samples from roughly 24 to 2,400, but the 100 bootstrap replicates for any given year are highly correlated (mean pairwise Pearson $r = 0.987$ for spawners, 0.991 for recruits, computed on log-transformed densities over valid cells), meaning the effective number of statistically independent training patterns remains approximately equal to the number of survey years. The implications of this constraint are discussed in Section 2.4.1.

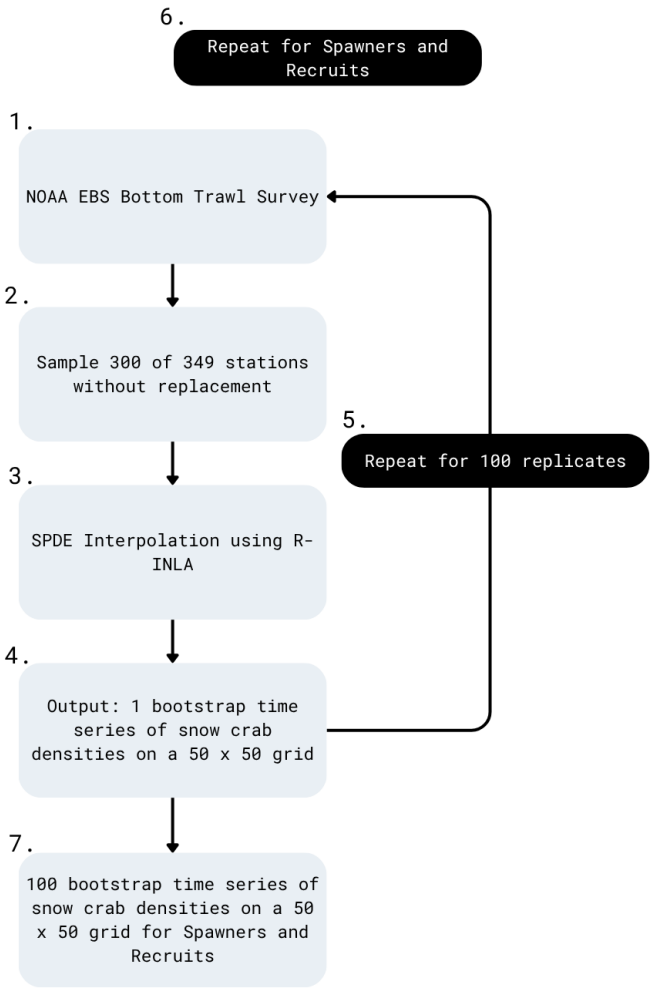


Figure 2.2. Bootstrap resampling and spatial interpolation procedure. (1) The NOAA EBS bottom trawl survey provides annual density observations at approximately 349 fixed stations. (2) For each bootstrap replicate, 300 stations are drawn without replacement from the full set. (3) An SPDE model implemented in R-INLA is fit to the subsampled stations for each survey year, interpolating station-level densities onto a regular 50 × 50 grid. (4) The output is a gridded time series of snow crab density across the survey domain. Steps 2 through 4 are repeated for 100 independent bootstrap replicates (5) and applied separately to the spawner and recruit subsets (6), yielding (7) 100 bootstrap realizations of the gridded spawner and recruit density time series for use in model training and evaluation.

2.2.3 Synthetic benchmark data

Three sets of simulated datasets were generated from a GMRF model to validate the ViT pipeline under "control" scenarios where the actual spawner-recruit relationship was known. This involved specifying covariance structures to emulate spatial patterns in spawner and recruit density that are correlated across space and time. Spatially correlated fields were sampled from a GMRF defined by a sparse precision matrix (Figure 2.3, steps 1-2). Temporal autocorrelation was introduced using

an AR(1) process (Figure 2.3, step 3), and the spawner and recruit fields were blended using a specified correlation coefficient and lag to produce the desired spawner-recruit relationship (Figure 2.3, step 3.c). Zero-inflation was applied by subtracting the 60th percentile of the transformed density field and setting negative values to zero, producing approximately 60% zero-valued cells, broadly consistent with the proportion of zero-density observations in the survey data (59.0%). Fields were then rescaled to match the observed mean density and clipped at the observed maximum (Figure 2.3, step 4). The full procedure was repeated for 100 bootstrap replicates (Figure 2.3, step 5) and across the three difficulty scenarios (Figure 2.3, step 6). Full GMRF simulation details are provided in Supplemental Section S.2.2. Temperature data were not included in the synthetic datasets.

The three datasets differed in the magnitude and complexity of the spawner-recruit relationship. Spawner-recruit correlation was set to 0.9 with no temporal autocorrelation and no lag in the "easy" scenario. Correlation remained 0.9 but temporal autocorrelation was set to 0.7 and a three-year lag was introduced in the "medium" scenario. Correlation was reduced to 0.8, temporal autocorrelation retained at 0.7, and the lag extended to five years in the "hard" scenario. The full simulation procedure is illustrated in Figure 2.3 and key parameters for each scenario are summarized in Table S.2.1.

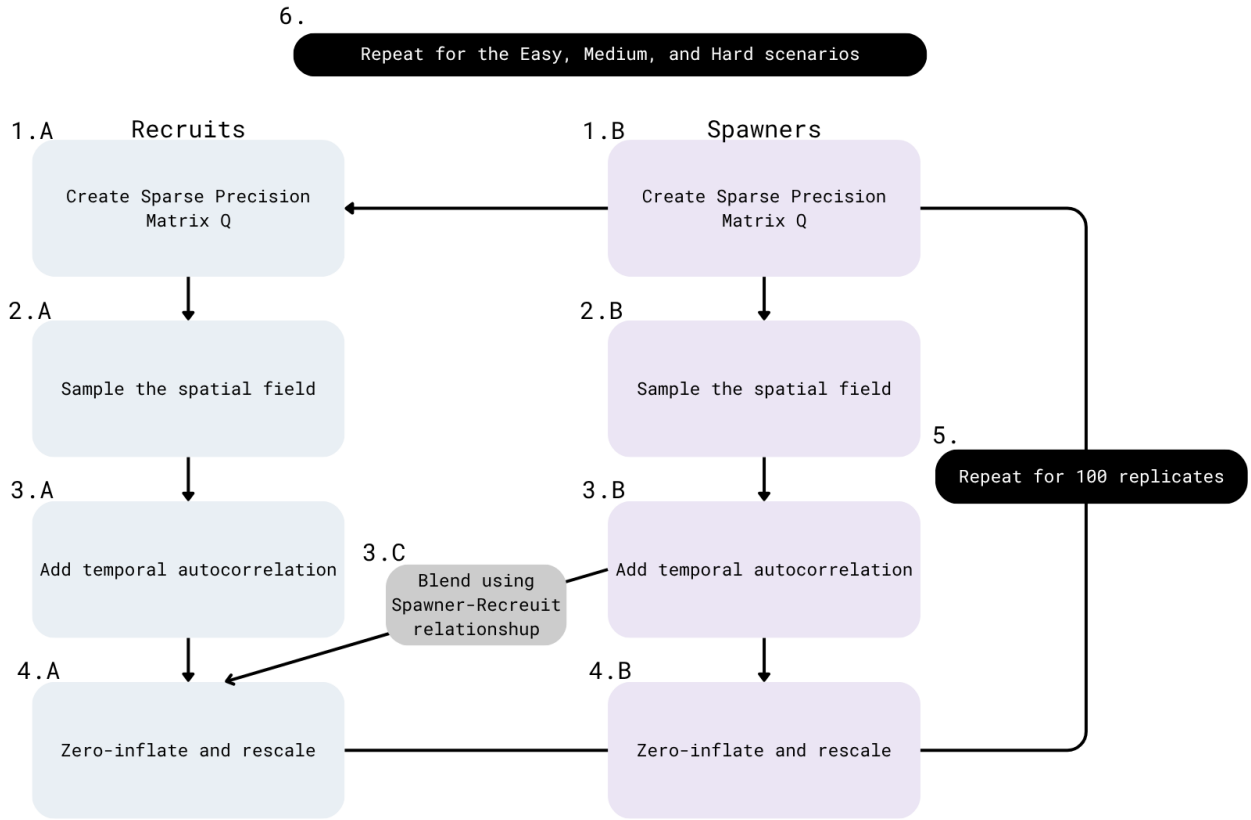


Figure 2.3. Synthetic data generation procedure. (1) A sparse spatial precision matrix **Q** is constructed on the 50 × 50 grid using a nearest-neighbor structure that encodes spatial correlation. (2) Spatially correlated fields are sampled from the resulting GMRF. (3) Temporal autocorrelation is introduced by evolving fields forward using an AR(1) process. (3c) Spawner and recruit fields are blended using a specified correlation coefficient and temporal lag to produce the desired spawner-recruit relationship. (4) Zero-inflation is applied by subtracting a density threshold and setting negative values to zero, followed by rescaling to match the observed mean density and maximum values. (5) Steps 1 through 4 are repeated for 100 bootstrap replicates. (6) The full procedure is repeated at three difficulty scenarios

(easy, medium, hard) by varying the spawner-recruit correlation, temporal autocorrelation, and lag structure (Table S.2.1).

2.2.4 Data preprocessing and temporal splits

All density values were log-transformed prior to analysis using $\log(1 + x)$, which compresses the heavy right tail of crab density distributions, reduces the influence of extreme values on gradient updates, and converts the multiplicative structure of count-like data to an approximately additive structure. While log(1 + x) transformation has been criticized as statistically inferior to likelihood-based alternatives such as negative binomial GLMs or Tweedie-distributed models (O'Hara & Kotze, 2010; Shono, 2008), implementing equivalent count-data likelihoods within a neural network training objective is non-standard and was not pursued here. The spatial mask was applied prior to transformation, setting density values for land and out-of-region cells to zero. Log transformation was applied rather than working in the original count scale for three reasons. First, crab density distributions are strongly right-skewed with a high proportion of zeros and occasional extreme values; MSE loss on the untransformed scale would be dominated by a small number of high-density cells, causing the model to prioritize accurate prediction of rare abundance peaks over the spatial pattern as a whole. Second, log-transformed counts have an approximately additive error structure that is better matched to the Gaussian assumptions underlying the MSE loss function. Third, the SPDE interpolation itself was performed on log-transformed densities, so training on the log scale preserves consistency with the data generating process used to produce the inputs. It is acknowledged that stock assessments typically operate in abundance units on the original scale; back-transformed predictions and the lognormal bias correction applied at evaluation time (Section 2.2.7.1) was intended to recover density estimates in the original units for reporting purposes. Temperature values were not log-transformed because they can take negative values.

Evaluating predictions against SPDE-interpolated fields rather than raw station observations was a deliberate design choice motivated by the bootstrap data augmentation procedure. Each bootstrap replicate subsamples 300 of the 349 available stations and fits an independent SPDE model, yielding 100 plausible realizations of the smooth latent density surface for each survey year. Grounding training and evaluation in these interpolated fields maintains internal consistency across the pipeline: the model learns to predict the same type of spatial object it receives as input. An alternative design of fitting the SPDE, passing gridded predictions through the ViT, and then evaluating predictions back at the original station locations against raw trawl observations would introduce an additional interpolation step and would require a decision about which station subset to evaluate against, since each bootstrap replicate withholds a different 49 stations. This would add considerable complexity without clear benefit given the training setup. The smoothed-field design does carry the consequence that both the model and the climatological baseline compete against a spatially regularized truth, the implications of which are discussed in Section 2.4.5. Evaluating ViT predictions directly against station-level observations remains a natural direction for future work, particularly if the bootstrap augmentation step were replaced or supplemented by a design that preserved a held-out set of raw observations.

Data were split into training, validation, and test sets by temporal ordering. For the actual EBS data, the 35 survey years excluding 2020 were partitioned into 24 training years (1988-2011), 8 validation years (2012-2019), and 3 test years (2021-2023), preserving chronological order to prevent data leakage. The same split was applied consistently across all 100 bootstrap replicates. The 2020 survey gap was handled through a per-sample validity flag, taking the value 0 for observations corresponding to 2020 and 1 otherwise. This flag was propagated into the loss

calculation (Section 2.2.7.1), ensuring that 2020 predictions contribute zero weight to training and validation without requiring truncation of the time series.

2.2.5 ViT architecture

The ViT was implemented in PyTorch (Paszke et al., 2019). It takes multiple spatial grids of spawners, recruits, and temperature as input and produces a 50 × 50 grid of predicted recruit density. Each of the up to 17 spatial grids is referred to as a channel (Table 2.1). The architecture is illustrated in Figure 2.4.

For each input channel grid, the model divides the 50 × 50 grid into 100 non-overlapping 5 × 5 patches (Figure S.2.1, step 2). Each patch is projected to a compact embedding vector using a learned linear transformation, layer normalization, and dropout (Figure S.2.1, step 3). Each channel has its own independent projection matrix, since the information each channel encodes differs (e.g., spawner density versus temperature). Per-channel embeddings are concatenated (Figure S.2.1, step 4) and projected to the model's working embedding dimension of 128 (Figure S.2.1, steps 5-6). This step is analogous to aggregating point observations into spatial units before analysis: it makes the problem computationally feasible while preserving local spatial structure. Full details are provided in Supplemental Section S.2.3.1.

A learned spatial embedding is added to each patch representation to provide explicit positional context (Figure 2.4, step 3.b), and a fixed sinusoidal temporal encoding derived from the prediction year index is added to supply temporal context (Figure 2.4, step 3.c). Full details are provided in Supplemental Section S.2.3.2.

The defining feature of this architecture is its self-attention mechanism, which acts as a data-adaptive spatial weighting scheme allowing every spatial patch to attend to every other patch simultaneously. Unlike traditional spatial models requiring pre-specified neighborhood structures or covariance functions, the transformer learns these weights directly from data, allowing it to discover non-local dependencies and long-range teleconnections without prior specification. The 100 patch embeddings are passed through a series of transformer layers (Figure 2.4, step 4), each applying multi-head self-attention (Figure S.2.2) followed by a feed-forward network (Figure S.2.3) with residual connections and layer normalization at each step (Figure S.2.4). The self-attention mechanism computes query (**Q**), key (**K**), and value (**V**) representations for each patch (Figure S.2.2, steps 2-3) and uses them to calculate an attention weight matrix encoding the relative influence of every patch on every other patch (Figure S.2.2, steps 4-7). Full details are provided in Supplemental Sections S.2.3.3 and S.2.3.4.

The patch-level representations from the final transformer layer are reshaped from a sequence of 100 vectors back to a 10 × 10 spatial grid (Figure S.2.5, step 2) and progressively upsampled through three convolutional stages to recover the full 50 × 50 spatial resolution (Figure S.2.5, steps 3-5). A final convolutional layer reduces the output to a single channel (Figure S.2.5, step 6), softplus ($ln(1 + e^x)$) activation enforces non-negativity (Figure S.2.5, step 7), and the spatial mask is applied to zero out land and out-of-region cells (Figure S.2.5, step 8), yielding the predicted recruit density grid (Figure S.2.5, step 9). Full details are provided in Supplemental Section S.2.3.5.

During training, the model processes samples in batches of 8 (meaning 8 data samples are given to the model at a time), which stabilizes optimization by averaging prediction errors across samples before updating weights. Training batches are shuffled randomly. Validation and test batches are processed in chronological order. One complete pass through all training samples constitutes an "epoch"; the model was trained for 20 epochs with validation loss monitored after each epoch (Section 2.2.8; Figure 2.1, steps B.1-B.3).

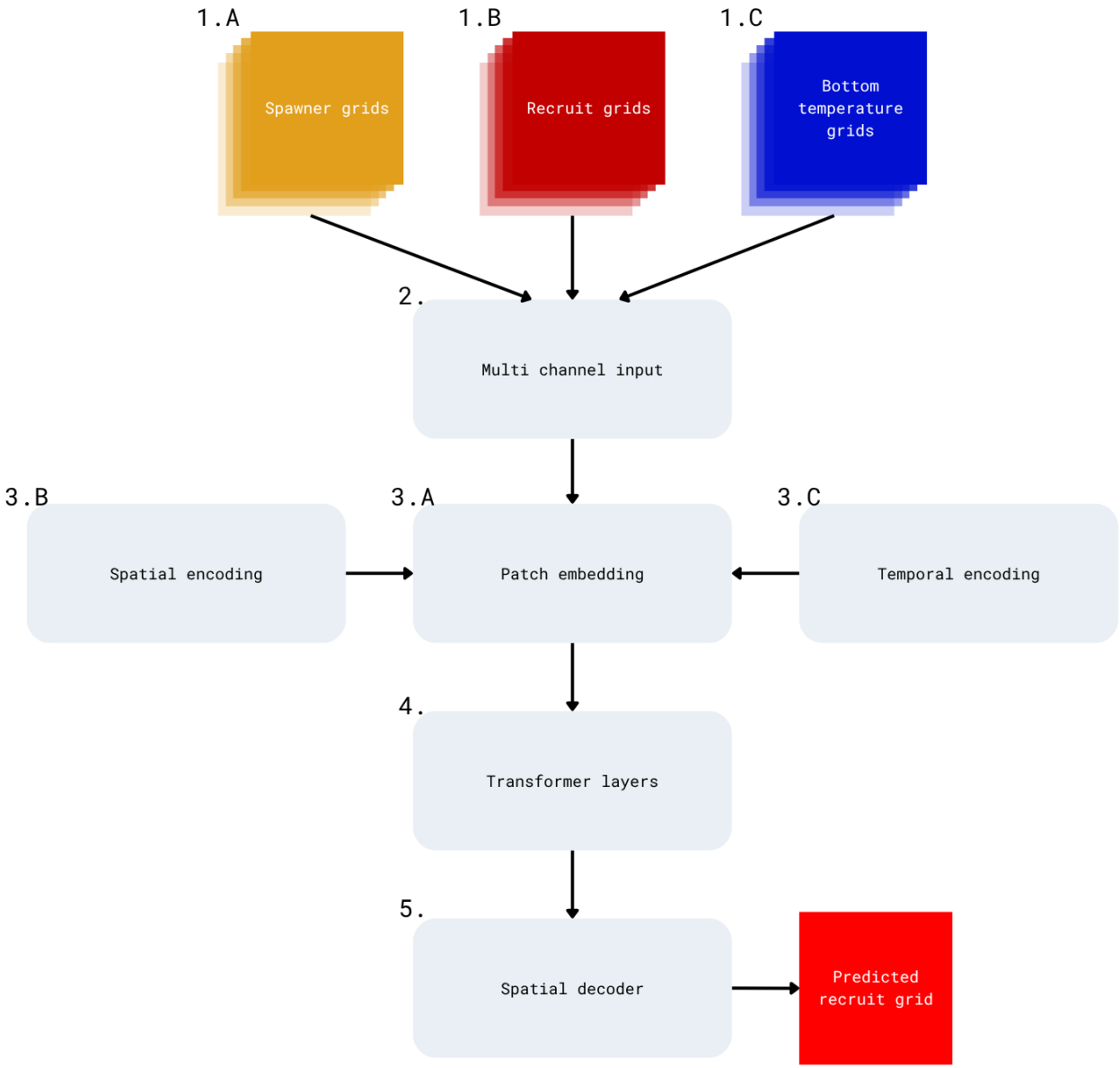


Figure 2.4. High-level ViT architecture. (1) Input channel grids are selected according to the model configuration: spawner grids (1a), recruit grids (1b), and bottom temperature grids (1c) are assembled into a multi-channel input tensor (2). (3.a) Each channel grid is divided into 100 non-overlapping 5 × 5 patches, and each patch is projected to a compact embedding vector via a learned linear transformation. Learned spatial encodings are added to each patch embedding to provide explicit positional context (3.b), and a fixed sinusoidal temporal encoding derived from the prediction year index is added to supply temporal context (3.c). (4) The resulting embeddings are passed through a series of transformer layers, each of which applies multi-head self-attention followed by a feed-forward network to model spatial relationships across all patch locations simultaneously. (5) A convolutional spatial decoder upsamples the patch-level representations back to the full 50 × 50 grid, producing a predicted recruit density map.

2.2.6 Temporal memory bank

A five-year historical window was selected to align with the estimated growth timeline of EBS snow crab, representing the approximate lag between spawning and consistent survey selectivity of the recruits. For each data type (spawners, recruits, and bottom temperature), the five years preceding year t were included as input channels, yielding up to 17 channels depending on

configuration: current-year spawner (channel 0), spawner history *t*-1 to *t*-5 (channels 1-5), recruit history *t*-1 to *t*-5 (channels 6-10), current-year temperature (channel 11), and temperature history *t*-1 to *t*-5 (channels 12-16). Current-year spawner density was included because the spatial distribution of egg-bearing females provides a direct indicator of recent spawning effort. Although recruits observed in the current year hatched approximately five years earlier, the current spawner field may still be informative because both adults and juveniles occupy the same cold pool habitat in the EBS (Fedewa et al., 2020), such that spawner and recruit distributions tend to co-vary through shared environmental preferences.

Channel	Content
0	Spawner (*t*)
1–5	Spawner (*t*-1 to *t*-5)
6–10	Recruit (*t*-1 to *t*-5)
11	Temperature (*t*)
12–16	Temperature (*t*-1 to *t*-5)

Table 2.1. Input channel configuration. Each channel corresponds to one spatial grid provided to the model as input. Channels 0 and 11 represent current-year observations and are excluded in one-year-ahead prediction mode. Channel availability by prediction mode is summarized in Table 2.2.

For years early in the time series where fewer than five years of historical data are available, missing channels are filled with zeros and flagged via a temporal mask vector of length 6 (index 0 for the current year, indices 1-5 for lookback years *t*-1 through *t*-5). Each input channel is mapped to its corresponding temporal slot, so that during patch embedding any channel whose slot is flagged as missing is multiplied by zero, preventing missing observations from influencing downstream computations. Zero-padded values are treated as missing rather than genuine zero-density observations. A similar masking mechanism handles the 2020 gap when it falls within the historical window.

Because samples in the first years of the validation and test splits lack a full five-year lookback window within that split, the final five years of the preceding split were carried forward as historical context. To illustrate how these steps work together: predicting recruitment for 2021, the first test year, the five preceding validation years are provided as historical context. The *t*-1 lookback channel corresponding to 2020 is flagged as missing by the temporal mask and its patch embedding is multiplied by zero, while the remaining valid years *t*-2 through *t*-5 are used normally.

2.2.6.1 Prediction modes

Three prediction modes differing in what information is available at prediction time were explored (summarized in Table 2.2). The key distinction is whether the current-year spawner density and temperature channels (channels 0 and 11) are available to the model.

- In now-cast mode, all 17 channels are available including current-year spawners and temperature (Figure 2.4, step 1; Table 2.2). Current-year temperature was included because it characterizes the thermal habitat available to juveniles at the time of the survey, which may influence their spatial distribution independently of spawning history.
- In one-year-ahead mode, the current-year spawner and temperature channels are excluded entirely from the input tensor rather than being zeroed, reducing the total input from 17 to

15 channels (Table 2.2). Exclusion rather than zeroing ensures the model cannot exploit any residual signal from masked channels, accurately reflecting the operational scenario in which the current year's survey has not yet been conducted.

- In the lag-5 mode, the training data were pre-aligned so that spawner and temperature observations from year *t*-5 were paired with recruit observations from year *t* (Table 2.2), with the alignment applied during data construction (Section 2.2.4).

The model architecture is identical across all prediction modes; only the temporal relationship between inputs and targets changes.

Model Name	**Data available to model**	**[*t*-10:*t*-6]**	**t-5**	***t*-4**	***t*-3**	***t*-2**	***t*-1**	**Target Year (*t*)**
Now-cast	Recruitment	-	✓	✓	✓	✓	✓	-
	Spawners	-	✓	✓	✓	✓	✓	✓
	Temperature	-	✓	✓	✓	✓	✓	✓
One-year Ahead	Recruitment	-	✓	✓	✓	✓	✓	-
	Spawners	-	✓	✓	✓	✓	✓	-
	Temperature	-	✓	✓	✓	✓	✓	-
5-year lag	Recruitment	-	✓	✓	✓	✓	✓	-
	Spawners	✓	-	-	-	-	-	-
	Temperature	✓	-	-	-	-	-	-

Table 2.2. Input channels available to the model under each prediction mode. Checkmarks indicate channels included as predictors; dashes indicate channels excluded. In the 5-year lag mode, spawner and temperature inputs are drawn from year *t*-5 rather than year *t*, reflecting the alignment of spawner observations with the estimated year of origin of the observed recruit cohort. Recruit history channels cover years *t*-1 through *t*-5 in all modes.

2.2.7 Sensitivity analysis

Given the fundamental data constraint described in Section 2.2, a range of sensitivity analyses was performed to ensure best model performance. These sensitivity analyses included using two loss functions during training, model size (number of parameters), and the configurations of channels available to the model.

2.2.7.1 MSE loss

The primary loss function was mean squared error (MSE), computed on log-transformed predictions and targets. For a single sample with spatial mask $M_{i,j}$ indicating valid cells, the per-sample loss is:

$$L_{MSE} = \frac{1}{N_{valid}} \sum_{i=1}^{N_{valid}} m_i \left(\hat{y}_i - y_i\right)^2$$

where $\hat{y}_i$ and y_i are the predicted and observed log-transformed recruit densities at grid cell i, and $N_{valid} = \sum_{i=1}^{n} m_i$is the number of valid cells, where n is the total number of cells. The batch loss is the mean of per-sample losses over valid samples (those not flagged as 2020 by the per-sample validity mask). Predictions were corrected for lognormal bias at evaluation time using the factor $exp(\mu_r + \sigma_r^2/2)$, where μ_r and $\sigma_r^2/2$ are the mean and variance of the training residuals in log space, so that the back-transformed prediction approximates the conditional mean density on the original scale rather than the geometric mean.

2.2.7.2 Tweedie loss

Models were also trained using the Tweedie deviance, which is specifically suited to non-negative continuous data with a high proportion of zeros. The Tweedie distribution with power parameter $1 < p < 2$ corresponds to a compound Poisson-Gamma distribution, placing a point mass at zero combined with a continuous distribution for positive densities. A power parameter of 1.2 was used and the model output $\hat{y}_i$is produced on the log1p scale (as in the MSE branch). Prior to computing the Tweedie loss, predictions were back-transformed to the original density scale via $\hat{\mu}_i = \text{expm1}(\hat{y}_i) = e^{\hat{y}_i} - 1$, while targets $y_i$are the observed densities on the original scale.. The per-sample Tweedie loss is:

$$L_{Tweedie} = \frac{1}{N_{valid}} \sum_{i=1}^{N_{valid}} m_i \left[\frac{-y_i \cdot \hat{\mu}_i^{1-p}}{1-p} + \frac{\hat{\mu}_i^{2-p}}{2-p} \right]$$

where $\hat{\mu}_i$ is the exponentiated prediction (lower-bounded to a small positive value for numerical stability), y_i is the observed density on the original scale, and m_i is the binary spatial mask.

The power parameter p was set to 1.2 following a limited search; an initial value of 1.5 was evaluated before settling on 1.2 based on validation loss. A more exhaustive grid search over p would be a natural refinement, though collaborative experience with similar ecological count data (M. Kapur, pers. comm.) suggests that treating p as a trainable parameter rather than a fixed hyperparameter yields limited additional benefit in practice. The current value of 1.2 places the distribution closer to the Poisson end of the compound Poisson-Gamma family, reflecting the high proportion of zero-density cells (~60%) relative to the positive tail, but the sensitivity of results to this choice was not formally evaluated and represents an avenue for future methodological refinement.

2.2.7.3 Model parameter size

The possibility of overfitting was a primary concern because the effective number of independent training patterns is small (~24 unique spatial configurations). The effects of this were examined by comparing models of two different sizes across all channel configurations and prediction modes. The base-size model uses an embedding dimension (the length of each patch's numerical representation) of 128, 8 attention heads (the number of attention heads allow the model to attend to different spatial relationships simultaneously, see Supplementary Section S.2.3.3 for more details), a feed-forward dimension of 512 (the dimension of the vectors used for non-linear 'learning', see Supplementary Section S.2.3.4 for more details), and 6 transformer layers (successive rounds of self-attention and information mixing, see Supplementary Section S.2.3.4 for more details), totaling approximately 1.3 million trainable parameters. The reduced-size model halves the number of attention heads to 4 and reduced the depth to 3 transformer layers, while keeping the embedding and feed-forward dimensions the same, reducing the total parameters to approximately 500,000. Reducing model capacity limits the number of patterns the model can

memorize, making it less likely to fit noise in the training data at the expense of generalizable signal. The two configurations were trained and evaluated using identical data splits and metrics, allowing direct comparison of whether the additional capacity of the full model improves generalization or merely increases overfitting.

2.2.7.4 Channel sensitivity

Different channel groups were systematically withheld to evaluate which types of historical information are most critical for recruitment prediction: spawners (current-year plus five-year history), recruits (five-year history only), and temperature (current-year plus five-year history). These were combined into seven distinct non-empty subsets (Table S.2.2). No structural changes to the ViT were required between runs because the model architecture dynamically adapts to the number of input channels at the patch embedding stage (Figure S.2.1, step 3). Each of the seven channel configurations was crossed with two model sizes, three prediction modes, and two loss functions. Combined with the synthetic benchmark datasets, this sensitivity analysis comprised 228 independent training runs.

2.2.8 Model selection

The model was trained using the AdamW optimizer, a gradient-based algorithm that iteratively adjusts model weights to minimize the loss function by computing partial derivatives of the loss with respect to each parameter (Loshchilov & Hutter, 2017). AdamW extends the Adam optimizer by incorporating weight decay, a regularization technique that adds a penalty proportional to the magnitude of each model parameter to the update step, discouraging the model from assigning excessively large weights to any individual input feature and thereby reducing overfitting. Training proceeded for 20 complete passes through the training dataset, referred to as epochs; each epoch exposes the model to every training sample once, after which weights are updated based on the accumulated gradient information (Figure 2.1, step B.3). Validation loss was evaluated after each epoch (Figure 2.1, step B.2) and the model checkpoint with the lowest validation loss was retained for evaluation on the held-out test set (Figure 2.1, step C.1).*2.2.9 Model evaluation and interpretability*

Model performance was assessed by comparing test set predictions (2021-2023) against observations in the survey area. Predicted values below the minimum observed crab density of 14 crabs km^{-2} were set to zero prior to computing metrics, as model predictions near the boundary of detectability are unreliable at very low densities. Final model ranking was based on test set performance.

Absolute accuracy was assessed using mean absolute error (MAE) in log1p space (Figure 2.1, step C.2a). Spatial pattern accuracy was assessed using Spearman's correlation between predicted and observed fields, providing a distribution-free measure robust to scale differences (Figure 2.1, step C.2b).

Throughout this study, model predictions and climatological baseline predictions were both evaluated against SPDE-interpolated spatial fields rather than raw station-level observations, meaning all methods compete on equal footing against the same smoothed truth. The implications of this for interpreting Spearman's correlation in particular are discussed in Section 2.4.5.

To enable comparison across metrics with different scales and directions, min-max normalization to MAE and Spearman's correlation was applied across all model configurations evaluated on the same dataset. Each metric was rescaled to [0, 1] as: $normalized = (x - x_{min})/(x_{max} - x_{min})$, where x_{min} and x_{max} are the minimum and maximum values observed

across configurations. The normalized score for MAE value was inverted (i.e., 1 − normalized) because a lower MAE indicates better performance. A composite score was then computed as the geometric mean of the two normalized metrics: $composite = \sqrt{MAE_{norm} \cdot Spearman_{norm}}$. Model configurations were ranked using this composite score. The geometric mean was chosen over the arithmetic mean because it penalizes configurations that perform poorly on either metric, favoring balanced performance over strong performance on one metric at the expense of the other.

To contextualize model performance, a climatological baseline was computed for both the actual and synthetic datasets. For each grid cell in a bootstrap timeseries, the mean recruit density was calculated across all valid training years after log1p transformation, then back-transformed to count space. This per-cell mean grid was used as a constant prediction for all test years, evaluated using the same spatial mask and year mask applied during model evaluation. The performance metrics across the 100 bootstraps were then averaged to calculate the mean climatological baseline performance scores.

2.2.10 Teleconnection study via Integrated Gradients

Integrated Gradients (IG; Sundararajan et al., 2017) were used to investigate which spatial regions and input channels drive the recruitment predictions from the ViT model (Figure 2.1, step C.2c). IG is a post-hoc attribution method that assigns an importance score to every input value based on its contribution to a model's prediction. These attribution maps also serve as a data-driven tool for identifying spatial teleconnections learned by the model without prior specification.

The method compares the model's prediction for an actual input against a baseline representing average historical conditions, defined as the grid cell mean of all valid non-2020 training samples across all bootstrap replicates. Attribution is computed by gradually interpolating between the baseline and the actual input in 50 equal steps, computing the model's sensitivity at each step, and averaging:

$$IG(x) = (x - x_{base}) \cdot \frac{1}{z} \sum_{i=1}^{z} \nabla_x f(x_{base} + \frac{i}{m}(x - x_{base}))$$

where x is the actual input, x_{base} is the baseline, $z = 50$ is the number of interpolation steps, and f is the model's total predicted recruitment (summed over valid cells). The resulting attribution map has the same dimensions as the input (17 channels × 50 × 50 grid cells) and satisfies a completeness property: the sum of all attributions equals exactly the difference between the model's prediction for the actual input and its prediction for the baseline. The attribution map fully accounts for why the model predicted more or less recruitment in a given year compared to what it would predict under average conditions. Gradients at each interpolation step were computed via backpropagation through the full model using PyTorch's automatic differentiation engine, which tracks all arithmetic operations performed on input tensors during the forward pass and applies the chain rule to compute exact partial derivatives of the model output with respect to each input value.

Attribution was computed independently for each bootstrap replicate and test year, then averaged across replicates to reduce noise. Spatially resolved attribution maps and channel-level importance summaries were produced for each of four highlighted model configurations (Figures 2.6-2.13). Aggregated maps summing absolute attribution across all channels were produced to identify spatial regions most influential to recruitment prediction regardless of which channel drove the effect. Total absolute attribution per channel was computed to assess the importance of the relative channel; computed by summing IG scores across all valid grid cells, producing one importance score per channel for each prediction year.

2.3 Results

2.3.1 Synthetic benchmark validation

Performance on three synthetic datasets spanning a gradient of difficulty scenarios was assessed to establish baseline capabilities before applying the model to actual survey data. Across all difficulty scenarios, models trained using the Tweedie loss consistently outperformed MSE-trained models in overall predictive skill (Table 2.3).

The climatological baseline achieved a mean spatial Spearman's correlation of 0.197, 0.163, and 0.119 across the easy, medium, and hard difficulty scenarios respectively, with baseline MAE values of 4,483, 4,274, and 3,900 crabs km^{-2}. The ViT model improved over this baseline in both metrics, across all prediction modes, confirming the model was able to learn a generalizable spatial structure rather than recovering a static long-run mean.

The ViT model accurately recovered the underlying spatial fields in the Easy scenario (spawner-recruit correlation of 0.9, no lag). The best-performing configuration achieved a Spearman's correlation of 0.791 and a composite score of 0.976 (Table 2.3). Withholding current-year observations caused Spearman's correlation to collapse to 0.190, which reflects the structure of the problem rather than a model failure: when the true relationship operates with zero lag, the current-year spawner field contains the primary predictive signal, and its removal leaves little recoverable information in the historical channels alone. This collapse therefore serves as an internal validation that the ViT model was exploiting genuine predictive structure rather than spurious correlations in the historical record.

The ViT model successfully learned the delayed spatial dependency in the Medium scenario (spawner-recruit correlation of 0.9, 3-year lag, temporal autocorrelation of 0.7). The best-performing now-cast configuration (base-size model, spawner and recruit channels, Tweedie loss) achieved a Spearman's correlation of 0.796 and a composite score of 0.944. Notably, the one-year-ahead configuration performed slightly better (composite score = 0.952), demonstrating that when the true signal operates on a multi-year lag, withholding current-year observations does not degrade accuracy and may reduce overfitting to short-term noise.

Spearman's correlation declined in the Hard scenario (spawner-recruit correlation of 0.8, 5-year lag, temporal autocorrelation of 0.7), as expected given the weaker underlying signal. The top now-cast configuration (reduced-size model, spawner and recruit channels, Tweedie loss) achieved a Spearman's correlation of 0.701 and a composite score of 0.980. The one-year-ahead configuration matched this closely (composite score = 0.973), confirming that the model learned the five-year lagged relationship rather than relying on current-year information. The reduced-size model's advantage in this scenario recurred throughout our analyses: as prediction difficulty increases and the effective signal-to-noise ratio decreases, smaller architectures generalized more reliably than larger ones (see Section 2.3.3).

Across all synthetic benchmarks, configurations restricted to recruit history channels alone performed poorly, confirming that the spatial autocorrelation of juvenile distributions alone is insufficient for recruitment prediction even when a true spawner-recruit relationship exists. This is highlighted quite starkly when performance is compared to the climatological baseline. This is most evident in the Spearman's correlation metric, which highlights the model's ability to learn relationships.

Model Size	Scenario	Channel	Prediction Mode	Loss	MAE	Spearman	Normalized MAE	Normalized Spearman	Composite

Climatological Baseline	Easy	N/A	N/A	N/A	4486	0.108	N/A	N/A	N/A
Reduced	Easy	Spawners	Now-cast	Tweedie	**3801**	**0.791**	1	0.953	0.976
Base	Easy	Spawners + Recruits	Now-cast	Tweedie	3859	0.791	0.999	0.953	0.976
Base	Easy	Spawners	Now-cast	Tweedie	3871	0.790	0.998	0.952	0.975
Reduced	Easy	Spawners	One-year-ahead	Tweedie	5891	0.190	0.67	0.049	0.218
Reduced	Easy	Spawners + Recruits	One-year-ahead	Tweedie	6035	0.184	0.965	0.041	0.199
Base	Easy	Spawners	One-year-ahead	Tweedie	6054	0.184	0.965	0.041	0.199
Climatological Baseline	Medium	N/A	N/A	N/A	4335	0.018	N/A	N/A	N/A
Base	Medium	Spawners + Recruits	Now-cast	Tweedie	3539	0.796	0.997	0.893	0.944
Reduced	Medium	Spawners + Recruits	Now-cast	Tweedie	**3339**	0.789	1	0.871	0.933
Reduced	Medium	Spawners	Now-cast	Tweedie	3809	0.782	0.994	0.849	0.919
Base	Medium	Spawners + Recruits	One-year-ahead	Tweedie	3383	**0.800**	0.999	0.906	0.952
Reduced	Medium	Spawners + Recruits	One-year-ahead	Tweedie	3416	0.793	0.999	0.883	0.939
Reduced	Medium	Spawners	One-year-ahead	Tweedie	3974	0.782	0.992	0.847	0.916
Climatological Baseline	Hard	N/A	N/A	N/A	3934	0.027	N/A	N/A	N/A
Reduced	Hard	Spawners + Recruits	Now-cast	Tweedie	3654	**0.701**	0.998	0.962	0.980
Base	Hard	Spawners + Recruits	Now-cast	Tweedie	3656	0.694	0.998	0.933	0.965
Reduced	Hard	Spawners	Now-cast	Tweedie	4195	0.673	0.992	0.840	0.912
Base	Hard	Spawners + Recruits	One-year-ahead	Tweedie	**3462**	0.697	1	0.947	0.973
Reduced	Hard	Spawners + Recruits	One-year-ahead	Tweedie	3896	0.621	0.995	0.925	0.959
Reduced	Hard	Spawners	One-year-ahead	Tweedie	4150	0.602	0.992	0.833	0.909

Table 2.3. The Climatological Baseline and the top three model configurations for each difficulty scenario and prediction mode for the synthetic benchmark datasets, ranked by composite score. Only now-cast and one-year-ahead modes were evaluated on synthetic data; the 5-year lag mode was evaluated on actual survey data only. Model size:

reduced (~500K parameters), and base (~1.3M parameters). Composite score is the geometric mean of normalized MAE and normalized Spearman's correlation (Section 2.2.9).

2.3.2 Predictive performance for EBS Snow Crab

The synthetic benchmarks confirmed that the model could learn spawner-recruit relationships across a range of lag structures and signal strengths under controlled conditions. Performance on actual EBS survey data introduced additional complications: a weaker and less certain underlying signal, an effective sample size of roughly 24 independent spatial configurations, and spatial complexity that the synthetic generative process did not attempt to replicate.

Model Size	Channel	Prediction Mode	Loss	MAE	Spearman	Normalized MAE	Normalized Spearman	Composite
Climatological Baseline	N/A	Now-Cast + One-year-ahead	N/A	500	0.807	N/A	N/A	N/A
Base	All	Now-cast	MSE	332	0.866	0.973	0.973	**0.973**
Base	Spawner + Temperature	Now-cast	MSE	**309**	0.860	0.986	0.938	0.962
Reduced	Temperature	Now-cast	MSE	376	**0.869**	0.947	0.968	0.958
Reduced	Recruit + Temperature	One-year-ahead	MSE	459	0.868	0.899	0.968	0.932
Base	Spawner + Temperature	One-year-ahead	MSE	381	0.855	0.944	0.918	0.931
Reduced	All	One-year-ahead	MSE	340	0.848	0.968	0.896	0.931
Climatological Baseline	N/A	5-year-lag	N/A	455	0.812	N/A	N/A	N/A
Base	Spawner + Temperature	5-year lag	MSE	416	**0.832**	0.924	0.837	0.879
Base	Spawner + Recruit	5-year lag	MSE	**401**	0.827	0.932	0.821	0.875
Reduced	Spawner	5-year lag	MSE	479	0.828	0.887	0.823	0.855

Table 2.4. The Climatological Baseline and the top three model configurations for each prediction mode on the actual EBS survey data, ranked by composite score. Model size: reduced (~500K parameters), and base (~1.3M parameters). Composite score is the geometric mean of normalized MAE and normalized Spearman rank correlation (Section 2.2.9).

Under now-cast mode, the base-size model using all 17 channels achieved a Spearman's correlation of 0.866 and an MAE of 332 crabs km^{-2}, the highest composite score across all actual-data configurations (Table 2.4). Performance for the base-size model in now-cast mode was consistent across test years, with Spearman's correlations of 0.864, 0.864, and 0.872 in 2021, 2022, and 2023, respectively, and MAE ranging from 245 to 449 crabs km^{-2} (Table 2.5). The 2022 test year showed the largest error for this configuration, consistent with the broader pattern of models underestimating the magnitude of the sharp population fluctuations that characterized the validation and test period. The temperature-only configuration ranked third overall despite relying

entirely on abiotic inputs, achieving a mean Spearman's correlation of 0.869 across test years, though MAE was more variable (66 to 570 crabs km^{-2}), reflecting the configuration's ability to rank spatial patterns reliably but its less consistent ability to calibrate absolute density.

Predictive skill declined only modestly when current-year spawner and temperature observations were withheld. The top one-year-ahead configuration (reduced-size model using recruit and temperature channels) achieved a composite score of 0.932 and Spearman's correlations of 0.894, 0.857, and 0.854 in 2021, 2022, and 2023 (Table 2.5). The small difference relative to now-cast performance is consistent with the pattern observed in the medium synthetic scenario, where multi-year lagged relationships reduced dependence on current-year inputs.

Predicting recruitment five years in advance, to account for the ontogenetic lag between spawning and being observed in the groundfish survey, led to the largest performance reduction. Despite this, spatial skill remained substantial: the best configuration (base-size model, spawner and temperature channels) achieved a composite score of 0.879 and Spearman's correlations of 0.824, 0.808, and 0.863 across the 2021, 2022, and 2023 test years (Table 2.5). The 2022 test year had the weakest performance, again consistent with greater prediction difficulty during years of rapid population change.

To provide a lower bound for interpreting model performance, two climatological baselines were created for the actual EBS data. One climatological baseline was for the now-cast and one-year-ahead models since they shared the same training data, and the other climatological baseline was for the 5-year lag model, since it had fewer training years. The per-cell mean of training years for the now-cast and one-year-ahead mode achieved a spatial Spearman's correlation of 0.807 and an MAE of 500 crabs km^{-2}. The climatological baseline for the 5-year lag mode achieved a Spearman's correlation of 0.812 and an MAE of 455 crabs km^{-2}. MAE provides more meaningful discrimination between models than Spearman's correlation for these data: the best-performing models reduced MAE by 33-38% relative to the baseline (from 500 to 309-332 crabs km^{-2}), representing genuine improvement in absolute prediction accuracy beyond what static climatology can achieve. The 5-year lag baseline performed comparably to the now-cast baseline in Spearman's correlation (0.812 vs. 0.807) but had a lower MAE (455 vs. 500 crabs km^{-2}), likely reflecting the reduced range of population densities in the shorter training window available to the lag model; the best 5-year lag configurations reduced MAE by only 9-14% relative to this baseline (from 455 to 401-416 crabs km^{-2}), a substantially smaller margin than the 33-38% reduction achieved by now-cast models, consistent with the fundamental difficulty of predicting recruitment five years in advance.

Spatial prediction maps and aggregate time series for a representative subset of configurations: the top-performing model in each prediction mode (now-cast, one-year-ahead, and 5-year lag), plus the temperature-only now-cast configuration, which achieved the third-highest composite score across all models despite relying solely on abiotic inputs, were examined to complement the quantitative metrics (Figure 2.5). Spatially, predicted recruit distributions were broadly consistent across all four configurations, generally capturing the northern extent of high-density areas and correctly excluding regions outside the range of snow crab. Configurations differed most in their representation of the southern population boundary, where lower densities and greater spatial uncertainty made precise delineation more difficult. The strong Spearman's correlations reported above reflect this shared ability to rank spatial patterns correctly.

Both the now-cast/one-year-ahead and the 5-year lag climatological baseline's spatial structure covers more of the survey area than the model predictions, and shows a deeper color, reflecting the higher density and wider ranging footprint of recruits in training years.

The aggregate time series, showing median summed predicted density across all 100 bootstraps alongside the observed median, revealed a consistent pattern: models tracked fluctuations in the training period closely but underestimated the magnitude of extreme changes in the validation period, particularly the sharp population spike observed during the validation years (Figure 2.5). Potential explanations for this pattern, including model overfitting to the training distribution, structural shifts in the population, and the inherent difficulty of extrapolating to density ranges not well represented during training, are addressed in the Discussion.

Test Year	Model Size	Channels	Mode	Loss	MAE (crabs/km^2)	MAE SD	Spearman	Spearman SD
2021	Base	All	Now-cast	MSE	300.9	15.1	0.8635	0.0177
2022	Base	All	Now-cast	MSE	449.1	77.7	0.8643	0.0156
2023	Base	All	Now-cast	MSE	244.6	37.5	0.8715	0.0097
2021	Base	Spawner + Temp	Now-cast	MSE	110.3	6.6	0.8435	0.0184
2022	Base	Spawner + Temp	Now-cast	MSE	449.2	80.3	0.8736	0.0136
2023	Base	Spawner + Temp	Now-cast	MSE	367	46.8	0.8632	0.0097
2021	Reduced	Temp only	Now-cast	MSE	65.6	4.2	0.8517	0.0149
2022	Reduced	Temp only	Now-cast	MSE	491.1	82.1	0.8579	0.0111
2023	Reduced	Temp only	Now-cast	MSE	570.2	49.2	0.8962	0.0076
2021	Base	Spawner + Temp	One-year-ahead	MSE	143.4	8.3	0.8588	0.0164
2022	Base	Spawner + Temp	One-year-ahead	MSE	524.6	82.8	0.8662	0.0165
2023	Base	Spawner + Temp	One-year-ahead	MSE	476.1	48.6	0.8391	0.0109
2021	Reduced	All	One-year-ahead	MSE	107.2	5.2	0.8711	0.0144
2022	Reduced	All	One-year-ahead	MSE	481.2	80	0.804	0.0144
2023	Reduced	All	One-year-ahead	MSE	433.3	46	0.8699	0.0094
2021	Reduced	Recruit + Temp	One-year-ahead	MSE	471.1	13.1	0.8944	0.016
2022	Reduced	Recruit + Temp	One-year-ahead	MSE	456.2	76.7	0.8568	0.0187
2023	Reduced	Recruit + Temp	One-year-ahead	MSE	450.5	47.6	0.8544	0.0117
2021	Base	Spawner + Temp	5-year lag	MSE	581.4	10.3	0.8243	0.0151

2022	Base	Spawner + Temp	5-year lag	MSE	350.3	52.2	0.8084	0.0115
2023	Base	Spawner + Temp	5-year lag	MSE	315.3	44.9	0.8626	0.0078
2021	Base	Spawner + Recruit	5-year lag	MSE	389.1	10.9	0.8238	0.0145
2022	Base	Spawner + Recruit	5-year lag	MSE	413.8	67.1	0.8186	0.0108
2023	Base	Spawner + Recruit	5-year lag	MSE	400.5	29.6	0.8399	0.0092
2021	Reduced	Spawners only	5-year lag	MSE	339	17.7	0.8314	0.0168
2022	Reduced	Spawners only	5-year lag	MSE	496.4	61.8	0.8176	0.0108
2023	Reduced	Spawners only	5-year lag	MSE	601	45.7	0.8349	0.0083
2021	Climatological baseline		Now-cast/One-year ahead		678	33.8	0.809	0.0114
2022	Climatological baseline		Now-cast/One-year ahead		520.6	53.0	0.7871	0.0084
2023	Climatological baseline		Now-cast/One-year ahead		301.5	28.8	0.8247	0.0077
2021	Climatological baseline		5-year lag		605	32.4	0.8108	0.0132
2022	Climatological baseline		5-year lag		483.9	54.4	0.7938	0.0088
2023	Climatological baseline		5-year lag		275	32.8	0.8312	0.0085

Table 2.5. Per-year predictive performance for the top three configurations in each prediction mode on the real EBS survey data. MAE in crabs km^{-2}. SD estimated across 100 bootstrap replicates.

Top Model Summary (2023)

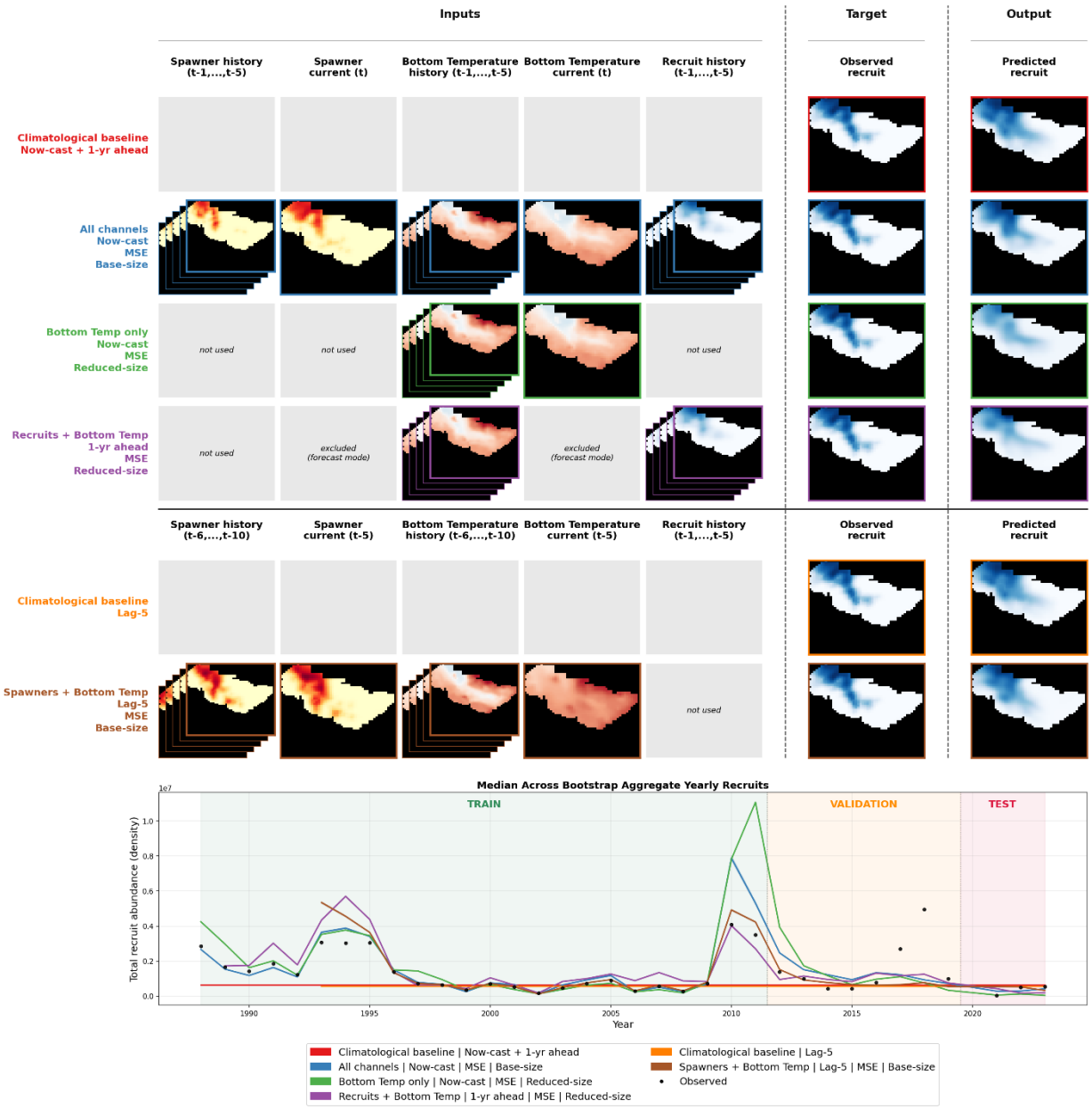


Figure 2.5. Summary of top model configurations evaluated for the 2023 test data. Upper panel: spatial input and output grids for each of the four highlighted model configurations and two climatological baselines (rows). From left to right, columns show spawner history (*t*-1 to *t*-5 for now-cast and one-year-ahead models; *t*-6 to *t*-10 for the lag-5 model), current-year spawner density (or the spawner field at *t*-5 for the lag model; shown as not applicable where excluded by prediction mode), bottom temperature history, current-year bottom temperature (excluded in one-year-ahead mode), recruit history (*t*-1 to *t*-5), observed recruit density, and model-predicted recruit density. Each row corresponds to one model configuration, identified by the border color used in both panels. Spatial input and output grids are shown for the 2023 test year. A single bootstrap replicate is displayed for each model. Lower panel: yearly aggregate predicted and observed recruit abundance across all splits. Lines show median predicted total recruitment across 100 bootstrap replicates; dots show median observed total recruitment. Shaded regions indicate training (green), validation (orange), and test (red) periods. Colors correspond to the four model configurations shown in the upper panel.

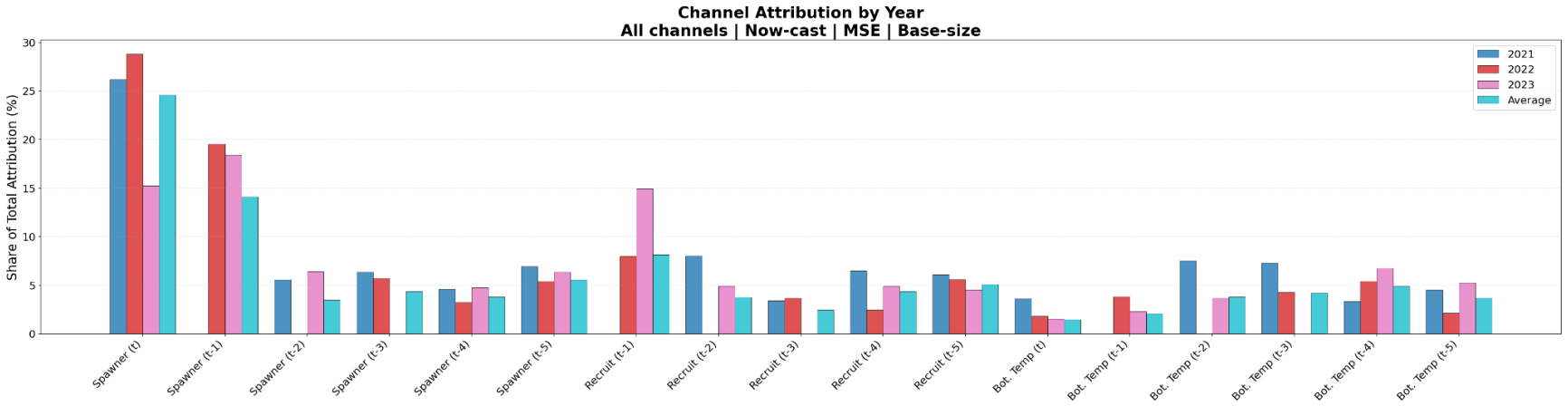


Figure 2.6. Channel-level attribution summary for the base-size, all-channel, now-cast, MSE model. Bar charts show the proportion of total absolute attribution assigned to each input channel for each of the three test years (2021, 2022, 2023) and the mean across test years. Higher values indicate greater influence of that channel on the model's predictions relative to the average baseline year. The 2020 survey was cancelled due to the COVID-19 pandemic, resulting in no observations for that year; channels dependent on 2020 survey data were set to zero, and the corresponding bars are absent from the attribution plots.

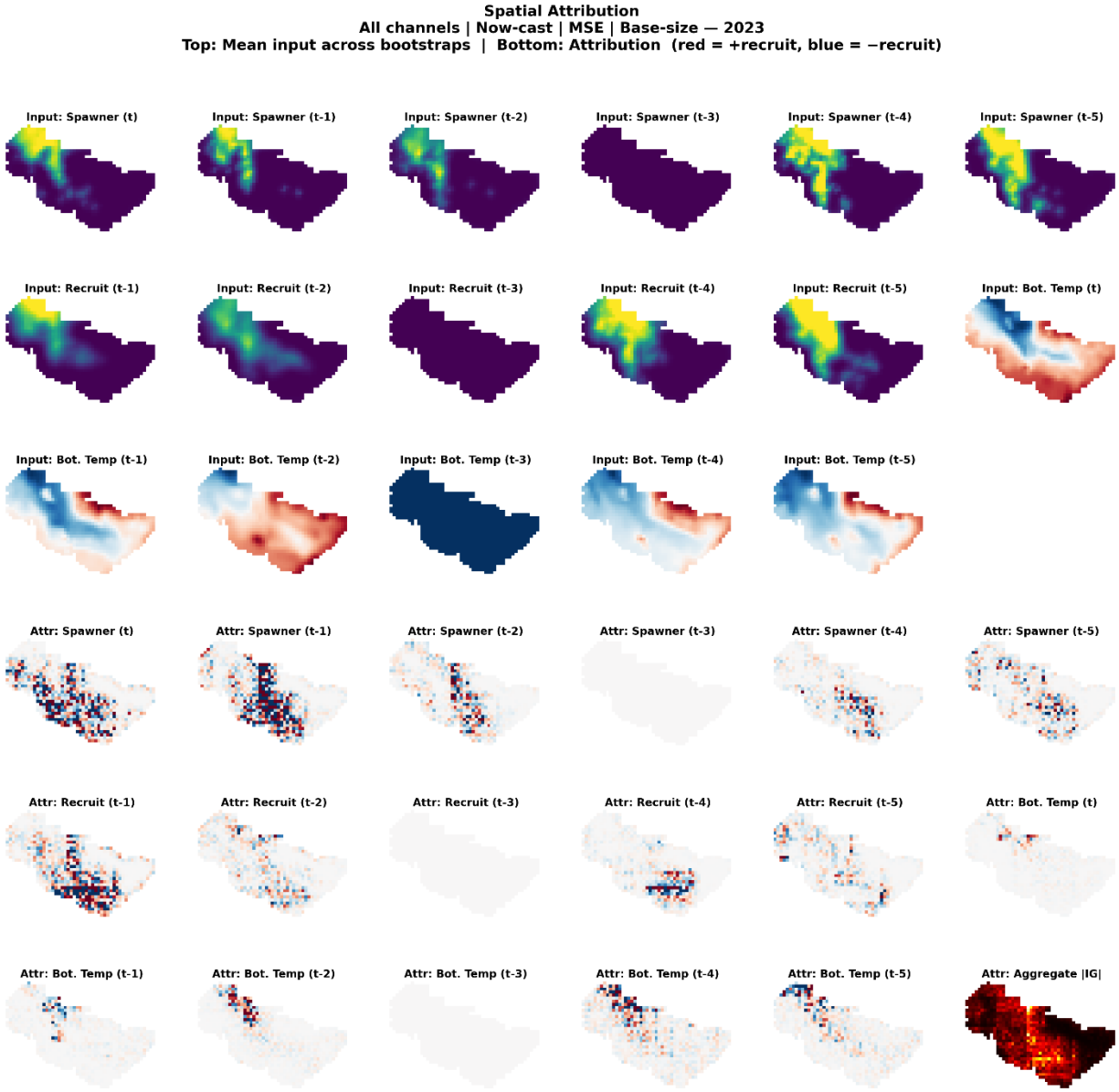

Spatial Attribution
All channels | Now-cast | MSE | Base-size — 2023
Top: Mean input across bootstraps | Bottom: Attribution (red = +recruit, blue = −recruit)
Input: Spawner (t)
Input: Spawner (t-1)
Input: Spawner (t-2)
Input: Spawner (t-3)
Input: Spawner (t-4)
Input: Spawner (t-5)
Input: Recruit (t-1)
Input: Recruit (t-2)
Input: Recruit (t-3)
Input: Recruit (t-4)
Input: Recruit (t-5)
Input: Bot. Temp (t)
Input: Bot. Temp (t-1)
Input: Bot. Temp (t-2)
Input: Bot. Temp (t-3)
Input: Bot. Temp (t-4)
Input: Bot. Temp (t-5)
Attr: Spawner (t)
Attr: Spawner (t-1)
Attr: Spawner (t-2)
Attr: Spawner (t-3)
Attr: Spawner (t-4)
Attr: Spawner (t-5)
Attr: Recruit (t-1)
Attr: Recruit (t-2)
Attr: Recruit (t-3)
Attr: Recruit (t-4)
Attr: Recruit (t-5)
Attr: Bot. Temp (t)
Attr: Bot. Temp (t-1)
Attr: Bot. Temp (t-2)
Attr: Bot. Temp (t-3)
Attr: Bot. Temp (t-4)
Attr: Bot. Temp (t-5)
Attr: Aggregate |IG|

Figure 2.7. Integrated Gradients spatial attribution maps for the base-size, all-channel, now-cast, MSE model applied to the 2023 test year. **Upper panel:** input channel grids used by the model, showing spawner density at times *t* through *t*-5 (channels 0-5), recruit history at *t*-1 through *t*-5 (channels 6-10), and bottom temperature at *t* through *t*-5 (channels 11-16). Spawner and recruit inputs are shown on a yellow-to-purple scale (low to high density); temperature inputs are shown on a blue-to-red scale (cool to warm). **Lower panel:** Integrated Gradients attribution maps for each corresponding input channel. Red indicates positive attribution (input values at that location increase predicted total recruitment relative to the baseline); blue indicates negative attribution (input values decrease predicted recruitment). Strong red attribution in a region indicates that above-average values there (relative to the baseline) pushed predicted recruitment upward; strong blue indicates the opposite. Channels with faint or near-zero attribution contributed little to the prediction for this year regardless of their input values. The aggregate map (bottom right) summarizes where spatial signal was concentrated across all channels combined. **Bottom right:** aggregate attribution map showing the sum of absolute attribution across all 17 channels, where darker values indicate grid cells that most strongly influenced the model's prediction regardless of direction. For the 2023 test year, the *t*-3 time step corresponds to the cancelled 2020 survey; channels for that year were set to zero, and the corresponding input and attribution panels are blank. Values shown are averaged across 100 bootstrap replicates for the 2023 test year.

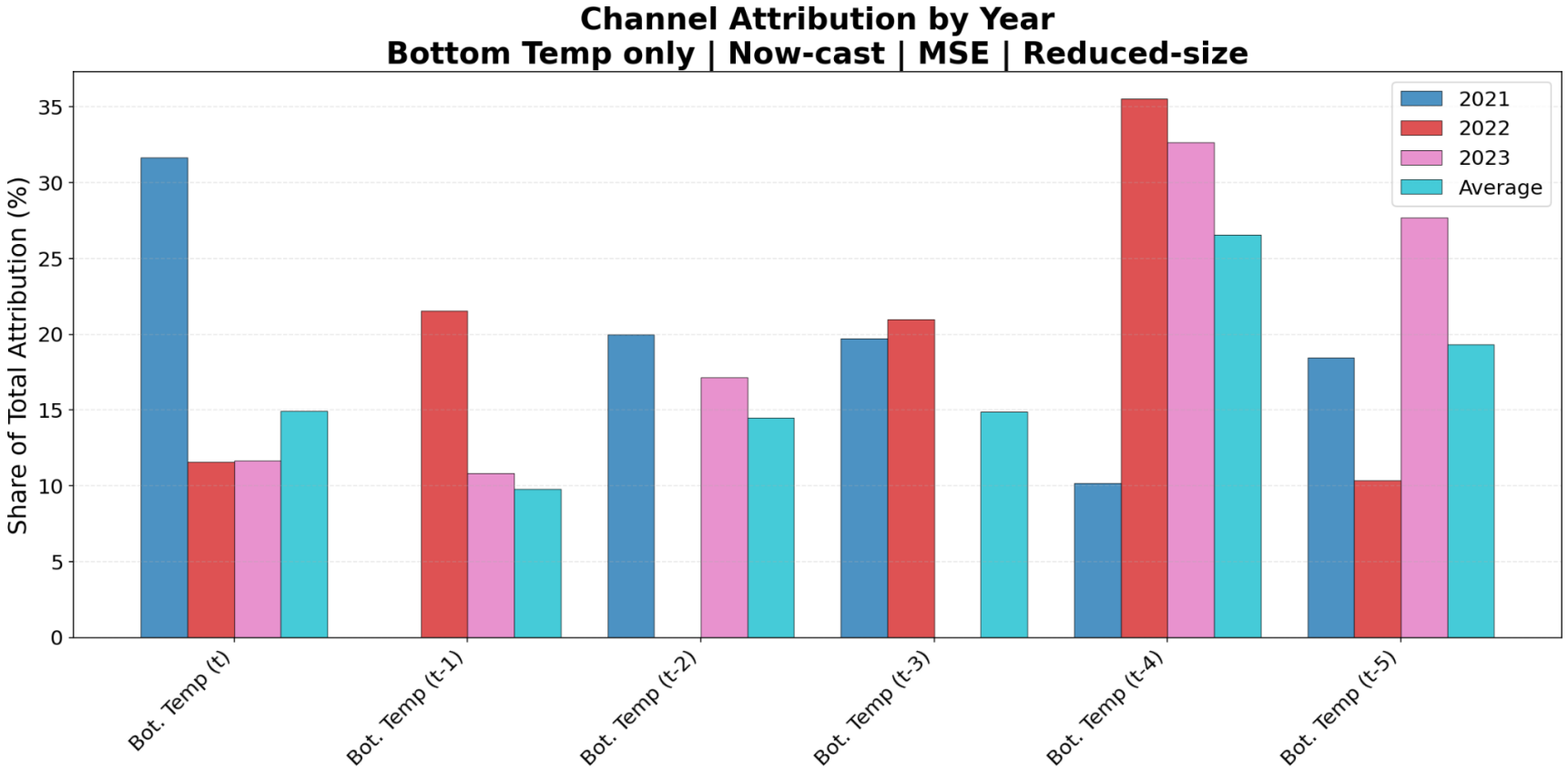


Figure 2.8. Channel-level attribution summary for the reduced-size, bottom temperature only, now-cast MSE model. Bar charts show the proportion of total absolute attribution assigned to each input channel for each of the three test years (2021, 2022, 2023) and the mean across test years. Higher values indicate greater influence of that channel on the model's predictions relative to the average baseline year. The 2020 survey was cancelled due to the COVID-19 pandemic, resulting in no observations for that year; channels dependent on 2020 survey data were set to zero, and the corresponding bars are absent from the attribution plots.

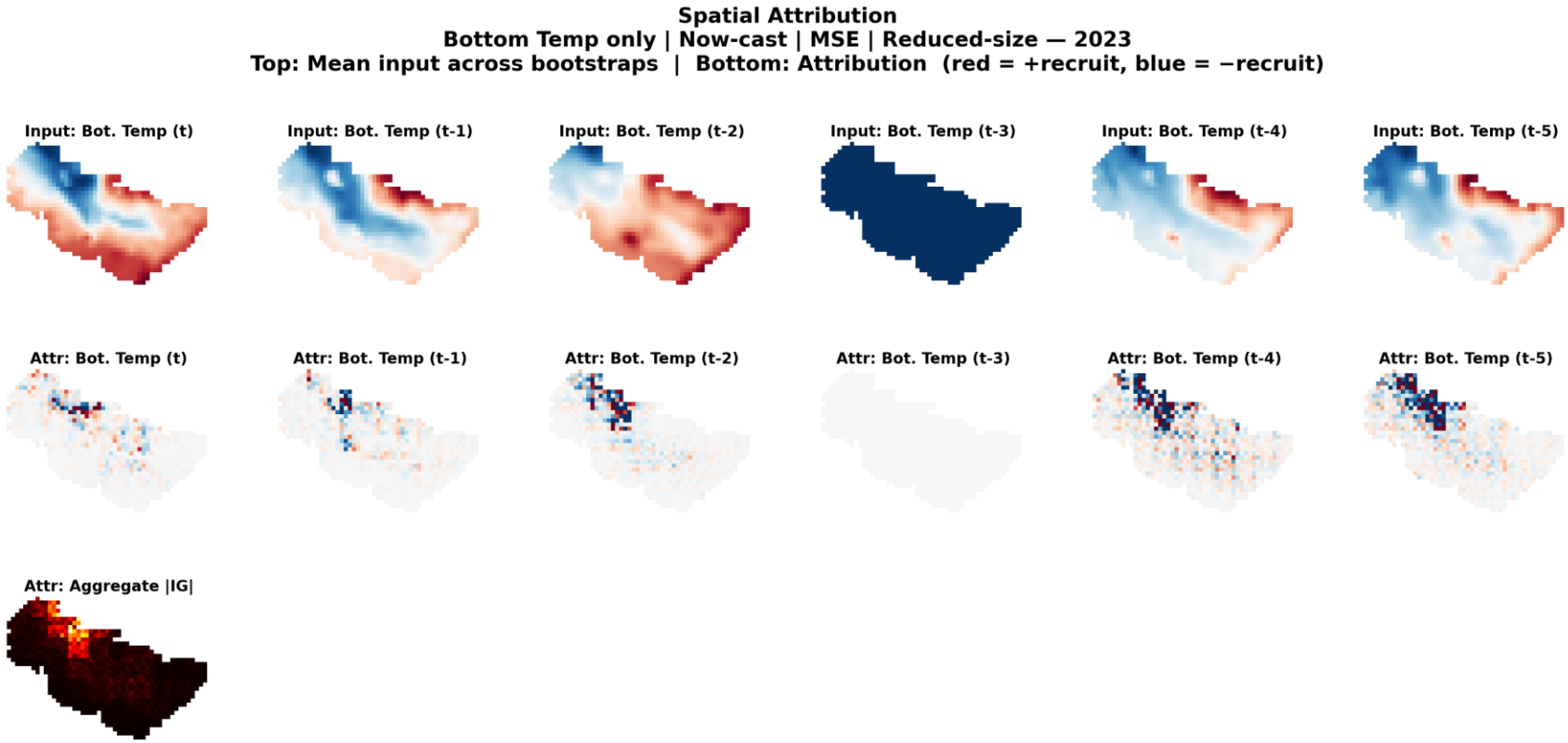


Figure 2.9. Integrated Gradients spatial attribution maps for the reduced-size, bottom temperature only, now-cast MSE model applied to the 2023 test year. **Upper panel:** input channel grids used by the model, showing bottom temperature at *t* through *t*-5 (channels 0-5). Temperature inputs are shown on a blue-to-red scale (cool to warm). **Lower panel:** Integrated Gradients attribution maps for each corresponding input channel. Red indicates positive attribution (input values at that location increase predicted total recruitment relative to the baseline); blue indicates negative attribution (input values decrease predicted recruitment). Strong red attribution in a region indicates that above-average values there (relative to the baseline) pushed predicted recruitment upward; strong blue indicates the opposite. Channels with faint or near-zero attribution contributed little to the prediction for this year regardless of their input values. The aggregate map (bottom right) summarizes where spatial signal was concentrated across all channels combined. **Bottom right:** aggregate attribution map showing the sum of absolute attribution across all 6 channels, where darker values indicate grid cells that most strongly influenced the model's prediction regardless of direction. For the 2023 test year, the *t*-3 time step corresponds to the cancelled 2020 survey; channels for that year were set to zero, and the corresponding input and attribution panels are blank. Values shown are averaged across 100 bootstrap replicates for the 2023 test year.

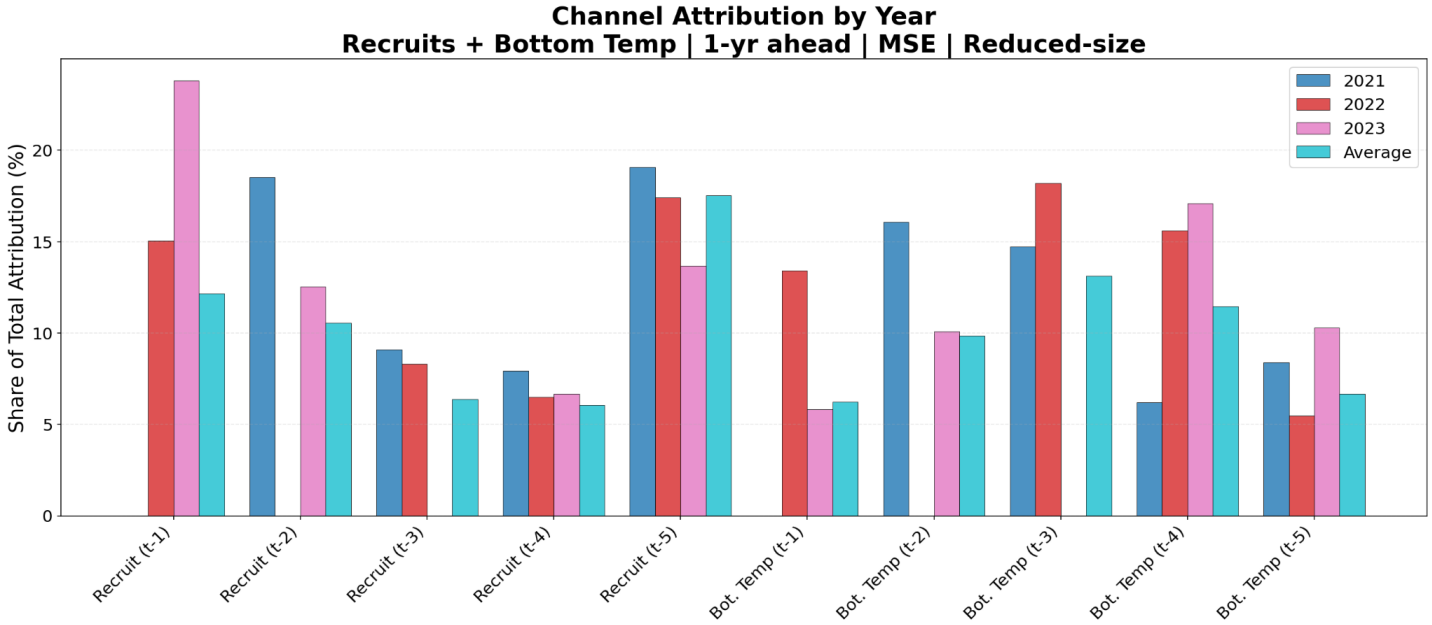


Figure 2.10. Channel-level attribution summary for the reduced-size, recruit and bottom temperature channels, one-year-ahead MSE model. Bar charts show the proportion of total absolute attribution assigned to each input channel for each of the three test years (2021, 2022, 2023) and the mean across test years. Higher values indicate greater influence of that channel on the model's predictions relative to the average baseline year. The 2020 survey was cancelled due to the COVID-19 pandemic, resulting in no observations for that year; channels dependent on 2020 survey data were set to zero, and the corresponding bars are absent from the attribution plots.

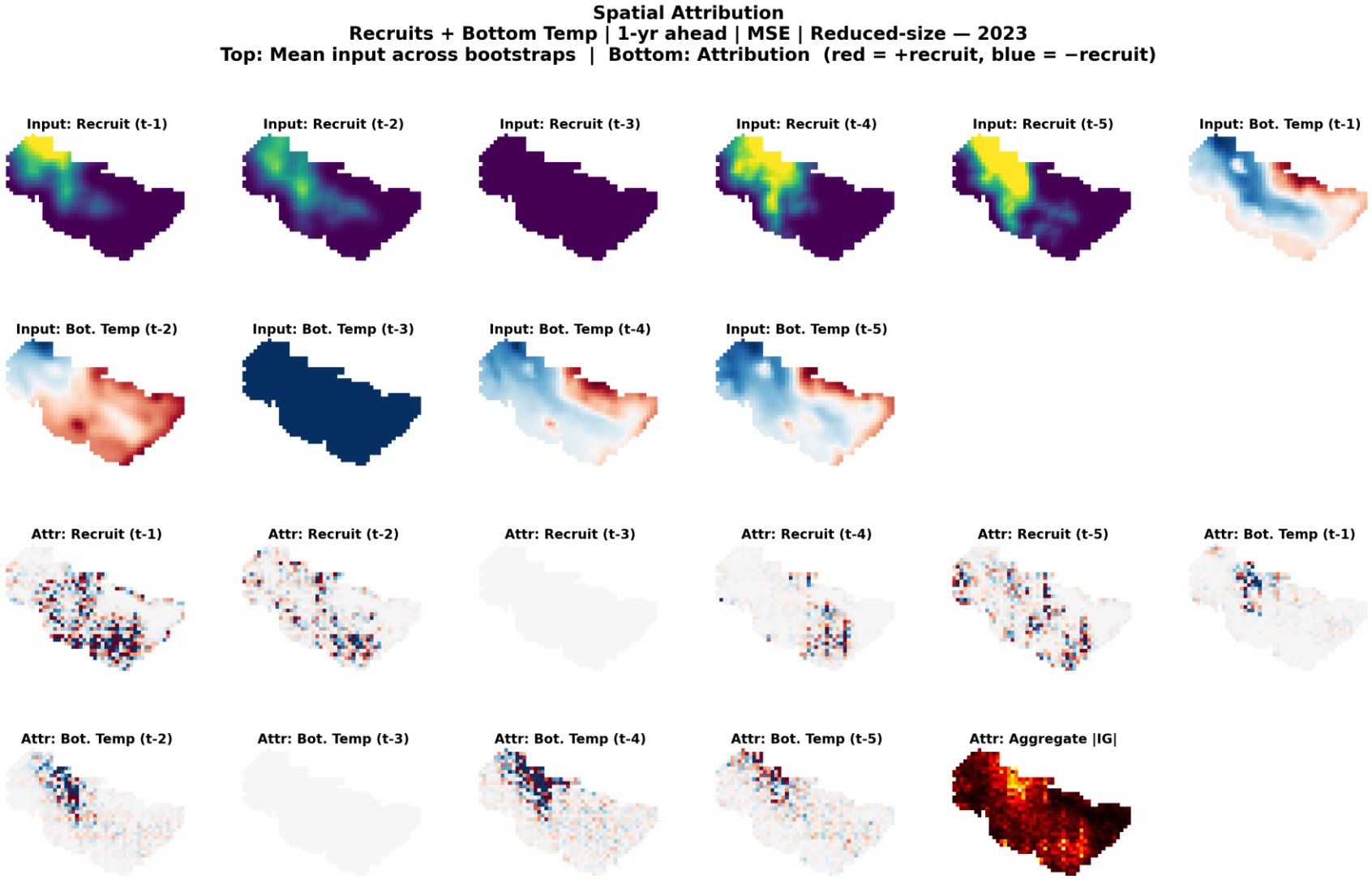


Figure 2.11. Integrated Gradients spatial attribution maps for the reduced-size, recruit and bottom temperature channels, one-year-ahead, MSE model applied to the 2023 test year. **Upper panel:** input channel grids used by the model, showing recruit history at *t*-1 through *t*-5 (channels 0-4), and bottom temperature at *t*-1 through *t*-5 (channels 5-9). Recruit inputs are shown on a yellow-to-purple scale (low to high density); temperature inputs are shown on a blue-to-red scale (cool

to warm). **Lower panel:** Integrated Gradients attribution maps for each corresponding input channel. Red indicates positive attribution (input values at that location increase predicted total recruitment relative to the baseline); blue indicates negative attribution (input values decrease predicted recruitment). Strong red attribution in a region indicates that above-average values there (relative to the baseline) pushed predicted recruitment upward; strong blue indicates the opposite. Channels with faint or near-zero attribution contributed little to the prediction for this year regardless of their input values. The aggregate map (bottom right) summarizes where spatial signal was concentrated across all channels combined. **Bottom right:** aggregate attribution map showing the sum of absolute attribution across all 10 channels, where darker values indicate grid cells that most strongly influenced the model's prediction regardless of direction. For the 2023 test year, the *t*-3 time step corresponds to the cancelled 2020 survey; channels for that year were set to zero, and the corresponding input and attribution panels are blank. Values shown are averaged across 100 bootstrap replicates for the 2023 test year.

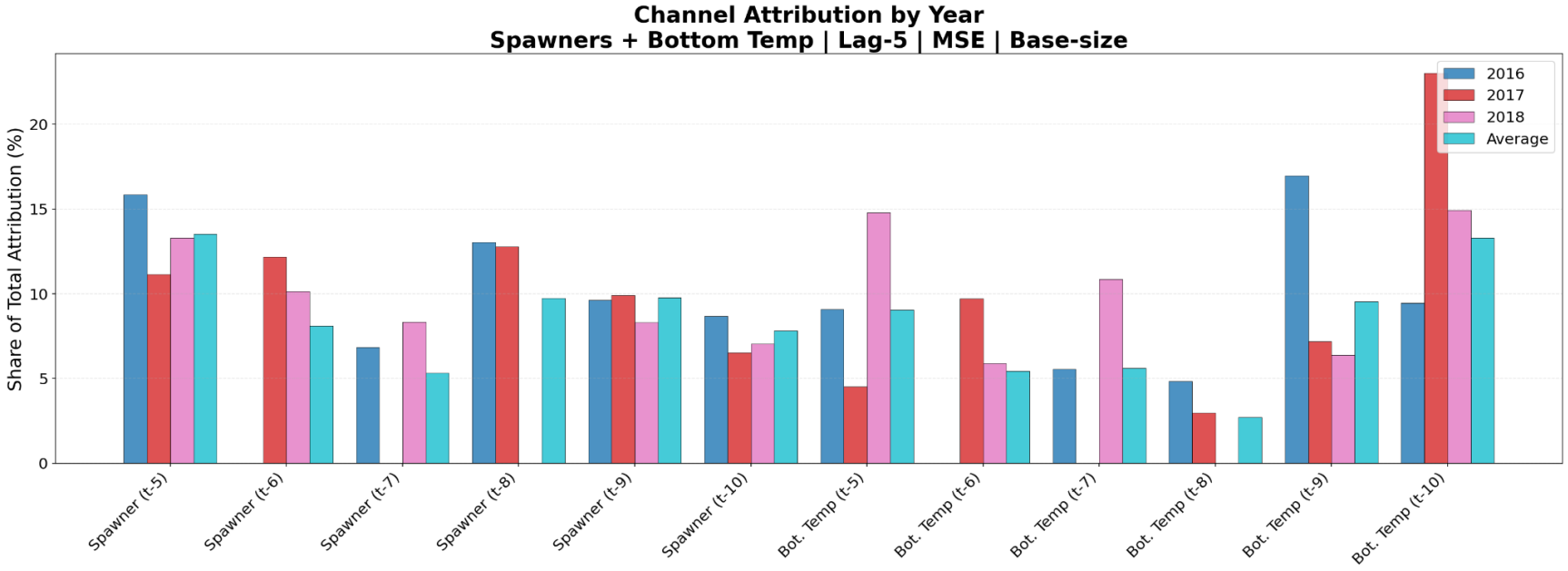


Figure 2.12. Channel-level attribution summary for the base-size, spawner and bottom temperature channels, 5-year lag, MSE model. Bar charts show the proportion of total absolute attribution assigned to each input channel for each of the three test years (2021, 2022, 2023) and the mean across test years. Higher values indicate greater influence of that channel on the model's predictions relative to the average baseline year. The 2020 survey was cancelled due to the COVID-

19 pandemic, resulting in no observations for that year; channels dependent on 2020 survey data were set to zero, and the corresponding bars are absent from the attribution plots.

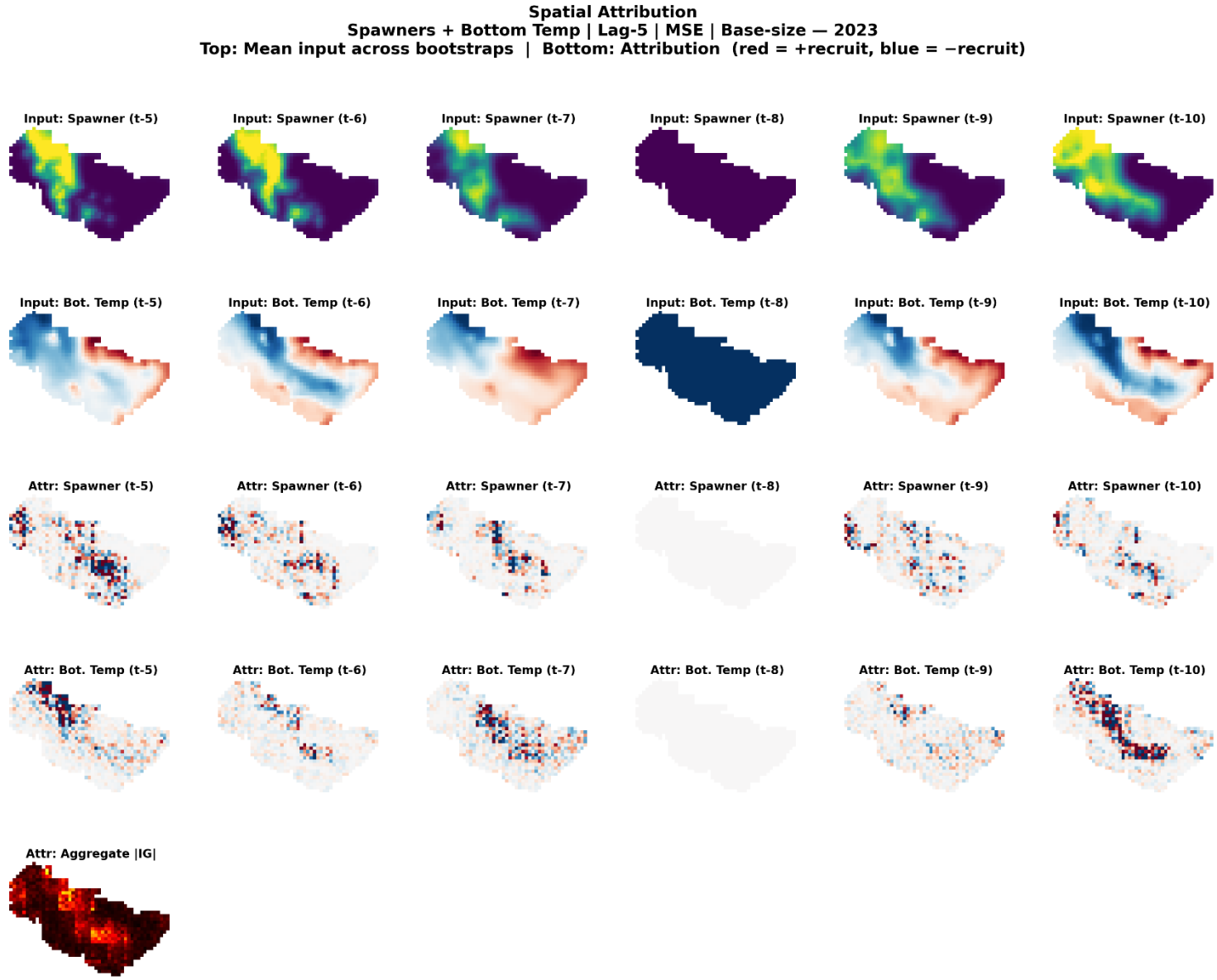

Spatial Attribution
Spawners + Bottom Temp | Lag-5 | MSE | Base-size — 2023
Top: Mean input across bootstraps | Bottom: Attribution (red = +recruit, blue = −recruit)
Input: Spawner (t-5)
Input: Spawner (t-6)
Input: Spawner (t-7)
Input: Spawner (t-8)
Input: Spawner (t-9)
Input: Spawner (t-10)
Input: Bot. Temp (t-5)
Input: Bot. Temp (t-6)
Input: Bot. Temp (t-7)
Input: Bot. Temp (t-8)
Input: Bot. Temp (t-9)
Input: Bot. Temp (t-10)
Attr: Spawner (t-5)
Attr: Spawner (t-6)
Attr: Spawner (t-7)
Attr: Spawner (t-8)
Attr: Spawner (t-9)
Attr: Spawner (t-10)
Attr: Bot. Temp (t-5)
Attr: Bot. Temp (t-6)
Attr: Bot. Temp (t-7)
Attr: Bot. Temp (t-8)
Attr: Bot. Temp (t-9)
Attr: Bot. Temp (t-10)
Attr: Aggregate |IG|

Figure 2.13. Integrated Gradients spatial attribution maps for the base-size, spawner and bottom temperature channels, 5-year lag, MSE model applied to the 2023 test year. **Upper panel:** input channel grids used by the model, showing spawner density at times *t*-5 through *t*-10 (channels 0-5), and bottom temperature at *t*-5 through *t*-10 (channels 6-10). Spawner inputs are shown on a yellow-to-purple scale (low to high density); temperature inputs are shown on a blue-to-red scale (cool to warm). **Lower panel:** Integrated Gradients attribution maps for each corresponding input channel. Red indicates positive attribution (input values at that location increase predicted total recruitment relative to the baseline); blue indicates negative attribution (input values decrease predicted recruitment). Strong red attribution in a region indicates that above-average values there (relative to the baseline) pushed predicted recruitment upward; strong blue indicates the opposite. Channels with faint or near-zero attribution contributed little to the prediction for this year regardless of their input values. The aggregate map (bottom right) summarizes where spatial signal was concentrated across all channels combined. **Bottom right:** aggregate attribution map showing the sum of absolute attribution across all 12 channels, where darker values indicate grid cells that most strongly influenced the model's prediction regardless of direction. For the 2023 test year, the *t*-3 time step corresponds to the cancelled 2020 survey; channels for that year were set to zero, and the corresponding input and attribution panels are blank. Values shown are averaged across 100 bootstrap replicates for the 2023 test year.

2.3.3 Training dynamics and model size

Given the fundamental data constraints of the survey time series (~24 independent spatial configurations), tracking the divergence between training and validation loss was critical for evaluating model generalization. All MSE-trained configurations exhibited substantial overfitting on the actual survey data: training loss decreased steadily across all 20 epochs, reaching final values of 0.11-0.17, while validation loss typically plateaued as early as epoch 4 and no later than epoch 16 (Table 2.6). The ratio between final validation and training loss ranged from 13x to 27x across configurations, confirming that the ViT model had far more capacity than the ~24 independent spatial patterns could support.

The base-size model (~1.3M parameters) consistently achieved lower training loss than the reduced-size model (~500K parameters), reaching a mean final training loss of 0.12 compared to 0.15, but this additional learning capacity did not improve generalization. Base-size configurations showed mean overfit ratios of 21.1x (range 16.6-26.6x), compared to 17.5x (range 12.8-21.3x) for reduced-size configurations (Table 2.6b).

Overfitting was most pronounced for the 5-year lag prediction mode, which represents the most data-constrained scenario (21 training years rather than 24). The base-size model with spawner and temperature channels reached an overfit ratio of 26.6x, the highest of any configuration, with best validation occurring at epoch 7. The base-size model with all channels had a qualitatively different trajectory: validation loss continued to decrease until epoch 13, suggesting that this particular channel combination required more training to learn the five-year lagged relationship, and benefited from the additional model capacity.

The post-hoc checkpoint selection procedure (Section 2.2.8) was essential for extracting usable predictions from these overfit models. Best validation epochs ranged from 1 to 16 across configurations for MSE-trained models on the actual data, with base-size models peaking later on average (epoch 8.1) than reduced-size models (epoch 5.4), and in most cases the selected checkpoint substantially outperformed the epoch-20 model on test data (Table 2.6b). For ecological datasets with effective sample sizes on the order of 20-30 independent observations, even a reduced Vision Transformer with ~500K parameters has far more capacity than the data can constrain, and both model sizes require validation-based checkpoint selection to prevent overfitting.

The divergence between training and validation trajectories is illustrated for four representative configurations in Figure 2.14. The top row shows that both model sizes on actual data exhibit the same fundamental pattern: validation loss bottoms out within the first several epochs and climbs or plateaus while training loss continues to fall. The bottom row of Figure 2.14 shows that neither synthetic configuration exhibited this behavior to the same degree, with training and validation losses tracking closely through epoch 20 regardless of model size.

Data	Criterion	N	Final Train Loss	Final Validation Loss	Best Validation Loss	Best Epoch	Overfit Ratio
Actual	MSE	40	0.11-0.17 (mean 0.14)	2.06-3.11 (mean 2.63)	1.73-2.89 (mean 2.31)	1-16 (mean 6.7)	12.8x-26.6x (mean 19.3x)
Actual	Tweedie	40	1428-1553 (mean 1490)	1883-2158 (mean 2006)	1855-2060 (mean 1938)	1-16 (mean 5.3)	1.22x-1.51x (mean 1.35x)
Synthetic	MSE	30	4.00-15.06 (mean 8.41)	4.46-20.52 (mean 9.18)	4.42-15.64 (mean 8.55)	5-15 (mean 10.4)	0.78x-1.58x (mean 1.06x)
Synthetic	Tweedie	30	3563-4364 (mean 3859)	3547-4385 (mean 3808)	3535-4322 (mean 3790)	7-20 (mean 13.9)	0.92x-1.03x (mean 0.99x)

Table 2.6. Training dynamics summary for all models on actual EBS survey data (n = 40 runs) and synthetic benchmark data (n = 30 runs). Overfit ratio is the ratio of final validation loss to final training loss. Best epoch is the epoch at which validation loss was minimized. Values shown as mean (range) across configurations within each grouping.

Data	Criterion	Model Size	Mean	Overfit Ratio Range	Mean	Best Epoch Range
Actual	MSE	40	21.1x	16.6x-26.6x	8.1	3-16
Actual	MSE	40	17.5x	12.8x-21.3x	5.4	1-12

Table 2.6b. Overfitting by model size for MSE-trained configurations on actual EBS survey data.

Clear bounds on model capacity emerged when segmenting the synthetic benchmarks by simulation difficulty. Reduced-size models consistently avoided overfitting across all difficulty tiers, maintaining final validation-to-training loss ratios between 0.78x and 1.09x, and in many cases training loss remained above validation loss for the entire 20-epoch run. The best validation epochs for the synthetic data ranged from 7 to 20, with several reduced-model configurations still improving at epoch 20, suggesting that longer training could have marginally improved performance. The contrast between the two data types is stark: the same architecture that showed negligible overfitting on 3,000 synthetic samples generated from a stationary process showed 13x-27x overfit ratios on the real survey data, driven entirely by real-world data scarcity rather than architectural excess.

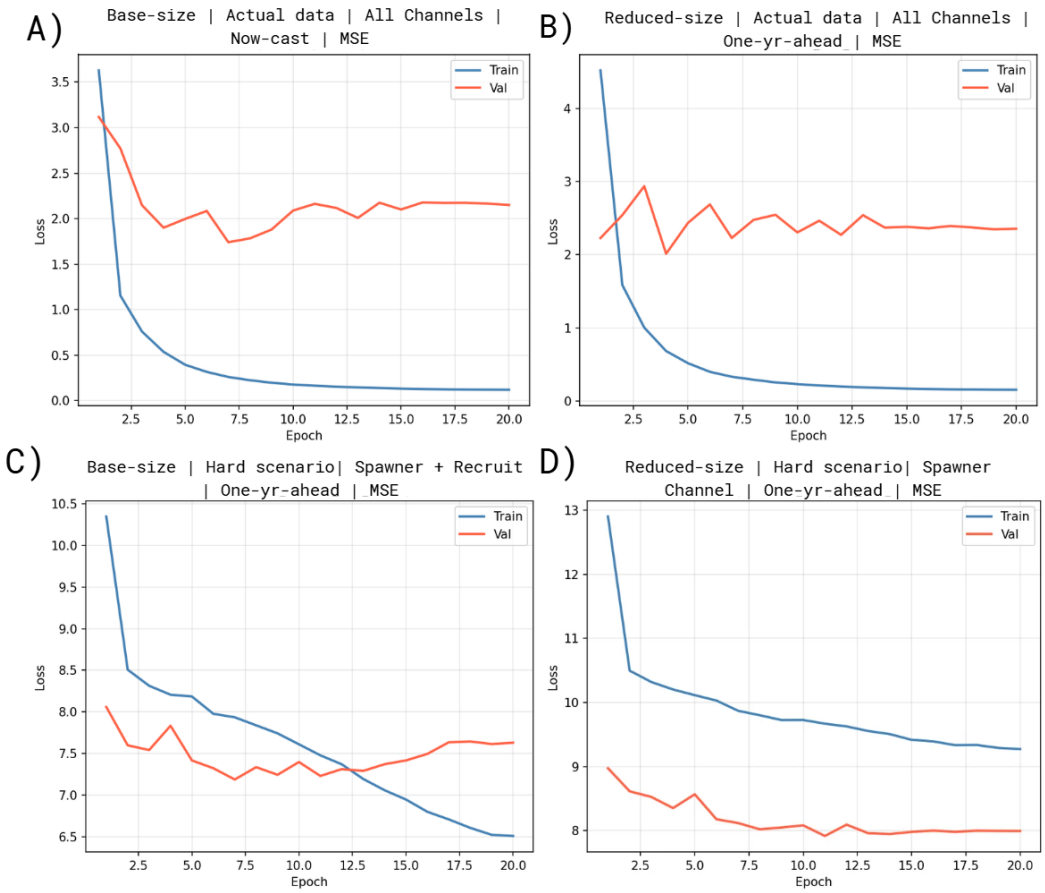


Figure 2.14. Training (blue) and validation (red) loss curves for four selected model configurations, contrasting actual EBS survey data (top row) and synthetic benchmark data (bottom row). All panels show MSE loss. (A) Actual survey data: base-size model, all channels, now-cast mode. Validation loss reached its minimum at epoch 7 and diverged steadily thereafter, yielding a final validation-to-training loss ratio of 18.0x. (B) Actual survey data: reduced-size model, all channels, one-year-ahead mode. Validation loss was minimized at epoch 4, resulting in an overfit ratio of 15.3x. Together, panels A and B show that neither model size resolved the overfitting problem on the actual data; the reduced model reached its best checkpoint sooner but ultimately overfit to a comparable degree. (C) Synthetic benchmark data (hard scenario): base-size model, spawner and recruit channels, one-year-ahead mode. Training and validation losses tracked closely throughout the 20-epoch run, with a final ratio of 1.17x, indicating negligible overfitting. (D) Synthetic benchmark data (hard scenario): reduced-size model, spawner channels only, one-year-ahead mode. Training loss remained above validation loss for the entire run (final ratio 0.86x), with no sign of overfitting through epoch 20. The contrast between rows illustrates the effect of effective sample size on model generalization: 3,000 synthetic training samples from a stationary generative process were sufficient to support both architectures through a full training run, whereas the approximately 24 independent spatial configurations in the real survey data could not constrain either model beyond the first few epochs. Note that y-axis scales are not comparable between panels, as MSE loss magnitudes differ between the real and synthetic datasets.

2.3.4 Spatial Attribution via Integrated Gradients

Integrated Gradients (IG) attribution (Sundararajan et al., 2017) was applied to the top-performing model configuration for each of the three prediction modes, and the bottom temperature-only model to understand which input channels and spatial regions most strongly influenced recruitment

predictions. Attribution was computed independently for each bootstrap replicate and test year, then averaged across replicates to produce mean attribution maps for each year. Results are presented as channel-level attribution summaries (share of total absolute attribution assigned to each of the 17 input channels; Figures 2.6, 2.8, 2.10, 2.12) and spatially resolved attribution maps for each channel (Figures 2.7, 2.9, 2.11, 2.13), where red in the attribution map figures indicates positive attribution (input values at that location push total predicted recruitment upward) and blue indicates negative attribution.

2.3.4.1 Now-cast mode

2.3.4.1.1 All channel model

The model's predictions were overwhelmingly driven by current-year spawner density in the now-cast mode base-size model using all channels and MSE loss. Spawner(*t*) accounted for approximately 25% of total attribution on average across the three test years, with spawner(*t*-1) contributing a further 15-18% (Figure 2.6). Attribution declined steeply for older spawner channels, with channels *t*-2 through *t*-5 each contributing 4-7%. Recruit history channels contributed modestly and approximately equally (4-8% each), while temperature channels were minor contributors distributed relatively evenly across lags (3-7% each).

The spatial attribution maps (Figure 2.7) for the base-size model with all channels revealed that the spawner channel attributions were concentrated in the central and southern shelf, broadly overlapping with the regions of highest variation in observed spawner density. The recruitment channel attribution showed that the lack of recruits in the southern shelf was most central to the model. Recruit(*t*-1) showed a mixture of positive and negative attributions, while recruit(*t*-5) displayed a notable cluster of negative attribution in the southern shelf region, suggesting that the model used longer-lag recruit history as a corrective or inhibitory signal rather than a straightforward positive predictor. The temperature channel attributions were spatially structured in a pattern consistent with the cold pool boundary. The aggregate attribution map, summing absolute attribution across all channels, identified a clear hotspot in the central shelf region consistent with known historical high-density, but now quite variable snow crab habitat.

2.3.4.1.2 Bottom temperature channel only

Attribution across the six bottom temperature channels was distributed relatively evenly, with temperature at *t*-4 accounting for the largest share (approximately 25%) and *t*-5 contributing roughly 20% (Figure 2.8). The dominance of older temperature lags over the current-year channel suggests the model was not simply tracking contemporary thermal conditions but was instead responding to thermal states that preceded the observed recruitment by several years, roughly corresponding to the period when the observed recruits would have been early juveniles. Spatially, attribution broadly followed the extent of the cold pool, with the greatest concentration in the northern shelf (Figure 2.9). The *t*-4 and *t*-5 channels showed the widest spatial footprint, covering a larger portion of the survey domain than recent-year channels.

2.3.4.2 One-year-ahead mode

Attribution was distributed relatively evenly across all input channels in the one-year-ahead configuration using recruit and bottom temperature channels, with recruit(*t*-5) carrying the largest share at approximately 18% and bottom temperature at year *t*-3 contributing 14% (Figure 2.10). The absence of current-year channels shifted the model's reliance toward the historical record.

This redistribution is notable because it demonstrates that the model did not simply rely on current-year spawner distributions as a spatial template for recruitment, instead, the model

compensated by drawing more heavily on recruit history, and historical temperature fields when current-year information was removed. The elevated importance of recruit(*t*-5) in this model is particularly striking because individuals observed as recruits at t-5 would reach sexual maturity and enter the spawning population at approximately time t, suggesting the model may have learned to use early abundance distributions as a proxy for the parental generation responsible for the current recruit cohort. If cohorts tend to occupy similar spatial distributions at equivalent life history stages, the spatial footprint of the previous cohort five years prior would serve as an informative prior for the current cohort's distribution, a form of spatial habitat fidelity the model appears to have recovered from the data without explicit supervision.

Spatially, attribution was concentrated in the central and southern shelf, the region of greatest observed variability in recruit density (Figure 2.11). This concentration was more pronounced than in the now-cast configurations, with the aggregate attribution map showing a particularly strong hotspot at the southern extent of the snow crab's range.

2.3.4.3 Five-year lagged mode

All channel labels below are expressed relative to the recruit year because the 5-year lag alignment pairs spawner observations at time *t* with recruit observations at time *t*+5. Temperature at year *t*-10 therefore corresponds to thermal conditions recorded ten years before the observed recruitment event, or five years before the paired spawning observation. Under the 5-year lag prediction scenario (base-size model, spawner and temperature channels), the temporal structure of attribution was concentrated in the spawner *t*-5 (14%), bottom temperature *t*-10 (14%), with spawner *t*-9, *t*-10, and bottom temperature t-9 around 8-10% (Figure 2.12). This shift toward historical channels is consistent with the one-year-ahead configuration, which similarly lacked access to same-year observations.

The most striking result in this model was the emergence of bottom temperature as a dominant predictor. Bottom temperature (*t*-10) was the single most heavily attributed channel on average at approximately 13% of total attribution, and for individual test years its share exceeded 20%. Spatially, attribution in the temperature channels showed pronounced alignment with the cold pool boundary, more so than in either of the shorter-horizon modes (Figure 2.13). Spawner channel attribution was concentrated along the shelf break, consistent with regions of historically high but variable spawner density.

2.4 Discussion

2.4.1 Data constraints in deep learning

The ViT architecture achieved meaningful spatial predictions of snow crab recruitment despite training on only approximately 24 independent spatial configurations, a sample size substantially below what vision transformer architectures typically require. Two design choices helped mitigate this constraint: data augmentation, and reduced model architectures. Bootstrap resampling increased the nominal training set to 2,400 samples, and while the high inter-replicate correlation ($r > 0.98$) means the effective number of independent patterns remained near 24, the interpolation-driven variation across replicates provided sufficient gradient signal to support optimization. The reduced model architecture (~500K parameters) consistently matched or outperformed the base model (~1.3M parameters) on the most challenging prediction tasks, confirming that aggressive capacity reduction is appropriate when effective sample size is small.

The contrast between actual and synthetic training dynamics quantifies this constraint precisely. Even the base-size model showed negligible overfitting on synthetic data with 3,000 samples from a stationary generative process, with validation-to-training loss ratios of 0.99 to 1.03.

In contrast, the same architecture led to ratios of 17 to 27 for the actual data. This gap is not attributable to architectural excess alone; it reflects the fundamental mismatch between model capacity and the number of independent spatial patterns available in a 36-year survey time series. A further asymmetry is that the synthetic data contain no observation error. Fields were sampled directly from the GMRF generative process, whereas real data carry trawl sampling variability propagated through SPDE interpolation. The effective signal-to-noise ratio in the real training data is therefore lower than the synthetic difficulty levels suggest, compounding the effect of data scarcity on generalization.

The post-hoc checkpoint selection procedure was essential for recovering usable predictions from these overfit models, with best validation epochs ranging from 1 to 16 across configurations. Year-level variation in test performance was substantial, with MAE SD across bootstrap replicates two to five times wider in 2022 than in 2021 or 2023 for nearly every configuration. This is consistent with the interpretation that checkpoint selection noise is compounded by genuine inter-annual unpredictability rather than model instability alone. Ecologists applying deep learning to similar monitoring datasets should expect this behavior and treat aggressive regularization, reduced architectures, and validation-based model selection as standard practice rather than optional refinements.

Future work using synthetic datasets with varying time series lengths could help establish how much data are needed to support different model capacities, providing practical guidance for applying this architecture to other monitoring programs with similarly constrained records.

2.4.2 Model configuration sensitivity

The sensitivity analyses revealed that model size and channel configuration had larger effects on performance than hyperparameter choices within the ranges tested. From an architectural standpoint, both the base and reduced model sizes produced competitive results, and in the case when access to compute power is limited, a reduced-size architecture should be explored as a strong alternative. A range of hyperparameter choices including learning rate schedulers, normalization strategies, regularization levels, and dropout rates were examined. The configurations evaluated here represent a practical but not exhaustive search of this space, and future work could identify combinations that better balance overfitting and predictive accuracy, particularly for the lag-5 prediction mode where training data were most limited.

A related finding concerns the low attribution assigned to the oldest spawner channel even in now-cast mode (Figure 2.6). Given that the recruit size class (45-55 mm carapace width) is assumed to be approximately five years old, one might expect spawner distribution five years prior to be the strongest predictor of current recruitment location. Instead, the model consistently assigned highest attribution to Spawner(t) and Spawner(t-1), with steep declines for older lags. Several factors likely contribute to this pattern.

First, there is substantial overlap between the spawner and recruit size classes within the same survey year. On average, 55.9% of mature females fell within the 45-55 mm recruit size class, meaning a majority of recruits were simultaneously represented in the spawner index. This means the current-year spawner channel already contains spatial information about where recruits are, making the now-cast attribution pattern partly a reflection of this co-occurrence rather than a causal spawning-to-recruitment pathway. Second, because the bottom trawl captures all size classes at a station simultaneously, any shared habitat preferences or seasonal aggregation patterns between size classes would inflate the apparent spatial association between same-year spawners and recruits. Third, the five-year lag is a population-level approximation; individual growth rates vary

with temperature and food availability, so no single historical spawner year maps cleanly onto the current recruit cohort (Sainte-Marie 1996).

This overlap motivates the one-year-ahead and lag-5 configurations as the more operationally meaningful prediction modes, since neither has access to current-year survey data and therefore avoids this source of ambiguity.

2.4.3 Model performance

Spearman's correlation and MAE were used as complementary metrics because they capture distinct dimensions of predictive skill. Spearman's correlation reflects the model's ability to correctly rank grid cells by relative recruitment intensity, while MAE measures absolute accuracy in density units relevant to total abundance estimation. A model that correctly identifies spatial hotspots but at the wrong magnitude, or one that estimates total abundance accurately but misplaces it geographically, would score well on one metric while failing the other. In practice, the two metrics address different management questions. Performance in terms of Spearman's correlation is most relevant when the goal is spatial: identifying survey locations with elevated recruitment potential, prioritizing areas for protection, or allocating survey or fishing effort ahead of the season. MAE performance matters most when model outputs feed directly into abundance estimation or harvest control rules, where errors in predicted density propagate into stock assessment calculations.

Across all three prediction modes, the model relied predominantly on same-year spawner distributions when current-year data were available, but compensated when these same-year inputs were withheld by upweighting recent recruit and spawner history, demonstrating that it had learned genuine temporal dynamics rather than simply copying the current-year spawner field. Temperature history at lags of four to five years emerged as the dominant predictor when predicting across the full five-year ontogenetic lag, consistent with the cold pool's role in structuring juvenile habitat during the years between spawning and survey selectivity. These attribution patterns are discussed in detail by prediction mode in the subsections below.

Two climatological baselines were constructed to provide mode-specific lower bounds for interpreting model performance: one for the now-cast and one-year-ahead modes (which share the same training data) and one for the 5-year lag mode (which trains on fewer years). The now-cast/one-year-ahead baseline achieved a spatial Spearman's correlation of 0.807 and an MAE of 500 crabs km^{-2}, while the 5-year lag baseline achieved a Spearman's correlation of 0.812 and a lower MAE of 455 crabs km^{-2}, the latter reflecting the reduced density range in the shorter training window. Both baselines confirm that stable spatial structure in SPDE-smoothed survey data accounts for the majority of rank correlation. Model performance on the actual data should therefore be interpreted primarily through MAE, where the best-performing now-cast configurations achieved a 34-38% reduction relative to their baseline, while the best 5-year lag configurations achieved a more modest 9-14% reduction, consistent with the fundamental difficulty of predicting recruitment five years in advance. Across all prediction modes, 2022 was consistently the most difficult test year: MAE was elevated and bootstrap SD was markedly wider than in 2021 or 2023 (Table 2.5), reflecting the sharp population fluctuation that year rather than any systematic model failure. The 2021 and 2023 test years showed tighter bootstrap distributions and lower absolute errors, suggesting the models were better calibrated to population states within the range of the training record.

The synthetic benchmarks provide an important perspective: the climatological baseline Spearman's correlation was near zero (0.12-0.20 across difficulty scenarios), and the model achieved Spearman's correlations of 0.70-0.80, demonstrating that the ViT architecture can extract

genuine predictive signal when the spatial mean is not itself informative. Taken together, these results suggest that the relatively modest improvements in Spearman's correlation over the baseline for the actual data reflect a property of the observational data rather than a ceiling on the model's capacity to learn spawner-recruit relationships. Therefore, the model has the ability to predict spatiotemporal teleconnections in theory, even if the snow crab dataset is not the best test case to showcase it. Other applications of this model architecture should be explored.

2.4.3.1 Now-cast top model

The top now-cast configuration: the base-size model using all 17 channels with MSE loss, achieved the highest composite score across all actual-data configurations and serves as a useful performance ceiling for the other prediction modes. Attribution was dominated by the current-year spawner channel, which is consistent with the size-class overlap described above. The now-cast mode has limited operational utility in the EBS snow crab system where spawners and recruits are surveyed simultaneously, but it could be more directly applicable in systems where spawner and recruit observations are collected at different times of year or through different survey programs (e.g., Pacific Halibut (*Hippoglossus stenolepis*) in the Gulf of Alaska). Performance was stable across 2021 and 2023, with MAE of 301 and 245 crabs km^{-2} respectively, but degraded in 2022 (MAE = 449 crabs km^{-2}), a year in which the population was undergoing change. The Spearman's correlation held up well even in 2022 (0.864), confirming that spatial rankings remained accurate even when absolute density estimates did not.

The pattern of negative attribution from distant recruit history, particularly in the southern shelf, may reflect density-dependent compensation: high juvenile abundance five years prior predicts lower current recruitment in the same region, consistent with intraspecific competition documented in snow crab populations (Szuwalski et al. 2026).

2.4.3.2 Bottom temperature only model

The third-best configuration overall was the temperature-only now-cast model, a reduced-size model trained with MSE loss that relies entirely on historical bottom temperature fields. For a modest reduction in composite score relative to the all-channel model, this configuration produces spatial recruitment predictions without any direct observation of crab density. This could be useful in years where trawl survey data are delayed, incomplete, or unavailable, although it is important to recognize that none of the configurations, including this one, performed well at capturing the absolute magnitude of population changes over time. Spatial rankings were reliable, and MAE was still 25% better than the climatological baseline.

The most heavily attributed temperature channel was for year *t*-4, followed by year *t*-5. These lags correspond roughly to the period when the observed recruits would have been early juveniles, suggesting the model was responding to thermal conditions during a critical early life history window rather than tracking contemporary temperatures. Attribution was spatially concentrated near the cold pool boundary, with years *t*-4 and *t*-5 showing the widest spatial footprint across the survey domain. For contexts where bottom temperature data are more readily available than trawl observations, a temperature-driven spatial prediction of this kind could serve as a practical interim tool. Future work with higher-frequency or spatially denser temperature records could extend this approach toward more continuous habitat forecasting. One caution is that the temperature-only configuration showed the most erratic year-level MAE of any top configuration: 66, 491, and 570 crabs km^{-2} in 2021, 2022, and 2023 respectively (Table 2.5). The 2021 result is strikingly low, but the 2022 and 2023 errors are the largest of any top configuration in those years, suggesting the

model's spatial rankings are reliable but its implied density magnitudes can drift substantially when temperature patterns alone are asked to carry the full predictive load.

2.4.3.3 One-year-ahead top model

The best one-year-ahead configuration used recruit history and bottom temperature channels with MSE loss in a reduced-size model. This is arguably the most operationally relevant prediction mode: it produces a spatial forecast of recruitment before the current year's survey is conducted, when management decisions often need to be made. Composite score declined only modestly relative to the now-cast (0.932 vs 0.973), and Spearman's correlation was comparable across most years (0.894, 0.857, and 0.854 in 2021, 2022, and 2023 versus 0.864, 0.864, and 0.872 for the top now-cast). The 2022 Spearman's correlation of 0.857 is the only year where the one-year-ahead configuration fell meaningfully below its now-cast counterpart, and it coincides with the same population discontinuity that elevated MAE across all modes in that year. That this gap appears in 2022 specifically, rather than uniformly, suggests the performance cost of withholding current-year data is not a fixed architectural penalty but depends on how much the population has departed from the historical spatial template.

The one-year-ahead configuration did not use spawner channels, relying instead on recruit history and temperature. The elevated attribution assigned to recruit(t-5) is notable, as this cohort represents the approximate parental generation of the current recruit class: individuals that were themselves recruits five years prior would be reaching reproductive maturity and contributing to the spawning stock that produced the observed cohort. The spatial distribution of this parental cohort may therefore carry information about where the current cohort is likely to concentrate. The practical implication is that spatial predictions of recruitment hotspots could be produced before each survey season using only historical recruit observations and temperature records, potentially informing the allocation of survey effort or the design of spatial management measures.

2.4.3.4 5-year lag top model

The lag-5 configuration pairs spawner and temperature observations at year *t*-5 with recruit observations at year *t*, reflecting the approximate lag between when they were spawned and when they would reliably be seen in the survey. The best-performing configuration used spawner and temperature channels in a base-size model with MSE loss. Predictive skill was lower than for the other modes, which is expected given the longer forecasting horizon and the reduction in usable training years from 24 to 21. Despite this, the model retained a Spearman's correlation of 0.832 on held-out test data, indicating meaningful spatial skill at a five-year horizon. Year-level results showed 2022 had the weakest Spearman (0.808) while 2021 had the highest MAE (581 crabs km^{-2}).

The most unexpected result in this mode was the emergence of bottom temperature as a dominant predictor at long lags. Bottom temperature at t-10 was the single most heavily attributed channel on average, with individual test years exceeding 20% of total attribution. These channels correspond to thermal conditions 9 to 10 years before the observed recruitment event, or 4 to 5 years before the paired spawning observation. Attribution across spawner channels was distributed more evenly over time than in shorter-horizon modes, with no single year dominant, suggesting the model learned that spawner distributions across multiple preceding years collectively carry predictive information about recruitment five years later.

2.4.4 Thermal teleconnections and the cold pool

Bottom temperature emerged as an important predictor across all prediction modes, and the spatial structure of temperature attribution was consistent across configurations. Attribution maps for temperature channels broadly followed the extent of the cold pool, with the greatest concentration along the cold pool boundary on the central and northern shelf. This spatial weighting is consistent with the well-documented role of the cold pool as a thermal barrier that determines the availability of suitable benthic habitat for juvenile snow crab (Fedewa et al. 2020). Suitable habitat expands when the cold pool extends further south and juvenile habitat becomes restricted to the northern shelf when the cold pool contracts. The ViT model appears to have internalized this relationship from the data, using temperature fields at biologically relevant lags to determine where spawning effort is likely to translate into successful recruitment.

The finding that temperature alone achieved a Spearman's correlation of 0.869, the highest of any now-cast configuration, further underscores the central role of thermal structure in shaping the spatial distribution of recruitment. Bottom temperature may serve as a spatial proxy for recruitment potential that is partly independent of spawner distribution because the cold pool boundary delineates suitable juvenile habitat. The higher MAE of the temperature-only configuration relative to the all-channel model indicates that while temperature reliably ranks locations by relative recruitment intensity, spawner information is needed to calibrate absolute density estimates.

Spatially, attribution in the lag-5 temperature channels showed pronounced alignment with the cold pool boundary, more so than in either of the shorter-horizon modes (Figure 2.13). Spawner channel attribution was concentrated along the shelf break, consistent with regions of historically high but variable spawner density. This spatial structure provides direct empirical support, recovered from data via the self-attention mechanism without any prior specification, for the hypothesis that cold pool extent governs juvenile habitat suitability and mediates the spatial relationship between spawning events and eventual recruitment.

The strong attribution to temperature at t-10 is harder to explain through direct habitat effects alone. One interpretation is that these lags correspond to the early life history of the parental generation. The EBS snow crab spawning population has a significant proportion of age-5 individuals, meaning many of the spawners observed at time t-5 were potentially themselves larvae approximately five years prior. Strong attribution to temperature at t-10 may therefore reflect the influence of early-life thermal conditions on the parental cohort's survival and subsequent abundance as spawners at time t-5, which in turn determines recruit abundance at t. Favorable bottom temperatures during the larval stage of the parental cohort would increase survival, producing a larger and potentially more reproductively productive cohort of spawners and ultimately more recruits five years later. However, if this signal operated strictly through spawner abundance at t-5, we would expect the model to weight spawner density at t-5 more heavily than it does. Whether the t-10 temperature signal reflects an independent thermal effect on parental cohort quality, temporal autocorrelation in cold pool extent, or a statistical artifact of the lag structure cannot be resolved with the current time series, but the consistency of the signal across test years suggests it warrants further investigation.

2.4.5 Limitations and assumptions

Several limitations should be considered when interpreting these results. While bootstrap resampling increased the nominal training set to 2,400 samples, the high inter-replicate correlation (r = 0.987) means the effective number of independent training patterns remains approximately 24. Generalization capacity is fundamentally bounded by the number of unique spatial configurations observed over the survey time series, not by the number of bootstrap replicates. The

test set covers only three years, meaning the reported test metrics are themselves noisy estimates of true generalization performance.

The model implicitly assumes stationarity in the spawner-recruit relationship across the training period. The EBS has experienced pronounced thermal regime shifts over the past four decades (Szuwalski & Punt, 2013), and given the tight coupling between temperature and spatial structure in this system through cold pool extent, both may be subject to non-stationarity across regimes. A model trained primarily on the cold phase may therefore not generalize well to the recent warm period represented by the test years. The early onset of overfitting in the training curves may partly reflect this. Regime-aware training strategies, such as upweighting recent observations or fitting separate models for distinct thermal periods, could address this but would further reduce the already small effective sample size.The high climatological baseline Spearman's correlation (0.807) reflects this stationarity in the spatial mean: the long-run average field is a good predictor precisely because habitat structure has been consistent across most of the training period. If regime shifts alter the spatial distribution of recruitment substantially, the baseline's advantage would erode along with that of the ViT model. This effect is not seen when we train on the synthetic data because these data do not have a long-term spatial structure.

The current lag-5 configuration aligns both spawner and temperature observations at year *t*-5 with recruits at year *t*, meaning temperature channels represent conditions concurrent with spawning rather than conditions experienced by the resulting larvae during development. A more biologically motivated alignment would pair spawner observations at year *t*-5 with temperature observations at years *t*-1 through *t*-5, capturing thermal conditions during the cohort's early life between spawning and survey recruitment. This configuration was not evaluated during the current study but represents a natural extension for future work.

The high climatological baseline Spearman's correlation reflects the dominant role of stable spatial structure in INLA-smoothed survey data. The SPDE interpolation process imposes spatial smoothness on station-level observations, making the geographic pattern of relative crab density largely consistent across years. Any prediction that correctly captures the long-run spatial mean will therefore score well on rank correlation regardless of temporal forecasting skill, making Spearman's correlation a weak discriminator between models on these data. Importantly, both the ViT model and the climatological baseline are evaluated against these same SPDE-smoothed targets, so the relative comparison between them remains valid on an equal-footing basis. However, because predictions are compared to smoothed targets, improvements in Spearman's correlation over the baseline should be interpreted as improvements in recovering year-specific smoothed spatial patterns, not as improvements in predicting raw trawl counts. This inflated baseline is a property of the smoothed observational data rather than the model or evaluation framework, as confirmed by the synthetic benchmarks where the climatological baseline Spearman's correlation was near zero. Ecologists applying this model to inform survey design or spatial management would ultimately want predictions validated against station-level observations, a step not taken here and representing a natural direction for future work.

2.4.6 Conclusion and implications for fishery management

This study demonstrates that Vision Transformers can be adapted for spatiotemporal ecological prediction under severe data constraints, and that their self-attention mechanism can generate biologically interpretable spatial relationships without prior specification. The ViT architecture here achieved Spearman's correlations of 0.83 to 0.87 across all three forecasting horizons on held-out test data, maintaining strong spatial pattern identification even when predicting five years

ahead. The learned attribution structure was consistent with known features of snow crab biology, including the importance of spawner spatial distribution, the role of the cold pool as a determinant of juvenile habitat, and the multi-year temporal structure linking spawning events to eventual recruitment.

For fisheries management, the one-year-ahead forecasting mode is the most directly applicable. It achieved spatial accuracy approximately 4 percentage points below the now-cast composite score despite withholding all current-year survey data, meaning spatial recruitment predictions could in principle be produced before each season's bottom trawl survey is completed. Such predictions could inform the allocation of survey effort toward areas of predicted high recruitment, support the design of spatial closures aimed at protecting recruitment hotspots, or contribute to spatially structured harvest strategies in a rebuilding context. The temperature-only configuration offers a related practical option for situations where trawl data are unavailable, at the cost of some reduction in absolute accuracy.

More broadly, this work provides a practical example of adapting a large vision model architecture to a small-data ecological problem. The key lessons are that model capacity should be matched to effective sample size rather than nominal sample count, that validation-based checkpoint selection is essential when overfitting is likely, and that the self-attention mechanism can identify multi-year lagged spatial relationships from observational data without requiring those relationships to be specified in advance. These lessons apply beyond snow crab to any ecological monitoring program where the time series is long enough to capture meaningful variation but too short to support unconstrained deep learning.

S.2 Supplementary Material

S.2.1 INLA SPDE interpolation details

Spatial interpolation of station-level crab density observations on to the 50 × 50 prediction grid was performed using the Stochastic Partial Differential Equation (SPDE) approach implemented in R-INLA (Lindgren et al., 2011; Rue et al., 2009). This approach approximates a continuously indexed Gaussian random field with Matérn covariance using a Gaussian Markov Random Field (GMRF) defined on a triangulated mesh, enabling computationally efficient spatial prediction using sparse matrix operations. A GMRF is a spatially correlated random field in which the precision matrix (the inverse of the covariance matrix) is sparse, encoding the property that each cell is conditionally independent of all other cells given only its immediate neighbors.

A constrained Delaunay triangulation mesh was constructed from the survey station coordinates using *fm_mesh_2d* (fmesher package, Lindgren 2026). The mesh parameters were: a minimum distance between mesh vertices of 30 km (cutoff), a maximum triangle edge length of 40 km in the interior domain and 100 km in the outer extension (max.edge), and a buffer extension of 50 km around the station locations plus an additional 150 km outer boundary (offset). The inner edge constraint ensures sufficient resolution to capture spatial variation between survey stations (which are spaced approximately 37 km apart on the 20 × 20 nm survey grid), while the outer extension and buffer prevent boundary artifacts in the spatial predictions.

A Matérn SPDE model with smoothness parameter $\alpha = 2$ (corresponding to a Matérn field with smoothness $\nu = 1$) was specified on this mesh using *inla.spde2.matern*. The following model was fit independently to the subsampled station observations for each bootstrap replicate and survey year:

$$y_i = \beta_0 + u(s_i) + \varepsilon_i$$

where $y_i = log(1 + d_i)$ is the log-transformed density at station i, β_0 is a global intercept, $u(s_i)$ is the spatial random effect at location s_i evaluated using the SPDE approximation, and ε_i is Gaussian observation error. The model was fit using the Gaussian approximation strategy with empirical Bayes integration for the hyperparameters (strategy = "gaussian", int.strategy = "eb"), which provides a fast and stable approximation suitable for the large number of model fits required (100 bootstraps × 35 years × 2 variables = 7,000 fits per dataset).

Predictions were generated at the centroids of valid grid cells using a projection matrix mapping the mesh nodes to the prediction locations. Predicted values were back-transformed to the original density scale using the inverse transformation $expm1(\hat{y}) = e^{\hat{y}} - 1$, with negative predictions and missing values set to zero. The resulting fields were placed on to the 50 × 50 padded grid using the row and column indices computed during grid construction (Section 2.1), with cells outside the valid survey footprint set to zero.

S.2.2 GMRF simulation details

A spatiotemporally autocorrelated GMRF model was used to generate synthetic benchmark data. A sparse precision matrix Q was constructed on the 50 × 50 grid with diagonal entries equal to κ^2 × (number of neighbors) and off-diagonal entries of $-\kappa^2$ for neighboring cells, where κ controls the spatial range of correlation (Lindgren et al. 2011). A spatial field was sampled by solving $\mathbf{Qx} = \mathbf{z}$ for a standard normal random vector $\mathbf{z}$. Temporal autocorrelation was introduced by blending the field at each time step with the previous year's field (coefficient ρ), preserving marginal variance. Synthetic density values were obtained by standardizing and exponentially transforming the GMRF samples, scaling to target mean densities, and applying a threshold to produce approximately 60% zeros. Spawner-recruit correlations were introduced by blending the recruit

field with an optionally lagged copy of the spawner field. Importantly, no observation error was added at any stage: each synthetic sample represents the true latent spatial process directly, with variability arising solely from the GMRF sampling procedure and AR(1) temporal evolution. This contrasts with the real EBS data, where station-level trawl observations carry sampling variability that is partially smoothed but not eliminated by the SPDE interpolation step. The synthetic benchmarks therefore represent an idealized scenario in which the signal-to-noise ratio is determined entirely by the specified spawner-recruit correlation and lag structure, without the additional noise introduced by sparse point sampling of a continuous spatial field.

Parameter	Easy	Medium	Hard
Grid size	50	50	50
Number of years	30	30	30
Number of bootstraps	100	100	100
Spatial kappa (κ)	0.3	0.3	0.3
Temporal Rho (ρ)	0.0	0.7	0.7
Recruitment Correlation	0.9	0.9	0.8
Lag (years)	0	3	5
Mean spawner density	3728	3728	3728
Mean recruit density	4514	4514	4514
Seed	2026	2026	2026

Table S.2.1. Parameters used to generate synthetic spawner and recruit density fields for the three difficulty scenarios. Spatial fields were sampled from a GMRF defined by the spatial range parameter κ, where smaller values produce longer-range spatial correlation. Temporal autocorrelation between consecutive years was controlled by the AR(1) parameter ρ. The spawner-recruit relationship was introduced by blending independent spatial fields with a correlation coefficient equal to the recruitment correlation parameter, with recruits lagged behind spawners by the specified number of years. Zero-inflation was applied by subtracting the 60^{th} percentile of the transformed density field and setting negative values to zero, then rescaling to match the target mean densities. The easy scenario represents a strong, contemporaneous spawner-recruit relationship with no temporal autocorrelation; the medium scenario introduces realistic temporal persistence and a three-year lag; the hard scenario reduces the spawner-recruit correlation and extends the lag to five years, reflecting the most biologically realistic configuration. All scenarios used identical spatial resolution, time series length, and number of bootstrap replicates to ensure comparability with the real EBS survey data.

S.2.3 ViT architecture details

S.2.3.1 Patch embedding

Rather than treating each of the 2,500 grid cells individually (which would create an intractably large sequence for attention computation), the 50 × 50 grid was divided into non-overlapping 5 × 5 patches, yielding 100 patches arranged in a 10 × 10 grid. This is analogous to grouping nearby grid cells into spatial blocks before analysis, similar to spatial aggregation in geostatistical or species distribution modelling.

Each patch contains 5 × 5 = 25 values per channel, giving a raw vector of 25 values per channel per patch. The fixed-size vector space in which all patches are represented and through which attention is computed, called the model's working dimension, is 128. This dimension is a hyperparameter that controls model capacity: all patch embeddings, regardless of how they are constructed, must be projected to this common size before entering the transformer. Each channel's

25-element patch vector is projected to an 8-dimensional per-channel embedding (computed as [128/17] = 8). The 17 per-channel embeddings are concatenated into a 136-element vector (17 × 8), which is then projected to the model's working dimension of 128. Using separate projections for each channel allows the model to learn a different representation for current spawners, spawner history, recruit history, and temperature, as these channels have fundamentally different ecological meanings. This two-stage process (project-then-combine) is analogous to constructing separate regression models for each predictor type and then combining their outputs in a meta-model.

Temporal masking is applied at this stage: for any channel flagged as missing (mask value = 0), the corresponding channel embedding is multiplied by zero before concatenation, ensuring that missing channels contribute nothing to the combined patch representation. After per-channel projection, each embedding is normalized using Layer Normalization, and dropout regularization (p = 0.2) is applied to reduce overfitting.

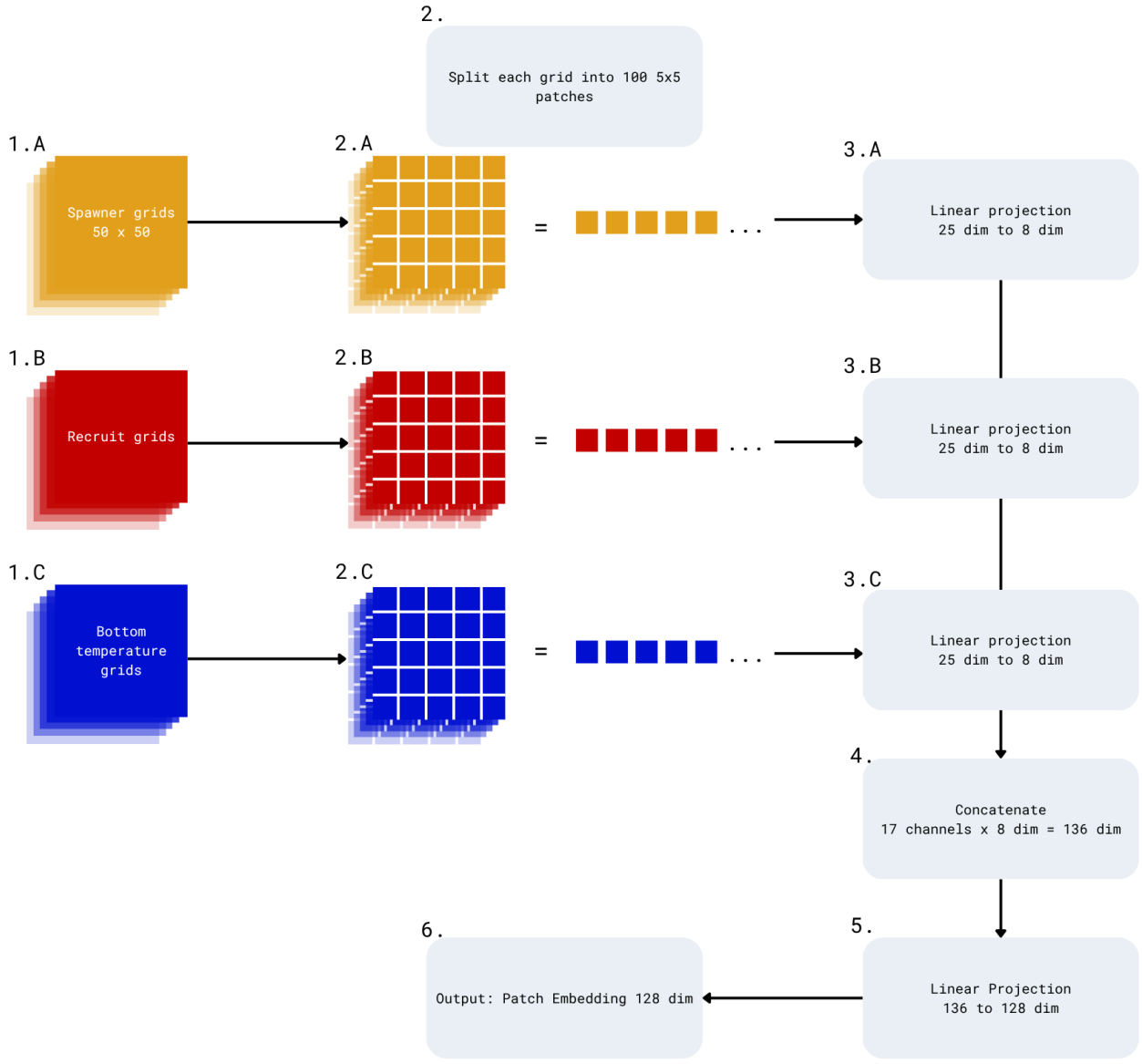


Figure S.2.1. Patch embedding procedure. (1) Input channel grids are assembled according to the model configuration. (2) Each 50 × 50 channel grid is divided into 100 non-overlapping 5 × 5 patches, yielding a 25-dimensional vector per

patch per channel. (3) Each channel's patches are projected independently to an 8-dimensional embedding via a learned linear transformation, layer normalization, and dropout. (4) The per-channel embeddings for all channels are concatenated, yielding a vector of dimension equal to the number of channels times 8 (e.g., 136 dimensions for the 17-channel configuration). (5) A final linear projection maps this concatenated representation to the model's working embedding dimension. (6) The output is a 100 × 128 matrix of patch embeddings, one 128-dimensional vector per spatial patch.

S.2.3.2 Positional and temporal encoding

After patch embedding, two types of contextual information were added to each patch's embedding vector.

- **Spatial (positional) encoding:** Each of the 100 patch positions was assigned a learned position vector of length 128, optimized jointly with the rest of the model during training. Rather than supplying fixed geographic coordinates, the model learns its own representation of spatial location directly from the data, analogous to estimating spatial random effects in a geostatistical model. These position vectors are added to the patch embeddings so that the model can distinguish patches from different grid locations when computing spatial attention weights (Dosovitskiy et al., 2021).
- **Temporal encoding:** Fixed sinusoidal encodings (Vaswani et al. 2017) were used rather than learned temporal embeddings because the time series is short (~24-35 years) and learned embeddings would risk overfitting to specific years. In addition, using a fixed embedding allows the model to generalize to years that were not seen in training, i.e., to make predictions in the future. Each prediction year was assigned a fixed sinusoidal encoding based on the year index:

$$PositionalEncoding(t, 2i) = sin\left(\frac{t}{10000^{2i/d}}\right)$$
$$PositionalEncoding(t, 2i+1) = cos\left(\frac{t}{10000^{2i/d}}\right)$$

  where t is the year index, $d = 128$ is the embedding dimension, and $i \in \{0, 1, ..., d/2 - 1\}$ indexes pairs of embedding dimensions, with sine applied to even dimensions and cosine applied to odd dimensions. The temporal encoding provides a unique temporal fingerprint for each year that encodes the relative distance between years without requiring the model to learn year-specific patterns, and allows the model to generalize to years not seen during training.

S.2.3.3 Multi-head self-attention

Self-attention can be understood as a data-adaptive weighting scheme: for each patch, it computes a weighted average of all other patches' representations, where the weights reflect how 'relevant' each other patch is to the current one. High attention weights between two patches indicate that information at one location is useful for predicting the representation at the other. All patch embeddings (a matrix of size 100 x 128) are projected through three separate learned weight matrices to produce three matrices: a query (**Q**), a key (**K**), and a value (**V**), each of length d_head = 128/num_heads (16 in the full model, 32 in the reduced model). The query represents what information a given patch is searching for from other patches; the key represents what information that patch offers; and the value contains the actual information to be transmitted. The attention weight between patches *i* and *j* is computed as the scaled dot product of their query and key vectors, normalized across all patches using the softmax function (which converts a set of raw scores into a probability distribution that sums to 1):

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \boldsymbol{softmax}(\frac{\mathbf{Q}\mathbf{K}^{\boldsymbol{T}}}{\sqrt{d_{head}}})\mathbf{V}$$

where **Q**, **K**, and **V** have dimensions [100 × d_head] (100 patches, each with a d_head-dimensional vector), and the scaling factor $\sqrt{\text{d\_head}}$ prevents the dot products from growing too large as the embedding dimension increases, which would cause the softmax to collapse into a winner-take-all distribution and impede learning. Using multiple attention heads (8 in the full model, 4 in the reduced model) allows the model to learn several different types of spatial relationships simultaneously. For example, one head might attend to nearby patches while another attends to distant patches in a different part of the shelf. Each head operates on a d_head-dimensional subspace, and their outputs are concatenated back to 128 dimensions and projected through a final linear layer.

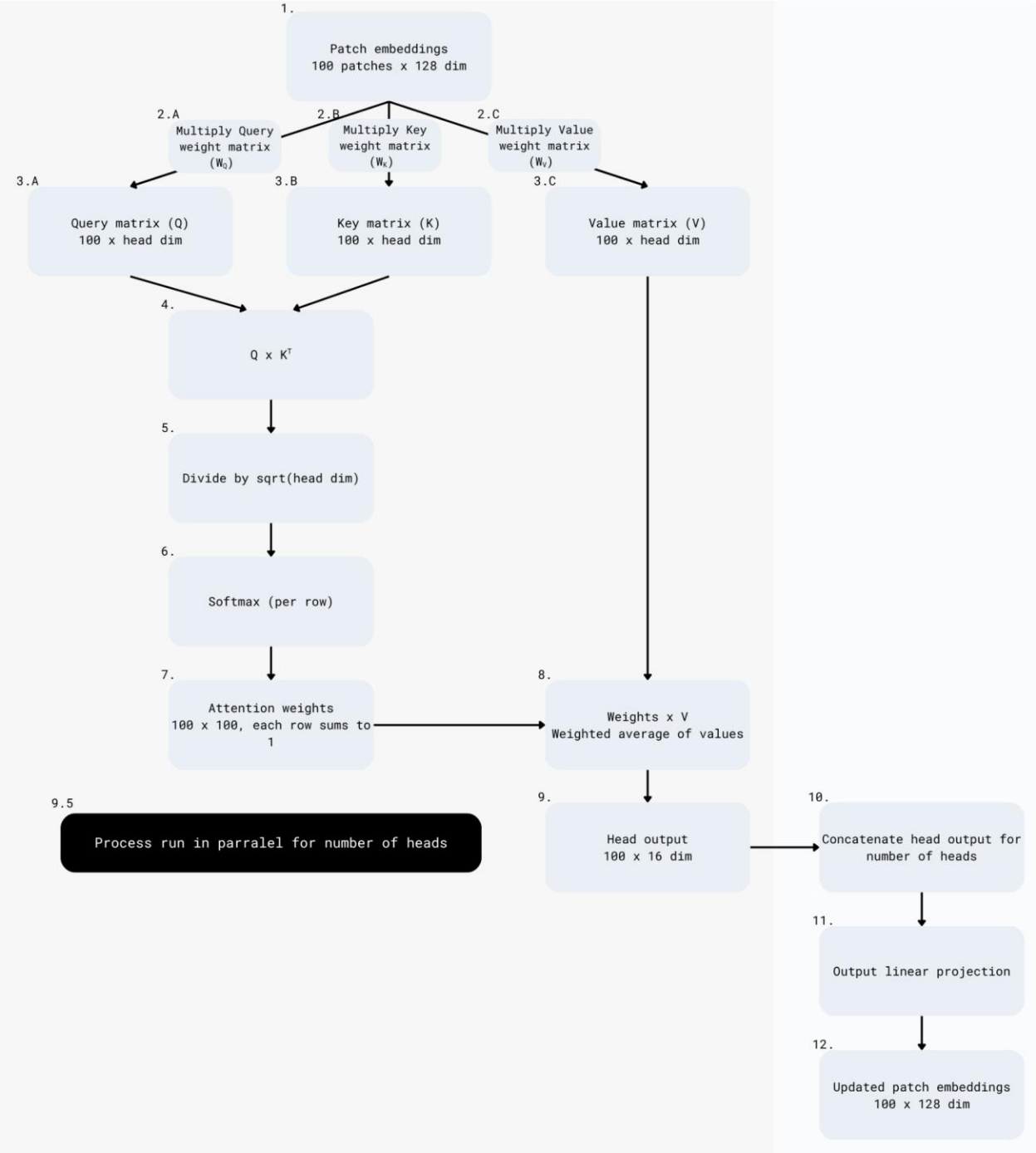


Figure S.2.2. Multi-head self-attention mechanism. (1) Input patch embeddings of dimension 100 × 128 are received from the patch embedding step. (2) Three learned weight matrices ($\mathbf{W_Q}$, $\mathbf{W_K}$, $\mathbf{W_V}$) are applied to produce (3) query, key, and value matrices, each of dimension 100 × head_dim, where head_dim = embedding_dim / num_heads. (4) The query matrix is multiplied by the transpose of the key matrix. (5) The result is divided by the square root of head_dim to stabilize gradients. (6) Softmax is applied row-wise to produce (7) an attention weight matrix of dimension 100 × 100, where each row sums to one and encodes the relative influence of every other patch on a given patch's representation. (8) The attention weights are multiplied by the value matrix to produce a weighted average of patch representations. (9) This constitutes the output of one attention head. The process is repeated for all attention

heads in parallel (9.5), and the outputs are (10) concatenated and (11) projected through a final linear layer to produce (12) updated patch embeddings of dimension 100 × 128.

S.2.3.4 Feed-forward network and transformer blocks

Following each self-attention operation, the updated patch embeddings are passed through a position-wise feed-forward network (FFN): a two-layer perceptron with GELU (Gaussian Error Linear Unit) activation. GELU is an activation function that, unlike the more commonly known ReLU (which simply sets negative values to zero), smoothly modulates inputs based on their magnitude (values that are more negative are scaled closer to zero, while positive values pass through approximately unchanged). This smooth gating generally improves training stability in transformer architectures (Hendrycks & Gimpel, 2016). Self-attention and the feed-forward network are combined into a Transformer Block using residual (skip) connections and Layer Normalization (pre-LN variant). The output of each transformer block is:

$$updated\ \mathbf{X} = \mathbf{X} + \text{Attention}(\text{LayerNorm}(\mathbf{X}))$$
$$updated\ \mathbf{X} = \mathbf{X} + \text{FFN}(\text{LayerNorm}(\mathbf{X}))$$

where $\mathbf{X}$ is the input to the block with dimensions [100 × 128] (100 patches, each with a 128-dimensional embedding). The addition operations (residual connections) ensure that each layer only needs to learn a refinement to its input rather than reconstructing the entire representation from scratch, which stabilizes training in deep networks.

The full model uses six transformer layers, which provides sufficient depth for the model to iteratively refine spatial representations across multiple rounds of attention. The reduced model uses three layers as a control for overfitting given the small effective sample size. The feed-forward dimension of 512 (4× the embedding dimension) follows the convention established in Vaswani et al. (2017), which found this expansion ratio to be effective across a range of tasks.

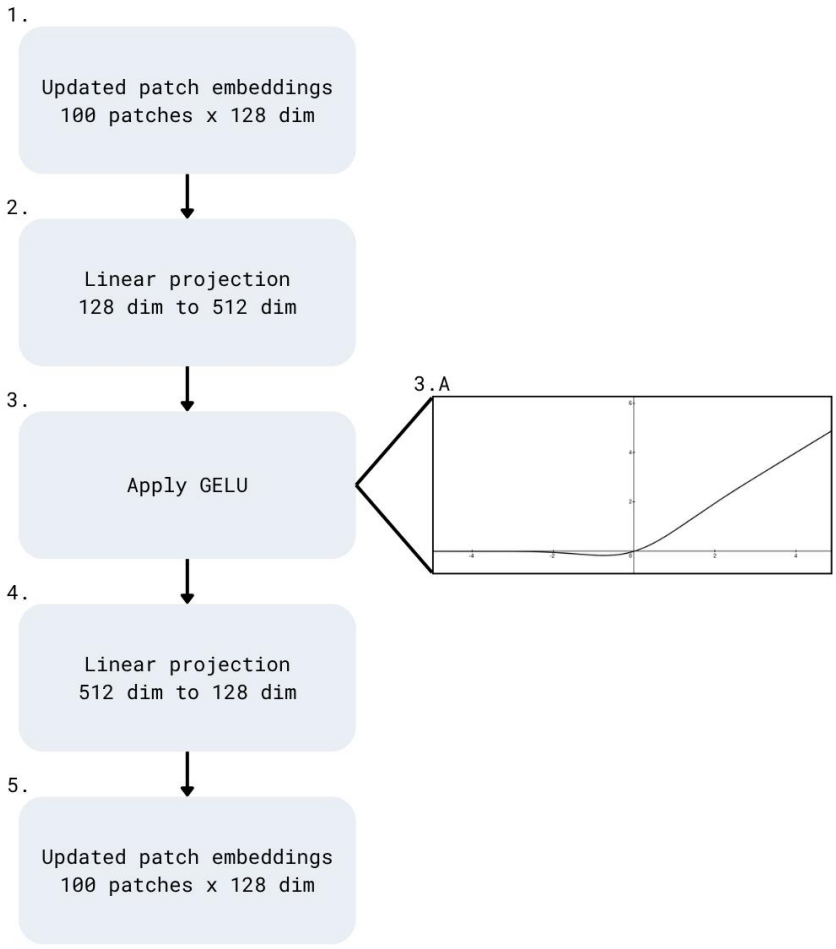


Figure S.2.3. Feed-forward network applied within each transformer layer. (1) Updated patch embeddings of dimension 100 × 128 are received from the self-attention step. (2) A linear projection expands the embedding dimension from 128 to 512. (3) A GELU nonlinear activation function is applied element-wise, introducing nonlinearity that allows the model to learn complex input-output mappings; the GELU function is shown in (3.a). (4) A second linear projection reduces the dimension from 512 back to 128. (5) The output is a set of updated patch embeddings of dimension 100 × 128, with richer representations than the input due to the nonlinear transformation.

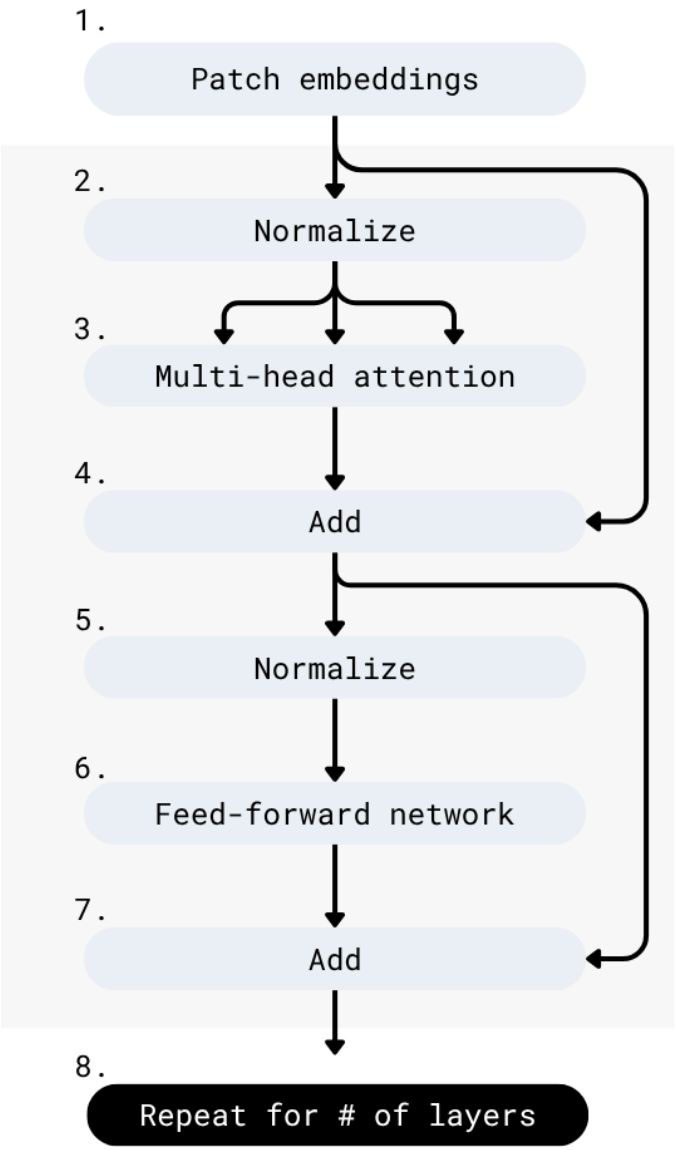


Figure S.2.4. Transformer layer structure. (1) Patch embeddings enter the transformer layer. (2) Layer normalization is applied. (3) Multi-head self-attention is computed on the normalized embeddings. (4) The attention output is added to the original patch embeddings via a residual connection. (5) Layer normalization is applied to the result. (6) The normalized representation is passed through the feed-forward network. (7) The feed-forward output is added to the result of step 4 via a second residual connection, producing the final output of the transformer layer. (8) This full process is repeated for each transformer layer in sequence, with the number of layers determined by the model configuration (6 layers for the base-size model, 3 for the reduced-size model).

S.2.3.5 Spatial decoder

After the transformer blocks, the 100 patch embeddings are reshaped into a 10 × 10 × 128 feature map and passed through a convolutional decoder that progressively upsamples to the full 50 × 50 output resolution. The decoder proceeds through three upsampling stages: (1) bilinear interpolation from 10 × 10 to 20 × 20, followed by a 3 × 3 convolution reducing channels from 128 to 64; (2) bilinear interpolation from 20 × 20 to 40 × 40, with a 3 × 3 convolution reducing from 64 to 32; (3) bilinear interpolation from 40 × 40 to 50 × 50, with a 3 × 3 convolution reducing from 32 to 16. A final 3 × 3 convolution maps from 16 channels to the single output channel, producing the

50 × 50 predicted recruit density field. Group Normalization (Wu & He, 2018) was used instead of the more common Batch Normalization because Batch Normalization's statistics become unstable with small batch sizes (this model's batch size is 8). Group Normalization computes normalization statistics within groups of channels for each sample independently, making it insensitive to batch size. Each convolutional layer uses reflection padding, which mirrors values at the edge of the feature map rather than padding with zeros. This avoids introducing artificial boundary artifacts near the coastline or edge of the survey area.

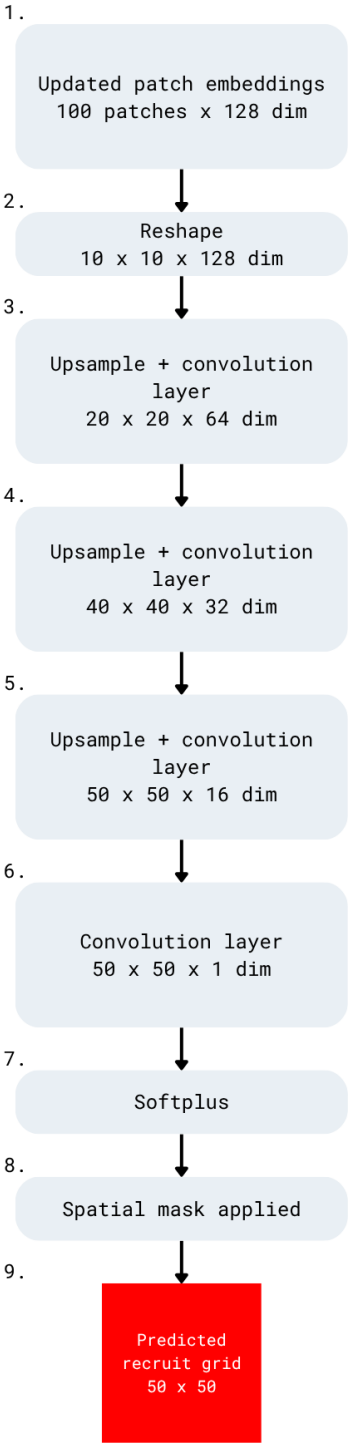


Figure S.2.5. Convolutional spatial decoder. (1) Patch embeddings of dimension 100 × 128 from the final transformer layer are received. (2) These are reshaped into a spatial tensor of dimension 10 × 10 × 128, recovering the two-dimensional patch grid. (3) Bilinear upsampling followed by a convolutional layer with GroupNorm and GELU activation produces a 20 × 20 × 64 feature map. (4) A second upsample and convolution step produces a 40 × 40 × 32 feature map. (5) A third upsample and convolution step produces a 50 × 50 × 16 feature map. (6) A final convolutional layer reduces the channel dimension to 1, yielding a 50 × 50 × 1 output. (7) A softplus activation enforces non-negativity of all predicted densities. (8) The binary spatial mask is applied, setting all out-of-region cells to zero. (9) The output is the predicted recruit density grid at 50 × 50 resolution.

S.2.3.6 Output Activation and Spatial Masking

Crab densities are non-negative by definition. To enforce this constraint, the decoder output is passed through a softplus activation function, $f(x) = log(1 + e^x)$ which is a smooth approximation to ReLU that maps the real line to strictly positive values without an upper bound. ReLU (Rectified Linear Unit) is the simplest non-negativity activation: it outputs zero for any negative input and passes positive inputs unchanged. However, ReLU has a discontinuous gradient at zero, which can cause training instability when many output values are near zero, as is common in the data for EBS snow crab where approximately 60% of grid cells have zero density. Softplus, defined as $softplus(x) = log(1 + e^x)$, is a smooth approximation that approaches zero for large negative inputs and approaches *x* for large positive inputs, providing well-defined gradients everywhere. This was particularly important given the high proportion of zero-density cells in the crab survey data. The spatial mask is applied as a final step, setting all out-of-region cells to exactly zero.

S.2.3.7 Weight initialization and bias warm-start

All linear layers were initialized using Xavier uniform initialization (Glorot & Bengio 2010), which scales initial weights to maintain approximately constant gradient variance through the forward and backward passes. Convolutional layers in the decoder were initialized with Kaiming (He) normal initialization (He et al. 2015), which compensates for the approximate halving of signal variance that occurs under ReLU-family activations (GELU in our case). Normalization layer scale and shift parameters were initialized to 1 and 0, corresponding to an identity transformation at the start of training.

The bias of the final decoder output convolution was initialized to the mean log-transformed recruit density over all valid (non-masked, non-2020) training pixels, preventing the model from predicting near-zero outputs early in training.

Without this initialization, the model would begin training by predicting values near zero everywhere, producing large initial gradients that can destabilize early training. By starting from the empirical mean, the model only needs to learn deviations from average conditions, which reduces the magnitude of early gradient updates and accelerates convergence.

S.2.3.8 Optimizer and learning rate schedule

Model parameters were updated using AdamW (Loshchilov & Hutter, 2017), which incorporates decoupled weight decay ($\lambda = 10^{-4}$) as a true L2 penalty on parameter magnitudes. AdamW was chosen over standard stochastic gradient descent (SGD) because adaptive learning rate methods generally converge faster on small datasets where each gradient estimate is noisy. Unlike the original Adam optimizer, AdamW applies weight decay directly to the parameters rather than incorporating it into the gradient, which provides more effective regularization, an important property given the small effective sample size. The base learning rate was set to 10^{-4}. The OneCycleLR scheduler (Smith and Topin 2019), which ramps the learning rate from an initial

value of ~1.2×10^{-5} to a maximum of 3×10^{-4} over the first 30% of training steps, then anneals to a final rate of 3×10^{-7} over the remaining 70% was applied. This warm-up and decay schedule helps the model establish productive gradient directions before making large update steps, then fine-tune near a minimum. Gradient clipping (max_norm = 1.0) was applied after each backward pass. Gradient clipping limits the magnitude of parameter updates at each training step. Without clipping, an unusually informative or noisy training batch can produce very large gradients that push parameters far from their current values, potentially destabilizing training. Stable training throughout all 20 epochs was ensured by capping the gradient norm at 1.0.

S.2.4 Channel sensitivity

A systematic channel sensitivity study varying which channel groups were provided to the model was conducted to evaluate which types of input information are most important for recruitment prediction. The three channel groups: spawners (current-year density plus 5-year history), recruits (5-year history only), and temperature (current-year plus 5-year history), were combined in all possible non-empty subsets, yielding seven channel configurations (Table S.2.2).

Configuration	Spawners	Recruits	Temperature	Channels (same-year)	Channels (one-year-ahead)	Channels (5-year lag)
All	✓	✓	✓	17	15	17
Spawners + Recruits	✓	✓	–	11	10	11
Spawners + Temperature	✓	–	✓	12	10	12
Recruits + Temperature	–	✓	✓	11	10	11
Spawners only	✓	–	–	6	5	6
Recruits only	–	✓	–	5	5	5
Temperature only	–	–	✓	6	5	6

Table S.2.2. Channel group configurations evaluated in the sensitivity analysis. Seven non-empty subsets of the three channel groups (spawners, recruits, temperature) were evaluated. The number of input channels varies by configuration and prediction mode because current-year spawner and temperature channels are excluded in one-year-ahead mode, reducing the channel count by two relative to now-cast and lag-5 modes.

Note that the recruit channel group contains no current-year channel by design; recruits at year *t* are the prediction target, so they cannot also serve as input. This means the recruits-only configuration is inherently identical across same-year and one-year-ahead modes (5 channels in both cases).

Each of the 7 channel configurations was crossed with 2 model sizes (full and reduced), 3 prediction modes (same-year, one-year-ahead, and lag-5), and 2 loss functions (MSE and Tweedie), yielding a total of 84 configurations per data type. For the actual EBS data, temperature

channels were available, giving the full 84 configurations. For the three synthetic datasets, temperature data were not generated, so only the 4 configurations that do not include temperature (all spawner/recruit combinations) were evaluated, crossed with the same model sizes, prediction modes, and loss functions, yielding 48 configurations per synthetic dataset. The complete configuration matrix across all data types contained 228 individual training runs.

The model architecture automatically adapts to the number of input channels: the patch embedding layer creates one independent linear projection per channel, and the channel-to-mask index mapping is recomputed for each configuration (see Section 2.2.6 and S.2.3.1). No other architectural changes were required across configurations.

Chapter 2 References

Cadrin, S. X. (2020). Defining spatial structure for fishery stock assessment. Fisheries Research, 221, 105397. https://doi.org/10.1016/j.fishres.2019.105397

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. Methods in Ecology and Evolution, 10, 1632–1644. https://doi.org/10.1111/2041-210X.13256

Cressie, N., & Wikle, C. K. (2011). Statistics for spatio-temporal data. Wiley.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Fedewa, E. J., Jackson, T. M., Richar, J. I., Gardner, J. L., & Litzow, M. A. (2020). Recent shifts in northern Bering Sea snow crab (Chionoecetes opilio) size structure and the potential role of climate-mediated range contraction. Deep-Sea Research Part II: Topical Studies in Oceanography, 181–182, 104878. https://doi.org/10.1016/j.dsr2.2020.104878

Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y., Li, M., & Yeung, D. Y. (2022). Earthformer: Exploring space-time transformers for earth system forecasting. Advances in Neural Information Processing Systems, 35, 25390–25403.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 9, 249–256.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778. https://doi.org/10.1109/CVPR.2016.90

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kapur, M. S., Adams, G., Lapeyrolerie, M., & Thorson, J. T. (2026). AI for fisheries science: Neural network tools for forecasting, spatial standardization, and policy optimization. bioRxiv. https://doi.org/10.64898/2026.03.13.711664

Lindgren, F. (2026). fmesher: Triangle meshes and related geometry tools (R package version 0.7.0). https://github.com/inlabru-org/fmesher

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B, 73, 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. Advances in Neural Information Processing Systems, 29.

Markowitz, E. H., Dawson, E. J., Wassermann, S., Rohan, S. K., Charriere, N. E., & Stevenson, D. E. (2024). Results of the 2022 eastern and northern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate fauna. NOAA. https://repository.library.noaa.gov/view/noaa/50154

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. Proceedings of the 40th International Conference on Machine Learning.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences, 115(25), E5716–E5725. https://doi.org/10.1073/pnas.1719367115

O'Hara, R.B. and Kotze, D.J. (2010), Do not log-transform count data. Methods in Ecology and Evolution, 1: 118-122. https://doi.org/10.1111/j.2041-210X.2010.00021.x

give me the citation for pytorch

8:21 PM

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 8026–8037.

Ricker, W. E. (1954). Stock and recruitment. Journal of the Fisheries Research Board of Canada, 11(5), 559–623.
Rohan, S. K., Barnett, L. A. K., & Charriere, N. (2022). Evaluating approaches to estimating mean temperatures and cold pool area from AFSC bottom trawl surveys of the eastern Bering Sea (NOAA Technical Memorandum NMFS-AFSC-456). https://doi.org/10.25923/1wwh-q418
Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B, 71, 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x
Sainte-Marie, B., Raymond, S., & Brêthes, J.-C. (1995). Growth and maturation of the benthic stages of male snow crab, Chionoecetes opilio (Brachyura: Majidae). Canadian Journal of Fisheries and Aquatic Sciences, 52(5), 903–924. https://doi.org/10.1139/f95-091
Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems, 28, 802–810.
Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. Proceedings of SPIE 11006, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, 1100612. https://doi.org/10.1117/12.2520589
Shono, H. (2008). Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research*, 93, 154–162.
Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, 70, 3319–3328.
Szuwalski, C. S., Alglave, B., Balstad, L. J., Olmos, M., Punt, A. E., Richar, J., Stockhausen, W., & Veron, M. (2026). Density dependence modulates climate change impacts on eastern Bering Sea snow crab. Journal of Applied Ecology, 63, e70316. https://doi.org/10.1111/1365-2664.70316
Szuwalski, C. S., & Punt, A. E. (2013). Regime shifts and recruitment dynamics of snow crab, Chionoecetes opilio, in the eastern Bering Sea. Fisheries Oceanography, 22(5), 345–354. https://doi.org/10.1111/fog.12026
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
Wu, Y., & He, K. (2018). Group normalization. Proceedings of the European Conference on Computer Vision, 3–19. https://doi.org/10.1007/978-3-030-01261-8_1

3.Discussion of Chapter 1 and 2

Taken together, the two chapters of this thesis represent a systematic investigation of the spatiotemporal teleconnection between snow crab spawners and recruits in the EBS, conducted across a range of model complexity, spatial resolution, and statistical approach. Chapter 1 explored whether dimensionality reduction methods grounded in classical statistics (EOFs, CF-GLLVMs, and spectral clustering) could extract a predictive teleconnection signal from station-level survey data, and Chapter 2 explored whether a modern deep learning architecture (ViT) could recover the same signal from spatially interpolated gridded fields, and whether its learned structure was biologically coherent. Both chapters reached a consistent answer: a meaningful spatial teleconnection exists, it is robust across the period of unprecedented population decline spanning 2021-2023, and it is detectable by methods ranging from algebraic matrix decomposition to multi-head self-attention.

A direct quantitative comparison of predictive performance across Chapters 1 and 2 is not possible, and attempting one would be misleading. The methods evaluated in Chapter 1 produce predictions at 349 discrete survey station locations, while those in Chapter 2 operate on a 50 × 50 km regular grid generated by SPDE interpolation of the same underlying station data. Not only are these fundamentally different prediction targets, but the data used to train the models are fundamentally different too. The SPDE interpolation process applied in Chapter 2 imposes spatial smoothing on station-level observations, producing gridded fields that are more spatially consistent across years than the raw data would suggest. This smoothing inflates the apparent predictive skill of even a simple climatological mean: the grid-level baseline Spearman's rank correlation of 0.807 in Chapter 2 arises in part because the interpolated fields share a stable spatial structure by construction, whereas the station-level baseline correlation of 0.696 in Chapter 1 reflects raw inter-annual variability without any imposed smoothness. Differences in reported Spearman's correlations between chapters therefore reflect differences in data representation as much as differences in model performance.

A further consequence of the SPDE interpolation is that improvements in Spearman's correlation over the baseline are a weaker discriminator of model quality in Chapter 2 than in Chapter 1. Mean absolute error (MAE), which measures prediction accuracy in density units and is not inflated by smoothing in the same way, provides more meaningful discrimination in Chapter 2: the best-performing ViT configurations reduced MAE by 33-38% relative to the climatological baseline, from 500 crabs km^{-2} to 309-332 crabs km^{-2}. This distinction between spatial pattern metrics and magnitude metrics applies generally: researchers applying these methods in other systems should be cautious about interpreting high Spearman's correlations as evidence of strong predictive teleconnection signal when the underlying data have been spatially smoothed.

All methods in this thesis face the same fundamental constraint: the EBS snow crab survey provides approximately 35 temporally independent spatial observations, while each observation consists of density measurements at 349 stations. This high-dimensionality constraint imposes difficulties in parameter estimation, as the degrees of freedom are severely limited. Each approach in this thesis addresses this constraint through a different form of dimensionality reduction, and the strengths and limitations of each approach are largely determined by the assumptions embedded in that reduction.

3.1 Strengths and weaknesses of the methods

3.1.1 The EOF-GAM pipeline

The EOF-GAM pipeline compresses spatial dimensionality algebraically, representing each year's spatial field as a linear combination of five orthogonal spatial patterns. The GAM then operates on these five temporal coefficients per dataset, reducing the problem to a ten-dimensional input (5 spawner factors and 5 recruit factors) rather than a 349-dimensional one. This is computationally efficient and requires no distributional assumptions, but it imposes orthogonality on the spatial patterns and requires complete data: missing station values must be imputed before singular value decomposition can be applied, introducing a potential source of bias in years with systematic coverage gaps.

The primary strengths of the EOF-GAM pipeline are its computational efficiency, conceptual transparency, and strong performance on hotspot identification: it achieved the highest top-10% Jaccard score (0.409) across all methods in Chapter 1, making it the most reliable tool for identifying high-recruitment areas even when absolute magnitude predictions are less precise. The method's interpretability is also an asset in management contexts. The spatial patterns extracted by single value decomposition can be visualized and communicated directly, and the GAM relationships between spawner and recruit factors can be inspected directly for biological interpretation.

The method's core weakness is that the EOF basis is fixed at training time, so the spatial modes are implicitly tuned to the dominant patterns of that era. If a regime shift alters the spatial structure of the spawner field between training and test years, the retained modes may no longer capture ecologically meaningful variation, and the GAM relationships built on those modes inherit the same misspecification. This exposure is shared to some degree by all Chapter 1 methods, but it is most acute for the EOF-GAM because the basis cannot adapt: unlike the CF-GLLVM, whose spatial factors are estimated within a probabilistic model that can at least quantify uncertainty about the dominant patterns, the EOF basis is a fixed algebraic solution with no analog to posterior updating. The failure is also hard to detect, since the pipeline can produce smooth, plausible predictions even when the underlying basis is no longer appropriate. Mean imputation for missing stations and the inability to propagate uncertainty through the full pipeline compound this fragility under non-stationary conditions.

3.1.2 The CF-GLLVM framework

The CF-GLLVM framework allows for more flexibility than the EOF-GAM pipeline by pairing dimensionality reduction within a formal statistical model. Spatial factors are estimated jointly with their uncertainty using a Matérn covariance structure implemented via the SPDE approximation, which handles missing data through the spatial correlation structure rather than requiring imputation. Fitting the full sensitivity analysis of 108 CF-GLLVM configurations required significant computational resources, and the model is sensitive to structural choices that are not always straightforward to make in advance. This is particularly the case for the spatiotemporal residual field and its interaction with the low-rank factor decomposition.

The CF-GLLVM framework achieved the highest Spearman's correlations (Linked: $\rho = 0.728$; SVC: $\rho = 0.733$) and F1 scores (Linked: 0.782; SVC: 0.767) of any method in Chapter 1, and provides the most complete statistical modeling of the spawner-recruit relationship. The spatially varying coefficient (SVC) formulation is particularly noteworthy: by allowing the strength of the spawner-recruit linkage to vary continuously across the EBS shelf, it can represent geographic heterogeneity in the teleconnection that neither the EOF-GAM nor spectral clustering can capture. However, the CF-GLLVM's sensitivity analysis revealed an important and counterintuitive

finding: when more complex spawner configurations (those incorporating spatiotemporal residual fields) were used as inputs, they consistently degraded out-of-sample performance, likely because the spatiotemporal surface absorbed predictive signal into in-sample residuals rather than propagating it through the teleconnection mechanism. This is a cautionary result for applications of GLLVM-based approaches in other systems: structural flexibility does not guarantee better prediction, and careful cross-validation of configuration choices is essential.

3.1.3 Spectral clustering

Spectral clustering takes a different approach: rather than compressing a continuous spatial field, it partitions the survey domain into a small number of discrete spatial regions. This reduces the prediction problem to a much smaller number of cluster-level time series, which GAMs can then connect across the spawner-recruit lag. The approach is interpretable and computationally light, and the resulting cluster boundaries may correspond to ecologically meaningful spatial units. However, the discretization loses within-cluster spatial heterogeneity, and the cluster assignments themselves are sensitive to hyperparameter choices that must be tuned from the same limited data used for prediction.

Spectral clustering performed at or below the station-mean climatology across all metrics in Chapter 1. Although the method identified spatially coherent regions with ecologically interpretable boundaries, the GAMs fitted to cluster-level time series did not extract sufficient predictive signal from the spawner-recruit relationship to outperform a simple historical mean. This result may reflect the loss of within-cluster spatial heterogeneity or it may indicate that the relationship between spawner and recruit cluster time series is too weak or noisy to be reliably identified from the available training data. Spectral clustering is therefore most appropriately used as an exploratory tool for understanding spatial structure and communicating results to non-specialist audiences, rather than as an operational forecasting method for this system.

3.1.5 The Vision Transformer

The Vision Transformer addresses the dimensionality problem through a combination of spatial aggregation, bootstrap data augmentation, and architectural regularizations. The 50 × 50 grid is divided into 5 × 5 patches, reducing the sequence length for attention computation to 100 patch embeddings rather than 2,500 individual cells. Bootstrap resampling of survey stations inflates the nominal training set from approximately 24 to 2,400 samples, but the 100 replicates for any given year are highly correlated (mean pairwise Pearson $r \approx 0.99$ for both spawners and recruits), so the effective number of statistically independent patterns remains approximately equal to the number of survey years. Choosing the model size to complement the low effective sample size was essential for obtaining generalizable predictions, particularly in harder prediction scenarios where the signal-to-noise ratio was lower.

The Vision Transformer demonstrated strong spatial pattern prediction and MAE on held-out test data from 2021-2023 (0.83-0.87 for Spearman's correlation; 309-332 crabs km^{-2} for MAE). Its self-attention mechanism was able to discover long-range spatial dependencies without prior specification, and the one-year-ahead forecasting mode achieved performance in line with the now-cast models, suggesting the model learned genuine multi-year temporal dynamics rather than propagating current-year spatial structure forward. The temperature-only configuration achieved the third-highest composite score despite using no crab density inputs, confirming that bottom temperature structure alone carries substantial spatial information about recruitment intensity. The ViT's principal limitations are its opacity without post-hoc attribution analysis, the risk of

overfitting to training data magnitude distributions, and the requirement for careful architectural tuning that may not be straightforward for practitioners without deep learning expertise.

3.2 Selecting a best method for use in ecological teleconnection studies

The methodological framework developed in this thesis is not specific to snow crab. Any ecological system where a leading population at one location or time influences a response population elsewhere or later, and where spatially explicit observational data are available over a multi-decadal monitoring period, can use the approaches described here. Candidate applications include other species with spatially structured spawner-recruit relationships and documented larval dispersal mechanisms, such as Pacific halibut (*Hippoglossus stenolepis*), Atlantic cod (*Gadus morhua*), or Dungeness crab (*Metacarcinus magister*); environmental-biological teleconnections in which remotely sensed abiotic conditions (sea surface temperature, chlorophyll, sea ice extent) predict biological responses at distant locations or delayed times; and terrestrial systems where spatially explicit monitoring data are available, such as bird breeding atlases, vegetation cover surveys, or remote sensing time series of habitat quality (e.g., Boudreau et al. 2017, Planque and Frédou 1999, Shanks and Roegner 2007).

The choice among methods ultimately reflects a tradeoff between statistical rigor, computational cost, and robustness to non-stationarity that no single method resolves cleanly. The appropriate choice depends on the primary prediction goal, the computational resources available, and the degree to which formal uncertainty quantification is required for decision-making. As a general guide: the EOF-GAM is most appropriate when computational constraints are binding, when hotspot identification is the primary management objective, or when a transparent, reproducible baseline is needed against which more complex methods can be benchmarked. The CF-GLLVM framework should be preferred when formal uncertainty quantification is required for management advice, when the spatial structure of the teleconnection is of scientific interest in its own right, or when the analyst has the computational resources and statistical expertise to tune the model configuration carefully. Spectral clustering is best reserved for exploratory analyses rather than operational recruitment forecasting. The Vision Transformer is most appropriate when the spatial domain is large and the relevant teleconnection structure is unknown in advance, when a pre-season forecast before survey completion is operationally useful, or when the analyst wishes to generate biological hypotheses through attribution analysis.

The key adaptations required to transfer these methods to a new system are largely data-structural: defining the leading and response populations or variables, specifying the expected temporal lag based on known biology or empirical exploration, and ensuring that the spatial resolution of the prediction grid is matched to the ecological scale of the teleconnection. The statistical methods in Chapter 1 require only that the data can be represented as a station-by-year matrix with consistent spatial coverage across time. The ViT in Chapter 2 requires an additional interpolation step to convert station data to a regular grid, which can be accomplished using the same SPDE approach described here or by other spatial interpolation methods depending on the density and configuration of the monitoring network.

3.3 Conclusions

This thesis demonstrates that the spatial teleconnection between snow crab spawners and recruits in the EBS is real, detectable by a range of statistical and machine learning methods, and robust across a period of unprecedented population change. The biological teleconnection studied here is spatially structured, where certain configurations of spawner distribution are more predictive of

recruit distribution than others, and the ViT found a temperature signal consistent with the role of the cold pool as a determinant of juvenile habitat suitability and a mediator of the spawner-recruit relationship. No single method dominates across all prediction objectives: the EOF-GAM offers the best hotspot identification with minimal computational cost; the CF-GLLVM provides the most complete uncertainty quantification and the highest correlation-based predictive skill; spectral clustering is most useful for exploratory spatial analyses; and the Vision Transformer achieves the strongest overall spatial pattern recovery while generating biologically interpretable attribution structures through Integrated Gradients analysis. Together, these methods provide a practical, scalable, and transferable framework for detecting and quantifying biological teleconnections in other species and ecosystems where spatially structured leading populations regulate responses in recipient populations.

Chapter 3 Discussion References

Boudreau, S. A., Shackell, N. L., Carson, S., & den Heyer, C. E. (2017). Connectivity, persistence, and loss of high abundance areas of a recovering marine fish population in the Northwest Atlantic Ocean. Ecology and Evolution, 7, 9739–9749. https://doi.org/10.1002/ece3.3495

Planque, B., & Frédou, T. (1999). Temperature and the recruitment of Atlantic cod (Gadus morhua). Canadian Journal of Fisheries and Aquatic Sciences, 56(11), 2069–2077. https://doi.org/10.1139/f99-114

Shanks, A. L., & Roegner, G. C. (2007). Recruitment limitation in Dungeness crab populations is driven by variation in atmospheric forcing. Ecology, 88, 1726–1737. https://doi.org/10.1890/06-1003.1